Four Facets of Forecast Felicity

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Rabanus Derr

aus Ulm

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	06.05.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Robert C. Williamson
2. Berichterstatterin:	Prof. Dr. Rediet Abebe
3. Berichterstatter:	Prof. Dr. Peter Grünwald

Thesis advisor: Prof. Dr. Robert C. Williamson Rabanus Derr

Four Facets of Forecast Felicity

Abstract

Forecasts are a central goal of machine learning. The current relevance of machine learning in technological development and society motivates a renewed examination of fundamental questions: What is the meaning or "goodness" of a prediction? What warrants the use of a quantified and/or statistical prediction? I argue that forecasts are speech acts, which as such *do* something. For example, a weather forecast *warns* of extreme precipitation. Accordingly, forecasts should be measured by how *felicitous* they are, i.e., whether they are successful in what they do. In this thesis, I examine *felicity conditions* for statistical forecasts in relation to the aforementioned questions. I will focus on two, possibly overlapping, types of prevalent statistical forecasts. The first type is based on modeling an aggregate which is relevant to the subject of the forecast. I show that notions of measurability and randomness are not universally given model components. Instead, these aspects are context-dependent modeling choices that determine the success of the prediction. The second type of forecast is based on the individual assignment of numerical values that are evaluated on an aggregate. I investigate the landscape of evaluation metrics with a focus on calibration and its relationship to regret. I demonstrate that there are two accounts of calibration currently conflated in literature. Furthermore, most evaluation metrics used in machine learning, including certain types of calibration and regret, can be expressed as the capital of a gambler betting against the forecaster. Beyond the aforementioned facets of *felicity conditions*, I point out commonly overlooked, open questions, particularly with regard to randomness, individuality, performativity, and communication of forecasts. A tabular questionnaire designed to help understand, create, and analyze *(in)felicitous* forecasts on a practical level concludes the work.

Betreuer: Prof. Dr. Robert C. Williamson Rabanus Derr

Four Facets of Forecast Felicity

Zusammenfassung

Vorhersagen sind ein zentrales Ziel des maschinellen Lernens. Die aktuell beobachtbare Relevanz des maschinellen Lernens in technologischer Entwicklung und Gesellschaft motiviert die erneute Untersuchung grundlegender Fragen: Was ist die Bedeutung oder Güte einer Vorhersage? Was rechtfertigt die Verwendung einer quantifizierten und/oder statistischen Vorhersage? Ich argumentiere, dass Vorhersagen Sprechakte sind, welche als solche etwas *tun*. Z. B. *warnt* eine Wettervorhersage vor extremen Niederschlägen. Dementsprechend sollten Vorhersagen daran gemessen werden wie *geglückt* sie sind, d.h. wie erfolgreich sie in ihrem Tun sind. In der vorliegenden Arbeit möchte ich *Glückskriterien* für statistische Vorhersagen in Verbindung mit den vorher genannten Fragen untersuchen. Hierzu, werde ich eine Unterscheidung zwischen zwei Arten von gängigen statistischen Vorhersagen vornehmen. Die erste Arte basiert auf der Modellierung einer Gesamtheit welche für das Subjekt der Vorhersagen von Relevanz ist. Ich zeige, dass Messbarkeit und Zufälligkeit nicht universell gegebene Modellbestandteile sind. Stattdessen sind diese Aspekte kontextbezogene Modellierungsentscheidungen, welche den Erfolg der Vorhersage bestimmen. Die zweite Art basiert auf der individuellen Zuweisung numerischer Werte, die auf einer Gesamheit bewertet werden. Ich untersuche Bewertungsmetriken mit einem Schwerpunkt auf Kalibrierung und deren Beziehung zu Regret. Ich zeige, dass es derzeit zwei Semantiken für Kalibrierung gibt, die es zu trennen gilt. Darüber hinaus lassen sich die meisten Bewertungsmetriken, die im maschinellen Lernen verwendet werden, einschließlich bestimmter Arten der Kalibrierung und des Regret, als Kapital eines Spielers ausdrücken, der gegen den Vorhersager wettet. Über die genannten Facetten von *Glückskriterien* hinaus verweise ich auf oft übersehene, offene Fragen insbesondere im Bezug auf Zufälligkeit, Individualität, Performativität und Kommunikation von Vorhersagen. Ein tabellarischer Fragebogen, der auf praktischer Ebene dabei helfen soll (nicht) *geglückte* Vorhersagen zu verstehen, zu erstellen und zu analysieren, rundet die Arbeit ab.

Content

Acknowledgments

There are about two reasons why I started my doctoral studies. I could continue studying and learning and get paid for it. Plus, there was the pragmatic option as Bob joined the University of Tübingen and Ulrike made me aware of this. After four years, the journey I started 2021 seems to come to an end.
This document tries to summarize in a concise format what filled the time in between. However, it mainly captures the academic part of it. It contains the miles walked during this journey.
Apparently, Tim Cahill wrote at some point "A journey is best measured in friends, not in miles".[1] Hence, I spent at least this single page[2] on the non-academic side of doing a doctoral study, "the friends". In particular, I want to say:

Thank you!

To my colleagues, Armando, Ben, Chris (!), Laura, Nan, Raphael, Wenkai, to my supervisor Bob, to Aaron and his group, Natalie, Ira, Yahav, Mirah, Varun, Sikata, Jessie and the surrounding Uruguayan and Persian gang with Turkish member, to my "academic" friends James, Julian, Min Jae, Rajeev, Jessie, to my friends in Tübingen and family around, and all the persons I certainly forgot.
A special **thanks** I'd like to direct to my supervisor Bob. I was very happy to learn and become a "Mini-Bob", as Samira once called me.
A second special **thanks** is addressed to Karina and all the penguins of the world. There is no water too cold or deep when you have a colony along with you.

[1] I have never read any book by Tim Cahill. But ChatGPT is sometimes good in finding quotes. I was actually searching for a different proverb whose author I didn't remember.

[2] According to Laura, I could be a little more sparing with my "thank you"s. So I'll keep it short.

Part 0

Prologue

0

The origin?

It was a hot summer afternoon when Penelope set off for the Lyceum. Since early morning, the heat from the sea had permeated the streets of Athens, enveloping the city as if under a bell jar. Aristotle had instructed Penelope to find Tyrtamus.

Penelope had asked her uncle if he conjectured it would rain tomorrow. Aristotle, knowing his young niece's thirst for knowledge but also her shrewdness, had not wanted to answer. He suspected one of her unusual ideas, which always surprised him, and did not want to be lured into a trap with this question. Penelope surely already had an answer ready. Furthermore, Tyrtamus had studied the signs of the weather more extensively than Aristotle.

"Tell me, Tyrtamus, will it rain tomorrow?" Penelope asked when she finally found her uncle's friend in the temple. "Let me study the signs," replied Tyrtamus, glancing out of a side door. "The plainest sign is that which is to be observed in the morning, when, before the sun rises, the sky appears reddened over; and it indicates rain, either on the same day, or generally within three days," he quoted from the great work he had contributed to, $\pi\epsilon\rho\iota\ \sigma\eta\mu\epsilon\iota\omega\nu$ (On Signs). "Furthermore, moisture harbors the seeds of rain," he continued, looking at the alleys, "It will rain. But you have asked me the question for another reason." " In the last year, I counted four days with rain and nine without rain which followed a day like today," Penelope replied without explaining herself first. "It will rain tomorrow if it is one of the four days." Barely surprised by the girl's unusual counting activity, Tyrtamus replied, "But child, where is the logic in your argument? Where is the truth?" "What do I care about the truth when the fisherman who fears the storm stays in the harbor and the farmer sows his seeds in the hope that the rain will make them grow?" [1]

[1] Tyrtamus lived at the same time as Aristotle. He is better known as Theophrastus, a close colleague and successor of Aristotle at the Peripatetic school of philosophy [Ierodiakonou, 2020]. Theophrastus is usually considered the author of "On Signs," but there is some controversy about this [Cronin, 1992]. "On Signs" is considered one of the first meteorological studies, probably intended as a practical guide to weather forecasting [Beardmore, 2013]. Penelope's statistical thinking must have been revolutionary for its time.

1

Motivation

1.1 Forecasts Are Everywhere

Forecasts are everywhere. Imagine your alarm driving you out of your cozy bed. You prepare your cereal, open the newspaper – well, suppose you are as old-fashioned as I am in this respect – and you'll find on a central page the weather *forecast*. Before you walk to the station to take the subway to work, you check the transporter's app, which already shows you how crowded the cars are *expected* to be [Deutsche Bahn AG, 2025]. While you are waiting for the subway, an announcement is made that the tracks on which the subway will arrive will change due to "*predictive* maintenance" [Carvalho et al., 2019]. Let's stop this example here. It is easy to imagine that our daily life is sprinkled with even more forecasts.

Historically, people have looked to the sky to see signs of the weather to come, at least since ancient times [Beardmore, 2013]. Hence, arguably forecasting has been a central element of human society for some time. Within the sciences, "forecasting" is a central goal that is usually complemented by "explaining" [Carnap, 1986]. The importance of "forecasting" in the sciences can be further substantiated by the existence of an "International Journal of Forecasting" [International Institute of Forecasters, 2025], respectively the role of "instrumentalism" as a strand of philosophical thoughts which roughly speaking puts "forecasts" to the center stage of science while denying their metaphysical content [Chakravartty, 2017].

The current wave of machine learning methods further amplifies the sections of life and research in which forecasts play a role (for a summary of use cases see for instance [Petropoulos et al., 2022]).[1] Large parts of machine learning are devoted to producing forecasts:[2] clas-

[1] At the Philosophy of Science Meets Machine Learning conference in Tübingen 2022, Jan-Willem Romeijn entitled his talk "Machine Learning, or: The Return of Instrumentalism".

[2] Forecasting, as Leo Breiman claims, is one of two central goals of statistics. Breiman classifies machine learning, using other terms, as statistics [Breiman, 2001].

sification, regression, online learning [Cesa-Bianchi and Lugosi, 2006, Shalev-Shwartz and Ben-David, 2014] and many more.

Even during the current advent of "generative models"[3] , such as generative pre-trained transformers (GPT) [Radford et al., 2018] or diffusion models [Sohl-Dickstein et al., 2015], forecasting remains at the core of machine learning.

First, several of those "generative methods" actually produce a probabilistic forecast before sampling from it, e.g., language generation models sequentially predict the next token. Second, samples are less informative than probabilistic forecasts, hence for certain purposes, e.g., forecasts on unemployment rate will hardly be replaced by sampling steps. Third, a series of samples can be used to reverse-engineer a probabilistic forecast, e.g., this can be helpful for methods such as diffusion models which do not directly provide probabilistic forecasts. In conclusion, forecasting remains highly significant as one of the goals of machine learning.

I hope I have now convinced the reader that forecasts are ubiquitous, central to current scientific practice and further (will be) promoted by the increasing use of machine learning methods. Hence, what is the problem with those forecasts as otherwise I would not have written this thesis about them?

1.2 Four Problems with Forecasts

Let us wind back to our imaginary self waking up and reading a weather forecast in the newspaper. We read on page 14, "Chance of rain today: 60%.". Ignoring how the forecast has been constructed, we might readily ask ourselves two questions:

What does the forecast mean? The formulation in terms of a percentage opens up the space for interpretation. Is the 60% here expressing a relative frequency of an event? Is it a subjective belief of a meteorologist expressed in numbers? I call this problem the *interpretation problem*.

What makes the forecast "good"? The fact that the forecast has been published in a newspaper gives us reason to belief that the forecast is not simply made up. But, how do we know it is "good"? In which way is the forecast "good" and is this "good" relevant for us? I coin this the *evaluation problem*.

Less obvious, and potentially not on the first thought, an awake reader – not me in the

[3]Note that "generative models" here do *not* refer to what has been classically considered a generative statistical model, i.e., a model in which probability distributions over parameters interest are inferred [MacKay, 2003, Bishop, 2006]. Instead, the term "generative model" here simply refers to a method which produces examples (apparently) sampled from a desired source distribution.

morning – might even go beyond that and wonder about the following even more fundamental questions:

Is a quantified forecast warranted? The forecast is given in form of a number. To be able to provide a number, it was necessary to quantify the event "rain today". Quantification presupposes measurability. Is the event measurable? In what sense is it measurable? This is what the *measurability problem* is about.

Is a statistical forecast warranted? [4] The fact that the rain forecast has been expressed in terms of 60% chance suggests that the forecaster had to deal with randomness (cf. [Eagle, 2021]). The attribute "rain" for today seems to be taken as being random in some sense. Is it justified to treat the attribute as random? How does the "statisticalness" of the forecast relate to the presumed randomness of the attribute? This is the *statisticalness problem*.

First, note that it was not necessary to refer to machine learning to ask those questions. Nevertheless, machine learning motivated the study of those questions, as it makes forecasts more prevalent in different and novel contexts. Second, there may be further questions an even more awake reader might ask. The list is not exhaustive, but it captures the central problems which motivate the studies which follow. Third, there is another problem I have not yet mentioned which will motivate the type of glasses I will make us wear within this work.

1.3 An Additional Problem – The Truth Fetish

What is the purpose of the forecast? Is 60% the true chance or probability that it rains today? Are the forecasts meant to resemble the truth?

I already nudged the question towards whether the purpose of the forecasts is "truth". But is "truth" a universal purpose? There are at least two superficial ways to understand "truth" in the context of forecasting. Truth is referring to an unobserved state of nature which materializes in an attribute or outcome. In that, truth is comparable to a disposition, a propensity of nature. It is the underlying generative process, e.g., described by a probability distribution. Or, truth is the actual attribute or outcome which realized, e.g., "it rains today". Without starting a debate on the metaphysical nature of truth, I question that truth, in either way, is the purpose of many forecasts.

In a social context forecasts surely influence the subject they are about. Think of the example of an app which provides forecasts on the degree of utilization of trains. In this

[4] In Section 3.2, I precisely define "statistical forecast".

case, the "true" forecast becomes contested. It is part of normative considerations on "how full *should* the train be". In particular, in these scenarios forecasts are often (in)directly linked to interventions. Hence, informing or defining the intervention is part of the purpose and the intention of the forecast. This means that the involved translation from forecast to action makes the "truth" of the forecast a secondary priority, irrelevant or ill-defined, e.g., in education [Liu et al., 2022] or criminal justice [Harcourt, 2007]. Finally, note that even beyond social scenarios forecasts serve more purposes. For instance, a weather forecast has the goal to *warn* people without the need that "truth" is the ultimate aim.

Motivating Forecast Felicity for Machine Learning Scholars The talk in machine learning is peppered with "true labels", "true distributions", "true probability" or "true forecasts". Among the 61 papers presented as Orals at NeurIPS 2024 there are 34 in which "true" appears.[5] Sure, not all of these uses are related to forecasts and their relation to truth, but the examples presented in Table 1.1 taken from the papers should be convincing.

Machine learning arguably has a truth fetish in its rhetoric and culture [Moss, 2021, page 119][Aroyo and Welty, 2015], unnecessarily so.[6] I'm not claiming that machine learning does not aim (or should not aim) for finding "truth", but I'm claiming that machine learning does not need to find "truth" to still be relevant and/or useful.[7] I suggest to center the purpose or aim of a forecast in the discussion on the adequateness or success of the forecast. I want to broaden the scope of the purposes or aims of forecasts beyond "truth". For instance, does the weather forecast successfully *warn*? I will use the term *felicity* borrowed from speech act theory [Austin, 1975] to say: a *felicitous* forecast is a forecast which achieves its purpose or aim in being a forecast [Franco, 2019]. A more detailed treatment of the relationship of speech act theory and forecasting will be provided in Chapter 3.

I demonstrate that *felicity* and *felicity conditions*, i.e., conditions under which a forecast is felicitous, are helpful for machine learning. First, they do not carry heavy semantical baggage in machine learning yet. Second, they refresh a debate in which universality and monolithicness, e.g., in terms of finding truth in data [Williamson, 2024], are often considered to be justified and desired goals. Felicity emphasizes relativization and contextualization.

[5] For producing the statistics, ChatGPT was used to code a retrieval and automated checking algorithm.

[6] In fact, Moss [2021] goes deeper, demonstrating that the practice of machine learning implicitly and explicitly claims to be neutral, objective, and inevitable. In that, roughly speaking, machine learning makes itself *being* truth. From this perspective, I would in fact disagree that the truth fetish is unnecessary. Rather, the truth fetish is the fuel in the hype engine of machine learning.

[7] This claim is analogous to the critique on mirrorism, the focus on mirroring the truth instead of aiming for pragmatically or epistemically useful theories, in the psychological science as expressed in [van Dongen et al., 2025].

Reference	Quote(s)
[Cai et al., 2024]	"This analysis assumes the existence of *true* correspondences between labels in the pretrained and downstream domains." (Emphasis added)
[Lin et al., 2024]	"Moreover, as illustrated in Fig. 8b bottom, FlipClass outperforms leading methods [55] by moving closer to the *true* class distribution, yielding higher and less biased accuracies across all classes." and "FlipClass demonstrates a better fit to the *true* distribution, whereas InfoSieve shows skewed predictions." (Emphasis added)
[Guo et al., 2024]	"A more recent line of work [...] considers causal inference under the assumption of independent causal mechanisms (ICM) [...], which postulates that distinct causal mechanisms of the *true* underlying generating process do not inform or influence one another." (Emphasis added)
[Xie et al., 2024]	"Recently, Darrous et al. [2021] proposed a latent heritable confounder MR method to estimate bi-directional causal effects with a hierarchical prior on the *true* parameters." (Emphasis added)
[Amortila et al., 2024]	"Following prior work [...], we assume that the latent states can be uniquely decoded from observations, but that the *true* decoder is unknown and must be learned." (Emphasis added)

Table 1.1: Exemplary quotes from papers presented as Orals at NeurIPS 2024 referring to "true parameters", "true distribution", "true underlying generating process" or others.

Motivating Forecast Felicity for Probabilists The reader may have wondered why I avoided to use the word "probability". For instance, the rain example deliberately says "chance". The reason is that I intent to not evoke any specific connotation or meaning of "probability".[8]

The four problems raised before are closely related to questions which have already been asked (and partially answered) for probabilities. However, there are at least four reasons which make the forecasting perspective appealing.

Relevance Forecasting is a frame which is very relevant to the modern debate on probabilities (cf. [Höltgen, 2024]). A central element of machine learning is forecasting and machine learning is currently one of the key applications of probability theory.

Demarcated Scope Forecasting makes it possible to ignore other uses or contexts in which probabilities play a role, e.g., logical probabilities [Carnap, 1986], statements about

[8] "Chance" seemed a more ambiguous and vague term to me.

beliefs and credences [De Finetti, 1974/2017], uses in statistics or probabilities as a tool to make scientific error calculations more efficient [Glymour, 2001], in which necessarily concepts, questions and statements get conflated.

Relation to Felicity As speech acts, forecasts bring all four problems into touch with felicity conditions. It will follow that none of the problems allows for a universal, decontextualized solution.

Contextualized Understanding of Probabilities Forecast felicity fruitfully contributes to a contextualized understanding of probabilities. In [Dawid, 2017], Dawid dissects the exemplary question, "What is "the risk" that Sam will die in the next 12 months?".[9] From the perspective of forecast felicity, such a question is ill-posed. A more appropriate formulation would be: what is a felicitous forecast for the attribute "dying in the next 12 months" for the individual Sam in a certain context? In particular, we first have to ask, what does the forecast *do*? Does it *set* the rate for a life insurance company? Does it *warn* Sam in order to stop smoking? Only after we have answered those questions we can assess the statistical felicity conditions. For instance, what is a felicitous type of forecast to calculate a rate which Sam is willing to pay, and the life insurance company is somewhat safe-guarded against high losses? What is a felicitous evaluation metric which Sam cares about to evaluate the forecast for his death, such that Sam would take the needed action, e.g., stop smoking, to prolongate his life? Forecast felicity pushes us to ask pragmatically and epistemically useful questions.

Goal of Thesis The goal of the thesis is to expose, highlight and structure implicit and explicit choices necessarily made by forecasters. These choices partially determine the success of the forecasting, hence the *forecast felicity*. I won't suggest any universally "right" choice. Felicity (conditions) will be the glasses to wear for this work. Relativization and contextualization will be the primary new shades to see.

[9] In his paper, the term "risk" is based upon "probability".

2

How to navigate?

2.1 Structure

The thesis is structured in four main parts. Part I clarifies concepts and the narrative. After a concise introduction of forecasts (Chapter 3) and a short primer on felicity (Chapter 4), I demonstrate that forecasts, as for instance provided by machine learning algorithms, are speech acts (Section 4.2). Based on this understanding, I reiterate and detail the four problems identified before (Chapter 5).

In Part II, I focus on a type of statistical forecast which can roughly be summarized as measuring the size of an aggregate (Chapter 6). This part is built upon the following two papers which study a definition of randomness and the problem of measurability.

FR Rabanus Derr and Robert C. Williamson. Fairness and Randomness in Machine Learning: Statistical Independence and Relativization. *New England Journal of Statistics in Data Science*, 3(1):55–72, 2025. – [Derr and Williamson, 2024] – Chapter 7

SP Rabanus Derr and Robert C. Williamson. Systems of Precision: Coherent Probabilities on Pre-Dynkin Systems and Coherent Previsions on Linear Subspaces. *Entropy*, 25(9):1283, 2023. – [Derr and Williamson, 2023b] – Chapter 8

Part III is based on two works on the structure of evaluation metrics. The type of forecasts investigated in this part are assignments of numerical values which are rationalized by the evaluation on relevant aggregates (Chapter 9).

TTC Rabanus Derr, Jessica Finocchiaro and Robert C. Williamson. Three Types of Calibration with Properties and their Semantic and Formal Relationships. *Journal of Machine Learning Research*, 27(110):1-52, 2026. – [Derr et al., 2026] – Chapter 10

EM Rabanus Derr and Robert C. Williamson. Evaluation Metrics As Averaged Outcomes of Fair Gambles. *preprint*, 2026. – [Derr and Williamson, 2025] – Chapter 11

In total, the four contributions in Part II and III cover the four facets of forecast felicity mentioned in the title. The four facets do *not* correspond one-to-one to the four problems.

In the last Part IV, I supplement and diversify the four facets by briefly examining the role of the individual in statistical forecasts, taking another look at randomness (Chapter 12), and transcending the boundaries of statistical forecast felicity towards communication and performativity (Chapter 13). This part highlights interesting avenues for future research.

Finally, the Epilogue V summarizes the thesis in a practical "4F-Questionnaire" designed to assist and support forecast makers and receivers of forecasts. It makes explicit choices and justification of choices to the end of felicitous forecasting.

2.2 Remarks on the Included Papers

The four papers listed above form the core of the thesis. The four papers are not included in the chronological order of their development. The projects were started in the order, FR, SP, EM and TTC. To each paper, I devoted a chapter which contains a personal note, a summary, a translation table of important terms and symbols and the manuscript.

The published papers are kept as in the original publication. The not yet accepted contribution is included as the latest arXiv version dated June 23, 2026. Only minor modifications were made, such as image size, table layout or section levels for appendices.

Instead of separate bibliographies, I decided to merge them in a single document. In single cases, I updated bibliographic entries, for example due to publication, so that they may now differ from the original citation.

There is *no* unified notation nor terminology across the papers. Hence, the terms and symbols used in Chapter 7, Chapter 8, Chapter 10 and Chapter 11 differ, can be inconsistent and are distinct to the *unified* notation and terminology in the other chapters. For instance, what I introduce as "attribute" in Chapter 3 is called "outcome" in Chapter 11. To facilitate the reading and understanding a table in each of the named chapters is provided. This table gives the correspondences between key terms and symbols.

In addition, I have added a personal note to each article. It seems a rather exotic practice to publish a short personal story about a scientific project within papers. Exceptions fortunately exist, [Khrennikov, 2018, Meng, 2024]. For the apparent neutrality we loose the story, thoughts, and motivations behind the work, which emphasize the environment and labyrinths in which papers are written. This motivated me to add a subjective section to every manuscript, which I hope will make it more accessible.

2.3 Additional Papers not Included in the Thesis

During my doctoral studies, I have been involved in further projects which are not included in this thesis. For the sake of completeness, I list them in the following.

(a) Rabanus Derr and Robert C. Williamson. The set structure of precision. *International Symposium on Imprecise Probability: Theories and Applications*, PMLR, 2023. [Derr and Williamson, 2023a]

(b) Christian Fröhlich, Rabanus Derr and Robert C. Williamson. Towards a strictly frequentist theory of imprecise probability. *International Symposium on Imprecise Probability: Theories and Applications*, PMLR, 2023. [Fröhlich et al., 2023]

(c) Christian Fröhlich, Rabanus Derr and Robert C. Williamson. Strictly frequentist imprecise probability. *International Journal of Approximate Reasoning*, 168:109148, 2024. [Fröhlich et al., 2024]

(d) Armando Jose Cabrera Pacheco, Rabanus Derr and Robert C. Williamson. Aggregating Algorithm and Axiomatic Loss Aggregation. *Transactions of Machine Learning Research*, 2025. [Cabrera Pacheco et al., 2024]

(e) Natalie Collina, Rabanus Derr and Aaron Roth. The Value of Ambiguous Commitment in Multi-Follower Games. *Games and Economic Behavior*, Under submission, 2025. [Collina et al., 2024]

Note that the paper "The set structure of precision" is a conference version of the included Paper SP.

Part I

Forecasts as Speech Acts

3

What are Forecasts?

Forecasts are everywhere, central to machine learning and "decorated" with some fundamental problems.[1] Furthermore, forecasts are usually considered to be *statements* for which we can at least assign a degree of truth, if not even judging them as being true or false. In this Part I, I would like to argue that forecasts are *speech acts* for which truth conditions are only partially relevant. A simple example would be that a warning system forecasting storms does not need to provide "true" forecasts, but it should be successful in warning. To this end, let me first clarify what I mean by a forecast.

3.1 Forecasts are Utterances About Unobserved Attributes

I define a forecast as an *utterance* about unobserved attributes. Even though the temporal structure often is a definitional part of forecasts, i.e., a forecast about attributes has to be made *before* the attributes are revealed, I consider this temporal structure of less importance. For instance, an entity which takes pictures and states the probability that the picture shows a dog, still does, in my conception, forecasting. The critical assumption I want to make is that the forecasting entity has not observed the attributes yet, no matter whether deliberately or not. Second, I'd like *attribute* to be interpreted relatively generally, such that "showing a dog" or "it rains" are attributes. I use "attribute" to emphasize that it is attributed to a certain subject of the forecast, such as a picture or a day. Third, I do *not* want to make any assumption on how the forecasts comes into being. What matters is the type of utterance.

[1] If the reader is not convinced yet, I'd suggest to either reiterate the previous chapter or skip to Chapter 5.

3.2 Statistical Forecasts are Numerical Utterances About Attributes of Individuals with Relationship to Aggregate

The forecasts I am concerned about are *statistical*. A *statistical forecast* is a numerical utterance about unobserved attributes of an individual ι which bears a relationship to an aggregate $\mho$ of which the individual is a part, i.e., $\iota \in \mho$. A realized attribute "attributes" a property to an individual $\varrho(\iota) = \wp$.[2] An "individual" should not necessarily be interpreted as related to a person in a society. Instead I mean by "individual" some kind of unit defined a priori. Then, an "aggregate" is simply a collection of such units.

It is not a priori clear whether a forecast is statistical or not. Consider the exemplary forecast "Chance of rain today: 60%". The formulation suggests that the numerical utterance about the individual "today" bears a relationship to an aggregate of days, e.g., the percentage is the fraction of days with similar weather conditions on which it rained. However, the percentage can as well be some numerical expression of a belief of an expert about today without any reference to an aggregate whatsoever. A mere forecast often cannot be claimed to be statistical or not. Either the relationship to an aggregate is made explicit or has to be assumed.

Origin of Statistical Forecasts As mentioned earlier, I do not restrict the definition of a statistical forecast by the process from which the forecast originates. Hence, statistical forecasts can be the result of a machine learning model [Vamathevan et al., 2019], an expert judgment [Core Writing Team, 2023, European Central Bank, 2025], a numerical computation [Hennig et al., 2022, MeteoSwiss], or something else.

Nevertheless, the processes and assumptions are highly relevant for the forecast's felicity, as I will discuss in the following Chapter 4 (cf., [Wang et al., 2024]). Further, even though the forecast might be the product of some process, I will try to convince the reader that the forecast should better be understood as the act of *forecasting*.

Two Types of Statistical Forecasts I broadly distinguish between two types of statistical forecasts. These types are neither meant to be exhaustive, nor to be mutually exclusive. The two types structure the discussion of felicity conditions. Certain felicity conditions only apply to certain types of forecasts.

A2I The numerical utterance is an (idealized) measurement of the size of an aggregate with respect to the attributes of interest, i.e., how large are the sub-aggregates which

[2] Even though I provide formal symbols for important terms used in the thesis, I don't aim for a completely mathematical treatment. The formalization is intended to facilitate understanding. I use exotic notation to avoid overlap with the notation used in the main manuscripts included in this thesis (Chapter 7, 8, 10 and 11)

share the attributes of interest. Formally, how large is $\{\iota \in \mathfrak{V} \colon \varrho(\iota) = \wp\}$ for different possible realization of the attribute ϱ.

For instance, the forecast "Chance of rain today: 60%" can be the fraction of days with similar weather conditions on which it rained. In this case, the forecast is a measurement using relative counting. Being measurements, such forecasts are quantities [Carnap, 1986, page 74].[3] Quantities usually follow more relational rules then just numerical values do, e.g., there exists an order relation and a "unit measure" [Carnap, 1986].

I2A The numerical utterance is arbitrarily assigned to the individual, but evaluated on an aggregate of which the individual is a part.

Those forecasts are not necessarily quantities. For instance, "Chance of rain today: 60%" can be an expert's belief, as before (Section 3.2). But, different than before, if the expert's belief is evaluated on an aggregate, e.g., on all days of this year, then it is a statistical forecast of Type I2A.

The distinction between the types largely corresponds to Philip Dawid's separation of the G2i (group to individual) and i2G (individual to group) approach [Dawid, 2017]. Dawid considers the G2i approach as "the attempt to make sense of individual risk when Probability is understood as a group phenomenon". He considers this approach to be doomed. I take a more moderate position in this work, extend the meaning of G2i to statistical forecasts beyond individual risks and probabilities, and discuss the felicity of such forecasts. Dawid places more hope in i2G in which "we take individual risk as a fundamental concept, and explore what that implies about the behaviour of group frequencies". Type I2A forecasts differ in this aspect. I don't consider them as being a fundamental concept. Instead, I argue that the forecasts can be felicitous by solely justifying the appropriateness of the individual forecasts on an overarching aggregate.

Forecasts and Probabilities Not every statistical forecast is a probability. For instance, the forecast "Average rain amount today: 4 mm" is not a probability but suggests to be statistical as it refers to an average of an unspecified aggregate.

Nevertheless, there are statistical forecasts which are probabilities. For instance, if the forecast "Chance of rain today: 60%" continues with "Chance of no rain today: 40%", then the forecast is a probability distribution on the event space "rain" and "no rain". It might then be, depending on the understanding of "probability", considered a probability. In fact, the way how I use the term "probability" within this thesis is that it should be

[3] I do not want to open an entire chapter on measurement theory. There are more ways to define quantities, e.g., see [Tal, 2020]. Notably, measurement theory and measure theory are usually treated separately. However, the statistical measurement of the size of aggregates here will be formalized in terms of measure theory.

stripped of its semantical load and simply be considered a mathematical probability distribution. Exceptions are clearly marked. Mathematically, a probability distribution is an object which follows the rules of calculus of standard probability theory following [Kolmogorov, 1933] (translated in [Kolmogorov, 1956]), i.e., countably additive probabilities on σ-algebras. Thus, both types of forecasts A2I and I2A can be probabilities, depending on the properties of the measurement or the value assignment, i.e., whether they form a probability distribution. In fact, probabilistic forecasts simply fulfill a certain type of *internal consistency*.

Internal Consistency Beyond the important relational aspect of individual to aggregate, statistical forecasts are numerical. In particular, they usually follow *internal consistency* criteria, e.g., rules of some calculus, which are either naturally derived from idealization arguments (Type A2I) or directly imposed (Type I2A). A simple example for an internal consistency criterion would be that a forecaster who assigns 60% chance to rain today, has to commit to 40% chance of no-rain today. Internal consistency guarantees a minimal soundness of the forecast without referring to the relationship of the forecast with the realized attributes.

More generally, a standard example for an imposed internal consistency is that the forecast is a probability distribution. The understanding of "probability distribution" being an internal consistency criterion is supported by de Finetti's development of coherence [De Finetti, 1974/2017]. De Finetti [1974/2017] provides a set of rationality axioms for assigning numbers to uncertain outcomes, which he calls "coherence".[4] Those axioms ensure that the numbers follow, up to finite additivity, the rules of calculus of probabilities as promoted in the seminal work by Kolmogorov.

There is no single choice of internal consistency for statistical forecasts. For instance, the theory of imprecise probability allows for weaker internal consistency criteria, usually called "coherence" as well [Walley, 1991, Augustin et al., 2014]. Hence, internal consistency of forecasts will be part of the felicity conditions for forecasts (Paragraph 4.2).

The felicity of the forecast with respect to internal consistency depends on the context. If the forecasts are meant to be used to calculate expected utilities under the forecast, probability calculus guarantees mathematical soundness and comparability of expected utilities. If the forecasts are meant to be used by defining the most likely attribute, the forecast does not need to provide internal consistency as long as the order of probabilities is correct. In simple words, internal consistency is not a universal requirement but should match the purpose and aim of the forecast.

[4] In fact, this type of coherence is sometimes referred to as coherence$_1$. De Finetti suggested another equivalent coherence$_2$ using Brier score [De Finetti, 1974/2017, Seidenfeld, 1985].

4

Forecast Felicity

The general aim of this Part I is to argue that (statistical) forecast are, in the interesting cases, *speech acts* and should be judged by *felicity conditions*. Remind the statistical forecast "Chance of rain today: 60%" in the newspaper. Is it relevant whether the "true" chance of rain today is 60%? Isn't it more relevant whether the readers of the newspaper trust the forecast, that the forecast is for the region in which the newspaper is published, that the forecasts have been tested for the last half a year and have been shown to nudge Peter to take his umbrella on those days when it rained? The forecast *informs* and this might include to simply round the number to 60%, even though the "true" chance would have been 61.435%. In this way, the forecast is a speech act. It is judged by felicity and not truth conditions. Given that we have "statistical forecasts" on our table, let me briefly summarize parts of speech act theory which are relevant to elaborate the argument.

4.1 A Primer on Speech Act Theory by Austin

In a seminal series of lectures, John Austin [Austin, 1975] scrutinizes the observation that not all utterances are *constantive statements*. A constantive statement is what can be judged to be true of false. Austin essentially noted that there exist utterances, e.g., "I promise I finish proof-reading the PhD thesis by December.", which are not really true nor false, and rather become true by uttering them.[1] He promotes the conception of an utterance as a *speech act*, e.g., the promise is made.

[1] Arguably, he was not the first to pay attention to this use of language [Mulligan, 1987, Smith, 1990, Grgic and Žagar, 2020].

Components of Speech Acts Austin argues that the distinction between *constantive* – true or false statement – and *performative*[2] – making by saying – is not helpful to understand speech acts [Austin, 1975, Lecture 11]. Instead, he pushes the focus on three sub-acts performed within every speech act [Austin, 1975, page 101].[3]

Locutionary Act The actual utterance and the meaning of the utterance, i.e. the act *of* saying something [Austin, 1975, page 99]. – My examiner said to me, "I promise I grade the PhD thesis by November." and meant committing to perform the act so described.

Illocutionary Act The act performed *in* saying something [Austin, 1975, page 99]. – My examiner made the promise to grade my PhD thesis by the last day of November.

Perlocutionary Act The (intentional) consequence or effect of the utterance [Austin, 1975, page 101]. – My examiner actually does the grading by the set time and thus calms me that the defense of my dissertation will happen in this year.

Felicity Conditions In a first part of his lectures, John Austin proposes that the question whether a constantive statement is true or false, is analogous to whether a performative utterance is happy or unhappy, i.e. the act fails or does not fail [Austin, 1975, page 14]. For instance, "The tree is green." can be judged to be true or false. While, "I name this method ABTSOTA[4].", either actually names for the uttering person and others this method "ABSOTA" or not. Austin then makes the point that the dichotomy on the judgment conditions is as blurry as the constantive-performative line [Austin, 1975, page 150]. Hence, he suggests to consider *felicity conditions* for any speech act.[5]

Along the lines of Austin, I understand a speech act to be *felicitous* if the speech act, which comprises of sub-acts as dissected before, is successful in being performed. The conditions which have to be satisfied to make the speech act felicitous are the *felicity conditions*. The felicity conditions mainly, but not exclusively, refer to the illocutionary act within the speech act. Let me provide some examples of felicity conditions for the utterance given above, "I promise I finish grading the PhD thesis by November.": (a) The utterance has to be clearly articulated. This refers to the locutionary act. (b) The person uttering has to be the examiner. Austin calls this felicity condition "misapplication" if it is not fulfilled [Austin, 1975, A.2 on page 15 and 18]. (c) The person uttering has to intent to finish the grading. The opposite is called "insincerity" by Austin [Austin, 1975, Γ.1 on page 15 and

[2] A performative utterance is an utterance such as "I warn you.". The warning is the act itself.
[3] Scholarship has suggested different separations of sub-acts, e.g., [Searle, 1969].
[4] Always Better Than State Of The Art.
[5] Admittedly, Austin does not use the term "felicity condition". Instead, he talks about the negative "infelicities" whenever a speech act does not succeed for a certain reason, i.e., whenever certain felicity conditions are not met.

18]. (d) The person uttering actually needs to do the grading [Austin, 1975, Γ.2 on page 15 and 18]. This felicity conditions is in fact equal to the truth condition of "I grade the PhD thesis by December.".

It fills large parts of Austin's lectures to discuss the overlap and extent of his suggested classification of felicity conditions [Austin, 1975, page 15 and 18]. He further disentangles truth conditions and felicity conditions. For our purpose it suffices to understand that truth conditions and felicity conditions tightly interact. In particular, felicity conditions usually encompass truth conditions. However, there is a crucial distinction between those truth and felicity conditions, which is contextualization respectively relativization.

Contextualization and Relativization Truth conditions are, given the assumption of unique truth, universal. However, felicity conditions – as the illocutionary and perlocutionary act – are bound to the source of the utterance, the audience, and other parts of the context [Austin, 1975, page 145]. In particular, any speech act is always performed in a context and has a illocutionary and perlocutionary act included. Felicity conditions extend the scope of true-false to success-failure in context.

In this work I will use contextualization and relativization exchangeably. Contextualization is often understood as the existence of universal validity claims, but an embedding of an utterance within a social, historical or cultural frame [Finlayson and Rees, 2023]. Hence, contextualization accepts that "1 plus 1 is 2." is universally true and "Human Rights" are universally right, but there is a need of understanding them in context. Relativization questions the existence of such universal validity claims [Baghramian and Carter, 2025]. My choice is simply taken by convenience. I don't want to make any claim about the existence of universal validity claims.

Given this background on speech act theory let me embed forecasts in this theory. The motivation for this move is that forecasts are speech acts rather than constantive statements as I will demonstrate in the following.

4.2 Interesting Forecasts Are Speech Acts

There are forecasts which are speech acts. Forecasts *warn* about extreme weather events [Deutscher Wetterdienst, 2025]. Forecasts *inform* school authorities about students' expected drop-out rate [U.S. Department of Education, 2016]. Forecasts *promise* to predict stock market prices in AI-driven portfolios [Kolanovic and Krishnamachari, 2017]. Forecasts *describe* how the future of coastal regions will look like [Vitousek et al., 2017]. Forecasts *approve* drugs for the next round of experimental studies [Vamathevan et al., 2019].

There are forecasts which are not speech acts. However, those forecasts are not being

uttered in any context. They are uninteresting. There is no pragmatic value in them. They *do* nothing.

Hence, (interesting) forecasts are *acts*, e.g., they warn, they inform, they promise, they describe, they approve. In this respect they are performative, or better formulated in Austin's terms: the forecast has an illocutionary force. Note that this type of (illocutionary) performativity differs from the type of (perlocutionary) performativity as understood in machine learning (Section 13.2).

Felicity Conditions for (Statistical) Forecasts Forecasts as constantive statements allow for judgment about their truth. Forecasts as general speech acts allow for judgments by felicity conditions. This framing is better-suited for at least three reasons.

Aims, Purpose and Intentions of Forecasts The sole evaluation in terms of true or false of forecasts falls short of the aims, purposes and intentions of forecasts. The argument made here is closely aligned to what Franco [2019] writes about a speech act perspective on science mainly focusing on explanations. Franco dissects different aims of science, e.g., explanations to different audiences. Simply put, the "interests and cognitive resources of the speaker and audiences" [Franco, 2019, page 1007] determine the success of the act of explaining and not necessarily the truth of the explanation. Science is typically taken to follow an ideal of approximating truth [Wimsatt, 2007]. Forecasts are closer to a certain use (in daily life) and such an ideal of truth is less dogmatic, machine learning apparently being an exception (Section 1.3).

Context-Dependence Felicity conditions are context-dependent. The same forecast might be felicitous in one and infelicitous in another context. For example, the weather forecast in the newspaper is felicitous only if the reader is in the addressed geographical region.

No Distracting Semantics Felicity conditions are not yet loaded with connotational baggage in machine learning. I would like to fill the picture of felicity conditions for that research field and promote its perspective on the aims and requirement for felicitous forecasting.

To further convince the reader about the added value in discussing forecast felicity instead of forecast truthfulness, let me provide a short non-exhaustive lists of potential felicity conditions. The forecast is felicitous if,

a) the forecaster has the authority to provide the forecast. – A weather forecast by the Deutsche Wetterdienst (German Weather Service) might be trusted, and has the authority, while some random weather agency might miss this authority.

b) the forecast is provided in a matching context, e.g., is relevant for the audience. – A weather forecast for La Paz, Bolivia is likely to not be relevant for a German citizen.

c) the forecast is understandable by the audience. – A weather forecast given to children in the kindergarten might not result in the desired preparation against an extreme rain.

d) the forecast bears a relationship to the truth. – A weather forecast which uses a lottery to choose the next prediction is not adequate for warning against extreme weather events.

e) the truth conditions are relevant for the audience. – A weather forecast which has been shown to be provide true forecasts during the night, is relevant only for persons who need the truth conditions to hold during the night.

f) the truth conditions are relevant for the purpose, use or aim of the forecast. – A weather forecast for extreme weather events should likely be tested for truth in a different way than a weather forecast for daily temperature.

g) the forecast is internally consistent. – A weather forecast which predicts clear sky and rain simultaneously seems of little use.

Concluding, the truth of the forecasts is clearly one, but not the only, part of this mosaic which make a forecast felicitous. The type of felicity conditions given are relatively abstract and largely do not care about the exact nature of the forecasts.

In particular, the given felicity conditions do not require the forecast to be *statistical*, i.e., that the forecast bears a relation to an aggregate. Those felicity conditions specific to statistical forecasts, i.e., which concern the relationship of the individual to the aggregate, I call *statistical felicity conditions*.[6] With regard to statistical felicity conditions, I would now like to reiterate and reinvestigate the four problems stated earlier wearing "felicity-glasses".

[6] In Chapter 13, I address dimensions of felicity conditions that go beyond statistical aspects.

5

A Felicitous Framing of the Four Problems

In Section 1.2, I asked four questions about a rain forecast in a newspaper: "Chance of rain today: 60%". Given the elaborations on the definition of forecast above we can consider this rain forecast as a statistical forecast. Let me give us a deeper dive into the questions raised before. In particular, the distinction between the two types of statistical forecasts disentangles important aspects of the questions.

5.1 What Does the Forecast Mean? – The Interpretation Problem

What does it mean to have a chance of rain today of 60%? BBC Weather states that "It means that out of 100 situations with similar weather, it should rain on 60 of those, and not rain on 40." (numbers changed) [BBC]. The German weather service (DWD) provides an interpretation along the same lines [Deutscher Wetterdienst]. The US National Weather Service (NWS) just refers to a further unspecified "probability that the forecast grid will receive at least $0.01''$ of rain." [National Weather Service]. The Australian Bureau of Meteorology states that "this value refers to the likelihood of the selected location receiving more than the minimum measurable amount of rainfall ($0.2mm$) over the 24 hours from midnight to midnight" [Bureau of Meteorology (Australia), 2018]. Meteo Swiss explains that the 60% chance refers to the relative number of numerical weather models which forecast rain for a certain region where the numerical weather models have been initialized differently [MeteoSwiss]. Finally, it would even be reasonable to consider the forecast to be a subjective expert's judgment [Murphy and Brown, 1984], hence the forecast is simply a numerical belief.

In the given examples, it is clear that the way the forecast was made tightly interacts with the meaning given to the numerical value. Hence, the *interpretation problem* is (partially) induced by an opaque forecast maker. For instance, the rain forecast is not explained in

detail in every new edition of the newspaper.[1]

Further, note that the *interpretation problem* for forecasts resembles the *interpretation problem* for probabilities [Hájek, 2023], i.e., is a probability a belief, a relative frequency, a propensity, is it objective or subjective? There is a silent premise that there is something like a given "probability" and it is on us to interpret it. Analogously, there is something like a given "forecast" and it is on us to interpret it.

Finally, the statistical forecasts usually bear a relationship to the aggregate of which the individual is a part. Hence, the question for the meaning of a forecast should take this into account.

It is for these three observations that the question, "What does the forecast mean?", is better replaced by: *what is the relationship of the individual, whose attributes are forecasted, to an aggregate?* Then, we can ask about the felicity of the relationship within the context of the forecast. It would thus be more appropriate to call the interpretation problem the *relation problem*. To avoid confusion, I will adhere to the *interpretation problem*.

None of my included papers is specifically treating the interpretation problem. However, there are two answers captured in the separation into Part II and Part III. The two types of statistical forecasts let us further refine the question. Given that the forecast is of Type A2I, i.e., the numerical value is a measurement of an aggregate relevant to the individual, we need to clarify:

1. What is the aggregate?
2. What is the measure of the aggregate?

The first question is essentially the reference class problem, i.e., what is a felicitous reference class. The second question requires the felicitous choice of a measure. I call this the *measurement resolution* for the interpretation problem (Chapter 6).

For forecasts of the Type I2A, i.e., the numerical value is assigned individually and evaluated on an aggregate, the following two questions are key:

1. What is the aggregate?
2. What is the evaluation?

Again, the first question is related to the reference class problem mentioned before. However, I contend that this reference class is different in kind. As the evaluation justifies the

[1] Interestingly, in [Gigerenzer et al., 2005] the study authors suggest to be more explicit about the reference class when giving rain forecasts such as "30% chance of rain tomorrow" in order to facilitate the understanding of the forecasts.

forecast, I use the term *evaluation resolution* for this answer to the interpretation problem (Chapter 9). It is closely related to the next problem.

5.2 What Makes the Forecast "Good"? – The Evaluation Problem

Forecasts should be "good", e.g., if the weather forecasts said "Chance of rain today: 60%" on the last ten days and it has not rained, the forecast seems not "good". Hence, a "good" forecast is consistent with an aggregate of observations. The consistency is checked by means of evaluation. In other words, "goodness", in my conception, is external consistency.

As observations are sometimes equated with "truth", external consistency is closely related to a truth condition (Section 1.3). It is rather immediate to argue that external consistency is a felicity condition (see Chapter 11). The "goodness" of a forecast certificates the usefulness, reliability and relation to the real world which contribute to the success of the forecasting. Evaluations are relative or contextual. The forecast informing a surgery requires a different evaluation, than a forecast of weather published in a newspaper. Depending on the context there is a choice to make, what is the relevant evaluation. To inform this decision, we need to understand:

1. How different notions of evaluation relate.
2. What the limits of such evaluations are.
3. What the meaning or aim of an evaluation is.

This choice of evaluation is real. For instance, in the technical report on the GPT-4 model by OpenAI the authors provide accuracy tables for certain benchmarks, alongside a diagram depicting the calibration score [OpenAI (2023), 2024]. Calibration score and accuracy are evaluation metrics which serve different purposes and are in a non-trivial relationship to each other (cf., Chapter 11). A report of both evaluation metrics seemed necessary to address different readers and their own aims and purposes. On the other hand, the EU AI Act [European Parliament and Council of the European Union, 2024] Article 15 demands "an appropriate level of accuracy" for high-risk AI systems. This demand is interpretable as being "good" with respect to a relevant evaluation. In particular, the evaluation itself is not defined, hence is a contextual felicity condition.

The thesis does *not* add to the literature which argues about the shift from evaluation of forecasts to evaluation on interventions, e.g., [Liu et al., 2022, Wang et al., 2024, Zezulka and Genin, 2025], neither to the challenge of AI evaluation or auditing, e.g., [Burden et al., 2025]. In this thesis, I will consider a formal perspective on the evaluation problem. The contributions add to the project on mapping the landscape of evaluation *metrics* which, e.g.,

has been pursued by [Williamson and Cranko, 2023] for proper scoring rules, [Waghmare and Ziegel, 2025] for parametrized scoring rules or [Gneiting, 2011] for loss functions. I use the term "metric" not in a mathematical sense. Instead, the use of the term is consistent with Chapter 11. An "evaluation metric" *measures* how "good" a forecast is with respect to realized observations.

In Chapter 10 we – this work has been pursued together with Jessica Finocchiaro and Robert C. Williamson – provide a formal and semantical relation structure for notions of calibrations used in machine learning. In Chapter 11 we – this work has been pursued together with Robert C. Williamson – summarize a large set of evaluation metrics, encompassing types of calibration and regret, into a game-theoretic setup, in which evaluation is understood as gambling.

Note that the evaluation problem mainly applies to forecast of the Type I2A. In particular, the evaluation plays the significant role of defining the forecast as *statistical*. Nevertheless, forecasts of Type A2I as well can be evaluated. However, these forecasts are, by definition, assumed to be "good" (Chapter 6).

5.3 Is a Quantified Forecast Warranted? – The Measurability Problem

Statistical forecasts of Type A2I are not only numerical, they are quantified i.e., numbers made through a process of measurement [Carnap, 1986, page 74]. A key feature of quantities is that they have a unit measure [Carnap, 1986, page 73]. This begs the question, how to choose the "unit"? Respectively, what is a felicitous choice of "unit"? Note that this question essentially repeats the "What is the (idealized) measurement of the aggregate?" in Section 5.1.

But even before that, quantities presuppose that the entity measured is *measurable*, i.e. that the entity allows to be measured. Do we rightfully assume measurability? Let me clarify for Type A2I forecasts: the measured entity is the aggregate of which the subject of forecast, i.e., the individual, is a part (Section 3.2). For instance, the statistical forecast "Chance of rain today: 60%" is actually the relative number of days on which it rained among the aggregate of Mondays. Can we count the number of days in the aggregate of Mondays?

The measurability of the aggregate is not as clear as it might appear on the first glance. In Chapter 8 we provide a list of examples, from quantum physics to decision theory, in which universal measurability is questioned. I want to rely on a simple intuitive argument here. If the precipitation has never been recorded, then the relative number of days on which it rained is not defined, i.e., not measurable. This case is obviously made up, but missing data is a huge issue, e.g., check the missing data files in the open data base of the German

weather service (DWD).[2] Then, how can measurability be a contextual felicity condition? Because, precipitation might have been recorded at some places and not at others. The relative number of rain days might be measurable for some geographical areas but not for all.

Hence, in order to provide a felicitous, quantified, statistical forecast, we do not only require a felicitous choice of measure, but as well a felicitous choice of measurability.

1. What is the structure of measurability? In other words, what are the aggregates which we allow to be measured?
2. Can we provide any information about non-measurable aggregates?
3. What is the meta-structure of the structures of measurability? For instance, what are the structures of measurability which allows more aggregates to be measurable than a certain reference structure?

In Chapter 8 we – this project has been pursued together with Robert C. Williamson – provide answers to those questions under the assumption that the measure fulfills some weak internal consistency called additivity.

Finally, note that forecasts of the Type I2A might not be quantities in a narrow sense. For instance, they can simply refer to classes, i.e., 1 refers to "rain" and 0 refers to "no rain", when doing weather classification. Respectively, they can be probability distributions which I consider as purely mathematical object and not an abstract measure of likelihood. The measurability problem is circumvent by assuming that the individual assignment of a numerical value is without any direct semantical relation, but justified by evaluation.

5.4 Is a Statistical Forecast Warranted? – The Statisticalness Problem

Commonly, statistical inference is about to "predict as yet unobserved random variables from the same random system that generated the data" [Cox, 2006, page 7]. Accordingly, a potentially obvious definition, distinct to mine, could be: a forecast is statistical if the attributes of the individual are random. As randomness is usually taken to be a universal and absolute concept [Eagle, 2021], the statisticalness of the forecast is then a matter of the forecasted subject.

But, the ubiquitous use of statistical forecasts in problems from text generation [OpenAI (2023), 2024] to resource allocation in social problems [Perdomo et al., 2025] induce a paradoxical question: are the phenomena under consideration random and hence ask for

[2] https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/daily/kl/historical/

a statistical treatment? For instance, are the GPT-forecasted tokens which are used to coherently fill in a missing word in a text random [OpenAI (2023), 2024]? Is the drop-out of a student from a high-school random [Perdomo et al., 2025]?

To clean up the ambiguity and vagueness around the concept "randomness", let me use operational and mathematical definitions of randomness. The only such definitions I am aware of refer to an aggregate, most often sequences (Chapter 7 and Chapter 12). For this reason, I need to presuppose that there is an aggregate to well-define randomness.

The aggregate is furthermore definitional for statistical forecasts. In my conception, *statisticalness* is that there exist a relationship to an aggregate. A numerical utterance about unobserved attributes of an individual without any reference to an aggregate is not *statistical*. By a well-defined, felicitous assumption of randomness, the statisticalness of the forecast is rationalized. Intuitively, if the aggregate is random, the individual attributes are uncertain and require a statistical forecast. I will argue that randomness is indeed a contextual, relative assumption. The assumption is a choice for which we need to answer:

1. What are the limits of what randomness can be?
2. What structure does the space of randomness have?

In Chapter 7, we – the work has been pursued together with Robert C. Williamson – we provide answers to those questions. We suppose that forecasts are the limiting relative frequency of a sequence, i.e., a specific instance of a Type A2I forecast. Then, we leverage a precise mathematical definition of randomness following Von Mises [1919], Von Mises and Geiringer [1964] to show that the choice of randomness is formed by epistemic, pragmatic and ethical considerations.

In Section 12.2, I provide a sketch of an argument which Robert C. Williamson and I have used in an earlier version of Chapter 11. It says that randomness is the dual concept to predictiveness. In particular, "good" forecasts on certain observations, in the sense of an evaluation (cf. Section 5.2), is equivalent to random observations with respect to the forecasts. This way, forecasts of Type I2A are implicitly guaranteeing a felicitous choice of randomness, as long as the evaluation has been chosen felicitously.

I'm convinced that the *statisticalness problem* still remains partially unsolved. It is for the depth and the remaining ambiguity of those questions that I will provide answers only in a demarcated frame (cf. Section 12.2).

5.5 Conclusion

Forecast are speech acts. Forecasts *do* something. A successful speech act is felicitous. Felicity, however, depends on the context. Thus, for a statistical forecast to be felicitous requires context-dependent formal choices, such as evaluation metrics, definitions of aggregates, randomness or measurability. The remaining thesis will investigate the space of felicitous choices brought up by the four problems and consider why it is a choice[3], how the choice is structured, and what the consequences of the choice are. These are three fundamental questions of relativization (or contextualization). First, the investigations will start with the premise that the statistical forecast is an (idealized) measurement of an aggregate (Forecast Type A2I). In a second part, the statistical forecast is individualized, but evaluated on an aggregate (Forecast Type I2A).

[3] This has partially been answered in this chapter already.

Part II

Statistical Forecast Felicity: Aggregate to Individual

6

The Measurement Resolution

At the beginning of the 20th century Hilbert's call to mathematically treat the axiomatization of physics provoked the modern development of the theory of probability [Hilbert, 1902, The Sixth Problem]. Hilbert considered probabilities to be part of physics [Gorban, 2018, Section 1.4]. Kolmogorov's seminal book "Grundbegriffe der Wahrscheinlichkeitsrechnung"[Kolmogorov, 1933] can be seen as an offspring of this development. It dominates our understanding and formal treatment of probabilities until today. Shortly before, von Mises responded in "Grundlagen der Wahrscheinlichkeitsrechnung" [Von Mises, 1919] in a way that was perhaps even closer to the problem raised by Hilbert. Von Mises aimed for a physical theory of probability about mass phenomena. First, an (idealized) quantity, how often does an event occur, gets measured. Then, the obtained measurement is treated by the calculus of probability. The result is a testable number about a mass phenomenon observable in the world [Von Mises, 1981, Gillies, 2010]. This physical thinking is still prevalent in some text books on probability, e.g., [Papoulis, 1965]. From von Mises' perspective, probability theory is a theory of measurable mass phenomena. It does not provoke the interpretation problem of probabilities already mentioned in Section 5.1 to which Kolmogorov's more abstract theory, potentially unwillingly and accidentally, contributed.

It is not by accident that my definition for a statistical forecast of the Type A2I reads as: the numerical value of the statistical forecast is an (idealized) measurement of the size of an aggregate with respect to the attributes of interest, i.e., how large are the sub-aggregates which share the attributes of interest. The value of the statistical forecast *is* the (idealized) measurement. The meaning of the obtained quantity is created through the felicitous measurement [Carnap, 1986, page 74]. This is an extreme operationalists' point of view to resolve the interpretation problem [Chang, 2021].[1] I call this the *measurement resolution*.

[1]This extreme position on operationalism would potentially not be endorsed by one of its key figures Percy W. Bridgman [Bridgman, 1954].

The *measurement resolution* leaves central questions open:

1. What is the aggregate?
2. What is the measure of the size of the aggregate?
3. What is the relevance of the idealized measurement of the aggregate to the individual?

Let me first start with a short answer to the latest question before we go deeper into the first two. I claim that the relevance of the measurement is given through the membership of the individual to the aggregate which is measured. Hence, the measurement is *not* a quantity obtained on the individual[2], but a quantity obtained only through the existence of the individual in the aggregate. Let's consider a "classical" example in which one counts the number of penguins which have the flu $\varrho(\iota) = 1$ on an ice floe. The entirety of penguins on the ice floe is $\mho_0$. Let's say, one out of 20 have the flu. Now, does the information that $\frac{1}{20}$th of the penguins on the floe have the flu tell us anything about Peggy $\iota \in \mho_0$, who I meet on the floe? No, that would be an ecological fallacy[3]. Peggy as well has no $\frac{1}{20}$th fraction of flu. But, Peggy is part of a group in which a penguin has the flu. This, I claim, makes the measurement relevant for her.[4]

6.1 The Felicitous Choice of Reference Class

Let's keep on looking at the penguin colony $\mho_1$ and consider more than the single ice floe. Taking into account all the neighbouring ice floes in the bay, we count 30 out of 300 penguins have the flu. When I picked Peggy on the single floe I could have used the $\frac{1}{20}$ as a forecast that she has the flu. Now, for the entire bay, Peggy is part of an aggregate in which $\frac{1}{10}$th of the penguins have the flu. What is the "right" aggregate?

The choice described here is essentially the reference class problem for probabilities [Hájek, 2007]. The reference class problem has been first noted in frequential theories of probability [Venn, 1876, Reichenbach, 1949]. For those theories, the probability of an event is equal to the limit of the relative frequency of an event occurring in a sequence, i.e. the limiting relative frequency (see Equation 6.1). The reference class problem asks: what is the "right" sequence to consider?

The reference class problem is usually considered a central logical gap in frequential theories of probability [Hájek, 1996, 2009]. There is no "right" reference class which assigns the

[2]There are philosophical positions which actually refer to the quantity as a quantity of the individual, a "single case probability" [Fetzer, 1979] or "individual risk" [Dawid, 2017].

[3]This emptiness of aggregate statements for individuals seemed to motivate Clark Glymour to call probabilities "The Science of Nothing" [Glymour, 2001].

[4]In the words of Dawid [2017], a forecast of the Type A2I is analogous to a "groupist" understanding of probability.

"true" probability to a singular event, i.e., a single case probability [Fetzer, 1979]. However, subsequently Hájek [2007] argued that the reference class problem applies in a related fashion to other (objectivist) interpretations of probability as well.

Our felicity glasses allow us to re-frame the reference class problem. I suggest to depart from finding the "right" reference class, as if we could find the "true" probability of an event in the right sequence. Instead, I would like to ask for a "felicitous" reference class. What is the reference class which serves the purpose and aim of the forecast? The choice might be informed by limited access to certain aggregates, justifications of relevance of other aggregates and stability of the estimate.

The contextualized approached to the reference class choice is actually not just hypothetical. The German weather service (DWD) declares that the chance of rain forecasts are to understood as "that in $[n]$ out of $[m]$ days for a forecasted weather situation it has rained at the specified place" (n and m added, own translation from [Deutscher Wetterdienst]). Hence, the reference class is chosen to be the set of days on which the same weather situation, according to some definition, has been forecasted. BBC weather states that "that out of $[m]$ situations with similar weather, it should rain on $[n]$ of those, and not rain on $[m - n]$" (n and m added, [BBC]). The reference class is defined by the set of days with similar weather, and not forecasted weather. There is no "right" choice of reference class for the forecast, but both can be felicitous.[5]

A valuable project for future research would be to further structure the choice of reference class. By which conditions can a reference class be defined? What are the quantifiable consequences of the choice?

Finally, note that the aggregate might be an idealization. For instance, if the forecast is the limiting relative frequency, the aggregate itself is a countably infinite sequence. We cannot observe this infinite sequence. But, like von Mises, we can consider the sequence as an idealization of individual attributes – one of which pertains to the individual we are forecasting about.

6.2 No Measure In, No Measure Out

When the aggregate is chosen, then how do we measure the size of the sub-aggregates which share the attributes of interest, i.e., $\{\iota \in \mho \colon \varrho(\iota) = \wp\}$? A simple and practical measure is relative counting. Given the aggregate we count the number of individuals which share the attribute of interest which we are interested in and divide by the size of the entirety of the

[5] I assume in this argument that the German weather service and BBC weather *actually* provide their forecasts in the way described. I cannot check wether this is the case. Actually, it seems more likely that the agencies do some model averaging and ensembling as described more in detail by Meteo Swiss [MeteoSwiss].

aggregate. The number is used to forecast the attribute of an individual. That is what I did for the penguin example before.

Note that the normalization of the measure is not crucial for a statistical forecast. I consider a forecast "1" which states the number of penguins with flu on the ice floe as well as a statistical forecast. Providing the forecast "1" about Peggy, would here simply refer to: she is part of an aggregate of penguins in which one penguin has the flu.

Furthermore, the aggregate might be idealized and hence larger than countable. For instance, limiting relative frequencies are measures of an idealized aggregate. Let $\mho = \mathbb{N}$ be an idealized aggregate and $\varrho(\iota) \in \{0, 1\}$ be a binary attribute for all $\iota \in \mho$. Every $\iota \in \mho$ can be interpreted as an individual penguin. The value $\varrho(\iota)$ is the attribute whether the penguin has the flu or not. Hence, the sequence $(\varrho(\iota))_{\iota \in \mho}$ is an idealized aggregate of the penguins in and around the bay. The limiting relative frequency, if it exists, is defined as,

$$\lim_{n \to \infty} \frac{|\{\iota \in \mho \colon \varrho(\iota) = 1, 1 \leq \iota \leq n\}|}{n}. \tag{6.1}$$

Hence, by definition the obtained number is the relative count of sick penguins, taken to infinity. The size of the idealized aggregate of penguins which share the attribute "flu" is the limiting relative frequency. This example is interesting for at least three reasons.

(a) The frequential theory of probability by von Mises is based upon this idealized measurement. It is the starting point for the investigations on the statisticalness problem in Chapter 7.

(b) The choice of measure directly imposes a certain internal consistency to the forecasts. In particular, the measure is finitely additive, i.e., given a finite amount of disjoint outcomes, e.g., "penguin flu", "penguin vaccinated", "penguin healthy, not vaccinated", we can add up the measurements to get the joint size [Von Mises, 1919, Schurz and Leitgeb, 2008].[6]

(c) Furthermore, it turns out that the not every attribute pattern on the aggregate allows for measurement. For instance, the following sequence 0100110000111100000000 . . . (exponential growth of 0 and 1 segments), does not, as one can show, have a limiting relative frequency [Fröhlich et al., 2023]. Hence, a sequence of penguins for which the flu is distributed according to this sequence cannot get an idealized measurement in this specific sense. We observe an instance of the measurability problem (Section 5.3, see details in Chapter 8).

[6]The measure is *not* countably additive though. That is, for a countable set of disjoint attributes, the measures do not add up. The counterexample is readily is explained. Consider the outcomes "being penguin i" for all $i \in \mathbb{N}$. This outcome is individualized for every penguin in the series. The relative size, i.e., the limiting relative frequency, of each of those outcomes within the aggregate is 0. But, the joint outcome, "being penguin" has clearly limiting relative frequency 1, which is not equal to the countable sum of 0s.

In fact, standard probability distributions following Kolmogorov's axiomatization can as well be understood as idealized measurement.[7] The aggregate defined by Kolmogorov is the set $\mho$, an idealized penguin colony. Every $\iota \in \mho$ refers to an individual penguin. A (measurable) subset $A \subseteq \mho$ is equivalent to a subset of the penguin colony having the flu. The probability of A defines the idealized measurement of the aggregate, in this case the relative size of how many penguin have the flu in the aggregate. In total, Kolmogorov requires to choose a probability space $(\mho, \Sigma, \mu)$, where Σ is a σ-algebra and μ a probability distribution. Hence, internal consistency and measurability are imposed choices by the theory. In addition, there is the requirement to choose a probability measure.

How did we circumvent this basic choice of measure for limiting relative frequencies and counting? We did *not*. First, counting is a particular measure which considers individuals as "units". Second, limiting relative frequency is simply one choice of extending the counting measure to countable aggregates [Schurz and Leitgeb, 2008]. A closer look on the definition of limiting relative frequency shows that for a fixed n one is simply counting the relative amount of $\varrho(\iota) = 1$ among the first n entries of the sequence. This number is then taken to infinity. Instead, it would have been possible to mix the sequence first and then do the relative counting on an ever increasing prefix. This would have result in a different measure. *No measure in, no measure out.* There is a choice of measure.

6.3 The Felicitous Choice of Measure

The choice of measure is part of the felicity conditions. For instance, relative counts on finite aggregates put equal weight on the contribution of an individual to the size of the aggregate. A rain forecast which is based on the historical record of days on which a similar weather situation has been forecasted and it rained, can potentially be adjusted by reweighting the days depending on how similar the forecasted weather situation was. Which reweighting, i.e., measure, is more appropriate depends on the circumstances. The felicity of the forecasts depends on a felicitous choice of measure.

6.4 How to Practically Measure?

There remains a practical question: how is it possible to count individuals sharing attributes on an aggregate, where the aggregate contains the relevant individual, without knowing the individual's attribute? If the attribute is known, then by definition there is no forecasting. There are at least three options.

[7] This perspective is relatively unusual. More often, probability spaces following Kolmogorov are used as abstract source of randomness (Chapter 7). An exception being [Dawid, 1985b, page 111].

(a) In certain situations it is possible to measure the size of an aggregate which shares an attribute without knowing the exact correspondence of attributes to individuals. For instance, when the flu causes penguin to produce differently colored excrement, then we can measure the amount of differently colored excrement on the ice floe when all penguins are fishing under water. Such a situation essentially makes the penguins on the ice-floe exchangeable. This scenario is not entirely made up. Waste water analysis for drug use is a comparable type of counting [Van Wel et al., 2016].

(b) Commonly, Type A2I forecasts are rationalized by the inductive assumption that an (idealized) measurement of an aggregate applies to an individual. I suggest to split the inductive assumption in two steps. (a) It is assumed that the individual about which the forecasts is uttered is part of an aggregate which has partially been observed already. (b) It is assumed that the individual would not change the measurement. Hence, this allows to make a measurement of the occurrence of an attribute relevant to an individual whose attribute is unknown. For instance, the German weather service forecasts the chance of rain by taking the relative count of days on which it rained among the days with similar forecasted weather conditions [Deutscher Wetterdienst]. Then, it is assumed that the measurement applies to the next day. To justify the inductive step it is common to fall back to idealized aggregates and measures, such as sequences and limiting relative frequencies, as it is clear that a single individual only marginally contributes to the obtained measure.

(c) The measurement can simply be considered hypothetical. In this scenario, the forecaster has to felicitously justify the hypothetical measurement.

Conclusion The *measurement resolution* responds to the interpretation problem. The cost for the resolution is paid by the burden to make a felicitous choices of a reference class and measure. Furthermore, the construction of Type A2I forecasts can render its evaluation unnecessary. The felicitous choice of measurement essentially implies the assumption of the "goodness" of the forecast (cf. Chapter 9).

In the following two chapters, I will turn to the subtle problem of measurability and the use of a statistical forecasts of Type A2I. In particular, I will takes us on a more detailed journey through von Mises theory of probability in the following, with which this chapter started.

7

Fairness and Randomness in Machine Learning: Statistical Independence and Relativization

Author Contributions

Robert C. Williamson (RW) approached Rabanus Derr (RD) with the idea of linking fairness in machine learning to randomness as formalized by Richard von Mises. RW supervised and steered this project. Rabanus Derr wrote the majority of the paper, while RW took over the internal reviewing and editing.

Note that different to the original, published work in the New England Journal of Statistics in Data Science, I used the spelling "von Mises" with a lowercase "v" unless the name appears at the beginning of the sentence.

The Personal Note

On a sunny day, it must have been some time around May 2021, I met Bob for the first time in person. A part from, "Oh, he is taller than I expected." – thanks to his Zoom-camera usually pointing right downwards to his chair – I remember that we essentially talked – or better said, I listened – about some Austrian physicist. This physicist named Richard von Mises, who lived and worked at the beginning of the last century, developed a theory of probabilities, but in a different way than I was used to. The atoms of this theory were sequences. Surprisingly, via some relative counting taking to the infinite the intuitive rules of calculus of probabilities re-appeared from "nothing". Further, Bob argued that von Mises theory is tightly related to fairness in machine learning and randomness.

In the months before I started my PhD with Bob, I cycled through Scandinavia. In the Norwegian landscapes, "randomness" echoed through my mind without hitting borders and fading in the void. There was a "need" to fill the space.

Back in October, my first task was to translate a piece called "Die Widerspruchsfreiheit des Kollektivbegriffs" [Wald, 1937] which essentially showed that the theory of Richard von Mises was consistent. I never did the full translation and just jumped from one reading to the next.

Before, I used to think probability theory was an extremely beautiful mathematical branch, very profound and rich, right on the fine line of being practical enough to motivate and theoretical enough to interest me. I have not anticipated the sea of meanings, concepts, and theories which lay hidden there, beyond the superficial glimpse I had previously gained.

Rather early, the main take of the work, "randomness is fairness", seemed clear to me. I was more sure about its truth, than I had good arguments for it. The details were messy, the text unpolished, the overall picture not well-structured and coherent. It took a lot of iterations until the work got in the shape it is now. It made me learn to write a paper. It was not always fun. It got me emotionally detached from the work. It got published without getting helpful reviews. It got the first bit of my thesis.

Still, I think that this work contains some of the parts that I understand least within my thesis. It includes parts where I fail to even formulate reasonable questions.

Summary: Fairness and Randomness in Machine Learning

Motivation Machine learning is used *in* society. Its uses has created and amplified problems with societal dimensions, such as manipulation, privacy, fairness. One of those problem topics grew into the field of "Fair Machine Learning" [Barocas et al., 2023].

Its community came up with mathematical criteria of fairness in order to constrain algorithms to fulfill them and check forecasts to be fair or not. The debate mainly revolved and revolves around how to implement the fairness definitions and which fairness definitions are the "right" one to use, their mutual incompatabilities, and their trade-offs. Little attention is paid to the fact that three major types of fairness criteria, *Independence*, *Sufficiency* and *Separation*, are built upon statistical independence and require some kind of grouping.

For instance, *Independence* demands that the probability of assigning a certain forecast to an individual is statistically independent to the membership of the individual in a certain sensitive group[1] [Calders and Verwer, 2010, Kamishima et al., 2012, Dwork et al., 2012].

It is far from obvious what the options to group and their consequences are [Hu and Kohler-Hausmann, 2020, Doh et al., 2025]. Similarly, it is unclear what statistical independence means when referring exclusively to a mathematical definition.

Contributions In this work, we contribute to the these less treated, but central questions. The main tool we use is a mathematical theory of probability introduced by Richard von Mises. He defines probabilities as limiting relative frequencies [Von Mises, 1919, Von Mises and Geiringer, 1964, Von Mises, 1981]. We argue that statistical independence "as is", which refers solely to a mathematical definition based on Kolmogorov's classical probability theory, is insufficient to semantically underpin definitions of fairness. However, von Mises' frequential theory offers one possible interpretation.

This perspective further implies that fairness understood as one of the criteria, *Independence*, *Separation* or *Sufficiency*, and randomness, as defined by von Mises, become equivalent. Different to Kolmogorov [1933], von Mises directly incorporated randomness via some independence criterion in his theory. Probabilities, for him, only exist for random mass phenomena. The equivalence implies that the set of groups for which predictions or actions are fair is equivalent to the set of groups to which they are random and vice-versa, hence showing a consequence of a certain grouping. Furthermore, the context-dependence and contestedness of fairness undermines the often-claimed universality of randomness.

This work was carried out before I began to frame my thesis based on *forecast felicity*. For this reason, the used terms and symbols in the subsequent sections differ from the ones used before. I summarize the key concepts and their "translations" in Table 7.1.

Tools and Sketch of Arguments Von Mises describes mass phenomena, which are the subjects of his probability theory, by sequences. For instance, let $(x_i)_{i\in\mathbb{N}}$ be the sequence of 0s and 1s which corresponds to the days on which it rained (1) or not rained (0) , i.e.,with

[1]The literature on fair machine learning often refers to "sensitive groups", which are socially salient groups defined for protection. Standard examples are race, gender or age.

Unified Notation & Terminology	Chapter 7
statistical forecast	prediction
individual, ι	element, individual, $i \in \mathbb{N}$
aggregate, $\mathfrak{U}$	sequence, $\mathbb{N}$
attribute, $\varrho(\iota) = \wp$	outcome, $x(i) = x_i \in \{0, 1\}$
measurement	limiting relative frequency "probability" if $(x_i)_{i \in \mathbb{N}}$ is collective

Table 7.1: Translation table – Fairness and Randomness in Machine Learning: Statistical Independence and Relativization

$x_i \in \{0, 1\}$ for all $i \in \mathbb{N}$. The probability of rain on those days of the sequences is the *limiting relative frequency* of 1s, if it exists,

$$\lim_{n \to \infty} \frac{|\{i \in \mathbb{N} \colon x(i) = 1, 1 \leq i \leq n\}|}{n},$$

and if the following crucial additional assumption holds: for a (at most countable) set of *selection rules* $\mathcal{S}$ the limiting relative frequency does not change when the outcomes are subselected by a selection rule. A selection rule is a sequence $(s_j)_{j \in \mathbb{N}} \in \mathcal{S}$ with $s_j \in \{0, 1\}$ for all $j \in \mathbb{N}$ and $s_j = 1$ for infinitely many $j \in \mathbb{N}$. The subselection condition is formalized by, for every $s \in \mathcal{S}$,

$$\lim_{n \to \infty} \frac{|\{i \in \mathbb{N} \colon x(i) = 1 \text{ and } s(i) = 1, 1 \leq i \leq n\}|}{|\{j \in \mathbb{N} \colon s(j) = 1, 1 \leq j \leq n\}|} = \lim_{n \to \infty} \frac{|\{i \in \mathbb{N} \colon x(i) = 1, 1 \leq i \leq n\}|}{n}.$$

Von Mises calls a sequence which has invariant limiting relative frequencies under a set of selection rules $\mathcal{S}$ *collective* (Definition 7.2). The collective is *random* with respect to $\mathcal{S}$.

Kolmogorov defines that two events are independent if their joint probability factorizes to the individual probabilities, roughly denoted $P(A \cap B) = P(A)P(B)$. In contrast, von Mises defines statistical independence between collectives and uses subselection again. Collective x is *independent* of collective y if the collective x has invariant limiting relative frequencies under the subselection defined by the collective y which can be treated as subselection rule as long as its limiting relative frequency is greater than 0. Hence, randomness is defined by statistical independence from subselection rules, and statistical independence is an interpretable concept defined between collectives.

We suggest to re-interpret definitions of fairness with von Mises statistical independence. That is, for instance, a sequence of binary predictions is fair with respect to a sensitive group, if the prediction is independent of the sensitive group belonging (Definition 7.4). The von Misesean reading would be, a collective of binary predictions is fair with respect to a sensitive group if the relative limiting frequency of predictions subselected by group membership

remains unchanged with respect to the relative limiting frequency of all predictions. In other words, the frequency of prediction "1" is f on all individuals for which a prediction has been given, and it is f on all individuals belonging to the sensitive group.

With this background it is rather immediate that fairness can be equated to randomness through von Mises theory. A collective is not only random with respect to the subselection rules, it is statistically independent to each selection rule. Hence, it is fair with respect to the subselection rules (Proposition 7.5).

The summarized implications are that demanding fairness is randomization and randomness is an ethical choice. In particular, there is no absolute and universal choice of randomness, because otherwise there would be an absolute and universal choice of groups against which to be fair. Nor is there perfect randomness, i.e., a collective which is random with respect to all subselection rules. Such a sequence necessarily is almost constant, with only finitely many exceptions of deviation. Even though, the investigation of fairness in machine learning motivated this work, the consequences mainly touch our understanding of randomness.

Relation to Forecast Felicity The felicity condition I would like to understand better through this work is not that the forecast is felicitous if it is fair. Instead, I would like to shed light on the relationship of forecast felicity and randomness. This paper guides towards some answer of the fourth problem, the *statisticalness problem* (Section 5.4).

Whenever the subject of the forecast is understood as a mass phenomenon, von Mises' formalization is appropriate as a model. A collective describes the forecasted attributes. Importantly, the collective as model implies the forecast to be of Type A2I.

The collective is an idealized aggregate of which the individual subject of the forecast is a part. The relative limiting frequency is a measurement. Exemplarily, the forecast "Chance of rain today: 60%" would be the result of the idealized measurement that today is part of an infinite sequence of days $\omega \in \Omega = \mathbb{N}$, where the limiting relative frequency of days on which it rained, i.e., $x(\omega) = 0$ for day $\omega \in \mathbb{N}$, is 0.6.

By definition a forecast of Type A2I is *statistical*. Furthermore, the forecast's felicity depends on the adequateness of the collective as formal description.

Central to the formal description is the definition of randomness. What forms the a felicitous definition of randomness, i.e., the set of selection rules?

Epistemic Considerations Which subselections do we know to (not) change the limiting relative frequency. For instance, the week days do not change the relative frequency of days with rain, but the temperature of the nearby ocean being above $10°C$ does. Thus, it is easy to justify that Wednesdays are ignorable for a weather forecast. Is might be harder to justify that ocean temperature above $10°C$ is ignorable.

Pragmatic Considerations Which subselections do we practically obtain and they do (not) change the limiting relative frequency. For instance, we might deliberately ignore and not measure whether the temperature of the ocean is larger than $10°C$, even though this might be a selection rule which changes the relative frequency of rainy days. For a coastal wind farmer such a choice could render the forecast useless, for a banker in Hamburg who checks whether she should bring her umbrella to the office it might not matter.

Ethical Considerations Our argument laid out in this work, adds a further aspect, the ethical consideration. For instance, the subselection by race *should not* lead to changing limiting relative frequencies when the days are replaced by persons and "rain/no-rain" is replaced by "pays back loan/does not pay back loan". The forecast is ethically inquired to make no difference, in terms of the limiting relative frequency, between races.

Hence, the choice of randomness to adequately model the mass phenomenon has to be relative to the context at hand. Randomness is a deliberate assumption of ignorance. The choice has to be felicitous with respect to the act and context of the forecast derived from the formal model.

Furthermore, the definition of randomness via a set of selection rules highlights at least three further aspects:

i) "Perfect" randomness is uninteresting. The sequences which have no changing limiting relative frequency with respect to *all* selection rules are (up to finitely many exceptions) constant.

ii) Randomness can be almost trivial. If the set of selection rules $\mathcal{S}$ is empty, then every sequence with converging limiting relative frequency is random.

iii) The set of selection rules $\mathcal{S}$ forms a pre-Dynkin system. Since a selection rule is a binary sequence, it can equivalently be expressed as a subset of the natural numbers $A \subseteq \mathbb{N}$. The set of selection rules, i.e., every selection rule is a set, is closed under complement and disjoint union. To illustrate, given two disjoint selection rules, e.g., Wednesdays and Thursdays in the above example, and both do not change the limiting relative frequency. Then, the selection rule "Wednesdays or Thursdays" does not change the limiting relative frequency. Hence, the space of von Mises' randomness definitions is the set of pre-Dynkin systems.

Abstract

Fair Machine Learning endeavors to prevent unfairness arising in the context of machine learning applications embedded in society. To this end, several mathematical fairness notions have been proposed. The most known and used notions turn out to be expressed in terms of statistical independence, which is taken to be a primitive and unambiguous notion. However, two choices remain (and are largely unexamined to date): what exactly is the meaning of statistical independence and what are the groups to which we ought to be fair? We answer both questions by leveraging Richard von Mises' theory of probability, which starts with data, and then builds the machinery of probability from the ground up. In particular, his theory places a relative definition of randomness as statistical independence at the center of statistical modelling. Much in contrast to the classically used, absolute i.i.d.-randomness, which turns out to be "orthogonal" to his conception. We show how von Mises' frequential modeling approach fits well to the problem of fair machine learning and show how his theory (suitably interpreted) demonstrates the equivalence between the contestability of the choice of groups in the fairness criterion and the contestability of the choice of relative randomness. We thus conclude that the problem of being fair in machine learning is precisely as hard as the problem of defining what is meant by being random. In both cases there is a consequential choice, yet there is no universal "right" choice possible.

7.1 Introduction

Under the name "Fair Machine Learning" researchers have attempted to tackle problems of injustice, fairness, discrimination arising in the context of machine learning applications embedded in society [Barocas et al., 2023]. Despite the variety of definitions of fairness and proposed "fair algorithms," we still lack a conceptual understanding of fairness in machine learning [Passi and Barocas, 2019, Scantamburlo, 2021]. What does it mean for predictions to be fair? How does the statistical frame influence fairness? Is there fair data and what would it look like? For instance more concretely, how does a population of individuals and their corresponding predictions look like if a provided definition of fairness is fulfilled?

We focus on a collection of widely used fairness notions which are based on statistical independence e.g., [Calders and Verwer, 2010, Hardt et al., 2016, Chouldechova, 2017], but examine them from a new perspective. Surprisingly, debates concerning these notions have not questioned the role and meaning of statistical independence upon which they are based. As we shall argue, statistical independence is far from being a mathematical concept linked to one unique interpretation (see §7.4.2). This paper, in contrast to much of the literature on fairness in machine learning, e.g., [Calders and Verwer, 2010, Dwork et al., 2012, Hardt et al., 2016, Chouldechova, 2017], investigates what many definitions of fairness take for granted: a well-defined and meaningful notion of statistical independence.

Another, less popular, strand of research investigates the role of randomness in machine learning [Steinwart et al., 2009]. The standard randomness notion, independently and often identically distributed data points, suffers the longstanding critique of being inadequate (cf. [Shafer, 2007]). Again, statistical independence lies at the foundation of this, hitherto unrelated to fairness, concept. (We will justify below our use of "randomness" as used here.)

At the core of both our observations is the unreflective use of a convenient mathematical theory of probability. Kolmogorov's axiomatization of probability theory, developed 1933 in his book [Kolmogorov, 1933] (translated in [Kolmogorov, 1956]), dominates most research in machine learning. As Kolmogorov explicitly stated, his theory was designed as a purely axiomatic, mathematical theory detached from meaning and interpretation. In particular, Kolmogorov's statistical independence lacks such reference. However, the modeling nature of machine learning and the arising ethical complications within machine learning applied in society ask for semantics of probabilistic notions.

In this work, we focus on statistical independence. We leverage a theory of probability axiomatized by von Mises [Von Mises and Geiringer, 1964] in order to obtain *meaningful* access to probabilistic notions. (In leaning on von Mises we are directly following the explicit advice of Kolmogorov [1956, page 3, footnote 4].) This theory construes probability

theory as "scientific"[2] (as opposed to purely mathematical) with the aim to describe the world and provide interpretations and verifiability [Von Mises and Geiringer, 1964, pages 1 and 14]. Von Mises' theory of probability provides a mathematical definition of statistical independence which describes statistical phenomena observable in the world. In particular, von Mises' statistical independence is mathematically, but not conceptually, related to Kolmogorov's definition.

In this paper, we, to the best of our knowledge, are the first to apply von Mises' randomness to machine learning and to interpret randomness in machine learning in a von Mises' way. The paper is structured as follows:

In Section 7.2, we outline our statistical perspective on machine learning. We present the "independent and identically distributed"-assumption (i.i.d.-assumption) as one commonly used choice for modeling randomness. The further occurrence of statistical independence as fundamental ingredient of fairness notions in machine learning (§7.3) pushes us to the question: "How to interpret statistical independence in (fair) machine learning?" Remarkably, "Independence" governs many discussions around fairness in machine learning without getting to a concrete meaning of this term. Its deeper semantics remain untouched even in the considerably exhaustive book by Barocas et al. [Barocas et al., 2023, Chapter 3, p. 13].

We first dissect Kolmogorov's widely used definition of statistical independence in Section 7.4 before we propose another mathematical notion following von Mises. Von Mises uses his notion of independence in order to define randomness. We contrast his definition and the i.i.d.-assumption in Section 7.5. This reveals a general typification of mathematical definitions of randomness which most importantly differ in the *absoluteness* respectively *relativity* to the problem under consideration (§7.5.2).

Finally, we leverage von Mises' definition of statistical independence to redefine three fairness notions from machine learning (§7.6). Against the background of von Mises' probability theory, we then link randomness and fairness both expressed as statistical independence (§7.7). Thereby, we reveal an unexpected hypothesis: randomness and fairness can be considered equivalent concepts in machine learning. Randomness becomes a *relative*, even an ethical choice. Fairness, however, turns out to be a modeling assumption about the data used in the machine learning system.

Due to the frequent use of the word "independence" with different meanings in this paper, we differentiate. By "independence" we mean an abstract concept of unrelatedness and non-influence [Simpson and Weine, 1989]. We use it interchangeably with "statistical in-

[2]We further use the term "scientific" in order to describe a theory modeling a phenomenon in the world providing interpretations and verifiability in the sense of von Mises (spelled out more concretely in Section 7.4.3). Any more detailed discussion, e.g., along the lines of Popper [2010], is out of the scope of this paper.

dependence" which emphasizes the probabilistic and statistical context. When referring to the later introduced, formal definitions of statistical independence following Kolmogorov or von Mises we explicitly state this. Finally, we assign "Independence" (capital "I") to one of the fairness criteria in machine learning which demands for statistical independence of predictions and sensitive attribute [Barocas et al., 2023, Räz, 2021]. Despite the abstract appearance of this work, we consider it as part of the project to make the very abstract notion of "independence" or "statistical independence" at least a bit more concrete — specifically, we describe independence in terms of samples, not abstractions such as countably additive probability measures. We ask (and try to assist) the reader to become aware of the implicit assumptions taken about the concept of "independence".

7.2 A Statistical Perspective on Machine Learning

Machine Learning ingests data and provides decisions or inferences. In this sense, at its core, it is statistics[3]. Adopting this perspective, we understand machine learning as "modeling data generative processes". Statistics, respectively machine learning, asks for properties of data generating processes given a collection of data [Wasserman, 2004, p. ix] [Bandyopadhyay and Forster, 2011, p. 1].

We assume that a data generative process occurs somehow in the world. We are confronted with a collection of numbers, the data, produced by the process and acquired by measurement. A "model" is a mathematical description of such a data generating process. This description should allow us to make predictions in an algorithmic fashion. Thus, we require, *inter alia*, a mathematical description of data — "a model for data collection" [Casella and Berger, 2002, p. 207]. What is arguably *the* standard model of data is stated in [Devroye et al., 1996, p. 11]:

> We shall assume in this book that $(X_1, Y_1), ..., (X_n, Y_n)$, the data, is a sequence of independent identically distributed (i.i.d.) random pairs with the same distribution[…].

Similar definitions can be found in many machine learning or statistics textbooks (e.g., [Casella and Berger, 2002, Def. 5.1.1]). Data (measurements from the world) is conceived of (mathematically) as a collection of random variables which share the same distribution and which are (statistically) independent to each other. Implicit in this definition is that the data indeed *has* a stable distribution. The assumed independence can be interpreted

[3] Machine learning could be viewed as classical statistics with a stronger focus on algorithmic realizations (cf. [Shalev-Shwartz and Ben-David, 2014, p. 6] or [Lafferty and Wasserman, 2006]). While there are ML approaches that do not seem statistical in the classical sense (e.g., worst case online sequence prediction), our general description still holds.

as presumption of randomness of the data. Each data point was "drawn independently"[4] from all others, with the obvious interpretation that each data point does not give a hint about the value of any other.

The i.i.d. assumption is two-fold: 1) the assumption of identical distributions of a sample, and 2) the mutual independence of points in the sample. The second assumption alone captures and pertains to randomness [Humphreys, 1977, Section 3]. However, since the use of i.i.d. is more common, we refer to this more specific assumption.

Even though many results in statistics and machine learning rely on the i.i.d. assumption (e.g., law of large numbers and central limit theorem in statistics [Casella and Berger, 2002], generalization bounds in statistical learning theory [Shalev-Shwartz and Ben-David, 2014] and computationally feasible expressions in probabilistic machine learning [Devroye et al., 1996]), it has always been subject of fundamental critique[5]. Other randomness definitions are rarely applied, but exceptions exist [Vovk, 1993, Tadaki, 2014].

In summary, statistical conclusions often rely on the i.i.d.-description of data. This description embraces a model of randomness making randomness a substantial assumption about the data in statistics and machine learning. Interestingly, statistical independence lies at the foundation of another, hitherto unrelated, concept: many fairness criteria in fair machine learning are expressed in terms of statistical independence.

7.3 Fair Machine Learning Relies on Statistical Independence

With the broad use of machine learning algorithms in many socially relevant domains, e.g., recidivism risk prediction or algorithmic hiring, machine learning algorithms turned out to be part of discriminatory practices [Chouldechova, 2017, Raghavan et al., 2020]. These revelations were accompanied by the rise of an entire research field, called "fair machine learning" (cf. [Calders and Verwer, 2010, Chouldechova, 2017, Barocas et al., 2023]). We do not attempt to summarize this large literature here. Instead, we simply take a snapshot of the most widely known fairness criteria in machine learning [Barocas et al., 2023, p. 45].

[4] Observe that as a mathematical theory of data this already leaves a lot to be desired: you will not find in any text a precise and constructive explanation of this process of "drawn independently". To be sure, the notion of "statistical independence" is well defined (see further below), as is a collection of random variables. But not the mysterious process of "drawing from". The closest we can come to such a description of mechanism is the indirect abstract version: the data are created (by the world) in a manner that the (mathematical theorem) of the law of large numbers holds, and that their empirical distribution converges to the distribution which was given in the first place.

[5] Nicely summarized by Glenn Shafer in a comment on [Gammerman and Vovk, 2007] "The i.i.d. case has also been central to statistics ever since Jacob Bernoulli proved the law of large numbers at the end of the 17th century, but its inadequacy was always obvious." [Shafer, 2007].

7.3.1 Three Fairness Criteria in Machine Learning

The three so called observational fairness criteria, which are expressed in terms of statistical independence, encompass a large part of fair machine learning literature: [6]

Independence demands that the predictions $\hat{Y}$ are statistically independent of group membership S in socially salient, morally relevant groups [Calders and Verwer, 2010, Kamishima et al., 2012, Dwork et al., 2012].

Separation is formalized as $\hat{Y} \perp S|Y$, i.e., the prediction $\hat{Y}$ is conditionally independent of the sensitive attribute S given the true label Y [Hardt et al., 2016].

Sufficiency is fullfilled if and only if $Y \perp S|\hat{Y}$ [Kleinberg et al., 2017].

For the sake of distinguishing between the fairness criteria "Independence" and statistical independence, we henceforth mark all fairness criteria by a leading capital letter. Each of the notions appear in a variety of ways and under different names [Barocas et al., 2023, p. 45ff]. From the perspective of ethics, the fairness criteria have been substantiated via loss egalitarianism [Binns, 2018, Williamson and Menon, 2019], absence of discrimination [Binns, 2018], affirmative action [Biddle, 2006, Räz, 2021] or equality of opportunity [Hardt et al., 2016].

Certainly, statistical independence is not equivalent to fairness in general (a constellation of concepts sharing a common name, and perhaps little else agreed by all) [Dwork et al., 2012, Hardt et al., 2016, Räz, 2021]. The nature of fairness has been discussed for decades, e.g., in political philosophy [Rawls, 1971], moral philosophy [Lippert-Rasmussen, 2006] and actuarial science [Abraham, 1985]. The "essentially contested" nature of fairness suggests that no universal, statistical criterion of fairness exists [Gallie, 1955]. How fairness should be defined is a context-specific decision [Scantamburlo, 2021, Hertweck et al., 2021].

Nevertheless, in order to incorporate fairness notions into algorithmic tools we require mathematical formalisations of fairness definitions. The three named criteria dominate most of the practical fair machine learning tools [Barocas et al., 2023, p. 45], presumably because their simple definitions make it easy to incorporate them in learning procedures in a pre-, in- or post-processing way [Barocas et al., 2023, Chapter 3, p. 20]. Regarding both the reductionist definition of fairness and the pragmatic justification, we emphasize that *our argumentation is solely with respect to the fairness criteria named above.*

The three fairness criteria are described as *group* fairness notions since each of the definitions is intrinsically relativized with respect to sensitive groups. The definition of sensitive groups

[6] Obviously, there are other fairness notions in machine learning which we have not listed above and which are not expressed as statistical independence, e.g., [Dwork et al., 2012, Kilbertus et al., 2017, Williamson and Menon, 2019].

substantially influences the notion of fairness. For instance, via custom categorization one can provide fairness by group-design (see [Menon and Williamson, 2018, Section H.3] for a detailed discussion of the question of choice of groups). In addition, the meaning of groups in a societal context influences the choice of groups as elaborated in [Hu and Kohler-Hausmann, 2020], and as explored in a long line of work in social psychology [Campbell, 1958, Tajfel, 1974, Hogg and Abrams, 1998, McGarty, 1999, Castano et al., 2002]. We contribute to the debate by drawing a connection between the choice of groups and the choice of randomness in §7.7.1.

7.3.2 Independence in mathematics and the world

Behind the formalization of fairness as statistical independence, there is an apparently rigid, mathematical definition of statistical independence. The fairness criteria presume that we have the machinery of probability theory at our disposal and the relationship of mathematics and the world is clear and unambiguous. However, as we elaborate further in the following, there is no single notion of "the" mathematical theory of probability [Fine, 1973]. Furthermore, it is not clear what it means to be statistically independent when talking of measurements in the world. Respectively, it is not obvious that the standard formulation of statistical independence is the right one to use.

The current definitions of fairness in machine learning fail and even hurt in practice, because debates on fairness notions take for granted a commonly agreed meaning of statistical independence. This common ground fatally does not exist given just the standard probability theory. Hence, given a specified scenario, e.g., hiring algorithm for public service in Kenya, fair machine learning research should enable founded debates on the meaning and purposefulness of deploying a specific notion of fairness, in particular those which require independence statements because their are algorithmically attractive. If the debate in Kenya would rely on everybody's - who are involved in the deployment process - intuitive understanding of statistical independence, it would result in a fatal misalignment of reasonings.

We perceive this work as part of the project to emphasize the substantive character formal notions of fairness should have. In other words, fairness is a societal and ethical concept that does not allow for the separation of machine learning as technical tool on one side and the idea of fairness in society on the other side. Debates on fairness have to consider machine learning *in* society.

Hence, if we desire statistical independence to capture a fairness notion applicable to the world, we ought to understand what the mathematical formulae signify in the world. Thus, in addition to the debate about the fairness criteria, a debate on the interpretation of statistical concepts in ethical context is required.

In this work, we contribute to the understanding by scrutinizing the standard definition of statistical independence. Motivated by the occurrence of statistical independence as fundamental ingredient in randomness as well as fairness in machine learning, we first detail the standard account due to Kolmogorov. What is statistical independence? How does statistical independence relate to an independence in the world?

7.4 Statistical Independence Revisited

To make any sense of phenomena in our world, we need to ignore large parts of it in order to avoid being overwhelmed. Hence, we usually assume or presume the phenomena of interest depends only on a few factors and to be *independent* of everything else [Naylor, 1981]. Thus the concept of independence is inherent in a variety of subjects ranging from causal reasoning [Spohn, 1980], to logic [Grädel and Väänänen, 2013], accounting [Cook, 2020], public law [Murchison, 1992] and many more. Independence, as we understand it, grasps the concept of incapability of an entity to be expressed, derived or deduced by something else [Simpson and Weine, 1989].

7.4.1 From Independence to Statistical Independence

Of special interest to us is the concept of independence in probability theories and statistics [Levin, 1980][Fine, 1973, Section IIF, IIIG and VH]. Independence in a probabilistic context should somehow capture the unrelatedness between the occurrence of events, as has been understood for centuries:

> Two Events are independent, when they have no connexion [sic] one with the other, and that the happening of one neither forwards nor obstructs the happening of the other. [De Moivre, 1738/1967, Introduction, p. 6].

Modern probability theory loosely follows this intuition as we see in the following.

7.4.2 Statistical Independence as we know it lacks semantics

Since the axiomatization of probability theory developed by Kolmogorov [1933] (translated in [Kolmogorov, 1956]), it displaced many other approaches. Mathematically, Kolmogorov's measure-theoretic axiomatization dominates all other mathematical formalizations to probability and related concepts.[7] In particular, his definition of statistical independence developed a well-accepted, ubiquitous notion. In a simple form it is given by[8]:

[7] Exceptions exist, e.g., in quantum theory [Gudder, 1979] or in statistics [Vovk, 1993].
[8] For a rigorous definition see Appendix 7.6.

Definition 7.1 (Simplified Statistical Independence following Kolmogorov)**.** *Two events A, B are statistically independent iff*

$$P(A \cap B) = P(A)P(B).$$

Independence plays a central role in Kolmogorov's probability theory: "measure theory ends and probability begins with the definition of independence" [Durrett, 2019, p. 37] (quoted in [Tao, 2015]). However, Kolmogorov's definition of independence is subtle and requires closer investigation.

We employ a small toy example in order to convey the semantic emptiness of Kolmogorov's definition: consider the experiment of throwing a die. The events under observation are $A = \{1, 2\}$, seeing one or two pips, respectively $B = \{2, 3\}$, seeing two or three pips. If the die were fair, so each face has equal probability $\frac{1}{6}$ to show up, the events A and B would turn out to not be independent. In contrast, if the die were loaded in a very special way $p_2 = p_3 = \frac{1}{2}, p_1 = p_4 = p_5 = p_6 = 0$, where p_i refers to the probability of seeing i pips, the events A and B would be independent following Kolmogorov's definition.

Thus, statistical independence, even though defined over events, manifests in the correspondence of how probabilities are mapped to events and why. The definition focuses on events. But, the crucial ingredient is the probability. Thus, there is no unique interpretation of statistical independence. A more detailed interpretation and meaning heavily depends on the interpretation of probability in the first place.

Given this observation, we may ask for a notion of independence in the world, which is somehow captured by Kolmogorov's definition. Kolmogorov himself underlined the avowedly mathematical, axiomatic nature of probability. His theory is in principle detached from any meaning of probabilistic concepts such as statistical independence in the world [Kolmogorov, 1956, p. 1]. He even questioned the validity of his axioms as reasonable descriptions of the world [Kolmogorov, 1956, p. 17].

However, one can possibly construct a notion of independence in world which is captured by Kolmogorov's definition. If one assumes one's calculations about one's beliefs on the happening of events are governed by the mathematical rules laid out by Kolmogorov, then Kolmogorov's definition of statistical independence captures one's (in-the-world) understanding of an independence of beliefs on the happening of events. This sketch of a purely subjectivist account to statistical independence neglects a justification for the choice of mathematical formulation and skips over any reference to an objective world. In conclusion, Kolmorogov's independence might capture a worldly concept. But, this independence in the world is not uniquely attached to Kolmogorov's definition.

There is a third major irritation arising from Kolmogorov's definition. As observed already

above, Kolmogorov treats events as the entities of independence. Though, against the background of De Moivre [1738/1967]'s intuition on statistical independence, we wonder about this focus. Statistical independence, as De Moivre [1738/1967] already emphasized, refers to altering "the happening of the event," but not the event itself. It is not the independence between the shown numbers of the die (whatever this means), but the independence of the processes how the number showed up (loading the die, throwing the die, etc.) which are captured by statistical independence.

This critique is not new. Von Mises already criticized the measure theoretical definition by Kolmogorov in his book [Von Mises and Geiringer, 1964, p. 36-39]. In summary, he argued that there is no interpretation of statistical independence of "single events." The unrelatedness, which the probabilistic notion of statistical independence is trying to capture, locates in the process of reoccuring events, but not single events themselves. More recently, Von Collani [2006] argued in a similar way. It is probability which brings the definition of statistical independence to life and it is the question what probability means and why we use it which links the mathematical definition to a concept in the real world.

In machine learning and statistics it is often presumed that the mathematical definition of independence captures a worldly concept. As we argued in this section, this link is far from being well-defined. However, if we consider machine learning as worldly data modeling, then the natural question arises: what do we model when we leverage Kolmogorov's statistical independence? What do we mean by independence of events? In order to circumvent these questions, we propose to look into another mathematical theory of probability. This theory was led by the idea of modeling statistical phenomena in the world.

7.4.3 Statistical Independence and a Probability Theory with inherent Semantics

Around 15 years before Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung* [Kolmogorov, 1933] (translated as [Kolmogorov, 1956]), von Mises proposed an earlier axiomatization of probability theory [Von Mises, 1919]. His less known theory approached the problem of a mathematical theory of probability through the lens of physics. Von Mises aimed for a "mathematical theory of repetitive events."

This aim included the emphasis on the link between real-world and mathematical idealization. In particular, he offers interpretability and verifiability of his theory [Von Mises and Geiringer, 1964, p. 1 and 14].[9] For interpretation he defined probabilities in a frequency-

[9] Von Mises' discussion pre-dates most of the modern work done in the philosophy of science. Popper [2010], for instance, was inspired by von Mises' conception of randomness and probability. While von Mises' definition of verifiability and interpretability is arguably somewhat ill-conceived, nevertheless, he works, in contrast to Kolmorogov's theory which is of purely syntactical nature, on a semantical project, where the connection of the real world and the mathematical formulation is of vital importance.

based way (see Definition 7.2). This inherently reflects the repetitive nature of the phenomena under description. By verifiability he referred to the ability to *approximately* verify the probabilistic statements made about the world [Von Mises and Geiringer, 1964, p. 45].

In summary, von Mises' theory, in our conception of machine learning, starts the "modeling of data generating processes" on an even more fundamental level then it is currently done via the use of Kolmogorov's axiomatization. His aim for a mathematical description of data-generating processes (the sequence of repetitive events) aligns to our perspective on machine learning as laid out earlier (cf. Section 7.2). With von Mises we obtain access to *meaningful* foundations for statistical concepts in machine learning. In particular, we redefine and reinterpret statistical independence in a von Misesean way. This suggests new perspectives on the problem of fair machine learning and the concepts of fairness and randomness themselves.

For the further discussion, we summarize the major ingredients of Von Mises' theory. Fortunately, it turns out that von Mises' notion of statistical independence, central to our discussion, is mathematically analogous to the well-known Kolmogorovian definition. Thus, one's intuition on statistical independence is refined but its mathematical applicability remains.

7.4.4 Von Mises' theory of Probability and Randomness in a Nutshell

Von Mises' axiomatization of probability theory is based on random sequences of events and the interpretation of probability as the limiting frequency that an event occurs in such a sequence [Von Mises, 1919]. These random sequences, called collectives, are the main ingredients of his theory. For the sake of simplicity, we stick to binary collectives. Thus, collectives are 0-1-sequences with certain randomness properties which define probabilities for the labels 0 and 1. Nevertheless, it is possible to define collectives respectively probabilities on richer label sets. Collectives can be extended up to a continuum [Von Mises and Geiringer, 1964, II.B].

For notational economy, we note that a sequence taking values in $\{0,1\}$, $(s_i)_{i\in\mathbb{N}}$ can be identified with a *function* $s\colon \mathbb{N} \to \{0,1\}$.

Definition 7.2 (Collective [Von Mises and Geiringer, 1964, p. 12])**.** *Let $\mathcal{S}$ be a set of sequences $s\colon \mathbb{N} \to \{0,1\}$ with $s(j) = 1$ for infinitely many $j \in \mathbb{N}$. In mathematical terms,* collectives *with respect to $\mathcal{S}$ are sequences $x\colon \mathbb{N} \to \{0,1\}$ for which the following two conditions hold.*

1. *The limit of relative frequencies of* 1*s,*

$$\lim_{n\to\infty} \frac{|\{i \in \mathbb{N}\colon x(i) = 1, 1 \leq i \leq n\}|}{n}$$

exists.[10] *If it exists, then the limit of relative frequencies of* 0*s exists, too. We define* p_1*, respectively* $p_0 = 1 - p_1$*, to be its value.*

2. *For all* $s \in \mathcal{S}$*,*

$$\lim_{n\to\infty} \frac{|\{i \in \mathbb{N}\colon x(i) = 1 \text{ and } s(i) = 1, 1 \leq i \leq n\}|}{|\{j \in \mathbb{N}\colon s(j) = 1, 1 \leq j \leq n\}|} = \lim_{n\to\infty} \frac{|\{i \in \mathbb{N}\colon x(i) = 1, 1 \leq i \leq n\}|}{n} = p_1.$$

We call p_0 the *probability* of label 0. Conversely, p_1 is the probability of label 1.[11] The existence of the limit (Condition 1) is a non-vacuous condition. One can easily construct sequences whose frequencies do not converge [Fröhlich et al., 2023].

The sequences $s \in \mathcal{S}$ are called *selection rules*. A selection rule selects the jth element of x whenever $s(j) = 1$.[12] Informally, a collective (w.r.t $\mathcal{S}$) is a sequence which has invariant frequency limits with respect to all selection rules in $\mathcal{S}$. We call any selection rule which does not change the frequency limit of a collective *admissible*. This invariance property of collectives is often called "law of excluded gambling strategy" [Von Mises and Geiringer, 1964]. When thinking of a sequence of coin tosses, a gambler is not able to gain an advantage by just considering specific selected coin tosses. The probability of seeing "heads" or "tails" remains unchanged.

Von Mises introduced the "law of excluded gambling strategy" with the goal to define randomness of a collective [Von Mises and Geiringer, 1964, p. 8]. A collective is called *random* with respect to $\mathcal{S}$. Consequently, von Mises integrated randomness and probability into one theory. In fact, admissibility of selection rules is equivalent to statistical independence in the sense of von Mises. But it is defined with respect to collectives instead of selection rules.

Definition 7.3 (Von Mises' Definition of Statistical Independence of Collectives [Von Mises and Geiringer, 1964, p. 30 Def. 2])**.** *A collective* x *with respect to* $\mathcal{S}_x$ *is called statistically*

[10] The mathematically inclined reader might notice the requirement for an order structure to obtain a definition of limit here. We use x as sequences with the standard order structure on the natural numbers. However, x can be generalized to be a net on more arbitrary base sets [Ivanenko, 2010] .

[11] Von Mises called p_0 chance as long as x is a sequence. When x is a collective, he called it probability.

[12] We sweep under the carpet a substantial difference in von Mises' definition of selection rules and our definition. Von Mises allowed the selection rules to "see" the first n elements of the collective when deciding whether to choose the $n + 1$th element [Von Mises and Geiringer, 1964, p. 9]. Our definition is more restrictive. We require the selected position to be determined before "seeing" the *entire* sequence. We focus on the ex ante nature of von Mises randomness but neglect his recursive formalism.

independent to the collective y with respect to $\mathcal{S}_y$ iff the following limits exist and

$$\lim_{n\to\infty} \frac{|\{i \in \mathbb{N}\colon x(i) = 1 \text{ and } y(i) = 1, 1 \leq i \leq n\}|}{|\{j \in \mathbb{N}\colon y(j) = 1, 1 \leq j \leq n\}|}$$
$$= \lim_{n\to\infty} \frac{|\{i \in \mathbb{N}\colon x(i) = 1, 1 \leq i \leq n\}|}{n} = p_1,$$

where $\frac{0}{0} := 0$. *When two collectives are independent of each other we write*

$$x \perp y.$$

In comparison to admissibility, the collective y adopts the role of a selection rule. It *is* in fact an admissible selection rule with the difference that a potentially finite number of elements in x are selected (cf. [Gillies, 2010, p. 120f]). Conversely, von Mises' randomness is statistical independence with respect to sequences with infinitely many ones and potentially no frequency limit. (For a general comparison between Kolmogorov's and von Mises' theory of probability see Appendix 7.9.1 and Table 7.3.)

7.4.5 Kolmogorov's Independence versus von Mises' Independence

What is the relationship between Kolmorogov's and von Mises' definition of statistical independence? On a conceptual level, the critique posed earlier, which questioned the meaning of statistical independence between events following Kolmogorov, gets resolved.

Von Mises adopted a strong frequential perspective on probabilities which clarifies the mapping from real world to mathematical definition. He idealized repetitive observations by infinite sequences and defined probabilities as limiting frequencies.[13] Von Mises' independence states that there is no difference in counting the frequency of occurrences of an event in the entirety of the sequences or in a subselected sequence. His independence forbids any statistical interference between processes described as sequences. No statistical patterns can be derived from one sequence by leveraging the other. Von Mises' definition formalizes the concept of statistical independence between processes of reoccuring events.

In contrast to Kolmogorov, von Mises' definition does not evoke the conceptual obscurity. His focus on idealized sequences of repetitive events restricts his definition of statistical independence to specific applications with the gain of clarity in the goal of the mathematical description. Von Mises' definition of independence makes statistical independence more concrete then Kolmogorov's definition does.

[13] Without the idealization, again the mathematical description would miss a link to a worldly phenomenon. The idealization in terms of infinite sequences is substantial. In fact, the legitimacy of this idealization is the subject of another debate [Hájek, 2009]. Nevertheless, the idealization taken in von Mises' framework is explicitly and transparently stated. Kolmogorov's axioms do not possess such a statement.

On a more formal level, Kolmogorov defined statistical independence via the factorization of measure (cf. Definition 7.6), whereas von Mises defined statistical independence via conditionalization of measures. The invariance of the frequency limit of a collective with regard to the subselection via another collective can be interpreted as the invariance of a probability of an event with regard to the conditioning on another event, i.e., "selecting with respect to" is "conditioning on" (cf. Theorem 7.9 and Theorem 7.10).

Mathematically, it turns out that Kolmogorov's definition and von Mises' definition are both special cases (modulo the measure zero problem in conditioning) of a more general form of measure-theoretic statistical independence. A selection rule with converging frequency limit is admissible (respectively, statistically independent), to a collective if and only if the two are statistically independent of each other in the sense of Kolmogorov, when generalized to finitely additive probability spaces (see Appendix 7.9.1 for a formal statement of this claim). Thus, we can replace the known definition of statistical independence by Kolmogorov with the definition by von Mises. Thereby, we give a specific meaning to statistical independence.

We have been motivated to dissect the notion of statistical independence for its central role in fair machine learning. Von Mises' definition drew us closer to a more transparent mathematical formalization of statistical independence for fairness notions in machine learning. However, our discussion of von Mises' theory skipped over a substantial part of his work so far. Von Mises included a definition of randomness in his theory of probability. Much in contrast to Kolmogorov: There is no definition of "randomness" in Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung* [Kolmogorov, 1933] (translated as [Kolmogorov, 1956]). Even more interestingly, von Mises' definition of randomness is stated in terms of statistical independence. The reader might notice that in Section 7.2 we already stumbled upon a heavily used notion of randomness in machine learning, which is expressed as statistical independence (i.i.d.). How do i.i.d. and von Mises' randomness relate to each other? How does the close connection between statistical independence and randomness complement our picture of the three fairness criteria from machine learning?

7.5 Randomness as Statistical Independence

The nature and definition of randomness seems as "random" as the term itself [Eagle, 2021, Ornstein, 1989, Muchnik et al., 1998, Volchan, 2002, Berkovitz et al., 2006]. Usually, a very broad distinction between two approaches to randomness is made: process randomness versus outcome randomness [Eagle, 2021]. In this work, we focus on outcome randomness and more specifically the role of randomness in statistics and machine learning.

Randomness is a modeling assumption in statistics (cf. Section 7.2). Upon looking into statistics and machine learning textbooks one often finds the assumption of independent

and identically distributed (i.i.d.) data points as the expression of randomness [Casella and Berger, 2002, p. 207] [Devroye et al., 1996, p. 4].

We adopt von Mises' differing account of randomness. The expression of randomness *relative* to the problem at hand, particularly in settings with data models such as statistics, turns out to be substantial.

7.5.1 Orthogonal Perspectives on Randomness as Independence in Machine Learning and Statistics

Von Mises defined a random sequence as a sequence which is statistically independent to a (pre-specified) set of selection rules respectively other sequences. In contrast, an i.i.d.-sequence consists of elements each statistically independent to all others.

Both definitions are stated in terms of statistical independence. But, the relationship of independence and randomness in terms of i.i.d. and in von Mises' theory differ substantially. Von Mises' randomness is stated relative with respect to a set of selection rules. Furthermore, it is stated between sequences, respectively collectives. Whereas, in an i.i.d. sequence randomness is expressed between random variables. The randomness definitions are in an abstract sense "orthogonal." We consider a concrete example for better understanding.

Horizontal Randomness. Let $\Omega = \mathbb{N}$ be a penguin colony. Let s, f be two attributes of a penguin, namely sex and whether a penguin has the penguin flu or not. Mathematically: $s\colon \Omega \to \{0,1\}$, $f\colon \Omega \to \{0,1\}$. So, penguins are individuals $n \in \Omega$ which we do not know individually, but we know some attributes of them. Suppose we are given a sequence $f(1), f(2), f(3), \ldots$ of flu values with existing frequency limit. This allows us to state randomness of f with respect to the corresponding sequence of sex values s, containing infinitely many ones and having a frequency limit, by: the sequence of sex values s is admissible on f. Respectively, s and f are statistically independent of each other. In the context of colony Ω a penguin having flu is random with respect to the sex of the penguin.

Vertical Randomness. This is different to the i.i.d.-setting in which each penguin $i \in \mathbb{N}$ obtains its own random variable $F_i\colon \Omega \to \{0,1\}$ on some probability space $(\Omega, \mathcal{F}, P)$. Here, F_i encodes whether penguin i has the penguin flu or not. The sequence $F_1, F_2, F_3, \ldots$ somehow represents the colony. The included random variables share their distribution and are statistically independent to each other. The attribute flu is not random with respect to the attribute sex here, but the penguins are random with respect to each other. The random variables are (often implicitly) defined on a standard probability space on Ω. The set Ω here does *not* model the colony. It shrivels to an abstract source of randomness and probability.

The choice of perspective, horizontal or vertical, on randomness expressed as statistical independence is a question of the data model. The two types of randomness definitions are distinct in a number of ways. For a summary see Table 7.2. Most importantly, horizontal randomness is inherently expressed *with respect to* some mathematical object. Vertical randomness lacks this explicit relativization. This typification of horizontal and vertical, mathematical definitions of randomness is actually more broadly applicable.

	Horizontal Randomness	**Vertical Randomness**
Data points are modelled as:	Evaluations of RVs	RVs
Math. definition of randomness of:	Sequences	Sequences of RVs
Explicit relativization:	Yes	No

Table 7.2: Typification of horizontal and vertical randomness ("RV"="random variable").

To the set of vertical randomness notions one can add: exchangeability [De Finetti, 1929], α-mixing, β-mixing [Steinwart et al., 2009] and possibly many more. The set of horizontal randomness notions is spanned up by an entire branch of computer science and mathematics: algorithmic randomness.

Algorithmic randomness poses the question whether a sequence is random or not. This question arose in [Von Mises, 1919] within the attempt to axiomatize probability theory [Bienvenu et al., 2009, p. 3]. In algorithmic randomness further definitions of random sequences have been proposed. For the sake of simplicity the considered sequences consist only of zeros and ones.

Four intuitions for random sequences crystallized [Porter, 2012, pp.280ff]: typicality, incompressibility, unpredictability and independence (see Appendix 7.9.3). For our purposes, the key point to note is that a random sequence is typical, incompressible, unpredictable or independent with respect to "something" (they are all relativised in some way). Each of these intuitions has been expressed in various mathematical terms. In particular, formalizations of the same intuitions are not necessarily equivalent, and formalizations of different intuitions sometimes coincide or are logically related (see Appendix 7.9.4).[14] We mainly stick to the intuition of independence in this paper. A random sequence is independent of "some" other sequences [Von Mises, 1919, Church, 1940].

7.5.2 Relative Randomness Instead of Absolute, Universal Randomness

The definition of randomness for sequences is inherently relative. Even though, the notion is relative with respect to "something," most of the effort has been spent on finding *the* set

[14]For an overview of algorithmic randomness see [Uspenskii et al., 1990, Muchnik et al., 1998, Volchan, 2002, Downey and Hirschfeldt, 2010].

of statistically independent sequences defining randomness [Church, 1940, Muchnik et al., 1998].[15]

Naively, one could attempt to define a random sequence as: a sequence is random if and only if it is independent with respect to *all* sequences. However, this approach is doomed to fail. There is no sequence fulfilling this condition except for trivial ones such as endless repetitions of zeros or ones (see Kamke's critique of von Mises' notion of randomness [Van Lambalgen, 1987]).

So instead, research focused on computability expressed in various ways (because it was felt by those investigating these matters that computability was somehow given, or more primitive, and thus a natural way to resolve the relativity of the notion of randomness). Intuitively, randomness is considered the antithesis of computability [Porter, 2012, p. 288]: something which is computable is not random. Something which is random is not computable. If we then informally update the definition above we obtain: a sequence is random if and only if it is independent with respect to all computable sequences [Church, 1940][16]. Analogous to *the* definition of computability [Porter, 2012, p. 165], this is taken as an argument for the existence of *the* definition of randomness [Porter, 2012, p. 287].

In our work, we argue towards a relativized conception of randomness in line with work by Porter [2012], Humphreys [1977] and Von Mises and Geiringer [1964][17]. A *relative* definition of randomness is a definition of randomness which is *relative* with respect to the problem under consideration.[18] In contrast, an *absolute and universal* definition of randomness would preserve its validity in all problems. It presupposes the existence of *the* randomness.

Relative randomness with respect to the problem which we want to describe aligns to von Mises' theory of probability and randomness. Von Mises emphasized the *ex ante* choice of randomness [Von Mises, 1981, p. 89] *relative* to the problem at hand [Von Mises and Geiringer, 1964, p. 12]. First, one formalizes randomness with respect to the underlying problem, then one can consider a sequence to be random or not. Otherwise, if we are given a sequence, it is easy to construct a set of selection rules, such that the sequence is random with respect to this set.[19] This, however, undermines the concept of randomness, which should capture the pre-existing typicality, incompressibility, unpredictability or independence of

[15] The analogous observation holds for all four intuitions (see Appendix 7.9.3 and [Muchnik et al., 1998]).

[16] To be precise, a random sequence following Church [1940] is independent to all *partially computable* selection rules following von Mises (see Footnote 12).

[17] Humphreys [1977] presented randomness as relativized to a probabilistic hypothesis or reference class. Porter [2012, p. 169] even postulated the "No-Thesis"-Thesis: any notion of randomness neither defines a well-defined collection of random sequences nor captures all mathematical conceptions of randomness; confer the logic of "essentially contested" concepts [Gallie, 1955], which, presumably, are unavoidably contested for the same reason.

[18] In machine learning, a problem under consideration is, for instance, animal classification via neural networks.

[19] This idea is transferable to other intuitions of randomness, for instance [Ville, 1939].

a sequence (cf. [Vovk, 1993, p. 321]). Von Mises' randomness intrinsically possesses a modeling character, similar to our needs in machine learning and statistics.

Given its role as modeling assumption in statistics, randomness lacks substantial justification to be expressed in any *absolute and universal* manner in this context. Neither are there reasons why computability[20] is the only mathematical, expressive way to encode one of the four intuitions of randomness. The i.i.d. assumption, an *absolute and universal* definition of randomness, does not fit this purpose. To appropriately model data we require adjustable notions of randomness. Otherwise, we restrict our modeling choice without reason or gain.[21]

Equipped with the interpretation of statistical independence as randomness we now return to our motivation for investigating statistical independence. ML-friendly fairness criteria are built upon statistical independence. In contrast to Kolmogorov, von Mises' statistical independence transparently refers to a concept of independence in the real world. To clarify the meaning of fairness expressed as statistical independence, we directly apply von Mises' independence to the fairness criteria listed in Section 7.3 in the following.

7.6 Von Mises' Fairness

With von Mises' definition of statistical independence we have a notion at our disposal which is conceptually focused on a more "scientific" perspective (i.e., making claims about the world) of statistical concepts. Since it is mathematically related to Kolmogorov's standard account of statistical independence, Kolmogorov's definition can, at many places, be easily replaced by von Mises' definition.

Let us denote the three presented fairness criteria in a von Mises' way (cf. Section 7.4.5).

Definition 7.4 (Fairness as Statistical Independence)**.** *A collective* $x\colon \mathbb{N} \to \{0,1\}$ *(with respect to a family of selection rules* $\mathcal{S}$*) is* fair with respect to a set of sensitive groups $\mathcal{G} = \{s^j\colon \mathbb{N} \to \{0,1\} | j \in J\}$ *if*

$$x \perp s^j \quad \forall j \in J$$

The 0-1-sequences s^j determine for each individual i whether it is part of the group or not (according to whether $s^j(i) = 1$ or $s^j(i) = 0$). We call these groups "sensitive," as these are the groups which are of moral and ethical concern. In philosophical literature these groups

[20] Computability is often taken (at least by computer scientists) as a purely mathematical notion, detached from the world. An alternate view, close in spirit to von Mises, is that computation is part of physics, and thus needs to be viewed in a *scientific*, and not merely *mathematical* manner [Deutsch, 2011].

[21] This is not entirely true, as specific computability notions of randomness and the i.i.d.-assumption deliver convergence results for random sequences, which can be used to guarantee low estimation error in the long run.

are often called "socially salient groups" [Altman, 2020, Loi et al., 2019].[22] We see that the connection between von Misesean independence and fairness arises from the observation that the set of sensitive groups $\mathcal{G}$ *is* a family of selection rules, so that if $\mathcal{G} \subseteq \mathcal{S}$, then indeed the collective x will be fair for $\mathcal{G}$.

Following von Mises' interpretation of independence, the given definition reads as follows: we assume we are in the idealized setting of infinitely many individuals with values x_i, e.g., binary predictions. The predictions are fair if and only if there is no difference in counting the frequency of 1-predictions in the entirety or in the sensitive group. (For an illustration see Appendix 7.6.) A proper conceptualization of fairness requires such immediate semantics, but a purely mathematical theory of probability cannot offer these (see Section 7.4.2).

Each of the three fairness criteria is captured in Definition 7.4; the choice of fairness criterion manifests in the collective under consideration:

Independence The collective $x\colon \mathbb{N} \to \{0,1\}$ consists of predictions; i.e., $\{0,1\}$ is the set of predictions.

Separation The collective $x\colon \mathbb{N} \to \{0,1\}$ is obtained via the subselection of predictions based on the sequence of true labels corresponding to the predictions.[23]

Sufficiency The true labels are subselected by predictions.

To enable intuitive access to the von Mises' notions of fairness we provide a toy example in Appendix 7.9.2. The three fairness criteria Independence, Separation and Sufficiency encompass a large part of fair machine learning [Barocas et al., 2023, p. 45]. Von Mises' statistical independence gives a consistent interpretation to all of them. In fact, von Mises' independence opens the door to further investigations. To this end, we we recapitulate the strong linkage between statistical independence and randomness in von Mises' theory.

7.7 The Ethical Implications of Modeling Assumptions

Machine learning methods try to model data in complex ways. Derived statements, such as predictions, then potentially get applied in society. In these cases one is obliged to ask which ought-state the machine learning model should reflect [Passi and Barocas, 2019, Räz, 2021]. To enable a justified choice, statistical concepts in machine learning require relations to the real world. Furthermore, modeling even requires understanding of the entanglement of societal and statistical concepts.

[22] Intersections of sensitive groups are not necessarily independent.
[23] "Selecting with respect to" is "conditioning on."

We proposed one specific *meaningful* definition of statistical independence which can be directly applied to the three observational fairness criteria from fair machine learning. In addition, this von Mises' independence is key to a relativized notion of randomness. Pulling these threads together, we are now able to establish the following link: *Randomness is fairness. Fairness is randomness.*

7.7.1 Randomness is Fairness. Fairness is Randomness.

The concepts fairness and randomness frequently appear jointly: Broome [1984] argues that a random allocation of goods is fair under certain conditions. Literature on sortition argues for just representation of society by random selection of people [Parker, 2011, Stone, 2011].[24] Bennett [2011, p. 633] even states that randomness encompasses fairness.

With von Mises' axiom 2 and Definition 7.4 we can now tighten the conceptual relationship of fairness and randomness. The proposition directly follows from the definition of randomness respectively fairness in the sense of von Mises.

Proposition 7.5 (Randomness is fairness. Fairness is randomness.)**.** *Let x be a collective with respect to $\emptyset$ (the empty set). It is fair with respect to a set of sensitive groups (0-1-sequences) $\{s^j\}_{j\in J}$, if and only if it is is random with respect to $\{s^j\}_{j\in J}$.*

The given proposition establishes a helpful link. It gives insights into both of the concepts. In particular, it substantiates the relativized conception of randomness in machine learning as it presents randomness as an ethical choice.

Randomness as Ethical Choice

Randomness in machine learning is a modeling assumption (Section 7.2). Fairness is an ethical choice.[25] In light of Proposition 7.5 randomness gets an ethical choice and fairness a modeling assumption. We now further detail this perspective.

We assume that we are given a fixed set of selection rules, which defines "the" randomness. As far-fetched as this may sound, if we, for example, accept the so called Martin-Löf randomness as *absolute and universal* definition, then we exactly do this and fix the set of selection rules to the partial computable ones (see Appendix 7.9.4). A sequence which is random with respect to this specified set of selection rules is fair with respect to the groups defined by the selection rules. Rephrased in terms of Martin-Löf randomness: a Martin-Löf random sequence is fair with respect to all partial computable groups. Only non-partial-computable groups (respectively sequences) can be discriminated against in this setting.

[24] Representativity here can be interpreted as typicality, thus one of the four intuitions for randomness.

[25] More specifically, the choice of operationalized fairness, one of the fairness criteria, and the choice of groups.

If we interpret statistical independence as fairness (Section 7.3), then fairness is as *absolute and universal* as randomness here. Where did the "essentially contested" nature of fairness [Gallie, 1955] leave the picture?

The set of admissible selection rules specifies the choice of sensitive groups, which indeed is a fraught and contestable choice [Menon and Williamson, 2018, Section H.3]. Thus each selection rule gets ethically loaded. Furthermore, the choice of collective, which we consider as random, fixes the fairness criterion. In summary, the determination of randomness is analogous to the determination of fairness.

However one defines randomness, it is an ethical choice. For symmetry reasons one can equivalently state in machine learning: fairness is a modeling assumption. The randomness assumption has an ethical, moral and potentially legal implication. We need non-mathematical, contextual arguments to each problem at hand which justify the adjustable and explicit randomness assumptions.

Given that randomness is an ethical choice, an absolute, universal conception of randomness counteracts any ethical debate in machine learning. Discussions about sexism, racism and other kinds of discrimination and injustice persist over time without ever arrogating the discovery of "the" fairness [Gallie, 1955]. But if "the" randomness as statistical independence would exist, then "the" fairness as statistical independence would be an accessible notion. For illustration, we reconsider Martin-Löf randomness. A Martin-Löf random sequence is independent, respectively fair, to the set of all partial computable selection rules. But, it is completely unclear what the ethical meaning of partial computable groups is. And, it remains unsolved whether the groups given by gender are partial computable, when we desire to be fair with respect to them. We conceive Proposition 7.5 as further counterargument to an absolute, universal definition of randomness. Randomness is, like fairness, better interpreted as a *relative* notion.

Further concluding, the equivalence of randomness and fairness highlights the deficiency of fairness notions in machine learning. The equivalence only holds due to the very reductionist perspective on fairness in fair machine learning. Despite their regular co-occurrence [Broome, 1984, Parker, 2011] [Bennett, 2011, p. 633], fairness and randomness are more multi-facetted and non-overlapping concepts as illustrated in Proposition 7.5.

Fairy Tales of Fairness: "Perfectly Fair" Data

With the relationship between fairness and randomness in mind, we now turn towards random data as primitive. Discussions in fair machine learning sometimes seemingly presume the existence of "perfectly fair" data (e.g., as highlighted in [Räz, 2021, p. 134]), as if fair machine learning merely tackles the cases where "perfectly fair" data is not available.

We interpret "perfectly fair" data as a collective with respect to all possible selection rules. The data does not depend on any (sensitive) group at all. In other words, "perfectly fair" data is "totally random" data. As we saw in Section 7.5.2 this is self-contradicting except of the trivial constant case. "Perfectly fair" data does not exist or is statistically useless.

7.7.2 Demanding Fairness is Randomization: Fair Predictors are Randomizers

In practice, it is often unreasonable to *assume* random or fair data as in Proposition 7.5. Instead one *demands* for fairness respectively randomness of predictions. In these settings, fair machine learning techniques are deployed to exhibit ex post fulfillment of fairness criteria.

We assume for the following discussion that the collective x consists of predictions, as in the fairness criteria Independence or Separation. Fair machine learning techniques enforce statistical independence of predictions and sensitive attributes. Rephrased, fair machine learning techniques actually introduce randomness post-hoc into the predictions. Thus, fair machine learning techniques can potentially be interpreted as randomization techniques.

Fairness-Accuracy Trade-Off — Another Perspective

We noticed that fair predictions are random predictions with respect to the sensitive attribute. In contrast, accurate predictions exploit all dependencies between given attributes and predictive goal, including the sensitive attributes. Thus, in fair machine learning morally wrongful discriminative potential of sensitive attributes is thrown away by purpose. On these grounds, it is not surprising that an increase in fairness respectively randomness (usually) goes hand in hand with a decrease in accuracy [Wick et al., 2020]. Randomization of predictions leads to the so called fairness-accuracy trade-off.

Concluding, via von Mises' axiomatization we established: *Randomness is fairness. Fairness is randomness.* Exploiting this new perspective, we unlock another perspective on fair predictors as randomizers, demonstrate the nonexistence of "perfectly fair" data and treat randomness as an ethical choice, which can be neither universal nor total. In particular, the "essentially contested" nature of fairness is tied to the "essentially *relative*" nature of randomness.

7.8 Conclusion

Fair machine learning attained an increasing interest in the last years. However, its conceptual maturity lags behind. In particular, the interplay between data, its mathematical

representation and their relation to fairness is encompassed by a veil of nescience. In this paper, we contribute towards a better understanding of randomness and fairness in machine learning.

We started from the most commonly used definition of statistical independence and questioned its representation due to a lack of semantics. Generally, we observe that in machine learning, as in statistics, probability and its related concepts should be interpreted as modeling assumptions about the world (of data). Von Mises aimed for exactly this "scientific" perspective on probability theory. We lean on his statistical independence, which clarifies the relation to the real world, and his definition of randomness, which is *relative* and orthogonal to the i.i.d. assumption, but similarly expressed as statistical independence. Then by the three fairness criteria in machine learning we obtain a further interpretation of independence, which we finally exploit to argue for a *relative* conception of randomness, randomness as an ethical choice in machine learning and fair predictors as randomizers.

7.8.1 Future Work: Approximate Randomness and Fairness, Randomness as Fairness via Calibration

Despite future conclusions in-between the topics fairness and randomness in other research subjects as machine learning, we claim that a significant dimension is missing in the present discussion. Practitioners usually deal with approximate versions of randomness, statistical independence or fairness. Yet, approximation spans another dimension of choice beset with pitfalls [Putnam, 1990/2011, Lohaus et al., 2020]. Several questions ranging from the choice of approximation to the interference of concepts arise. Future work should detail the implications of this choice.

Second, we conjecture that "*Randomness is Fairness. Fairness is randomness.*" can be substantiated via the intuition of unpredictability. Starting from [Shafer and Vovk, 2019] definition of unpredictability randomness, which is closely related to the calibration idea presented in [Dawid, 2017], we can bridge to fairness as calibration as given in [Chouldechova, 2017]. A recent work by Cynthia Dwork and collaborators in fact show a formal link between pseudo-randomness and fairness as calibration [Dwork et al., 2023]. This work, however, still misses a more thorough discussion of the concepts of individual versus group fairness in machine learning [Binns, 2020]. As a subproblem, which is contained therein, the categorization into (sensitive) groups in fair machine learning deserves its own work.

Third, regarding a more thorough definition of statistical independence within the fairness criteria, we are convinced that a subjectivist interpretation of probability might reveal yet another perspective on the problem. We assume that the interplay between different interpretations of probability and ethical concepts such as fairness still leaves room for many important investigations.

Fourth, there are certainly more frameworks to give an interpretation and concretization to current notions in (fair) machine learning (cf. [Gillies, 2010, Fine, 1973]).

Last but not least, we already referred to sortition literature and random allocation. The somewhat different relation between fairness and randomness in this literature leads us to speculate that further fruitful discussions between the two concepts may develop.

In the jungle of statistical concepts such as probability, uncertainty, randomness, independence etc. further relations to social and ethical concepts wait to be brought to light. And machine learning research should care:

> The arguments that justify inference from a sample to a population should explicitly refer to the variety of non-mathematical considerations involved. [Battersby, 2003, p. 11]

Acknowledgements

Many thanks to Christian Fröhlich, Eric Raidl, Sebastian Zezulka, Thomas Grote and Benedikt Höltgen for helpful discussions and feedback. Futhermore, the authors thank all participants of the Philosophy of Science meets Machine Learning Conference 2022 in Tübingen, for all helpful comments and debates.

The authors appreciate and thank the International Max Planck Research School for Intelligent System (IMPRS-IS) for supporting Rabanus Derr.

Funding

This work was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645; it was also supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center.

7.9 Appendix

7.9.1 Generalized von Misesean Probability Theory

In this appendix we outline a theory of probability subsuming that of Kolmogorov and von Mises.

Kolmogorov's notion of Independence

Kolmogorov axiomatized probability theory in his book [Kolmogorov, 1956] in a measure theoretical way. He defined a probability space $(\Omega, \mathcal{F}, P)$ as a measure space with base set Ω, σ-algebra $\mathcal{F}$ and normalized measure P. Events are elements of $\mathcal{F}$, i.e., subsets of Ω, which obtain a probability via P. Statistical independence is defined as a specific assignment of probability to an intersection event.

Definition 7.6 (Kolmogorov's Definition of Statistical Independence of Events [Kolmogorov, 1956, p. 9 Def. 1])**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two events $A, B \in \mathcal{F}$ are called statistically independent iff*

$$P(A \cap B) = P(A)P(B).$$

As we highlighted above, Kolmogorov's axiomatization is, despite its success, not the only mathematical theory of probability. Specifically, one can weaken the structure of the probability space and still work with concepts such as statistical independence, expectation, conditioning (e.g., [Gudder, 1969, Chichilnisky, 2010, Narens, 2016]).

Finitely Additive Probability Space

We introduce a weaker measure structure, which we call finitely additive probability space. Interestingly, this weaker structure includes the axiomatization of Kolmogorov and von Mises as special cases. We define a finitely additive probability space modified from [Rao and Rao, 1983, Def. 2.1.1 (7)]) as

Definition 7.7 (Finitely Additive Probability Space)**.** *The tuple $(N, \mathcal{A}, \nu)$ is called a* finitely additive probability space *for a base set N, a set of measurable sets $\mathcal{A} \subset \mathcal{P}(N)$ containing the empty set $\emptyset \in \mathcal{A}$ and a finitely additive probability measure $\nu : \mathcal{A} \to [0, 1]$ satisfying the following conditions:*
(1) $\nu(\emptyset) = 0$,
(2) if $A_1, A_2, A_1 \cup A_2 \in \mathcal{A}$ and $A_1 \cap A_2 = \emptyset$ then $\nu(A_1 \cup A_2) = \nu(A_1) + \nu(A_2)$.

Observe that this definition does not impose any structural restrictions on the set of subsets $\mathcal{A}$.

Kolmogorov's probability space is certainly a specific finitely additive probability space in our sense, as every set σ-algebra contains the empty set and every countably additive probability is finitely additive.

Analogously, von Mises implicitly uses a finitely additive probability space. This space is given by $(\mathbb{N}, \mathcal{A}_{\text{vM}}, \nu_{\text{vM}})$, where $\mathbb{N}$ are the natural numbers and $\mathcal{A}_{\text{vM}}, \nu_{\text{vM}}$ are defined in

the following.

First, we consider the finitely additive base measure

$$\nu_{\text{vM}}(A) := \lim_{n\to\infty} \frac{|A \cap \mathbb{N}_n|}{n},$$

where $A \subset \mathbb{N}$ and $\mathbb{N}_n = \{1, ..., n\}$. ν_{vM} is called the "natural density" in the number theory literature [Nathanson, 2008, p. 256]. From this definition is not clear whether the given limit exists. Thus, we define the set of measurable sets by

$$\mathcal{A}_{\text{vM}} := \{A : \nu_{\text{vM}}(A) \text{ exists}\},$$

which is called the "density logic" in [Pták, 2000]. It is a pre-Dynkin-system [Schurz and Leitgeb, 2008].

We generalize Kolmogorov's definition of statistical independence of events to finitely additive probability spaces.

Definition 7.8 (Statistical Independence of Events on a Finitely Additive Probability Space)**.** *Let $(N, \mathcal{A}, \nu)$ be a finitely additive probability space. Two measurable sets $A, B \in \mathcal{A}$ are independent iff*

1. $A \cap B \in \mathcal{A}$
2. $\nu(A \cap B) = \nu(A)\nu(B)$

Observe that the first condition is naturally fulfilled in Kolmogorov's σ-algebra. In the case of von Mises' density logic, this constraint is strict. A pre-Dynkin-system is not closed under arbitrary intersections.

Von Mises' Admissibility and Kolmogorov's Independence are Analogues

We now reconsider the definition of collectives and selection rules. Both are $0, 1$-sequences on the natural numbers, with the restriction that selection rules contain infinitely many 1's and their frequency limit does not have to exist. Both sequences can be interpreted as indicator functions on the natural numbers. So for $x\colon \mathbb{N} \to \{0, 1\}$ and $s\colon \mathbb{N} \to \{0, 1\}$ we write $X = x^{-1}(1)$ and $S = s^{-1}(1)$ for the corresponding subsets of the natural numbers.

We want to show that von Mises' admissibility condition is equivalent to the given definition of statistical independence on finitely additive probability spaces. Actually, the equivalence only holds for the slightly restricted case in which the selection rule itself possesses converging frequencies.

Theorem 7.9 (Admissibility in von Mises setting implies Statistical Independence on Finitely Additive Probability Spaces)**.** *Let x be a collective with respect to s. Suppose furthermore that s has a converging frequency limit. Then X and S, the indexed sets corresponding to the collective (respectively selection rule) on the finitely additive probability space $(\mathbb{N}, \mathcal{A}_{\mathrm{vM}}, \nu_{\mathrm{vM}})$ are statistically independent.*

Proof. We know that

$$\begin{aligned}&\lim_{n\to\infty} \frac{|\{i : x(i) = 1 \text{ and } s(i) = 1, 1 \leq i \leq n\}|}{|\{j : s(j) = 1, 1 \leq j \leq n\}|}\\ &= \lim_{n\to\infty} \frac{|\{i : x(i) = 1, 1 \leq i \leq n\}|}{n}\end{aligned}$$

which we can rewrite as

$$\lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{|S \cap \mathbb{N}_n|} = \lim_{n\to\infty} \frac{|X \cap \mathbb{N}_n|}{n}.$$

This gives

$$\begin{aligned}\nu_{\mathrm{vM}}(X \cap S) &= \lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{n}\\ &= \lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{|S \cap \mathbb{N}_n|} \frac{|S \cap \mathbb{N}_n|}{n}\\ &= \lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{|S \cap \mathbb{N}_n|} \lim_{n\to\infty} \frac{|S \cap \mathbb{N}_n|}{n}\\ &= \lim_{n\to\infty} \frac{|X \cap \mathbb{N}_n|}{n} \lim_{n\to\infty} \frac{|S \cap \mathbb{N}_n|}{n}\\ &= \nu_{\mathrm{vM}}(X)\nu_{\mathrm{vM}}(S),\end{aligned}$$

by help of a standard result for the multiplication of sequence limits [Liskevich, 2014, Theorem 3.1.7]. □

Theorem 7.10. *Let X and S be two statistically independent events on the finitely additive probability space $(\mathbb{N}, \mathcal{A}_{\mathrm{vM}}, \nu_{\mathrm{vM}})$ with $\nu_{\mathrm{vM}}(S) > 0$, then the corresponding collective x, indicator function of X, has the admissible selection rule s, indicator function of S.*

Proof. It is given that

$$\begin{aligned}&\lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{n} = \nu_{\mathrm{vM}}(X \cap S)\\ &= \nu_{\mathrm{vM}}(X)\nu_{\mathrm{vM}}(S) = \lim_{n\to\infty} \frac{|X \cap \mathbb{N}_n|}{n} \lim_{n\to\infty} \frac{|S \cap \mathbb{N}_n|}{n}.\end{aligned}$$

Furthermore $\nu_{\text{vM}}(S) > 0$ implies that the corresponding selection rules selects infinitely many elements and $\lim_{n\to\infty} \frac{n}{|S\cap\mathbb{N}_n|} = \frac{1}{\nu_{\text{vM}}(S)}$.

This implies

$$\begin{aligned}
\lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{|S \cap \mathbb{N}_n|} &= \lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{n} \frac{n}{|S \cap \mathbb{N}_n|} \\
&= \lim_{n\to\infty} \frac{|X \cap S \cap \mathbb{N}_n|}{n} \lim_{n\to\infty} \frac{n}{|S \cap \mathbb{N}_n|} \\
&= \lim_{n\to\infty} \frac{|X \cap \mathbb{N}_n|}{n} \lim_{n\to\infty} \frac{|S \cap \mathbb{N}_n|}{n} \lim_{n\to\infty} \frac{n}{|S \cap \mathbb{N}_n|} \\
&= \lim_{n\to\infty} \frac{|X \cap \mathbb{N}_n|}{n}.
\end{aligned}$$

□

We note some caveats regarding the preceding discussion.

1. We require the frequency limit for s to exist in Theorem 7.9. So the sequence S is not an entirely general selection rule.

2. The condition $\nu_{\text{vM}}(S) > 0$ in Theorem 7.10 ensures that we do not condition on measure zero events. Furthermore, it guarantees at this point, that the indicator function of S is a selection rule containing infinitely many ones. Besides that, its frequency limit exists.

3. Even though given here only for the binary case. The argumentation can be extended to continuum labeled collectives [Von Mises and Geiringer, 1964, II.B].

4. In our discussion, we focus on admissibility instead of statistical independence in the sense of von Mises, since our argument finally focuses on the randomness interpretation of statistical independence. Admissibility and von Mises' statistical independence are, despite the objects on which they are defined, equivalent.

5. Statistical independence in von Mises' setting and Kolmogorov's setting are not equivalent. They are special cases of a more general form of statistical independence. So they are mathematically analogous under mild conditions. But we emphasize the conceptual difference. Von Mises refused to define statistical independence between mere events [Von Mises and Geiringer, 1964, p. 35]. Instead he demanded statistical independence to be defined between collectives underlining that independence is a concept about aggregates, the collectives, not single occurring events.

7.9.2 Von Mises' Notions of Fairness in Practice

In practice we never observe infinite sequences, which are the building blocks of von Mises' theory. Nevertheless, infinite sequences can be seen as idealizations of finite, observable sequences. We make the following crucial and debatable assumption here: *Given a sequence of sufficient length, the frequencies observed in this sequences are arbitrarily close to the limit of the frequency in the continuation of this sequence.*

This assumption is critical for at least three parts:

1. We assume convergence (cf. [Fröhlich et al., 2023]).
2. In finite data we can only observe fractional frequencies, but not irrational ones (converging frequencies do fill the entire [0,1]).
3. The frequency limit of any infinite sequence is not governed by any frequency observed on a finite sequence.

For further simplification, we constrain ourselves to $0, 1$-outcomes and $0, 1$-decisions. In the following, we call 1 as outcome or decision "positive", comparable to a setting where the 1-label corresponds to "getting a loan" and 0 to "not getting a loan".

We consider the following simple toy example: let $X = \mathbb{R}, Y = \{0, 1\}$. The data points of the 0 label are distributed according to a Gaussian distribution with mean -1 and standard deviation 1. The data points of the positive labels are distributed according to a Gaussian distribution with mean 1 and standard deviation 0.5. We consider the simple logistic probabilistic predictor $p(x) = \frac{e^{cx+a}}{1+e^{cx+a}}$ with $c = 2$ and $a = -0.2$ and $x \in X$. We consider two groups: group A consists of all data points with values between -1 and -2, group B consists of all data points with values between smaller than -1.5 or greater than 1.5. In this case, the input and sensitive groups are highly correlated. Figure 7.1 and Figure 7.2 illustrate the example.

Since the predictor is probabilistic we have to introduce a threshold $q \in [0, 1]$ which derives a $0, 1$-decision from the prediction. The following plots show how the frequency of positive outcomes respectively positive decisions depending on the notion of fairness changes for different decision-thresholds.

Independence We compute the frequency of positive decisions for each group and the entirety. Independence requires these frequencies to be equal. See Figure 7.3.

Separation We compute the frequency of positive decisions given that the outcomes were positive for each group and the entirety. Separation requires these frequencies to be equal. See Figure 7.4.

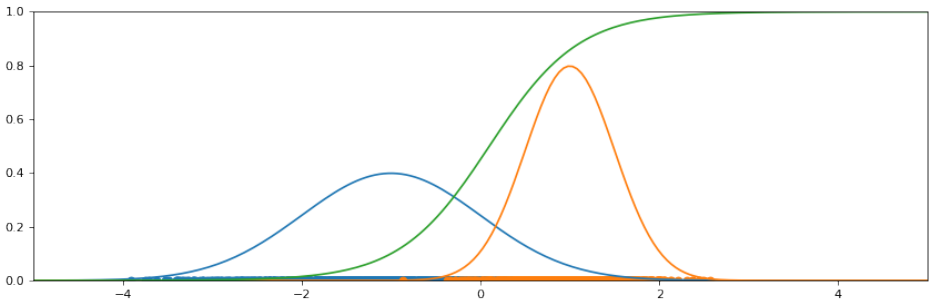


Figure 7.1: Generating Distribution and Predictor of the Toy Example

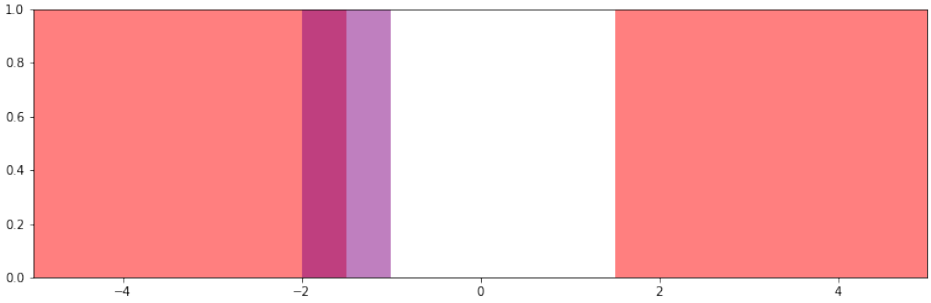


Figure 7.2: Groups in the Toy Example

Sufficiency We compute the frequency of positive outcomes given that the decision were positive for each group and the entirety. Sufficiency requires these frequencies to be equal. See Figure 7.5.

What do we learn from this toy example:

1. Our suggested notions have an empirical counterpart. Even though, the assumption taken above is critical. Much in contrast to most literature in fair machine learning, however, we (can) specify the idealization assumption explicitly. The regular setup hides many of the concerns behind a curtain of acceptance.

2. We observe that there is nothing special about our definitions of fairness. The frequential approach is intuitive and simplifies communication of technical results about machine learning fairness in societal debates.

3. We did not allude to specific algorithms which guarantee accurate predictions under fairness constraints. All such existing algorithms can be deployed in order to provide fair predictions following our definitions. Our notions are reinterpretations, not entirely new setups.

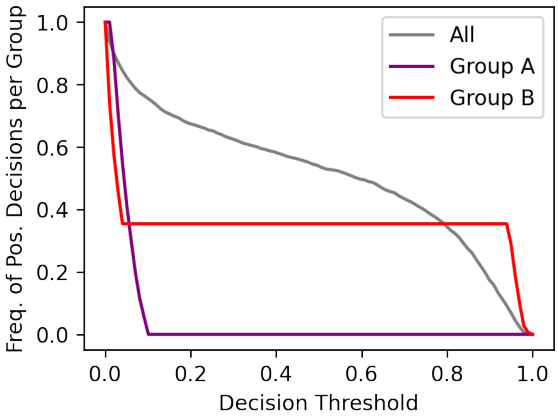


Figure 7.3: Frequency of Positive Decisions per Group depending on the Decision-Threshold

4. There is still a further, important, debate missing in this work. As pointed out in the conclusion of the paper, further conceptual clarification has to account for approximations of fairness notions to further enable principled argumentation in fairness debates.

7.9.3 Four Intuitions of Randomness

Typicality A sequence is called random if it shares "some" characteristics of any possible sequence. Martin-Löf and Schnorr formalized this idea via statistical tests [Martin-Löf, 1966, Schnorr, 1977].

Incompressibility The information contained in a random sequences is (approximately) as large as the sequence itself. The sequences cannot be compressed by "some" procedure [Kolmogorov, 1965, Chaitin, 1966, Schnorr, 1972, Levin, 1973].

Unpredictability In a random sequence one can know all foregoing elements without being able to predict by "some" procedure the next element [Ville, 1939, Muchnik et al., 1998, Shafer and Vovk, 2019, Frongillo and Nobel, 2020, De Cooman and De Bock, 2022].

Independence A random sequence is independent of "some" other sequences [Von Mises, 1919, Church, 1940].

7.9.4 Relations between Randomness Definitions

We briefly outline some of the known relationships between various notions of randomness.

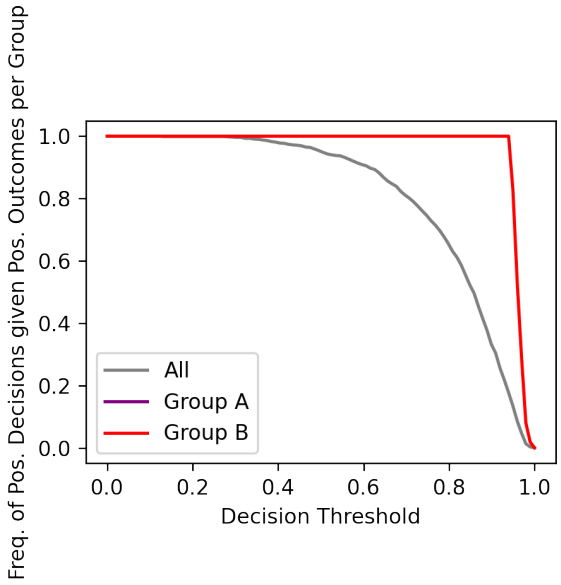


Figure 7.4: Frequency of Positive Decisions given Positive Outcomes per Group depending on the Decision-Threshold

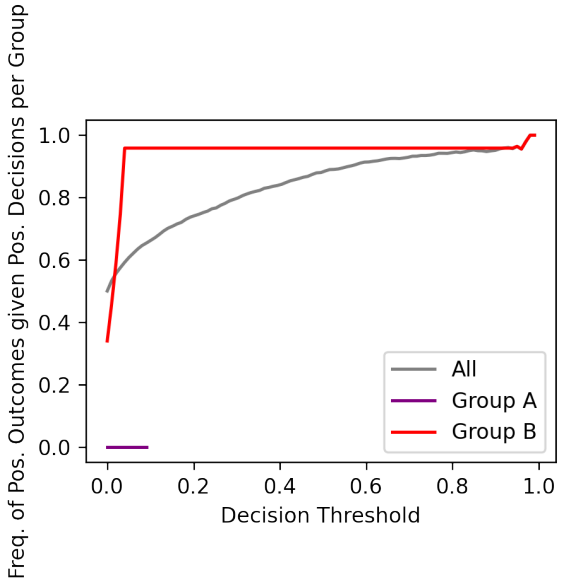


Figure 7.5: Frequency of Positive Outcomes given Positive Decisions per Group depending on the Decision-Threshold

On one hand, there are mathematical expressions capturing the same randomness intuitions meanwhile being mathematically distinct to each other.

For instance, the definition of a typical sequence following Martin-Löf [Martin-Löf, 1966] implies Schnorr's definition of typicality [Schnorr, 1971, 1977] but not vice versa [Uspenskii et al., 1990, p. 143]. Cooman and De Bock's imprecise unpredictability randomness [De Cooman and De Bock, 2022] is strictly more expressive than Vovk and Shafer's unpredictability randomness [Shafer and Vovk, 2019, Section 1.1] [De Cooman and De Bock, 2022, Theorem 37].

On the other hand, there are mathematical expressions capturing differing randomness intuitions meanwhile being necessary, sufficient or even equivalent to each other. For instance, the Levin-Schnorr theorem, simultaneously proven by Levin [Levin, 1973] and Schnorr [Schnorr, 1972, 1977], established an equivalence of typicality following Martin-Löf [Martin-Löf, 1966] and incompressibility following Levin [Levin, 1973] and Schnorr [Schnorr, 1972] based on the idea of [Kolmogorov, 1965] (e.g., see in [Volchan, 2002, Theorem 5.3] and references therein). Cooman and de Bock expressed Schnorr [Schnorr, 1971, 1977] and Martin-Löf randomness [Martin-Löf, 1966] in terms of an unpredictability approach [De Cooman and De Bock, 2022]. Muchnik showed in [Muchnik et al., 1998] that all incompressible sequences, again following Levin and Schnorr's prefix-complexity approach [Schnorr, 1972, Levin, 1973] are unpredictable in his sense.

In Ville's thesis Ville [1939] he attempted to generalize the idea of "excluding a gambling strategy" (see Section 7.4.4) via a game-theoretic approach. He showed [Ville, 1939, p. 76] that for any set of admissible selection rules $\mathcal{S}$ he could construct a gambling strategy, more exactly a capital process associated to a gambling strategy, which captures the same randomness definition. The opposite direction is, however, impossible [Ville, 1939, p. 39]. But Ambos-Spies et al. Ambos-Spies et al. [1996] introduced a weaker form of Ville [1939]'s randomness as unpredictability which is equivalent to Church's randomness Church [1940] as independence [Downey and Hirschfeldt, 2010, Section 12.3]. Finally, van Lambalgen observed an abstract interpretation of "random with respect to something" as "independent to" in [Van Lambalgen, 1990].

A prototypical, absolute and universal Notion of Randomness

One often referred *absolute and universal* definition of randomness is Martin-Löf's typicality approach [Buss and Minnes, 2013]. Martin-Löf's randomness as typicality has been equivalently formalized in terms of incompressibility and unpredictability (see Section 7.9.4). Furthermore, a sequence, which is Martin-Löf random [Martin-Löf, 1966], is statistically independent to all partial computable selection rules [Church, 1940] [Tadaki, 2014, Theorem 11]. Contrarily, not every collective with respect to all partial computable selection rules

is a Martin-Löf random sequence [Blum, 2009, p. 193]. So Martin-Löf's definition is linked to all four intuitions.

7.9.5 Kolmogorov's versus von Mises' Probability Theories in a Table

See Table 7.3.

	Kolmogorov	Von Mises
fundamental structure	probability space $(\Omega, \mathcal{F}, P)$	collectives $(x_i)_{i\in\mathbb{N}}$, implicitly $(\mathbb{N}, \mathcal{A}_{\text{vM}}, \nu_{\text{vM}})$
base set	(almost arbitrary) set Ω	$\mathbb{N}$
set of events	σ-algebra $\mathcal{F}$	Dynkin-system $\mathcal{A}_{\text{vM}}$
probability	finite positive measure	limit of frequency sequence
probability measure	countably additive	finitely additive ν_{vM}
randomness	no explicit mathematical definition	explicit mathematical definition
statistical independence	factorization of joint distribution	frequency limit doesn't change under subselection
data model	each data point a random variable X_i on $(\Omega, \mathcal{F}, P)$	each data point an element in a collective $(x_i)_{i\in\mathbb{N}}$

Table 7.3: Summary of main difference between Kolmogorov's and von Mises' probability theories.

7.9.6 Penguin Colony Example for Fair Collective

See Figure 7.6.

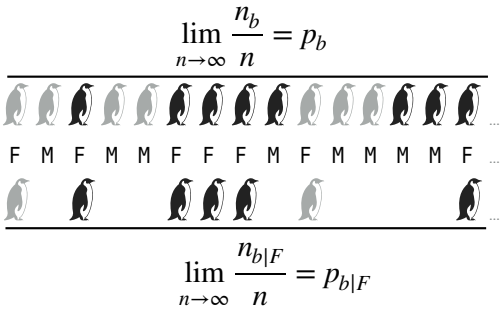


Figure 7.6: Example for a subselection and fair collective. n_b denotes the number of black penguins among the first n-penguins. Blackness of penguins is distributed fairly with respect to sex if $p_b = p_{b|F}$.

8

Systems of Precision: Coherent Probabilities on Pre-Dynkin Systems and Coherent Previsions on Linear Subspaces

Author Contributions

The idea of the paper was collaboratively conceptualized by Robert C. Williamson (RW) and Rabanus Derr (RD). RW supervised and steered this project. He wrote short paragraphs (e.g., the link to intersectionality in the introduction) and took over the internal editing. Rabanus Derr wrote the majority of the paper.

The Personal Note

The frequential theory of probability by von Mises drew my attention beyond the first project on fairness and randomness. While "Fairness and Randomness", i.e., Chapter 7, had finally been submitted and accepted for review, I had already crawled into a new rabbit hole. The comparison of von Mises theory with the theory by Kolmogorov revealed a set of differences. One of them being the set of events to which a probability is assigned. Kolmogorov's theory, built upon measure theory, defines this set to be a σ-algebra. Von Mises derives this set from collectives. It forms a pre-Dynkin system, i.e., a set system which is closed under complement and disjoint union.

After a little research I discovered that pre-Dynkin systems and the closely related Dynkin systems have occurred in a wide range of research fields under various of names – I found at least 12 names. I was fascinated that those strings have never been pulled together.

During the same time I learned more and more about imprecise probability (IP). IP questions whether a single probability distribution suffices as universal descriptive tool for uncertainties. One flavor of IP extends single probability distribution to sets of probability distributions. Under certain regularity conditions this set of probability distributions can be represented by a lower and upper probability, comparable to lower and upper bounds. Surprisingly, a regularity condition called *coherence* coincides with a concept of *extendability* hidden in quantum probability. I stumbled over extendability when searching for more cousins of (pre-)Dynkin-systems. I was truly fascinated and excited to find a link between topics which apparently nobody has observed before.

Nevertheless, I definitely got more and more insecure about what to answer when people asked me: "And, what is the use of this? What should I do differently when doing machine learning?". Well, it didn't really help to see that at my first conference abroad, the ELLIS (European Laboratory for Learning and Intelligent Systems) Doctoral Symposium 2022 in Alicante, Spain, I presented the poster to notice that people joined mainly for my penguin illustration but didn't really take an interest in the set systems which *are* implicitly used in machine learning, e.g., [Eban et al., 2014]. I needed to move forward.

Rather luckily, the SIPTA (Society for Imprecise Probability: Theories and Applications) announced their call for papers with perfecti timing, Bob took notice of that and suggested to submit. It was during the last steps to finish that Bob suggested to start the paper with the question: What is the set of events on which an imprecise probability is precise? The answer to the question formed the title. The ISIPTA (International Symposium on Imprecise Probability: Theory and Applications) almost a year after Alicante, 2023 in Oviedo, Spain, was an entirely different experience. Researchers were actually interested! They were (mostly) not machine learning researchers, but maybe for the better.

We extended the conference version to include an analogous development for imprecise expectations and systems of random variables. The longer version got rejected by the Internal Journal for Approximate Reasoning (IJAR), but we were invited to contribute to a special issue by Entropy by two editors, physicists, to whose work we referred to.

Now, I'm uncertain about whether the set structure of precision serves well as a starting point for expressing uncertainty. Quantum probability never seemed to get traction. It feels a bit like that set structures are a dead-end. The tools are weak, hence they reoccur or get re-invented here and there. But the tools are weak, hence they seem to not allow for a strong theoretical development. I'm still unclear about the exact applicability that the machine learning researchers around me were referring to. Maybe, I'm wrong and I just don't see it yet.

Summary: Systems of Precision

Motivation In order to describe uncertainties scholarship has moved beyond standard probability distribution. Imprecise probabilities (IP) weaken the rules of probability calculus [Walley, 1991, Augustin et al., 2014]. The field of imprecise probability has (sometimes independent) developed into a fruitful forest of mathematical treatment of uncertainty in finance, e.g., [Delbaen, 2002, Föllmer and Schied, 2011], decision theory, e.g., [Gilboa and Schmeidler, 1989, Narens, 2016, Zhang, 2002], engineering, e.g., [Faes and Moens, 2020], statistics, e.g., [Walley, 1991, Martin and Liu, 2013] and philosophy, e.g., [Schoenfield, 2012, Konek, forthcoming, Isaacs et al., 2022]. However, the simple question of on which events are imprecise probabilities precise seemed to have obtained no attention.

On the other hand, standard probability theory presumes that probabilities are always defined on an entire σ-algebra. But, literature on decision theory [Epstein and Zhang, 2001, Zhang, 2002], quantum physics [Gudder, 1969, 1979, Pták, 2000], frequential probability theory [Schurz and Leitgeb, 2008, Fröhlich et al., 2023], and machine learning [Eban et al., 2014] observed that probabilities are sometimes only defined on fewer events than a σ-algebra presupposes. Mainly, scholars noted that the intersection of arbitrary events does not necessarily lead to a new event with a precise probability ascribed. That observation contradicts the classical picture. The set structure of a σ-algebra, central to Kolmogorov's axiomatization, allows for countable intersections.

Hence, what can we say about the weaker event structure? What can we say about probabilities for events which are not in this weaker event structure? What is the relationship of this structure to the set of events on which an imprecise probability is precise? Overall, how is uncertainty captured in the "set structure of precision"?

This work has been published before I began to think about *forecast felicity*. Therefore,

the terms and formulations might differ to the surrounding chapters. The key terms are "translated" in Table 8.1.

Unified Notation & Terminology	Chapter 8
statistical forecast	–
individual, ι	element, $\omega \in \Omega$
aggregate, $\mho$	base set, Ω
attribute, $\varrho(\iota) = \wp$	set, elements of set systems, events, $A \subseteq \Omega$
measurement	finitely additive probability μ

Table 8.1: Translation table – Systems of Precision: Coherent Probabilities on Pre-Dynkin Systems and Coherent Previsions on Linear Subspaces

Contribution We show that the answer to both strands of questions revolve around *(pre-)Dynkin systems*. Pre-Dynkin systems are set systems, i.e., sets of sets (respectively sets of events), that contain the empty set, are closed under complement and finite disjoint union. Dynkin system are furthermore closed under countable disjoint union.

The set structure of precision for imprecise probabilities are, under some conditions on the imprecise probabilities, pre-Dynkin systems (respectively Dynkin systems, if further continuity conditions are added). The restriction of the imprecise probability to the set structure of precision leads to a standard finitely (respectively countably) additive probability measure. This measure is defined on a (pre-)Dynkin system instead of a σ-algebra.

On the other hand, (pre-)Dynkin systems are the minimal structure of events on which finitely (respectively countably) additive probabilities are defined. A precise probability defined on a (pre-)Dynkin system is extendable in different ways. One of the ways requires the probability to be *coherent*, a regularity condition central to imprecise probability. We show that coherence is equivalent to *extendability* a regularity condition introduced in quantum probability.

Finally, we span up a duality structure between the set of probability distributions which are precise on a pre-Dynkin system and the pre-Dynkin system. In particular, we show that such a set of probability distributions, under some additional restriction, identifies the pre-Dynkin system. With this dual structure we close the loop between set structures and probabilities on set structures.

An analogous treatment of imprecise expectations and random variables completes the work.

Tools and Sketch of Arguments Let us formally introduce pre-Dynkin systems and finitely additive probabilities to intuitively explain the findings. We leave details and the

treatment of Dynkin systems with countably additive probabilities, i.e., standard probabilities, to the full paper (Chapter 8).

A set system $\mathcal{D} \subseteq 2^{\Omega}$[1] on some set Ω is a *pre-Dynkin system* (Definition 8.1) if and only if all of the following conditions hold:

(a) $\emptyset \in \mathcal{D}$,

(b) $D \in \mathcal{D}$ implies $D^c := \Omega \setminus D \in \mathcal{D}$

(c) $C, D \in \mathcal{D}$ with $C \cap D = \emptyset$ implies $C \cup D \in \mathcal{D}$.

A *finitely additive probability measure on a pre-Dynkin system* $\mathcal{D}$ (Definition 8.9) is a function $\mu \colon \mathcal{D} \to [0, 1]$ if and only if it fulfills the following two conditions:

(a) *Normalization*: $\mu(\emptyset) = 0$ and $\mu(\Omega) = 1$.

(b) *Additivity*: let $A, B \in \mathcal{D}$ and $A \cap B = \emptyset$. Then, $\mu(A \cup B) = \mu(A) + \mu(B)$.

A finitely additive probability measure μ on a pre-Dynkin system is *extendable* if and only if there exists a finitely additive probability measure ν on 2^{Ω} such that $\nu|_{\mathcal{D}} = \mu$ (Definition 8.15).

Note that a probability measure defined as above but on an arbitrary set system can by default be extended to a finitely additive probability measure on the smallest pre-Dynkin system which encompasses the arbitrary set system. Hence, the restriction to pre-Dynkin systems is without loss of generality (Theorem 8.10 and Proposition 8.25).

To get an intuition for the findings let us look at the set of probability measures which coincide with a reference probability measure μ on a pre-Dynkin system $\mathcal{D}$,

$$\mathcal{P} := \{\nu \in \Delta \colon \nu|_{\mathcal{D}} \mu\},$$

where Δ is the set of all finitely additive probability measures on 2^{Ω}. If $\mathcal{P} \neq \emptyset$, then μ is extendable, hence coherent (Theorem 8.19). Suppose μ is extendable (hence coherent). Define $\underline{\mu}_{\mathcal{D}}(A) := \inf_{\nu \in \mathcal{P}}(A)$ and $\overline{\mu}_{\mathcal{D}}(A) := \inf_{\nu \in \mathcal{P}}(A)$ for all $A \in 2^{\Omega}$. Then, all sets $A \subseteq \Omega$ such that $\underline{\mu}_{\mathcal{D}}(A) = \overline{\mu}_{\mathcal{D}}(A)$ form a pre-Dynkin system, not necessarily $\mathcal{D}$ (Theorem 8.10 and Proposition 8.31). The reason is that measure zero sets impose additional constraints, which make the probabilities coincide on more sets. If furthermore, $|\Omega| < \infty$ and $\mu > 0$, then the pre-Dynkin system is equal to $\mathcal{D}$ (Theorem 8.37).

[1] We denote the power set of Ω by 2^{Ω}.

Relation to Forecast Felicity The *measurability problem* (Section 5.3) motivates the study of the set structure of precision in the context of forecast felicity.

Suppose that the forecasts under consideration are of Type A2I. More specifically, we suppose that the forecast is a finitely additive probability on a pre-Dynkin system (cf. Chapter 6). Finitely additive probabilities on pre-Dynkin systems can be understood as idealized measurements, much like a standard probability distribution on σ-algebra (Section 6.2). For this, consider the set Ω to be an abstract aggregate. An individual is an element $\omega \in \Omega$. An event, or set, $A \subseteq \Omega$ is a subset of the aggregate which shares the attribute of interest for the forecast. The set A defines an binary indicator function which corresponds to the attribution function ϱ (see Table 8.1). The probability of A, $\mu(A)$, if defined, is an idealized measurement of the size of individuals which share the attribute.

Exemplarily, let Ω be the aggregate of days. The set A is the subset of rainy days. The idealized measurement of the size of A, i.e., $\mu(A)$, is the forecast that it rains today. In this case, $\mu(A)$ is literally a probability "measure". But is $\mu(A)$ defined, respectively is $A \in \mathcal{D}$ where $\mathcal{D}$ is the pre-Dynkin system on which the probability is defined?

To justify that the forecast is such an idealized measurement, one must assume that the relevant attributes, i.e., events, are part of the pre-Dynkin system. In other words, the assumption of *measurability*, i.e., being in the pre-Dynkin system, is required.

The prerequisite of measurability is often swept under the carpet. For instance in machine learning, hardly any paper pays attention to the fact that measurability can break. However, the series of works in different research fields (see the paragraph on Motivation above or Section 8.1 in the original paper) show it *does* break.

The re-occuring pattern in those examples is that if there are two overlapping events which are attributed a precise probability it is not a priori clear whether their intersection can get assigned a precise probability. For instance, if the size of rainy days $\mu(A)$ is well-defined and the size of warm days (temperature $> 20°C$) is well-defined $\mu(B)$, then $\mu(A \cap B)$ might not well-defined. There may be numerous reasons. (a) The intersection is an extremely rare event for which we practically cannot obtain any precise estimate. (b) The intersection is practically not measurable with the devices at hand, e.g., because rain measurement disturbs temperature measurement. This is indeed the case for social measurements, in which if one question gets asked, the answers to a second question are distorted [Trueblood and Busemeyer, 2011]. (c) The intersection is physically impossible to measure. It follows that intersectability, hence measurability, is a relative and contextual choice.

To demonstrate that measurability can be (in)felicitously assumed and provided, reconsider the two events A (days with rain) and B (days with temperature $> 20°C$). A rain forecast in a newspaper might deliberately ignore the intersection of rainfall and temperature measurement and do "as if" rainfall and temperature do not interfere. An agricultural forecast

should be more accurate and, accordingly, provide independent, simultaneous measurements and statistics for rainfall and temperature. Respectively, if not possible, the forecast should make explicit the shortcoming with respect to the interference between the two events. A felicitous forecast requires a felicitous assumption of measurability.

Our elaborations in Chapter 8 provide a structure of the choice of measurability – the lattice of pre-Dynkin systems. As we argued pre-Dynkin systems are the relevant set systems to consider. They are ordered depending on their fine-grainedness from the trivial $\{\emptyset, \Omega\}$ pre-Dynkin system to the full power set 2^{Ω}. Furthermore, we show that the assumption of measurability can be circumvent by providing bounds to the precise forecasts.

Abstract

In the literature on imprecise probability, little attention is paid to the fact that imprecise probabilities are precise on a set of events. We call these sets *systems of precision*. We show that, under mild assumptions, the system of precision of a lower and upper probability form a so-called (pre-)Dynkin system. Interestingly, there are several settings, ranging from machine learning on partial data over frequential probability theory to quantum probability theory and decision making under uncertainty, in which, a priori, the probabilities are only desired to be precise on a specific underlying set system. Here, (pre-)Dynkin systems have been adopted as systems of precision, too. We show that, under extendability conditions, those pre-Dynkin systems equipped with probabilities can be embedded into algebras of sets. Surprisingly, the extendability conditions elaborated in a strand of work in quantum probability are equivalent to coherence from the imprecise probability literature. On this basis, we spell out a lattice duality which relates systems of precision to credal sets of probabilities. We conclude the presentation with a generalization of the framework to expectation-type counterparts of imprecise probabilities. The analogue of pre-Dynkin systems turns out to be (sets of) linear subspaces in the space of bounded, real-valued functions. We introduce partial expectations, natural generalizations of probabilities defined on pre-Dynkin systems. Again, coherence and extendability are equivalent. A related but more general lattice duality preserves the relation between systems of precision and credal sets of probabilities.

8.1 Introduction

Scholarship in imprecise probability largely focuses on the imprecision of probabilities. However, imprecise probability models often lead to *precise* probabilistic statements on certain events or gambles, i.e., bounded, real-valued functions. In this work, we follow a hitherto untaken route investigating the *system of precision*, i.e., the set structure on which an imprecise probability is precise. (We elaborate the exact definition of imprecise probabilities and expectation used here in Section 8.3 and Section 8.6). It turns out that (pre-)Dynkin systems (as shown in Appendix 8.8.2, (pre-)Dynkin systems appear under plenty of names) describe the set of events with precise probabilities (cf. Section 8.3). This event structure is a neglected object in the literature on imprecise probability. In particular, it constitutes a parametrized choice somewhat "orthogonal" to the standard. Roughly stated, existing approaches to imprecise probability generalize the probability measure μ_σ in a classical probability space $(\Omega, \mathcal{F}_\sigma, \mu_\sigma)$. Following Kolmorogov's classical setup, Ω is the base set, $\mathcal{F}_\sigma$ a σ-algebra and μ_σ a countably additive probability on $\mathcal{F}_\sigma$. Approaches to imprecise probability often do not even presuppose an underlying measure space (e.g., [Walley, 1991]). However, they are often linked to finitely additive measure spaces $(\Omega, \mathcal{F}, \mu)$, where μ is a finitely additive probability, and $\mathcal{F}$ is an algebra of sets (sometimes called a field). We start by generalizing $\mathcal{F}_\sigma$ from a σ-algebra to a pre-Dynkin system.

This suggestion is practically motivated: what do the following scenarios have in common?

(a) A machine learning algorithm has access to a restricted subset of attributes. It cannot jointly query all attributes simultaneously. This is called "learning on partial, aggregated information" [Eban et al., 2014]. The reasons might be manifold: for privacy preservation, "not-missing-at-random" features, restricted data base access for acceleration or multi-measurement data sets.

(b) Quantum physical quantities, e.g., location and impulse, are (statistically) incompatible [Gudder, 1979].

(c) A preference ordering on a set of acts gives rise to precise beliefs on a set of events, whereas this belief is not necessarily precise for intersections of such events [Epstein and Zhang, 2001, Zhang, 2002].

In all of these scenarios, there does not exist a precise probability over all attributes and events. Or, there is no such precise probability accessible. Two attributes might each on their own exhibit a precise probabilistic description, while a joint precise probabilistic description does not exist. On a more fundamental level, no intersectability is provided. A precise probabilistic description of two events does *not* imply that the intersection of those events possesses a precise probability. The set system for the description of the events with

precise probabilities which independently turned up in the various, previously mentioned fields of research is, again, the (pre-)Dynkin system.

The question of intersectability (or "intersectionality") is of considerable interest in the social sciences where it is used as a label to describe the problem of the *joint* effect of various individual attributes on social outcomes [Cole, 2009, Shields, 2008, Weldon, 2008]. That this notion of intersectionality has something to do with set systems is clear already from the fact that the Venn diagram pictured in Figure 8.1 is used as an illustration both for the Wikipedia articles on hypergraphs [Anonymous, 2023a] (another name for a set system [Berge, 1989]) and intersectionality [Anonymous, 2023b].

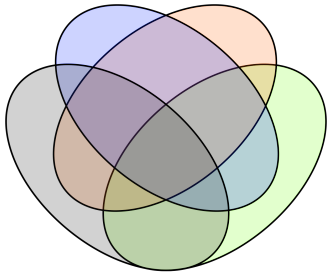


Figure 8.1: Exemplary illustration of Venn diagram for four sets (By RupertMillard, CC BY-SA 3.0).

Needless to say, the concept as used in the social sciences is rich, complex, and somewhat vague, which is not necessarily held to be a weakness: "at least part of its success has been attributed to its vagueness" [Hancock, 2013] (p. 260). Our interest is in under what circumstances precise probabilities can be ascribed to events; we speculate that such formal results may well contribute to a deeper empirical understanding of social intersectionality, without resorting to fuzzy logic [Hancock, 2007] with its potential lack of operational definition as argued for by Cooke [2004].

By rethinking the domain of probability measures, one might wonder about the origins of Kolmogorov's σ-algebra as *the* set system for events which possess probabilities. This links back to the old problem of measurability [Elstrodt, 2018] (pp. 1–5). The measurability problem is the mathematical problem to assign a uniform measure to all subsets of a continuum. Giuseppe Vitali showed in 1905 that this problem is not solvable for countably additive measures [Elstrodt, 2018] (p. 5) (from [Vitali, 1905]). Hence, more restricted set systems such as the σ-algebra arose. Isaacs et al. [2022] reconsidered this century-old discussion to argue for rationality of imprecise probabilities. We take their argument even further. Inside the borders of mathematical measurability, the set of events which ought to be assigned probabilities is a modeling choice. Measurability is a modeling tool. We show that it is naturally parametrized by the set of (pre-)Dynkin systems.

All of the preceding considerations bring us to the main question of this paper: *What is the*

system of precision and how does it relate to an imprecise probability on "all" events? We approach this question from three perspectives.

1. First, we show that, under mild assumptions, a pair of lower and upper probabilities assign precise probabilities, i.e., lower and upper probability coincide, to events which form a pre-Dynkin system or even a Dynkin system.

2. Second, we define probabilities on pre-Dynkin systems in accordance with the literature on quantum probability, in particular [Gudder, 1969]. We argue that probabilities on pre-Dynkin systems, as well as their inner and outer extension, exhibit few desirable properties, e.g., subadditivity cannot be guaranteed. Hence, extendability, the ability to extend a probability from a pre-Dynkin system to a larger set structure, turns out to be crucial, as it implies coherence of the probability defined on the pre-Dynkin system. This observation links together the research from probabilities defined on weak set structures [Gudder, 1969, Zhang, 2002, Schurz and Leitgeb, 2008] to imprecise probabilities [Walley, 1991, Augustin et al., 2014]. Furthermore, extendability guarantees the existence of a nicely behaving, so-called coherent extension. We finally show that the inner and outer extension of a probability defined on a pre-Dynkin system is always more pessimistic than its corresponding lower and upper coherent extension.

3. Last, we develop a duality theory between pre-Dynkin systems on a predefined base measure space and their respective credal sets of probabilities. The credal sets consist of all probabilities which coincide with the pre-defined measure on a pre-Dynkin system. A so-called Galois connection links together the containment structure on the set of set systems with the containment structure on the set of credal sets.

We conclude our presentation with a generalization to expectation-type counterparts of imprecise probabilities in Section 8.6. These are often called previsions, e.g., in [Walley, 1991]. Our main question thus generalizes to: *what is the system of precision and how does it relate to an imprecise expectation on "all" gambles?* In this case, by "system of precision" we mean the set of gambles on which a lower and upper expectation coincide.

1. First, we propose a generalization of a finitely additive probability defined on a pre-Dynkin system. More concretely, we define *partial expectations* which correspond to expectation functionals which are only defined on a set of linear subspaces of the space of all gambles. However, on those linear subspaces, they behave like "classical" (finitely additive) expectations.

2. Second, we show that under some properties, imprecise expectations are precise on a linear subspace of the linear space of gambles. (cf. Section 8.3)

3. Third, we present a natural generalization of extendability for partial expectations, which again turns out to be equivalent to coherence of the partial expectation.

4. Last, analogous to the lattice duality (A lattice is a poset with pairwise existing minimum and maximum. The duality is expressed via an antitone lattice isomorphism.) described in Section 8.5, we present a lattice duality for linear subsets of the space of gambles and credal sets which define coherent lower and upper previsions.

In summary, our work makes contributions in-between the research field of imprecise probabilities, probabilities defined on general set structures, and partially defined expectation functionals. Part of this work has been presented at the International Symposium on Imprecise Probabilities: Theories and Applications under the title "The Set Structure of Precision" [Derr and Williamson, 2023a]. The following version is a more complete and exhaustive presentation of this conference version. We included all omitted proofs of the conference version. We elaborated the content of Section 8.5. We added an entire section about the generalization to expectation-type counterparts of probabilities on pre-Dynkin systems (Section 8.6). We presented relations to different research areas in more detail. We shortly discussed countably additive probabilities on Dynkin systems in the appendix, as we put emphasis on finitely additive probabilities in the main text. Before we begin the structural investigation of pre-Dynkin systems, we first introduce the used notation and fix the mathematical framework.

8.1.1 Notation and Technical Details

As we deal with a lot of sets, sets of sets, and rarely even sets of sets of sets in this paper, we agree on the following notation: sets are written with capital Latin or Greek letter, e.g., A or Ω. Sets of sets are denoted $\mathcal{A}$. Sets of sets of sets obtain the notation $\mathfrak{A}$. As usual, $\mathbb{R}$ is reserved for the set of real numbers, $\mathbb{N}$ for the natural numbers. The power set of a set A is written as 2^A.

In the course of this work, we require the notions of σ-algebras and algebras (of sets). An algebra is a subset of 2^Ω which contains the empty set and is closed under complement and finite union. A σ-algebra is an algebra which is closed under countable union [Williams, 1991] (Definition 1.1). (Our notion of an algebra should not be confused with the notion of an algebra over a field.) (Probability) measures are denoted by lowercase Greek letters, e.g., μ, ν and ψ, except for σ. Generally, we use "σ" to emphasize the countable nature of a mathematical object. This becomes clear when we define Dynkin systems (Definition 8.1). Other functions are denoted by lowercase Latin letters, e.g., f and g.

Regarding the technical setting, we roughly follow the setup of [Walley, 1991] (§3.6 and Appendix D). For a summary, see Table 8.2. Let Ω be an arbitrary set. In several examples

Table 8.2: Summary of important notations used.

$\Omega, 2^\Omega$	Base set and its power set
$[n]$	Set $\{1, \dots, n\}$
$\mathcal{D}$	Pre-Dynkin system on Ω (Definition 8.1)
$\mathcal{D}_\sigma$	Dynkin system on Ω (Definition 8.1)
$\mathrm{D}(\mathcal{A})$	Pre-Dynkin hull of a set system $\mathcal{A} \subseteq 2^\Omega$ (Definition 8.1)
μ	Finitely additive probability defined on $\mathcal{D}$ (Definition 8.9)
μ_*, μ^*	Inner respectively outer extension (Proposition 8.12)
$\underline{\mu}_\mathcal{D}$, $\overline{\mu}_\mathcal{D}$	Lower respectively upper coherent extension (Corollary 8.20)
$M(\mu, \mathcal{D})$	Credal set of μ on $\mathcal{D}$ (Corollary 8.20)
ν	Finitely additive probability defined on 2^Ω
ψ	Fixed, finitely additive probability defined on 2^Ω
χ_A	Indicator function of the set $A \subset \Omega$
Δ	Set of finitely additive probability measures on 2^Ω, set of linear previsions
$m_\psi \colon 2^{2^\Omega} \to 2^\Delta$	Credal set function (Definition 8.23)
$m_\psi^\circ \colon 2^\Delta \to 2^{2^\Omega}$	Dual credal set function (Definition 8.28)
$\overline{\mathrm{co}}$	Convex, Weak* Closure
$\mathrm{B}(\Omega)$	Set of real-valued, bounded functions on Ω
$\mathrm{ba}(\Omega)$	Set of bounded, signed, finitely additive measures on 2^Ω
E	Partial Expectation (Definition 8.45)
$\mathrm{S}(\Omega, \mathcal{A})$	Linear space of simple gambles on the set system $\mathcal{A}$
$\mathrm{B}(\Omega, \mathcal{F}_\sigma)$	Linear space of bounded, $\mathcal{F}_\sigma$-measurable functions
$\underline{\mathbb{E}}$	Coherent lower prevision (Definition 8.49)
$\overline{\mathbb{E}}$	Coherent upper prevision (Definition 8.49)
ν	Linear prevision defined on $\mathrm{B}(\Omega)$ (equivalent to ν above)
ψ	Fixed, linear prevision defined on $\mathrm{B}(\Omega)$ (equivalent to ψ above)
$m_\psi \colon 2^{\mathrm{B}(\Omega)} \to 2^\Delta$	Generalized credal set function (Definition 8.53)
$m_\psi^\circ \colon 2^\Delta \to 2^{\mathrm{B}(\Omega)}$	Generalized dual credal set function (Definition 8.54)

$\Omega = [n]$, where $[n]$ denotes the set $\{1, \dots, n\}$. The set $\mathrm{B}(\Omega)$ is defined as the set of all real-valued, bounded functions on Ω. We call those functions *gambles*. For instance, χ_A, the indicator function of $A \subseteq \Omega$, is in $\mathrm{B}(\Omega)$. The supremum norm, $\|f\|_{\sup} \coloneqq \sup_{\omega \in \Omega} |f(\omega)|$ makes $\mathrm{B}(\Omega)$ a topological linear vector space [Hildebrandt, 1934]. With $\mathrm{ba}(\Omega)$, we denote the set of all bounded, signed, finitely additive measures on 2^Ω. In fact, $\mathrm{ba}(\Omega)$ is the topological dual space of $\mathrm{B}(\Omega)$. So, in particular, every continuous linear functional $\nu \in \mathrm{B}(\Omega)^*$ can be identified with a bounded, signed, finitely additive measure [Hildebrandt, 1934]. (As the linear functional is defined on a normed space, continuity and boundedness are equivalent.) For this reason, we use, with minor abuse of notation, the same notation for bounded, signed, finitely additive measures and for continuous linear functionals in the dual space of $\mathrm{B}(\Omega)$, i.e., we write $\nu(f) = \int f d\nu$. The dual space $\mathrm{ba}(\Omega)$ gets equipped with the weak* topology, i.e., the weakest topology which makes all evaluation functionals of the form $f^* \in \mathrm{ba}(\Omega)^*$ such that $f^*(\nu) \coloneqq \int f d\nu$ for some $f \in \mathrm{B}(\Omega)$ continuous (for more details see,

e.g., [Schechter, 1997] (p. 758)). With $\Delta \subseteq \mathrm{ba}(\Omega)$, we denote the convex, weak*-closed subset of finitely additive probability measures. The set Δ plays a major role in Walley's theory of previsions, as the measures in Δ are in one-to-one correspondence to his linear previsions [Walley, 1991] (Theorem 3.2.2). The operator $\overline{\mathrm{co}}$ is the convex, weak* closure on the space $\mathrm{ba}(\Omega)$. We further introduce the following two notations: let $\mathcal{F} \subseteq 2^\Omega$ be an algebra. Then, $\mathrm{S}(\Omega, \mathcal{F}) \subseteq \mathrm{B}(\Omega)$ denotes the linear subspace of simple functions on $\mathcal{F}$, i.e., scaled and added indicator functions of a finite number of disjoint sets (cf. [Rao and Rao, 1983] (Definition 4.2.12)). Let $\mathcal{F}_\sigma \subseteq 2^\Omega$ be a σ-algebra. Then, $\mathrm{B}(\Omega, \mathcal{F}_\sigma) \subseteq \mathrm{B}(\Omega)$ denotes the linear subspace of all bounded, real-valued, $\mathcal{F}_\sigma$-measurable functions. Equipped with these notions and tools we are ready for a first preliminary question.

8.2 What Is a (Pre-)Dynkin System?

In this work, the main objects under consideration are pre-Dynkin systems and Dynkin systems. A (pre-)Dynkin system is a set system on Ω. It contains the empty set, is closed under complement and (countable) disjoint union. More formally:

Definition 8.1 ((Pre-)Dynkin system)**.** *We say $\mathcal{D} \subseteq 2^\Omega$ is a* pre-Dynkin system *on some set Ω if and only if all of the following conditions hold:*

(a) $\emptyset \in \mathcal{D}$,

(b) $D \in \mathcal{D}$ *implies* $D^c \coloneqq \Omega \setminus D \in \mathcal{D}$

(c) $C, D \in \mathcal{D}$ *with* $C \cap D = \emptyset$ *implies* $C \cup D \in \mathcal{D}$.

We call $\mathcal{D}_\sigma \subseteq 2^\Omega$ a Dynkin system *if and only if the conditions (a), (b) and*

(c') *let* $\{D_i\}_{i\in\mathbb{N}} \subseteq \mathcal{D}_\sigma$, *if for all* $i, j \in \mathbb{N}$ *with* $i \neq j$ *it holds* $D_i \cap D_j = \emptyset$ *then* $\bigcup_{i\in\mathbb{N}} D_i \in \mathcal{D}_\sigma$,

are fulfilled.

Observe that every Dynkin system is a pre-Dynkin system. We will denote pre-Dynkin systems by the use of $\mathcal{D}$, in contrast to $\mathcal{D}_\sigma$ for Dynkin systems. This should not be confused with $\mathrm{D}(\mathcal{A})$ for $\mathcal{A} \subseteq 2^\Omega$, which is the intersection of all pre-Dynkin systems which contain $\mathcal{A}$, i.e., the smallest pre-Dynkin system containing $\mathcal{A}$. (For $\mathcal{A} = \emptyset$ we define $\mathrm{D}(\mathcal{A}) = \{\emptyset, \Omega\}$.) In other words, $\mathrm{D}(\mathcal{A})$ is the *pre-Dynkin hull* generated by $\mathcal{A}$. The following short lemma will be helpful in later proofs.

Lemma 8.2 (Closedness under Set Difference)**.** *Let $\mathcal{D} \subseteq 2^\Omega$ be a pre-Dynkin system, if $A, B \in \mathcal{D}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{D}$.*

Proof. Let $A, B \in \mathcal{D}$ and $A \subseteq B$. Then, $(B \setminus A)^c = B^c \cup A$. Since $\mathcal{D}$ is closed under complement and disjoint union $B^c \cup A \in \mathcal{D}$. But, then again, the complement, $B \setminus A$, is in $\mathcal{D}$. □

In classical probability theory, Dynkin systems appear as a technical object required for the measure-theoretic link between cumulative distribution functions and probability measures (cf. [Williams, 1991] (Proof of Lemma 1.6)). In particular, every σ-algebra, the well-known domain of probability measures, is a Dynkin system. Thus, all statements within this work are generalizations of classical probability theoretical results. We give a short example of a pre-Dynkin system, which is not an algebra in the following. This example gets reused to illustrate forthcoming statements.

Example 8.3. *The smallest pre-Dynkin system, up to isomorphisms,which is not an algebra can be defined on* $\Omega_4 \coloneqq \{1, 2, 3, 4\}$. *It is given by* $\mathcal{D}_4 \coloneqq \{\emptyset, 12, 34, 13, 24, \Omega_4\}$, *where we write* 12 *as a shorthand for* $\{1, 2\}$.

Pre-Dynkin and Dynkin systems naturally arise in probability theory. For instance, the set of all subsets $A \subseteq \mathbb{N}$, such that the natural density $\mu(A) = \lim_{n\to\infty} \frac{|A \cap [n]|}{n}$ exists is a pre-Dynkin system $\mathcal{D}_{\mathbb{N}}$ but not an algebra [Schurz and Leitgeb, 2008]. It is sometimes called the density logic [Pták, 2000] and constitutes the foundation of von Mises' century-old frequential theory of probability [Von Mises, 1919] (refined and summarized in [Von Mises and Geiringer, 1964]). Intriguingly, this was used as an example by Kolmogorov [1927/1929] of a measure defined on a restricted set system for which it is desired to extend the measure to the power set $2^{\mathbb{N}}$ (cf. § 8.4); see the discussion in [Khrennikov, 2009b] (pp. 11–14), who observed (page 14) that "the main problem is *non-uniqueness* of an extension" and that such extended measures are impossible to verify from observed frequencies, because the relative frequencies do not converge for events in $2^{\mathbb{N}} \setminus \mathcal{D}_{\mathbb{N}}$. The non-uniqueness is naturally handled in the present paper by working with lower and upper previsions (or lower and upper probabilities) (cf. [Fröhlich et al., 2023]).

Another class of Dynkin systems occurs in so-called marginal scenarios [Cuadras et al., 2002]. Marginal scenarios are settings in which marginal probability distributions for a subset of a set of random variables are given, but not the entire joint distribution. This restricted "joint measurability" of the involved random variables can be expressed via Dynkin systems [Gudder, 1984] (Example 4.2), [Vorob'ev, 1962].

Pre-Dynkin systems are so helpful because they structurally align with finitely additive probability measures. The same statement holds for Dynkin systems and countably additive probabilities. If we know the probability of an event, then we know the probability of the complement; i.e., the event does not happen. If we know the probability of several events which are disjoint, then we know the probability of the union, which is just the sum.

Probabilities following their standard definition go hand in hand with Dynkin systems. We see this observation manifested in many following statements.

Remarkably, (pre-)Dynkin systems appeared under a variety of names (cf. Appendix 8.8.2). Fundamental to all its regular, independent occurences in many research areas is the need for a set structure which does not allow for arbitrary intersections.

8.2.1 Compatibility

(Pre-)Dynkin systems are not necessarily closed under intersections. However, when the intersection of two sets (events) is contained in the (pre-)Dynkin system, we call the two events *compatible*.

Definition 8.4 (Compatibility)**.** *Let A, B be elements in a pre-Dynkin system $\mathcal{D}$, then A and B are* compatible *if and only if $A \cap B \in \mathcal{D}$.*

This definition follows the definitions given in, e.g., [Gudder, 1969, 1973, 1984]. (It should not be confused with the very similar, and sometimes equivalent, notion of commutativity in logical structures [Narens, 2016] (Definition 14) (cf. Appendix 8.8.5).) Compatibility in pre-Dynkin systems is a symmetric relation, but it is not necessarily transitive. Furthermore, it is complement inherited, i.e., if A, B are compatible in a pre-Dynkin system, then so are A, B^c [Gudder, 1979] (Lemma 3.6). Lastly, compatibility, even though expressed as intersectability, i.e., "closed under intersection", can be equivalently expressed as unifiability, i.e., "closed under union".

Lemma 8.5 (Cup gives Cap gives Cup)**.** *Let $\mathcal{D}$ be a pre-Dynkin system and $A, B \in \mathcal{D}$. Then*

$$A \cap B \in \mathcal{D} \Leftrightarrow A \cup B \in \mathcal{D}$$

Proof. Using Lemma 8.2 for pre-Dynkin systems, we can quickly see that the following two decompositions give the desired equivalence:
For the "$\Rightarrow$"-direction: $A \cup B = (A \setminus (A \cap B)) \cup B$. The fact $A, A \cap B, B \in \mathcal{D}$ implies $(A \setminus (A \cap B)) \cup B \in \mathcal{D}$.
For the "$\Leftarrow$"-direction: $A \cap B = A \setminus ((A \cup B) \setminus B)$. The fact $A, A \cup B, B \in \mathcal{D}$ implies $A \setminus ((A \cup B) \setminus B) \in \mathcal{D}$. (A related result for Dynkin systems is given in [Gudder, 1969] (5.1).) □

Example 8.6. *We reconsider the set Ω_4 and pre-Dynkin system $\mathcal{D}_4$ from Example 8.3. The elements 12 and 34 are intersectable $12 \cap 34 = \emptyset \in \mathcal{D}_4$ and unifiable $12 \cup 34 = \Omega_4 \in \mathcal{D}_4$. The elements 12 and 13 are not intersectable $12 \cap 13 = 1 \notin \mathcal{D}_4$ and not unifiable $12 \cup 13 = 123 \notin \mathcal{D}_4$.*

The term "compatibility" underlines that closedness under intersection gets loaded with further meaning in the context of probability theory. As we define in the next section, $\mathcal{D}$ is the set of "measurable" events, i.e., events which get assigned a probability. Hence, two events, A, B, are called compatible if and only if a precise joint probabilistic description, i.e., a precise probability of $A \cap B$, exists. For a more thorough discussion of the nature of compatibility (and its cousin commutativity) we point to the literature on quantum probability, e.g., [Khrennikov, 2009a] (Definition 3.12), or [Rivas, 2019].

Compatibility is not only a property of elements in pre-Dynkin systems. One can take compatibility as a primary notion; i.e., one requires the statements of Lemma 8.5 and [Gudder, 1979] (Lemma 3.6) to hold. Then, a set structure which contains the empty set and the entire base set and is equipped with this notion of compatibility is a pre-Dynkin system [Khrennikov, 2009b] (Definition 5.1). (It is called semi-algebra in [Khrennikov, 2009b] (Definition 5.1).)

Interestingly, the assumption of arbitrary compatibility is fundamental to most parts of probability theory. σ-algebras, the domain of probability measures, are exactly those Dynkin systems in which all events are compatible with all others [Gudder, 1973] (Theorem 2.1). Algebras are exactly those pre-Dynkin systems in which all events are compatible with all others. Surprisingly, it turns out that, as well, all pre-Dynkin systems can be dissected into such "blocks" of full compatibility. Every pre-Dynkin system consists of a set of maximal algebras, which we call *blocks*. In particular, maximality here stands for the following: there is no algebra contained in $\mathcal{D}$ such that some $\mathcal{A}_i$ is a strict sub-algebra of this algebra. Similar and related results can be found in [Katriňák and Neubrunn, 1973, Šipoš, 1978, Brabec, 1979, Vallander, 2016].

Theorem 8.7 (Pre-Dynkin Systems Are Made Out of Algebras)**.** *Let $\mathcal{D}$ be a pre-Dynkin system on Ω. Then, there is a unique family of maximal algebras $\{\mathcal{A}_i\}_{i\in I}$ such that $\mathcal{D} = \bigcup_{i\in I} \mathcal{A}_i$. We call these algebras the* blocks *of $\mathcal{D}$.*

Proof. For the proof, we require the definition of a compatible subset of $\mathcal{D}$. A subset $\mathcal{A} \subseteq \mathcal{D}$ is compatible if all elements are completely compatible, i.e., any finite intersection of elements in $\mathcal{A}$ is contained in $\mathcal{D}$. This is indeed a stronger requirement than pairwise compatibility (cf. Definition 8.4). Certainly, every subset $\mathcal{A} \subseteq \mathcal{D}$ is compatible if and only if every finite subset of $\mathcal{A}$ is compatible. Hence, compatibility is a property of so-called finite character [Schechter, 1997] (Definition 3.46). Then, Tuckey's lemma (e.g., [Schechter, 1997] (Theorem 6.20.AC5)) guarantees that any compatible subset of $\mathcal{D}$ is contained in a maximal compatible subset. Since every element $D \in \mathcal{D}$ is in at least one compatible subset, e.g. $\{\emptyset, D, D^c, \Omega\} \subseteq \mathcal{D}$, the (unique) set of maximal compatible subsets $\{\mathcal{A}_i\}_{i\in I}$ covers the entire pre-Dynkin system $\mathcal{D}$. It remains to show that the maximal compatible subsets are algebras. Consider a maximal compatible subset $\mathcal{A}_i$. First, $\emptyset \in \mathcal{A}_i$ as $\emptyset$ is

compatible to all sets in 2^Ω. Second, $\mathcal{A}_i$ is closed under finite intersection, otherwise there would exist a finite combination of elements $A_1, \dots, A_n \subseteq \mathcal{A}_i$ such that $A_\cap := \bigcap_{j=1}^n A_j \in \mathcal{D}$, but $A_\cap \notin \mathcal{A}_i$. Then, one can easily see that $\mathcal{A}_i \cup \{A_\cap\}$ would be a compatible subset which strictly contained $\mathcal{A}_i$. This is impossible, since $\mathcal{A}_i$ is maximal. Finally, $\mathcal{A}_i$ is closed under complement. Consider $A \in \mathcal{A}_i$, we show that $\mathcal{A}_i \cup \{A^c\}$ is again a compatible subset. Let $A_1, \dots, A_n \subseteq \mathcal{A}_i$ be an arbitrary finite collection of subsets, then $A \cap \bigcap_{j=1}^n A_n \in \mathcal{D}$, hence $A^c \cap \bigcap_{j=1}^n A_n \in \mathcal{D}$ [Gudder, 1979] (Lemma 3.6). By maximality of $\mathcal{A}_i$, we then know $A^c \in \mathcal{A}_i$. □

Example 8.8. *The pre-Dynkin system $\mathcal{D}_4$ of Example 8.3 consists of the two algebras, $\{\emptyset, 12, 34, \Omega_4\}$ and $\{\emptyset, 13, 24, \Omega_4\}$.*

Theorem 8.7 simplifies several follow-up observations. Instead of pre-Dynkin systems, we can equivalently consider a set of algebras. However, not every union of algebras is a pre-Dynkin system. If these algebras form a compatibility structure, i.e., a set of maximal π-systems, which are non-empty set systems closed under finite intersections, then their union is a pre-Dynkin system (Definition 8.59 and Theorem 8.60 in Appendix 8.8.1). Analogous results for Dynkin systems and σ-algebras exist and are given in Appendix 8.8.4. In summary, pre-Dynkin systems are set structures which do not allow for arbitrary intersections but can be split into maximal intersectable subsets, their *blocks*.

8.2.2 Probabilities on Pre-Dynkin Systems

We require a notion of probability on a pre-Dynkin system to elaborate the relationship of imprecise probability and the system of precision in the following. Probabilities are classically defined on σ-algebras. We generalize this definition as e.g., stated in [Williams, 1991] (p. 18f) to pre-Dynkin systems.

Definition 8.9 (Probability Measure on a Pre-Dynkin System)**.** *Let $\mathcal{D}$ be a pre-Dynkin system. We call a function $\mu\colon \mathcal{D} \to [0,1]$ a* finitely additive probability measure on $\mathcal{D}$ *if and only if it fulfills the following two conditions:*

(a) Normalization*: $\mu(\emptyset) = 0$ and $\mu(\Omega) = 1$.*

(b) Additivity*: let $A, B \in \mathcal{D}$ and $A \cap B = \emptyset$. Then, $\mu(A \cup B) = \mu(A) + \mu(B)$.*

If condition (b) can be extended to countable subsets of $\mathcal{D}$, then we call μ countably additive probability measure on $\mathcal{D}$*: let $I \subseteq \mathbb{N}$ and $\{A_i\}_{i \in I}$ such that $A_i \in \mathcal{D}$ for all $i \in I$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, $i, j \in I$ and $\bigcup_{i \in I} A_i \in \mathcal{D}$. Then $\mu(\bigcup_{i \in I} A_i) = \sum_{i \in I} \mu(A_i)$.*

For the sake of readability, we use "probability" and "probability measure" exchangeably. Probabilities on pre-Dynkin systems are monotone, i.e., for $A, B \in \mathcal{D}$, if $A \subseteq B$, then

$\mu(A) \leq \mu(B)$. This can be seen when applying Lemma 8.2 and Definition 8.9. But, in contrast to a probability defined on a σ-algebra, a probability on a pre-Dynkin system is not necessarily *modular*, i.e., for $A, B \in \mathcal{D}$, $\mu(A)+\mu(B) = \mu(A \cup B)+\mu(A \cap B)$ [Denneberg, 1994] (p. 16). (It is, however, possible to define modular probabilities on pre-Dynkin systems. This leads to a fixed parametrization of probability functions already on simple examples [Navara and Pták, 1998] (p. 125).) It is that sophisticated interplay of set structure and probability function which leads us through this paper. In particular, why should we consider pre-Dynkin systems?

8.3 Imprecise Probabilities Are Precise on a Pre-Dynkin System

As we now demonstrate, pre-Dynkin systems are, under mild assumptions, the systems of precision. To make this formal, we solely require a normed, conjugate pair of lower and upper probability which fulfill super (resp. sub)-additivity and possibly a continuity assumption.

Theorem 8.10 (Imprecise Probability Induces a (Pre-)Dynkin System)**.** *Let* $\ell \colon 2^\Omega \to [0,1]$ *and* $u \colon 2^\Omega \to [0,1]$ *be two set functions, for which all the following properties hold:*

(a) Normalization: $u(\emptyset) = \ell(\emptyset) = 0$.

(b) Conjugacy: $u(A) = 1 - \ell(A^c)$ *for* $A, A^c \in 2^\Omega$.

(c) Subadditivity of u*: for* $A, B \in 2^\Omega$ *such that* $A \cap B = \emptyset$*, then* $u(A \cup B) \leq u(A) + u(B)$.

(d) Superadditivity of ℓ*: for* $A, B \in 2^\Omega$ *such that* $A \cap B = \emptyset$*, then* $\ell(A \cup B) \geq \ell(A) + \ell(B)$.

Then u *and* ℓ *define a finitely additive probability measure* $\mu \coloneqq u|_{\mathcal{D}} = \ell|_{\mathcal{D}}$ *on the pre-Dynkin system* $\mathcal{D} \coloneqq \{A \in 2^\Omega \colon \ell(A) = u(A)\} \subseteq 2^\Omega$*. If either* u *fulfills*

(e) Continuity from below: for $A_n \in 2^\Omega$ *with* $A_n \subseteq A_{n+1}$ *such that* $\bigcup_{n=1}^\infty A_n = A \in 2^\Omega$*, then*
$\lim_{n\to\infty} u(A_n) = u(A)$,

or ℓ *fulfills*

(e') Continuity from above: for $A_n \in 2^\Omega$ *with* $A_{n+1} \subseteq A_n$ *such that* $\bigcap_{n=1}^\infty A_n = A \in 2^\Omega$*, then*
$\lim_{n\to\infty} \ell(A_n) = \ell(A)$,

then u *and* ℓ *define a countably additive probability measure* $\mu_\sigma \coloneqq u|_{\mathcal{D}_\sigma} = \ell|_{\mathcal{D}_\sigma}$ *on the Dynkin system* $\mathcal{D}_\sigma \coloneqq \{A \in 2^\Omega \colon \ell(A) = u(A)\} \subseteq 2^\Omega$.

Proof. We start proving the first part of the theorem. Let

$$\mathcal{D} \coloneqq \{A \in 2^{\Omega} \colon \ell(A) = u(A)\}. \tag{8.1}$$

We show that $\mathcal{D}$ is a pre-Dynkin system. First, $\emptyset \in \mathcal{D}$ by assumption (a). Second, let $D \in \mathcal{D}$. Then, $u(D^c) = 1 - \ell(D) = 1 - u(D) = \ell(D^c)$ by the conjugacy relation. Third, let $\{A_i\}_{i \in I} \subseteq \mathcal{D}$ for finite $I \subseteq \mathbb{N}$ such that $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

$$\begin{aligned}
\sum_{i \in I} \ell(A_i) &\overset{(d)}{\leq} \ell\left(\bigcup_{i \in I} A_i\right) \\
&\overset{(\star)}{\leq} u\left(\bigcup_{i \in I} A_i\right) \\
&\overset{(c)}{\leq} \sum_{i \in I} u(A_i) \\
&\overset{\text{Equation (8.1)}}{=} \sum_{i \in I} \ell(A_i).
\end{aligned}$$

For $(\star)$, observe that $\ell(A) \leq u(A)$ for all $A \in 2^{\Omega}$, since

$$\ell(A) + \ell(A^c) \leq \ell(A \cup A^c) = 1 = u(A \cup A^c) \leq u(A) + u(A^c),$$

and thus,

$$\begin{aligned}
&\ell(A) + \ell(A^c) \leq u(A) + u(A^c) \\
\Leftrightarrow\ &\ell(A) + 1 - u(A) \leq u(A) + 1 - \ell(A) \\
\Leftrightarrow\ &\ell(A) \leq u(A).
\end{aligned}$$

Concluding, we define $\mu \coloneqq \ell|_{\mathcal{D}} = u|_{\mathcal{D}}$ for which it is trivial to show that it is a finitely additive probability on $\mathcal{D}$.

For the second part, we first notice that continuity from below and from above are equivalent for conjugate set functions on set systems which are closed under complement [Denneberg, 1994] (Proposition 2.3). Next, we show that subadditivity of u and continuity from below (of u) imply σ-subadditivity of u: for $\{A_i\}_{i \in I} \subseteq 2^{\Omega}$ such that $I \subseteq \mathbb{N}$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$ with $i, j \in I$ then $u\left(\bigcup_{i \in I} A_i\right) \leq \sum_{i \in I} u(A_i)$. In the case that I is finite, subadditivity of u is provided by assumption. For infinite I, we can construct an increasing sequence of

sets, namely $B_j = \bigcup_{1 \leq i \leq j} A_i$, so that $B_j \subseteq B_{j+1}$. Furthermore, $\bigcup_{j=1}^{\infty} B_j = \bigcup_{i \in I} A_i$. Thus,

$$\begin{aligned} u\left(\bigcup_{i \in I} A_i\right) &= u\left(\bigcup_{j=1}^{\infty} B_j\right) \\ &\overset{(e)}{=} \lim_{j \to \infty} u\left(B_j\right) \\ &= \lim_{j \to \infty} u\left(\bigcup_{1 \leq i \leq j} A_i\right) \\ &\overset{(d)}{\leq} \lim_{j \to \infty} \sum_{1 \leq i \leq j} u\left(A_i\right) \\ &= \sum_{i \in I} u\left(A_i\right). \end{aligned}$$

The same argument holds analogously for superadditivity and continuity from above of ℓ which is implied by continuity from below and the conjugacy relationship [Denneberg, 1994] (Proposition 2.3). In summary, the proof of the first part can then be applied again, now without the restriction that $I \subseteq \mathbb{N}$ is finite. Instead, it potentially is countable. □

Example 8.11. *Remember,* $\Omega_4 = \{1, 2, 3, 4\}$. *We define* $\ell \colon 2^{\Omega_4} \to [0, 1]$ *by* $\ell(1) = 0$, $\ell(2) = 0.3$, $\ell(3) = 0$, $\ell(4) = 0.3$, $\ell(12) = 0.5$, $\ell(34) = 0.5$, $\ell(13) = 0.2$, $\ell(24) = 0.8$, $\ell(14) = 0.3$, $\ell(23) = 0.3$, $\ell(123) = 0.5$, $\ell(124) = 0.8$, $\ell(134) = 0.5$, $\ell(234) = 0.8$, $\ell(\Omega_4) = 1$ *and* $u \colon 2^{\Omega_4} \to [0, 1]$ *by* $u(A) = 1 - \ell(A^c)$. *It is easy to show that* ℓ *and* u *fulfill the assumptions (a), (b), (c) and (d) in Theorem 8.10. The imprecise probabilities* u *and* ℓ *coincide on* $\{\emptyset, 12, 34, 13, 24, \Omega_4\}$, *the pre-Dynkin system described in Example 8.3. The example is illustrated in Figure 8.2. In this figure* $\ell = \underline{\mu}_{\mathcal{D}_4}$ *and* $u = \overline{\mu}_{\mathcal{D}_4}$.

In summary, imprecise probabilities are, under mild assumptions, precise on a pre-Dynkin system or even a Dynkin system. (The events in the system of precisions are called *unambiguous events* in the decision theory literature [Epstein and Zhang, 2001].) This, importantly, is also the case if the system of precision is strictly larger than the trivial pre-Dynkin systems $\{\emptyset, \Omega\}$. Exemplarily, a pair of conjugate, coherent lower and upper probability (e.g., [Walley, 1991] (§2.7.4)) fulfills the conditions (a)–(d). However, in several cases (e.g., distorted probability distributions), imprecise probabilities are just precise on the *system of certainty*, i.e., the events which possess 0 or 1 probability (Proposition 8.69). Concluding, the system of precision is a pre-Dynkin system $\mathcal{D} \subseteq 2^{\Omega}$. What if we first define precise, finitely additive probabilities on a pre-Dynkin system, i.e., we fix a system of precision? We can then ask for "imprecise probabilities" deduced from this probability which are defined on a larger set structure, e.g., an algebra in which the pre-Dynkin system is contained.

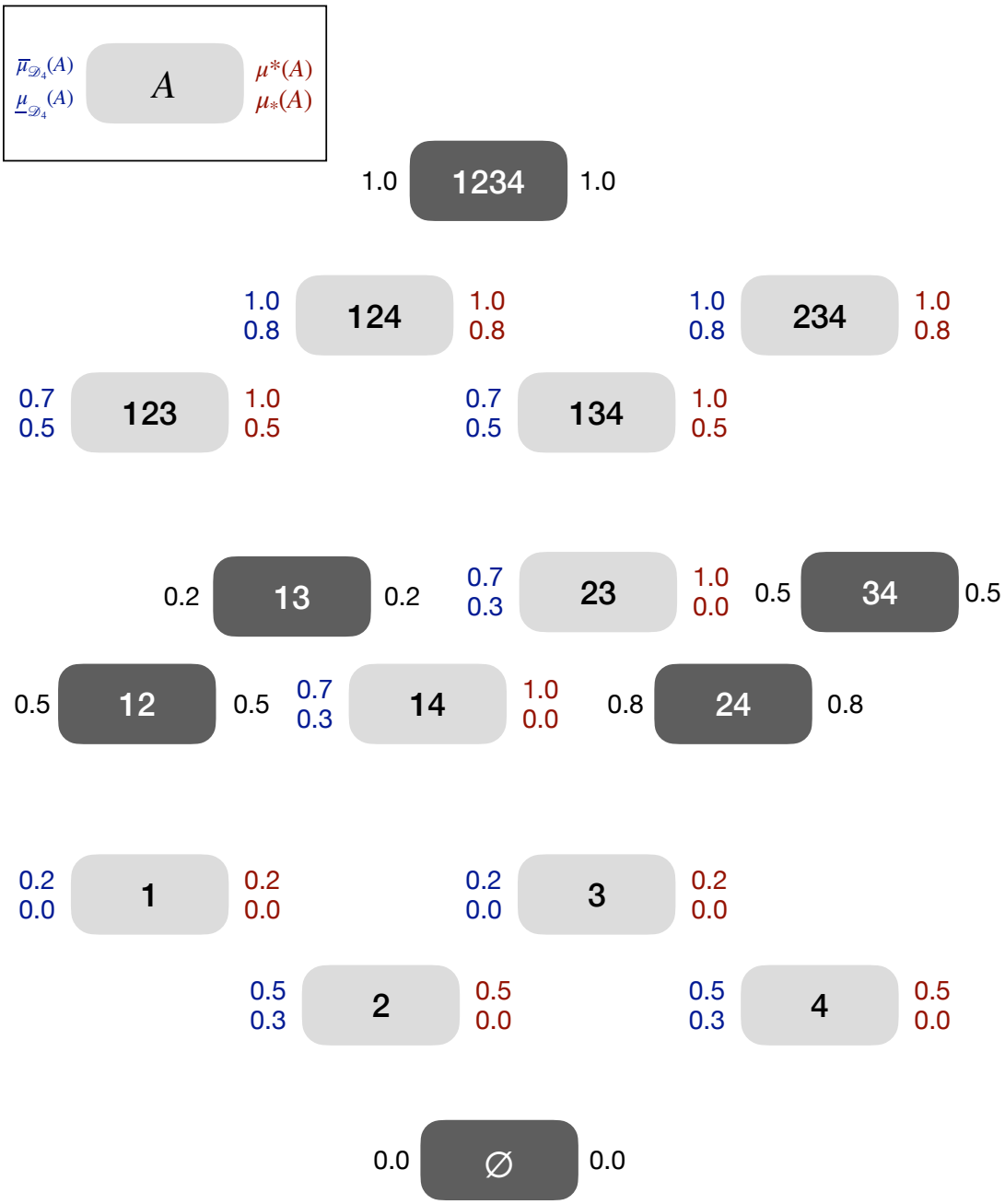


Figure 8.2: Illustration of the running example. The dark elements are contained in the pre-Dynkin system $\mathcal{D}$ on $\Omega = \{1, 2, 3, 4\}$. The lower and upper coherent extension, respectively, the inner and outer extension, are denoted at the sides of the elements in the set system as shown in the example in the left upper corner. Elements in $\mathcal{D}$ possess a precise probability.

8.4 Extending Probabilities on Pre-Dynkin Systems

Precise probabilities on pre-Dynkin system naturally arise in many, distinct, applied scenarios as we argued in the Introduction (Section 8.1). However, we acknowledge that the definition of probabilities on pre-Dynkin systems is mathematically cumbersome. The possibilities to prove standard theorems is very limited as the approaches by Gudder [1973, 1979], Gudder and Zerbe [1981] demonstrate. However, if we consider a probability defined on a pre-Dynkin system as an imprecise probability on a larger set system with a fixed system of precision, we possibly obtain a richer, mathematical toolkit to work with. In this case the larger set system preferably is an algebra in which the pre-Dynkin system is contained. It remains to clarify how we construct the imprecise probability from the precise probability on the pre-Dynkin sytem.

8.4.1 Inner and Outer Extension

A simple but, as we show, unsatisfying solution is the use of an inner and outer measure extension. It does not rely on imposing any conditions on the probability defined on the pre-Dynkin system. We pay for this generality with the few properties that we can derive for the obtained extension.

Proposition 8.12 (Inner and Outer Extension [Zhang, 2002] (Lemma 2.2))**.** *Let $\mathcal{D}$ be a pre-Dynkin system on Ω and μ a finitely additive probability measure on $\mathcal{D}$. The* inner probability measure

$$\mu_*(A) := \sup\{\mu(B) \colon A \supseteq B \in \mathcal{D}\},\ \forall A \in 2^\Omega,$$

and outer probability measure

$$\mu^*(A) := \inf\{\mu(B) \colon A \subseteq B \in \mathcal{D}\},\ \forall A \in 2^\Omega,$$

define $\mu_*, \mu^* : 2^\Omega \to [0,1]$, i.e., all of the following conditions are fulfilled:

(a) Normalization: $\mu^*(\emptyset) = 0$, $\mu_*(\Omega) = 1$.

(b) Conjugacy: $\mu^*(A) = 1 - \mu_*(A^c),\ \forall A \in 2^\Omega$.

(c) Monotonicity: for $A, B \in 2^\Omega$, if $A \subseteq B$, then $\mu^*(A) \leq \mu^*(B)$.

Furthermore, μ_* is superadditive, for $A, B \in 2^\Omega$ if $A \cap B = \emptyset$, then $\mu_*(A \cup B) \geq \mu_*(A) + \mu_*(B)$. But μ^* is not generally subadditive.

Example 8.13. *For $\mathcal{D}_4$, as in Example 8.11, let $\mu\colon \mathcal{D}_4 \to [0,1]$ be defined as $\mu(\emptyset) = 0, \mu(12) = 0.5$, $\mu(34) = 0.5, \mu(13) = 0.2, \mu(24) = 0.8, \mu(\Omega) = 1$. The inner and outer extension of μ on $\mathcal{D}_4$ is $\mu_*(\emptyset) = 0, \mu_*(1) = \mu_*(2) = \mu_*(3) = \mu_*(4) = 0, \mu_*(12) = 0.5, \mu_*(34) = 0.5$, $\mu_*(13) = 0.2, \mu_*(24) = 0.8, \mu_*(14) = 0, \mu_*(23) = 0, \mu_*(123) = 0.5, \mu_*(124) = 0.8$, $\mu_*(134) = 0.5, \mu_*(234) = 0.8, \mu_*(\Omega_4) = 1$ and $\mu^* = 1 - \mu_*$. The inner and outer extension are not coherent (Definition 8.18). In particular, the outer extension is not subadditive: $\mu^*(14) = 1 - \mu_*(23) = 1 > 0.2 + 0.5 = (1 - \mu_*(234)) + (1 - \mu_*(123)) = \mu^*(1) + \mu^*(4)$. The example is illustrated in Figure 8.2.*

In conclusion, the inner and outer extension provides an imprecise probability, which is not necessarily coherent (cf. Definition 8.18) and it does not fulfill the conditions required for Theorem 8.10 to post hoc guarantee that the set of precision is a pre-Dynkin system. We remark that there exist normalized, conjugate, monotone superadditive but not subadditive pairs of probabilities, hence possibly inner and outer probabilities as defined here, whose system of precision is not a pre-Dynkin system (see Example 8.14). For this reason we now explore another, more powerful extension method.

Example 8.14. *Let $\Omega_3 := \{1,2,3\}$. The probability pair defined by $\underline{\nu}(\emptyset) = \overline{\nu}(\emptyset) = 0$, $\underline{\nu}(1) = \overline{\nu}(1) = 0.2$, $\underline{\nu}(2) = \overline{\nu}(2) = 0.2$ and $\underline{\nu}(3) = 0, \overline{\nu}(3) = 0.6$ is precise on $\{\emptyset, 1, 2, 23, 13, \Omega_3\}$, which is obviously not a pre-Dynkin system.*

8.4.2 Extendability and Its Equivalence to Coherence

In the following, we try to entirely embed pre-Dynkin systems equipped with a probability into larger algebras. Then, we extend the probability defined on the pre-Dynkin system in all possible ways to probabilities on the algebra. It turns out that this embedding is only possible under certain conditions on the probability defined on the pre-Dynkin system. We call this condition *extendability*. For the sake of generality, we focus on the extension of finitely additive probabilities from pre-Dynkin systems to algebras here. We treat countably additive probabilities, Dynkin systems and σ-algebras in Appendix 8.8.4. In addition, all results until Section 8.4.3 can be formulated in more general terms for non-structured set systems. For the sake of simplicity, we remain within the setting of probabilities defined on pre-Dynkin systems in this work.

Extendability is the property that a probability measure defined on a pre-Dynkin system can be extended to a probability measure on an algebra containing the pre-Dynkin system. Formally:

Definition 8.15 (Extendability)**.** *Let $\mathcal{D}$ be a pre-Dynkin system on Ω. We call a finitely additive probability measure μ on $\mathcal{D}$* extendable *to 2^Ω if and only if there is a finitely additive probability measure $\nu\colon 2^\Omega \to [0,1]$ such that $\nu|_{\mathcal{D}} = \mu$.*

We defined extendability with respect to the power set 2^Ω. In fact, any relativization to an arbitrary sub-algebra of 2^Ω is equivalent. A finitely additive probability defined on $\mathcal{D}$ is extendable to any sub-algebra of 2^Ω which contains $\mathcal{D}$ if and only if it is extendable to 2^Ω [Rao and Rao, 1983] (Theorem 3.4.4).

The definition is non-vacuous [Gudder, 1984, De Simone et al., 2007]. For instance, a probability measure on a pre-Dynkin System is not generally extendable to a measure on the generated algebra (e.g., Example 3.1 in [Gudder, 1984]). If a probability is extendable, its extension is in general non-unique.

Extendability of probabilities on (pre-)Dynkin systems has already been part of discussions in quantum probability since 1969 [Gudder, 1969] up to more current times [De Simone and Pták, 2010]. It as well has been of interest in frequential probability theory [Schurz and Leitgeb, 2008] (Theorem 2). Several necessary and/or sufficient conditions on the structure of $\mathcal{D}$ and/or the values of μ are known [Gudder, 1984, De Simone et al., 2007, De Simone and Pták, 2010]. We present here a sufficient and necessary condition discovered by Horn and Tarski [1948] and restated in [Rao and Rao, 1983] (Theorem 3.2.10). In fact, Theorem 8.16 can be stated for a more general definition of probabilities on arbitrary set systems e.g., [Rao and Rao, 1983] (Theorem 3.2.10), [Simone and Pták, 2006] (Proposition 2.2). For the sake of simplicity, we restrict this result to pre-Dynkin systems and probabilities defined on pre-Dynkin systems.

Theorem 8.16 (Extendability Condition [Rao and Rao, 1983] (Theorem 3.2.10))**.** *Let $\mathcal{D}$ be a pre-Dynkin system on Ω. A finitely additive probability measure μ on $\mathcal{D}$ is extendable to 2^Ω if and only if*

$$\sum_{k=1}^{m} \chi_{B_k}(\omega) - \sum_{j=1}^{n} \chi_{A_j}(\omega) \geq 0, \ \forall \omega \in \Omega \implies \sum_{k=1}^{m} \mu(B_k) - \sum_{j=1}^{n} \mu(A_j) \geq 0 \tag{8.2}$$

for all finite families of sets in $\mathcal{D}$: $A_1, \ldots, A_n, B_1, \ldots, B_m \in \mathcal{D}$.

Example 8.17. *For $\mathcal{D}_4$ as in Example 8.11 let $\mu \colon \mathcal{D}_4 \to [0,1]$ be defined as $\mu(\emptyset) = 0, \mu(12) = 0.5, \mu(34) = 0.5, \mu(13) = 0.2, \mu(24) = 0.8, \mu(\Omega) = 1$. The probability μ on $\mathcal{D}_4$ meets the extendability condition.*

Extendability proves to be more than a helpful mathematical property for embedding pre-Dynkin systems and their respective probabilities into algebras. Whether a probability defined on $\mathcal{D}$ can be extended to a probability on 2^Ω is directly connected to the question whether the probability measure on $\mathcal{D}$ is coherent in the sense of [Walley, 1991] (pp. 68, 84) or not. Coherence is a minimal consistency requirement for probabilistic descriptions which has been introduced in the fundamental work of De Finetti [1974/2017] and developed by Walley [1991]. Shortly summarizing, an incoherent imprecise probability is tantamount to

an irrational betting behavior, thus the name. Thus, extendability is, besides its mathematical convenience, a desirable property of probabilities in pre-Dynkin settings.

We adapt here the definition of coherence of previsions in [Walley, 1991] (Definition 2.5.1) to probabilities.

Definition 8.18 (Coherent Probability)**.** *Let $\mathcal{A} \subseteq 2^\Omega$ be an arbitrary collection of subsets. A set function $\underline{\nu}\colon \mathcal{A} \to [0,1]$ is a* coherent lower probability *if and only if*

$$\sup_{\omega\in\Omega} \sum_{i=1}^{j} (\chi_{A_i}(\omega) - \underline{\nu}(A_i)) - m(\chi_{A_0}(\omega) - \underline{\nu}(A_0)) \geq 0,$$

for any non-negative $n, m \in \mathbb{N}$ and any $A_0, A_1, \dots A_n \in \mathcal{A}$. If $\mathcal{A}$ is closed under complement, the conjugate coherent upper probability *is given by $\overline{\nu}(A) \coloneqq 1 - \underline{\nu}(A^c)$ for all $A \in \mathcal{A}$. If furthermore $\overline{\nu}(A) = \underline{\nu}(A)$ for all $A \in \mathcal{A}$, we call $\nu \coloneqq \overline{\nu}$ a* coherent additive probability*.*

At first sight, the Horn-Tarski condition given in Theorem 8.16 and the coherence condition presented here already appear similar. This becomes even more apparent in Walley's reformulation of coherence for additive probabilities [Walley, 1991] (Theorem 2.8.7). In the following, we show that this superficial similarity is indeed based on a rigorous link. Surprisingly, Walley did not mention Horn and Tarski's work in his foundational book.

Theorem 8.19 (Extendability Equals Coherence)**.** *Let $\mathcal{D}$ be a pre-Dynkin system on Ω. A finitely additive probability measure μ on $\mathcal{D}$ is extendable to 2^Ω if and only if it is a coherent additive probability on $\mathcal{D}$.*

Proof. If μ is a coherent additive probability on $\mathcal{D}$, then the linear extension theorem [Walley, 1991] (Theorem 3.4.2) applies. Hence, a coherent additive probability $\nu\colon 2^\Omega \to [0,1]$ exists, such that $\nu|_\mathcal{D} = \mu$. In particular, ν is a finitely additive probability following Definition 8.9 on 2^Ω [Walley, 1991] (Theorem 2.8.9).

For the converse direction, we observe that if μ possess an extension following Definition 8.15, then such an extension is a finitely additive probability on 2^Ω following Definition 8.9. Hence, Walley [1991, Theorem 2.8.9] guarantees that the extension is a coherent additive probability (Definition 8.18). Any restriction to a subdomain $\mathcal{D} \subseteq 2^\Omega$ keeps the probability coherent and additive. □

The linear extension theorem in Walley [Walley, 1991] (Theorem 3.4.2) used here is a generalization of de Finetti's fundamental theorem of probability [De Finetti, 1974/2017] (Theorems 3.10.1 and 3.10.7). De Finetti's theorem is furthermore interesting, as he explicitly states that a coherent additive probability defined on an arbitrary collection of sets can be extended in a precise way (so lower and upper probability coincide) to some sets. De

Finetti does not characterize this collection. Our Theorem 8.10, however, gives an answer to this question: the collection forms a pre-Dynkin system. It is possible to recover the statement by an argumentation using credal sets. Roughly, credal sets of coherent lower probabilities consist of all dominating coherent additive probabilities. In the case here, extendability guarantees that the credal set of a probability defined on a pre-Dynkin system is non-empty. Furthermore, Walley [1991] introduced another slightly weaker concept than coherence, called "avoiding sure loss" which is equivalent to non-empty credal sets [Walley, 1991] (Corollary 3.3.4). Coherence and avoiding sure loss are equivalent for probabilities defined on pre-Dynkin systems [Troffaes and De Cooman, 2014] (Theorem 4.12).

Theorem 8.19 provides a missing link between two strands of work: on the one hand, probabilities on pre-Dynkin systems and related weak set structures have been closely investigated in foundational quantum probability theory [Gudder, 1969, 1979] and decision theory [Epstein and Zhang, 2001, Zhang, 2002]. On the other hand, coherent probabilities are central to imprecise probability, in particular, the more general formulations of coherent previsions and risk measures [Walley, 1991, Delbaen, 2002, Pelessoni and Vicig, 2003].

The reader familiar with the literature on imprecise probability might well not be surprised by the equivalence of extendability and coherence. We still think that this link is indeed valuable to be spelled out explicitly here. The concept of extendability and coherence have been developed separately in two communities with different goals in focus. Coherence tries to capture "rational" betting behavior [De Finetti, 1974/2017, Walley, 1991]. Extendability links to what is sometimes called "quantum weirdness". For readers familiar with quantum probability, we sketch the relationship between extendability and the concepts of compatibility and contextuality in the following.

Extendability, Compatibility and Contextuality

Extendability in quantum theory tightly interacts with a series of properties and concepts which pervade discussions about the "specialness" of quantum theory in comparison to other classical physical theories: compatibility, contextuality, hidden variables and more. To be concrete, two measurements are compatible if, for any initial state, usually represented as a probability distribution (cf. [Gudder, 1979]), there exists a joint measurement such that a fixed joint distribution for both measurement outcomes exists, whose marginals are the distribution of the single measurement [Busch et al., 2012, Xu and Cabello, 2018]. We remark that this notion of compatibility is related, but not directly, to our Definition 8.4. If measurements are incompatible, then there are potentially still states such that a joint distribution of measurements exists. Only in the cases that no joint distribution of measurements exists, i.e., extendability is not provided, a measuring observer observes *contextual* behavior [Xu and Cabello, 2018]. Translated to the language of imprecise probability, contextuality

amounts to non-coherence of a probabilistic description. Compatibility, in contrast, is a structural notion. If any finitely additive probability on a pre-Dynkin system is extendable, then the pre-Dynkin system, very roughly, resembles compatible measurements. We are indeed not the first to notice intriguing links between imprecise probability and concepts therein to quantum mechanics. Benavoli and collaborators recovered the four postulates of quantum mechanics with desirability as a starting point [Benavoli et al., 2016] (Desirability is a relatively general framework for imprecise probability [Walley, 2000].)

Extendability and Marginal Problem

A particularly helpful special case of probabilities defined on pre-Dynkin systems are marginal scenarios. Marginal scenarios, as we already stated in Section 8.2, are settings in which several (classical) marginal probability distributions, for instance, for three random variables p_X, p_Y, p_Z, are given, but a joint distribution $p_{X,Y,Z}$ is not specified. We note that we do not define random variables, marginals and probability distributions rigorously here. Instead, we argue on a vague level to convey the intuition. Such marginal scenarios can be expressed via probabilities on pre-Dynkin systems as laid out in detail in [Vorob'ev, 1962, Kellerer, 1964] and [Gudder, 1984] (Example 4.2). Central to marginal scenarios is the relation structure of the marginals. For instance, $p_{X,Y}, p_Z$ as well as $p_{X,Y}, p_{Y,Z}$ or $p_{X,Y}, p_{Y,Z}, p_{X,Z}$ form marginal scenarios. The so-called marginal problem for marginal scenarios now asks whether for all instantiations of the marginal scenario there exists a joint distribution, e.g., for all $p_{X,Y}, p_Z$ a joint distribution $p_{X,Y,Z}$ exists. In contrast, for some assignments $p_{X,Y}, p_{Y,Z}, p_{X,Z}$ there does not exist a corresponding $p_{X,Y,Z}$. (The attentive reader might have noticed the similarity to the notion of compatibility, for good reasons [Budroni et al., 2022] (§V.B.2). Marginal scenarios are used to represent multi-measurement settings and compatibility among the measurements.) This question was, some while ago, asked for probabilities on finite spaces [Vorob'ev, 1962], countably additive probabilities [Kellerer, 1964], finitely additive probabilities (cf. [Maharam, 1972]) and recently for even more general probability models—sets of desirable gambles [Miranda and Zaffalon, 2018, Casanova et al., 2022]. A recurring theme in all those studies is the so-called running intersection property which characterizes all solvable marginal problems. In other words, the running intersection property guarantees that *every* instantiation of the marginal distributions is *extendable*. But there exist marginal problems for which only *specific* instantiations of the marginal distributions allow for extendability.

8.4.3 Coherent Extension

A probability on a pre-Dynkin system $\mathcal{D}$, even when extendable, only allows for probabilistic statements on $\mathcal{D}$ itself. However, extendability guarantees that a "nice" embedding into a

larger system of measurable sets exists. More specifically, extendability expressed in terms of credal sets provides a well-known tool for the worst-case extension of a probability from a pre-Dynkin system to a larger algebra.

If a finitely additive probability on a pre-Dynkin system is extendable, then we can obtain lower and upper probabilities of events which are not in the pre-Dynkin system but on a larger algebra. We follow the idea of natural extensions, e.g., as described by [Walley, 1991] (p. 136). In particular, [Walley, 1991] (Theorem 3.3.4 (b)) directly applies as long as a probability on a pre-Dynkin system is extendable.

Corollary 8.20 (Coherent Extension of Probability)**.** *Let $\mathcal{D}$ be a pre-Dynkin system on Ω. For a finitely additive probability measure μ on $\mathcal{D}$ we define the* credal set

$$M(\mu, \mathcal{D}) \coloneqq \{\nu \in \Delta \colon \nu(A) = \mu(A),\ \forall A \in \mathcal{D}\}\,.$$

If μ on $\mathcal{D}$ is extendable to 2^Ω, then $\forall A \in 2^\Omega$,

$$\underline{\mu}_{\mathcal{D}}(A) \coloneqq \inf_{\nu \in M(\mu,\mathcal{D})} \nu(A), \quad \overline{\mu}_{\mathcal{D}}(A) \coloneqq \sup_{\nu \in M(\mu,\mathcal{D})} \nu(A).$$

define a coherent lower respectively upper probability on 2^Ω.

Example 8.21. *The coherent extension of μ on $\mathcal{D}_4$ as defined in Example 8.17 is $\underline{\mu}_{\mathcal{D}_4} = \ell$ and $\overline{\mu}_{\mathcal{D}_4} = u$ where, ℓ and u are defined as in Example 8.11 (cf. [Walley, 1991] (p. 122)). Figure 8.2 illustrates the coherent extensions. Even though coherent, $\underline{\mu}_{\mathcal{D}_4}$ is neither supermodular nor submodular:*

$$\underline{\mu}_{\mathcal{D}_4}(12) + \underline{\mu}_{\mathcal{D}_4}(13) = 0.5 + 0.2 > 0.5 + 0.0 = \underline{\mu}_{\mathcal{D}_4}(123) + \underline{\mu}_{\mathcal{D}_4}(1),$$
$$\underline{\mu}_{\mathcal{D}_4}(1) + \underline{\mu}_{\mathcal{D}_4}(2) = 0.0 + 0.3 < 0.5 + 0.0 = \underline{\mu}_{\mathcal{D}_4}(12) + \underline{\mu}_{\mathcal{D}_4}(\emptyset).$$

This implies that as well $\overline{\mu}_{\mathcal{D}_4}$ is neither supermodular nor submodular [Denneberg, 1994] (Proposition 2.3).

These lower and upper probabilities allow for at least two interpretations: We can assume that a precise probability on a pre-Dynkin systems $\mathcal{D} \subseteq 2^\Omega$ just reveals its values on $\mathcal{D}$, but is actually defined over 2^Ω. Then the lower and upper probability constitute lower and upper bounds of the precise "hidden probability" on 2^Ω, which is solely accessible on $\mathcal{D}$. On the other hand, we can even reject the existence of such precise "hidden probability". Then lower and upper probability *are* the inherently imprecise probability of an event in 2^Ω but not in $\mathcal{D}$. (As remarked by Walley [1991] (p. 138), De Finetti [1974/2017] surprisingly only considered the first mentioned interpretation.)

The obtained lower and upper probabilities represent the imprecise interdependencies be-

tween all events of precise probabilities. We illustrate this statement: in the variety of updating methods in imprecise probability we pick the generalized Bayes' rule [Walley, 1991] (§6.4) to exemplarily compute the conditional probability of two events for the coherent extension of a probability from a pre-Dynkin system. For $A, B \in \mathcal{D}$ such that $\mu(B) > 0$, the generalized Bayes' rule gives [Walley, 1991] (Theorem 6.4.2):

$$\overline{\mu}_{\mathcal{D}}(A|B) \coloneqq \sup_{\nu \in M(\mu,\mathcal{D})} \frac{\nu(A \cap B)}{\nu(B)} = \frac{\sup_{\nu \in M(\mu,\mathcal{D})} \nu(A \cap B)}{\mu(B)} = \frac{\overline{\mu}_{\mathcal{D}}(A \cap B)}{\mu(B)}, \qquad \forall A, B \in \mathcal{D}$$

We can easily rearrange the above as $\overline{\mu}_{\mathcal{D}}(A \cap B) = \overline{\mu}_{\mathcal{D}}(A|B)\mu(B)$. In this case the imprecision of the probability of the intersected event is *purely* controlled by the conditional probability $\overline{\mu}(A|B)$ and *not* by the marginal, which is precise. So, the imprecision captured by the lower and upper probabilities locates solely in the interdependency of the events. We remark that Dempster's rule gives the same conditional probability here [Dempster, 1967].

8.4.4 Inner and Outer Extension Is More Pessimistic Than Coherent Extension

We have presented two extension methods for probabilities defined on pre-Dynkin systems. We relate the methods in the following. In the case of an extendable probability we can guarantee the following inequalities to hold.

Theorem 8.22 (Extension Theorem—Finitely Additive Case)**.** *Let $\mathcal{D}$ be a pre-Dynkin system on Ω and μ a finitely additive probability on $\mathcal{D}$ which is extendable to 2^Ω. Then,*

$$\mu_*(A) \leq \underline{\mu}_{\mathcal{D}}(A) \leq \overline{\mu}_{\mathcal{D}}(A) \leq \mu^*(A), \qquad \forall A \in 2^\Omega.$$

Proof. Since $\mathcal{D} \subseteq 2^\Omega$, we easily obtain

$$\begin{aligned}
\mu_*(A) &= \sup\{\mu(B) \colon A \supseteq B \in \mathcal{D}\} \\
&= \sup\left\{\underline{\mu}_{\mathcal{D}}(B) \colon A \supseteq B \in \mathcal{D}\right\} \\
&\leq \sup\left\{\underline{\mu}_{\mathcal{D}}(B) \colon A \supseteq B \in 2^\Omega\right\} \\
&= \underline{\mu}_{\mathcal{D}}(A),
\end{aligned}$$

for all $A \in 2^\Omega$. The other inequalities follow by the conjugacy of inner and outer measure and lower and upper coherent extension. □

In words, Theorem 8.22 states that the inner and outer extension is more "pessimistic" than the coherent extension. We use "pessimistic" in the sense of giving a looser bound for the probabilities assigned to elements not in the pre-Dynkin system $\mathcal{D}$ but in 2^Ω. We

remark that Walley has shown that for a probability defined on a set algebra, inner and lower coherent extension, as well as outer and upper coherent extension, coincide [Walley, 1991] (Theorem 3.1.5). Walley even characterizes the sets for which a unique coherent extension exists [Walley, 1991] (Theorem 3.1.9). In Appendix 8.8.4, we demonstrate analogous inequalities for countably additive probabilities on Dynkin systems.

8.5 The Credal Set and Its Relation to Pre-Dynkin System Structure

In the earlier parts of the paper, we derived pre-Dynkin systems as the system of precision for relatively general imprecise probabilities. Then, we showed that, under extendability conditions, a precise probability on a pre-Dynkin system gives rise to a coherent imprecise probability on an encompassing algebra. In other words, imprecise probabilities can be "mapped" to pre-Dynkin systems and vice versa. We concretize these mappings in the following. This manifestation then reveals structure in the interplay between the systems of precision, i.e., pre-Dynkin systems, and coherent imprecise probabilities. In particular, we argue that the order structure of pre-Dynkin systems can be mapped to the space of finitely additive probabilities. This provides a (lattice) duality for coherent imprecise probabilities with precise probabilities on pre-Dynkin systems. More concretely, the duality allows for the interpolation from imprecise probabilities which are precise on "all" events to imprecise probabilities which are precise only on the empty set and the entire set.

In the following discussion, we assume, in addition to the technicalities presented in Section 8.1.1, that a fixed finitely additive probability on 2^Ω, which we call ψ, is given. The finitely additive probability ψ with the algebra 2^Ω and the base set Ω constitute our "base measure space" analogous to the choice of a base measure space in the theory of coherent risk measures [Delbaen, 2002]. The probability ψ can be interpreted as the pivot point around which we define increasingly imprecise coherent lower and upper probabilities. The choice of ψ particularly influences the sets of measure zero, which will play a major role in characterizing so-called bipolar-closed set systems. In comparison to the previous sections, we use ψ instead of μ as "reference measure" to emphasize the difference that μ was defined on a relatively arbitrary pre-Dynkin systems $\mathcal{D}$ on Ω, while ψ is defined and fixed on the algebra 2^Ω on Ω.

8.5.1 Credal Set Function Maps From Pre-Dynkin Systems to Coherent Probabilities

Equipped with a reference measure ψ we define the credal set function. The name arises due to its close link to the credal set as defined in Corollary 8.20.

Definition 8.23 (Credal Set Function)**.** *Let Δ be the set of all finitely additive probabilities on 2^Ω. For a fixed, finitely additive probability $\psi \in \Delta$, we call*

$$m_\psi \colon 2^{2^\Omega} \to 2^\Delta, \quad m_\psi(\mathcal{A}) \coloneqq \{\nu \in \Delta \colon \nu(A) = \psi(A),\ \forall A \in \mathcal{A}\},$$

the credal set function*.*

Example 8.24. *Let $\Omega_4 = \{1,2,3,4\}$ as in Example 8.3. With abuse of notation, we define the probability $\psi \colon 2^{\Omega_4} \to [0,1]$ via its corresponding point on the simplex $\psi \in \Delta$, $\psi_1 = 0.2, \psi_2 = 0.3, \psi_3 = 0.5, \psi_4 = 0$. It follows, e.g., $m_\psi(\{12,3\}) = \{\nu \in \Delta \colon \nu_1 + \nu_2 = 0.5, \nu_3 = 0.5\}$. In the subsequent examples, we implicitly assume the here defined ψ.*

We stress that although not notated explicitly, the credal set function depends upon the choice of ψ. For a fixed ψ on 2^Ω, the credal set function m_ψ maps a subset of the algebra 2^Ω to the set of all finitely additive probabilities which coincide with ψ on this subset. It should be noticed that by definition of ψ, $m_\psi(\mathcal{A}) \neq \emptyset$ for every non-empty $\mathcal{A} \subseteq 2^\Omega$, because $\psi \in m_\psi(\mathcal{A})$ for every $\mathcal{A} \subseteq 2^\Omega$.

This defined mapping now simplifies our discussion about how pre-Dynkin systems and imprecise probabilities correspond. For instance, one can easily see that the extreme case $\mathcal{A} = 2^\Omega$ corresponds to $m_\psi(\mathcal{A}) = \{\psi\}$ and $\mathcal{A} = \emptyset$ to $m_\psi(\mathcal{A}) = \Delta$. More generally, we observe the following two properties of the credal set function.

Proposition 8.25 (Credal Set Function is Invariant to Pre-Dynkin Hull)**.** *Let m_ψ be the credal set function. For any $\mathcal{A} \subseteq 2^\Omega$,*

$$m_\psi(\mathcal{A}) = m_\psi(\mathrm{D}(\mathcal{A})).$$

Proof. We need to show that

$$\{\nu \in \Delta \colon \nu(A) = \psi(A),\ A \in \mathcal{A}\} = \{\nu \in \Delta \colon \nu(A) = \psi(A),\ A \in \mathrm{D}(\mathcal{A})\}.$$

The set inclusion of the right hand side in the left hand side is trivial. For the reverse direction, consider an element $\nu \in \Delta$ such that $\nu(A) = \psi(A)$ for $A \in \mathcal{A}$. Let

$$\mathcal{H} \coloneqq \{A \in \mathrm{D}(\mathcal{A}) \colon \nu(A) = \psi(A)\}.$$

By Theorem 8.10 $\mathcal{H}$ is a pre-Dynkin system. Since $\mathcal{A} \subseteq \mathcal{H}$ we know $\mathrm{D}(\mathcal{A}) \subseteq \mathcal{H}$. Hence, $\nu(A) = \mu(A)$ for $A \in \mathrm{D}(\mathcal{A})$. This gives the desired inclusion. We remark that for $\mathcal{A} = \emptyset$ the equality still holds, since $\mathrm{D}(\emptyset) = \{\emptyset, \Omega\}$. □

Example 8.26. *The credal set* $m_\psi(\{12,3\}) = \{\nu \in \Delta \colon \nu_1 + \nu_2 = 0.5, \nu_3 = 0.5\}$ *given in Example 8.24 nicely illustrates Proposition 8.25:*

$$m_\psi(\{12,3\}) = \{\nu \in \Delta \colon \nu_1 + \nu_2 = 0.5, \nu_3 = 0.5, \nu_4 = 0\} = m_\psi(\{\emptyset, 12, 3, 4, 34, 123, 124, \Omega_4\}).$$

Proposition 8.27 (Credal Set Function Maps to Weak*-Closed Convex Sets)**.** *Let* m_ψ *be the credal set function. For every non-empty* $\mathcal{A} \subseteq 2^\Omega$, $m_\psi(\mathcal{A})$ *is weak***-closed convex.*

Proof. The reference probability ψ is by definition coherent. Hence, for all non-empty $\mathcal{A} \subseteq 2^\Omega$, the set $m_\psi(\mathcal{A})$ is the set of all $\nu \in \Delta$ which dominate ψ. This set is, by Theorem 3.6.1 in [Walley, 1991], weak*-closed and convex. □

In words, the credal set on some set system coincides with the credal set on its generated pre-Dynkin system. And the credal set of probabilities is always weak*-closed and convex. Proposition 8.25 allows us to work with credal sets of arbitrary set systems instead of the entire pre-Dynkin system. Thus, it resembles the well-known π-λ-Theorem, which is fundamental to classical probability theory [Williams, 1991] (Lemma A.1.3). On the other hand, this result justifies our focus on pre-Dynkin systems instead of arbitrary set systems. We do not lose generality when considering pre-Dynkin systems instead of non-structured sets of sets.

Proposition 8.27 guarantees that the images of the credal set function behave "nicely". Specifically, these weak*-closed convex sets correspond to coherent previsions, i.e., generalizations of coherent probabilities as already stated by Walley [1991] (Theorem 3.6.1). We elaborate this observation in Section 8.5.4. In conclusion, credal set functions map pre-Dynkin systems to coherent probabilities. What about the reverse mapping?

8.5.2 The Dual Credal Set Function

The following is a natural definition of a dual credal set function. We justify this name by Proposition 8.29 below.

Definition 8.28 (Dual Credal Set Function)**.** *Let* Δ *be the set of all finitely additive probabilities on* 2^Ω*. Fix a finitely additive probability* ψ *on* 2^Ω*. We call*

$$m_\psi^\circ \colon 2^\Delta \to 2^{2^\Omega}, \quad m_\psi^\circ(Q) \coloneqq \{A \in 2^\Omega \colon \nu(A) = \psi(A), \forall \nu \in Q\}$$

the dual credal set function.

The dual credal set function also depends upon ψ, but we do not notate this explicitly. The dual credal set function maps an arbitrary set of finitely additive probability measures on 2^Ω

to the (largest) set of events on which all contained probabilities coincide. We remark that each set of finitely additive probability measures can be linked to an imprecise probability.

We suggestively called the antagonist to the credal set function the "dual credal set function". The duality appearing here is a well-known fundamental relationship between partially ordered sets: a Galois connection. A *Galois connection* is a pair of mappings $f\colon X \to Y$ and $f^\circ\colon Y \to X$ on partially ordered sets $(Y, \leq)$ and $(X, \leq)$, which preserves order structure (cf. Corollary 8.30). More formally, f and f° are a Galois connection if and only if for all $x \in X, y \in Y$, $x \leq f(y) \Leftrightarrow y \leq f^\circ(x)$ [Birkhoff, 1940] (§V.8). Galois connections, even though they do *not* form an order isomorphism, induce a lattice duality. We exploit this lattice duality to provide an order-theoretic interpolation from no compatibility at all to full compatability. For this purpose, we first establish the Galois connection.

Proposition 8.29 (Galois Connection by (Dual) Credal Set Function)**.** *The credal set function m_ψ and the dual credal set function m_ψ° form a Galois connection.*

Proof. m_ψ and m_ψ° form a Galois connection if and only if $\mathcal{A} \subseteq m_\psi^\circ(Q) \Leftrightarrow Q \subseteq m_\psi(\mathcal{A})$ [Birkhoff, 1940] (§V.8). First, we show the left to right implication. We assume $\mathcal{A} \subseteq m_\psi^\circ(Q)$, i.e. every $\nu \in Q$ coincides with ψ on $\mathcal{A}$. Hence,

$$\begin{aligned}\nu \in Q &\Rightarrow \nu(A) = \psi(A), \forall A \in \mathcal{A} \\ &\Rightarrow \nu \in \{\nu' \in \Delta\colon \nu'(A) = \psi(A), \forall A \in \mathcal{A}\} = m_\psi(\mathcal{A}).\end{aligned}$$

In case of the right to left implication, we suppose $Q \subseteq m_\psi(\mathcal{A})$. Thus,

$$\begin{aligned}A \in \mathcal{A} &\Rightarrow \nu(A) = \psi(A), \forall \nu \in Q \\ &\Rightarrow A \in \{A' \in 2^\Omega\colon \nu(A') = \psi(A'), \forall \nu \in Q\} = m_\psi^\circ(Q).\end{aligned}$$

□

The mappings involved in the Galois connection are antitone, i.e., they reverse the order structure from domain to codomain. Their pairwise application is extensive, i.e., the image of an object contains the object. In summary, the following rules of calculation hold:

Corollary 8.30 (Rules for (Dual) Credal Set Function)**.** *Let m_ψ be the credal set function and m_ψ° be the dual credal set function. For arbitrary $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A} \subseteq 2^\Omega$ and $Q_1, Q_2, Q \subseteq \Delta$,*

$$\begin{aligned}&\mathcal{A}_1 \subseteq \mathcal{A}_2 \Rightarrow m_\psi(\mathcal{A}_2) \subseteq m_\psi(\mathcal{A}_1), \quad Q_1 \subseteq Q_2 \Rightarrow m_\psi^\circ(Q_2) \subseteq m_\psi^\circ(Q_1), && \textit{(antitone)} \\ &Q \subseteq m_\psi(m_\psi^\circ(Q)), \quad \mathcal{A} \subseteq m_\psi^\circ(m_\psi(\mathcal{A})), && \textit{(extensive)} \\ &m_\psi(\mathcal{A}) = m_\psi(m_\psi^\circ(m_\psi(\mathcal{A}))), \quad m_\psi^\circ(Q) = m_\psi^\circ(m_\psi(m_\psi^\circ(Q))), && \textit{(pseudo-inverse).}\end{aligned}$$

Proof. [Birkhoff, 1940] (§V.7 and V.8) □

Proposition 8.29 provides a tool to further investigate the dual credal set function. The reader might have noticed the similarity of the dual credal set function and the main question of Section 8.3: given a lower and upper probability, on which set systems do both coincide? In fact, we obtain an analogous result to Theorem 8.10, and again an imprecise probability is mapped to the set of events on which it is precise.

Proposition 8.31 (Dual Credal Set Function Maps to Pre-Dynkin Systems)**.** *Let m_ψ° be the dual credal set function. For all non-empty $Q \subseteq \Delta$, $m_\psi^\circ(Q)$ is a pre-Dynkin system.*

Proof. We show the statement by establishing the equality $\mathrm{D}(m_\psi^\circ(Q)) = m_\psi^\circ(Q)$. Trivially, $\mathrm{D}(m_\psi^\circ(Q)) \supseteq m_\psi^\circ(Q)$. Furthermore,

$$\mathrm{D}(m_\psi^\circ(Q)) \overset{C.8.30}{\subseteq} m_\psi^\circ(m_\psi(\mathrm{D}(m_\psi^\circ(Q)))) \overset{P.8.25}{=} m_\psi^\circ(m_\psi(m_\psi^\circ(Q))) \overset{C.8.30}{=} m_\psi^\circ(Q).$$

□

Proposition 8.32 (Dual Credal Set Function is Invariant to Weak*-Closed Convex Hull)**.** *Let m_ψ° be the dual credal set function. For any $Q \subseteq \Delta$, $m_\psi^\circ(Q) = m_\psi^\circ(\overline{\mathrm{co}}\ Q)$.*

Proof. By Corollary 8.30, $m_\psi^\circ(Q) = m_\psi^\circ(m_\psi(m_\psi^\circ(Q)))$. Furthermore, $Q \subseteq m_\psi(m_\psi^\circ(Q))$. Via, Proposition 8.27 we obtain $\overline{\mathrm{co}}\ Q \subseteq m_\psi(m_\psi^\circ(Q))$. Hence, the result follows. □

Example 8.33. *Let $Q = \{\nu\}$ where we identify the probability ν on 2^{Ω_4} with an element $\nu \in \Delta_4$. For instance, $\nu_1 = 0, \nu_2 = 0.5, \nu_3 = 0, \nu_4 = 0.5$. It is easy to see that $m_\psi^\circ(Q) = \mathrm{D}(\{12, 3\})$ as given in Example 8.26.*

In summary, credal set functions map pre-Dynkin systems to weak*-closed convex credal sets. Dual credal set functions map (weak*-closed convex) credal sets to pre-Dynkin systems. In addition, the two functions form a Galois connection. In fact, every Galois connection defines closure operators, i.e., extensive, monotone and idempotent maps [Schechter, 1997] (Definition 4.5.a). The closure operators are defined as the sequential application of the credal set function and the dual credal set function to subsets of Δ or 2^Ω. In symbols: $\mathcal{A} \mapsto m_\psi^\circ(m_\psi(\mathcal{A})), Q \mapsto m_\psi(m_\psi^\circ(Q))$. In particular, these closure operators define bipolar-closed sets.

8.5.3 Bipolar-Closed Sets

Bipolar-closed sets are sets $\mathcal{A} \subseteq 2^\Omega$ such that $\mathcal{A} = m_\psi^\circ(m_\psi(\mathcal{A}))$, respectively, $Q \subseteq \Delta$ such that $Q = m_\psi(m_\psi^\circ(Q))$. Most importantly, the bipolar-closed sets form two antitone isomorphic lattices ordered by set inclusion [Birkhoff, 1940] (Theorem V.8.20). This relationship gives us a lattice duality between set systems and credal sets of probabilities. See Figure 8.3 for an illustration of bipolar-closed sets and the Galois connection.

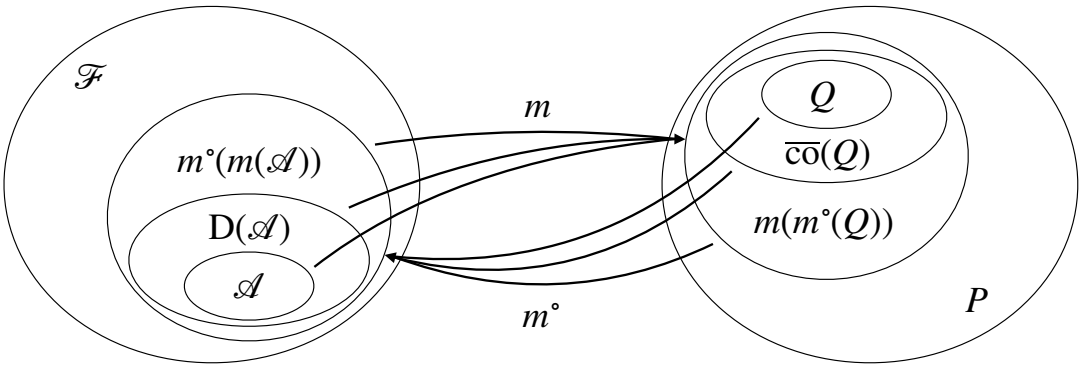


Figure 8.3: Galois connection between the lattice of pre-Dynkin systems and the set of credal sets. In the illustrated case, we have $m_\psi^\circ(Q) = m_\psi^\circ(m_\psi(m_\psi^\circ(\mathcal{A})))$, respectively, $m_\psi(\mathcal{A}) = m_\psi(m_\psi^\circ(m_\psi(Q)))$. The set containment on both sides follows from Proposition 8.25, Corollary 8.30 and Proposition 8.32.

Example 8.34. *The pre-Dynkin system* $\mathrm{D}(\{12, 3\})$ *already discussed in Example 8.26 and Example 8.33 is a bipolar-closed set. In contrast, the set* $\{12, 3\}$ *cannot be a bipolar-closed set, as it is not a pre-Dynkin system.*

More precisely, bipolar-closed sets in the set of finitely additive probability distributions are weak*-closed convex (Proposition 8.27). These map to bipolar-closed subsets of 2^Ω, which are pre-Dynkin systems (Proposition 8.31). All of the stated properties of bipolar-closed sets are necessary. But are they sufficient?

Sufficient Conditions for Bipolar-Closed Sets

In the search for sufficient conditions for bipolar-closed sets we focus on bipolar-closed subsets of 2^Ω. We leave bipolar-closed subsets of Δ to future investigations. Already, bipolar-closed subsets of 2^Ω are surprisingly difficult to characterize. We show that bipolar-closed sets are exactly the pre-Dynkin systems if the reference probability measure has

support on all elements of a finite set. To this end, let us start with a simple implication of Proposition 8.25 and Corollary 8.30.

Corollary 8.35. *Let m_ψ be the credal set function and m_ψ° be the dual credal set function. For an arbitrary subset $\mathcal{A} \subseteq 2^\Omega$ we have*

$$\mathrm{D}(\mathcal{A}) \subseteq m_\psi^\circ(m_\psi(\mathcal{A})).$$

Proof. By Proposition 8.25, $m_\psi^\circ(m_\psi(\mathcal{A})) = m_\psi^\circ(m_\psi(\mathrm{D}(\mathcal{A})))$. By Corollary 8.30, the statement follows. □

This corollary gives rise to the follow-up question: under which circumstances does $\mathrm{D}(\mathcal{A}) = m_\psi^\circ(m_\psi(\mathcal{A}))$? As the following theorem demonstrates, this question is closely connected to the sets of measure zero of the base probability ψ and its problems (cf. [Rota, 2001]).

Proposition 8.36 ("Closedness" under Measure Zero Sets)**.** *Let m_ψ be the credal set function and m_ψ° be the dual credal set function. Let $\mathcal{A} \subseteq 2^\Omega$. For any $A \in m_\psi^\circ(m_\psi(\mathcal{A}))$, if $\psi(A) = 0$, then all $B, C \in 2^\Omega$ such that $B \subseteq A$ and $C \supseteq A^c$ are in $m_\psi^\circ(m_\psi(\mathcal{A}))$.*

Proof. Let $A \in m_\psi^\circ(m_\psi(\mathcal{A}))$ for arbitrary $\mathcal{A} \subseteq 2^\Omega$ with $\psi(A) = 0$. Consider two sets $B, C \in 2^\Omega$ such that $B \subseteq A$, respectively, $C \supseteq A^c$. Then, $\psi(B) = 0$ and $\psi(C) = 1$. Furthermore, for any $\nu \in m_\psi(m_\psi^\circ(m_\psi(\mathcal{A})))$, we have $\nu(B) = 0$ (respectively, $\nu(C) = 1$). Since $m_\psi(m_\psi^\circ(m_\psi(\mathcal{A}))) = m_\psi(\mathcal{A})$, we obtain $B, C \in m_\psi^\circ(m_\psi(\mathcal{A}))$. □

A pre-Dynkin system $\mathrm{D}(\mathcal{A})$ can only coincide with $m_\psi^\circ(m_\psi(\mathcal{A}))$ if subsets of measure zero are included. Thus, Proposition 8.36 provides a further necessary condition for bipolar-closed subsets of 2^Ω. Yet, it turns out that the sets of measure zero as well can give a sufficient condition for bipolar-closed sets at least in a finite setting.

Theorem 8.37 (Pre-Dynkin Hull is Bipolar-Closure in Finite, Discrete Setting)**.** *Let $\Omega = [n]$. Fix a finitely additive probability ψ on 2^Ω such that $\psi(A) > 0$ for every $A \in 2^\Omega \setminus \{\emptyset\}$. Let $\mathcal{D} \subseteq 2^\Omega$ be a pre-Dynkin system. Then,*

$$\mathcal{D} = m_\psi^\circ(m_\psi(\mathcal{D})).$$

Proof. If $\mathcal{D} = 2^{[n]}$ the result follows directly. Thus, from now on we assume $\mathcal{D} \subsetneq 2^{[n]}$. The set inclusion $\mathcal{D} \subseteq m_\psi^\circ(m_\psi(\mathcal{D}))$ is given by Corollary 8.35. We show $\mathcal{D} \supseteq m_\psi^\circ(m_\psi(\mathcal{D}))$ by proving that for every $B \in 2^{[n]} \setminus \mathcal{D}$ there exists $\nu \in m_\psi(\mathcal{D})$ such that $\nu(B) \neq \psi(B)$.

Let us consider an arbitrary $B \in 2^{[n]} \setminus \mathcal{D}$. Without loss of generality (Lemma 8.63) we can decompose

$$B = B_\mathcal{D} \cup A,$$

where $B_{\mathcal{D}} \in \mathcal{D}$ and $A \notin \mathcal{D}$ is a weak atom with respect to $\mathcal{D}$ (Definition 8.62). Thus, we can leverage Lemma 8.64: there exists $\nu \in m_\psi(\mathcal{D})$ such that $\nu(A) \neq \psi(A)$. Thus,

$$\nu(B) = \nu(B_{\mathcal{D}}) + \nu(A) = \psi(B_{\mathcal{D}}) + \nu(A) \neq \psi(B_{\mathcal{D}}) + \psi(A).$$

It follows that $B \notin m_\psi^\circ(m_\psi(\mathcal{D}))$, concluding the proof. □

We emphasize that the statement does not hold as long as there are non-empty subsets of Ω which get assigned zero probability.

Example 8.38 (Bipolar-Closure is Strictly Greater Than Pre-Dynkin Hull)**.** *Let* $\Omega_3 = \{1, 2, 3\}$ *and* $\psi(\{1\}) = 1, \psi(\{2\}) = 0, \psi(\{3\}) = 0$ *constitute a base probability space. We have* $\mathrm{D}(\{1\}) = \{\emptyset, 1, 23, \Omega_3\}$*, but* $m_\psi^\circ(m_\psi(\{1\})) = 2^\Omega$*.*

Whether this theorem can be extended to more general sets Ω is an open question. Seemingly, proofs along the line of Theorem 8.37 are doomed to fail, since one cannot argue via probabilities on atoms of Ω.

8.5.4 Interpolation From Algebra to Trivial Pre-Dynkin System

In probability theory there is a choice to be made regarding which events should get assigned probabilities [Kolmogorov, 1927/1929] (p. 52). This significant choice has (mathematically) been standardized to form a (σ-)algebra (cf. standard probability space). But, already Kolmogorov, the "father" of modern probability theory, emphasized that this choice is not universal but should depend on the problem at hand. More recently, Khrennikov [2016] argued that a more appropriate probabilistic modeling should appeal to weaker domains for probabilities to, for instance, represent physical observations such as quantum phenomena.

In particular, it cannot always be taken for granted that all events are compatible with all others, as implied by a (σ-)algebra (cf. Section 8.2.1). For instance, von Mises' axiomatization of probability inherently reflects potential incompatible events in terms of a pre-Dynkin system [Von Mises and Geiringer, 1964, Schurz and Leitgeb, 2008]. In other words, there is a choice to be made about the system of precision. Which sets should be compatible to each other, and which should not? How do the choices of the systems of precision relate to each other?

We neglect, without loss of generality, arbitrary systems of precision and focus on pre-Dynkin systems (cf. Proposition 8.25). The range of choices is captured by the system of pre-Dynkin systems.

Proposition 8.39 (Set of Pre-Dynkin Systems is a Lattice)**.** *The set* $\mathfrak{D} \coloneqq \{\mathcal{D} \subseteq 2^\Omega \colon \mathcal{D} =$

$\mathrm{D}(\mathcal{D})\}$ *is ordered by set inclusion. Furthermore,* $(\mathfrak{D}, \subseteq)$ *is a lattice with*

$$\bigvee_{i=1}^{n} \mathcal{D}_i = \mathrm{D}\left(\bigcup_{i=1}^{n} \mathcal{D}_i\right)$$
$$\bigwedge_{i=1}^{n} \mathcal{D}_i = \bigcap_{i=1}^{n} \mathcal{D}_i.$$

Proof. On the one hand, it is easy to show that the intersection of pre-Dynkin systems forms a pre-Dynkin system again. On other hand, the smallest pre-Dynkin system which contains a finite set of pre-Dynkin systems is by definition the pre-Dynkin system generated by the union over all elements in this finite set of pre-Dynkin systems. □

Example 8.40. *Let* Ω_4 *be as defined in Example 8.24. The minimal element in* $\mathfrak{D}$ *then is* $\{\emptyset, \Omega_4\}$*. The maximal element is* 2^{Ω_4}*. For the sake of brevity, we omit all further elements in* $\mathfrak{D}$ *and remain with the observation that* $\mathcal{D}_4$ *of Example 8.3 and* $\mathrm{D}(\{12, 3\})$ *are elements of* $\mathfrak{D}$*.*

The lattice $\mathfrak{D}$ spans a range of choices from $\mathcal{D} = 2^{\Omega}$, i.e., complete compatibility and only a single probability distribution in its credal set, namely $m_\psi(2^{\Omega}) = \{\psi\}$, to $\mathcal{D} = \{\emptyset, \Omega\}$, i.e., no compatibility and the entire space of probability distributions constitute its credal set $m_\psi(\{\emptyset, \Omega\}) = \Delta$. How "close" $\mathcal{D}$ is to the algebra 2^{Ω} determines how "classical" the credal set behaves. In other words, $(\mathfrak{D}, \subseteq)$ parametrizes a family of credal sets. Thus, it parametrizes coherent probabilities. The knob of compatibility can be turned from trivially nothing ($\{\emptyset, \Omega\}$), to everything (2^{Ω}). How does the "amount of compatibility" of the pre-Dynkin system map to the credal sets? Or, e.g., given two pre-Dynkin systems on which a probability is defined, what is the credal set of the union of these systems?

Proposition 8.41 (Lattice of Dynkin Systems and Credal Sets)**.** *The credal set function* m_ψ *(Definition 8.23) together with the lattice* $(\mathfrak{D}, \subseteq)$ *provides a parametrized family of credal sets for which hold* $(\forall \mathcal{D}_1, \mathcal{D}_2 \in \mathfrak{D})$*:*

$$m_\psi(\mathcal{D}_1 \vee \mathcal{D}_2) = m_\psi(\mathcal{D}_1 \cup \mathcal{D}_2) = m_\psi(\mathcal{D}_1) \cap m_\psi(\mathcal{D}_2)$$
$$m_\psi(\mathcal{D}_1 \wedge \mathcal{D}_2) = m_\psi(\mathcal{D}_1 \cap \mathcal{D}_2) \supseteq m_\psi(\mathcal{D}_1) \cup m_\psi(\mathcal{D}_2).$$

Proof. Concerning the first equality, we observe that for arbitrary $\mathcal{D}_1, \mathcal{D}_2 \in \mathfrak{D}$

$$m_\psi(\mathcal{D}_1 \vee \mathcal{D}_2) = m_\psi(\mathrm{D}(\mathcal{D}_1 \vee \mathcal{D}_2)) = m_\psi(\mathcal{D}_1 \cup \mathcal{D}_2),$$

by Proposition 8.25. Consequently,

$$\begin{aligned} m_\psi(\mathcal{D}_1 \cup \mathcal{D}_2) &= \{\nu \in \Delta \colon \nu(D) = \psi(D),\ \forall D \in \mathcal{D}_1 \cup \mathcal{D}_2\} \\ &= \{\nu \in \Delta \colon \nu(D) = \psi(D),\ \forall D \in \mathcal{D}_1\} \cap \{\nu \in \Delta \colon \nu(D) = \psi(D),\ \forall D \in \mathcal{D}_2\} \\ &= m_\psi(\mathcal{D}_1) \cap m_\psi(\mathcal{D}_2). \end{aligned}$$

The second line follows by the definition of infimum on the lattice of pre-Dynkin systems and simple set containment: $m_\psi(\mathcal{D}_1) \subseteq m_\psi(\mathcal{D}_1 \cap \mathcal{D}_2)$ and $m_\psi(\mathcal{D}_2) \subseteq m_\psi(\mathcal{D}_1 \cap \mathcal{D}_2)$. □

Unfortunately, the mentioned interpolation is slightly improper. It turns out that there are pre-Dynkin systems $\mathcal{D}_1 \neq \mathcal{D}_2$ such that $m_\psi(\mathcal{D}_1) = m_\psi(\mathcal{D}_2)$.

Example 8.42 (Non-Injectivity of Credal Set Function)**.** *Let* $\Omega_3 = \{1, 2, 3\}$ *and* $\psi(\{1\}) = 1, \psi(\{2\}) = 0, \psi(\{3\}) = 0$ *constitute a base probability space, cf. Example 8.38. Then, obviously* $m_\psi(\mathrm{D}(\{1\})) = m_\psi(2^\Omega)$*, but* $\mathrm{D}(\{1\}) \neq 2^\Omega$*.*

The reason for this collision of credal sets is that not every pre-Dynkin system $\mathcal{D} \subseteq 2^\Omega$ is a bipolar-closed set.

Proposition 8.43 (Credal Set Function is Injective on Bipolar-Closed Sets)**.** *Let* m_ψ *be the credal set and* m_ψ° *be the dual credal set function. Let* $\mathcal{D}_1, \mathcal{D}_2 \in \mathfrak{D}$ *be pre-Dynkin systems, which are bipolar-closed. If* $\mathcal{D}_1 \neq \mathcal{D}_2$*, then* $m_\psi(\mathcal{D}_1) \neq m_\psi(\mathcal{D}_2)$*.*

Proof. We prove the claim by contraposition:

$$m_\psi(\mathcal{D}_1) = m_\psi(\mathcal{D}_2) \Rightarrow m_\psi^\circ(m_\psi(\mathcal{D}_1)) = m_\psi^\circ(m_\psi(\mathcal{D}_2)) \Rightarrow \mathcal{D}_1 = \mathcal{D}_2.$$

□

Hence, it is reasonable to focus on the set of bipolar-closed sets contained in 2^Ω. We define the *set of interpolating pre-Dynkin systems* $\mathfrak{C} \coloneqq \{\mathcal{C} \subseteq 2^\Omega \colon \mathcal{C} = m_\psi^\circ(m_\psi(\mathcal{C}))\}$. We know that $\mathfrak{C} \subseteq \mathfrak{D}$ (Proposition 8.31) and $\mathfrak{C}$ is even a lattice contained in $\mathfrak{D}$.

Proposition 8.44 (Lattice of Bipolar-Closed Sets and Credal Sets)**.** *Let* m_ψ *be the credal set function and* m_ψ° *be the dual credal set function. The set of interpolating pre-Dynkin systems* $\mathfrak{C}$ *equipped with the* $\subseteq$*-ordering forms a lattice:*

$$\begin{aligned} \bigvee_{i=1}^n \mathcal{C}_i &= m_\psi^\circ\left(m_\psi\left(\bigcup_{i=1}^n \mathcal{C}_i\right)\right) \\ \bigwedge_{i=1}^n \mathcal{C}_i &= \bigcap_{i=1}^n \mathcal{C}_i. \end{aligned}$$

In particular, this lattice $\mathfrak{C}$ *is antitone isomorphic to the lattice of bipolar-closed sets in* 2^{Δ}*. It holds (We denote the composition of functions with* $\circ$*.):*

$$m_\psi(\mathcal{C}_1 \vee \mathcal{C}_2) = m_\psi(\mathcal{C}_1 \cup \mathcal{C}_2) = m_\psi(\mathcal{C}_1) \cap m_\psi(\mathcal{C}_2)$$
$$m_\psi(\mathcal{C}_1 \wedge \mathcal{C}_2) = m_\psi(\mathcal{C}_1 \cap \mathcal{C}_2) = (m_\psi \circ m_\psi^\circ)\,(m_\psi(\mathcal{C}_1) \cup m_\psi(\mathcal{C}_2))\,.$$

Proof. By Theorem V.8.20 in [Birkhoff, 1940], $\mathfrak{C}$ is a lattice and m_ψ an antitone lattice isomorphism on $\mathfrak{C}$. The equations hold by simple manipulations

$$m_\psi(\mathcal{C}_1 \vee \mathcal{C}_2) = m_\psi(m_\psi^\circ(m_\psi(\mathcal{C}_1 \cup \mathcal{C}_2))) = m_\psi(\mathcal{C}_1 \cup \mathcal{C}_2) = m_\psi(\mathcal{C}_1) \cap m_\psi(\mathcal{C}_2),$$

and

$$m_\psi(\mathcal{C}_1 \wedge \mathcal{C}_2) = m_\psi(m_\psi^\circ(m_\psi(\mathcal{C}_1 \cap \mathcal{C}_2))) = (m_\psi \circ m_\psi^\circ)\,(m_\psi(\mathcal{C}_1) \cup m_\psi(\mathcal{C}_2))\,.$$

□

Note that $\mathfrak{C}$ is not generally a sublattice of $\mathfrak{D}$. It is a lattice contained in the lattice $\mathfrak{D}$, but the closure operator for the supremum is distinct. Interestingly, the order structure which both lattices, $\mathfrak{D}$ and $\mathfrak{C}$, induce on the set of credal sets via m_ψ is identical. Every set $m_\psi(\mathcal{A})$ for arbitrary $\mathcal{A} \subseteq 2^\Omega$ is bipolar-closed (Corollary 8.30). Hence, the lattice of bipolar-closed sets in 2^Δ is the domain of m_ψ for elements in $\mathfrak{D}$ and $\mathfrak{C}$. In other words, the lattices $\mathfrak{D}$ and $\mathfrak{C}$ provide one and the same parametrized family of credal sets, thus one and the same parametrized family of imprecise probabilities.

In comparison to other parametrized families of imprecise probability, such as distortion risk measures [Wirch and Hardy, 2001], which heavily rely on convex analysis, the duality used here is structurally weaker. Lattice isomorphisms give a glimpse of structure to the involved dual spaces. Convex dualities as exploited in [Fröhlich and Williamson, 2024] are far more informative, but apparently not able to handle the structural knob which we presented in this work: the set of sets which get assigned precise probabilities. Nevertheless, a natural question arises from this lattice duality: how does this lattice duality relate to a convex duality? We leave this question open to further research. A first attempt to an answer is discussed in Appendix 8.8.3, where we link the parametrized family of distorted probabilities to the pre-Dynkin system family of imprecise probabilities.

8.6 A More General Perspective—The Set of Gambles With Precise Expectation

To this point, we have exclusively focused on probabilities and set systems of events to which we assign probabilities. In fact, there is a more general story to be told. In the literature on imprecise probability, focus often lies on expectation-type functionals instead of probabilities and on sets of gambles (bounded functions from the base set Ω to the real numbers) instead of sets of events. One can easily see that the latter is more general and can recover the former. Indicator functions of events are gambles. An expectation-type functional evaluated on an indicator gamble of an event corresponds to a generalized probability of the event. The converse direction, i.e., recovering a unique expectation-type functional from an imprecise probability, however, is not always possible [Walley, 1991] (§2.7.3). In the following, we reiterate several questions which we asked in the preceding sections for probabilities and set systems.

8.6.1 Partial Expectations Generalize Finitely Additive Probabilities on (Pre-)Dynkin Systems

We propose the following definition of partial expectation and show afterwards that it is a natural generalization of finitely additive probabilities defined on (pre-)Dynkin systems.

Definition 8.45 (Partial Expectation)**.** *Let $\{L_i\}_{i\in I}$ be a non-empty family of linear subspaces of* $\mathrm{B}(\Omega)$*. We call $E\colon \bigcup_{i\in I} L_i \to \mathbb{R}$ a* partial expectation *if and only if all of the following conditions are fulfilled:*

(a) for any $i \in I$ and for all $f, g \in L_i$, then $E(f+g) = E(f)+E(g)$, (Partial Linearity),

(b) for any $i \in I$ and any $f \in L_i$, then $E(f) \geq \inf f$, (Coherence).

We remark that for this definition, we leveraged the requirements for a linear prevision (Definition 8.49) on a linear space given in [Walley, 1991] (Theorem 2.8.4). In other words, a partial expectation is a functional which is defined on a union of linear subspaces and behaves like a "classical" (finitely additive) expectation on each of the subspaces but not necessarily on all simultaneously. It is a linear prevision when restricted to one of the subspaces L_i (cf. Definition 8.49).

There is a one-to-one correspondence of linear previsions and coherent additive probabilities. For every coherent additive probability ν defined on an algebra $\mathcal{A}$, there is a unique linear prevision, which we equivalently denote ν, the set of all $\mathcal{A}$-measurable gambles, which agrees with the probability ν on the indicator gambles of the sets in $\mathcal{A}$ [Walley, 1991] (Theorem 3.2.2). The following Proposition exploits this correspondence. A finitely additive

probability defined on a pre-Dynkin system relates one-to-one to a partial expectation which is defined on the set of linear spaces induced by the simple gambles on the blocks of the pre-Dynkin system. To this end, we introduce the following two notations: let $\mathcal{F} \subseteq 2^\Omega$ be an algebra. Then, $\mathrm{S}(\Omega, \mathcal{F}) \subseteq \mathrm{B}(\Omega)$ denotes the linear subspace of simple gambles on $\mathcal{F}$, i.e., scaled and added indicator gambles of a finite number of disjoint sets (cf. [Rao and Rao, 1983] (Definition 4.2.12)). Let $\mathcal{F}_\sigma \subseteq 2^\Omega$ be a σ-algebra. Then $\mathrm{B}(\Omega, \mathcal{F}_\sigma) \subseteq \mathrm{B}(\Omega)$ denotes the linear subspace of all bounded, real-valued, $\mathcal{F}_\sigma$-measurable gambles.

Proposition 8.46 (Finitely Additive Probability on Pre-Dynkin System and its Partial Expectation)**.** *Let $\mathcal{D} \subseteq 2^\Omega$ be a pre-Dynkin system. Let $\mu\colon \mathcal{D} \to [0,1]$ be a finitely additive probability defined on the pre-Dynkin system with block structure $\{\mathcal{A}_i\}_{i\in I}$. Then, μ is in one-to-one correspondence to a partial expectation $E\colon \bigcup_{i\in I} \mathrm{S}(\Omega, \mathcal{A}_i) \to \mathbb{R}$ defined on the union of linear spaces of simple gambles induced by all blocks $\mathcal{A}_i$ of $\mathcal{D}$.*

Proof. By Theorem 8.7, we can decompose the pre-Dynkin system $\mathcal{D}$ into a set of blocks $\{\mathcal{A}_i\}_{i\in I}$. Since, $\mathcal{A}_i \subseteq 2^\Omega$ for all $i \in I$, each of the blocks induces a the linear subspace of simple gambles $\mathrm{S}(\Omega, \mathcal{A}_i) \subseteq \mathrm{B}(\Omega)$. Given a finitely additive measure $\mu\colon \mathcal{D} \to [0,1]$, we now define $E\colon \bigcup_{i\in I} \mathrm{S}(\Omega, \mathcal{A}_i) \to \mathbb{R}$ by

$$E(f) \coloneqq \int f d\mu|_{\mathcal{A}_i} \text{ if } f \in \mathrm{S}(\Omega, \mathcal{A}_i).$$

For every $i \in I$, $E|_{\mathrm{S}(\Omega,\mathcal{A}_i)}$ is a linear prevision in one-to-one correspondence to the finitely additive probability $\mu|_{\mathcal{A}_i}$ [Walley, 1991] (Theorem 3.2.2). Hence, conditions (a) and (b) in Definition 8.45 are met [Walley, 1991] (Theorem 2.8.4). For $f \in \mathrm{S}(\Omega, \mathcal{A}_i) \cap \mathrm{S}(\Omega, \mathcal{A}_j)$ $(i \neq j, i, j \in I)$, we know that $f \in \mathrm{S}(\Omega, \mathcal{A}_i \cap \mathcal{A}_j)$ (Lemma 8.65), hence,

$$\int f d\mu|_{\mathcal{A}_i} = \int f d\mu|_{\mathcal{A}_i \cap \mathcal{A}_j} = \int f d\mu|_{\mathcal{A}_j}.$$

Thus, E is well-defined and there is no other partial expectation which agrees with μ on the indicator gambles of the sets in $\mathcal{D}$. □

The attentive reader might have noticed that we defined the partial expectation in Proposition 8.46 on very specific linear subspaces of $\mathrm{B}(\Omega)$, namely the linear subspaces of simple gambles. In fact, the statement would still hold when enlarging the linear subspaces of simple gambles $\mathrm{S}(\Omega, \mathcal{A}_i)$ for every $i \in I$ to linear subspaces of functions which are "convergence in measure"-approximated by gambles in $\mathrm{S}(\Omega, \mathcal{A}_i)$. For more details, we refer the reader to [Rao and Rao, 1983] (Definition 4.4.5 and Corollary 4.4.9).

But, it is not the case that we can extend the definition to all sets of bounded, $\mathcal{A}_i$-measurable functions, i.e., bounded functions whose pre-images of sets in the smallest algebra which con-

tains all open sets of the real numbers are contained in $\mathcal{A}_i$. For an algebra $\mathcal{A} \subseteq 2^\Omega$, the set of $\mathcal{A}$-measurable gambles is not necessarily a linear subspace of $\mathrm{B}(\Omega)$ [Walley, 1991] (p. 129).

This is different for a σ-algebra $\mathcal{A}_\sigma \subseteq 2^\Omega$. The set of bounded, $\mathcal{A}_\sigma$-measurable functions $\mathrm{B}(\Omega, \mathcal{A}_\sigma)$ forms a linear subspace of $\mathrm{B}(\Omega)$ [Walley, 1991] (p. 129). Here, measurability is defined as the pre-image of every Borel-measurable set in $\mathbb{R}$ is in $\mathcal{A}$. In this case, the set of linear spaces on which the partial expectation is defined is given by all bounded, measurable functions on the σ-blocks.

Proposition 8.47 (Finitely Additive Probability on Dynkin Systems and its Partial Expectation)**.** *Let $\mathcal{D}_\sigma \subseteq 2^\Omega$ be a Dynkin system on the base set Ω. Let $\mu\colon \mathcal{D}_\sigma \to [0,1]$ be a finitely additive probability defined on the Dynkin system. Then μ is in one-to-one correspondence to a partial expectation $E\colon \bigcup_{i\in I} \mathrm{B}(\Omega, \mathcal{A}_i) \to \mathbb{R}$ defined on the union of linear spaces of measurable gambles induced by all σ-blocks $\mathcal{A}_i$ of $\mathcal{D}_\sigma$.*

Proof. By Theorem 8.71, we can decompose the Dynkin system $\mathcal{D}_\sigma$ into a set of σ-blocks $\{\mathcal{A}_i\}_{i\in I}$. Since, $\mathcal{A}_i \subseteq 2^\Omega$ for all $i \in I$, each of the σ-blocks induce a linear subspace of $\mathrm{B}(\Omega)$, which we denote as $\mathrm{B}(\Omega, \mathcal{A}_i)$. Given a finitely additive measure $\mu\colon \mathcal{D}_\sigma \to [0,1]$, we now define $E\colon \bigcup_{i\in I} \mathrm{B}(\Omega, \mathcal{A}_i) \to \mathbb{R}$ by

$$E(f) \coloneqq \int f d\mu|_{\mathcal{A}_i} \text{ if } f \in \mathrm{B}(\Omega, \mathcal{A}_i).$$

For every $i \in I$, $E|_{\mathrm{B}(\Omega,\mathcal{A}_i)}$ is a linear prevision in one-to-one correspondence to the finitely additive probability $\mu|_{\mathcal{A}_i}$ [Walley, 1991] (Theorem 3.2.2). Hence, conditions (a) and (b) in Definition 8.45 are met [Walley, 1991] (Theorem 2.8.4).

For $f \in \mathrm{B}(\Omega, \mathcal{A}_i) \cap \mathrm{B}(\Omega, \mathcal{A}_j)$ $(i \neq j, i, j \in I)$, we know that $f \in \mathrm{B}(\Omega, \mathcal{A}_i \cap \mathcal{A}_j)$ (Lemma 8.66); hence,

$$\int f d\mu|_{\mathcal{A}_i} = \int f d\mu|_{\mathcal{A}_i \cap \mathcal{A}_j} = \int f d\mu|_{\mathcal{A}_j}.$$

Thus, E is well-defined and there is no other partial expectation which agrees with μ on the indicator gambles of the sets in $\mathcal{D}$. □

It remains to emphasize that there are partial expectations defined on families of linear subspaces which are not induced by finitely additive probabilities on Dynkin systems. A simple example is given by a linear space which does not contain the constant gamble corresponding to the indicator gamble of the set Ω. Hence, the definition of a partial expectation is indeed a generalization of the definition of a finitely additive probability on a pre-Dynkin system.

Under the name "partially specified probabilities", Lehrer [2012] introduced a closely related

notion to our partial expectation. Lehrer, however, assumed that there is by definition an underlying probability distribution over the entire base set (or better said, a σ-algebra on the base set). Hence, his partially specified probabilities are by definition extendable (see Definition 8.50), a fact, which he implicitly exploited by re-defining the natural extension following [Walley, 1991] (Lemma 3.1.3 (e)) of partially specified probabilities [Lehrer, 2007] (§3.2). Lehrer did not ask for the structure of the set of gambles with precise expectations, nor did he draw any connection to Walley's work, nor did he link his "partially specified probabilities" to finitely additive probabilities on pre-Dynkin systems.

8.6.2 System of Precision—The Space of Gambles With Precise Expectations

In Section 8.3, we have shown that imprecise probabilities are precise on (pre-)Dynkin systems. The natural analogue of this *set structure of precision* is the *space of gambles with precise expectation*, which actually forms a linear subspace. The space of gambles with precise expectation is known as a set of ambiguity-free or unambiguous gambles in the finance and decision theory literature [De Castro and Chateauneuf, 2011].

Theorem 8.48. *(Imprecise Expectations Are Precise on a Linear Subspace of Precise Gambles) Let* $\mathrm{B}(\Omega)$ *be the linear space of bounded, real-valued functions on* Ω*. Let* $L\colon \mathrm{B}(\Omega) \to \mathbb{R}$ *and* $U\colon \mathrm{B}(\Omega) \to \mathbb{R}$ *be two functionals, for which all the following properties hold:*

(a) Normalization: $L(\chi_\Omega) = U(\chi_\Omega) = 1$.

(b) Conjugacy: $U(f) = -L(-f)$ *for* $f \in \mathrm{B}(\Omega)$.

(c) Subadditivity of U*: for* $f, g \in \mathrm{B}(\Omega)$*, we have* $U(f+g) \leq U(f) + U(g)$.

(d) Superadditivity of L*: for* $f, g \in \mathrm{B}(\Omega)$*, we have* $L(f+g) \geq L(f) + L(g)$.

(e) Positive Homogeneity: for $\alpha \in [0, \infty)$ *and* $f \in \mathrm{B}(\Omega)$*, we have* $L(\alpha f) = \alpha L(f)$ *and* $U(\alpha f) = \alpha U(f)$.

Then L *and* U *coincide on the linear space* $\mathcal{S} \coloneqq \{f \in \mathrm{B}(\Omega)\colon L(f) = U(f)\} \subseteq \mathrm{B}(\Omega)$*, the* space of gambles with precise expectation*, which contains all constant gambles.*

Proof. We define

$$\mathcal{S} \coloneqq \{f \in \mathrm{B}(\Omega)\colon L(f) = U(f)\}, \tag{8.3}$$

and show that $\mathcal{S}$ forms a linear subspace of $\mathrm{B}(\Omega)$. First, let $f, g \in \mathcal{S}$, then

$$L(f) + L(g) \overset{(d)}{\leq} L(f+g) \overset{(\star)}{\leq} U(f+g) \overset{(c)}{\leq} U(f) + U(g) \overset{\text{Eq. 8.3}}{=} L(f) + L(g).$$

For $(\star)$ observe that $L(f) \leq U(f)$ for all $f \in \mathrm{B}(\Omega)$, since

$$L(f) + L(-f) \leq L(0) \overset{(a),(e)}{=} 0 \overset{(a),(e)}{=} U(0) \leq U(f) + U(-f),$$

we have

$$L(f) + L(-f) \leq U(f) + U(-f) \Leftrightarrow L(f) - U(f) \leq U(f) - L(f) \Leftrightarrow L(f) \leq U(f).$$

Second, let $f \in \mathcal{S}$ and $\alpha \in \mathbb{R}$. If $\alpha \geq 0$, then

$$L(\alpha f) \overset{(e)}{=} \alpha L(f) \overset{Eq.8.3}{=} \alpha U(f) \overset{(e)}{=} U(\alpha f).$$

Otherwise,

$$L(\alpha f) \overset{(b)}{=} -U(-\alpha f) \overset{(e)}{=} \alpha U(f) \overset{Eq.8.3}{=} \alpha L(f) \overset{(b)}{=} -\alpha U(-f) \overset{(e)}{=} U(\alpha f).$$

Third, $\mathcal{S}$ contains all constant gambles by (a), (b) and (e). Hence, we have shown that $\mathcal{S}$ forms a linear subspace of $\mathrm{B}(\Omega)$ which contains all constant gambles. □

The choice of properties for the lower and upper expectation functional is not arbitrary. We tried to resemble the properties involved in the analogous statement for lower and upper probabilities (Theorem 8.10). One can easily check that a lower and upper expectation with the given properties (a)–(d) forms a lower and upper probability as required in Theorem 8.10 if the expectation is restricted to indicator gambles. However, we added property (e), positive homogeneity.

Without the property of positive homogeneity, the resulting set of gambles with precise expectations would not form a proper linear subspace, as then one can only guarantee closedness of $\mathcal{S}$ under rational multiplication. The condition of positive homogeneity "fills up" the gaps with all real-scaled functions. Instead of positive homogeneity one can as well demand a continuity assumption of the lower and upper functional L and U, e.g., [Walley, 1991] (Property (l) Theorem 2.6.1).

It is known that the system of precision of coherent lower and upper probability is a linear space [Seidenfeld, 2023]. The derivation here slightly generalizes this to not necessarily coherent lower and upper functionals L and U. We emphasize that coherent previsions (see Definition 8.49) fulfill all of the demanded properties [Walley, 1991] (Theorem 2.6.1).

In Theorem 8.10, we did not only derive the structure of the system of precision for events, we as well showed that the restriction of the lower and upper probability to this system give us a probability defined on a (pre-)Dynkin system. In the statement here, we can define a similar restriction $P\colon \mathcal{S} \to \mathbb{R}$ with $P \coloneqq L|_{\mathcal{S}} = U|_{\mathcal{S}}$. However, P is not necessarily coher-

ent *and* not necessarily a partial expectation. In other words, we cannot directly recover a partial expectation on subdomain of $\mathrm{B}(\Omega)$ induced by a lower and upper expectation functional. For probabilities, this recovery was possible.

But, we can sidestep this restriction. Any pair of lower and upper expectations, as we defined them here, induce a unique lower and upper probability. The resulting finitely additive probability on the set structure of precision gives rise to a partial expectation (Proposition 8.46) on a set of linear subspaces contained in the space of gambles with precise expectation $\mathcal{S}$ of L and U.

A converse construction, however, is not possible. There is no unique lower and upper expectation functional with the given properties associated with a lower and upper probability fulfilling the axioms of Theorem 8.10 [Walley, 1991] (§2.7.3). Concluding, lower and upper expectation as defined here are not the "perfect" analogues of lower and upper probabilities.

This as well explains the mismatch between systems of precision for probabilities and expectations. A lower and upper expectation fulfills the properties of a lower and upper probability but not vice versa. Hence, only weaker statements about the system of precision are possible for probabilities. As a result, the analogue of the set structure of precision, a pre-Dynkin system, is the space of gambles with precise expectations, a *single* linear subspace. In Definition 8.45, however, we equated pre-Dynkin systems with *sets* of linear subspaces. In this case, a one-to-one correspondence between a finitely additive probability on a pre-Dynkin system and a generalized expectation, concretely a partial expectation, can be established. Hence, the analogy of pre-Dynkin systems and linear subspaces of gambles depends on the correspondence of probability and expectation.

8.6.3 Generalized Extendability Is Equivalent to Coherence

Partial expectations are, as we have shown, a natural generalization of finitely additive probabilities on pre-Dynkin systems. Hence, it is not far-fetched to ask for definitions of coherence and extendability again, now in the more general context. It turns out that the same story can be re-told on a more general scale: The definition of coherent probabilities (Definition 8.18) is in fact just the reduction of the following definition of a coherent prevision to indicator gambles.

Definition 8.49 (Coherent Prevision [Walley, 1991] (Definition 2.5.1))**.** *Let $L \subseteq \mathrm{B}(\Omega)$ be an arbitrary subset of the linear space of bounded functions. A functional $\underline{\mathbb{E}}\colon L \to \mathbb{R}$ is a* coherent lower prevision *if and only if*

$$\sup_{\omega\in\Omega} \sum_{i=1}^{j} (f_i(\omega) - \underline{\mathbb{E}}(f_i)) - m(f_0(\omega) - \underline{\mathbb{E}}(f_0)) \geq 0,$$

for non-negative $n, m \in \mathbb{N}$ *and* $f_0, f_1, \dots f_n \in L$*. If* $L = -L$*, the conjugate* coherent upper prevision *is given by* $\overline{\mathbb{E}}(f) \coloneqq -\underline{\mathbb{E}}(-f)$ *for all* $f \in L$*. If, furthermore,* $\underline{\mathbb{E}}(f) = \overline{\mathbb{E}}(f)$ *for all* $f \in L$*, we call* $\nu \coloneqq \underline{\mathbb{E}} = \overline{\mathbb{E}}$ *a* linear prevision.

This definition of coherent previsions is substantiated by consistency of gamblers regarding their betting behavior on gambles with uncertain outcome, e.g., [Walley, 1991] (§2.3.1). Importantly, the rather opaque but general definition of coherence can be simplified greatly for coherent previsions defined on linear subspaces of $\mathrm{B}(\Omega)$. Theorem 2.5.5 in [Walley, 1991] shows that coherence for lower previsions on linear subspaces can be expressed as super-additivity, positive homogeneity, and accepting sure gains (see [Walley, 1991] (Definition 2.3.3)).

Having introduced the notion of a coherent prevision, we now envisage the link between partial expectations and coherent previsions. We introduced extendability for finitely additive probabilities on pre-Dynkin systems as a useful property. It guarantees that the probability can "nicely" be embedded into "larger" finitely additive probability which is defined on an encompassing algebra. Hence, the analogue for partial expectations is straightforward.

Definition 8.50 (General Extendability)**.** *A partial expectation* $E\colon \bigcup_{i\in I} L_i \to \mathbb{R}$ *is* extendable *if and only if there exists a partial expectation* $E'\colon \mathrm{B}(\Omega) \to \mathbb{R}$ *such that* $E'|_{\bigcup_{i\in I} L_i} = E$*.*

Interestingly, the extendability condition provided in Theorem 8.16 has a (more general) cousin adapted to the setting of gambles instead of events.

Proposition 8.51 (Extendability Condition for Previsions)**.** *(cf. [Maharam, 1972] (Theorem 6.1).) Let* $\{L_i\}_{i\in I}$ *be a non-empty family of linear subspaces of* $\mathrm{B}(\Omega)$*. A* partial expectation $E\colon \bigcup_{i\in I} L_i \to \mathbb{R}$ *is extendable if and only if for every finite collection of functions* $f_1, \dots, f_n \in \{L_i\}_{i\in I}$*,*

$$\sum_{i=1}^{n} f_i \geq 0 \implies \sum_{i=1}^{n} E(f_i) \geq 0.$$

Proof. It seems that Theorem 6.1 [Maharam, 1972] is equivalent to our statement. However, there is a subtlety which we want to argue here is indeed irrelevant. Extendability of a partial expectation requires the existence of a positive, *normed*, linear functional on $\mathrm{B}(\Omega)$, whose restriction on the according linear subspaces coincides with the partial expectation. Theorem 6.1 in [Maharam, 1972] only guarantees that a positive, linear functionals exists. But, normedness of such functional is automatically given if $\chi_\Omega \in L_i$ for some $i \in I$. Otherwise, we extend the partial expectation E to $E'\colon \{\alpha\chi_\Omega \colon \alpha \in \mathbb{R}\} \bigcup_{i\in I} L_i \to \mathbb{R}$ such

that

$$E'(f) \coloneqq \begin{cases} \alpha \text{ if } f \in \{\alpha\chi_\Omega \colon \alpha \in \mathbb{R}\} \\ E(f) \text{ otherwise.} \end{cases}$$

Then again, Theorem 6.1 [Maharam, 1972] applies. □

Against the background that extendability and coherence define the same concept for finitely additive probabilities on pre-Dynkin systems, the resulting equivalence of extendability and coherence for partial expectations is of little surprise.

Proposition 8.52 (Extendability is Equivalent to Coherence)**.** *Let $\{L_i\}_{i\in I}$ be a non-empty family of linear subspaces of $\mathrm{B}(\Omega)$. The partial expectation $E\colon \bigcup_{i\in I} L_i \to \mathbb{R}$ is extendable if and only if E is a linear prevision, i.e., is coherent.*

Proof. If E is a linear prevision on $\bigcup_{i\in I} L_i$, then there is a linear prevision $E'\colon \mathrm{B}(\Omega) \to \mathbb{R}$ such that $E'|_{\bigcup_{i\in I} L_i} = E$ [Walley, 1991] (Theorem 3.4.2). Conversely, if E is an extendable partial expectation, then its extension is obviously a linear prevision, and hence it is coherent. The restriction of a coherent linear prevision to any subset of gambles is coherent (and linear). □

8.6.4 A Duality Theory for Previsions and Families of Linear Subspaces

In Section 8.5, we step by step spelled out an order relationship between the set structure of precision and credal sets, a model for (coherent) imprecise probabilities. Naturally the presented generalization begs the question whether a related relationship between credal sets and the spaces of gambles with precise expectations exists. We answer affirmatively. We redefine the credal and dual credal set function and shortly discuss its analogous properties. Again, we require a "reference measure". In this case, it is a fixed linear prevision ψ on the space of all gambles $\mathrm{B}(\Omega)$, which is indeed in one-to-one correspondence to a finitely additive probability measure on 2^Ω.

Definition 8.53 (Generalized Credal Set Function)**.** *Let Δ be the set of linear previsions on the Banach space $\mathrm{B}(\Omega)$. For a fixed linear prevision $\psi \in \Delta$, we call*

$$m_\psi \colon 2^{\mathrm{B}(\Omega)} \to 2^\Delta, \quad m_\psi(\mathcal{G}) \coloneqq \{\nu \in \Delta \colon \nu(g) = \psi(g), \forall g \in \mathcal{G}\},$$

the generalized credal set function.

Definition 8.54 (Generalized Dual Credal Set Function)**.** *Let Δ be the set of linear previsions on the Banach space $\mathrm{B}(\Omega)$. For a fixed linear prevision $\psi \in \Delta$, we call*

$$m_\psi^\circ \colon 2^\Delta \to 2^{\mathrm{B}(\Omega)}, \quad m_\psi^\circ(\mathcal{Q}) \coloneqq \{g \in \mathrm{B}(\Omega) \colon \nu(g) = \psi(g), \forall \nu \in \mathcal{Q}\},$$

the generalized dual credal set function*.*

Why can we call those functions "generalized"? Simply because any system of sets is equivalently represented as its set of indicator gambles which span their own linear space of simple gambles, i.e., linear combinations of indicator gambles.

The generalized credal set function maps, as the credal set function in Definition 8.23, to weak*-closed, convex subsets of Δ. The generalized dual credal set function, however, reveals a first subtlety. It maps to linear subspaces of $\mathrm{B}(\Omega)$. The dual credal set function following Definition 8.28 mapped to pre-Dynkin systems. In Proposition 8.46 and Proposition 8.47, families of linear subspaces were the analogues of (pre-)Dynkin systems. Here, a single linear subspace is the analogue of a pre-Dynkin system. For a first step towards an explanation of this asymmetry, see Section 8.6.2. Finally, the pair of functions constitute a Galois connection.

Proposition 8.55 (Properties of Generalized (Dual) Credal Set Function)**.** *Let m_ψ be a generalized credal set function and m_ψ° be a generalized dual credal set function. All the following properties hold:*

(a) The generalized credal set function m_ψ maps to weak*-closed, convex sets.

(b) The generalized dual credal set function m_ψ° maps to a linear subspace.

(c) The generalized credal set function m_ψ and generalized dual credal set function m_ψ° form a Galois connection.

Proof. (a) We have fixed ψ to a linear prevision. Hence, it is coherent. For any $\mathcal{G} \subseteq \mathrm{B}(\Omega)$, $m_\psi(\mathcal{G})$ is the set of all linear previsions which dominate ψ on $\mathcal{G}$. Theorem 3.6.1 in [Walley, 1991] then states that this set is weak*-closed and convex.

(b) Let $\mathcal{Q} \subseteq \Delta$.

Additivity Let $f, g \in m_\psi^\circ(\mathcal{Q})$. Then, for all $\nu \in \Delta$,

$$\nu(f+g) = \nu(f) + \nu(g) = \psi(f) + \psi(g) = \psi(f+g),$$

i.e., $f + g \in m_\psi^\circ(\mathcal{Q})$.

Homogeneity Let $f \in m_\psi^\circ(\mathcal{Q})$ and $\alpha \in \mathbb{R}$. Then, for all $\nu \in \Delta$,

$$\nu(\alpha f) = \alpha\nu(f) = \alpha\psi(f) = \psi(\alpha f),$$

i.e., $\alpha f \in m_\psi^\circ(\mathcal{Q})$. For homogeneity we need the easy fact that a linear prevision is not only positive homogeneous but generally homogeneous. For this consider a linear prevision ν and any gamble $f \in \mathrm{B}(\Omega)$ with $\alpha < 0$, then $\nu(\alpha f) = -\nu(-\alpha f) = \alpha\nu(f)$.

(c) The two functions constitute a Galois connection (cf. Proposition 8.29), $\mathcal{G} \subseteq m_\psi^\circ(\mathcal{Q}) \Leftrightarrow \mathcal{Q} \subseteq m_\psi(\mathcal{G})$. To this end, we show the left to right implication,

$$\nu \in \mathcal{Q} \Rightarrow \nu(g) = \psi(g), \forall g \in \mathcal{G} \Rightarrow \nu \in m_\psi(\mathcal{G}),$$

and the right to left implication,

$$g \in \mathcal{G} \Rightarrow \nu(g) = \psi(g), \forall \nu \in \mathcal{Q} \Rightarrow g \in m_\psi^\circ(\mathcal{Q}).$$

This concludes the proof. □

Galois connections possess a series of helpful properties [Birkhoff, 1940] (§V.7 and V.8). For instance, they give rise to a bipolar-closure operator. A non-empty subset $\mathcal{Q} \subseteq \Delta$ is bipolar-closed if and only if $\mathcal{Q} = m_\psi(m_\psi^\circ(\mathcal{Q}))$. A non-empty subset $\mathcal{G} \subseteq \mathrm{B}(\Omega)$ is bipolar-closed if and only if $\mathcal{G} = m_\psi^\circ(m_\psi(\mathcal{G}))$. Furthermore, Proposition 8.55 provides necessary conditions for bipolar-closed sets. For instance, a bipolar-closed set $\mathcal{G} \subseteq \mathrm{B}(\Omega)$ is a linear subspace. But is every such linear subspace a bipolar-closed set? No.

Example 8.56. *Let* $\Omega_2 \coloneqq \{1, 2\}$*, then* $\mathrm{B}(\Omega_2) = \{\alpha_1\chi_{\{1\}} + \alpha_2\chi_{\{2\}} \colon \alpha_1, \alpha_2 \in \mathbb{R}\}$*. Hence, linear functionals on* $\mathrm{B}(\Omega_2)$ *are defined via their behavior on the basis. Let* $\psi(\chi_{\{1\}}) = 1$*. Then,* $\{\alpha_1\chi_{\{1\}} \colon \alpha_1 \in \mathbb{R}\} \subseteq \mathrm{B}(\Omega_2)$ *is a linear subspace, but it is, as one can easily check, not bipolar-closed, because the demand for normalization of any linear prevision* ν *which coincides with* ψ *on* $\chi_{\{1\}}$ *requires* $\nu(\chi_{\{2\}}) = 0$*.*

Again, it seems to be more intricate than expected to characterize bipolar-closed sets. For pre-Dynkin systems and finitely additive measures we already collected some first hints that sets of measure zero play an important role in the characterization of bipolar-closed sets. In the case of linear subspaces and linear previsions, we observe a similar "combinatorial restriction". In order to improve understanding, let us replace 2^Δ by $2^{\mathrm{ba}(\Omega)}$ in Definition 8.53, Definition 8.54 and Proposition 8.55 (The proposition still holds.),which is equivalent to stating that linear previsions are not necessarily normalized, nor positive. Then, by leveraging the Hahn–Banach-type Theorem 1.5.14 in [Rao and Rao, 1983], one can easily see

that linearity of a subset $\mathcal{G} \subseteq \mathrm{B}(\Omega)$ is not only a necessary, but as well a sufficient condition for bipolar-closedness for those modified "credal set functions". Thus, the restriction to actual linear previsions makes the characterization of bipolar-closed sets more complex. A compelling, more exhaustive answer still waits to be found.

Analogous to the discussion in Section 8.5.4, it is possible to provide a lattice duality and interpolation scheme via the generalized (dual) credal set functions. Instead of the lattice of pre-Dynkin systems $(\mathfrak{D}, \subseteq)$, the interpolation is directed by the lattice of linear subspaces $(\mathcal{L}, \subseteq)$ of $\mathrm{B}(\Omega)$. As commonly known, the lattice of linear subspaces has the two operations $L_1 \wedge L_2 \coloneqq L_1 \cap L_2$ and $L_1 \vee L_2 \coloneqq \mathrm{lin}(L_1 \cup L_2)$. We denote the linear span with lin. Its minimal element is the trivial zero vector linear subspace $\{0\}$. Its maximal element is the entire space of all gambles $\mathrm{B}(\Omega)$. Due to higher generality of the here-presented (dual) credal set function, the interpolation provided is more fine-grained than for the previously given interpolation by pre-Dynkin systems. The following set containment (trivially) holds:

$$\{m_\psi(\mathcal{D})\colon \mathcal{D} \in \mathfrak{D}\} \subseteq \{m_\psi(L)\colon L \in \mathcal{L}\},$$

where $m_\psi(\mathcal{D}) = m_\psi(\{\chi_D\colon D \in \mathcal{D}\})$. In other words, the lattice of pre-Dynkin systems is "contained" in the lattice of of linear subspaces. However, as for pre-Dynkin systems the interpolation via linear subspaces is improper. The reason for this is again that not every linear subspace is bipolar-closed. By restriction to linear, bipolar-closed subspaces one can clean up the setup. For details we refer to Section 8.5.4. We do not make explicit the detailed reiteration of the same argument here.

In summary, we confirmed our findings of Section 8.5 extended to previsions. The generalized dual lattice setup underlines the structural consistency between the system of precision and its corresponding imprecise probability.

8.7 Conclusions and Open Questions

In this paper, we have explicated relations between the systems of precision and imprecise probabilities (respectively expectations). First, we have shown that the system of precision forms a pre-Dynkin system (respectively, a linear subspace). This structural insight raises a series of follow-up questions: how does the system of precision of a coherent prevision relate to the set of desirable gambles of this prevision? How does the preference ordering change the set structure of precision for the corresponding beliefs? What is the role of coherence with respect to the system of precision?

Second, we defined finitely additive probabilities on pre-Dynkin systems. The equivalence of extendability and coherence of such probabilities strengthens the link between quantum probability and imprecise probability. We speculate that further insights can be obtained by

exploiting this relationship. In addition, the generalization of finitely additive probabilities on pre-Dynkin systems to partial expectations directly opens the door to machine learning applications. In robust machine learning, the expected risk minimization framework is extended to more general expectation functionals. Partial expectation can, possibly after more computational investigations, deliver the desired robustness against dependencies in specific domains, such as privacy preservation, "not-missing-at-random" features, restricted data base access or multi-measurement data.

Finally, we developed a duality theory of systems of precision and imprecise probabilities (respectively expectations). A Galois connection defines a parametrized family of imprecise probabilities which follow an order structure provided by the lattice of pre-Dynkin systems (respectively the lattice of linear subspaces).

In modern statistics, especially in machine learning, probabilistic statements are increasingly tailored to individuals. Individual probabilistic statements, however, require justification. One can interpret probabilities on pre-Dynkin systems as probabilities which do not allow for such statements in the first place. One could perceive this fact as a weakness. We, in contrast, embrace its strength, when for ethical, legislative or other reasons, individualistic assignments are harmful, unjustifiable, forbidden or not desirable. We provide a first, rough interpolation scheme via the lattice duality. It demonstrates the space of adjustability of probabilistic assumptions in real-world scenarios. The involved pre-Dynkin systems are mathematical definitions of levels of group resolution. The question of how to choose such set-systems is related to the questions of intersectionality.

Several fundamental, technical questions remain open: how does the lattice duality imposed by pre-Dynkin systems or linear spaces relate to other dualities, such as convex duality, exploited in the field of imprecise probability. Can one easily characterize the bipolar-closed sets? Why is there no clear analogy between pre-Dynkin systems and linear subspaces?

We leave this collection of intriguing questions open to future work contributing to an understanding of the system of precision and the imprecise probability model.

Funding

This work was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC number 2064/1—Project number 390727645.

Acknowledgements

The authors thank the International Max Planck Research School for Intelligent System (IMPRS-IS) for supporting Rabanus Derr. Many thanks to Christian Fröhlich for helpful discussions and feedback. We thank all anonymous reviewers on a first version for the detailed and helpful feedback. In particular, with their help, the statements and proofs of Theorem 8.19 and Theorem 8.22 have been cleaned and shortened substantially.

8.8 Appendix

8.8.1 Lemmas and Proofs

Compatibility Structure

We have shown in Theorem 8.7 that every pre-Dynkin system can be decomposed into a union of algebras. A trivial follow-up question remains to be answered: is every union of algebras a pre-Dynkin system? The general answer is no.

Example 8.57 (Union of Algebras is not Necessarily Pre-Dynkin System)**.** *Let* $\Omega = \{1,2,3\}$*. Then,* $\mathcal{A}_1 \coloneqq \{\emptyset, 1, 23, \Omega\}$ *and* $\mathcal{A}_2 \coloneqq \{\emptyset, 12, 3, \Omega\}$ *are algebras on* Ω*. However,* $\mathcal{A}_1 \cup \mathcal{A}_2$ *do not form a pre-Dynkin system, as e.g.,* $1 \cup 3 = 13 \notin \mathcal{A}_1 \cup \mathcal{A}_2$*.*

However, we can give sufficient conditions under which the union over a set of algebras form a pre-Dynkin system. To this end, we have to introduce the so-called compatibility structure, which is made out of π-systems.

Definition 8.58 (π-System)**.** *Let* Ω *be an arbitrary base set. A* π*-system* $\mathcal{I}$ *is a non-empty subset of* 2^Ω *such that for arbitrary* $A_1, \ldots, A_n \in \mathcal{I}$*,* $\bigcap_{1 \leq i \leq n} A_i \in \mathcal{I}$*.*

Definition 8.59 (Compatibility Structure)**.** *A set of* π*-systems* $\{\mathcal{I}_i\}_{i \in I}$ *is called a* compatibility structure *of* $\mathcal{A}$ *if and only if the following two conditions hold:*

(a) $\bigcup_{i \in I} \mathcal{I}_i = \mathcal{A}$

(b) For every π*-system* $\mathcal{I} \subseteq \mathcal{A}$ *there exist* $i \in I$ *such that* $\mathcal{I} \subseteq \mathcal{I}_i$*.*

A "compatibility structure" is a system of set systems. We call it compatibility structure, because every finite intersection of elements in a π-system is compatible with every other finite intersection of elements from the same π-system. The π-systems in the compatibility structure are not necessarily disjoint. Given a set of π-systems, the union of these π-systems is not necessarily a compatibility structure. In fact, there are possible incompatibilities between the π-systems (e.g., Example 8.57). It turns out that being a compatibility structure is a sufficient condition for a set of algebras to form a pre-Dynkin system.

Theorem 8.60 (Union of Algebras is Pre-Dynkin System if Compatibility Structure)**.** *Let $\{\mathcal{A}_i\}_{i\in I}$ be a family of algebras. If $\{\mathcal{A}_i\}_{i\in I}$ is a compatibility structure, then $\mathcal{A}_\cup \coloneqq \bigcup_{i\in I} \mathcal{A}_i$ is a pre-Dynkin system.*

Proof. We show that $\mathcal{A}_\cup$ is a pre-Dynkin system if it is the union of the compatibility structure. First, it contains the empty set. Second, if $A \in \mathcal{A}_\cup$ there is an i such that $A \in \mathcal{A}_i$. Then $A^c \in \mathcal{A}_i$, thus $A^c \in \mathcal{A}_\cup$. Third, let $\{A_j\}_{j\in[n]}$ be a subset of $\mathcal{A}_\cup$ with $n \in \mathbb{N}$ such that $A_k \neq \emptyset$, $A_l \neq \emptyset$ and $A_k \cap A_l = \emptyset$ for all $k \neq l; k, l \in [n]$. Thus, $\{A_j\}_{j\in[n]} \cup \{\emptyset\}$ is closed under finite intersection. Hence, it is a π-system. By definition of a compatibility structure, there is an $i \in I$ such that $\{A_j\}_{j\in[n]} \cup \{\emptyset\} \subseteq \mathcal{A}_i$. Since $\mathcal{A}_i$ is an algebra, it is closed under disjoint union. Thus, $\bigcup_{j\in[n]} A_j \in \mathcal{A}_i \subseteq \mathcal{A}_\cup$. □

Supremum of a Chain of Algebras Is an Algebra

The following lemma is used to prove that every pre-Dynkin system can be dissected into algebras (Theorem 8.7).

Lemma 8.61 (Supremum of a Chain of Algebras is an Algebra)**.** *Let $(\{\mathcal{A}_i\}_{i\in I}, \subseteq)$ be a non-empty chain of algebras. Then it has an upper bounding algebra $\mathcal{A}_{\text{sup}}$.*

Proof. Let $\mathcal{A}_{\text{sup}} = \bigcup_{i\in I} \mathcal{A}_i$. Trivially, $\mathcal{A}_i \subseteq \mathcal{A}_{\text{sup}}$ for every $i \in I$. It remains to show that $\mathcal{A}_{\text{sup}}$ is an algebra. It is non-empty by construction. Consider any finite subset $\{A_1, \dots, A_n\} \subseteq \mathcal{A}_{\text{sup}}$. Without loss of generality, there exist $j \in I$ such that $\{A_1, \dots, A_n\} \subseteq \mathcal{A}_j$. Thus,

$$\bigcup_{i=1}^{n} A_i \in \mathcal{A}_j \subseteq \mathcal{A}_{\text{sup}}.$$

Furthermore, for every $A \in \mathcal{A}_{\text{sup}}$, there exist $j \in I$ such that $A \in \mathcal{A}_j$. Thus,

$$A^c \in \mathcal{A}_j \subseteq \mathcal{A}_{\text{sup}},$$

which concludes the proof. □

Sufficient Conditions for Bipolar-Closed Sets

In order to give sufficient conditions for bipolar-closed sets we introduce the following notion.

Definition 8.62 (Weak Atom with Respect to Pre-Dynkin System)**.** *A* weak atom *with respect to a pre-Dynkin system $\mathcal{D}$ on Ω is a set $A \subseteq \Omega$ such that every $B \in \mathcal{D}$ with $B \subseteq A$ is either $B = \emptyset$ or $B = A$.*

Weak atoms always exist, for instance the empty set is a weak atom with respect to any pre-Dynkin system $\mathcal{D}$ on any non-empty set Ω.

Lemma 8.63 (Decomposition into Pre-Dynkin Set and Atom)**.** *Let $\mathcal{D} \subseteq 2^\Omega$ be a pre-Dynkin system. Every element $B \in 2^\Omega \setminus \mathcal{D}$ can be expressed as a disjoint union of an element $B_\mathcal{D} \in \mathcal{D}$ and a weak atom $A \in 2^\Omega \setminus \mathcal{D}$ with respect to $\mathcal{D}$.*

Proof. We define $B_\mathcal{D} \subseteq B$ as a maximal set $B_\mathcal{D} \in \mathcal{D}$ such that for all $H \in \mathcal{D}$ with $B_\mathcal{D} \subseteq H \subseteq B$ we have $H = B_\mathcal{D}$. By Zorn's lemma, such an element always exists. To see this, consider the poset of all subsets of B which are in $\mathcal{D}$ ordered by set inclusion. Then, we decompose $B \setminus B_\mathcal{D} = A$. Clearly, $A \neq \emptyset$. Furthermore, $A \notin \mathcal{D}$ and there is no $H \in \mathcal{D}$ such that $H \subseteq A$ except of the empty set, otherwise $B_\mathcal{D}$ would not have been maximal in our sense. Both statements follow from the closedness of $\mathcal{D}$ under disjoint union. Thus, A is a weak atom with respect to $\mathcal{D}$. □

With these definitions at hand we can show the following crucial lemma.

Lemma 8.64 (Weak Atoms Are Not in the Bipolar-Closed Sets – Finite, Discrete Setting)**.** *Let $\Omega = [n]$. We fix a finitely additive probability ψ on 2^Ω such that $\psi(F) > 0$ for every $F \in 2^\Omega \setminus \{\emptyset\}$. Let $\mathcal{D} \subseteq 2^\Omega$ be a pre-Dynkin system. Then, for every weak atom A with respect to $\mathcal{D}$ such that $A \notin \mathcal{D}$, we have*

$$A \notin m_\psi^\circ(m_\psi(\mathcal{D})),$$

where m_ψ and m_ψ° are defined following Definition 8.23 and 8.28, respectively.

Proof. We show the Lemma by constructing, for every weak atom $A \subseteq \Omega$ with respect to $\mathcal{D}$ such that $A \notin \mathcal{D}$, a probability measure on 2^Ω which coincides with ψ on the pre-Dynkin system $\mathcal{D}$ but differs to ψ on the event A. Observe, weak atoms $A \notin \mathcal{D}$ exist as long as $\mathcal{D} \neq 2^\Omega$ (by Lemma 8.63). Second, the weak atom A contains an element $i \in A$ such that $\{i\} \notin \mathcal{D}$. Otherwise, $A \in \mathcal{D}$.

So, we fix an arbitrary weak atom A with respect to $\mathcal{D}$ such that $A \notin \mathcal{D}$. Due to the finiteness of $[n]$ we can represent every measure $\nu \in \Delta$ as $\nu_1, \ldots, \nu_n$ with the constraints $0 \leq \nu_k \leq 1$ and $\sum_{k=1}^n \nu_k = 1$. With this in mind, we construct a probability $\nu \in \Delta$ such that $\nu(B) = \psi(B)$ for all $B \in \mathcal{D}$ but $\nu(A) \neq \psi(A)$.

To this end, we define ν on all elements in $[n]$. We choose

$$\nu_i = \psi_i + \epsilon$$

for the $i \in A$ specified earlier, with $\epsilon := \min_{k \in [n]} \psi_k$. By assumption, $\epsilon > 0$.

By Theorem 8.7 we know that we can decompose $\mathcal{D}$ into sub-algebras. In fact, we can decompose $\mathcal{D}$ into a finite set of sub-algebras as $\mathcal{D}$ is finite itself. Furthermore, set algebras on finite sets are build upon a finite set of atoms [Rao and Rao, 1983] (Remark 1.1.17.(2)). Those atoms form a partition of the finite set; hence, each set algebra on a finite set is in one-to-one correspondence to a partition (cf. [Rao and Rao, 1983] (Proposition 1.1.18)). Hence, $\mathcal{D}$ determines a family of partitions corresponding to its blocks. In detail, $\mathcal{D} = \bigcup_{o\in O} \mathcal{A}_o$ for set algebras $\mathcal{A}_o$ and finite O (Theorem 8.7). Then, we define $\{\mathcal{B}_o\}_{o\in O}$ as the corresponding set of partitions. In particular, $\mathrm{D}(\mathcal{B}_o) = \mathcal{A}_o$. Thus, $\mathcal{D} = \bigcup_{o\in O} \mathrm{D}(\mathcal{B}_o) = \mathrm{D}(\bigcup_{o\in O} \mathcal{B}_o)$, because

$$\bigcup_{o\in O} \mathrm{D}(\mathcal{B}_o) \subseteq \mathrm{D}(\bigcup_{o\in O} \mathcal{B}_o) \subseteq \mathrm{D}(\mathcal{D}) = \mathcal{D} = \bigcup_{o\in O} \mathcal{A}_o = \bigcup_{o\in O} \mathrm{D}(\mathcal{B}_o).$$

For every partition $\mathcal{B}_o$ we have one $B_o \in \mathcal{B}_o$ such that $i \in B_o$. Obviously, $B_o \cap A \neq \emptyset$. But as well, $B_o \cap A^c \neq \emptyset$. Otherwise, $B_o \subseteq A$, thus $B_o = \emptyset$, because A is a weak atom which is not contained in the pre-Dynkin system $\mathcal{D}$. Thus, there is at least $j_o \in B_o$ with $j_o \notin A$. For one such j_o, we define the probability

$$\nu_{j_o} \coloneqq \psi_{j_o} - \frac{\epsilon}{|\{j_o \colon o \in O\}|}.$$

By choice of ϵ and $1 \leq |\{j_o \colon o \in O\}| \leq \infty$ we have $\nu_{j_o} \geq 0$. Observe, we divide ϵ by the number of atoms on which we decrease the probability to guarantee normalization of ν. Finally, for all $l \in [n] \setminus (\{j_o \colon o \in O\} \cup A)$, we set $\nu_l = \psi_l$. This assignment leads to the following conclusions: $\nu(B_o) = \psi(B_o)$ for every $o \in O$. More generally, for all $B \in \bigcup_{o\in O} \mathcal{B}_o$, we have $\nu(B) = \psi(B)$.

The probability distribution ν is uniquely determined by the probability on the partitions, because $\mathcal{D} = \mathrm{D}(\bigcup_{o\in O} \mathcal{B}_o)$ (see above) and the π-λ-Theorem [Williams, 1991] (Lemma A.1.3). So, $\nu \in m_\psi(\mathcal{D})$. But, $\nu(A) = \nu(\{i\}) + \nu(A \setminus \{i\}) = \psi(\{i\}) + \psi(A \setminus \{i\}) + \epsilon \neq \psi(A)$. Thus, $A \notin m_\psi^\circ(m_\psi(\mathcal{D}))$. □

Intersection of Linear Subspaces

Lemma 8.65 (Intersection of Simple Gamble Spaces)**.** *Let $\mathcal{A}, \mathcal{B} \subseteq 2^\Omega$ be two algebras. Then,*

$$\mathrm{S}(\Omega, \mathcal{A}) \cap \mathrm{S}(\Omega, \mathcal{B}) = \mathrm{S}(\Omega, \mathcal{A} \cap \mathcal{B}).$$

Proof. It is clear that $\mathrm{S}(\Omega, \mathcal{A} \cap \mathcal{B}) \subseteq \mathrm{S}(\Omega, \mathcal{A})$ and $\mathrm{S}(\Omega, \mathcal{A} \cap \mathcal{B}) \subseteq \mathrm{S}(\Omega, \mathcal{B})$, which makes the right to left set inclusion obvious. For the left to right inclusion, observe that a func-

tion $f \in \mathrm{S}(\Omega, \mathcal{A}) \cap \mathrm{S}(\Omega, \mathcal{B})$ is a linear combination of indicator gambles of disjoint events $A_1, \ldots, A_n \in \mathcal{A}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and a linear combination of indicator gambles of disjoint events $B_1, \ldots, B_m \in \mathcal{B}$, $\beta_1, \ldots, \beta_m \in \mathbb{R}$. Without loss of generality, we can demand that all $\alpha_1, \ldots, \alpha_n$ (respectively $\beta_1, \ldots, \beta_m$) have to be pairwise distinct, because we can replace indicator gambles scaled with the same factor by the appropriately scaled indicator function of the disjoint union of the events. Then, there is $i \in [n]$ for every $j \in [m]$ such that $\alpha_i = \beta_j$ and vice versa. More importantly, there is $i \in [n]$ for every $j \in [m]$ such that $A_i = B_j$ and vice versa. Hence, $\{A_1, \ldots, A_n\} = \{B_1, \ldots, B_m\} \in \mathcal{A} \cap \mathcal{B}$. □

Lemma 8.66 (Intersection of Measurable Gamble Spaces)**.** *Let $\mathcal{A}_\sigma, \mathcal{B}_\sigma \subseteq 2^\Omega$ be two σ-algebras. Then,*

$$\mathrm{B}(\Omega, \mathcal{A}_\sigma) \cap \mathrm{B}(\Omega, \mathcal{B}_\sigma) = \mathrm{B}(\Omega, \mathcal{A}_\sigma \cap \mathcal{B}_\sigma).$$

Proof. It is clear that $\mathrm{B}(\Omega, \mathcal{A}_\sigma \cap \mathcal{B}_\sigma) \subseteq \mathrm{B}(\Omega, \mathcal{A}_\sigma)$ and $\mathrm{B}(\Omega, \mathcal{A}_\sigma \cap \mathcal{B}_\sigma) \subseteq \mathrm{B}(\Omega, \mathcal{B}_\sigma)$, which makes the right to left set inclusion obvious. For the left to right inclusion, observe that for a function $f \in \mathrm{B}(\Omega, \mathcal{A}_\sigma) \cap \mathrm{B}(\Omega, \mathcal{B}_\sigma)$, the preimage of any Borel-measurable set in $\mathbb{R}$ is contained in $\mathcal{A}_\sigma$ and $\mathcal{B}_\sigma$. □

8.8.2 Names of Pre-Dynkin Systems and Dynkin Systems

See Table 8.3 for a list of names for pre-Dynkin systems. Table 8.4 summarizes a list of names for Dynkin systems.

Table 8.3: A summary of names for pre-Dynkin systems found in the literature.

pre-Dynkin system [Schurz and Leitgeb, 2008]
additive-class [Rao and Rao, 1983] (p. 2)
concrete logic [Ovchinnikov, 1999, De Simone et al., 2007]
partial field [Godowski, 1981]
quantum-mechanical algebra [Suppes, 1966]
semi-algebra [Khrennikov, 2009b] (p. 13)
set-representable orthomodular poset [Pták, 1998]

Table 8.4: A summary of names for Dynkin systems found in the literature.

Dynkin system [Jun et al., 2011]
d-system [Williams, 1991] (p. 193)
λ-class [Chow and Teicher, 1988] (p. 7)
quantum-mechanical σ-algebra [Suppes, 1966]
σ-class [Gudder, 1984]

8.8.3 Credal Sets of Pre-Dynkin System Probabilities — Credal Sets of Distorted Probabilities

To the best of the authors' knowledge, there has not been any attempt to parametrize a family of imprecise probabilities via the set of induced precise probabilities. In fact, a much better known class of imprecise probabilities is parametrized via distortion functions [Wirch and Hardy, 2001]. We use distorted probability functions as they regularly occur as examples of imprecise probabilities [Walley, 1991, Wirch and Hardy, 2001, Fröhlich and Williamson, 2024]. In particular, there is a one-to-one correspondence of distorted probabilities as defined in the following and so-called spectral risk measures, an important class of coherent upper previsions often used in economics and finance [Fröhlich and Williamson, 2024].

Definition 8.67 (Credal Set of Distorted Probability)**.** *Let* $\gamma\colon [0,1] \to [0,1]$ *be a concave, increasing function with* $\gamma(0) = 0$ *and* $\gamma(1) = 1$*. Let* ψ *denotes a finitely additive probability on an algebra* 2^Ω *on* Ω*. We overload the definition of a credal set*

$$M(\psi, \gamma) \coloneqq \{\nu \in \Delta\colon \nu(F) \leq \gamma(\psi(F)),\ \forall F \in 2^\Omega\}.$$

How does the credal set of probabilities for distorted probabilities relate to the credal set of probabilities for a probability defined on a Dynkin system?

Lemma 8.68 (Distortion Lemma)**.** *Let* $\gamma\colon [0,1] \to [0,1]$ *be a concave, increasing function with* $\gamma(0) = 0$ *and* $\gamma(1) = 1$*. If* γ *is not the identity function, then* $\gamma(x) > x$ *for all* $x \in (0,1)$*.*

Proof. As γ is concave, it is, in particular, quasi-concave. Thus, $x \mapsto \frac{\gamma(x)}{x}$ is decreasing [Rubshtein et al., 2016] (Definition 10.1.1). This gives the following inequalities:

$$\frac{\gamma(x)}{x} \geq \frac{\gamma(x')}{x'} \geq \frac{\gamma(1)}{1} = 1,$$

for $0 < x \leq x'$. This implies, if there were $x \in (0,1)$ such that $\gamma(x) = x$, then γ would be the identity function. We excluded this by assumption, and thus it follows $\gamma(x) > x$ for all $x \in (0,1)$. □

In the following Proposition, we show that the set of events on which all measures of a credal set $M(\psi, \gamma)$ coincide forms a pre-Dynkin system. Actually, it is the *system of certainty*, i.e., the set of all events which get assigned either the value 0 or the value 1. We reuse the notation of the dual credal set function m°_ψ which, as we noted earlier, maps a set of probabilities to the set structure on which those probabilities coincide.

Proposition 8.69 (Events of Precise Probability for Distorted Probabilities)**.** *Let* $\gamma\colon [0,1] \to [0,1]$ *be a concave, increasing function with* $\gamma(0) = 0$ *and* $\gamma(1) = 1$*, which is not the identity function. Let* ψ *denote a finitely additive probability on an algebra* 2^Ω *on* Ω*. Let*

$M(\psi, \gamma)$ be the credal set (Definition 8.67) and m°_ψ the dual credal set function (Definition 8.28). Let $\mathcal{F}_0 \coloneqq \{F \in 2^\Omega \colon \psi(F) = 0\}$ denote the set of all measure zero sets and $\mathcal{F}_{01} \coloneqq \{F \in 2^\Omega \colon \psi(F) = 0$ or $\psi(F) = 1\}$ denote the set of all measure zero or one sets. Then,

$$m^\circ_\psi(M(\psi, \gamma)) = \mathrm{D}(\mathcal{F}_0) = \mathcal{F}_{01}.$$

Proof. We show the equality of all sets via a circular set containment.

(a) $\mathcal{F}_{01} \subseteq \mathrm{D}(\mathcal{F}_0)$
If $F \in 2^\Omega$ such that $\psi(F) = 0$, then clearly $F \in \mathcal{F}_0 \subseteq \mathrm{D}(\mathcal{F}_0)$. If $F \in 2^\Omega$ such that $\psi(F) = 1$, then $\psi(F^c) = 0$. Thus, $F^c \in \mathcal{F}_0$ because pre-Dynkin systems are closed under complement $F \in \mathrm{D}(\mathcal{F}_0)$.

(b) $\mathrm{D}(\mathcal{F}_0) \subseteq m^\circ_\psi(M(\psi, \gamma))$
First, we write out

$$m^\circ_\psi(M(\psi, \gamma)) = \{F \in 2^\Omega \colon \nu(F) = \psi(F),\ \forall \nu \in \{\nu' \in \Delta \colon \nu'(F) \leq \gamma(\psi(F)),\ \forall F \in 2^\Omega\}\}$$

Since $\gamma(0) = 0$, it is easy to see that $\mathcal{F}_0 \subseteq m^\circ_\psi(M(\psi, \gamma))$. By Proposition 8.31, we know that $m^\circ_\psi(M(\psi, \gamma))$ is a pre-Dynkin system. Thus, $\mathrm{D}(\mathcal{F}_0) \subseteq m^\circ_\psi(M(\psi, \gamma))$.

(c) $m^\circ_\psi(M(\psi, \gamma)) \subseteq \mathcal{F}_{01}$
We show this set inclusion via contraposition. If $F \in 2^\Omega$ has measure $\psi(F) \in (0, 1)$, then $F \notin m^\circ_\psi(M(\psi, \gamma))$. For this we have to argue that there is a measure $\nu_F \in M(\psi, \gamma)$ for every $F \in 2^\Omega$ with $\psi(F) \in (0, 1)$ such that $\nu_F(F) \neq \psi(F)$.

Observe that $\gamma \circ \psi$ defines a normalized, monotone, submodular set function on 2^Ω [Denneberg, 1994] (p. 17). Furthermore, any normalized, monotone, submodular set function induces a translation equivariant, monotone, positively homogeneous and subadditive functional $L_{\gamma\circ\psi}$ on all $f \in \mathrm{B}(\Omega)$ such that $L_{\gamma\circ\psi}(\chi_F) = (\gamma \circ \psi)(F)$ for all $F \in 2^\Omega$ [Huber, 1981] (p. 260), [Walley, 1991] (p. 130), [Denneberg, 1994] (Proposition 5.1, Theorem 6.3). Hence, $L_{\gamma\circ\psi}$ is a coherent upper prevision [Walley, 1991] (p. 65). Thus, Walley's extreme point theorem applies [Walley, 1991] (Theorem 3.6.2 (c)). (Even though this theorem is stated in terms of coherent lower previsions, it applies to coherent upper previsions, too. The weak*-compactness of the credal set $M(\psi, \gamma)$, which is given by the coherence of $L_{\gamma\circ\psi}$ and [Walley, 1991] (Theorem 3.6.1), is crucial.) For any function $f \in \mathrm{B}(\Omega)$, in particular any χ_F with $F \in 2^\Omega$, there is a linear prevision $f \mapsto \langle f, \nu\rangle$ with $\nu \in \Delta$ on $\mathrm{B}(\Omega)$ dominated by $L_{\gamma\circ\psi}$ such that $\langle \chi_F, \nu\rangle = L_{\gamma\circ\psi}(\chi_F)$. More concretely, for any $F \in 2^\Omega$ there is a $\nu_F \in M(\psi, \gamma)$ such that $\nu_F(F) = (\gamma \circ \psi)(F)$. If $\psi(F) \in (0, 1)$, then Lemma 8.68 applies and

$\nu_F(F) = (\gamma \circ \psi)(F) = \gamma(\psi(F)) > \psi(F)$ gives the desired inequality. In conclusion, there is no $F \in 2^\Omega$ with measure $\psi(F) \in (0,1)$ such that $F \in m^\circ_\psi(M(\psi,\gamma))$. This implication finalizes the proof.

□

We call the set $\mathcal{F}_{01}$ a *system of certainty* for ψ. Hence, the system of precision of a distorted probability is the system of certainty. Since the proposition clarifies the relation from distorted imprecise probability to probability on Dynkin systems, the reverse question immediately follows: when is $M(\psi|_\mathcal{D}, \mathcal{D}) \subseteq M(\psi,\gamma)$? In other words, given an extendable probability defined on a pre-Dynkin system, what is a distortion function of an extension of this probability such that the credal set of the former is contained in the credal set of the latter. The simple example below gives an instantiation of this problem for which it is easy to find a solution. However, the problem is harder for more general cases.

Example 8.70. *Let* $\Omega = \{1,2\}$ *and* $\psi(1) = \psi(2) = 0.5$. *Let* $\mathcal{D} = \{\emptyset, \Omega\}$. *Then, for the admittedly extreme distortion function* $\gamma(0) = 0$, *otherwise* $\gamma(a) = 1$, *for* $a \in (0,1]$ *the credal set* $M(\psi,\gamma) = \Delta$ *is the set of all probability measures. Hence,* $M(\psi|_\mathcal{D}, \mathcal{D}) = \Delta \subseteq M(\psi,\gamma)$.

8.8.4 Dynkin Systems and Countably Additive Probability

In the main text, we presented the majority of the results for the general case: pre-Dynkin systems and finitely additive probabilities. We did so, as we emphasize clearly here, not for the sake of mathematical generality. There are probabilistic problems which demand for the use of finitely additive probabilities, e.g., von Mises frequentistic notion of probability [Von Mises and Geiringer, 1964, Schurz and Leitgeb, 2008]. As pre-Dynkin systems and finitely additive probabilities walk hand in hand, so too do Dynkin systems and countably additive probabilities. It is possible to strengthen some results when we assume Dynkin systems and countably additive probabilities. Furthermore, countably additive probabilities are more familiar to students of probability theory. Finitely additive probabilities still eke out an exotic living [Rao and Rao, 1983, Kadane et al., 1986, Chichilnisky, 2010]. Nevertheless, they are central in large parts of literature on imprecise probability [Walley, 1991, Augustin et al., 2014].

Dynkin Systems

Many of the following results are analogous to the statements for pre-Dynkin systems. The subsequent theorem is analogous to Theorem 8.7. We remark, however, that we cannot use the same proof-technique as in Theorem 8.7 because the generalization of Lemma 8.61 to

σ-algebras does not hold. Hence, we use a different technique to prove Theorem 8.71 for Dynkin systems and σ-algebras.

Theorem 8.71 (Dynkin Systems are made out of σ-Algebras)**.** *Let $\mathcal{D}_\sigma$ be a Dynkin system on an arbitrary set Ω. Then, there is a unique family of maximal σ-algebras $\{\mathcal{A}_i\}_{i\in I}$ such that $\mathcal{D}_\sigma = \bigcup_{i\in I} \mathcal{A}_i$. We call these σ-algebras the* σ-blocks *of $\mathcal{D}_\sigma$.*

Proof. Since Dynkin systems are pre-Dynkin systems, Theorem 8.7 guarantees that $\mathcal{D}_\sigma$ is constituted of a set of algebras $\{\mathcal{A}_i\}_{i\in I}$. Each algebra $\mathcal{A}_i$ is closed under finite intersection. Thus, it is a "compatible collection" following the terms of Gudder [1973]. It follows that $\mathcal{A}_i$ is contained in a sub-σ-algebra of $\mathcal{D}_\sigma$ by Theorem 2.1 in [Gudder, 1973]. As any sub-σ-algebra is an algebra and $\mathcal{A}_i$ is maximal, i.e., there is no algebra contained in $\mathcal{D}$ such that $\mathcal{A}_i$ is a strict sub-algebra of this algebra, $\mathcal{A}_i$ itself is a σ-algebra. Hence, $\mathcal{A}_i$ are the σ-blocks of $\mathcal{D}_\sigma$. □

Not every union of σ-algebras is a Dynkin system (cf. Appendix 8.8.1). But, if a union of σ-algebras forms a compatibility structure, then it is a Dynkin system.

Theorem 8.72 (Union of σ-Algebras is Dynkin system if Compatibility Structure)**.** *Let $\{\mathcal{A}_i\}_{i\in I}$ be a family of σ-algebras on an arbitrary Ω. If $\{\mathcal{A}_i\}_{i\in I}$ is a compatibility structure, then $\mathcal{A}_\cup \coloneqq \bigcup_{i\in I} \mathcal{A}_i$ is a Dynkin system.*

Proof. We show that $\mathcal{A}_\cup$ is a Dynkin system if it is the union of the compatibility structure. First, it contains the empty set. Second, if $A \in \mathcal{A}_\cup$ there is an i such that $A \in \mathcal{A}_i$. Then $A^c \in \mathcal{A}_i$, thus $A^c \in \mathcal{A}_\cup$. Third, let $\{A_j\}_{j\in J}$ be a subset of $\mathcal{A}_\cup$ with $J \subseteq \mathbb{N}$ such that $A_k \neq \emptyset$, $A_l \neq \emptyset$ and $A_k \cap A_l = \emptyset$ for all $k \neq l; k, l \in J$. Thus, $\{A_j\}_{j\in J} \cup \{\emptyset\}$ is closed under finite intersection. It is thus a π-system. By definition of a compatibility structure there is an $i \in I$ such that $\{A_j\}_{j\in J} \cup \{\emptyset\} \subseteq \mathcal{A}_i$. Since $\mathcal{A}_i$ is a σ-algebra, it is closed under countable disjoint union. Thus, $\bigcup_{j\in J} A_j \in \mathcal{A}_i \subseteq \mathcal{A}_\cup$. □

Technical Setup

In order to work on firm ground when introducing countably additive probabilities, we change our basic technical setup. We summarize the used notations in Table 8.5. First, we assume now that Ω is a Polish space, that is, a separable completely metrizable topological space [Huber, 1981] (p. 20). We denote by $\mathcal{F}_\sigma$ the Borel-σ-algebra with respect to the given topology. With $\mathrm{ca}(\Omega, \mathcal{F}_\sigma)$, we denote the space of all finite, signed countably additive measures on $(\Omega, \mathcal{F}_\sigma)$ (cf. [Huber, 1981] (p. 20)).

We define $\Delta_\sigma \subseteq \mathrm{ca}(\Omega, \mathcal{F}_\sigma)$ as the set of all probability measures on $(\Omega, \mathcal{F}_\sigma)$. Every measure in Δ_σ is regular [Huber, 1981] (p. 20). The set $\mathrm{ca}(\Omega, \mathcal{F}_\sigma)$ is equipped with the total

Table 8.5: Summary of used notations in Appendix 8.8.4.

Ω	Polish Space
$\mathcal{D}_\sigma$	Dynkin system on Ω (Definition 8.1)
μ_σ	Countably additive probability defined on $\mathcal{D}_\sigma$ (Definition 8.9)
$\sigma(\mathcal{A})$	σ-algebra hull of set system $\mathcal{A} \subseteq 2^\Omega$
$\mathcal{F}_\sigma$	Borel-σ-algebra on Ω
$\text{ca}(\Omega, \mathcal{F}_\sigma)$	Set of bounded, signed, countably additive measures on $\mathcal{F}_\sigma$
$\Delta_\sigma(\Omega, \mathcal{F}_\sigma)$ resp. Δ_σ	Set of countably additive probability measures on $\mathcal{F}_\sigma$
$M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)$	σ-Credal set of μ_σ on $\mathcal{D}_\sigma$ (Proposition 8.77)
$\underline{\mu}_{\mathcal{D}_\sigma}$, $\overline{\mu}_{\mathcal{D}_\sigma}$	Lower respectively upper coherent σ-extension (Proposition 8.79)

variation distance $d_{TV}(\nu, \nu') \coloneqq \sup_{A \in \mathcal{F}_\sigma} |\nu(A) - \nu'(A)|$ [Aliprantis and Border, 2006] (p. 366). We emphasize that this is a different topology in comparison to the topology used above. The weak* topology is weaker than the topology induced by the total variation distance.

Lastly, we denote the smallest σ-algebra which contains $\mathcal{A}$ with $\sigma(\mathcal{A})$. We say that $\sigma(\mathcal{A})$ is the σ-algebra hull of $\mathcal{A}$.

Dynkin Probability Spaces

The definition of probability measures on pre-Dynkin systems directly applies to Dynkin systems, too. Analogous to Kolmogorov's probability space, we can now leverage Definition 8.9 for a definition of a Dynkin probability space.

Definition 8.73 (Dynkin Probability Space [Gudder, 1969] (p. 296))**.** *The triple* $(\Omega, \mathcal{D}_\sigma, \mu_\sigma)$ *is called a* Dynkin probability space *if and only if (a)* Ω *is a non-empty base space, (b)* $\mathcal{D}_\sigma$ *is a Dynkin system on* Ω *and (c)* μ_σ *is a countably additive probability measure on* $\mathcal{D}_\sigma$ *following Definition 8.9.*

Dynkin probability spaces have been called "quantum probability spaces" in quantum probability theory [Gudder, 1969]. Dynkin probability spaces generalize Kolmogorov's probability spaces. If $\mathcal{D}_\sigma$ were a σ-algebra, then the Dynkin probability space would become a classical probability space. Theorem 8.71 provides another interesting link between Dynkin and classical probability spaces.

Proposition 8.74. *Every Dynkin probability space* $(\Omega, \mathcal{D}_\sigma, \mu_\sigma)$ *defines a collection of classical probability space* $\{(\Omega, \mathcal{A}_i, \mu_i)\}_{i \in I}$ *where* $\mathcal{D}_\sigma = \bigcup_{i \in I} \mathcal{A}_i$ *and* $\mu_i \coloneqq \mu_\sigma|_{\mathcal{A}_i}$ *are* consistent*, i.e.,* $\mu_i(A) = \mu_j(A)$ *for any* $A \in \mathcal{A}_i \cap \mathcal{A}_j$ *and* $i, j \in I$*.*

Such "multi-Kolmogorov" probability spaces have similarly been formalized in [Khrennikov, 2009a] (p. 32) or [Vorob'ev, 1962] (p. 154). The multiplicity of probability spaces has been

interpreted as a collection of contexts. Each context possesses its own classical probability space. (In quantum physics, we obtain classical behavior in single context and quantum behavior across context. In machine learning, each marginal scenario gets equipped an own probability space, i.e., an own context (cf. [Cuadras et al., 2002, Eban et al., 2014]).) Again, as for pre-Dynkin systems and finitely additive probabilities, the question of embedding Dynkin probability spaces into classical probability spaces arises.

Conditions for Extendability for Countably Additive Probabilities

Definition 8.75 (σ-Extendability)**.** *Let $\mathcal{F}_\sigma$ be the Borel-σ-algebra on a Polish space Ω and $\mathcal{D}_\sigma$ a Dynkin system contained in this σ-algebra. With Δ_σ we denote the set of all countably additive probability measures on $(\Omega, \mathcal{F}_\sigma)$. We call a countably additive probability μ_σ on $\mathcal{D}_\sigma$* σ-extendable *if and only if there is a countably additive probability measure $\nu \in \Delta_\sigma$ such that $\nu|_{\mathcal{D}_\sigma} = \mu_\sigma$.*

We call this extendability "σ" to emphasize the difference to Definition 8.15 and the properties of countably additive probability measures. Extendability of Dynkin probability spaces, as well as for finitely additive probabilities on pre-Dynkin systems, has already been part of discussions in quantum probability since 1969 [Gudder, 1969] up to more current times [De Simone and Pták, 2010]. Most of the results, e.g., [Gudder, 1984, De Simone et al., 2007, De Simone and Pták, 2010], are only stated in terms of finitely additive probability measure and/or apply only on a structurally restricted class of Dynkin systems. A theorem due to Maharam [1972] in the literature on marginal probabilities can be tweaked to give a universal sufficient and necessary criterion for σ-extendability.

Theorem 8.76 (σ-Extendability Condition for Countably Additive Probability)**.** *Let $\mathcal{F}_\sigma$ be the Borel-σ-algebra on a Polish space Ω and $\mathcal{D}_\sigma$ a Dynkin system contained in this σ-algebra such that $\sigma(\mathcal{D}_\sigma) = \mathcal{F}_\sigma$. Let μ_σ be a countably additive probability on $\mathcal{D}_\sigma$. For each σ-block $\mathcal{A}_i$ of $\mathcal{D}_\sigma$ we assume that $\mu_i := \mu_\sigma|_{\mathcal{A}_i}$ is inner regular, i.e., if $\mu(G) = \sup\{\mu(K) \colon K \subseteq G, K$ compact and measurable$\}$ [Schechter, 1997] (p. 808). The probability μ_σ is σ-extendable if and only if*

$$\sum_{j=1}^{n} f_j(\omega) \geq 0, \ \forall \omega \in \Omega \implies \sum_{j=1}^{n} \int_\Omega f_j(\omega) d\mu_{i_j}(\omega) \geq 0$$

for all finite families of measurable simple gambles $\{f_j\}_{j\in[n]} \subseteq S(\Omega, \mathcal{D}_\sigma)$.

Proof. Proposition 8.74 separates the Dynkin probability space into a set of Kolmogorov probability spaces. Then, Theorem 8.1 in [Maharam, 1972] applies. As technical requirements, we emphasize that Ω is a Hausdorff space and all restrictions μ_i are inner regu-

lar. Thus, there exists a probability measure on the algebra generated by $\mathcal{D}_\sigma$. Since the probability measure is countably additive, Caratheodory's Extension Theorem [Williams, 1991] (Theorem 1.7) states that it can be uniquely extended to the Borel-σ-algebra $\mathcal{F}_\sigma = \sigma(\mathcal{D}_\sigma)$. □

Similar and related results can be found in [Vorob'ev, 1962, Kellerer, 1964, Horn and Tarski, 1948]. In particular, every so-called marginal scenario can be represented as a countably additive probability on a Dynkin system (cf. [Vorob'ev, 1962], [Gudder, 1984] (Example 4.2)). Marginal scenarios often arise in practical setups. They are defined as settings in which for a collection of random variables only specific partial joint distributions of the random variables are given. For those scenarios there are purely combinatorial conditions on the Dynkin system sufficient for the extendability [Vorob'ev, 1962, Kellerer, 1964].

Credal Set of Countably Additive Probabilities on Dynkin Systems

The credal set of probabilities which we defined in Corollary 8.20 contains finitely additive probabilities, some of which might not be countably additive. For this reason we redefine the credal set for countably additive probabilities. The credal set is the set of all countably additive probabilities which coincide with the reference probability μ_σ on $\mathcal{D}_\sigma$.

Proposition 8.77 (σ-Credal Set for Probabilities on Dynkin Systems)**.** *Let $\mathcal{F}_\sigma$ be the Borel-σ-algebra on a Polish space Ω and $\mathcal{D}_\sigma$ a Dynkin system contained in this σ-algebra. Let μ_σ be a countably additive probability on $\mathcal{D}_\sigma$ and Δ_σ the set of all countably additive probabilities on $\mathcal{F}_\sigma$. We call*

$$M_\sigma(\mu_\sigma, \mathcal{D}_\sigma) \coloneqq \{\nu \in \Delta_\sigma \colon \nu(A) = \mu_\sigma(A),\ \forall A \in \mathcal{D}_\sigma\},$$

the σ-credal set of μ_σ on $\mathcal{D}_\sigma$. *If $M_\sigma(\mu_\sigma, \mathcal{D}_\sigma) \neq \emptyset$, then it is closed and convex.*

Proof. **Convexity** Let $\{\nu_i\}_{i\in[n]} \subseteq M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)$ and $\alpha_i \in [0,1]$ for all $i \in [n]$ with $\sum_{i=1}^n \alpha_i$, then $\sum_{i=1}^n \alpha_i\nu_i \in M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)$, because

$$\sum_{i=1}^n \alpha_i\nu_i(A) = \sum_{i=1}^n \alpha_i\mu_\sigma(A) = \mu_\sigma(A), \qquad \forall A \in \mathcal{D}_\sigma.$$

Closedness We assumed $\mathrm{ca}(\Omega, \mathcal{F}_\sigma)$ to be equipped with the total variation distance. Hence, of a set $Q \subseteq \mathrm{ca}(\Omega, \mathcal{F}_\sigma)$ can be identified via the convergence of sequences in Q [Munkres, 2014] (Lemma 21.2).

Suppose $\nu_1, \nu_2 \ldots \in M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)$ is a sequence of probability measures such that

$\lim_{n\to\infty} \nu_n \overset{TV}{=} \nu$. This directly implies set-wise convergence,

$$\nu(A) = \lim_{n\to\infty} \nu_n(A) = \lim_{n\to\infty} \mu_\sigma(A) = \mu_\sigma(A), \qquad \forall A \in \mathcal{D}_\sigma.$$

It follows that $\nu \in M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)$.

□

In fact, one can even guarantee weak closure of the σ-credal set, since it is convex [Wojtaszczyk, 1991] (Theorem II.A.4). Rather obviously, the following corollary holds.

Corollary 8.78 (σ-Extendability and σ-Credal Set)**.** *Let $\mathcal{F}_\sigma$ be the Borel-σ-algebra on a Polish space Ω and $\mathcal{D}_\sigma$ a Dynkin system contained in this σ-algebra. The countably additive probability μ_σ on $\mathcal{D}_\sigma$ is σ-extendable if and only if $M_\sigma(\mu_\sigma, \mathcal{D}_\sigma) \neq \emptyset$.*

Thus, we can redefine the lower and upper coherent extension if the Dynkin probability space is continuously extendable. This extension is then derived from a set of countably additive probability measures.

Proposition 8.79 (Coherent σ-Extension of Probability)**.** *Let $\mathcal{F}_\sigma$ be the Borel-σ-algebra on a Polish space Ω and $\mathcal{D}_\sigma$ a Dynkin system contained in this σ-algebra. Assume the countably additive probability μ_σ on $\mathcal{D}_\sigma$ is σ-extendable. Then,*

$$\underline{\mu}_{\mathcal{D}_\sigma}(A) \coloneqq \inf_{\nu \in M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)} \nu(A), \quad \overline{\mu}_{\mathcal{D}_\sigma}(A) \coloneqq \sup_{\nu \in M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)} \nu(A), \qquad \forall A \in \mathcal{F}_\sigma,$$

define a coherent lower (respectively, upper) probability on $\mathcal{F}_\sigma$ in the sense of [Walley, 1991].

Proof. First, we notice the every countably additive probability is also finitely additive. Via Theorem 8.19 and application of Theorem 3.3.4 (b) in [Walley, 1991], we directly obtain the result. □

The coherent extension is one of two methods of measure extension presented in this paper. Again, we ask how the coherent σ-extension relates to the inner and outer measure construction.

Corollary 8.80 (Extension Theorem – Countably Additive Case)**.** *Let $\mathcal{F}_\sigma$ be the Borel-σ-algebra on a Polish space Ω and $\mathcal{D}_\sigma$ a Dynkin system contained in this σ-algebra. Suppose the countably additive probability μ_σ on $\mathcal{D}_\sigma$ is σ-extendable. Then,*

$$\mu_*(A) \leq \underline{\mu}_{\mathcal{D}_\sigma}(A) \leq \overline{\mu}_{\mathcal{D}_\sigma}(A) \leq \mu^*(A), \qquad \forall A \in \mathcal{F}_\sigma.$$

Proof. Since $M_\sigma(\mu_\sigma, \mathcal{D}_\sigma) \subseteq M(\mu_\sigma, \mathcal{D}_\sigma)$, clearly

$$\inf_{\nu \in M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)} \nu(A) \geq \inf_{\nu \in M(\mu_\sigma, \mathcal{D}_\sigma)} \nu(A), \qquad \forall A \in \mathcal{F}_\sigma,$$
$$\sup_{\nu \in M_\sigma(\mu_\sigma, \mathcal{D}_\sigma)} \nu(A) \leq \sup_{\nu \in M(\mu_\sigma, \mathcal{D}_\sigma)} \nu(A), \qquad \forall A \in \mathcal{F}_\sigma.$$

Hence, the analogous Theorem 8.22 gives the result. □

8.8.5 From Set Systems to Logical Structures and Back

In the year 1936, Stone demonstrated a celebrated representation result of logical algebras [Stone, 1936]. In particular, he showed that any Boolean algebra can be equivalently represented by an algebra of sets and vice versa. This representation results opened the door to many generalizations of the domain of probabilities (see [Narens, 2016] for a nice summary). For instance, Boolean algebras can be replaced by weaker logical structures, such as orthomodular lattices. Interestingly, modular ortholattices and orthomodular lattices have been investigated as representational structures for quantum theoretic descriptions and measurements [Birkhoff and von Neumann, 1936, Husimi, 1937]. Furthermore, it turned out that an analogue to Stone's representation holds for some orthomodular lattices. Some orthomodular lattices are isomorphic to pre-Dynkin systems [Godowski, 1981]. On the other hand, every (pre-)Dynkin system is isomorphic to an (σ-)orthomodular poset, a order-theoretic generalization of an orthomodular lattice (cf. [Brabec, 1979]). In summary, the investigation of probabilities on general logical structures parallels our work presented here. We, however, have stuck to set structures as the domain of probabilities.

Part III

Statistical Forecast Felicity: Individual to Aggregate

9

The Evaluation Resolution

Besides the frequential nature of von Mises theory of probability [Von Mises, 1919], his definition of randomness (see Chapter 7) provoked many discussions [Khinchin, 2015]. A major point of critique, the consistency of his theory, i.e., the existence of collectives which are random with respect to a certain set of selection rules was at least solved 18 years after the publication of [Von Mises, 1919] by Abraham Wald [Wald, 1937]. A further critique, however, showed the need for extending the tools of von Mises theory. Von Mises' notion of randomness was not able to capture the law of the iterated logarithm and furthermore could not rule out certain other phenomena which were considered to be non-random [Bienvenu et al., 2009].

Subsequently, Jean Ville suggested to depart from frequencies and added "betting against odds" to the theory to deal with this critique [Ville, 1939].[1] He moved away from selection rules and instead worked with betting strategies, respectively the development of a bettor's capital called *martingale*. He let bettors check whether they can win against a sequence to define whether the sequence can be called random or not. Jean Ville formalized in a different way what von Mises was after, a "law of excluded gambling system".

Ville's exposition proved fruitful in introducing *(super-)martingales* to the theory of probability and advanced the field of algorithmic randomness in its search to define randomness [Bienvenu et al., 2009]. However, it took more than half a century before his game-theoretic approach was taken more literally and triggered developments in probability theory [Shafer and Vovk, 2019] and statistics [Ramdas et al., 2023].

Conceptually, game-theoretic probability and statistics does not need to formulate assumptions on the origin or meaning of probabilistic assignments. Instead, this field puts focus on

[1] Interestingly, Ville originally seemed to aim for improvement von Mises theory. His thesis is, potentially due to the views of his supervisors, framed as a critique [Bienvenu et al., 2009].

the testing of those assignments in order to pragmatically justify them without a claim to any metaphysical content [Dawid, 2004, Shafer, 2024]. This strand of ideas motivated the definition and study of forecasts of Type I2A.

Type I2A forecasts resolve the interpretation problem (Section 5.1) by evaluation, cf., *evaluation resolution*. There is no meaning in the forecast beyond the pragmatic value. The forecast is usable for certain circumstances. The evaluation guarantees the usefulness of the forecast.

Related ideas occur in [Gillies, 2010] and [Höltgen, 2024]. Gillies [2010] suggests to consider falsification of probabilistic statements as the way to justify probabilities. Then, there is no a priori rationalization of probabilistic assignments required. Höltgen [2024] circumvents the interpretation problem of probabilities by asking what makes probabilistic forecasts on uncertain systems useful. In particular, this work embraces, as my thesis, that forecasting is one of the central use-cases of probabilities nowadays.

On a high-level we move from Richard von Mises in the last chapters to Jean Ville in the upcoming chapters. A key difference in the theories and hence in the type of forecasts is: the Type A2I forecasts and its related measurement resolution (Chapter 6) ignores the individuality of the subject of forecast in the following sense. The (idealized) measure of the aggregate which share the random attribute is equal for all individuals within the aggregate. This is different for Type I2A forecasts. The numerical value assigned to the individuals can differ within an aggregate. But, the assignment requires the evaluation on an aggregate (cf. Section 12.1).

The aggregate, which can be understood as a reference class (cf. Type I2A forecasts), serves a new purpose. Within an aggregate the individuals do not need to be exchangeable in them being related objects, e.g., "repeated trials of the same event" [Dawid, 1985a], instead within the aggregate the individuals need to be exchangeable in term of the performance of the forecasting method. The "goodness" of the forecasts on the aggregate is meaningful for an individual in the aggregate. The aggregate does not need to be substantially equal, but evaluation-wise equal.

The distinction made between the two types of reference class is analogous to the distinction between object- and meta-induction [Schurz, 2017]. Object-induction is what one can induce from some objects to others. While meta-induction is what method is best for inducing from objects to other objects. Exemplarily, with object-induction I transfer the average rain of the last days to the next day. With meta-induction I transfer the performance of a rain forecasting method on the last days to the next day.

Clearly, one central choice for felicitous Type I2A forecasts is the aggregate. Actually, this choice is very known and common. Scholars choose their benchmarks to promote their protein folding forecasts [Jumper et al., 2021]. Scholars prove that their weather forecasting

methods have been performing well in the past [Bureau of Meteorology (Australia), 2025]. Scholars leverage adversarial example to show whether their image forecasting method is robust [Szegedy et al., 2013, Goodfellow et al., 2014].

In this work we leave out the structure of the choice of aggregate, e.g., the design of benchmarks [v. Kistowski et al., 2015, Reuel-Lamparth et al., 2024, Hardt, 2025], or its relationship to inductive principles. However, another central choice, the evaluation metric, will be the topic for the next chapters.

10

Three Types of Calibration with Properties and their Semantic and Formal Relationships

Author Contributions

The concept and results of the paper have been developed in mutual exchange mainly between Jessica Finocchiaro (JF) and Rabanus Derr (RD). At a turning point of the project when we decided to change directions towards the study of the "elements of calibration", Robert C. Williamson (RW) suggested leading questions moving forward. The exact attribution of contributions for the writing are hard to disentangle. JF was closely supervising the writing process and suggested changes on different levels, up to writing entire paragraphs. RW took over the internal editing. RD wrote most parts of the text.

The Personal Note

It is hard to define when the story of this paper should begin.

Was it when Benedikt Höltgen joined our research group? He brought the interest in calibration to the group. By "calibration" I mean the statistical connotation of the word, the property of a forecast to reflect the actual statistics of an aggregate of outcomes.

Was it when I moved to Philadelphia for a five months long research stay at the University of Pennsylvania. Several of the students in the group of Aaron Roth and Michael Kearns, who I visited, studied in great detail how to achieve calibrated forecasts in online prediction scenarios.

Was it when I visited James Bailie, a friend of mine whom I got to know during the ISIPTA conference 2023, in Cambridge, Massachusetts. Asking Bob whom to visit in the Boston area, he referred me to Jessica Finocchiaro. Jessie was a PostDoc at Harvard University that time. She very kindly invited me to give a talk entitled "4 Facets of Evaluating Predictions". That is were the title of the thesis originates. The content of the talk, however, was covering parts of the project presented in Chapter 11.

Jessie and I vibed, we thought about doing some project together. Our first ideas, e.g., elicitation of decision maker's preferences to provide useful schemes to evaluate forecasts, largely didn't survive. I was back in Tübingen again, and Bob joined our endeavor. We converged to sharing confusion and interest in what this mysterious "multi-calibration", "omniprediction", "swap-multicalibration", "agnostic learning", and what not, actually meant. The list of newly related notions circulating around calibration kept growing month by month. We were convinced that it should be possible to clean up the mess of definitions a bit by generalizing and using the tools of property elicitation. This was a natural choice as Jessie has academically grown up with them. Plus, those tools were already used in the "4 Facets of Evaluating Predictions"-project.

It turned out that a clean generalization was, even after more than 100 assumptions of smoothness, not in our reach. Omniprediction and its cousins didn't want to be treated in a framework which generalized the type of forecasts.

We hopped from narrative to narrative in order to collect as many of our thoughts and insights about calibration under one theme. One can now find a big chunk, but not all what we encountered, in "Three Types of Calibration with Properties and their Semantic and Formal Relationships". The manuscript summarizes the academic side of the story whose beginning is yet to be defined.

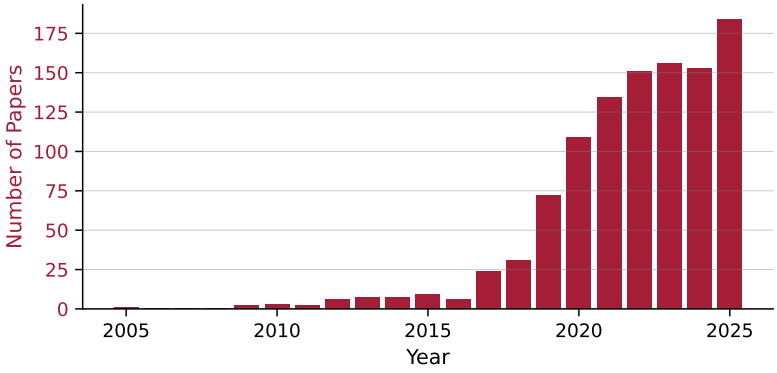


Figure 10.1: Number of query results, each being a paper, for the title search term "calibration" in the categories cs.LG (Computer Science – Machine Learning) and stat.ML (Statistics – Machine Learning) per year on arXiv. Since the API of arXiv provided different query results for different batch sizes (51,67, 103, 121), I have chosen the maximum among those numbers.

Summary: Three Types of Calibration

Motivation In the 1950s, 60s and 70s, a branch of metereology and statistics aimed for finding mathematical functions which captured adequacy when comparing forecasts and actual realized weather events [Brier, 1950, Murphy and Epstein, 1967, Murphy and Winkler, 1977]. One of the suggested evaluation criteria read as follows, if on the days on which the forecaster predicted "Chance of rain: $p\%$", it actually rained $p\%$ of them, for all $p \in [0, 100]$, the forecast is "calibrated" or "reliable" [Murphy and Epstein, 1967]. Calibration made its way through statistics [Scheffe, 1973, Dawid, 1984, 1985a, Gneiting and Katzfuss, 2014, Gneiting and Resin, 2023] and during the last decade obtained an increasing interest in the machine learning community (Figure 10.1)[1]. First, calibration is motivated as it arguably guarantees "equal meaning across groups" when enforced on them [Kleinberg et al., 2017, Hu, 2025]. Second, calibration promises usefulness of forecasts without the need to specify the down-stream decision maker [Foster and Vohra, 1997, 1999, Noarov and Roth, 2024, Roth and Shi, 2024].

For these motivations, scholarship has come up with dozens of differing definitions of calibration departing from the original definition as given above (cf. Chapter 10). Their formal relationships are part of ongoing investigations [Zhao et al., 2021, Błasiok et al., 2023, Garg et al., 2024]. Furthermore, the aims of the definitions partially differ. In this way, there is a more critical, but less prominent question beyond the formal relationships: what are the semantic relationships of notions of calibration?

[1]The code for the retrieval has been produced with the help of ChatGPT.

Contributions We – the work has been pursued together with Jessica Finocchiaro and Robert C. Williamson – suggest to focus on two accounts of calibration which we name *self-realization* and *precise loss estimation*. Most existing notions of calibration can be understood through one or both of these accounts.

Forecasts which are calibrated in a *self-realizing* way fulfill the following condition: on the instances to which the forecasts assigned a certain summary statistic, the summary statistic of those instances is equal to the forecasted one. Hence, the forecasts realize themselves in a certain way. This self-referential aspect make this account attractive. For instance, it is used to argue that a decision maker who consistently responds to a forecast will not regret the choice of action [Foster and Vohra, 1999, Noarov et al., 2023, Noarov and Roth, 2024, Roth and Shi, 2024]. Or, if the forecasts inform several agents, *self-realization* guarantees that the agents will behave in equilibrium, they expect to see a pattern of actions which will actually materialize [Foster and Vohra, 1997, Kakade and Foster, 2008, Foster and Hart, 2018, 2021].

The account of *precise loss estimation* adds another aspect to what makes a forecast useful. A useful forecast does not only inform which action to take, but as well provides an estimation of its consequences. In other words, a calibrated forecast following the *precise loss estimation*-account provides an estimate of the consequence when the decision maker takes the best action when informed with a forecast. The exact formulation requires the formalization of a decision maker as a utility maximizing agent. The consequences are the expected utilities.

Using the abstraction of properties, i.e., functions from probability distributions to value sets such as the mean or the median, we show that forecasts which are full probability distributions and fulfill an appropriate notion of calibration can serve both the purpose of self-realization and of precise loss estimation. However, we identify both accounts as distinct and independently reachable ideals. It is possible to provide predictions which are self-realizing but not precise loss estimators and vice-versa. Hence, we provide a tree structure of semantic and formal relationships which we prove to hold not only in the exact but as well in the approximate case under additional smoothness assumptions. Finally, we show that omniprediction, a newly occurring learning paradigm [Gopalan et al., 2022b,a, Globus-Harris et al., 2023, Deng et al., 2023, Gopalan et al., 2024b, Garg et al., 2024], exploits a more generally usable characteristic of self-realization calibration. Being calibrated is a necessary condition for the optimal minimization of a multitude of loss functions, as long as the loss functions share their ideal minimizer, i.e., elicit the same property.

Tools and Sketch of Argument The starting point for our discussion will be the vanilla form of calibration. To avoid confusion, the reader is asked to fully detach from the

interpretation of the forecast as being a measurement here. In particular, the semantics given to a probability space in Section 7 do *not* fit the semantics of the underlying probability space as used here. In fact, I will adopt the often used, but sloppy, way of defining random variables without explicitly writing out the abstract probability space underneath the definitions. For a translation of terms and symbols see Table 10.1. Let X be a random

Unified Notation & Terminology	Chapter 10
statistical forecast	prediction, $f(x) \in \mathbb{R}$
individual, ι	$x \in \mathcal{X}$
aggregate, $\mho$	input set, $\mathcal{X}$
attribute, $\varrho(\iota) = \wp$	outcome, label, $y \in \mathcal{Y}$
measurement	–

Table 10.1: Translation table – Three Types of Calibration with Properties and their Semantic and Formal Relationships

variable on $\mathcal{X} \subseteq \mathbb{R}$ and Y a random variable on $\mathcal{Y} = \{0, 1\}$ distributed jointly according to D. Let $f \colon \mathcal{X} \to [0, 1]$ be a measurable function. It represents a forecast of the probability $D_{Y|X=x}(Y = 1) = \mathbb{E}[Y|X = x]$. The forecaster f is *vanilla calibrated* on the distribution D if and only if, for all $\gamma \in \mathrm{im}f$[2],

$$\mathbb{E}_{(X,Y)\sim D}[Y|f(X) = \gamma] = \gamma.$$

Let's give this formula rain-forecasting semantics. The random variable Y is stating "no rain" $Y = 0$ or "rain" $Y = 1$. While X is a time point, an individual, e.g., a day. Hence $f(X)$ is the probability of rain on day X. If the actual probability of rain $\mathbb{E}_{(X,Y)\sim D}[Y|f(X) = \gamma] = D_{Y|f(X)=\gamma}(Y = 1)$ on the days which share the forecast γ, is γ, then the rain forecaster is calibrated.

As it turns out, it is not necessarily relevant that the forecaster is calibrated on all of its forecasted values. Let's imagine a decision maker who is about to carry or not carry the umbrella through the day. The decision maker is described by the loss Table 10.2. Assuming that the decision maker is an expected utility maximizer, the decision maker will take the umbrella as long as the rain forecasts predicts a probability higher than 0.5. Formally,

[2]I denote the image of a function f as $\mathrm{im}f$.

DM with ℓ	R	$\neg R$
U	1	1
$\neg U$	2	0

Table 10.2: Loss (ℓ) of decision maker (DM) observing rain (R) or no rain ($\neg R$), and having the actions carrying the umbrella (U) or not carrying the umbrella ($\neg U$).

$\Gamma\colon [0,1] \to \{U, \neg U\}$ is the decision function such that $\Gamma(0.6) = U$, $\Gamma(0.5) = \neg U$.

Hence, a desired guarantee called *distribution calibration with respect to* Γ is: on the days on which the decision maker (not) took its umbrella, the actual average rainfall was equal to the average forecast on those days. Formally,

$$\mathbb{E}_{(X,Y)\sim D}[Y|\Gamma(f(X)) = A] = \mathbb{E}_{(X,Y)\sim D}[f(X)|\Gamma(f(X)) = A], \quad A \in \{U, \neg U\}.$$

It is a rather immediate consequence that vanilla calibration implies *distribution calibration with respect to* Γ (Proposition 10.4).

Another appropriate formulation called Γ-*calibration* is: if the decision maker would have known the average rain situation on the days on which the decision maker has (not) taken the umbrella, they did (not) take the umbrella anyway. Formally,

$$\Gamma\left(\mathbb{E}_{(X,Y)\sim D}[Y|\Gamma(f(X)) = A]\right) = A, \quad A \in \{U, \neg, U\}.$$

Γ-*calibration* formalizes *self-realization*. The decision taken by the decision maker are consistent in the following sense. On the days when the decision maker took the umbrella with them, their loss increased if they did not take the umbrella with them, and vice versa. This is called *no swap-regret* (cf. Proposition 10.7 and Proposition 10.28). The former *distribution calibration with respect to* Γ implies the later Γ-*calibration* (Proposition 10.6 and Proposition 10.25). Notably, it is possible to provide Γ-calibrated forecasts which are not distribution calibrated, e.g., when the forecaster simply recommend the best action to take without giving any probabilistic information.

Another reasonable position is that the forecasts should fulfill,

$$\mathbb{E}_{(X,Y)\sim D}[\ell(Y, \Gamma(f(X)))] = \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y}, \Gamma_\ell(f(X))].$$

In words, as long as the decision maker follows the decision function Γ, the expected loss inferred from the forecaster f is equal to the actual realized average loss. To remind the reader, Γ is the the best response by assumption of expected utility maximization. This criterion is rather weak, but gives at least an average estimate of the consequences of the best response action. We call it *decision calibration* borrowed from [Zhao et al., 2021]. *Decision calibration* formalizes *precise loss estimation*. Again, decision calibration is implied by distribution calibration (Proposition 10.12 and Proposition 10.34). However, the reverse direction and the relationship to Γ-calibration is more complicate (Proposition 10.11 and Section 10.6.3).

In the full paper (Chapter 10) we extend the statements and definitions to arbitrary functions Γ, then called *properties*, and the approximate case. In contrast to the previous

projects, felicity conditions and different notions of calibration are immediately related.

Relation to Forecast Felicity The evaluation problem (Section 5.2) asks for a felicitous choice of evaluation of forecasts. Our provided semantic segmentation of calibration and the implication structure sketched before (depicted in Figure 10.2 and Figure 10.7) helps to structure the choice.

For instance, a *vanilla calibrated* probabilistic rain forecast for today, such as "Chance of rain today: 60%", would (a) guarantee decision makers to not regret their action instead of swapping to a different one, and (b) let the decision makers estimate consistently their expected loss when responding to the forecasts as if they were true.

However, imagine instead of "Chance of rain today: 60%" a simple umbrella symbol corresponding to a 0 or 1. This forecast would suggest an action for the reader of the newspaper. This forecast can be *self-realizing*. It can serve the aim and purpose it is made for.

On the other hand, a farmer might not be happy about the replacement, as she would rather try to estimate her actual loss in Euros if the mowed grass starts to rot due to the rain. An agricultural forecast requires different semantics of the notion of calibration.

Note that in these scenarios the choice of the forecast interferes with the choice of evaluation. It turns out that the choice of forecast is part of felicity considerations. The farmer *needs* a probability distribution instead of a mere umbrella symbol to pursue calculations.

Abstract

Fueled by discussions around "trustworthiness" and algorithmic fairness, calibration of predictive systems has regained scholars' attention. The vanilla definition and understanding of calibration is, simply put, on all days on which the rain probability has been predicted to be p, the actual frequency of rain days was p. However, the increased attention has led to an immense variety of new notions of "calibration." Some of the notions are incomparable, serve different purposes, or imply each other. In this work, we provide two accounts which motivate calibration: self-realization of forecasted properties and precise estimation of incurred losses of the decision makers relying on forecasts. We substantiate the former via the reflection principle and the latter by actuarial fairness. For both accounts we formulate prototypical definitions via properties Γ of outcome distributions, e.g., the mean or median. The prototypical definition for self-realization, which we call Γ-*calibration*, is equivalent to a certain type of swap regret under certain conditions. These implications are strongly connected to the omniprediction learning paradigm. The prototypical definition for precise loss estimation is a modification of *decision calibration* adopted from Zhao et al. [2021]. For binary outcome sets both prototypical definitions coincide under appropriate choices of reference properties. For higher-dimensional outcome sets, both prototypical definitions can be subsumed by a natural extension of the binary definition, called *distribution calibration* with respect to a property. We conclude by commenting on the role of groupings in both accounts of calibration often used to obtain multicalibration. In sum, this work provides a semantic map of calibration in order to navigate a fragmented terrain of notions and definitions.

10.1 Introduction

Calibration has increasingly gained interest since it seems to provide a mathematical criterion of "trustworthy"[3] predictions [Noarov and Roth, 2024] and it is a major component of studies on algorithmic fairness [Chouldechova, 2017, Hébert-Johnson et al., 2018]. Furthermore, the advent of capable, deep learning techniques gave rise to investigations of calibration of general deep neural networks [Guo et al., 2017] and large language models [Cruz et al., 2024, OpenAI (2023), 2024, Kalai and Vempala, 2024].

While calibration has become a quantity of interest in empirical studies [OpenAI (2023), 2024, Perdomo et al., 2025, Cruz et al., 2024], theoretical works came up with dozens of new (and old) definitions of calibration investigating particular relationships between them. To name a few: Γ-calibration [Noarov and Roth, 2023], decision calibration [Zhao et al., 2021], U-calibration [Kleinberg et al., 2023], class-wise calibration [Kull et al., 2019], confidence calibration [Guo et al., 2017], Global Interpretable Calibration Index [Cabitza et al., 2022], distance to calibration [Błasiok et al., 2023], smooth calibration [Foster and Hart, 2018]. The "jungle" of current notions of calibration is difficult to navigate.

As a result, wide-spread use of the term "calibration" has blurred the boundaries of what is meant by it. However, all usages have in common that calibration is either a criterion to judge predictions in light of obtained data or the process of fulfilling such a criterion. In this work we stick to the former (cf. [Höltgen and Williamson, 2023]). Our goal is *not* to provide an exhaustive list of notions of calibration.[4] Instead, we follow our main question: **What is the abstract purpose of calibration?** This way we provide a semantic map and guide through key notions of calibration and their central relationships.

Calibration is often contrasted with pure predictive accuracy [Cruz et al., 2024, Van Calster et al., 2019, Seidenfeld, 1985]. While predictive accuracy guarantees the "usefulness" of predictions, calibration guarantees the "trustworthiness" of the predictions. The "usefulness" narrative is readily justified. Expected risk minimization, the core principle behind a large portion of learning techniques, is the negative analogue of expected utility maximization. The lower the risk, the more useful the predictions are, when measured with the corresponding loss (respectively negative utility) function.

The "trustworthiness"-narrative, however, requires a more detailed explanation. We identified two accounts of calibration which could justify it:

[3] We hesitate to attempt a concise definition of trustworthiness in this work for the ambiguity and vagueness of its purpose and meaning. Instead, we write "trustworthiness" to emphasize to the reader that the term itself is and should be loaded with more than the formal intuitions for some facets of trustworthiness presented in this work.

[4] After trying this for some time, we gave up on this journey due to the immense variety and subtleties of the suggested variants of calibration.

Self-realization: For the instances where some value c was forecasted, the actual outcomes can be summarized to a value close to c.

Precise Loss Estimation: The forecasted values let one provide estimates of incurred losses (for certain loss functions) which are close to the actual materialized losses.

While these accounts are not an exhaustive list, these two paradigms account for the majority of calibration motivations. We found a third account of calibration in the literature which focuses on the approximate equivalence of means. Here, predictions are calibrated if the average of outcomes is equal to the average of predictions on certain subgroups or even individuals [Dawid, 1985a, Zhao et al., 2020, Luo et al., 2022, Höltgen and Williamson, 2023]. This account is somewhat located between the narratives of "usefulness" and "trustworthiness." Note that all accounts extend on a certain facet of the binary vanilla calibration definition (Definition 10.1) which has historically motivated previous definitions.

To uncover the two central accounts of this work, we rewrite current definitions of calibration using the language of properties [Osband, 1985, Fissler and Ziegel, 2016, Lambert et al., 2008]. Properties, simply put, are functions from distributions on the outcome set $\mathcal{Y}$ to some value set $\mathcal{R}$, e.g., $\mathcal{R} = \mathbb{R}$ or $\mathcal{R} = \Delta(\mathcal{Y})$. We distinguish between *optimization-level* and *decision-level* properties. Optimization-level properties specify the actual predicted entity, e.g., full distribution, mean, α-quantile, or best action. Decision-level properties most often correspond to downstream uses of the optimization-level property, such as decision makers informing their discrete action via a prediction about outcome probabilities, or ranking of classes being used to deduce the top-k most likely outcomes. Decision-level properties borrow their name since they generalize maximum expected utility decision makers. Notably, decisions are often discrete, and therefore hard to optimize directly, necessitating the distinction between (often continuous) optimization-level properties and (often discrete) decision-level properties. Properties subsume not only utility-based decision makers, but rather arbitrary statistical properties of distributions as well, such as quantiles[5], ratios of expectations, or class marginals, to name a few. Furthermore, the elicitability and identifiability of properties as detailed in Section 10.3 relate properties to loss functions and their optimization criterion. This makes properties a useful vehicle to study calibration in the abstract.

Our contributions are summarized as follows:

1. We subsume many existing definitions under three core *types* of definitions of calibration: distribution calibration with respect to a (decision-level) property Φ (Definition 10.2), property calibration with respect to an abstract property Γ (Definition 10.5) and decision calibration with respect to a set of loss functions $\mathcal{L}$ (Definition 10.9).

[5]Calibrated quantiles are relevant for conformal predictions [Jung et al., 2023].

2. We show that all types of definitions of calibration collapse in the case of binary outcome sets and appropriate choices of Φ, Γ and $\mathcal{L}$ (Proposition 10.11). More generally, we argue that distribution calibration is the central "parent" notion which implies both of the others (Proposition 10.6 and Proposition 10.12), as summarized in Figure 10.2. When generalizing to the approximate case, see Propositions 10.25 and 10.34 and Figure 10.7 in the appendix.

3. We elaborate on property calibration as a prototypical definition for the account of self-realization. We show it can equivalently be expressed as a type of swap regret (Proposition 10.7 and Proposition 10.28). Furthermore, we prove that self-realization is inherited by *refined* properties— those properties Φ that can be defined by composing a function with some original property Γ (Proposition 10.8 and Proposition 10.31), a characteristic of the notion which is exploited in omniprediction (cf. Proposition 10.33).

4. We contextualize decision calibration as a prototypical notion of precise loss estimation. In particular, we show it requires simultaneous precise Bayes risk estimation for several loss functions of interest (Proposition 10.15). Furthermore, we dissect the relationship between self-realization and precise loss estimation, concluding that these accounts are incommensurable (Section 10.6.3).

5. We provide a non-exhaustive categorization of existing notions of calibration in the three types in Table 10.3 and Table 10.4.

The distinction of the three types of definitions of calibration and their relationship to the presented accounts of self-realization and precise loss estimation has, to the best of our knowledge, not been made in the literature. Properties have been used to formalize calibration in [Gneiting and Resin, 2023] and [Noarov and Roth, 2023]. Some of the formal relationships have been discussed already in literature (e.g., distribution calibration implies decision calibration has been argued for in [Zhao et al., 2021]), but not within the more general picture of properties.

In this work, we pay attention on how to define calibration beyond binary outcomes. We largely ignore another strand of work on calibration notions which focus on how to approximate (binary) Vanilla calibration. Different choices (e.g., "distance to calibration" [Błasiok et al., 2023], "calibration decision loss" [Hu and Wu, 2024], "cutoff calibration error" [Rossellini et al., 2025], "smooth calibration" [Foster and Hart, 2018], "projected smooth calibration" [Gopalan et al., 2024a]) yield different decision-theoretic and empirical approximability guarantees. Rossellini et al. [2025], Haghtalab et al. [2024] give a more in-depth, axiomatic discussion about calibration error metrics. Our treatment is largely agnostic to the choice of metrics to capture the approximation, though the relationship between prop-

erties and the existence of calibration metrics satisfying normative axioms is an interesting line of future work.

Figure 10.2 graphically summarizes the two strands which we will follow in this work. The figure depicts the idealized "perfect" calibration case. For approximate notions the implications require additional assumptions (Figure 10.7). Clearly, we don't expect the reader to yet make sense of the definitions and implications.

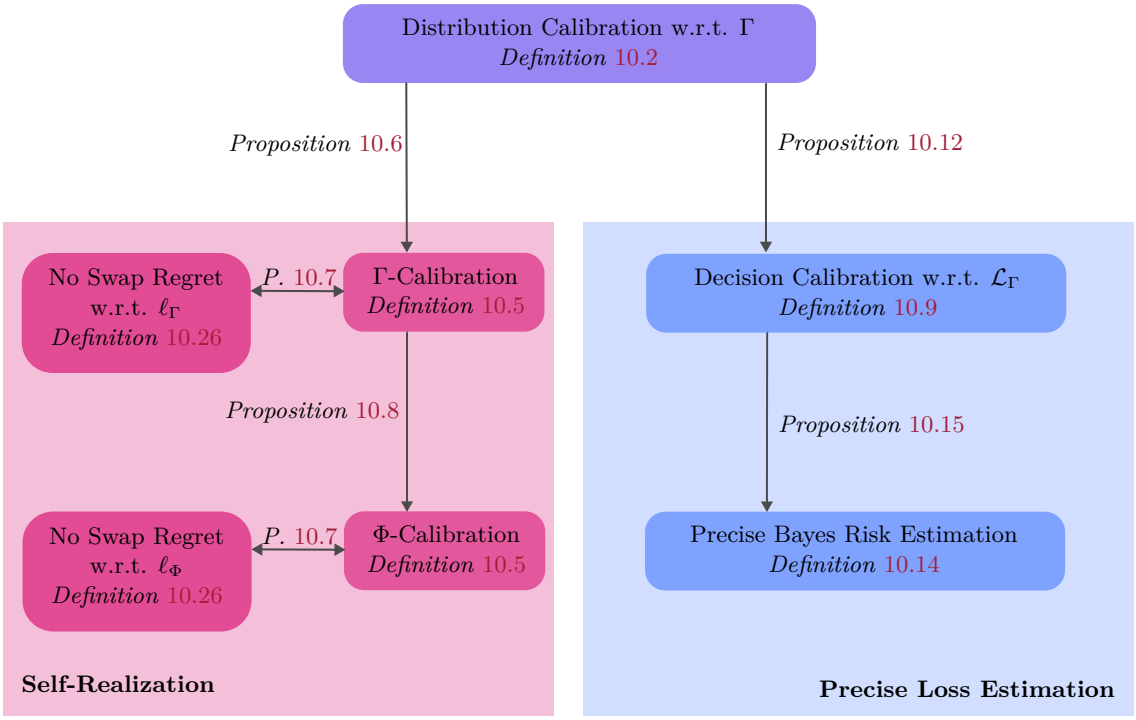


Figure 10.2: Relationships between Notions of Calibration. Implications under perfect calibration, finite $\mathcal{Y}$ and elicitable property Γ and Φ. The three types of calibration are marked in different colors. The abstract accounts of calibration are shaded.

10.2 An Exemplary Motivation

Let us walk through an example to get an intuitive understanding of the subtleties of different formalizations of calibration.

Perdomo et al. [2025] study the Dropout Early Warning System (DEWS) implemented at Wisconsin Public Schools. Based on state level testing in the 8th grade, the system predicted on-time high school graduation likelihood. Hence, the DEWS predictions themselves, i.e., the *optimization-level property*, are probability forecasts on a binary outcome set. As the authors observe, the forecasts are not perfectly calibrated following *binary vanilla calibration* (Definition 10.1), i.e., the graduation rate of students who get a certain prediction p is not

p.

The predictions of the DEWS are given to the department authorities who define three risk categories "high risk," "medium risk," and "low risk" for dropout. The risk category determines the measures to be taken to increase the graduation likelihood. In other words, the risk categories form the *decision-level property*. Are the *predictions calibrated conditioned on the risk categories*, i.e., is the average graduation rate among all persons who are categorized as "high risk" (or "medium risk" or "low risk") equal to the average prediction score of those persons? This property, which we call *distribution calibration with respect to the risk categories* (Definition 10.2), is not fulfilled as immediately derived from [Perdomo et al., 2025, Figures 1 and 2].

But, the average graduation rate among all persons who are categorized as "high risk" is lower than the average graduation rate in all other categories (for the other categories respectively). In other words, the "high risk" category self-realizes. Hence, the predictions are *property calibrated* (Definition 10.5).

Finally, self-realization of the decision-level property is not necessarily the intention of calibration. Instead one can ask whether the department authorities can estimate their expected loss in mis-assigning students to the wrong group. For this one would need to introduce a loss function, whose minimizer is the risk category assignment. Note that several such loss functions exist. Depending on the choice of the loss function, the predictions are precisely estimating the loss, i.e., the predictions are *decision calibrated* (Definition 10.9), or not.

Exactly, how calibrated the DEWS predictions are is irrelevant for the lessons we want to convey in this section. Some of the notions of calibration are fulfilled to a lesser degree than others. We use this example to demonstrate that there is no "right" definition of calibration. This leaves a choice open which type and definition of calibration to consider in a contextualized problem. This choice has semantics and implications which we disentangle in the following.

10.3 Formal Setup

Data Distribution Let $(\Omega, \Sigma, \lambda)$ be a standard probability space. Let $\mathcal{X}$ (and $\mathcal{Y}$) be an input set (and outcome set) respectively. For the sake of simplicity, we assume that those sets are finite dimensional Euclidean spaces or finite sets, equipped with the standard Borel-σ-algebra $\mathcal{B}(\mathcal{X})$ (respectively $\mathcal{B}(\mathcal{Y})$). We define two random variables $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \mathcal{Y}$. The distribution induced on $\mathcal{X} \times \mathcal{Y}$ by the joint random variable (X, Y) is denoted D, and called the *data distribution*. The set of all data distributions is denoted $\Delta(\mathcal{X} \times \mathcal{Y})$. For the expected value of a measurable function $g\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ we write

$\mathbb{E}_{(X,Y)\sim D}[g(X,Y)]$ or simply $\mathbb{E}_D[g(X,Y)]$.

The marginals are denoted D_X (respectively D_Y). We write push-forward measures such as for a measurable map $f\colon \mathcal{X} \to \mathbb{R}$ as $D_{f(X)}$. Conditional distributions, e.g., conditional probability of Y given X are written as $D_{Y|X\in A}$, where $A \in \mathcal{B}\mathcal{X}$ and $D_X(A) > 0$. For the corresponding conditional expectation for a measurable function $g\colon \mathcal{Y} \to \mathbb{R}$ we write,

$$\mathbb{E}_D\left[g(Y)|X \in A\right] \coloneqq \mathbb{E}_{Y\sim D_{Y|X\in A}}\left[g(Y)\right].$$

We pay attention to not conditioning on measure 0 events. For that reason, we refer to the support of a measure, $\operatorname{supp} D_{f(X)}$, which is the set $\{r \in \mathbb{R}\colon D_{f(X)}(\{r\}) > 0\}$. By $\mathcal{P}$ we mean a subset of $\Delta(\mathcal{Y})$ the set of all probability distributions on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. We call a data distribution D *regular with respect to* $\mathcal{P}$ if and only if for every $A \in \mathcal{B}(\mathcal{X})$ it is true that $D_{Y|X\in A} \in \mathcal{P}$. Finally, we denote the Iverson-brackets as $[\![\cdot]\!]$, i.e., $[\![A]\!] = 1$ if A is true, $[\![A]\!] = 0$ otherwise.

PROPERTIES Let $\mathcal{R}$ be a set of property values equipped with a metric m; i.e., $(\mathcal{R}, m)$ forms a metric space. Generally, we see three common choices of value sets: (i) if $\mathcal{R} \subseteq \mathbb{R}$ is an interval, we set m to be the absolute difference, (ii) if $|\mathcal{R}| < \infty$, we set m to be the discrete metric. Finally, (iii) if $\mathcal{R} = \Delta(\mathcal{Y})$, we set m to be the total variation distance. Furthermore, $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$ is a measurable space for $\mathcal{B}(\mathcal{R})$ being the Borel-σ-algebra induced by the topology given via m. A measurable function $\Gamma\colon \mathcal{P} \to \mathcal{R}$ is called a *property*.[6] Without loss of generality we assume that Γ is surjective. To distinguish between optimization-level and decision-level properties we sometimes use Γ and Φ respectively. In Section 10.6, we additionally use Θ to denote Bayes risk properties (defined later, Definition 10.13).

A property is *continuous* iff Γ is continuous with respect to the total variation distance on $\mathcal{P}$ and the metric m on $\mathcal{R}$. In particular, a property is *Lipschitz continuous* iff Γ is Lipschitz continuous with respect to the total variation distance on $\mathcal{P}$ and the metric m on $\mathcal{R}$.

A property $\Gamma\colon \mathcal{P} \to \mathcal{R}$ is called *elicitable* if there exists a loss function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ which is measurable in the first variable for all fixed $\gamma \in \mathcal{R}$ in the second variable and such that,

$$\Gamma(P) \in \arg\min_{\gamma\in\mathcal{R}} \mathbb{E}_{Y\sim P}[\ell(Y,\gamma)],$$

for all $P \in \mathcal{P}$. In the case ℓ elicits Γ, we say the loss ℓ is $\mathcal{P}$-consistent with respect to Γ. Overloading notation, we sometimes write $\ell(P,\gamma) = \mathbb{E}_{Y\sim P}[\ell(Y,\gamma)]$. An elicitable property can be understood as a "best response" for a minimum expected loss decision maker.

A property $\Gamma\colon \mathcal{P} \to \mathcal{R}$ is called *identifiable* if there exists an identification function $V\colon \mathcal{Y} \times$

[6] Note that measurability can be guaranteed by inducing a σ-algebra on $\mathcal{P}$ or simply referring to the Borel-σ-algebra on $\Delta(\mathcal{Y})$ induced by the total variation distance, in particular if $\mathcal{Y}$ is finite.

$\mathcal{R} \to \mathbb{R}$ which is measurable in both variables, and

$$\Gamma(P) = \gamma \Leftrightarrow \mathbb{E}_{Y\sim P}[V(Y,\gamma)] = 0,$$

for all $P \in \mathcal{P}$. We overload notation and write $V(P,\gamma) = \mathbb{E}_{Y\sim P}[V(Y,\gamma)]$.

A measurable function $f\colon \mathcal{X} \to \mathcal{R}$ is called a Γ-*predictor* for the property $\Gamma : \mathcal{P} \to \mathcal{R}$. Properties, as well as property predictors can be simply stacked to vectors, e.g., the distribution predictor $f\colon \mathcal{X} \to \Delta(\mathcal{Y})$ is a $(\Gamma_y)_{y\in\mathcal{Y}}$-predictor, where Γ_y denotes the property: probability of outcome $y \in \mathcal{Y}$.

10.4 Distribution Calibration

Calibration has been well-studied long before its evolution as a "trustworthiness" criterion or "fairness" criterion (as multi-calibration) in machine learning [Murphy and Epstein, 1967, Murphy and Winkler, 1977, Dawid, 1982, DeGroot and Fienberg, 1983].[7][8] Calibration was considered part of a canon of "goodness"-measures of (meteorological) forecasts [Murphy and Epstein, 1967, DeGroot and Fienberg, 1983]. In particular, calibration captured the following intuition: if it rains a p-proportion of times on the days on which the prediction is p for rain, the predictions are calibrated. Definition 10.1 makes this formal.

Definition 10.1 (Binary Vanilla Calibration)**.** *Let $\mathcal{X} \subseteq \mathbb{R}$ be an input set and $\mathcal{Y} = \{0,1\}$ a binary outcome set. A predictor $f\colon \mathcal{X} \to [0,1]$ predicting the mean, i.e., the probability $D_{Y|X=x}(Y=1) = \mathbb{E}[Y|X=x]$, is* calibrated *on a distribution $D \in \Delta(\mathcal{X}\times\mathcal{Y})$ if and only if, for all $\gamma \in \operatorname{supp} D_{f(X)}$*[9]*,*

$$\mathbb{E}_{(X,Y)\sim D}[Y|f(X) = \gamma] = \gamma.$$

Consider $f\colon \mathcal{X} \to \mathcal{P}$ to be a full distribution forecast for a finite $\mathcal{Y}$. It is not always relevant, nor attainable to provide full distribution estimates which are calibrated conditioned on the distributional predictions. In particular, in high-dimensional class probability prediction problems it is not realistic to achieve reasonable calibration guarantees [Roth and Shi, 2024, Noarov and Roth, 2024]. The number of events to condition on grows exponential in the dimensionality of the outcome set $\mathcal{Y}$. Hence, scholarship suggested conditioning on "decision-events" [Noarov et al., 2023, Roth and Shi, 2024], marginals [Kull et al., 2019],

[7] In some of the older literature the terms "reliability" and "validity" were (inconsistently) used instead of "calibration".

[8] Murphy and Winkler [1977] provide the first calibration plots, objective frequency versus prediction.

[9] We refer to the support of the push-forward measure to exclude that we condition on measure zero sets. In other words, our calibration conditions focus only on substantially often predicted values. That has the side-effect that a predictor which predicts each value only on measure zero sets is directly calibrated. Note that this is a theoretical artifact which does not play a role for any empirical distribution D.

or top-labels [Guo et al., 2017, Gupta and Ramdas, 2021]. In other words, the number of conditioning events is reduced by focusing only on the "relevant" ones.[10]

We subsume all those notions via a specified property. For the sake of readable notation we use Γ as the symbol for the property here, even though it can refer to a decision-level or optimization-level property. For instance, Γ could be the map which maps all distributions to one of their marginals (cf. [Kull et al., 2019]) or best responses with respect to some utility function (cf. [Noarov et al., 2023]).

Definition 10.2 (Distribution Calibration with Respect to Γ)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$, where $\mathcal{Y}$ is finite. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. The predictor f is* distribution calibrated with respect to Γ *on D if for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,*[11]

$$\mathbb{E}_D[[\![Y = y]\!] | \Gamma \circ f(X) = \gamma] = \mathbb{E}_D[f_y(X) | \Gamma \circ f(X) = \gamma], \quad \forall y \in \mathcal{Y},$$

where $f_y(x) \in [0,1]$ denotes the y-component of the prediction $f(x) \in \Delta(\mathcal{Y})$ for $x \in \mathcal{X}$. The predictor f is $\alpha(\gamma)$-approximate distribution calibrated with respect to Γ *on D if for every $\gamma \in \operatorname{im}\Gamma \circ f$, $\alpha(\gamma) \in \mathbb{R}$,*

$$|\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \Gamma \circ f(X) = \gamma]| \leq \alpha(\gamma), \quad \forall y \in \mathcal{Y}.$$

We illustrate Definition 10.2 in Figure 10.3, where we demonstrate exactly (left) and approximately (right) satisfying distribution calibration with respect to the property $\Gamma(p) = \arg\max_y p_y$ representing the mode.

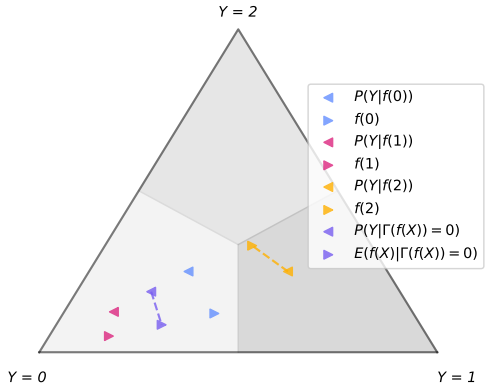
Note that, in principle, $\mathcal{Y}$ does not need to be finite. However, it simplifies notation and facilitates understanding. We provide definitions and results for general $\mathcal{Y}$ in Appendix 10.9.1. Note that if $\mathcal{Y}$ is infinite, most predictions actually directly refer to properties and not the full distribution (e.g., in Gaussian Process Regression, the mean and covariance are estimated, which are properties of the full distribution).

10.4.1 Distribution Calibration is Inherited

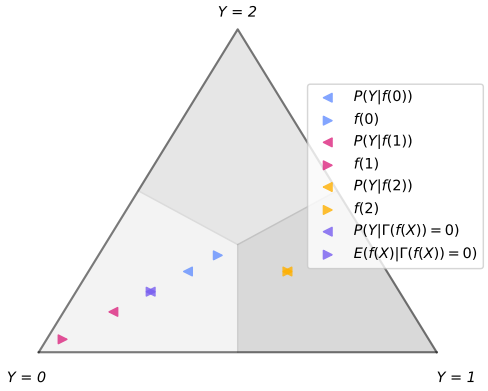
Distribution calibration with respect to some property is naturally inherited by *refined properties*. A property Φ is refined by another property Γ, if there exists a mapping which applied on Γ gives Φ.

[10] An interesting perspective on the conditioning events is given in [Grünwald et al., 2020, Section 2.3]. The author essentially describes calibrated predictors as "safe" summaries of a certain set of probability distributions, in our case the observed distribution D. The choice of conditioning events is then tantamount to a choice of uses for which the predictor is a "safe probability".

[11] Since $\mathcal{Y}$ is finite, we have $\mathcal{P} \subseteq \mathbb{R}^{|\mathcal{Y}|}$. Hence, the distribution calibration condition is equivalent to $D_{Y|\Gamma \circ f(X) = \gamma} = \mathbb{E}_D[f(X) | \Gamma \circ f(X) = \gamma]$.



(a) Approximate distribution calibration with respect to Γ.



(b) Perfect distribution calibration with respect to Γ.

Figure 10.3: Illustration of distribution calibration. The outcome set is defined as $\mathcal{Y} = \{0, 1, 2\}$, the input set $\mathcal{X} = \{0, 1, 2\}$. We define $\Gamma(P) = \arg\max_{y \in \mathcal{Y}} P(Y = y)$. The level sets of Γ are drawn in different shades of gray. The left-directing markers denote the true conditional distribution for different choices of $x \in \mathcal{X}$. The right-directing markers denote the predicted distribution by a predictor f. The purple markers are convex combinations of the blue and red markers, where the convex combination is defined through the marginal distribution on $\mathcal{X}$ which is in our case fixed to be uniform. The dashed lines highlight the deviation from the true outcome distribution conditioned on a value of $\Gamma \circ f$ versus the expected forecast conditioned on a value of $\Gamma \circ f$. Only the forecasts change when comparing Figure 10.3b versus Figure 10.3a.

Definition 10.3 (Property Refinement [Frongillo and Kash, 2021, Definition 12])**.** *A property $\Phi \colon \mathcal{P} \to \mathcal{R}'$ is called* refined by Γ*, if $\Gamma \colon \mathcal{P} \to \mathcal{R}$ is a property such that there exists a function $\phi \colon \mathcal{R} \to \mathcal{R}'$ with $\Phi = \phi \circ \Gamma$.*[12]

Proposition 10.4 (Distribution Calibration is Inherited)**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$, where $\mathcal{Y}$ is finite. Let $f \colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor . Suppose, furthermore, that for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$.*[13] *If the predictor f is $\alpha(\gamma)$-approximate distribution calibrated with respect Γ on D, then it is $\alpha'(c)$-approximate distribution calibrated with respect to $\Phi \coloneqq \phi \circ \Gamma$ for all $\phi \colon \mathcal{R} \to \mathcal{R}'$ where $\alpha'(c) \coloneqq \sup_{\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)} \colon \phi(\gamma) = c} \alpha(\gamma)$.*

[12] Property refinement is intricately related to the notion of *indirect property elicitation* [Frongillo and Kash, 2021, Finocchiaro et al., 2024, 2021].

[13] This technical assumption is necessary to exclude problems with aggregations of measure zero sets related to our definitional choice that our calibration conditions only hold almost everywhere. The assumption is essentially equivalent to assuming that f maps to at most countably many different values. The assumption re-appears in Proposition 10.8 and Proposition 10.9.2.

Proof. By assumption, for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

$$|\mathbb{E}_D[[\![Y = y]\!] - f_y(X) | \Gamma \circ f(X) = \gamma]| \leq \alpha(\gamma), \quad \forall y \in \mathcal{Y}.$$

Hence, in particular, for all $c \in \operatorname{supp} D_{\Phi \circ f(X)}$ and all $y \in \mathcal{Y}$,

$$\begin{aligned}
\alpha'(c) &= \sup_{\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)} \colon \phi(\gamma) = c} \alpha(\gamma) \\
&\geq \mathbb{E}_{G \sim D_{\Gamma \circ f(X)}} [\alpha(G) \mid \phi(G) = c] \\
&= \mathbb{E}_{G \sim D_{\Gamma \circ f(X)}} [|\mathbb{E}_D[[\![Y = y]\!] | \Gamma \circ f(X) = G] - \mathbb{E}_D[f_y(X) | \Gamma \circ f(X) = G]| \mid \phi(G) = c] \\
&\geq \left| \mathbb{E}_{G \sim D_{\Gamma \circ f(X)}} [\mathbb{E}_D[[\![Y = y]\!] | \Gamma \circ f(X) = G] - \mathbb{E}_D[f_y(X) | \Gamma \circ f(X) = G] \mid \phi(G) = c] \right| \\
&\geq |\mathbb{E}_D[[\![Y = y]\!] | \Phi \circ f(X) = c] - \mathbb{E}_D[f_y(X) | \Phi \circ f(X) = c]| \,.
\end{aligned}$$

Note that the supremum in the first line is well-defined, because for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. □

Under mild geometric conditions on the predictions, perfect distribution calibration with respect to all elicitable binary properties implies distribution calibration (Appendix 10.9.3). This statement can be understood as a reverse implication to Proposition 10.4. Figure 10.4 illustrates the idea of property refinement and inheritance of distribution calibration.

Distribution calibration with respect to Γ forms the entry point to the two accounts of calibration mentioned earlier (cf. Figure 10.2 and Figure 10.7). Distribution calibration is close to the account of mean-matching briefly mentioned in the introduction. However, this account puts more focus on groups and individuals, i.e., information provided through the input $\mathcal{X}$ (cf. [Höltgen and Williamson, 2023]).

10.5 Γ-Calibration as Self-Realization of Predictions

Central to vanilla calibration is the self-referential aspect in the conditioning, i.e., given the predicted Bernoulli distribution has parameter γ, the actual distribution of outcomes is Bernoulli-distributed with parameter γ. In this section we generalize calibration around this perspective.[14]

As observed by Jung et al. [2021, 2023] and Gupta et al. [2022], the expectation operator

[14] The self-referential aspect is similar in nature to a widely discussed principle for deference of belief in philosophy, the *reflection principle* [Molinari, 2023]. Van Fraassen [1984] introduced the reflection principle to argue that reflection of subjective beliefs of future selves is a requirement for rationality of the agent. In a footnote, Seidenfeld [1985, p. 276] suggest a definition of "subjective calibration" which resembles our Definition 10.5 and the reflection principle as defined in [Van Fraassen, 1984]. Unfortunately, we were not able to identify the work the author refers to.

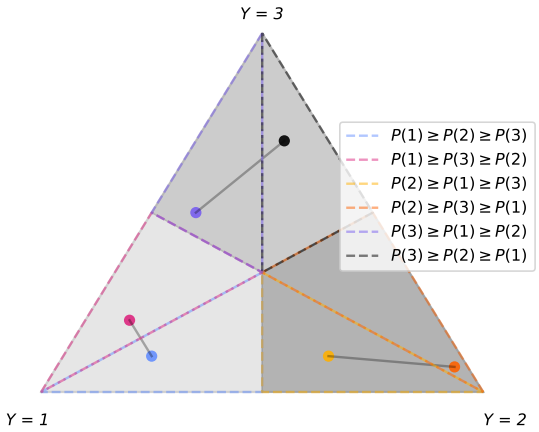


Figure 10.4: Illustration of property refinement and inheritance of distribution calibration. The outcome set is defined as $\mathcal{Y} = \{0, 1, 2\}$, the input set $|\mathcal{X}| = 6$. We define $\Gamma(P) = (y_1, y_2, y_3)$ such that $P(Y = y_1) \geq P(Y = y_2) \geq P(Y = y_3)$ and $\Phi(P) = \arg\max_{y \in \mathcal{Y}} P(Y = y)$. The level sets of Γ have colored boundaries listed in the legend. The level sets of Φ are drawn in different shades of gray. The property Γ refines Φ. We assume that f is a distributional predictor whose predictions are marked as dots. The color of the dots represents $\Gamma \circ f(x)$. The lines indicate the convex combination of predictions happening when conditioning on $\Phi \circ f(X)$ instead of $\Gamma \circ f(X)$. Since the level sets of Φ are all convex the line is always contained *within* a single level set.

in Definition 10.1 can be replaced by moments, respectively quantile functions. From these works, Noarov and Roth [2023] distilled a general definition of calibration with respect to properties. Different to Noarov and Roth [2023] we ignore groupings based on input X. Hence, we consider Γ-calibration and not "multi"-Γ-calibration.

Definition 10.5 (Γ-Calibration [Noarov and Roth, 2023])**.** *Suppose $(\mathcal{R}, m)$ is a metric space. Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let $f \colon \mathcal{X} \to \mathcal{R}$ be a Γ-predictor. Then, f is* Γ-calibrated *on D if for every $\gamma \in \operatorname{supp} D_{f(X)}$,*

$$\Gamma(D_{Y|f(X)=\gamma}) = \gamma.$$

The Γ-predictor f is $\alpha(\gamma)$-approximate Γ-calibrated *on D if for every $\gamma \in \operatorname{supp} D_{f(X)}$,*

$$m(\Gamma(D_{Y|f(X)=\gamma}), \gamma) \leq \alpha(\gamma).$$

Central to the definition is the self-realization aspect. The property of the distribution of outcomes, on which the predictor f predicted γ, actually is γ. The reader familiar with

[Noarov and Roth, 2023] might question the use of a general property Γ in Definition 10.5. For a discussion see Appendix 10.9.4.

Gneiting and Resin [2023] introduced "T-calibration" (Definition 2.7 therein), which is a measure-theoretic definition of Γ-calibration. The authors already note that certain properties Γ are not sensible for calibration following [Noarov and Roth, 2023, Definition 3.2], but only give necessary not sufficient conditions for the sensibility (cf. Appendix 10.9.4).

In practice it is rarely the case that a Γ-predictor is perfectly Γ-calibrated. However, there are algorithms, such as [Noarov and Roth, 2023, Algorithm 1] which post-hoc approximately calibrate Γ-predictors (for identifiable Γ). In that context, remember our conventions for m listed in Section 10.3, e.g., if $\mathcal{R}$ is an interval on the real line, then m is set to be the absolute difference for simplicity. However, the choice of metric m is worthy of its own line of inquiry (cf. Section 10.1).

In addition, observe that in many existing definitions the authors commit to a specified aggregation of $\alpha(\gamma)$ over all $\gamma \in \mathrm{im} f$. For instance, [Noarov and Roth, 2023] consider the expected squared value, $\alpha := \mathbb{E}_{\gamma \sim D_{f(X)}}[\alpha(\gamma)^2]$. For a summary of such aggregations see [Garg et al., 2024]. We illustrate Definition 10.5 in Figure 10.5.

10.5.1 Distribution Calibration Implies Γ-Calibration

Distribution calibration with respect to a property Γ almost immediately implies Γ-calibration. The argument naturally seems correct, but it requires careful treatment of convex combination of distributions on which distributional predictions with equal property values have been made. For the sake of readability, we restrict Proposition 10.6 to finite $\mathcal{Y}$. The extension to more general $\mathcal{Y}$ requires more technical machinery. We give the statement in the appendix (Proposition 10.22).

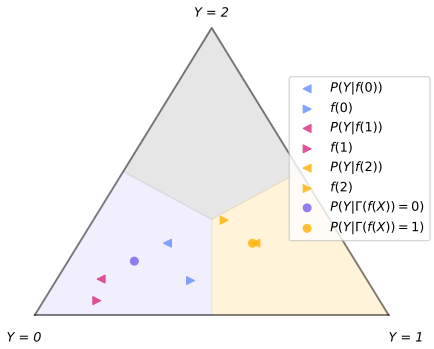
Proposition 10.6 (Distribution Calibration with Respect to Γ implies Γ-Calibration)**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$ be a property with convex level sets and D a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite, regular wrt. $\mathcal{P}$. Let $f \colon \mathcal{X} \to \mathcal{P}$ be a predictor which is distribution calibrated with respect to Γ. Then, $\Gamma \circ f$ is Γ-calibrated.*

Proof. We have to show that for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

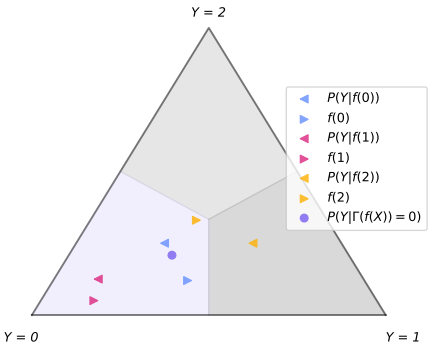
$$\Gamma(D_{Y|\Gamma \circ f(X)=\gamma}) = \gamma. \tag{10.1}$$

Let $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$ be arbitrary. By distribution calibration with respect to Γ,

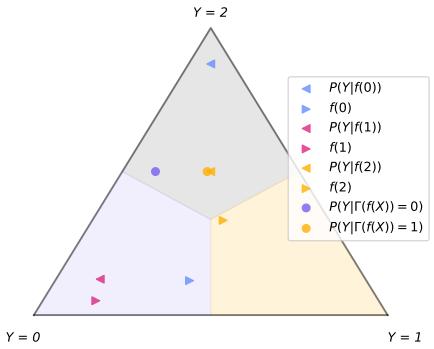
$$D_{Y|\Gamma \circ f(X)=\gamma} = \mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma].$$



(a) Perfect Γ-calibration.



(b) Perfect Γ-calibration II.



(c) Approximate Γ-calibration.

Figure 10.5: Illustration of Γ-calibration. The outcome set is defined as $\mathcal{Y} = \{0, 1, 2\}$, the input set $\mathcal{X} = \{0, 1, 2\}$. We define $\Gamma(P) = \arg\max_{y \in \mathcal{Y}} P(Y = y)$. The level sets of Γ are colored respectively drawn in different shades of gray. The left-directing markers denote the true conditional distribution for different choices of $x \in \mathcal{X}$. We assume that f is a distributional predictor (right-directing markers) which is then fed into Γ. The dots denote the true outcome distribution conditioned on a value of $\Gamma \circ f$. If the dot is in the level set of the same color, then the prediction is perfectly Γ-calibrated. Note that Figure 10.5a uses the same predictions and outcome distributions as Figure 10.3a showing that predictions could be perfect Γ-calibrated while only being approximately distribution calibrated with respect to Γ. Figure 10.5b shows that Γ-calibration does not necessarily require the true conditional distributions to all live in the correct level set. Only the predictor f is changed from Figure 10.5a to Figure 10.5b. Figure 10.5c then changes the true conditional distribution compared to Figure 10.5a but fixes f, which lead to the violation of the perfect Γ-calibration constraint.

Lemma 10.42 applied to D, f and $A := \{x \in \mathcal{X} : \Gamma \circ f(x) = \gamma\}$ gives,

$$\Gamma(\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]) = \gamma,$$

hence Equation 10.1 follows. □

The implication can, in general, *not* be extended to a bi-implication. Figure 10.3a and Figure 10.5a are illustrations for the same predictor, which is approximately distribution calibrated but perfectly Γ-calibrated. The intuition is that Γ-calibration makes the grid for testing calibration coarser.

The above Proposition 10.6 holds for arbitrary properties with convex level sets. In the approximate case, we show that the implication continues to hold if the property Γ is Lipschitz continuous (Appendix 10.9.2). Note that discrete properties Γ don't fulfill this condition, because a slight mismatch of the average predictions can lead to a catastrophic mismatch in the corresponding Γ-values.

10.5.2 Γ-Calibration is Equivalent to Low Swap Regret

Property calibration can be equivalently reformulated for loss (respectively identification) functions, if the property is elicitable (respectively identifiable).[15]

Proposition 10.7 (Perfect Γ-Calibration via Loss Function or Identification Function)**.** *Let $\Gamma : \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let f be Γ-calibrated on D.*

(a) Suppose Γ is elicitable with ℓ, then f is Γ-calibrated on D if and only if, for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$\mathbb{E}_D\left[\ell(Y, \gamma)|f(X) = \gamma\right] - \min_{\hat{\gamma} \in \operatorname{im}\Gamma} \mathbb{E}_D\left[\ell(Y, \hat{\gamma})|f(X) = \gamma\right] = 0.$$

(b) Suppose Γ is identifiable with V, then f is Γ-calibrated on D if and only if, for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$\mathbb{E}_D\left[V(Y, \gamma)|f(X) = \gamma\right] = 0.$$

Proof. By definition of elicitability and identifiability. □

[15] Already in [Davis, 2016] the terms "calibration" and "identification" got linked to each other. However, their Definition 4.2 of calibration is largely disconnected from the use of the term in current machine learning literature.

Statement (a) is essentially a "no swap regret" statement. The predictor f is compared with the best prediction $\hat{\gamma}^*$ on the sets on which the predictor f predicted a value γ. If statement (a) is not perfectly but approximately fulfilled, we call it "low swap regret". No swap regret is different from precise loss estimation as we argue in Section 10.6. The no swap regret transfers to the approximate case, i.e., low swap regret. For that, certain regularity assumptions on the property Γ and its identification function V are needed. In Appendix 10.9.2 we argue for the general correspondence and link it to the work on swap multicalibration and swap-agnostic learning [Gopalan et al., 2024b].

10.5.3 Γ-Calibration is Inherited

There is a commonplace gap between *optimization-level* Γ and *decision-level* properties Φ which trades off continuity for a granularity unnecessary for decision-making. Given this gap, we have to ask if self-realization with respect to Γ extends to a related Φ. In the sequel, we show that self-realization, as Γ-calibration, is inherited by *refined* properties (Definition 10.3).[16] That way, calibrated forecasts for optimization-level properties provide calibrated estimates for decision-level properties. In the language of decision making, Γ-calibration guarantees low swap regret for downstream decision makers.[17]

Proposition 10.8 (Γ-Calibration is Inherited by Refined Properties)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property, D a regular data distribution on $\mathcal{X} \times \mathcal{Y}$ with respect to $\mathcal{P}$. Suppose that $\mathcal{Y}$ is finite and f is a Γ-predictor which is Γ-calibrated on D. Suppose, furthermore, that for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. Then, for every property Φ with convex level sets and which is refined by Γ, the Φ-predictor $\phi \circ f$, where ϕ is defined following Definition 10.3, is Φ-calibrated on D.*

Proof. We have to show that,

$$\Phi(D_{Y|\phi\circ f(X)=v}) = \phi \circ \Gamma(D_{Y|\phi\circ f(X)=v}) = v,$$

for all $v \in \operatorname{supp} D_{\phi\circ f(X)}$. Fix $v \in \operatorname{supp} D_{\phi\circ f(X)}$ for the following. Let us define the probabilistic predictor $h\colon \mathcal{X} \to \mathcal{P}$ where $x \mapsto D_{Y|f(X)=f(x)}$. The mapping is measurable by definition, as f is measurable by assumption. Let $A \coloneqq \{x \in X\colon \phi \circ f(x) = v\}$. Since, f is Γ-calibrated, and for all $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$, it holds,

$$\Gamma(h(x)) = \Gamma(D_{Y|f(X)=f(x)}) = f(x)$$

[16] The argument seems more involved then necessary. This is a consequence of carefully arguing about convex combinations of outcome distributions analogous to Proposition 10.6.

[17] This provides a major motivation for self-realization in the first place [Noarov and Roth, 2024].

for all $x \in A$. Hence, in particular,

$$\Phi(h(x)) = v,$$

for all $x \in A$. By the law of total expectation, we obtain, for all $y \in \mathcal{Y}$,

$$\begin{aligned} D_{Y|\phi \circ f(X)=v}(Y = y) &= \mathbb{E}_D[[\![Y = y]\!]|\phi(f(X)) = v] \\ &= \mathbb{E}_{G \sim D_{f(X)}}\left[\mathbb{E}_D[[\![Y = y]\!]|f(X) = G]|\phi(G) = v\right] \\ &= \mathbb{E}_{G \sim D_{f(X)}}\left[\mathcal{D}_{Y|f(X)=G}(Y = y)|\phi(G) = v\right]. \end{aligned}$$

It follows by definition of h, A and because $\mathcal{Y}$ is finite, $D_{Y|\phi \circ f(X)=v} = \mathbb{E}_{D_X}[h(X)|X \in A]$. Applying Lemma 10.42 to h, D and A gives,

$$\Phi(D_{Y|\phi \circ f(X)=v}) = \Phi(\mathbb{E}_{\sim D_X}[h(X)|X \in A]) = v.$$

□

The above statement can be extended to more general $\mathcal{Y}$ beyond finiteness (Proposition 10.23).

Given that the Γ-predictor f is *not* perfectly calibrated, we show two approaches to elaborate the inheritance statement for approximately calibrated f in the appendix. First, we show that if Γ is Lipschitz, a relaxed inheritance statement continues to hold (Proposition 10.31). b) Second, we argue that if the decision-level property Φ is elicited by a Lipschitz loss function ℓ, a relatively weak assumption, then approximate Γ-calibration implies low swap regret with respect to this loss function ℓ (Proposition 10.33). In particular, the second approach is contextualized within the paradigm of "omniprediction" [Gopalan et al., 2022b] (Section 10.9.2).

The inheritance of self-realization is central to the rationale of using calibration. It provides low swap regret guarantees for a large class of loss functions (Proposition 10.33). Each such loss function corresponds to an elicitable property, a "decision-level" property. For instance, this decision-level property can model a person who is about to decide whether to take or not to take their umbrella with them given a probabilistic forecast on rain. However, self-realization can have largely decoupled effects on the usefulness of the predictions for the "decision-level" properties, respectively the decision makers. A calibrated prediction can be more useful to one than to another decision maker (Appendix 10.9.7).

In contrast with the account laid out above, Zhao et al. [2021] motivate calibration differently. They consider a predictor to be calibrated, if it can precisely[18] estimate the average loss incurred to the individuals. Formally, the intuition can be captured by a type of loss outcome indistinguishability [Gopalan et al., 2022a] following [Zhao et al., 2021].

Definition 10.9 (Decision Calibration (Modified from Zhao et al. [2021]))**.** *Let Γ be an elicitable property and $\mathcal{L}$ be a set of $\mathcal{P}$-consistent loss functions $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ with respect to Γ. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$ and $f\colon \mathcal{X} \to \mathcal{P}$ a distribution predictor. The predictor f is* β-approximate decision calibrated for $\mathcal{L}$*, if for all $\ell \in \mathcal{L}$,*

$$\left|\mathbb{E}_{(X,Y)\sim D}[\ell(Y, \Gamma(f(X))) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y}, \Gamma(f(X))]]\right| \leq \beta.$$

In the original definition [Zhao et al., 2021, Definition 2] the loss function ℓ and the elicited property Γ can be independent of each other. For instance, the loss function might penalize deviations of $f(X)$ from the mean, while Γ maps $f(X)$ to the median. That is a semantic mismatch. We were not able to identify a convincing justification for this choice, hence we neglect this degree of freedom for our purposes. Hence, we suggest a definition which generalizes "decision outcome indistinguishability" beyond the binary [Gopalan et al., 2022a].

In our definition we follow the presentation by Fröhlich and Williamson [2024]. They back the intuition of "calibration via precise loss estimation" from an actuarial point of view referring to the ideal of *actuarial fairness* [Arrow, 1978]:

> The forecaster should offer actuarially fair insurance for the uncertain loss [...]. Actuarial fairness here means that in the long run, the forecaster neither loses nor profits from offering insurance under the data model. [Fröhlich and Williamson, 2024, p. 12]

In other words, both parties get a fair deal. The forecaster could, without getting bankrupt, ask for a certain premium to all its decision makers based on the forecast and in turn cover all losses the decision maker might encounter because of the uncertain outcomes. Decision calibration fits to the narrative of "trustworthiness" via precise loss estimates encouraged by Kirchhof et al. [2024] and Yoo and Kweon [2019] and which can be traced back to at least [Rukhin, 1988]. Note in those papers precise loss estimation is usually desired for every individual. In contrast, we focus on average precise loss estimation here. Decision

[18] We deliberately refer to "precise" loss estimates as we focus on the internal consistency of the loss estimates and not the actual closeness to the truth (accuracy) [Everitt, 2002, p. 295].

calibration enforced on sub-groups based on input X bridge between the two extremes: individual precise loss estimation versus average precise loss estimation (cf. Section 10.7).

Decision calibration, like all other so far named notions of calibration, grows on the ground of binary vanilla calibration (cf. Proposition 10.11). We show that on binary outcome sets decision calibration with respect to the set of *simple loss functions*, also known as "cost-weighted misclassification losses", is equivalent to binary vanilla calibration. Simple loss functions play a key role in defining the spectrum of proper loss functions for binary outcome sets, [Schervish, 1989] (for a modern proof see [Reid and Williamson, 2011, Theorem 16] and an independent rediscovery [Kleinberg et al., 2023]).

Definition 10.10 (Simple Loss Function)**.** *Let $q \in [0,1]$. A loss function $\ell_q \colon \{0,1\} \times \{a,b\} \to \mathbb{R}$ is called* simple *if it has the form,*

$$\ell_q(y,c) = q[\![y = 0, c = a]\!] + (1-q)[\![y = 1, c = b]\!].$$

The loss ℓ_q elicits the simple binary property,

$$\Gamma_q \colon \Delta(\mathcal{Y}) \to \{a,b\}; P \mapsto \begin{cases} a \text{ if } P(Y=1) > q \\ b \text{ if } P(Y=1) \leq q. \end{cases}$$

We overload the notation of a simple loss function with its proper loss cousin: $\ell_q \colon \{0,1\} \times [0,1] \to \mathbb{R}$, $\ell_q(y,p) = \ell_q(y, \Gamma_q(p))$.

The equivalence of decision calibration and binary vanilla calibration has already been discussed in [Zhao et al., 2021, Theorem 1]. However, their proof is hard to follow and makes heavy use of the independent choice of the decision function and the loss function, which we hesitated to commit to (see discussion below Definition 10.9). Hence, our statement requires a weaker definition of decision calibration. On the other hand, our result only holds on binary outcome sets and requires predictors with finitely many output values, arguably a mild assumption for practical settings.[19]

Proposition 10.11 (Decision Calibration Equivalent to Vanilla Calibration for Binary Outcome Set)**.** *Let $\mathcal{Y} = \{0,1\}$ and suppose $\mathcal{X}$ is arbitrary. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f \colon \mathcal{X} \to [0,1]$ is a mean-predictor, i.e., it predicts the probability of $Y = 1$, and suppose that $|\mathrm{im} f| < \infty$ and $D_X(f(x) = v) > 0$ for all $v \in \mathrm{im} f$. Then, f is calibrated if, and only if, f is decision calibrated with respect to all simple loss functions $\{\ell_q \colon q \in [0,1]\}$.*

Proof. First, we consider two different cases. Let $f_{\inf} \coloneqq \inf_{\gamma \in \mathrm{im} f} \gamma$.

[19] For an intuitive argument why Proposition 10.11 should hold see the example provided in [Höltgen, 2024, Section 2.2].

(a) If $f_{\inf} > 0$, then Lemma 10.45 applies, hence f *matches the averages*, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$.

(b) If $f_{\inf} = 0$, then Lemma 10.46 applies because $\mathrm{im}f$ is finite, hence f *matches the averages*, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$.

Lemma 10.44 can be used to show that for all $q \in [0,1]$, $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) \leq q] = 0$ and $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) > q] = 0$. It remains to show that for all $v \in \mathrm{im}f$, $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) = v] = 0$. If $v = 0$, this is already given. For every $v \in \mathrm{im}f \setminus \{0\}$ there exists $v' \in \mathrm{im}f \cup \{0\}$ such that $v' < v$ (because $\mathrm{im}f$ is finite). We have,

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) \leq v] \\
&= P_D(f(X) \leq v')\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) \leq v'] \\
&\quad + P_D(f(X) = v)\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) = v] \\
&= P_D(f(X) = v)\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) = v] = 0,
\end{aligned}$$

which gives the desired result. □

In light of Proposition 10.11, the "precise loss estimation"-account is another generalization of binary vanilla calibration whose semantic focus is distinct from the self-realization discussed in § 10.5. Furthermore, precise loss estimation via decision calibration is also a child of the general notion of distribution calibration.

10.6.1 Distribution Calibration Implies Decision Calibration

Rather unsurprisingly, distribution calibration (Definition 10.2) implies decision calibration (Definition 10.9). This holds not only in the perfect case but as well in the case of approximate calibration if the loss function corresponding to the property of interest is bounded (Appendix 10.9.2). Proposition 10.12 (respectively the approximate version Proposition 10.34) has been proven in [Zhao et al., 2021, Proposition 2]. However, we were not able to follow all steps of the proofs. For this reason we provide a full argument here.

Proposition 10.12 (Distribution Calibration Implies Decision Calibration)**.** *Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite, regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor which is distribution calibrated with respect to Γ on D. Then, f is decision calibrated with respect to $\mathcal{L}_\Gamma \coloneqq \{\ell\colon \ell$ is $\mathcal{P}$-consistent loss function for $\Gamma\}$.*

Proof. The predictor f is *distribution calibrated with respect to* Γ on D if for every $\gamma \in$

$\operatorname{supp} D_{\Gamma(f(X))}$,

$$D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) = \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma\circ f(X)=\gamma], \quad \forall y\in\mathcal{Y}, \tag{10.2}$$

where $f_y(x)\in[0,1]$ denotes the y-component of the prediction $f(x)\in\Delta(\mathcal{Y})$ for $x\in\mathcal{X}$.

Let $\ell\in\mathcal{L}_\Gamma$ be arbitrary. For every $\gamma\in\operatorname{supp} D_{\Gamma(f(X))}$, we have,

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X)=\gamma\right] \\
&= \mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma)|\Gamma\circ f(X)=\gamma\right] - \mathbb{E}_{(X,Y)\sim D}\left[\mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X)=\gamma\right] \\
&= \sum_{y\in\mathcal{Y}} D_{Y|\Gamma\circ f(X)=\gamma}(Y=y)\ell(y,\gamma) - \mathbb{E}_{X\sim D_X}\left[\sum_{y\in\mathcal{Y}} f_y(X)\ell(y,\gamma)\,\middle|\,\Gamma\circ f(X)=\gamma\right] \\
&= \sum_{y\in\mathcal{Y}} \left(D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) - \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma\circ f(X)=\gamma]\right)\ell(y,\gamma) \\
&= 0,
\end{aligned}$$

by Equation (10.2). Hence, it follows,

$$\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\Gamma(f(X))) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\Gamma(f(X)))]\right] = 0.$$

□

Is the reverse direction true? Proposition 10.11 provides an affirmative answer in the case of binary outcomes. However, the proof is not immediately extendable beyond the binary. The question remains open for future work.

10.6.2 Precise Bayes Risk Estimation

Analogous to self-realization we obtain a clearer picture of the purpose of decision calibration when rewriting in terms of properties. To this end, we introduce Bayes pairs. A Bayes pair unifies the minimizer (elicited property) and the minimization value (Bayes risk) of a loss function.[20] Hence, following the account of precise loss estimation the Bayes risk is of central importance here. Being calibrated in the sense of precise loss estimation, is then tantamount to precise estimation of the Bayes risk.

Definition 10.13 (Bayes Pair [Embrechts et al., 2021])**.** *Let $\ell\colon\mathcal{Y}\times\mathcal{R}\to\mathbb{R}$ be a loss*

[20] The importance of eliciting Bayes pairs is also studied by [Fissler and Ziegel, 2016, Frongillo and Kash, 2021, Finocchiaro et al., 2021].

function. The property pair (Φ_ℓ, Θ_ℓ) *is called a Bayes pair, if, for all* $P \in \mathcal{P}$,[21]

$$\Phi_\ell(P) = \arg\min_{\phi \in \mathcal{R}} \mathbb{E}_{Y \sim P}[\ell(Y, \phi)],$$
$$\Theta_\ell(P) = \min_{\phi \in \mathcal{R}} \mathbb{E}_{Y \sim P}[\ell(Y, \phi)].$$

By definition, Φ_ℓ is elicitable. That the Bayes risk Θ_ℓ is elicitable on level sets of Φ_ℓ is less obvious. To see this, observe that there exists a $\Phi_\ell^{-1}(\phi)$-consistent loss function for Θ_ℓ for every $\phi \in \mathrm{im}\Phi_\ell$ [Embrechts et al., 2021]. We proceed by defining precise Bayes risk estimators.

Definition 10.14 (Precise Bayes Risk Estimator)**.** *Let* $\ell \colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ *be a loss function with corresponding Bayes pair* (Φ_ℓ, Θ_ℓ)*. Let* D *be a regular data distribution. A* (Φ_ℓ, Θ_ℓ)*-predictor* $f = (g, h)$ *is a* β-precise Bayes risk estimator *if,*

$$|\mathbb{E}_D[\ell(Y, g(X)) - h(X)]| \leq \beta.$$

This definition is extremely weak. The predictor h estimates the loss of g but only has to fit it on average. Additionally, there is *no* demand placed on the quality of g. The definition can be significantly strengthened if precise Bayes risk estimation is demanded on subgroups or individuals (cf. Section 10.7). Note that such Bayes risk predictors exist in current literature. For instance, they are called "uncertainty module" [Kirchhof et al., 2024] or "loss prediction module" [Yoo and Kweon, 2019]. It follows rather immediately that a decision calibrated full probability predictor is a precise Bayes risk estimator.

Proposition 10.15 (Decision Calibration Implies Precise Bayes Risk Estimation)**.** *Let* $\mathcal{L}$ *be a set of* $\mathcal{P}$*-consistent loss functions* $\ell \colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ *with corresponding Bayes pairs* (Φ_ℓ, Θ_ℓ)*. Let* D *be a data distribution on* $\mathcal{X} \times \mathcal{Y}$ *regular wrt.* $\mathcal{P}$ *and* $f \colon \mathcal{X} \to \mathcal{P}$ *a* β*-approximate decision calibrated predictor for* $\mathcal{L}$*. Then, for all* $\ell \in \mathcal{L}$ *the predictor* $(\Phi_\ell \circ f, \Theta_\ell \circ f)$ *is a* β*-precise Bayes risk estimator.*

Proof. We can rewrite the criterion of decision calibration as,

$$\left|\mathbb{E}_D[\ell(Y, \Phi_\ell(f(X))) - \mathbb{E}_{\hat{Y} \sim f(X)}[\ell(\hat{Y}, \Phi_\ell(f(X)))]]\right| = |\mathbb{E}_D[\ell(Y, \Phi_\ell(f(X))) - \Theta_\ell(f(X))]| \leq \beta. \tag{10.3}$$

□

[21] Under the restriction that $\mathcal{R}$ is compact both properties are guaranteed to be measurable [Aliprantis and Border, 2006, Theorem 18.19].

10.6.3 When Self-Realization Implies Precise Loss Estimation

In the sections before we explored the landscape of calibration as self-realization and as precise loss estimation. We introduced two prototypical notions of calibration using properties for each of the two worlds: property calibration and precise Bayes risk estimation. In the following section, we present several results and examples which shed light on the complicated relationship between the notions.

If we have access to a predictor which predicts a Bayes pair and is property-calibrated with respect to the involved Bayes risk property, then the predictor is a precise Bayes risk estimator (Proposition 10.16). That is, property calibration for the property (Γ, Θ) implies precise Bayes risk estimation.

Proposition 10.16 (Self-Realization Implies Precise Loss Estimation)**.** *Let (Φ_ℓ, Θ_ℓ) be a Bayes pair corresponding to a loss function ℓ and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. If a (Φ_ℓ, Θ_ℓ)-predictor $f = (g, h)$ is $\alpha(\phi, \theta)$-approximate (Φ_ℓ, Θ_ℓ)-calibrated, then f is a α-precise Bayes risk estimator, where $\alpha \coloneqq \mathbb{E}_{(\phi,\theta)\sim Q}[|\alpha(\phi, \theta)|]$ and $Q \coloneqq D_{(g(X),h(X))}$*

Proof.

$$
\begin{aligned}
&|\mathbb{E}_D[\ell(Y, g(X)) - h(X)]| \\
&= |\mathbb{E}_{(\phi,\theta)\sim Q}[\mathbb{E}_{(X,Y)\sim D|g(X)=\phi,h(X)=\theta}[\ell(Y, \phi)] - \theta]| \\
&\leq \mathbb{E}_{(\phi,\theta)\sim Q}[|\Theta_\ell(D_{Y|g(X)=\phi,h(X)=\theta}) - \theta|] \\
&\leq \mathbb{E}_{(\phi,\theta)\sim Q}[|\alpha(\phi, \theta)|] = \alpha.
\end{aligned}
$$

□

The implication cannot be generally extended into a bi-implication as the following example shows. In this example, we construct a Bayes pair predictor, i.e., a predictor which predicts the first and second component of a Bayes pair, which estimates the Bayes risk precisely. However, this predictor is not calibrated with respect to the first or second component of the Bayes pair. A instantiation of this example would be a $(\mathbb{M}, \mathbb{V})$-predictor, mean and variance predictor, which precisely estimates the Bayes risk, but is not $(\mathbb{M}, \mathbb{V})$-calibrated, nor is the first component of the predictor $\mathbb{M}$-calibrated, i.e., mean-calibrated, or the second component $\mathbb{V}$-calibrated, i.e., variance-calibrated. In that sense, self-realization is "stronger" than precise loss estimation.

Example 10.17 (On the Necessity of Self-Realization for Precise Loss Estimation)**.** *The following example is based on an arbitrarily chosen Bayes pair. The idea is that a predictor of the Bayes pair which swaps the elicited value, i.e., the first component of the Bayes pair,*

can still precisely estimate the Bayes risk. In other words, the predictor is biased, but still precisely estimate its error.

Let (Φ_ℓ, Θ_ℓ) be the Bayes pair corresponding to the loss function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$. Let (Φ_ℓ, Θ_ℓ) be the Bayes pair corresponding to the loss function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$. Suppose that there exist $A \subseteq \mathcal{X}$, such that $1 > D_X(A) > 0$ and

$$\begin{aligned}
\Phi_\ell(D_{Y|X\in A}) &= \phi_A, & \Theta_\ell(D_{Y|X\in A}) &= \mathbb{E}_{D_Y}[\ell(Y,\phi_A)|X \in A],\\
\Phi_\ell(D_{Y|X\in \overline{A}}) &= \phi_{\overline{A}}, & \Theta_\ell(D_{Y|X\in \overline{A}}) &= \mathbb{E}_{D_Y}[\ell(Y,\phi_{\overline{A}})|X \in \overline{A}],
\end{aligned}$$

where $\phi_A \neq \phi_{\overline{A}}$.

Let (f, g) be a stacked predictor of the Bayes pair (Φ_ℓ, Θ_ℓ). Suppose that,

$$f(x) = \begin{cases} \phi_{\overline{A}} \text{ for } x \in A \\ \phi_A \text{ for } x \in \overline{A}. \end{cases},$$

$$g(x) = \begin{cases} \mathbb{E}_{D_Y}[\ell(Y,\phi_{\overline{A}})|X \in A] & \text{for } x \in A \\ \mathbb{E}_{D_Y}[\ell(Y,\phi_A)|X \in \overline{A}] & \text{for } x \in \overline{A}. \end{cases}.$$

We can directly show that (f, g) is a precise loss estimator,

$$\begin{aligned}
&\mathbb{E}_D[\ell(Y, f(X)) - g(X)]\\
&= D_X(A)\mathbb{E}_{D_Y}[\ell(Y, f(X)) - g(X)|X \in A] + D_X(\overline{A})\mathbb{E}_{D_Y}[\ell(Y, f(X)) - g(X)|X \in \overline{A}]\\
&= D_X(A)\mathbb{E}_{D_Y}[\ell(Y, \phi_{\overline{A}}) - g(X)|X \in A] + D_X(\overline{A})\mathbb{E}_{D_Y}[\ell(Y, \phi_A) - g(X)|X \in \overline{A}]\\
&= 0.
\end{aligned}$$

On the other hand, f is clearly not Φ_ℓ-calibrated, because,

$$\begin{aligned}
\Phi_\ell(D_{Y|f(X)=\phi_{\overline{A}}}) &= \Phi_\ell(D_{Y|X\in A}) = \phi_A \neq \phi_{\overline{A}},\\
\Phi_\ell(D_{Y|f(X)=\phi_A}) &= \Phi_\ell(D_{Y|X\in \overline{A}}) = \phi_{\overline{A}} \neq \phi_A.
\end{aligned}$$

Furthermore, g is not Θ_ℓ-calibrated, because

$$\begin{aligned}
\Theta_\ell(D_{Y|g(x)=\mathbb{E}_{D_Y}[\ell(Y,\phi_{\overline{A}})|X\in A]}) &= \Theta_\ell(D_{Y|X\in A})\\
&= \mathbb{E}_{D_Y}[\ell(Y,\phi_A)|X \in A] < \mathbb{E}_{D_Y}[\ell(Y,\phi_{\overline{A}})|X \in A],
\end{aligned}$$

and,

$$\begin{aligned}\Theta_\ell(D_{Y|g(x)=\mathbb{E}_{D_Y}[\ell(Y,\phi_A)|X\in\overline{A}]}) &= \Theta_\ell(D_{Y|X\in\overline{A}}) \\ &= \mathbb{E}_{D_Y}[\ell(Y,\phi_{\overline{A}})|X\in\overline{A}] < \mathbb{E}_{D_Y}[\ell(Y,\phi_A)|X\in\overline{A}],\end{aligned}$$

by (strict) consistency of the loss function ℓ.

However, precise loss estimation always requires one to predict not only the property of interest but as well the corresponding Bayes risk. For instance, a pure mean predictor f as in Example 10.17 can be perfectly calibrated but still misses the information for being a precise Bayes risk estimator. In that sense, Γ-calibration and precise Bayes risk estimation are incommensurable as they are properties of different predictors. There is a type mismatch. This explains why distribution calibration, which requires full prediction estimates, subsumes both: all properties of a distribution including the Bayes risks are refined by the full distribution.

Nevertheless, Example 10.17 can be strengthened. If $\mathcal{Y}$ is binary, then the mean predictor identifies the full distribution. But even then, there exist predictors which are perfectly decision calibrated with respect to the squared loss (which implies precise Bayes risk estimation cf. Proposition 10.15), but still are not mean calibrated. We argue in the following, that decision calibration is *too* weak to imply property calibration (cf. Proposition 10.11). More concretely, we show in the following lemma that for any continuous loss function which elicits the mean there is a constant mean-predictor which is decision calibrated with respect to the loss function, independent of the data distribution. Hence, depending on the data distribution the constant mean-predictor is not mean-calibrated, as we argue afterwards.

Lemma 10.18 (Existence of Constant Decision Calibrated Mean Predictor)**.** *Let $\mathcal{Y} = \{0,1\}$ and $\mathcal{X} = \mathbb{R}$. For any continuous proper loss function $\ell\colon \{0,1\}\times[0,1]\to\mathbb{R}$ which elicits the mean on $\mathcal{Y}$, there is a (constant) mean predictor $f\colon \mathcal{X}\to[0,1]$ which is decision calibrated[22] for $\{\ell\}$ on every data distribution D on $\mathcal{X}\times\mathcal{Y}$.*

Proof. First, we argue that there exist $q^*\in[0,1]$ such that $\ell(0,q^*) = \ell(1,q^*)$. This is true, because of properness, for all $q\in[0,1]$,

$$\begin{aligned}\ell(0,0) &\le \ell(0,q)\\ \ell(1,1) &\le \ell(1,q).\end{aligned}$$

[22] We abuse the term decision calibration here a bit. Decision calibration following Definition 10.9 is defined for probabilistic predictors. In our present case, the predictor is a mean predictor. However, the mean fully identifies a distribution because we are in a binary setting.

Which implies that $h\colon q \mapsto \ell(1,q) - \ell(0,q)$, fulfills,

$$h(0) = \ell(1,0) - \ell(0,0) \geq 0$$
$$h(1) = \ell(1,1) - \ell(0,1) \leq 0.$$

By continuity of h, because of continuity of ℓ, it follows by the intermediate value theorem, that there exists $q^* \in [0,1]$ such that $h(q^*) = 0$.
Second, let us rewrite the decision calibration condition, for data distribution D,

$$\begin{aligned}&\mathbb{E}_{(X,Y)\sim D}[\ell(Y, f(X) - \mathbb{E}_{\hat{Y}\sim(1-f(X),f(X))}[\ell(\hat{Y}, f(X)]] \\ &= \mathbb{E}_{X\sim D_X}[\mathbb{E}_{Y\sim D_Y}[\ell(Y, f(x)|X=x] - \mathbb{E}_{\hat{Y}\sim(1-f(x),f(x))}[\ell(\hat{Y}, f(x)|X=x]].\end{aligned}$$

Define $p(x) \coloneqq D_{Y|X=x}(\{1\})$ and focus on the inner part of the expectation conditioned on $X = x$,

$$\begin{aligned}\mathbb{E}_{Y\sim D_Y}[\ell(Y, f(x))|X=x] &= p(x)\left(\ell(1, f(x)) - \ell(0, f(x))\right) + \ell(0, f(x)) \\ \mathbb{E}_{Y\sim(1-f(X),f(X))}[\ell(Y, f(x))|X=x] &= f(x)\left(\ell(1, f(x)) - \ell(0, f(x))\right) + \ell(0, f(x)).\end{aligned}$$

That is, the inner difference is,

$$\left(p(x) - f(x)\right)\left(\ell(1, f(x)) - \ell(0, f(x))\right),$$

which is 0 if $f(x) = q^*$. This proves the statement. □

It is now a matter of a simple construction for a specific data distribution D to show that a mean-predictor is decision calibrated but not mean-calibrated.

Example 10.19 (Decision-calibrated Mean-Predictor Which is Not Mean-Calibrated)**.** *Let $\mathcal{Y} = \{0,1\}$ and $\mathcal{X} = \mathbb{R}$. Pick $f\colon \mathcal{X} \to [0,1]$ to be a constant predictor $f(x) = q^*$ which is decision calibrated for the continuous proper loss function $\ell\colon \{0,1\} \times [0,1] \to \mathbb{R}$. Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ with $\mathbb{E}_{Y\sim D_Y}[Y] = q \neq q^*$. Then, clearly f is* not *mean-calibrated,*

$$\mathbb{M}(D_{Y|f(X)=q^*}) = q \neq q^*.$$

10.7 What about groups?

Certain frameworks like fair machine learning demand calibration on subgroups defined through the input space $\mathcal{X}$. Hence, one can understand calibration in fair machine learning, as asking for "equally good self-realization of predictions on sensitive subgroups" or "equally precise loss estimation on sensitive subgroups". The notions of calibration we considered

can easily be extended to subgroups. For instance, by taking a supremum over subgroups in Definition 10.5 one recovers the notion of multi-calibration used in [Noarov and Roth, 2024]. Respectively, a precise Bayes risk estimator following Definition 10.14 can be strengthened by checking for precise Bayes risk estimation on all relevant subgroups. Note that the choice of groups on which a certain notion of calibration ought to hold is *independent* of the choice of the notion itself.

In the extreme case, those subgroups could be enforced to the level of individuality [Zhao et al., 2020, Luo et al., 2022]. Then, "trustworthiness" meets "usefulness" again. The Bayes optimal predictor is the only predictor which achieves "trustworthiness" on all individuals. In particular, it is possible to equalize "trustworthiness" on arbitrary subgroups without losing on "usefulness". However, it is not possible to equalize "usefulness" on arbitrary subgroups without losing on "trustworthiness" [Barocas et al., 2023, Proposition 4 & 5].

Concluding, all considerations made on the semantics of the three different types of calibration considered in this work plus their formal relationships directly transfer to the fairness setting with multiple groups.

10.8 Conclusion

This work started with the promise to provide semantical structure in the chaotic world of calibration notions. We consider as our main contribution to semantically and formally separate two concerns calibration wants to solve: self-realization and precise loss estimation. Both accounts differently motivate a certain type of calibration. However, they can be bridged by distribution calibration.

How do existing notions relate to the three different types? Table 10.3 and Table 10.4 summarize a list of existing notions and their categorization in the three different types. Note that we make a distinction between formal strict derivatives, i.e., our suggested notion generalizes the notion in the list, and definitions which follow the type of definition more broadly.

What is the learning for a practitioner? We have *not* provided any new algorithm nor have we suggested *the* notion of calibration. But, our work shapes the debate on what is the right notion. It highlights implications (see Figure 10.2 and Figure 10.7), incommensurabilities (see Section 10.6.3) and equivalences (Proposition 10.36, Proposition 10.25 and Proposition 10.11). If those relationships are ignored, this may lead to unforeseen, unintentional consequences or may unnecessarily complicate a prediction problem[23]. Our work helps to define a set of questions which are relevant to formalizing a real-world prediction

[23]For instance, in the example given in Section 10.2 any calibration beyond the ranking is unnecessary, since the policy is bound to the rankings only [Perdomo et al., 2025].

Formally Strict Derivative	Name	Reference
Distribution Calibration	class-wise calibration	[Kull et al., 2019]
	confidence calibration	[Guo et al., 2017]
	event-conditional unbiasedness (without groupings via $\mathcal{X}$)	[Noarov et al., 2023]
Γ-Calibration	Γ-calibration (without groupings via $\mathcal{X}$)	[Noarov and Roth, 2023]
	T-calibration	[Gneiting and Resin, 2023]
	quantile calibration	[Kuleshov and Deshpande, 2022]
Decision Calibration	decision calibration	[Zhao et al., 2021]
	decision outcome indistinguishability	[Gopalan et al., 2022a]

Table 10.3: Categorization of existing notions of calibration in the three different types following by which type the notion is formally generalized. This table does not claim to contain a complete list of all definitions of calibration.

Comparable Type of Definition	Name	Reference
Distribution Calibration	distribution calibration	[Song et al., 2019]
	calibration safety	[Grünwald et al., 2020]
Γ-Calibration	marginal coverage calibration	[Räth and Ludwig, 2025]
	calibration safety	[Grünwald et al., 2020]
Decision Calibration	loss outcome indistinguishability	[Gopalan et al., 2022a]
	IP-calibration (without groupings via $\mathcal{X}$)	[Fröhlich and Williamson, 2024]

Table 10.4: Categorization of existing notions of calibration in the three different types following a broader comparability of the account. This table does not claim to contain a complete list of all definitions of calibration.

problem. What are the predictions we can or should get? What and who are the agents which use the predictions? Is one of the two accounts laid out desired as quality criterion of the predictions? Finally, our work advocates leaving the formal distinctions between evaluation metrics behind (Proposition 10.28) and re-focus on the actual purpose of evaluation: estimation of usefulness, self-realization or guarantees on the estimation of usefulness. Particularly, calibration is only a part of this bigger picture.

Acknowledgements

The authors are very grateful to Benedikt Höltgen and Rajeev Verma for feedback on an earlier draft of this work. Thanks to the International Max Planck Research School for Intelligent System (IMPRS-IS) for supporting Rabanus Derr. Thanks to James Bailie for

giving a reason to visit Cambridge, MA where Jessie and Rabanus met the first time. Part of the research happened when Rabanus was visiting Aaron Roth supported by a DAAD IFI-Stipend. Rabanus Derr and Robert C. Williamson were funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645; they were also supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center. Jessie Finocchiaro was supported in part by the National Science Foundation under Award No. 2202898. We thank the reviewers for helpful and constructive feedback.

10.9 Appendix

10.9.1 Distribution Calibration for General $\mathcal{Y}$

Our definition of distribution calibration (Definition 10.2) with respect to a property Γ asks for $\mathcal{Y}$ to be finite. This is *not* a logically required part of the definition and rather simplifies the presentation. We can define distribution calibration on arbitrary $\mathcal{Y}$ by referring to measurable sets $A \in \mathcal{B}(\mathcal{Y})$. We restrict all definitions and statements in this section to the perfect case.

Definition 10.20 (Distribution Calibration with Respect to Γ on General $\mathcal{Y}$)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$, regular wrt. $\mathcal{P}$. Suppose that $\mathcal{P} \subseteq W_{TV}$ is a subset of a Banach space equipped with the total variation distance. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. The predictor f is* distribution calibrated with respect to Γ *on D if for every $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$ and $A \in \mathcal{B}(\mathcal{Y})$,*

$$D_{Y|\Gamma \circ f(X)=\gamma}(A) = \mathbb{E}_D[f_A(X)|\Gamma \circ f(X) = \gamma]. \tag{10.4}$$

In the following we extend Proposition 10.4, Proposition 10.6, Proposition 10.8 and Proposition 10.12 to general $\mathcal{Y}$ using the new definition of distribution calibration. We summarize the extension in Table 10.5. All statements are given for perfect calibration. Note that with the additional statements of this section the implications in Figure 10.2 hold for general $\mathcal{Y}$ (Figure 10.6). To prove the implication structure we need to extend our technical machinery a bit. The technical ideas are heavily inspired by [Noarov and Roth, 2023].

Some Background on Bochner Integration In the course of the following statements we use *Bochner integration*. Bochner integration is a straight-forward generalization of Lebesgue integration. Hitherto, we were referring to distributions on finite $\mathcal{Y}$, e.g., $f(x) \in \mathcal{P} \subseteq \mathbb{R}^{|\mathcal{Y}|}$ and computed expectations such as $\mathbb{E}_{D_X}[f(X)]$. Given that $\mathcal{Y}$ is infinite, $f(x)$ is not in a Euclidean space anymore. Hence, we need to pay attention whether

Statement	Finite $\mathcal{Y}$	General $\mathcal{Y}$
Distribution calibration is inherited	Proposition 10.4	Proposition 10.21
Distribution calibration implies property calibration	Proposition 10.6	Proposition 10.22
Property calibration is inherited	Proposition 10.8	Proposition 10.23
Distribution calibration implies decision calibration	Proposition 10.12	Proposition 10.24

Table 10.5: Summary of extended statements from $\mathcal{Y}$ to general $\mathcal{Y}$.

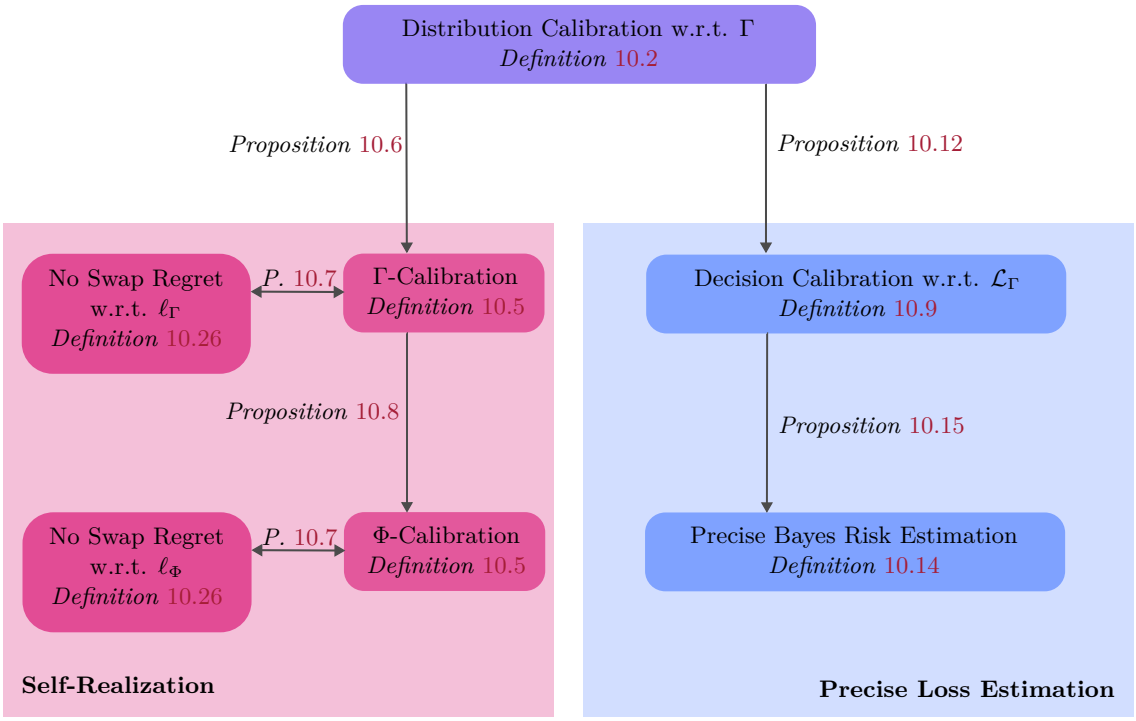


Figure 10.6: Relationships between Notions of Calibration. Implications under perfect calibration, *infinite* $\mathcal{Y}$ and elicitable property Γ and Φ. The three types of calibration are marked in different colors. The abstract accounts of calibration are shaded.

$\mathbb{E}_{D_X}[f(X)]$ is well-defined. For that reason, we assume that $\mathcal{P} \subseteq W_{TV}$, where W_{TV} is a Banach space of measures equipped with the total variation distance. By assumption f is measurable and it is easy to see that, $\mathbb{E}_D[\|f(X)\|_{TV}] \leq 1 < \infty$, hence, $\mathbb{E}_{D_X}[f(X)] \in W_{TV}$ is a well-defined Bochner integral [Diestel and Uhl, 1977, II.2 Theorem 2]. By definition we have, for all $A \in \mathcal{B}(\mathcal{Y})$,

$$\mathbb{E}_{D_X}[f(X)](A) = \mathbb{E}_{D_X}[f_A(X)],$$

where $f_A(x)$ denotes the probability assigned to the measurable set A by the probability measure $f(x)$. That is we can rewrite the distribution calibration condition (Equation 10.4)

to

$$D_{Y|\Gamma\circ f(X)=\gamma} = \mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]. \tag{10.5}$$

Proposition 10.21 (Distribution Calibration is Inherited (General $\mathcal{Y}$))**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. Suppose, that for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. If the predictor f is distribution calibrated with respect to Γ on D, then it is distribution calibrated with respect to $\Phi \coloneqq \phi \circ \Gamma$ for all $\phi\colon \mathcal{R} \to \mathcal{R}'$.*

Proof. Pick any $v \in \operatorname{supp} D_{\Phi\circ f(X)}$. By definition of the conditional distribution, we have,

$$D_{Y|\Phi\circ f(X)=v}(A) = \mathbb{E}_{G\sim D_{\Gamma\circ f(X)}}[D_{Y|\Gamma\circ f(X)=G}(A)|\phi(G) = v].$$

As f is distribution calibrated with respect to Γ on D, we have,

$$D_{Y|\Gamma\circ f(X)=G}(A) = \mathbb{E}_{X\sim D_X}[f_A(X)|\Gamma \circ f(X) = G],$$

for all $G \in \operatorname{supp} D_{\Gamma\circ f(X)}$. In particular, that is for all G such that $D_{G\sim\Gamma\circ f(X)}(\phi(G) = v) > 0$, because for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. Hence,

$$D_{Y|\Phi\circ f(X)=v}(A) = \mathbb{E}_{G\sim D_{\Gamma\circ f(X)}}[\mathbb{E}_{X\sim D_X}[f_A(X)|\Gamma \circ f(X) = G]|\phi(G) = v],$$

Finally,

$$D_{Y|\Phi\circ f(X)=v}(A) = \mathbb{E}_{X\sim D_X}[f_A(X)|\Phi \circ f(X) = v].$$

□

Proposition 10.22 (Distribution Calibration with Respect to Γ implies Γ-Calibration (General $\mathcal{Y}$))**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property with convex level sets and D a data distribution on $\mathcal{X}\times\mathcal{Y}$, regular wrt. $\mathcal{P}$. Suppose that $\mathcal{P} \subseteq W_{TV}$ is a subset of a Banach space equipped with the total variation distance. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a predictor which is distribution calibrated with respect to Γ. Then, $\Gamma \circ f$ is Γ-calibrated.*

Proof. We have to show that for all $\gamma \in \operatorname{supp} D_{\Gamma\circ f(X)}$,

$$\Gamma(D_{Y|\Gamma\circ f(X)=\gamma}) = \gamma. \tag{10.6}$$

Fix any such γ. Based on Corollary 8 in [Diestel and Uhl, 1977] reproduced in [Noarov and

Roth, 2023], we have,

$$\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma] \in \overline{\text{co}} f(\{x \in X \text{ s.t. } : \Gamma \circ f(x) = \gamma\}).$$

In particular,

$$\Gamma(\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]) \in \Gamma(\overline{\text{co}} f(\{x \in X \text{ s.t. } : \Gamma \circ f(x) = \gamma\})) = \{\gamma\},$$

because Γ has convex level sets. It follows by Equation 10.5,

$$\Gamma(D_{Y|\Gamma \circ f(X)=\gamma}) = \Gamma(\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]) = \gamma.$$

□

Proposition 10.23 (Γ-Calibration is Inherited by Refined Properties (General $\mathcal{Y}$))**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$ be a property, D a regular data distribution on $\mathcal{X} \times \mathcal{Y}$ with respect to $\mathcal{P}$. Suppose that f is a Γ-predictor which is Γ-calibrated on D. Suppose, furthermore, that for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. Then, for every property Φ with convex level sets and which is refined by Γ, the Φ-predictor $\phi \circ f$, where ϕ is defined following Definition 10.3, is Φ-calibrated on D.*

Proof. We have to show that,

$$\Phi(D_{Y|\phi \circ f(X)=v}) = \phi \circ \Gamma(D_{Y|\phi \circ f(X)=v}) = v,$$

for all $v \in \operatorname{supp} D_{\phi \circ f(X)}$. Fix $v \in \operatorname{supp} D_{\phi \circ f(X)}$ for the following. Let us define the probabilistic predictor $h \colon \mathcal{X} \to \mathcal{P}$ where $x \mapsto D_{Y|f(X)=f(x)}$. The mapping is measurable by definition, as f is measurable by assumption. Let $A \coloneqq \{x \in X \colon \phi \circ f(x) = v\}$. Since, f is Γ-calibrated, and for all $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$, it holds,

$$\Gamma(h(x)) = \Gamma(D_{Y|f(X)=f(x)}) = f(x)$$

for all $x \in A$. Hence, in particular,

$$\Phi(h(x)) = v,$$

for all $x \in A$. By the law of total expectation, we obtain, for all $B \in \mathcal{B}(\mathcal{Y})$,

$$\begin{aligned} D_{Y|\phi \circ f(X)=v}(Y \in B) &= \mathbb{E}_D[[\![Y \in B]\!]|\phi(f(\tilde{X})) = v] \\ &= \mathbb{E}_{G \sim D_{f(X)}}\left[\mathbb{E}_D[[\![Y \in B]\!]|f(X) = G]|\phi(G) = v\right] \\ &= \mathbb{E}_{G \sim D_{f(X)}}\left[\mathcal{D}_{Y|f(X)=G}(Y \in B)|\phi(G) = v\right]. \end{aligned}$$

It follows by definition of h, A and properties of the Bochner integral, $D_{Y|\phi\circ f(X)=v} = \mathbb{E}_{D_X}[h(X)|X \in A]$. Based on Corollary 8 in [Diestel and Uhl, 1977] reproduced in [Noarov and Roth, 2023], we have,

$$\mathbb{E}_D[h(X)|X \in A] \in \overline{\text{co}}h(A).$$

In particular,

$$\Phi(\mathbb{E}_D[h(X)|X \in A]) \in \Phi(\overline{\text{co}}(A) = \{v\},$$

because Φ has convex level sets. □

Proposition 10.24 (Distribution Calibration Implies Decision Calibration (General $\mathcal{Y}$))**.** *Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$, regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor which is distribution calibrated with respect to Γ on D. Then, f is decision calibrated with respect to $\mathcal{L}_\Gamma \coloneqq \{\ell\colon \ell$ is a bounded, $\mathcal{P}$-consistent loss function for $\Gamma\}$.*

Proof. Let $\ell \in \mathcal{L}_\Gamma$ be arbitrary. Because ℓ is bounded, the following expectations are well-defined. For every $\gamma \in \operatorname{supp} D_{\Gamma\circ f(X)}$, we have,

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X) = \gamma\right] \\
&= \mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma)|\Gamma\circ f(X) = \gamma\right] - \mathbb{E}_{(X,Y)\sim D}\left[\mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X) = \gamma\right] \\
&= \int_{y\in\mathcal{Y}} \ell(y,\gamma) dD_{Y|\Gamma\circ f(X)=\gamma} - \mathbb{E}_{X\sim D_X}\left[\int_{y\in\mathcal{Y}} \ell(y,\gamma) df(X) \,\middle|\, \Gamma\circ f(X) = \gamma\right] \\
&= 0,
\end{aligned}$$

The last equality holds, because, by [Diestel and Uhl, 1977, II.2 Theorem 6] and distribution calibration,

$$\begin{aligned}
\mathbb{E}_{X\sim D_X}\left[\int_{y\in\mathcal{Y}} \ell(y,\gamma) df(X) \,\middle|\, \Gamma\circ f(X) = \gamma\right] &= \int_{y\in\mathcal{Y}} \ell(y,\gamma) d\mathbb{E}_{X\sim D_X}\left[f(X) \mid \Gamma\circ f(X) = \gamma\right] \\
&= \int_{y\in\mathcal{Y}} \ell(y,\gamma) dD_{Y|\Gamma\circ f(X)=\gamma}.
\end{aligned}$$

It follows, legitimately ignoring $\gamma \in \text{im}\Gamma\circ f$ such that $D_X(\Gamma\circ f(X) = \gamma) = 0$,

$$\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\Gamma(f(X))) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\Gamma(f(X)))]\right] = 0.$$

□

10.9.2 Approximate Calibration

In the above, we left approximate versions of calibration relatively untouched. In the following section, we extend several statements beyond perfect calibration. Figure 10.7 summarizes the findings of this appendix analogously to Figure 10.2 for the perfect calibration case.

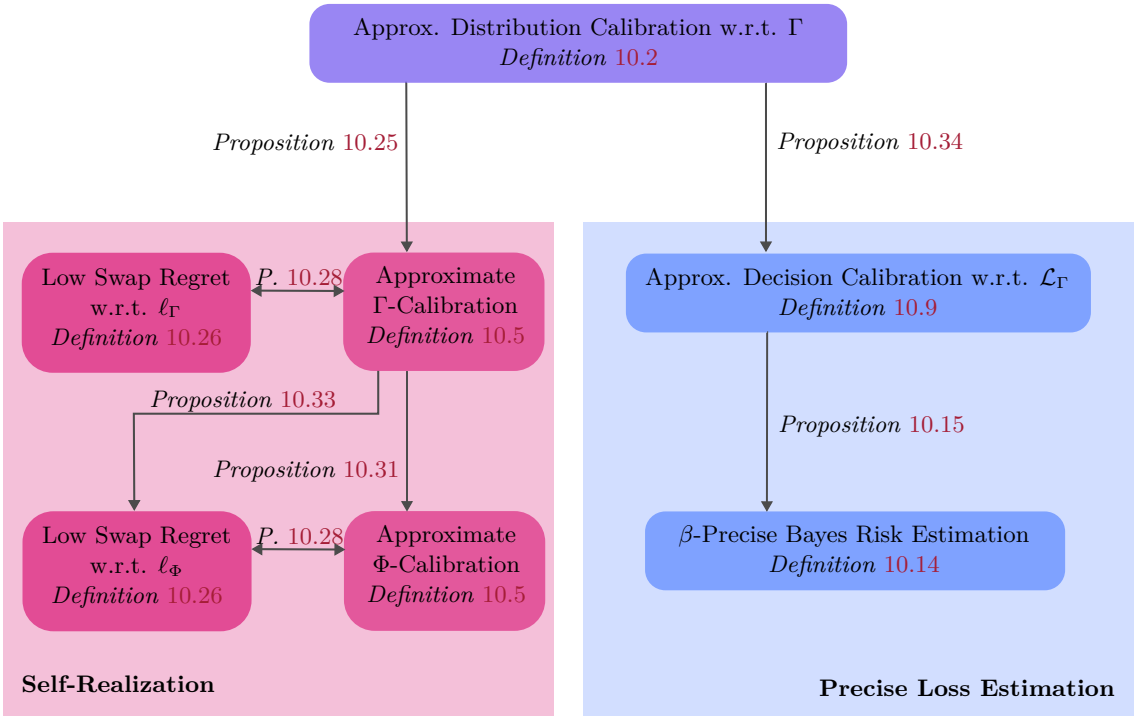


Figure 10.7: Relationship between Approximate Notions of Calibration. Implications Under Approximate Calibration and Finite $\mathcal{Y}$. Further conditions are stated in the referenced propositions. Those conditions particularly contain Lipschitz and Smoothness assumptions. The three types of calibration are marked in different colors. The abstract accounts of calibration are shaded.

Approximate Distribution Calibration Implies Approximate Γ-Calibration

In Proposition 10.6 we have shown that distribution calibration implies property calibration for all properties with convex level sets. We extend this statement to approximate versions of calibration in the following. To this end, we assume that the property Γ is Lipschitz continuous.

Proposition 10.25 (Approximate Distribution Calibration w.r.t. Γ implies approximate Γ-Calibration for Lipschitz continuous Γ)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ with $\mathcal{R} \subseteq \mathbb{R}$ be a K-Lipschitz property with convex level sets and D a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite,*

regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor that is $\alpha(\gamma)$-approximately distributionally calibrated with respect to Γ. Then, $\Gamma \circ f$ is $|\mathcal{Y}|K\alpha(\gamma)$-approximately Γ-calibrated.

Proof. Since Γ is real-valued, the metric m is the absolute difference. We have to show that for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

$$|\Gamma(D_{Y|\Gamma \circ f(X)=\gamma}) - \gamma| \leq |\mathcal{Y}|K\alpha(\gamma). \tag{10.7}$$

Pick any $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$. Lemma 10.42 applied to D, f and $A \coloneqq \{x \in \mathcal{X}\colon \Gamma \circ f(x) = \gamma\}$ gives,

$$\Gamma(\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]) = \gamma.$$

Second, the total variation distance between $D_{Y|\Gamma \circ f(X)=\gamma}$ and $\mathbb{E}_D[f(X)|\Gamma \circ f(X) = \gamma]$ is bounded above for all $\gamma \in \operatorname{supp} D_{\Gamma \circ f(X)}$,

$$\sup_{A \subseteq \mathcal{Y}} \left| \sum_{y \in A} D_{Y|\Gamma \circ f(X)=\gamma}(Y = y) - \mathbb{E}_D[f_y(X)|\Gamma \circ f(X) = \gamma] \right| \leq |\mathcal{Y}|\alpha(\gamma)$$

where $f_y(X)$ denotes the y-component of the prediction $f(X) \in \Delta(\mathcal{Y})$. This follows from the fact that f is $\alpha(\gamma)$-approximate distribution calibrated with respect to Γ. Hence, by applying Γ and exploiting the Lipschitzness of Γ, we obtain the desired Equation (10.7). □

Approximate Γ-Calibration is Equivalent to Swap Regret

Γ-calibration is equivalent to swap regret not only in the perfect case, but as well in the approximate case. To this end, we formally introduce a distributional definition of swap regret similar in spirit to [Cesa-Bianchi and Lugosi, 2006, p. 91] and introduce additional regularity assumptions on Γ and its identification function V.

Definition 10.26 (Swap Regret)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be an elicitable property with $\mathcal{P}$-consistent scoring function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let $f\colon \mathcal{X} \to \mathcal{R}$ be a Γ-predictor. The Γ-predictor f has $\beta(\gamma)$-bounded* swap regret *on D if for every $\gamma \in \operatorname{supp} D_{f(X)}$,*

$$\mathbb{E}_D\left[\ell(Y, \gamma)|f(X) = \gamma\right] - \min_{\hat{\gamma} \in \operatorname{im}\Gamma} \mathbb{E}_D\left[\ell(Y, \hat{\gamma})|f(X) = \gamma\right] \leq \beta(\gamma).$$

One of the crucial insights in [DeGroot and Fienberg, 1983] was that mean-calibration re-

lates to swap regret for the squared loss function.[24] In this section, we more generally show that if the property, which ought to be calibrated, is identifiable with a certain regular identification function, approximate calibration is equivalent to low swap regret. In particular, this extends the equivalence spelled out in Proposition 10.7 to the approximate case in Proposition 10.28, which requires certain regularity conditions on the identification function given in Definition 10.27.

Definition 10.27 (Conditions on Identification Functions)**.** *Let $\mathcal{R} \subseteq \mathbb{R}$ be an interval and $\Gamma\colon \mathcal{P} \to \mathcal{R}$ an identifiable property with identification function $V\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$. Let $\gamma, \gamma' \in \mathrm{im}\Gamma$. The identification function V is called*

oriented *if and only if for all $P \in P$, $V(P, \gamma) > 0 \Leftrightarrow \gamma > \Gamma(P)$.*

locally non-constant *if and only if there exists $N > 0$ such that for all $P \in \mathcal{P}$, $N|\gamma - \Gamma(P)| \leq |V(P, \gamma)|$.*

locally Lipschitz *if and only if there exists $M > 0$ such that for all $P \in \mathcal{P}$, $|V(P, \gamma)| \leq M|\gamma - \Gamma(P)|$.*

Let us shortly discuss the regularity conditions. Orientedness is arguably a weak assumption. Finocchiaro and Frongillo [2018] have shown that in finite dimensions, i.e., $|\mathcal{Y}| < \infty$, there exists a reweighting of any identification function, such that the reweighted identification function is oriented and identifies the same property.[25] Furthermore, we obtain orientedness for free in the main theorem of [Steinwart et al., 2014]. The non-constantness and Lipschitzness assumptions have been considered, in its non-local variant, in the context of composite properties [Noarov and Roth, 2023, Assumption 5.2 and 5.3]. Note that $N \leq M$ by definition, which intuitively captures the idea that an identification function cannot be more non-constant than Lipschitz-smooth. In Appendix 10.9.5 we provide several examples of properties beyond the mean which have identification functions which fulfill all of the above properties.

Proposition 10.28 (Approximate Γ-Calibration Equivalent to Low Swap Regret)**.** *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ with $\mathcal{R} \subseteq \mathbb{R}$ be an identifiable property with oriented, bounded identification function $V\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ measurable in both its inputs with respect to the standard Borel-σ-algebra on the sets $\mathcal{Y}, \mathcal{R}$ and $\mathbb{R}$. Then, Γ is elicitable with $\mathcal{P}$-consistent scoring function $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$,*

$$\ell(y, \gamma) \coloneqq \int_{\gamma_0}^{\gamma} V(y, r)dr + \kappa(y), \tag{10.8}$$

[24] Globus-Harris et al. [2023] and Gopalan et al. [2024b] extend this relationship to multi-calibration and swap agnostic learners.

[25] Finocchiaro and Frongillo [2018] show that the reweighted identification function is monotone increasing, which implies orientedness.

for some $\gamma_0 \in \mathrm{im}\Gamma$ *and* $\kappa\colon \mathcal{Y} \to \mathbb{R}$ *having a finite expected value with respect to all* $P \in \mathcal{P}$. *Let* f *be a* Γ*-predictor and* D *a regular data distribution.*

1. *Suppose* V *is locally Lipschitz on* $\mathcal{P}$ *with parameter* M. *If* f *is* $\alpha(\gamma)$*-approximately* Γ*-calibrated, then* f *has* $\beta(\gamma) \coloneqq \frac{M}{2}\alpha(\gamma)^2$*-bounded swap regret.*
2. *Suppose* V *is locally non-constant on* $\mathcal{P}$ *with parameter* N. *If* f *has* $\beta(\gamma)$*-bounded swap regret, then* f *is* $\alpha(\gamma) \coloneqq \sqrt{\frac{2}{N}\beta(\gamma)}$*-approximate calibrated.*

Proof. Lemma 10.29 does almost all of the heavy lifting. It remains to apply Equation 10.10 to $D_{Y|f(X)=\gamma}$ for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$|\ell(D_{Y|f(X)=\gamma}, \gamma) - \ell(D_{Y|f(X)=\gamma}, \Gamma(D_{Y|f(X)=\gamma}))| \leq \frac{M}{2}\left(\gamma - \Gamma(D_{Y|f(X)=\gamma})\right)^2 \leq \frac{M}{2}\alpha(\gamma)^2,$$

to obtain the first statement. For the second statement, we rewrite Equation 10.11 for $D_{Y|f(X)=\gamma}$ for all $\gamma \in \operatorname{supp} D_{f(X)}$,

$$|\gamma - \Gamma(D_{Y|f(X)=\gamma})| \leq \sqrt{\frac{2}{N}|\ell(D_{Y|f(X)=\gamma}, \gamma) - \ell(D_{Y|f(X)=\gamma}, \Gamma(D_{Y|f(X)=\gamma}))|} \leq \sqrt{\frac{2}{N}\beta(\gamma)}.$$

□

Lemma 10.29 (Identification Function Defines Consistent Loss Functions)**.** *Let* $\Gamma\colon \mathcal{P} \to \mathcal{R}$ *with* $\mathcal{R} \subseteq \mathbb{R}$ *be an identifiable property with oriented, bounded identification function* $V\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ *as in Proposition 10.28. Then,* Γ *is elicitable with* $\mathcal{P}$*-consistent scoring function* $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$,

$$\ell(y, \gamma) \coloneqq \int_{\gamma_0}^{\gamma} V(y, r)dr + \kappa(y), \tag{10.9}$$

for some $\gamma_0 \in \mathrm{im}\Gamma$ *and* $\kappa\colon \mathcal{Y} \to \mathbb{R}$ *having a finite expected value with respect to all* $P \in \mathcal{P}$.

(a) If furthermore, V *is locally Lipschitz with parameter* M*, then* ℓ *is locally Hölder-smooth with parameters* $(\frac{M}{2}, 2)$*, i.e., for all* $\gamma \in \mathcal{R}$,

$$|\ell(P, \gamma) - \ell(P, \Gamma(P))| \leq \frac{M}{2}\left(\gamma - \Gamma(P)\right)^2. \tag{10.10}$$

(b) If furthermore, V *is locally non-constant with parameter* N*, then* ℓ *is locally anti Hölder-smooth with parameters* $(\frac{N}{2}, 2)$*, i.e., for all* $\gamma \in \mathcal{R}$,

$$\frac{N}{2}\left(\gamma - \Gamma(P)\right)^2 \leq |\ell(P, \gamma) - \ell(P, \Gamma(P))|. \tag{10.11}$$

Proof. The first statement has been proved in [Steinwart et al., 2014]. We provide a reiteration of the argument for the sake of completeness, and because we will reuse some equations for the additional statements.

For any $P \in \mathcal{P}$, we want to show that $\mathbb{E}_{Y\sim P}[\ell(Y,\gamma)] > \mathbb{E}_{Y\sim P}[\ell(Y,\Gamma(P))]$ for all $\gamma \in \mathcal{R}$, $\gamma \neq \Gamma(P)$. Let us first consider the case that $\gamma > \Gamma(P)$. Hence,

$$\begin{aligned}
&\mathbb{E}_{Y\sim P}[\ell(Y,\gamma)] - \mathbb{E}_{Y\sim P}[\ell(Y,\Gamma(P))] \\
&= \mathbb{E}_{Y\sim P}[\ell(Y,\gamma) - \ell(Y,\Gamma(P))] \\
&= \mathbb{E}_{Y\sim P}\left[\int_{\Gamma(P)}^{\gamma} V(y,r)dr\right].
\end{aligned}$$

Since V is measurable and bounded in both variables we can apply Fubini's theorem, this gives

$$\mathbb{E}_{Y\sim P}\left[\int_{\Gamma(P)}^{\gamma} V(y,r)dr\right] = \int_{\Gamma(P)}^{\gamma} \mathbb{E}_{Y\sim P}[V(y,r)]dr, \tag{10.12}$$

finally, since V is oriented,

$$\int_{\Gamma(P)}^{\gamma} \mathbb{E}_{Y\sim P}[V(y,r)]dr = \int_{\Gamma(P)}^{\gamma} V(P,r)dr > 0.$$

The case $\gamma < \Gamma(P)$ follows analogously.

(a) Let $\gamma > \Gamma(P)$, then,

$$\begin{aligned}
|\ell(P,\gamma) - \ell(P,\Gamma(P))| &= \left|\int_{\Gamma(P)}^{\gamma} V(P,r)dr\right| \\
&= \int_{\Gamma(P)}^{\gamma} |V(P,r)|dr \\
&\leq \int_{\Gamma(P)}^{\gamma} M|r-\Gamma(P)|dr \\
&= \frac{M}{2}(\gamma-\Gamma(P))^2.
\end{aligned}$$

The analogous computation with swapped signs holds for $\gamma < \Gamma(P)$.

(b) Let $\gamma > \Gamma(P)$, then,

$$\begin{aligned}|\ell(P,\gamma) - \ell(P,\Gamma(P))| &= \left|\int_{\Gamma(P)}^{\gamma} V(P,r)dr\right| \\ &= \int_{\Gamma(P)}^{\gamma} |V(P,r)|dr \\ &\geq \int_{\Gamma(P)}^{\gamma} N|r - \Gamma(P)|dr \\ &= \frac{N}{2}(\gamma - \Gamma(P))^2.\end{aligned}$$

The analogous computation with swapped signs holds for $\gamma < \Gamma(P)$.

□

Hence, for the right choice of loss function, Γ-calibration and low swap regret can be used equivalently. Along the lines of [Gopalan et al., 2024b, Theorem 3.3], we extend the equivalence relationship to group-wise definitions of calibration via [Noarov and Roth, 2023, Definition 2.10] in Appendix 10.9.6. However, we do *not* achieve a direct generalization of [Gopalan et al., 2024b, Theorem 3.3 (1.) swap multicalibration $\iff$ (3.) swap-agnostic learner] by going from the mean (as in their work) to more general identifiable properties. Hence, we provide a bridge from calibration for general identifiable properties to swap regret notions, going beyond [Noarov and Roth, 2023]. Nevertheless, the mean is a good example to illustrate Proposition 10.28.

Example 10.30. *Let Γ be the mean and $V(y,\gamma) \coloneqq \gamma - y$ its identification function. In this case, $N = M = 1$. The mean is elicited by the squared loss,*

$$\ell(y,\gamma) = \frac{1}{2}(y-\gamma)^2 = \int_0^{\gamma} V(y,r)dr + \frac{1}{2}y^2.$$

Hence, a Γ-predictor f on a regular data distribution D has $\frac{1}{2}\alpha(\gamma)^2$ swap regret if, and only if it is $\alpha(\gamma)$-approximately calibrated (cf. [Gopalan et al., 2024b, Theorem 3.3]).

Inheritance of Approximate Γ-Calibration

For the approximate analogue of Proposition 10.8 we require the refined property to be Lipschitz.

Proposition 10.31 (Approximate Γ-Calibration is Inherited by Refined Properties)**.** *Let Γ be a property, D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$, f a Γ-predictor and $\Phi = \phi \circ \Gamma$ a property refined by Γ with convex level sets. Assume further that ϕ is Lipschitz*

(with constant K), $\mathcal{Y}$ is finite and for every $x \in \mathcal{X}$, $f(x) \in \operatorname{supp} D_{f(X)}$. If f is $\alpha(\gamma)$-approximately Γ-calibrated on D, then $\phi \circ f$ is $C\alpha'(\phi(v))$-approximately Φ-calibrated, where $\alpha'(v) := \sup_{\gamma \in \phi^{-1}(v)}$.

Proof. It is given that,

$$|\Gamma(D_{Y|f(X)=\gamma}) - \gamma| \leq \alpha(\gamma),$$

from which simply follows,

$$|\phi \circ \Gamma(D_{Y|f(X)=\gamma}) - \phi(\gamma)| \leq K\alpha(\gamma).$$

Then, by Proposition 10.8,

$$|\phi \circ \Gamma(D_{Y|\phi \circ f(X)=v}) - v| \leq K\alpha'(v),$$

where $\alpha'(v) = \sup_{\gamma \in \phi^{-1}(v)} \alpha(\gamma)$. □

The Lipschitz-condition on the property is strong. For instance, it rules out discrete properties. However, we will argue that in the case of elicitable properties we can still give guarantees about self-realization in terms of low swap regret (Proposition 10.33). This statement requires an arguably mild regularity condition on the loss function.

Definition 10.32 (Locally Outcome Lipschitz Loss Function)**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$. Suppose a loss function $\ell \colon \mathcal{Y} \times \mathcal{R}' \to \mathbb{R}$ elicits a property $\Phi \colon \mathcal{P} \to \mathcal{R}'$. It is called* locally outcome Lipschitz with respect to Γ *with constant B, if for all $v \in \operatorname{im}\Phi$, $\gamma \in \operatorname{im}\Gamma$, $P \in \Gamma^{-1}(\gamma)$ and $P' \in \mathcal{P}$,*

$$|\ell(P, v) - \ell(P', v)| \leq Bm(\Gamma(P'), \gamma). \tag{10.13}$$

If $\mathcal{Y}$ is finite, Γ is the full distribution, i.e., $\mathcal{R} = \Delta(\mathcal{Y})$ and ℓ a bounded function, then ℓ is locally outcome Lipschitz (Appendix, Lemma 10.43). Furthermore, the assumptions made in [Gopalan et al., 2024b] imply Definition 10.32 for Γ being the mean on a binary outcome set.

Proposition 10.33 (Low Swap Regret Guarantee for Refined Properties)**.** *Let $\Gamma \colon \mathcal{P} \to \mathcal{R}$ be a property, D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$ and f a Γ-predictor which is Γ-calibrated on D. Let $\Phi \colon \mathcal{P} \to \mathcal{R}'$ be a property which is refined by Γ, i.e., $\Phi \coloneqq \phi \circ \Gamma$ for some $\phi : \mathcal{R} \to \mathcal{R}'$.*

If Φ is elicitable with loss function ℓ, locally outcome Lipschitz with respect to Γ with constant B, and if f is $\alpha(\gamma)$-approximately Γ-calibrated, then the swap regret of $\phi \circ f$ with respect to

ℓ is bounded above by $2B\mathbb{E}_{\gamma\sim D_{f(X)}}\left[\alpha(\gamma)|\phi\circ\gamma=v\right]$ for all $v\in\operatorname{supp}D_{\phi\circ f(X)}$.

Proof. Let $P\in\Gamma^{-1}(\gamma)$, by Definition 10.32 Equation (10.13), for all $\gamma\in\operatorname{supp}D_{f(X)}$,

$$\begin{aligned}
&\mathbb{E}_D\left[\ell(Y,\phi(\gamma))-\ell(Y,\Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)})|f(X)=\gamma\right]\\
&=\ell(D_{Y|f(X)=\gamma},\phi(\gamma))-\ell(D_{Y|f(X)=\gamma},\Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)}))\\
&=\ell(D_{Y|f(X)=\gamma},v)-\ell(P,\Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)}))\\
&\quad+\ell(P,\Phi(D_{Y|\phi\circ f(X)=v}))-\ell(D_{Y|f(X)=\gamma},\Phi(D_{Y|\phi\circ f(X)=v}))\\
&\leq\ell(D_{Y|f(X)=\gamma},v)-\ell(P,\Phi(D_{Y|\phi\circ f(X)=v}))+Bm(\Gamma(D_{Y|f(X)=\gamma}),\gamma)|\\
&\leq\ell(D_{Y|f(X)=\gamma},v)-\ell(P,v)+\ell(P,v)-\ell(P,\Phi(D_{Y|\phi\circ f(X)=v}))\\
&\quad+Bm(\Gamma(D_{Y|f(X)=\gamma}),\gamma)\\
&\leq\ell(P,v)-\ell(P,\Phi(D_{Y|\phi\circ f(X)=v}))+2Bm(\Gamma(D_{Y|f(X)=\gamma}),\gamma)\\
&\leq 2B\alpha(\gamma).
\end{aligned}$$

It follows that, for all $v\in\operatorname{supp}D_{\phi\circ f(X)}$,

$$\begin{aligned}
&\mathbb{E}_D\left[\ell(Y,\phi(\gamma))-\ell(Y,\Phi(D_{Y|\phi\circ f(X)=\phi(\gamma)})|\phi\circ f(X)=v\right]\\
&\leq 2B\mathbb{E}_{\gamma\sim D_{f(X)}}\left[\alpha(\gamma)|\phi\circ f(X)=v\right].
\end{aligned}$$

□

An analogous proof holds for distributional calibration with respect to Γ and a locally outcome Lipschitz loss function with respect to Γ. The proof is an adapted version of parts of the proof of the main theorem in [Gopalan et al., 2024b].

The inheritance of self-realization is closely connected to "(swap) omniprediction" introduced by [Gopalan et al., 2022b, 2024b]. Informally, a (swap) omnipredictor provides predictions which allow for a simple post-processing to achieve low regret against a comparison base line for a large variety of losses. In particular, (swap) omnipredictors involve "for all" quantifiers about a set of losses.

Non-Refining (Swap) Omniprediction Proposition 10.28 already involves an "omni"-type statement. The statement holds for all loss functions defined following Equation (10.8). Yet, there is no property refinement necessary.

Refining (Swap) Omniprediction The inheritance rule makes use of the refinement function ϕ. This refinement function generalizes the post-processing function k_ℓ in [Gopalan et al., 2022b]. In particular, Proposition 10.33 can be plugged together for a set of loss

functions. For instance, let f be a Γ-calibrated Γ-predictor on a data distribution D. Then, those predictions guarantee low swap regret for all loss functions which are locally outcome Lipschitz with respect to Γ. Hence, this is an "omni"-type regret statement, related to the main theorem in [Gopalan et al., 2024b]. In there, the authors focused on a very specific set of losses which are convex, Lipschitz and locally outcome Lipschitz with respect to the mean. However, their main theorem is generalized to multi-calibration with respect to groups. We push aside this further complexity in this work (cf. Section 10.7). Hence, our statement is less general in the scope of groupings, but more general in the scope of sets of loss functions and predicted properties.[26]

Approximate Distribution Calibration Implies Approximate Decision Calibration

Subject to mild assumptions on the boundedness of the loss function ℓ which elicits the property of interest Γ, Proposition 10.12 can be generalized from perfect to approximate calibration.

Proposition 10.34 (Approximate Distribution Calibration Implies Approximate Decision Calibration)**.** *Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite, regular wrt. $\mathcal{P}$ Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor $\alpha(\gamma)$-approximately distribution calibrated with respect to Γ on D. Then, f is $C\mathbb{E}_{\gamma\sim D_{\Gamma\circ f(X)}}[\alpha(\gamma)]$-approximate decision calibrated with respect to $\mathcal{L}_{\Gamma,C} \coloneqq \{\ell\colon \ell$ is $\mathcal{P}$-consistent loss function for Γ and $\sup_{r\in\mathcal{R}} \sum_{y\in\mathcal{Y}} |\ell(y,r)| \leq C < \infty\}$.*

Proof. The predictor f is $\alpha(\gamma)$-*approximate distribution calibrated with respect to* Γ on D if for every $\gamma \in \operatorname{supp} D_{\Gamma\circ f(X)}$,

$$\left|D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) - \mathbb{E}_D[f_y(X)|\Gamma\circ f(X)=\gamma]\right| \leq \alpha(\gamma), \quad \forall y \in \mathcal{Y},$$

where $f_y(x) \in [0,1]$ denotes the y-component of the prediction $f(x) \in \Delta(\mathcal{Y})$ for $x \in \mathcal{X}$.

Let $\ell \in \mathcal{L}_{\Gamma,C}$ be arbitrary but fixed. For all $\gamma \in \operatorname{supp} D_{\Gamma\circ f(X)}$, we have,

$$\begin{aligned}
&\left|\mathbb{E}_{(X,Y)\sim D}\left[\ell(Y,\gamma) - \mathbb{E}_{\hat{Y}\sim f(X)}[\ell(\hat{Y},\gamma)]|\Gamma\circ f(X)=\gamma\right]\right| \\
&= \left|\sum_{y\in\mathcal{Y}} \left(D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) - \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma\circ f(X)=\gamma]\right)\ell(y,\gamma)\right| \\
&\leq \sum_{y\in\mathcal{Y}} \left|D_{Y|\Gamma\circ f(X)=\gamma}(Y=y) - \mathbb{E}_{X\sim D_X}[f_y(X)|\Gamma\circ f(X)=\gamma]\right| |\ell(y,\gamma)| \\
&\leq \alpha(\gamma)\sum_{y\in\mathcal{Y}} |\ell(y,\gamma)| \leq C\alpha(\gamma).
\end{aligned}$$

[26] Note that [Lu et al., 2025] as well consider generalizations of the set of loss functions.

The result follows by taking the expectation over $\gamma \sim D_{\Gamma \circ f(X)}$. For detailed steps from the first to the second term see the proof of Proposition 10.12. □

10.9.3 Distribution Calibration with Respect to all Binary Properties Implies Full Distribution Calibration

It follows relatively immediately that distribution calibration with respect to the full distribution implies distribution calibration with respect to any property. The reverse direction requires one to think about a *set* of properties. We show that if a forecast is perfectly distribution calibrated with respect to all binary, elicitable properties, it is distribution calibrated with respect to the full distribution. This statement requires some technical-looking assumption on separability via hyperplanes. If the number of different predicted forecasts is finite, which is arguably a mild assumption in practice, then this separability assumption holds.

Lemma 10.35. *If $|\mathrm{im} f| < \infty$, then for every $p \in \mathrm{im} f$ there exists a hyperplane $H = \{x \in \mathcal{R}^{|\mathcal{Y}|} \colon \langle a, x \rangle = b\}$ such that $p \in H$ and $0 < \epsilon \leq \inf_{h \in H} \|h - q\|$ for every $q \in \mathrm{im} f \setminus \{p\}$.*

Proof. The proof follows by contradiction. Let us assume that for some point $p \in \mathrm{im} f$ there is no hyperplane such that $p \in H$ and $0 < \epsilon \leq \inf_{h \in H} \|h - q\|$ for every $q \in \mathrm{im} f \setminus \{p\}$. That is, for every hyperplane H such that $p \in H$ there is $q_H \in \mathrm{im} f \setminus \{p\}$ with $\inf_{h \in H} \|h - q\| < \epsilon$ for every $\epsilon > 0$. It follows that such $q_H \in H$, because $\inf_{h \in H} \|h - q\| = 0$. Now, since there are uncountably many such hyperplanes H but $|\mathrm{im} f| < \infty$ we get into a contradiction. □

Proposition 10.36 (Recovering Distribution Calibration by Distribution Calibration with respect to all Binary, Elicitable Properties)**.** *Let D be a regular data distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is finite. Let $f \colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor. Suppose that $p_D(f(X) = q) > 0$ for all $q \in \mathrm{im} f$.*

If the predictor f is distribution calibrated with respect to all binary elicitable properties $\Phi \colon \Delta(\mathcal{Y}) \to \{0, 1\}$, then it is distribution calibrated with respect to the full distribution on D.

Proof. Let us pick an arbitrary $p \in \mathrm{im} f$. We show that there exist two elicitable, binary properties $\underline{\Phi}_p, \overline{\Phi}_p$ such that,

$$\underline{\Phi}_p(q) = \overline{\Phi}_p(q), \quad \forall q \in \mathrm{im} f \setminus \{p\}$$
$$\underline{\Phi}_p(p) \neq \overline{\Phi}_p(p).$$

Then, we can argue that there is $a \in \{0, 1\}$ such that

$$\mathbb{E}_D[[\![Y = y]\!] - f_y(X)|\underline{\Phi}_p \circ f(X) = a] = 0$$
$$\mathbb{E}_D[[\![Y = y]\!] - f_y(X)|\overline{\Phi}_p \circ f(X) = a] = 0,$$

where, with some $c \in [0, 1]$,

$$\begin{aligned}&\mathbb{E}_D[[\![Y = y]\!] - f_y(X)|\overline{\Phi}_p \circ f(X) = a] \\ &= c\mathbb{E}_D[[\![Y = y]\!] - f_y(X)|\underline{\Phi}_p \circ f(X) = a] + p_D(f(X) = p)\mathbb{E}_D[[\![Y = y]\!] - f_y(X)|f(X) = p].\end{aligned}$$

Hence, because $p_D(f(X) = p) > 0$,

$$\mathbb{E}_D[[\![Y = y]\!] - f_y(X)|f(X) = p] = 0.$$

Since $p \in \mathrm{im}f$ was picked arbitrarily, this shows the claim.

Nevertheless, it remains to argue that there exist such two elicitable, binary properties $\underline{\Phi}_p, \overline{\Phi}_p$ as demanded above. We show this by the prooving the existence of two hyperplanes such that all points $q \in \mathrm{im}f$ except for p stay on the same side of both hyperplanes, but p switches the side. As proven in [Lambert, 2019, Theorem 1], properties are elicitable if and only if the level sets form a power diagram, i.e., the intersection of the simplex $\Delta(\mathcal{Y}) \subseteq \mathbb{R}^{|\mathcal{Y}|}$ with a Voronoi diagram. For our purpose, it suffices to know that a hyperplane in $\mathbb{R}^{|\mathcal{Y}|}$ defines a binary, elicitable property. In fact, every binary elicitable property can be expressed through a corresponding hyperplane. Let

$$\underline{\Phi}_p(q) = \begin{cases} 0 \text{ if } \langle a, q\rangle > b - \frac{\epsilon\|a\|}{2} \\ 1 \text{ otherwise.} \end{cases}.$$

and

$$\overline{\Phi}_p(q) = \begin{cases} 0 \text{ if } \langle a, q\rangle > b + \frac{\epsilon\|a\|}{2} \\ 1 \text{ otherwise.} \end{cases}.$$

Then,

$$\underline{\Phi}_p(p) = 0 \neq 1 = \overline{\Phi}_p(p),$$

but for all $q \in \mathrm{im}f \setminus \{p\}$,

$$\underline{\Phi}_p(p) = \overline{\Phi}_p(p),$$

because if $\underline{\Phi}_p(q) = 0$ (the analogous argument holds for $\underline{\Phi}_p(q) = 1$), then

$$\langle a, q \rangle > b - \frac{\epsilon \|a\|}{2},$$

which implies,

$$\langle a, q \rangle \geq b + \|a\|\epsilon > b + \frac{\epsilon \|a\|}{2}$$

by formula of distance to a hyperplane,

$$\epsilon \leq \inf_{h \in H} \|h - q\| = \frac{|\langle a, q \rangle - b|}{\|a\|} \qquad \Leftrightarrow \|a\|\epsilon + b \leq \langle a, q \rangle.$$

□

10.9.4 A Note on Sensibility for Calibration

The reader familiar with [Noarov and Roth, 2023] might question the use of a general property Γ in Definition 10.5. The authors introduce "sensibility" for calibration which requires that the true property predictor is calibrated with respect to Definition 10.5. As the authors show therein, a property is only sensible for calibration if and only if the level sets (pre-images) of property values are convex. Steinwart et al. [2014] in turn, have proven that if the property is strictly locally non-constant, a minor technical assumption, and continuous, then convexity of the level sets of the property is equivalent to elicitability (respectively identifiability). If the property is discrete and its image finite, and the outcome set is finite, then convex level sets are not sufficient for the elicitability of Γ [Lambert, 2019, Pg. 12].[27] Hence, even if we require Γ to be sensible to calibration, Definition 10.5 extends to discrete properties with convex level sets, which are potentially neither identifiable nor elicitable. This justifies the definition of Γ-calibration with respect to any Γ with convex level sets. Furthermore, even if a property Γ is not sensible for calibration, a Γ-predictor could still be Γ-calibrated. However, in this case optimal individual prediction via the true property predictor and calibration are *not* compatible. Nevertheless, Γ-calibration is still a fulfillable criterion.

10.9.5 Examples of Properties with Regular Identification Functions

A priori it is unclear whether there exist interesting, non-trivial, identifiable properties with identification functions fulfilling the assumptions put forward in Definition 10.27. For this reason we provide a short list of examples.

[27] Instead the level sets have to form a Voronoi diagram [Lambert, 2019, Theorem 1].

Example 10.37. Mean *An easy example for a property with a monotone, locally non-constant, locally Lipschitz identification function is the mean, with identification function* $V(y, \gamma) \coloneqq (y - \gamma)$*, because*

$$\mathbb{E}_{Y \sim P}[(Y - \gamma)] = \Gamma(P) - \gamma.$$

Quantiles *One identification function for a* τ*-quantile is* $V_\tau(y, \gamma) \coloneqq (1-\tau)[\![y < \gamma]\!] - \tau[\![y > \gamma]\!]$*. For a loss function, which elicits the* τ*-quantile see e.g., [Rockafellar and Uryasev, 2002]. Assuming that* P *is atomless,*

$$\begin{aligned} V(P, \gamma) &= \mathbb{E}_{Y \sim P}[(1-\tau)[\![Y < \gamma]\!] - \tau[\![Y > \gamma]\!]] \\ &= (1-\tau)\mathbb{E}_{Y \sim P}[[\![Y < \gamma]\!]] - \tau\mathbb{E}_{Y \sim P}[[\![Y > \gamma]\!]] \\ &= (1-\tau)P(Y < \gamma) - \tau(1 - P(Y \leq \gamma)) \\ &= (1-\tau)P(Y < \gamma) - \tau(1 - P(Y < \gamma) - P(Y = \gamma)) \\ &= (1-\tau)P(Y < \gamma) - \tau(1 - P(Y < \gamma)) \\ &= P(Y < \gamma) - \tau. \end{aligned}$$

Hence, the identification function is monotone. For local Lipschitzness we have to assume that the cumulative density function of P *is Lipschitz with parameter* L *around the* τ*-quantile of* P*. For instance, this is true for* β*-distributions. Then, we get*

$$|P(Y < \gamma) - \tau| = |P(Y < \gamma) - P(Y < \gamma^*)| \leq L|\gamma - \gamma^*|,$$

where γ^* *is the* τ*-quantile. Finally, if the cumulative distribution function induced through* P *is locally non-constant with parameter* $N > 0$ *around the* τ*-quantile of* P*, then via the same argument it follows that the identification function is locally non-constant (e.g.,* β*-distribution). In summary, the identification function* V_τ *for the* τ*-quantile inherits the continuity properties from the cumulative distribution function of the probability distribution. Hence, the assumption is fulfilled when restricting the space of possible conditional probability distributions defined through the regular data distribution* D*.*

Ratios of Expectations *Let* $g, h \colon \mathcal{Y} \to \mathbb{R}$ *be measurable functions and* $N < h(y) < M$ *for all* $y \in \mathcal{Y}$*. The ratio of expectations* $\Gamma(P) = \frac{\mathbb{E}_P g(Y)}{\mathbb{E}_P h(Y)}$ *is identified by,*

$$V(y, \gamma) = h(y)\gamma - g(y),$$

because

$$V(P, \gamma) = \mathbb{E}_P[h(Y)]\gamma - \mathbb{E}_P[g(Y)] = 0$$

if and only if,

$$\frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]} = \gamma.$$

Furthermore, V is oriented, locally Lipschitz with M,

$$\begin{aligned}
|V(P, \gamma)| &= |\mathbb{E}_P[h(Y)]\gamma - \mathbb{E}_P[g(Y)]| \\
&= \left|\mathbb{E}_P[h(Y)]\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&= |\mathbb{E}_P[h(Y)]|\left|\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&\leq M\,|\gamma - \Gamma(P)|\,.
\end{aligned}$$

and locally non-constant with N,

$$\begin{aligned}
|V(P, \gamma)| &= |\mathbb{E}_P[h(Y)]\gamma - \mathbb{E}_P[g(Y)]| \\
&= \left|\mathbb{E}_P[h(Y)]\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&= |\mathbb{E}_P[h(Y)]|\left|\left(\gamma - \frac{\mathbb{E}_P[g(Y)]}{\mathbb{E}_P[h(Y)]}\right)\right| \\
&\geq N\,|\gamma - \Gamma(P)|\,.
\end{aligned}$$

Mean and Variance *The identification function of the mean M is $V(y, \gamma) \coloneqq (y - \gamma)$. On the v-level set of the mean the variance $\mathbb{V}$ can be identified with $V_v(y, \gamma) \coloneqq y^2 - v^2 - \gamma$. Note that for $P \in M^{-1}(v)$,*

$$V_v(P, \gamma) = \mathbb{E}_{Y\sim P}[Y^2 - v^2 - \gamma] = \mathbb{V}(P) - \gamma.$$

Quantile and CVar *The identification function of a τ-quantile Q_τ is given above. The* CVar_τ *is identified on the v-level set of the τ-quantile with $V_v(y, \gamma) = v + \frac{1}{1-\tau}\max(0, y -$*

$v) - \gamma$. *Note that for* $P \in Q_\tau^{-1}(v)$

$$\begin{aligned} V_v(P, \gamma) &= \mathbb{E}_{Y \sim P}\left[v + \frac{1}{1-\tau}\max(0, Y - v)\right] \\ &= v + \frac{1}{1-\tau}\mathbb{E}_{Y \sim P}[\max(0, Y - v)] - \gamma \\ &= \mathrm{CVar}_\tau(P) - \gamma, \end{aligned}$$

The second last line follows from [Rockafellar and Uryasev, 2002, Theorem 10].

10.9.6 Calibration-Swap Regret Bridge under Groups

We call $c\colon \mathcal{X} \to \{0, 1\}$ a *group* if c is measurable and $D(c(X) = 1) > 0$. The set $\mathcal{C}$ denotes a collection of groups.

Definition 10.38 ($\mathcal{C}$-Robust Γ-Swap-Learner)**.** *Let $\mathcal{C}$ be a collection of groups and Γ a real-valued, identifiable property with oriented, locally Lipschitz (M) and locally non-constant (N) identification function V. Let ℓ be a loss function induced through V following Equation (10.8). Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. A Γ-predictor f is a β-approximate $\mathcal{C}$-robust Γ-swap-learner on D, iff, for all $c \in \mathcal{C}$,*

$$\mathbb{E}_D\left[\ell(Y, f(X))|c(X) = 1\right] \leq \min_{h \in \mathcal{M}_f(\mathcal{X}, \mathcal{R})} \mathbb{E}_D\left[\ell(Y, h(X))|c(X) = 1\right] + \frac{\beta}{\mathbb{E}_D\left[c(X)\right]},$$

where $\mathcal{M}_f(\mathcal{X}, \mathcal{R})$ is the set of all measurable functions from $(\mathcal{X}, \sigma(f))$ to $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$, where $\sigma(f)$ is the σ-algebra induced by the Γ-predictor f and $\mathcal{B}(\mathcal{R})$ the Borel-σ-algebra on $\mathcal{R}$.

The definition of a $\mathcal{C}$-robust swap-learner inspired by minimax regret in distributional robust optimization. Formalized as an optimization problem the regret here is a special case of a minimax regret optimization problem as in [Agarwal and Zhang, 2022]. The class of functions $\mathcal{F}$ in their Equation (1) is $\mathcal{M}_f(\mathcal{X}, \mathcal{R})$ in our case and the set of probability distributions $\mathcal{P}$ in their Equation (1) corresponds to the conditional probabilities on $\mathcal{X} \times \mathcal{Y}$ induced by conditioning on the groups $\mathcal{C}$ in our case. The predictor $f \in \mathcal{M}_f(\mathcal{X}, \mathcal{R})$.

Proposition 10.39 (l^2-Multicalibration implies Robust Swap-Learner)**.** *Let $\mathcal{C}$ be a collection of groups and Γ a real-valued, identifiable property with oriented, locally Lipschitz (M) and locally non-constant (N) identification function V. Let ℓ be a loss function induced through V following Equation (10.8). Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let f be α-approximately $(\mathcal{C}, \nu)$-multicalibrated following Definition 15 in [Noarov and Roth, 2023]. Then, f is a $\frac{M}{2}\alpha$-approximate $\mathcal{C}$-robust Γ-swap-learner.*

Proof.

$$
\begin{aligned}
\alpha &\geq \sup_{c\in C} \mathbb{E}_{X\sim D_X}[c(X)] \cdot \mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}[|\Gamma(D_{Y|f(X)=\gamma,c(X)=1}) - \gamma|^2] \\
&\overset{P10.28}{\geq} \frac{2}{M} \sup_{c\in C} \mathbb{E}_{X\sim D_X}[c(X)] \cdot \\
&\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}} \left[\mathbb{E}_D\left[\ell(Y,\gamma)|f(X)=\gamma, c(X)=1\right] - \min_{\hat{\gamma}\in \mathrm{im}\Gamma} \mathbb{E}_D\left[\ell(Y,\hat{\gamma})|f(X)=\gamma, c(X)=1\right] \right] \\
&= \frac{2}{M} \sup_{c\in C} \mathbb{E}_D\left[c(X)\ell(Y,f(X))\right] \\
&\quad - \mathbb{E}_{X\sim D_X}[c(X)] \cdot \mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}}[\min_{\hat{\gamma}\in \mathrm{im}\Gamma} \mathbb{E}_D\left[\ell(Y,\hat{\gamma})|f(X)=\gamma, c(X)=1\right]] \\
&= \frac{2}{M} \sup_{c\in C} \mathbb{E}_D\left[c(X)\ell(Y,f(X))\right] - \min_{h\in \mathcal{M}_f(\mathcal{X},\mathcal{R})} \mathbb{E}_D\left[c(X)\ell(Y,h(X))\right].
\end{aligned}
$$

The final step is a consequence of conditional expectations and the following argument. For every $c \in \mathcal{C}$, applying [Rockafellar and Wets, 2009, Theorem 14.60] with $T = \{x \in \mathcal{X} \colon c(x) = 1\}$ and $\mathcal{A} = \mathcal{B}(\mathcal{X}) \cap T$ gives,

$$
\begin{aligned}
&\mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}} \left[\min_{\hat{\gamma}\in \mathrm{im}\Gamma} \mathbb{E}_D\left[\ell(Y,\hat{\gamma})|f(X)=\gamma, c(X)=1\right] \right] \\
&\qquad = \mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}} \left[\min_{\alpha\in\mathbb{R}} \mathbb{E}_D\left[\ell(Y,\alpha)|f(X)=\gamma, c(X)=1\right] \right] \\
&\qquad = \min_{h\in \mathcal{M}_f(\mathcal{X},\mathcal{R})} \mathbb{E}_{\gamma\sim D_{f(X)|c(X)=1}} \left[\mathbb{E}_D\left[\ell(Y,h(X))|f(X)=\gamma, c(X)=1\right]\right] \\
&\qquad = \min_{h\in \mathcal{M}_f(\mathcal{X},\mathcal{R})} \mathbb{E}_D\left[\ell(Y,h(X))|c(X)=1\right].
\end{aligned}
$$

□

Proposition 10.40 (Robust Swap-Learner implies l^2-Multicalibration)**.** *Let $\mathcal{C}$ be a collection of groups and Γ a real-valued, identifiable property with oriented, locally Lipschitz (M) and locally non-constant (N) identification function V. Let ℓ be a loss function induced through V following Equation (10.8). Let D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Let f be a β-approximate $\mathcal{C}$-robust Γ-swap-learner. Then, f is $\frac{2L^2}{N}\beta$-approximately $(\mathcal{C}, \nu)$-multicalibrated following Definition 15 in [Noarov and Roth, 2023].*

Proof.

$$
\begin{aligned}
\beta &\geq \sup_{c \in C} \mathbb{E}_D \left[c(X)\ell(Y, f(X))\right] - \min_{h \in \mathcal{M}_f(\mathcal{X},\mathcal{R})} \mathbb{E}_D \left[c(X)\ell(Y, h(X))\right] \\
&= \sup_{c \in C} \mathbb{E}_D \left[c(X)\ell(Y, f(X))\right] \\
&\quad - \mathbb{E}_{X \sim D_X}[c(X)] \cdot \mathbb{E}_{\gamma \sim D_{f(X)|c(X)=1}} \left[\min_{\hat{\gamma} \in \mathrm{im}\Gamma} \mathbb{E}_D \left[\ell(Y, \hat{\gamma}) | f(X) = \gamma, c(X) = 1\right]\right] \\
&= \sup_{c \in C} \mathbb{E}_{X \sim D_X}[c(X)] \cdot \\
&\mathbb{E}_{\gamma \sim D_{f(X)|c(X)=1}} \left[\mathbb{E}_D \left[\ell(Y, \gamma) | f(X) = \gamma, c(X) = 1\right] - \min_{\hat{\gamma} \in \mathrm{im}\Gamma} \mathbb{E}_D \left[\ell(Y, \hat{\gamma}) | f(X) = \gamma, c(X) = 1\right]\right] \\
&\overset{P10.28}{\geq} \frac{N}{2} \sup_{c \in C} \mathbb{E}_{X \sim D_X}[c(X)] \cdot \mathbb{E}_{\gamma \sim D_{f(X)|c(X)=1}} \left[|\Gamma(D_{Y|f(X)=\gamma, c(X)=1}) - \gamma|^2\right],
\end{aligned}
$$

reversing the steps of the proof of Proposition 10.39. □

10.9.7 Self-Realization Does Not Imply Equal to Usefulness for All

In the following appendix we provide a short example to show that calibration does not imply that the low swap regret guarantee is equally useful for all. Note that conceptually the statements could seem strange to the reader. The predictor in our case has access to the nature's outcome. This surely is unrealistic. Nevertheless, it does not undermine the argument that a perfectly calibrated predictor exists which is of different usefulness to different downstream decision makers.

For the example we reuse *simple loss functions* (Definition 10.10. They describe binary-valued elicitable properties on binary outcome sets. The optimal achieveable risk, Bayes risk, for a distribution $P \in \Delta(\mathcal{Y})$ for $\mathcal{Y} = \{0, 1\}$ is given by,

$$\mathcal{B}_q(P) \coloneqq \mathbb{E}_{Y \sim P}[\ell_q(Y, \Phi_q(P))].$$

Lemma 10.41 (Calibration Does not Guarantee Cost Parity for Arbitrary Downstream Decision Makers with Equal Bayes Risk)**.** *Let $\mathcal{X}$ be finite (with size T), $\mathcal{Y} = \{0, 1\}$. Let ℓ_c and ℓ_d be simple losses with $c, d \in (0, 1)$ and $d < c$. For every choice of c, d there exists a data distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ regular wrt. $\mathcal{P}$ with the $\mathcal{X}$-marginal being the uniform distribution and $Q \in \Delta(\mathcal{Y})$ being a constant conditional distribution $D_{Y|X=x}$ such that*

(i) the Bayes risks are equal $\mathcal{B}_c(Q) = \mathcal{B}_d(Q)$.

(ii) and there exists a calibrated predictor $p \colon \mathcal{X} \to \Delta(\mathcal{Y})$ such that,

$$|\mathbb{E}_{(X,Y)\sim D}[\ell_c(Y, \Phi_c(p(X)))] - \mathbb{E}_{(X,Y)\sim D}[\ell_d(Y, \Phi_d(X))]| \geq C,$$

for some $C \in \mathbb{R}$.

Proof. We first define the stationary conditional distribution Q on the binary outcome set via $q \in [0,1]$. Let $q = \frac{1}{\frac{1-c}{d}+1}$. Observe $d < \frac{1}{\frac{1-c}{d}+1} < c$.
It remains to check the Bayes risk condition,

$$\begin{aligned}
\mathcal{B}_c(Q) &= (1-c)q \\
&= \frac{1-c}{\frac{1-c}{d}+1} \\
&= d\left(1 - \frac{1}{\frac{1-c}{d}+1}\right) \\
&= d(1-q) = \mathcal{B}_d(Q).
\end{aligned}$$

Now, let us define the predictor p.

We define the following prediction function for every $\omega \in \Omega$: if $y(\omega) = 0$, then $p(X(\omega)) = f$ for some $f \in [0,1]$ such that $d < f < q$. If $y(\omega) = 1$, then with probability

$$x \coloneqq \frac{(1-q)f}{q(1-f)},$$

the prediction is $p(X(\omega)) = f$, otherwise $p(X(\omega)) = 1$. Since $f < q$ and $d, c \in (0,1)$,

$$0 < x = \frac{(1-q)f}{q(1-f))} < 1.$$

For the second, note that all predictions $p(X(\omega)) = 1$ are calibrated by definition. For the predictions $p(X(\omega)) = f$ we have

$$\begin{aligned}
\mathbb{E}_{Y\sim Q}[Y|p(X) = f] &= \frac{qx}{(1-q)+qx} \\
&= \frac{q\frac{(1-q)f}{g(1-f)}}{(1-q)+q\frac{(1-q)f}{q(1-f)}} \\
&= \frac{\frac{(1-q)f}{1-f}}{\frac{(1-q)(1-f)}{1-f} + \frac{(1-q)f}{1-f}} \\
&= \frac{\frac{(1-q)f}{1-f}}{\frac{(1-q)}{1-f}} \\
&= f.
\end{aligned}$$

So, we can finally show the statement

$$\begin{aligned}
&|\mathbb{E}_{(X,Y)\sim D}[\ell_c(Y, \Phi_c(p(X)))] - \mathbb{E}_{(X,Y)\sim D}[\ell_d(Y, \Phi_d(X))]| \\
&= |\mathbb{E}_{X\sim D_X}[qx(1-c) - (1-q)d]| \\
&= (1-q)\left|\frac{f}{1-f}(1-c) - d\right| \\
&\geq C,
\end{aligned}$$

because $\frac{f}{1-f}(1-c) - d \neq 0$ if $f < q$. □

In other words, the predictions, even though they are calibrated and hence fulfilling some notions of omniprediction (cf. Section 10.5.3), do not guarantee that the predictions are equally useful for every decision maker. This holds despite the forecasting task itself being equally difficult as measured by the Bayes risk.

10.9.8 Proofs and Lemmas

Lemma 10.42. *Let $\Gamma\colon \mathcal{P} \to \mathcal{R}$ be a property with convex level sets and D a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $\mathcal{Y}$ is finite and define $A \subseteq \mathcal{X}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a distributional predictor such that $\Gamma(f(x)) = \gamma$ for all $x \in A$. We have,*

$$\Gamma(\mathbb{E}_D[f(X)|X \in A]) = \gamma.$$

Proof. Because $|\mathcal{Y}| < \infty$, $f(x) \in \mathbb{R}^{|\mathcal{Y}|}$ for all $x \in A$. Hence, $\mathbb{E}_D[f(X)|X \in A] \in \mathbb{R}^{|\mathcal{Y}|}$. It remains to show that $\mathbb{E}_D[f(X)|X \in A]$ is in the level set $\Gamma^{-1}(\gamma) \subseteq \mathbb{R}^{|\mathcal{Y}|}$. For that, we use the convex indicator function $I\colon \mathcal{P} \to \mathbb{R}$ with $d \mapsto [\![d \in \Gamma^{-1}(\gamma)]\!]$

$$\begin{aligned}
I(\mathbb{E}_D[f(X)|X \in A]) &= [\![\mathbb{E}_D[f(X)|X \in A] \in \Gamma^{-1}(\gamma)]\!] \\
&\leq \mathbb{E}_D\left[[\![f(X) \in \Gamma^{-1}(\gamma)]\!]|X \in A\right] \\
&= 0,
\end{aligned}$$

by Jensen's inequality. Hence, $\mathbb{E}_D[f(X)|X \in A] \in \Gamma^{-1}(\gamma)$. □

Lemma 10.43. *Let $\mathcal{Y}$ be finite and Γ be the full distribution property, and thus $\mathcal{R} = \Delta(\mathcal{Y})$. If $\ell\colon \mathcal{Y} \times \mathcal{R} \to \mathbb{R}$ is a bounded loss function, then ℓ is locally outcome Lipschitz with respect to Γ.*

Proof. Let Φ be the property elicited through ℓ. For all $v \in \mathrm{im}\Phi$, $\gamma \in \mathrm{im}\Gamma$, $P \in \Gamma^{-1}(\gamma)$ and

$P' \in \mathcal{P}$,

$$
\begin{aligned}
&|\ell(P, v) - \ell(P', v)| \\
&= |\sum_{y \in \mathcal{Y}} (P(Y = y) - P'(Y = y))\ell(y, v)| \\
&\leq C|\sum_{y \in \mathcal{Y}} (P(Y = y) - P'(Y = y))| \\
&\leq C \sum_{y \in \mathcal{Y}} |(P(Y = y) - P'(Y = y))| \\
&\leq Cm(\Gamma(P'), \gamma),
\end{aligned}
$$

because m is the total variation distance. □

Lemma 10.44 (When Decision Calibration is Equivalent to Calibration on Decision Sets)**.** *Let $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X}$ arbitrary and D be a data distribution on $\mathcal{X} \times \mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f \colon \mathcal{X} \to [0, 1]$ is a mean-predictor, i.e., it predicts the probability of $Y = 1$. Furthermore, suppose f matches the average, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)] = 0$. Let ℓ_q be a simple loss function with parameter $q \in [0, 1]$. In addition, suppose $P_D(f(X) \leq q) > 0$. Then, f is calibrated on the decision sets of ℓ_q, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) \leq q] = 0$ and $\mathbb{E}_{(X,Y)\sim D}[Y - f(X)|f(X) > q] = 0$[28], if and only if, f is decision calibrated with respect to the loss function ℓ_q.*

Proof. Let us rewrite the decision calibration term,

$$
\begin{aligned}
&\mathbb{E}_{X\sim D_X}[\mathbb{E}_{Y\sim D_{Y|X=x}}[\ell_q(Y, f(x))] - \mathbb{E}_{Y\sim f(X)}[\ell_q(Y, f(x))]] \\
&= \mathbb{E}_{X\sim D_X}[P(Y = 0|X = X)q[\![f(X) > q]\!] + P(Y = 1|X = X)(1 - q)[\![f(X) \leq q]\!] \\
&\quad - (1 - f(X))q[\![f(X) > q]\!] - f(X)(1 - q)[\![f(X) \leq q]\!]] \\
&= \mathbb{E}_{X\sim D_X}[(P(Y = 1|X = x) - f(X))q[\![f(X) > q]\!] \\
&\quad + (f(X) - P(Y = 1|X = x))(1 - q)[\![f(X) \leq q]\!]] \\
&= q\mathbb{E}_{X\sim D_X}[(P(Y = 1|X = x) - f(X))[\![f(X) > q]\!]] \\
&\quad + (1 - q)\mathbb{E}_{X\sim D_X}[(f(X) - P(Y = 1|X = x))[\![f(X) \leq q]\!]] \\
&=: L(q).
\end{aligned}
$$

[28] If f matches the average, the first assumption implies the second.

Note that the following two equations hold,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))[\![f(X)>q]\!]] \\
&=\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&=\mathbb{E}_{(X,Y)\sim D}[(Y-f(X))|f(X)>q]P_D(f(X)>q),
\end{aligned}$$

and,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))[\![f(X)\leq q]\!]] \\
&=\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))|f(X)\leq q]P_D(f(X)\leq q) \\
&=\mathbb{E}_{(X,Y)\sim D}[(f(X)-Y)|f(X)\leq q]P_D(f(X)\leq q).
\end{aligned}$$

With these reformulations at hand, it is easy to see that calibration on decision sets of ℓ_q implies decision calibration with respect to the loss function ℓ_q, because,

$$\begin{aligned}
L(q) &= q\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&\quad +(1-q)\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))|f(X)\leq q]P_D(f(X)\leq q) \\
&= q0P_D(f(X)>q)+(1-q)0P_D(f(X)\leq q)=0.
\end{aligned}$$

The reverse implication requires a bit more reasoning. We argue via contraposition. Let us assume that $\mathbb{E}_{(X,Y)\sim D}[Y-f(X)|f(X)\leq q]\neq 0$. Then,

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim D}[Y-f(X)] \\
&=\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&\quad +\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)\leq q]P_D(f(X)\leq q) \\
&=0.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)>q]P_D(f(X)>q) \\
&=-\mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)\leq q]P_D(f(X)\leq q),
\end{aligned}$$

respectively,

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))[\![f(X)>q]\!]] \\
&=\mathbb{E}_{X\sim D_X}[(f(X)-P(Y=1|X=x))[\![f(X)\leq q]\!]].
\end{aligned}$$

Now,

$$\begin{aligned} L(q) &= \mathbb{E}_{X\sim D_X}[(P(Y=1|X=x)-f(X))|f(X)\leq q]P_D(f(X)\leq q) \\ &\neq 0, \end{aligned}$$

because $P_D(f(X)\leq q)>0$ by assumption. □

Lemma 10.45 (Matching Averages by Non-Extremal Predictions I)**.** *Let $\mathcal{Y}=\{0,1\}$ and $\mathcal{X}$ arbitrary and D be a data distribution on $\mathcal{X}\times\mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f\colon\mathcal{X}\to[0,1]$ is a mean-predictor, i.e., it predicts the probability of $Y=1$. Furthermore, we assume there exists $f_{\inf}\coloneqq\inf\operatorname{im} f>0$. Then, f matches the average, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y-f(X)]=0$, if and only if f is decision calibrated with respect to the loss function $\ell_{\frac{f_{\inf}}{2}}$.*

Proof. Observe,

$$\mathbb{E}_{(X,Y)\sim D}[Y-f(X)]=\mathbb{E}_{(X,Y)\sim D}\left[(P(Y=1|X=x)-f(X))[\![f(X)>\frac{f_{\inf}}{2}]\!]\right].$$

and

$$\begin{aligned} L\left(\frac{f_{\inf}}{2}\right) &= \frac{f_{\inf}}{2}\mathbb{E}_{X\sim D_X}\left[(P(Y=1|X=x)-f(X))[\![f(X)>\frac{f_{\inf}}{2}]\!]\right] \\ &\quad+\left(1-\frac{f_{\inf}}{2}\right)\mathbb{E}_{X\sim D_X}\left[(f(X)-P(Y=1|X=x))[\![f(X)>\frac{f_{\inf}}{2}]\!]\right] \\ &= \frac{f_{\inf}}{2}\mathbb{E}_{(X,Y)\sim D}[Y-f(X)]. \end{aligned}$$

Since $\frac{f_{\inf}}{2}>0$, $L(f_{\inf})=0$ if and only if $\mathbb{E}_{(X,Y)\sim D}[Y-f(X)]=0$. □

Lemma 10.46 (Matching Averages by Non-Extremal Predictions II)**.** *Let $\mathcal{Y}=\{0,1\}$ and $\mathcal{X}$ arbitrary and D be a data distribution on $\mathcal{X}\times\mathcal{Y}$ regular wrt. $\mathcal{P}$. Suppose that $f\colon\mathcal{X}\to[0,1]$ is a mean-predictor, i.e., it predicts the probability of $Y=1$. Furthermore, assume $f_{\inf}\coloneqq\inf\operatorname{im} f=0$, and $|\operatorname{im} f|<\infty$, hence there exists $v\in[0,1]$ such that $v>0$ but $v<v'$ for all $v'\in\operatorname{im} f\setminus\{0\}$. Then, f matches the average, i.e., $\mathbb{E}_{(X,Y)\sim D}[Y-f(X)]=0$, if and only if f is decision calibrated with respect to the loss functions ℓ_v and ℓ_0.*

Proof. It holds $L(0)=0$ and $L(v)=0$. In detail,

$$L(0)=\mathbb{E}_{X\sim D_X}\left[(f(X)-P(Y=1|X=x))[\![f(X)=0]\!]\right]=0.$$

$$
\begin{aligned}
L(v) &= v\mathbb{E}_{X\sim D_X}\left[(P(Y=1|X=x)-f(X))[\![f(X)>v]\!]\right] \\
&\quad + (1-v)\,\mathbb{E}_{X\sim D_X}\left[(f(X)-P(Y=1|X=x))[\![f(X)=0]\!]\right] \\
&= v\mathbb{E}_{X\sim D_X}\left[(P(Y=1|X=x)-f(X))[\![f(X)>v]\!]\right] = 0.
\end{aligned}
$$

Hence, since $v>0$, $\mathbb{E}_{X\sim D_X}\left[(P(Y=1|X=x)-f(X))\mathbf{1}\left[f(X)>v\right]\right]=0$. Consequently,

$$
\begin{aligned}
&\mathbb{E}_{X\sim D_X}\left[(P(Y=1|X=x)-f(X))[\![f(X)>v]\!]\right] \\
&\quad + \mathbb{E}_{X\sim D_X}\left[(f(X)-P(Y=1|X=x))[\![f(X)=0]\!]\right] \\
&= \mathbb{E}_{(X,Y)\sim D}[Y-f(X)] = 0.
\end{aligned}
$$

□

11

Evaluation Metrics As Averaged Outcomes of Fair Gambles

Author Contributions

Rabanus Derr (RD) pitched the idea of "available gambles" at the yearly retreat of the Foundations of Machine Learning Systems group. Then, the paper's content got refined in exchange with Robert C. Williamson (RW). In particular, RW pointed RD to the concept of "felicity" from speech act theory. RD did the writing. RW took over the internal reviewing.

The Personal Note

The FMLS group has the wonderful tradition – it can be called that way as it happened for three times already – of a yearly retreat for about four days. In the atmosphere of Black Forest we take time to update each other on the next projects to pursue, obtain feedback, play some fun ML-Hype-Powerpoint-Karaoke and simply chat about this and that. During the retreat of 2023, I presented an idea which boiled down to two observations. First, a property of a distribution like a mean, which are called *elicitable* because they can be elicited by expected loss minimization, defines a credal set. A credal set is a closed, convex set of probability distributions. The construction is: the set of probability distributions which share the mean is convex and, under mild restrictions, closed. Second, the set of random variables whose expected values under all those distributions is equal or lower than zero, is equal to the set of random variables which are smaller or equal to $y \mapsto \alpha(y - m)$ for some $\alpha \in \mathbb{R}$, where y is the unknown and m the mean.

What provoked those observations? Partially the mysterious "calibration" (see Section 10) and partially the intriguing game-theoretic probability and statistics. Game-theoretic probability aims for establishing probability theory on the ground of game protocols and betting strategies instead of measure theory [Shafer and Vovk, 2019]. Tests in game-theoretic statistics are taken as literal bets on hypothesis.

With "The Game" at its core, that is how I called the evaluation protocol, I tried to link together four facets of forecast evaluation: calibration-type of evaluations, loss- or regret-type of evaluations, randomness and predictiveness. The mathematical framework using dual L^p-spaces was fun to sort out. But, the amount of mathematical and non-mathematical concepts stirred in a single pot hardly convinced any reader of a draft (thanks to Aaron Roth) or audience of a presentation (thanks to the audience at Harvard University where Jessie invited me).

The project was called "Four Facets of Forecast Felicity". Sounds familiar? Bob suggested this alliterative title and pointed me to *felicity*. It was first the alliteration that convinced me. Then, I got persuaded by forecast felicity as a concept.

After several of iterations the four facets got shrunk to two. Additionally, the mathematical framework switched. Luckily, the project got a new title. Hence, the former title was left for the thesis.

The resulting work is, I think, the one I'm most attached to. I think it is a nice contribution. It is mathematical and conceptual and shapes my thinking about evaluation of forecasts. Many of the existing relations are captured. The mix of techniques and literature was a lot of fun. I would call it the most rabanus paper.

Game-theoretic probability and statistics is still one of the topics I'm very excited about.

I am convinced that the developments are underappreciated. Meanwhile, there are simple fundamental questions unsolved, e.g., the role of non-negativity of supermartingales in the theories [Koolen et al., 2025], the power of supermartingale to falsify and verify (!) certain statistical hypothesis, the role of falsifiability of statistical hypothesis versus the actual falsification [Olszewski and Sandroni, 2011].

Summary: Forecast Evaluation

Motivation As articulated in the Summary 10, calibration has regained a lot of interest in the machine learning community. But surely, calibration is not the only quality criterion used for forecasts. Loss functions, regret types and deviates numerically certify the "goodness" of forecasts with respect to realized observations at least as long as calibration [Brier, 1950, Murphy and Epstein, 1967].

Most of the new developments in machine learning share the interest in defining evaluation *objectives* of forecasts and then providing algorithms which computationally solve them [Bartlett et al., 2006, Jung et al., 2021, Globus-Harris et al., 2023, Deng et al., 2023, Gopalan et al., 2024b, Garg et al., 2024]. Less prominently, evaluations can certify forecasts.

Hence, evaluation metrics are measures to the end of certifying the "goodness" of forecasts. The contextualization of evaluation as forecast's felicity demands us to ask (cf. Section 5.2): What is the structure of the choices of evaluation metrics? How felicitously choose an evaluation metric? How do calibration-type and regret-type evaluation metrics relate? What are reasonable meta-criteria evaluation metrics should fulfill?

Contributions The presented answers require a dip into game-theoretic thinking in probability and statistics [Shafer and Vovk, 2019, Ramdas et al., 2022]. As before the terminology is not streamlined with the surrounding chapters. Hence, I provide a summary of the most important notations and terms in Table 11.1.

Unified Notation & Terminology	Chapter 11
statistical forecast	forecast, $v_t \in \mathcal{V}$
individual, ι	round, individual, $t \in T$
aggregate, $\mho$	input set, $T = \{1, \dots, n\}$
attribute, $\varrho(\iota) = \wp$	outcome, $y_t \in \mathcal{Y}$
measurement	–

Table 11.1: Translation table – Evaluation Metrics As Averaged Outcomes of Fair Gambles

We suggest to start with an *evaluation protocol* in which first, the forecaster suggests a forecast, then a gambler provides a gamble, i.e., a mapping from potential outcomes to

some currency, and finally a nature reveals the realized attribute or outcome. The protocol is repeated for every instance, i.e., individual. We then show that many existing evaluation metrics can be considered "as if" they would be the resulting capital of the gambles' outcomes, when the gambler would have played under certain intuitive conditions (Proposition 11.4 and Section 11.5).

The central condition placed on the gambler is that it is *fair*. That means that the gambler is only allowed to play gambles which the forecaster, under its (probabilistic) forecast, literally expects to lead to a non-positive value of the gamble. Hence, forecaster expects the gambler to loose money. It turns out that this condition is equivalent to demanding that a forecaster cannot gain any capital whenever the forecaster would know the realized attribute or outcome in advance (Proposition 11.6). Note that no capital means, that the corresponding evaluation metric would stay zero or tends to negative. Fairness is, that we argue, a reasonable meta-criterion for any evaluation metric.

In addition, we equip the gambler with its own *belief* about the next realized attribute or outcome. We assume that the gambler behaves rationally with respect to its belief. Depending on the alignment of this belief with the actual realized outcomes or attributes, the gambler's capital is easier or harder to be kept low from the perspective of the forecaster.

Finally, the gambler is artificially *restricted* in its available gambles. This way it is prevented that the gambler can simply increase its stakes on every instance, which would render previous values of gambles ignorable (cf. Saint Petersburg paradox [Anonymous author, 2025]).

Beliefs and *restrictions* form two axis which span up the space of evaluation metrics. With appropriate choices we can recover existing evaluation metrics (Section 11.5). Furthermore, within this structure we show general bounds between metrics (Proposition 11.42 and Proposition 11.45). *Beliefs* and *restrictions* form the two knobs to turn when tuning for a felicitous evaluation metric.

Tools and Sketch of Argument The full paper treats forecasts to be properties of probability distributions, e.g., the mean, the median, a quantile, the full distribution, which are elicitable (respectively identifiable) (Definition 11.7 and Definition 11.8). For intuition, a simple probabilistic forecasting setting with binary outcomes, e.g., rain forecasting, is sufficient.

Let $T \coloneqq \{1, \dots, n\}$ be an aggregate of individuals, e.g., days. Let $\mathcal{Y} \coloneqq \{0, 1\}$ be a binary outcome set. The evaluation protocol is as follows. For every individual $t \in T$,

1. Forecaster forecasts the probability of $Y = 1$, $v_t \in [0, 1]$.
2. Gambler gambles $G_t \in \mathbb{R}^{\{0,1\}}$.

3. Nature reveals outcome $y_t \in \mathcal{Y}$.

The gambler's capital is the averaged realized value of all played gambles, i.e., $K \coloneqq \frac{1}{|T|}\sum_{t\in T} G_t(y_t)$. Note that a gamble G_t is a mapping from the potential outcomes $\{0,1\}$ to the reals, hence, a point in $\mathbb{R}^2$.

The gambler is *fair*, if $\mathbb{E}_{v_t}[G_t] \leq 0$ for all $t \in T$. If $v_t = \delta(y_t)$, i.e., the forecaster is clairvoyant and puts Dirac measures on the realized outcomes, then fairness is equivalent to the gambler obtaining capital $K \leq 0$ (Proposition 11.6).

Let us consider two illustrative instantiations of a restriction and a belief of gambler. First, restrict the gambler A to play the following gambles $G_t(y) = (y-v)^2 - (y-c)^2$ for all $t \in T$, where $c \in [0,1]$ is some constant. By definition of the gambles, the gambler A plays fairly. Furthermore, assume that the gambler A is clairvoyant, i.e., it knows the revealed forecast in advance. The resulting capital of the gambler A is the Brier score [Brier, 1950] (Proposition 11.35),

$$K = \frac{1}{|T|}\sum_{t\in T}(y_t - v_t)^2.$$

Second, assume that the gambler B is restricted to play gambles such that $\|G_t\|_1 \leq 1$, i.e., the gamble is in the l_1-norm ball. If the gambler B has the averaged belief $b_t = \frac{1}{|T_\gamma|}\sum_{t'\in T_\gamma}\delta(y'_t)$ where $T_\gamma \coloneqq \{t \in T \colon v_t = \gamma\}$. That is, the gambler B has a true belief about the average outcome on the individuals on which γ was predicted. The resulting capital of the gambler B is the expected calibration score, e.g., [Gopalan et al., 2022a, Definition 3.1],

$$K = \sum_{\gamma\in\{v_t\colon t\in T\}} \frac{|T_\gamma|}{|T|}\left|\frac{1}{|T_\gamma|}\sum_{t\in T_\gamma} y_t - \gamma\right|.$$

To provide an intuition how the two scores relate, first note that if we scale up the norm ball times four, the gambler B is less restricted than gambler A. However, gambler A has a more refined belief about the outcomes. For this reason, the two scores are incomparable, as they do not match along one of the axis.

We furthermore show that a gambler A who would be allowed to arbitrarily scale its gambles could, in principle, achieve the same scores as a gambler B who could arbitrarily scale its gambles (Corollary 11.25). Hence, in the asymptotic restriction regime both gambler are equally "powerful", when neglecting their difference in beliefs.

Relation to Forecast Felicity Forecast evaluation is part of the felicity conditions. We argue that forecast evaluation is gambling. Thus, a forecast is felicitous if the forecasts

cannot be gambled against successfully by a fair gambler under certain conditions.

As for calibration (Section 10), there is no single right choice of evaluation. Imagine a rain forecaster. Is it required to be calibrated or achieve a low Brier score? This is a discussion as old as the use of probabilistic forecasts [Murphy and Epstein, 1967].

The evaluation protocol provides a rich structure of evaluation metrics with two main features interesting for felicitous forecasts. (a) An evaluation metric corresponding to an unfair gambler can arguably not guarantee felicity. (b) The choice of evaluation metric is governed by two axis, the restriction and belief of the gambler.

The gambles *test*[1] the forecasts. A high resulting capital of the gambler does not necessarily correspond to an actual currency, but it is a measure of evidence.

The evaluation problem is potentially the most prominent among the problems given in the introduction. It is most easily identifiable as a felicity condition, hence a relative choice.

[1] In fact, the gambles actually test the forecasts as being statistical hypothesis, cf. game-theoretic statistics and e-values Shafer et al. [2011], Grünwald et al. [2020], Ramdas et al. [2023]. For an elaboration see Section 11.6.1.

Abstract

Machine learning is about forecasting. When the forecasts come with an evaluation metric the forecasts become useful. What are reasonable evaluation metrics? How do existing evaluation metrics relate? In this work, we provide a general structure which subsumes many currently used evaluation metrics in a two-dimensional hierarchy, e.g., external and swap regret, loss scores, and calibration scores. The framework embeds those evaluation metrics in a large set of single-instance-based comparisons of forecasts and observations which respect a meta-criterion for reasonable forecast evaluations which we term "fairness". In particular, this framework sheds light on the relationship on regret-type and calibration-type evaluation metrics showing a theoretical equivalence in their ability to evaluate, but practical incomparability of the obtained scores.

11.1 Introduction

In the current practices of machine learning, the evaluation of forecasts has become a cornerstone of scientific progress [Blum and Hardt, 2015, Hardt, 2025]. For instance, OpenAI reports loss scores and calibration scores for GPT-4 on different tasks [OpenAI (2023), 2024]. The inventors of the GraphCast weather forecasting model provide several metrics, among them root mean squared error [Lam et al., 2023]. The introduction of AlphaFold was accompanied by a series of evaluation metrics, e.g., accuracy or absolute difference, used to discuss the quality of the protein folding forecasts [Senior et al., 2020]. On the more formal side machine learning scholars have come up with dozens of evaluation metrics: different types of (proper) loss functions [Williamson and Cranko, 2023], a multitude of calibration notions [Derr et al., 2026, Gopalan and Hu, 2025], and numerous types of regret [Cesa-Bianchi and Lugosi, 2006]. **What do all these evaluation metrics share? In which regard do the evaluation metrics differ? How can they be viewed from a unified perspective?**

Among the used metrics, two main types have crystallized, one which we call *regret-type* and one which we call *calibration type*. Regret-type evaluation metrics compare a loss incurred by a predictor with the loss of some comparison predictor. Regret-type evaluation metrics require a loss function and a comparison baseline, which is either explicitly or implicitly given.[2] Regret-type evaluation metrics include external regret [Cesa-Bianchi and Lugosi, 2006, page 80], swap regret [Cesa-Bianchi and Lugosi, 2006, page 91], and loss scores (cf. Table 11.2).

Calibration-type evaluation metrics (traditionally) compare average outcomes to average forecasts on certain instances [Dawid, 1985a, 2017, Höltgen and Williamson, 2023]. Those metrics have been generalized beyond the average, as e.g., mean and variance calibration [Jung et al., 2021], quantile calibration [Jung et al., 2023] or property calibration [Gneiting and Resin, 2023, Noarov and Roth, 2023, Derr et al., 2026]. This raises the questions: **What is the relationship between the regret-type and calibration-type evaluation metrics? What is the structure of the regret-type and the calibration-type evaluation metrics?** (cf. Section 11.6.2)

Towards answering the posed questions we commit to a general framework which we call an *evaluation protocol*. This protocol is based upon the game-theoretic probability setup developed in [Shafer and Vovk, 2005, 2019]. Roughly summarized, the evaluation protocol consists of a nature, a forecaster and a set of gamblers. For every forecast, (fair)[3] gambles by the gamblers are played and evaluated on the actual outcome revealed by nature. For each

[2] For instance, loss scores have an implicit baseline which is the predictor which knows the truth and hence often, depending on the definition of the loss function, incurs 0 loss.

[3] A gamble is fair if the forecaster literally expects the gamble to not be favorable for the gambler.

gambler, the average realized value of the gambles is summarized as its capital. The capital is the obtained score. It shows how much the forecasts and outcomes match, the bigger the capital the worse they match. Examples for scores are the expected calibration error, false positive rate, false negative rate, accuracy, mean squared error, multicalibration, bias in the large, external regret, swap regret, Brier score, and the log loss. With our machinery we will argue that all those (and many more) **scores obtained by evaluation metrics are averaged realized values of (fair) gambles**.

Warm-Up Let's shortly illustrate the evaluation protocol in a little example. Let $(y_t)_{t\in T}$ be a tuple of binary outcomes, i.e., $y_t \in \{0,1\}$ for all $t \in T$. Furthermore, let $(v_t)_{t\in T}$ be a tuple of probabilistic forecasts, $v_t \in [0,1]$ for all $t \in T$. For every instance $t \in T$, a forecast v_t is made, and let us suppose there is a gambler, who chooses to play the gamble $G_t \colon y \mapsto (y - v_t)^2 - (y - y_t)^2$. Such a gamble is *fair*, that is,

$$\mathbb{E}_{v_t}[G_t] = (1 - v_t)G_t(0) + v_t G_t(1) \leq 0.$$

In other words, the forecaster expects the gambler to not win. Hence, if indeed the gambler achieves a large capital, it disproves the abilities of the forecaster. Crucially, the gambler, when playing G_t makes use of the actual outcome y_t (see the definition). This (hypothetical) information is *not* accessible to the forecaster in this fictive play.

Now, when we actually evaluate the gambles on the binary outcomes, and average, we obtain,

$$\frac{1}{n}\sum_{t\in T} G_t(y_t) = \frac{1}{n}\sum_{t\in T}(y_t - v_t)^2 - (y_t - y_t)^2 = \frac{1}{n}\sum_{t\in T}(y_t - v_t)^2, \tag{11.1}$$

which is the *Brier score* an evaluation metric introduced in meteorology [Brier, 1950] and used to evaluate class probability estimates.

It is the case that besides the Brier score, all kinds of calibration, e.g., expected calibration error, and regret-type evaluation metrics can be embedded in a comparable fashion. In particular, all those metrics respect that the gambler only plays *fair* gambles. However, to generally reveal the underlying gambling structure it is necessary to move beyond binary outcomes and simple probabilistic forecasts.

11.1.1 Paper Outline

First, we summarize our contribution and clarify relevant terms and the scope of our work in the current section. Then, we introduce the formal setup (Section 11.2). We require some machinery from convex analysis in infinite dimensional spaces. The complexity of the

mathematical tools is needed to accurately reflect a wide range of evaluation metrics used in settings from classification to regression. We continue by constructing the evaluation protocol based on game protocols used in game-theoretic probability. We characterize the set of evaluation metrics which are representable in the evaluation protocol. Furthermore, we adapt a "sanity" condition for the gamblers which we call *fairness* (Definition 11.5) from [Shafer and Vovk, 2019].[4] Essentially, a gambler is fair if it only plays *available* gambles, i.e., gambles which the forecaster assumes, in expectation, to lead to a non-positive outcome. We show that this criterion is equivalent to guaranteeing that whenever the forecaster is "clairvoyant", i.e., exactly forecasts the true outcome, the gambler will not increase its capital above 0 (Proposition 11.6).

The general evaluation protocol allows forecasts to correspond to sets of probability distributions, cf. imprecise probabilities [Augustin et al., 2014]. We demarcate a subset of such forecasts using the concepts of elicitablity and identifiability from statistics, e.g. [Gneiting, 2011, Steinwart et al., 2014]. Equipped with these tools, we present our contributions which can be separated into two main parts (Section 11.4 and Section 11.5). Finally, we contextualize our findings in the literature on evaluation of forecasts (Section 11.6).

Contributions – Section 11.4 First, we develop a duality which allows us to fully characterize the set of available gambles. The central results of this part are summarized in Figure 11.1.

1. We show that *(unscaled) regret gambles* and *calibration gambles* (cf. Section 11.4.2), which are central to the recovery results in Section 11.5, are both available (Proposition 11.10 and Proposition 11.11). Furthermore, we fully characterize the set of available gambles in case the forecasts are aligned to elicitable (Theorem 11.13) or identifiable Properties (Theorem 11.15). In particular, *scaled regret gambles* (respectively *calibration gambles*) approximately dominate all available gambles. These statements are based upon a general polar duality between closed, convex sets of probability distributions, i.e., credal sets, and convex cones of gambles, i.e., available gambles (Theorem 11.22).[5]

2. The characterizations lead to the insight that scaled regret gambles are, in principle, equally powerful as calibration gambles. Both those types dominate the same set of available gambles (Corollary 11.25).

[4] In [Shafer and Vovk, 2019], the authors use the term "available" gambles, but don't formally define it.

[5] This duality is known for different technical setups, e.g., [Föllmer and Schied, 2011, Benavoli et al., 2017], however, not for the $C(\mathcal{Y})$-ca$(\mathcal{Y})$-pairing, cf. Section 11.4.5.

Contributions – Section 11.5 Second, we further constrain the gamblers to recover existing evaluation metrics and reveal a two-dimensional hierarchy of evaluation metrics.

1. For the recovery of existing evaluation metrics, we introduce two further conditions on the gamblers, *restrictiveness* and *rationality*. A restricted gambler is bounded to gambles in a certain set. Intuitively, there is a capital constraint set on the gambler. For rationality, we equip the gambler with a belief about the outcome, comparable to an alternative hypothesis of what the gambler actually believes is true, and demand that the gambler optimizes the choice of gamble with respect to the expected value under this belief. By tuning restriction and rationality, we obtain the evaluation metrics summarized in Table 11.2. For instance, the capital of a set of gamblers with access to all gambles with a bounded norm and a (true) belief about the average outcome on the subsets in which a certain value has been predicted is aggregated to calibration scores (Proposition 11.38). The two dimensional structure, restriction and belief, leads to the conclusion that standard calibration scores and standard loss scores are incommensurable in the sense that they differ in both the dimensions.

2. We provide a hierarchy along the dimension of restrictions via set-containment (Proposition 11.42) and along the dimension of beliefs via refinement (Proposition 11.45) which generalize known inequalities between calibration and swap regret (Corollary 11.43), between regret notions (Corollary 11.47) and between calibration notions (Corollary 11.48).

11.1.2 Clarification of Terms and Scope

Evaluation as Certificate The question of how one should evaluate forecasts has always been part of debates since the beginning of machine learning [DeGroot and Fienberg, 1983, Schervish, 1989, Gneiting, 2011, Williamson and Cranko, 2023]. However, large parts of machine learning literature concentrate on evaluation of forecasts as "modeling an objective", e.g., [Bartlett et al., 2006, Jung et al., 2021, Gupta and Ramdas, 2022, Deng et al., 2023]. We focus on evaluation of forecasts as "providing a certificate of quality", a perspective shared by the classical literature on the evaluation of meteorological forecasts [Brier, 1950, Murphy and Epstein, 1967]. In particular, the evaluation of forecasts is critical for the value of a forecast. Forecasts on their own are of limited value. When they come with an accompanying guide how to evaluate them, they become useful. The evaluation of forecasts *make* the forecasts, cf. [Dawid, 2017, Höltgen, 2024, Perdomo and Recht, 2025].

	Restriction on gambler	
Belief of gambler	Regret Gambles $\mathcal{L}_\gamma$	Norm Ball $\mathbb{B}_\alpha$
Average outcome	External Regret (Proposition 11.31)	Bias in the large (Proposition 11.36)
Average outcome on prediction-groups	Swap Regret (Proposition 11.33)	Calibration Score (Proposition 11.38)
Average outcome on subgroups	Group-Wise Swap Regret (Proposition 11.34)	Multicalibration Score (Proposition 11.39)
Exact knowledge of true outcome	Zeroed Loss Score (Proposition 11.35)	Individual Calibration Score (Proposition 11.41)

Table 11.2: Comparison of belief and restrictions of gambler and the resulting evaluation metrics. The corresponding propositions where the recovery is shown are referenced below the metric. We neglect details about the type of aggregation, the outcome set $\mathcal{Y}$ and the forecasted property Γ. A *prediction-group* is a set of instances which share the predicted value γ. A *subgroup* is a set of instances which share the predicted value γ and a certain attribute S.

Clarification of Terms To be clear, the test for consistency of forecasts with observations is what we call *evaluation*.[6] The actual mapping from forecasts and outcomes to some real number is what we call an *evaluation metric* and the resulting number the *score*.

No Algorithmic Solution In this work, we suggest an abstract evaluation protocol to map the landscape of *evaluation metrics*. We do not propose any algorithmic solution to give forecasts optimizing an evaluation metric.[7] Even though the evaluation-protocol is motivated by game-theoretic probability and statistics [Ramdas et al., 2023], we do not give claims about equilibria or convergence.

Helpful Untruth The evaluation protocol is not meant to be *real*, in the sense that the protocol is actually carried out. Instead, the evaluation protocol is a fictive play, a "helpful untruth" [Vaihinger, 1911]. By proceeding "as if" there were gamblers who play

[6] This is what has been called "empirical evaluation" by Murphy and Epstein [1967]. The same authors name another type of evaluation, which they refer to as "operational", concerned with the value of the forecast to the user.

[7] Nevertheless, we mention the work by Abernethy et al. [2011] and the summary by Perdomo and Recht [2025] which highlight the versatility of online learning algorithms via Blackwell approachability (respectively defensive forecasting) for a large classes of evaluation metrics treated in this work. In particular, [Perdomo and Recht, 2025] makes a related point to ours, in emphasizing the *defining* role of the evaluation for the forecasts.

to test forecasts, we *rec*over existing evaluation metrics and *dis*cover their structural relationships.[8]. For instance, we will equip the gamblers with knowledge about the actual outcomes, knowledge which the forecaster might not have. Clearly, this cannot be true in an actual evaluation scenarios, but it helps to understand what different evaluation metrics do.

11.2 Formal Setup

Let $\mathcal{Y}$ be a compact and metrizable topological space. We call $\mathcal{Y}$ the *outcome set*. Let $\Sigma(\mathcal{Y})$ be the Borel-σ-algebra on $\mathcal{Y}$. The set of signed measures with bounded variation on $\Sigma(\mathcal{Y})$ is denoted as $\text{ca}(\mathcal{Y})$. The set of continuous real-valued functions on $\mathcal{Y}$ is denoted $C(\mathcal{Y})$. Note that every such function is measurable. Furthermore, every such function is integrable with respect to a measure $\phi \in \text{ca}(\mathcal{Y})$, which defines a bilinear mapping,

$$\langle \cdot, \cdot \rangle \colon \text{ca}(\mathcal{Y}) \times C(\mathcal{Y}) \to \mathbb{R}; \quad \langle \phi, f \rangle \coloneqq \int f d\phi.$$

The space $C(\mathcal{Y})$ (respectively $\text{ca}(\mathcal{Y})$) can be normed by the supremum norm $\|f\|_\infty \mapsto \sup_{y \in \mathcal{Y}} f(y)$ (respectively total variation). The space $\text{ca}(\mathcal{Y})$ is isomorphic to the norm dual of $C(\mathcal{Y})$, i.e., every linear continuous mapping $C(\mathcal{Y}) \to \mathbb{R}$ can be written as $f \mapsto \langle \phi, f \rangle$ [Aliprantis and Border, 2006, Theorem 14.15]. Hence, the two spaces $C(\mathcal{Y})$ and $\text{ca}(\mathcal{Y})$ together with the bilinear mapping form a dual pairing [Aliprantis and Border, 2006, Definition 5.90].

The normed space $C(\mathcal{Y})$ (respectively $\text{ca}(\mathcal{Y})$) can be re-topologized in $\sigma(C(\mathcal{Y}), \text{ca}(\mathcal{Y}))$ topology, i.e., the topology which makes all evaluation functionals continuous. An evaluation function is defined as $\phi^* \colon C(\mathcal{Y}) \to \mathbb{R}$, $f \mapsto \langle \phi, f \rangle$, for some $\phi \in \text{ca}(\mathcal{Y})$. Analogously, the space $\text{ca}(\mathcal{Y})$ allows for the $\sigma(\text{ca}(\mathcal{Y}), C(\mathcal{Y}))$ topology, i.e., the topology which makes all evaluation functionals, $f^* \colon \text{ca}(\mathcal{Y}) \to \mathbb{R}$, $\phi \mapsto \langle \phi, f \rangle$ for $f \in C(\mathcal{Y})$, continuous. The topologies $\sigma(C(\mathcal{Y}), \text{ca}(\mathcal{Y}))$ and $\sigma(\text{ca}(\mathcal{Y}), C(\mathcal{Y}))$ are weak topologies in the sense of [Schechter, 1997, p. 758].

The norm-topology and the weak-topology relate to each other as summarized in the following lemma. Set containment of topologies can be understood as fine-graining or strengthening a topology. If topology $\mathcal{T}_1$ is contained in another topology $\mathcal{T}_2$, then all open sets of $\mathcal{T}_1$ are open sets in $\mathcal{T}_2$ [Aliprantis and Border, 2006, p. 24].

Lemma 11.1 (A Hierarchy of Topologies on $C(\mathcal{Y})$)**.** *Let $C(\mathcal{Y})$ be a Banach space defined by*

[8] In this regard, we follow the arguments by Vaihinger [1911] and Appiah [2017] that deliberately wrong assumptions can help to deepen our understanding of the world.

the supremum norm $\|\cdot\|_\infty$ *and* $\mathrm{ca}(\mathcal{Y})$ *the paired space. We have the following containment,*[9]

$$\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y})) \subseteq \sigma(\|\cdot\|_\infty).$$

If $S \subseteq C(\mathcal{Y})$ *is convex and closed with respect to* $\sigma(\|\cdot\|_\infty)$*, then it is closed with respect to* $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$*.*

Proof. The set containment is a special case of [Schechter, 1997, 28.13.b]. The second statement follows from [Schechter, 1997, Theorem 28.14.a], since $C(\mathcal{Y})$ with the norm-topology is a locally convex topological vector space [Schechter, 1997, p. 688]. □

We refer to the set of all non-positive gambles

$$C(\mathcal{Y})_{\leq 0} \coloneqq \{f \in C(\mathcal{Y})\colon f \leq 0\} = \{f \in C(\mathcal{Y})\colon \sup f \leq 0\},$$

i.e., that is the negative orthant. For notational convenience we introduce the following shortcut when A is a subset of $C(\mathcal{Y})$ or $\mathrm{ca}(\mathcal{Y})$,

$$\mathbb{R}_{\geq 0}A \coloneqq \{ra\colon r \in \mathbb{R}_{\geq 0}, a \in A\},$$

which is called the *cone* generated by A. It is closed under multiplication with a positive scalar. If A is convex, i.e., closed under convex combinations, then $\mathbb{R}_{\geq 0}A$ is a *convex cone* (Lemma 11.52).

The subset of probability measures, i.e., $\phi \in \mathrm{ca}(\mathcal{Y})$ such that $\phi(A) \geq 0$ for all $A \in \Sigma(\mathcal{Y})$ and $\mu(\mathcal{Y}) = 1$, on $\Sigma(\mathcal{Y})$ is denoted as $\Delta(\mathcal{Y}) \subseteq \mathrm{ca}(\mathcal{Y})$. If $\phi \in \Delta(\mathcal{Y})$, we define the expectation operator,

$$\mathbb{E}_\phi[f] \coloneqq \langle \phi, f \rangle.$$

For elements in $\mathrm{ca}(\mathcal{Y})$ we use, by default, greek letters, $\phi, \psi \in \mathrm{ca}(\mathcal{Y})$. A special case of a probability distribution is the Dirac-distribution on an element $y \in \mathcal{Y}$ denoted as $\delta(y) \in \Delta(\mathcal{Y})$, which gives $\mathbb{E}_{\delta(y)}[f] = f(y)$. For elements in $C(\mathcal{Y})$, termed "gambles", we use latin letters $f, g \in C(\mathcal{Y})$.

The setup we chose, i.e., the pairing of $C(\mathcal{Y})$ and $\mathrm{ca}(\mathcal{Y})$ spaces, is not the most general framework in which our statements hold. We conjecture that a more general pairing of locally convex topological vector spaces would lead to the same results, given meaningful definitions for properties of distributions exist. The reason for this is that the crucial Bipolar Theorem P3 holds in those more general cases. However, since those general spaces might

[9] We denote the topology induced by the supremum norm as $\sigma(\|\cdot\|_\infty)$.

include the often unfamiliar finitely additive probability measures we decided to simplify statements by restricting ourselves to the $C(\mathcal{Y})$ and $\mathrm{ca}(\mathcal{Y})$ pairing. On the other hand, we need to work with abstract Banach spaces, since we do *not* want to restrict our statements to hold for finite $\mathcal{Y}$. This way our setup encompass regression tasks. If $\mathcal{Y}$ would be finite the topologies of the Banach spaces would collapse to the Euclidean topology and the convexity argument could be carried out in finite dimensional Euclidean spaces. This step mainly simplifies the topological parts of the proofs, i.e., closure of sets, but not the duality statements.

We have summarized the most important definitions used in this paper in Table 11.3. Not all of the denoted concepts have already been introduced, but will be introduced subsequently.

11.3 The Evaluation Protocol

The *evaluation protocol* is at the core of this study. We first give an intuitive introduction of it, which we formalize in Definition 11.2. The evaluation protocol is illustrated in Figure 11.2. Then, we shortly discuss the definition of Forecaster, before we characterize which evaluation metrics can be represented by the evaluation protocol. It turns out that further restrictions need to be set on the gamblers to demarcate reasonable representable evaluation metrics. This will lead us to the characterization of available gambles in the next section.

The evaluation protocol is conducted in (potentially concurrent) rounds, indexed by the round parameter $T \coloneqq \{1, \dots, n\}, n \in \mathbb{N}$. It mainly revolves around three entities: Nature, Forecaster and a set of gamblers.[10] Roughly summarized, the protocol does "as if" Forecaster forecasts the next outcome of Nature, Gambler gambles against this forecast, and lastly Nature reveals an outcome. The gamble is evaluated on the outcome. Finally, the gambles' values are aggregated to Gambler's capital. Semantically, Forecaster's goal is to keep Gambler's capital low, while Gambler's job is to cast doubt on the forecast and increase its own gain. We generally assume that positive outcome values of gambles are "good" and negative are "bad". Hence, we take on Gambler's perspective.

Definition 11.2 (Evaluation Protocol)**.** *We call* $(\mathcal{Y}, \Gamma, T, v, G^I, y)$ evaluation protocol*, where*

(a) $\mathcal{Y}$ *is an outcome space.*

(b) $\Gamma \colon Q \to 2^{\mathcal{V}}$ *is a property (Section 11.3.1) and some* $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$*-closed, convex set of distributions* $Q \subseteq \Delta(\mathcal{Y})$ *and* $\mathcal{V}$ *some set.*

(c) $T \coloneqq \{1, \dots, n\}, n \in \mathbb{N}$ *is a set of instances.*

(d) $v \colon T \to \mathcal{V}$ *is a forecaster.*

(e) $y \colon T \to \mathcal{Y}$ *is a nature.*

(f) $G^I \coloneqq \{G^i \colon i \in I\}$ *for some finite index set* I*, where* $G^i \colon T \to C(\mathcal{Y})$*.*

For every $t \in T$*,*

i) Forecaster forecasts property Γ*, i.e.,* $v_t \in \mathcal{V}$*.*

ii) informed by the forecast, for every $i \in I$*, the gambler* G^i *plays the gamble* $G^i_t \in C(\mathcal{Y})$*.*

[10] We henceforth use a first capital letter when referring to a gambler, a forecaster or a nature, when using the word as name.

iii) Nature reveals an outcome $y_t \in \mathcal{Y}$.

The capital[11] *of the ith gambler is the average realized value of all played gambles, i.e., for all $i \in I$,*

$$K(G^i) \coloneqq \frac{1}{|T|} \sum_{t \in T} G_t^i(y_t).$$

Note that we explicitly make a distinction between generic gambles $g \in C(\mathcal{Y})$, and gambles played by the ith gambler G^i at instance t, which is denoted $G_t^i \in C(\mathcal{Y})$. When there is no index i, then the gambler is denoted G, and its gambler at instance t, G_t.

11.3.1 Forecaster Forecasts a Property Value

What do we mean by "Forecaster forecasts property Γ"? We were motivated to use the construction via Γ because there are plenty of types of forecasts in machine learning. Forecasts in machine learning include full distribution estimates over classes (class probability estimation), means (regression), quantiles (quantile regression), and many more. All those forecasts share that they are properties of a distribution on the outcome set.

Let Q always denote a $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed, convex set of distributions $Q \subseteq \Delta(\mathcal{Y})$ The set Q is the domain of a *property* $\Gamma \colon Q \to 2^{\mathcal{V}}$[12] which is a map to some subset of all possible property values $\mathcal{V}$, e.g., the mean, the median, the mode. We follow the exposition of Gneiting [2011, §2.2] allowing the property to map to *sets* of property values.[13] The property value set $\mathcal{V}$ is deliberately general, as it could be equal to the real line $\mathbb{R}$, e.g., for means (cf. [Gneiting, 2011]), to a finite-dimensional vector space $\mathbb{R}^d$ (cf. [Frongillo and Kash, 2015]), a discrete set $\{1, \ldots, n\}$ (cf. [Lambert, 2019]) or the set of all distributions $\Delta(\mathcal{Y})$ (cf. [Gneiting and Raftery, 2007, Williamson and Cranko, 2023]).

We say that a forecaster *forecasts a property* Γ or *is aligned to property* Γ, if the forecast $v_t \in \mathcal{V}$ provides an estimate of the property Γ for every $t \in T$ [Gneiting, 2011].[14] What matters to us is that a forecasted property value corresponds to a set of probability distributions which possess the announced property value. This is formalized using the pre-image of a property value.

[11] For the sake of readability, we do not make explicit the dependencies of the exact type of the capital $K(G^i)$, which clearly depends on more than only the gambler.

[12] We denote the power set of a set A as 2^A.

[13] Thereby, we allow for an elicitable property with non-strict consistent scoring functions (cf. Definition 11.7). For example, both possible elements in $\{0, 1\}$ are the median of a uniform Bernoulli distribution on this set.

[14] Within the assumption that the forecaster exactly provides its own estimate of the property, we hide the fact that the forecaster can freely choose the predicted value among the set of possible property values. In principle, the forecaster could be bound to a certain hypothesis class which disallows certain predictions.

Definition 11.3 (Level Set of Property)**.** *Let $\gamma \in \mathcal{V}$. The* level set *of a property $\Gamma\colon Q \to 2^{\mathcal{V}}$ is defined as*

$$\Gamma^{-1}(\gamma) \coloneqq \Gamma^{-1}(\{\gamma\}) = \{\phi \in Q\colon \gamma \in \Gamma(\phi)\}.$$

We assume level sets to be non-empty in the following, i.e., $\Gamma(Q) = \mathcal{V}$.

Hence, for every forecast $\gamma \in \mathcal{V}$ there is a corresponding, implicit, set of probability distributions $\Gamma^{-1}(\gamma) \subseteq \Delta(\mathcal{Y})$. We call such a set *forecasting set*. By the generality of the definition of a property, every non-empty set $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ is a forecasting set for some property Γ. If the property Γ is not of interest, we neglect to mention the property and directly refer to the forecasting set $\mathcal{P}$.

Concluding, a forecaster forecasts a property Γ if the forecasts provide an estimate of the property Γ. This estimate is in the set $\mathcal{V}$. Equivalently, one can understand a forecaster as forecasting a set of probability distributions in $\Delta(\mathcal{Y})$ which share a property value. This abstraction step from a single probability distribution to a set of probability distributions will be shown to be necessary to fully capture the variety of evaluation metrics and linking calibration and regret.

11.3.2 The Protocol is Based on a Game-Theoretic Probability Protocol

Our evaluation protocol is closely related to Protocol 6.11 and 6.12 in [Shafer and Vovk, 2019]. However, in contrast to theirs:

Modifications Their forecasts are closed, convex sets of probability distributions. Our forecasts are aligned to properties. They aggregate the capital via addition. We aggregate the capital via average[15]. Furthermore, Shafer and Vovk [2019] leave the gambler with a restricted set of options (called "offer"). So far, we did not incorporate this constraint. However, we will introduce this additional constraint in Section 11.3.4.

Extensions They consider a single gambler, while we treat several gamblers at once. Where we make no assumptions on the exchange of information between Forecaster, the gamblers and Nature, they assume that the involved agents have access to the past moves of all agents.

Purpose We are focusing on different conditions put on a gambler G in order to recover and understand existing evaluation metrics. They aim for a recovery of measure-theoretic probability statements via game-theoretic tools.

[15] Actually, this was mainly a choice of convenience as it makes some of the recovery results (Section 11.5.1) more streamlined.

11.3.3 More Than Single-Instance-Based Evaluation Metrics can be Represented by the Protocol

Having introduced the evaluation protocol, a first simple question is which evaluation metrics are representable via the evaluation protocol?

We understand a general *evaluation metric* as a map, for some finite $T \coloneqq \{1, \dots, n\}$, from outcomes and corresponding forecasts to some real number, $\mathcal{Y}^T \times \mathcal{V}^T \to \mathbb{R}$. We call an evaluation metric *representable in an evaluation protocol* if there exists $(\mathcal{Y}, \Gamma, T, v, G^I, y)$ following Definition 11.2, such that some aggregation $\mathbf{A} \colon \mathbb{R}^{|I|} \to \mathbb{R}$ of the capitals of the gamblers is equal to the evaluation metric. In other words, the representable evaluation metric is "as if" there were gamblers who aggregate their capital.[16] The set of representable evaluation metrics is readily characterized. Essentially, the argument is the (un)currying of the the gamblers' functions.

Proposition 11.4 (Characterization of Representability)**.** *Let $T \coloneqq \{1, \dots, n\}$, $\mathcal{Y}$ be an outcome set and $\mathcal{V}$ a set of property values. An evaluation metric $m \colon \mathcal{Y}^T \times \mathcal{V}^T \to \mathbb{R}$ is* representable in an evaluation protocol*, if and only if, there exists a set of functions $f_t^i \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ indexed by $t \in T$ and $i \in I$ for some finite I and an aggregation $\mathbf{A} \colon \mathbb{R}^{|I|} \to \mathbb{R}$, such that for all $y \in \mathcal{Y}^T$ and $v \in \mathcal{V}^T$,*

$$m((y_t)_{t\in T}, (v_t)_{t\in T}) = \mathbf{A}\left[\left(\frac{1}{|T|}\sum_{t\in T} f_t^i(y_t, v_t)\right)_{i\in I}\right].$$

The proof is deferred to Appendix 11.8.1. In particular, every *single-instance-based evaluation metric* is representable (Appendix 11.8.1, Proposition 11.50). A *single-instance-based evaluation metric* (cf. [Sandroni, 2003, Fortnow and Vohra, 2009]) is a metric which is an aggregation of comparisons between forecast and outcome for each instance.

Despite the generality of evaluation metrics which are representable by the protocol, there exist evaluation metrics which do not fit into this form. In [Sandroni, 2003] the authors consider general evaluation metrics, what they call "tests", which have access to all forecasts and outcomes at once and can compare and analyze predictions and outcomes across rounds. Furthermore, the gamblers only gamble against the actual forecasts for the outcomes and do not have the power to ask for forecasts for unrealized outcomes.[17] In oracle-access outcome indistinguishability (respectively code-access outcome indistinguishability) the "distinguisher", which are (computationally implementable) test functions, have access

[16] We do not delve deeper in the discussion for reasonable aggregation functionals, see [Fröhlich and Williamson, 2024, Höltgen and Williamson, 2023, Cabrera Pacheco et al., 2024] for the discussion of aggregations for evaluation metrics.

[17] In this regard, the forecast evaluation is *prequential* [Dawid and Vovk, 1999], respectively has no oracle-access [Dwork et al., 2021].

to predictions made for unobserved instances (respectively access to the description of the circuit which produced the predictions) [Dwork et al., 2021]. The evaluation protocol is a powerful, but not all-encompassing framework for evaluation metrics.

11.3.4 Fairness of the Gambler is a Minimal Criterion for Reasonable Evaluation Metrics

It is easy to see that the capital K, and hence the aggregated capitals, is an essentially meaningless quantity so far. For instance, a gambler can simply provide a constant positive function as gamble, which does not provide any information about the quality of the forecasts on the respective outcomes. For this reason, let us introduce a first, elementary, minimal criterion to which the gamblers have to adhere.

We define a *fair* gambler as one which only plays gambles *available* to it. Intuitively, a gamble is available if and only if the forecaster (literally) expects the gamble to incur a loss to the gambler for all probabilities in the forecasting set. In other words, Forecaster allows Gambler to play all gambles which Forecaster assumes to be not desirable.[18] We borrow the concept of availability from [Shafer and Vovk, 2019, Protocol 6.11 / 6.12].

Definition 11.5 (Fair Gambler, Availability, Set of Available Gambles)**.** *In an evaluation protocol* $(\mathcal{Y}, \Gamma, T, v, G^I, y)$*, the* i*th* $(i \in I)$ *gambler* G^i *is* fair*, if and only if, for all* $t \in T$*,* G^i_t *is* available*, i.e.,*

$$\sup_{\phi \in \Gamma^{-1}(v_t)} \mathbb{E}_\phi[G^i_t] \leq 0.$$

For the set of available gambles *at time* $t \in T$ *we write,*

$$\mathcal{G}_{\Gamma^{-1}(v_t)} := \left\{ g \in C(\mathcal{Y}) \colon \sup_{\phi \in \Gamma^{-1}(v_t)} \mathbb{E}_\phi[g] \leq 0 \right\}$$

A gambler which violates the availability criterion has an unfair advantage in increasing its capital against Forecaster. Hence, if we interpret the gambler's capital as an evaluation metric, this metric would not serve the purpose to guarantee "good" forecasts, since the resulting capital would most likely not relate to a quality of the forecasts.

The availability criterion is a necessary, but not sufficient meta-criterion for reasonable forecast evaluators in our protocol. Assuming that the forecaster is "clairvoyant", i.e., it does know the next outcome of Nature, we would expect any reasonable evaluation metric to

[18] *Availability* is negative "desirability"[Shafer and Vovk, 2019, p. 131], a concept used in the literature on imprecise probability [Walley, 2000]. Desirable gambles are all gambles for which the forecaster assumes that playing those will not lead to a loss, i.e., choosing a desirable gamble and evaluating this gamble at an outcome of Nature will lead to a "benign" outcome.

state that the forecasts are perfect. The following proposition shows that this is equivalent to fairness of the gambler.

Proposition 11.6 (Sanity Check)**.** *Let $(\mathcal{Y}, \Gamma, T, v, G^I, y)$ be an evaluation protocol. Let v be "clairvoyant", i.e., $v_t \coloneqq \Gamma(\delta(y_t))$*[19]*. The capital of any gambler G^i with $i \in I$ is lower or equal to zero, i.e., $K(G^i) \leq 0$, if and only if, the gambler is fair.*

Proof. Fix any gambler G^i for some $i \in I$. We first show that the capital of the gambler is upper bounded, if the gambler only plays available gambles. The set of gambles available to the gambler is strongly restricted: for any $G_t^i \in \mathcal{G}_{\Gamma^{-1}(v_t)}$ it holds, $\mathbb{E}_{\delta(y_t)}[G_t^i] \leq 0$, which is equivalent to $G_t^i(y_t) \leq 0$. Hence,

$$K(G^i) = \frac{1}{|T|} \sum_{t \in T} G_t^i(y_t) \leq 0.$$

We show the reverse direction by contraposition. Assume that the capital of the ith gambler is larger than zero. Then there exists a gamble G_t^i, such that $G_t^i(y_t) > 0$. However, this implies that $\mathbb{E}_{\delta(y_t)}[G_t^i] > 0$, from which follows that $G_t^i \notin \mathcal{G}_{\Gamma^{-1}(v_t)}$. □

Fairness of the gambler can thus be interpreted as Type-I error freeness. The null hypothesis, i.e., the forecast, cannot be rejected via available gambles, if the null hypothesis is true. In fact, this interpretation is well-founded, as availability is tightly related to game-theoretic hypothesis testing (Section 11.6.1).

However, a fair gambler does not necessarily guarantee the quality of forecasts either. A forecast can still be uninformative, even though several fair gamblers didn't obtain a large capital. For instance, the forecaster who predicts a property whose level set comprises of all distributions on $\mathcal{Y}$, i.e., a vacuous forecast, will enforce any fair gambler to play non-positive gambles (Lemma 11.53). Hence, the gambler's capital stays non-positive $K \leq 0$. However, the forecasts might not necessarily be "good", in terms of describing the outcomes by Nature appropriately.

The difficulty of evaluating arbitrary set-valued probabilistic forecasts is a known issue and part of a longer debate on scoring rules for imprecise probabilities (cf. Appendix 11.8.4). In conclusion, availability is necessary, but not sufficient to guarantee "good" forecasts. It is a meta-criterion for reasonable evaluation metrics representable by our evaluation protocol.

[19] We assume that $\delta(y_t) \in Q$ for all $t \in T$ to guarantee that the forecaster is well-defined.

11.4 Characterization of Available Gambles

The evaluation protocol is a general tool to describe evaluation metrics (Section 11.3.3) based on protocols from game-theoretic probability (Section 11.3.2). The *fairness* criterion on gambler is a minimal sanity check, that every reasonable evaluation metric should fulfill (Section 11.3.4). However, it remains unclear whether existing evaluation metrics are representable and fulfill the fairness criterion? Towards answering this question, we will first respond to: **What is the potential of fair gamblers to disprove Γ-aligned forecasters? Which gambles *can* be played?**

In other words, we ask for a characterization of the set of available gambles. For general Γ this question can only be answered to a limited extent. Essentially, it is possible to argue that the set of available gambles is equal for some level set of Γ and the closed, convex hull of the very same level set (Section 11.4.6). For a more exhaustive description we have to restrict ourselves to certain properties Γ which fulfill additional constraints.

As we will argue, *elicitable* (respectively *identifiable*) Γ do not only allow for clean characterizations of the corresponding sets of available gambles, but as well form the two pillars on which regret-type (respectively calibration-type) evaluation metrics are based. The statements and relationships are summarized in Figure 11.1.

For instance, we will show that *(unscaled) regret gambles* such as,

$$\{y \mapsto (y-\gamma)^2 - (y-c)^2 \colon c \in \mathcal{V}\}, \tag{11.2}$$

are available, if the property Γ, which is forecasted, is *elicited* by the loss function $\ell \colon y \mapsto (y-c)^2$ (cf. Section 10.1). We are interested in *regret gambles* as they mimic the structure of regret-type evaluation metrics such as external or swap regret (Section 11.5.2). And as it turns out, what elicitability is for regret-type evaluation metrics, identifiability is for calibration-type evaluation metrics [Noarov and Roth, 2023]. Let us formally introduce those concepts.

11.4.1 Elicitable and Identifiable Properties

We have argued that machine learning systems generally forecast properties (denoted Γ), i.e., mappings from a set of distributions to subsets of property values. We have given examples, such as the mean, the median, or a full class probability estimate. Those specific properties are furthermore *elicitable* (respectively *identifiable*). Elicitability guarantees that the property is the solution of an expected loss minimization problem [Osband, 1985]. A property which is identifiable can be characterized by a so-called identification function [Osband, 1985, Gneiting, 2011].

Definition 11.7 (Elicitability and Consistent Scoring Function)**.** *Let* $\Gamma\colon Q \to 2^{\mathcal{V}}$ *be a property. A function* $\ell\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ *is called a* consistent scoring function *for* Γ *if, for all* $\gamma \in \mathcal{V}$, $\ell_\gamma\colon y \mapsto \ell(y,\gamma) \in C(\mathcal{Y})$, *and*

$$\mathbb{E}_\phi[\ell_\gamma] \leq \mathbb{E}_\phi[\ell_c]$$

for all $\phi \in Q, \gamma \in \Gamma(\phi), c \in \mathcal{V}$. *We can equivalently write*

$$\Gamma(\phi) \subseteq \arg\min_{c\in\mathcal{V}} \mathbb{E}_\phi[\ell_c].$$

A property which has a consistent scoring function is called elicitable.

Definition 11.8 (Identifiability and Identification Function)**.** *Let* $\Gamma\colon Q \to 2^{\mathcal{V}}$ *be a property. A function* $\nu\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ *is called an* identification function *of* Γ *if, for all* $\gamma \in \mathcal{V}$, $\nu_\gamma\colon y \mapsto \nu(y,\gamma) \in C(\mathcal{Y})$, *and*

$$\mathbb{E}_\phi[\nu_\gamma] = 0 \Leftrightarrow \gamma \in \Gamma(\phi),$$

for all $\phi \in Q$. *A property which has an identification function is called* identifiable.

The prototypical example of an elicitable and identifiable property is the mean. A corresponding scoring function is $\ell(y,\gamma) = (y-\gamma)^2$. An identification function is $\nu(y,\gamma) = y-\gamma$. Another standard example is the τ-pinball loss which elicits the τ-quantile [Gneiting, 2011]. Furthermore, proper scoring rules [Williamson and Cranko, 2023] are consistent scoring functions. They elicit the entire distribution [Savage, 1971]. The corresponding property is the identity function, i.e., $\Gamma_{id}\colon Q \to 2^Q, \phi \mapsto \{\phi\}$.

When $\mathcal{V} = \mathbb{R}$, elicitable properties are, neglecting technicalities, identifiable and vice-versa [Lambert et al., 2008, Steinwart et al., 2014]. This is not necessarily true for more general $\mathcal{V} \subseteq \mathbb{R}^d$ [Fissler and Ziegel, 2016]. It is a matter of simple computations to show that elicitable properties and identifiable properties have convex level sets. This result goes back at least to [Osband, 1985].[20] Furthermore, the corresponding level sets are closed. Hence, if the forecaster provides a property value $v \in \mathcal{V}$ for an elicitable or identifiable property Γ, then implicitly this corresponds to a closed, convex forecasting set $\Gamma^{-1}(v) \subseteq \Delta(\mathcal{Y})$.[21]

Proposition 11.9 (Convex and Closed Level Sets)**.** *For all* $\gamma \in \mathcal{V}$ *the level set* $\Gamma^{-1}(\gamma) \subseteq \Delta(\mathcal{Y})$ *of an elicitable or identifiable property* Γ *is credal, i.e., convex and* $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$*-closed.*

The proof of the statement is deferred to Appendix 11.8.2. The statement has been shown

[20] For finite $\mathcal{Y}$ and finite $\mathcal{V}$ level sets are not only convex but are the result of a Voronoi-partition intersected with the simplex [Lambert, 2019].

[21] We call such sets *credal sets* (Definition 11.20) for reasons explained in Section 11.4.6.

under a variety of differing technical setups. We provide a proof as we have not found any such statement for the dual pairing between $\mathrm{ca}(\mathcal{Y})$ and $C(\mathcal{Y})$. The proof idea is not novel. The details are our contribution.

11.4.2 Warm-Up: Regret Gambles and Calibration Gambles are Available

Elicitable properties are minimizers of loss functions. Identifiable properties are identified by a zero-condition. What do those characteristics of properties have to do with regret and calibration?

In the following warm-up, we shortly argue that *(unscaled) regret gambles*, for some $\gamma \in \mathcal{V}$,

$$\mathcal{L}_\gamma \coloneqq \{\ell_\gamma - \ell_c \colon c \in \mathcal{V}\} \tag{11.3}$$

and *calibration gambles*, for some $\gamma \in \mathcal{V}$,

$$V_\gamma \coloneqq \{\alpha \nu_\gamma \colon \alpha \in \mathbb{R}\}, \tag{11.4}$$

are *available* if the forecasted property Γ is *elicitable* (respectively *identifiable*). Those structured gambles, i.e., unscaled regret gambles and calibration gambles, are the atoms of regret-type and calibration-type notions, as we will argue exhaustively in Section 11.5. However, we will need a more detailed understanding of *all* available gambles before we can fully recover used evaluation metrics.

Proposition 11.10 ((Unscaled) Regret Gambles are Available)**.** *Let $\Gamma \colon Q \to 2^{\mathcal{V}}$ be an elicitable property with consistent scoring function $\ell \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$. For every $\gamma \in \mathcal{V}$ such that $\Gamma^{-1}(\gamma) \neq \emptyset$,*

$$\mathcal{L}_\gamma \subseteq \mathcal{G}_{\Gamma^{-1}(\gamma)}.$$

Proof. The statement is a direct consequence of the consistency of the scoring function ℓ. For every $g \in \mathcal{L}_\gamma$, $g = \ell_\gamma - \ell_c$ for some $c \in \mathcal{V}$, and so,

$$\sup_{\phi \in \Gamma^{-1}(\gamma)} \mathbb{E}_\phi[g] = \sup_{\phi \in \Gamma^{-1}(\gamma)} \mathbb{E}_\phi[\ell_\gamma - \ell_c] = \sup_{\phi \in \Gamma^{-1}(\gamma)} \mathbb{E}_\phi[\ell_\gamma] - \mathbb{E}_\phi[\ell_c] \leq 0,$$

as $\mathbb{E}_\phi[\ell_\gamma] \leq \mathbb{E}_\phi[\ell_c]$ for all $\phi \in \Gamma^{-1}(\gamma)$, see Definition 11.7. □

Proposition 11.11 (Calibration Gambles are Available)**.** *Let $\Gamma \colon Q \to 2^{\mathcal{V}}$ be an identifiable property with identification function $\nu \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$. For every $\gamma \in \mathcal{V}$ such that $\Gamma^{-1}(\gamma) \neq \emptyset$,*

$$V_\gamma \subseteq \mathcal{G}_{\Gamma^{-1}(\gamma)}.$$

Proof. The statement is a consequence of the properties of the identification function ν. For every $g \in V_\gamma$, $g = \alpha\nu_\gamma$ for some $\alpha \in \mathbb{R}$, and so,

$$\sup_{\phi \in \Gamma^{-1}(\gamma)} \mathbb{E}_\phi[g] = \sup_{\phi \in \Gamma^{-1}(\gamma)} \mathbb{E}_\phi[\alpha\nu_\gamma] = \sup_{\phi \in \Gamma^{-1}(\gamma)} \alpha\mathbb{E}_\phi[\nu_\gamma] = 0,$$

as $\mathbb{E}_\phi[\nu_\gamma] = 0$ for all $\phi \in \Gamma^{-1}(\gamma)$, see Definition 11.8. □

Hence, unscaled regret gambles and calibration gambles are in the arsenal of a fair gambler to discredit a forecaster aligned to an elicitable (respectively identifiable) property. Nevertheless, the boundaries of availability offer more freedom than the suggested gambles above. In fact, for forecasts corresponding to elicitable (respectively identifiable) properties we can characterize the full set of available gambles. This way we argue that in principle the gamblers can leverage regret-gambles or calibration-gambles equivalently, i.e., with the same ability to test forecasts, when the regret-gambles are allowed to be scaled.

11.4.3 Characterizing Available Gambles of Forecasts of Elicitable Properties

If the property Γ is elicitable, then, so we argue, *scaled regret gambles*,

$$\mathcal{S}_\gamma \coloneqq \{\beta(\ell_\gamma - \ell_c) \colon \beta \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\}, \tag{11.5}$$

dominate, up to approximation, all available gambles.

In order to provide a full characterization of available gambles for forecasts aligned to elicitable properties we have to introduce the concept of a superprediction set. This concept originates from the literature on proper scoring rules [Kalnishkan and Vyugin, 2002, Kalnishkan et al., 2004, Dawid, 2007].

Definition 11.12 (Superprediction Set)**.** *Let $\ell \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ be a scoring function. We call*

$$\operatorname{spr}(\ell) \coloneqq \{g \in C(\mathcal{Y}) \colon \exists c \in \mathcal{V}, \ell_c \leq g\},$$

the superprediction set *of ℓ.*

The superprediction set consists of all gambles which incur no less loss than for some $c \in \mathcal{V}$. Positive values have to be interpreted as "bad" here. Superprediction sets largely describe the behavior of loss functions [Williamson and Cranko, 2023].

For our characterization we require convexity of the superprediction set. Convexity of the superprediction set is not a strong assumption for consistent scoring functions. For instance, every consistent scoring function induces a proper scoring rule, e.g., [Gneiting,

2011, Theorem 3], which, in finite dimensions, has a surrogate proper scoring rule with a convex superprediction set with same conditional Bayes risk [Williamson and Cranko, 2023, Section 3.4]. The proper scoring rule and its surrogate are equivalent almost everywhere [Williamson et al., 2016, Proposition 8]. On the other hand, it is possible to directly convexify the superprediction set by taking the convex closure of $\mathcal{V}$ and extending the scope of ℓ (cf. "randomized acts" in Section 3.2 in [Grünwald and Dawid, 2004]).

With this tool at hand we can provide a full characterization of the set of available gambles in the case that the property Γ is elicitable. The proof is provided in Section 11.4.7. It requires the introduction of more concepts from the literature on imprecise probabilities (Section 11.4.6).

Theorem 11.13 (Available Gambles of Elicitable Property-Forecasts)**.** *Let $\Gamma\colon Q \to 2^{\mathcal{V}}$ be an elicitable property with scoring function $\ell\colon \mathcal{Y}\times\mathcal{V} \to \mathbb{R}$ which has a convex superprediction set. For a fixed $\gamma \in \mathcal{V}$ such that $\Gamma^{-1}(\gamma) \neq \emptyset$, we define*

$$\mathcal{H}_{\ell_\gamma} \coloneqq \mathrm{cl}\{g \in C(\mathcal{Y})\colon g \leq \beta(\ell_\gamma - \ell_c) \text{ for some } \beta \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\}. \tag{11.6}$$

Then,

$$\mathcal{H}_{\ell_\gamma} = \mathcal{G}_{\Gamma^{-1}(\gamma)}.$$

Hence, the set of available gambles can be, up to approximation in $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$-topology, expressed as all gambles which are upper bounded by some *scaled regret gambles* (Equation (11.5)). It turns out that the closure is crucial here. Even in the finite dimensional case in a simple example, it is easy to show that $\{g \in C(\mathcal{Y})\colon g \leq \alpha(\ell_\gamma - \ell_c)$ for some $\alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\}$ is not closed.

Example 11.14. *Let us $\mathcal{Y} \coloneqq \{0,1\}$ and Γ be the mean with consistent scoring function $\ell(y,\gamma) \coloneqq (y-\gamma)^2$. In this simple setup, the mean identifies the distribution. For simplification we assume that Forecaster states $v = 0.4$, i.e., the forecasting set is the singleton set with the distribution $(0.6, 0.4) \in \Delta(\mathcal{Y})$.*

First, we observe that in our simplified setup $C(\mathcal{Y}) \equiv \mathbb{R}^2, \mathrm{ca}(\mathcal{Y}) \equiv \mathbb{R}^2$, $\leq$ is the coordinate-wise order and the considered topology is the Euclidean topology. For a detailed discussion of those simplifications see the proof of Corollary 11.16. We further leverage an analogous representation of the set

$$\{g \in C(\mathcal{Y})\colon g \leq \alpha(\ell_\gamma - \ell_c) \text{ for some } \alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\} = \mathcal{S}_\gamma + C(\mathcal{Y})_{\leq 0},$$

where $C(\mathcal{Y})_{\leq 0} \coloneqq \{g \in \mathbb{R}^2\colon g \leq 0\}$ is the negative orthant. We show that S_γ is open and

argue that this implies $S_\gamma + C(\mathcal{Y})_{\leq 0}$ is open. Some calculations give

$$\mathcal{S}_{0.4} = \left\{\alpha(0.16 - c^2, 0.16 - (1-c)^2) \colon c \in [0,1], \alpha \in \mathbb{R}_{\geq 0}\right\}.$$

Furthermore, solving for $c \in [0,1]$ and $\alpha \in \mathbb{R}$ one notices $(-0.4, 0.6) \notin \mathcal{S}_{0.4}$. It would require $c = 0.4$ with $\alpha = \infty$. However, for every $\epsilon > 0$, we can give a point $\ell_\epsilon \in S_{0.4}$ such that $\|(-0.4, 0.6) - \ell_\epsilon\|_2 \leq \frac{1}{\sqrt{2}}\epsilon$. For instance, such a point is given by,

$$\ell_\epsilon \coloneqq \beta_\epsilon(0.16 - c_\epsilon^2, 0.36 - (1-c_\epsilon)^2) \tag{11.7}$$

for $c_\epsilon \coloneqq 0.4 + \epsilon$ and $\beta_\epsilon \coloneqq \frac{1}{2\epsilon}$. Concluding, the set $\mathcal{S}_{0.4}$ is arbitrarily close to $(-0.4, 0.6)$ but $(-0.4, 0.6)$ is not included. The Minkowski-sum with $C(\mathcal{Y})_{\leq 0}$ does not change this fact, since it only adds points which are smaller. However, the points we gave are approximating $(-0.4, 0.6)$ from below. See Figure 11.3 for an illustration.

Identifiable properties are closely related to elicitable properties [Steinwart et al., 2014]. Hence, we pursue the natural follow-up question of the previous section: what is the set of available gambles in the case the forecasts are aligned to an identifiable property?

11.4.4 Characterizing Available Gambles of Forecasts of Identifiable Properties

We show that if Γ is an identifiable property with identification function $\nu \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$, the set of available gambles for the property-induced forecasting set is defined by all gambles $g \leq \alpha\nu_\gamma$ for some $\alpha \in \mathbb{R}$, where $\gamma \in \mathcal{V}$ is the forecasted property, or it can be approximated by such g. In other words, every available gamble is (approximately) dominated by a *calibration gamble*.

Again, the proof is deferred to later (Section 11.4.8). The characterization follows analogous steps in comparison to the proof of Theorem 11.13.

Theorem 11.15 (Available Gambles of Identifiable Property-Forecasts)**.** *Let $\Gamma \colon Q \to 2^{\mathcal{V}}$ be an identifiable property with identification function $\nu \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$. For a fixed $\gamma \in \mathcal{V}$, we define*[22]

$$\mathcal{H}_{\nu_\gamma} \coloneqq \mathrm{cl}\{g \in C(\mathcal{Y}) \colon g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\}. \tag{11.8}$$

Then,

$$\mathcal{H}_{\nu_\gamma} = \mathcal{G}_{\Gamma^{-1}(\gamma)}.$$

Unlike the characterization in Theorem 11.13, the closure of the set of gambles upper

[22] Closure is taken with respect to $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$-topology.

bounded by calibration gambles can be neglected in finite dimensions. We can show that $\mathcal{H}_{\nu_\gamma}$ is closed, if $|\mathcal{Y}| < \infty$.

Corollary 11.16 (Theorem 11.15 in Finite Dimensions)**.** *Let $\mathcal{Y}$ be finite. Let $\Gamma\colon Q \to 2^{\mathcal{V}}$ be an identifiable property with identification function $\nu\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$. For a fixed $\gamma \in \mathcal{V}$,*

$$\mathcal{H}_{\nu_\gamma} = \{g \in C(\mathcal{Y})\colon g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\} = \mathcal{G}_{\Gamma^{-1}(\gamma)}.$$

11.4.5 Proving the Characterizations

Let us now turn to proving the stated theorems. In the next three subsections, we first show that for a general forecasting set $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ (Section 11.3.1) the set of available gambles forms a structured convex cone called an *offer*. We then demonstrate that if the forecasting set is convex and closed, then the offer is in one-to-one correspondence with the forecasting set, i.e., the forecasting set characterizes the offer, and the offer characterizes the forecasting set. Finally, we leverage the obtained results to characterize the set of available gambles for elicitable and identifiable properties.

The proof ideas for the statements in Section 11.4.6 are not new (e.g., [Föllmer and Schied, 2011, Benavoli et al., 2017]). However, we are not aware of any work which has proved Theorem 11.18 and Theorem 11.22 for the $C(\mathcal{Y})$-ca$(\mathcal{Y})$-duality.

11.4.6 The Set of Available Gambles is a Structured Convex Cone Called Offer

Every set of available gambles possess a specific structure lent from convex analysis. They form what we call an *offer*. We borrow this term from [Shafer and Vovk, 2019, § 6.4] who use it for an analogous structure to which we compare below.

Definition 11.17 (Offer)**.** *We call a set $\mathcal{G} \subseteq C(\mathcal{Y})$ an* offer *if all of the following conditions are fulfilled:*

O1. *If $g, f \in \mathcal{G}$, then $g + f \in \mathcal{G}$.* *(Additivity)*

O2. *If $f \in \mathcal{G}$ and $\alpha \geq 0$, then $\alpha f \in \mathcal{G}$.* *(Positive Homogeneity)*

O3. *If $g \in \mathcal{G}$, then $\inf g \leq 0$.* *(Prohibiting Sure Gains)*

O4. *Let $g \in C(\mathcal{Y})$, if $\sup g \leq 0$, then $g \in \mathcal{G}$.* *(Default Availability)*

O5. *The set $\mathcal{G}$ is closed with respect to the $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$ topology.* *(Closure)*

Closure, additivity and positive homogeneity demand that an offer $\mathcal{G}$ is a closed, convex cone containing the zero element in $C(\mathcal{Y})$. Default availability then implies that all non-positive gambles are available. "Prohibiting Sure Gains" ensures that Gambler does not safely increase its capital via a gamble.[23]

Definition 11.17 is closely related to the definition of offer [Shafer and Vovk, 2019, § 6.4]. First, our definition works for gambles in $C(\mathcal{Y})$-spaces, while Shafer and Vovk [2019] consider more generally functions from $\mathcal{Y}$ to the extended real line.[24] Their axiom G1 is equivalent our O1. Their axiom G2 is equivalent to our O2. Their axiom G4 is equivalent to our O3. Their axiom G3 together with G2 implies our O4. Our O4 together with O1 implies their axiom G3. Finally, their axiom G0 and G5 are closure conditions analogous to our O5 adapted to the respective function spaces. As detailed in [Shafer and Vovk, 2019, § 6.7] offers are the sign-flipped counterparts of sets of desirable gambles [Walley, 1991].

Theorem 11.18 (Available Gambles Form an Offer)**.** *Let $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ be a forecasting set. The set,*

$$\mathcal{G}_{\mathcal{P}} := \left\{ g \in C(\mathcal{Y}) \colon \sup_{\phi \in \mathcal{P}} \mathbb{E}_\phi[g] \leq 0 \right\} \subseteq C(\mathcal{Y}),$$

is an offer.

In order to provide the proof, we introduce so-called *upper expectations* which are closely related to forecasting sets.

Definition 11.19 (Upper Expectation)**.** *A functional $\overline{\mathbb{E}} \colon C(\mathcal{Y}) \to (-\infty, \infty]$ for which the following axioms hold is called an* upper expectation*.*

UE1. *If $f_1, f_2 \in C(\mathcal{Y})$, then $\overline{\mathbb{E}}[f_1 + f_2] \leq \overline{\mathbb{E}}[f_1] + \overline{\mathbb{E}}[f_2]$. (Subadditivity)*

UE2. *If $f \in C(\mathcal{Y})$ and $c \in [0, \infty)$, then $\overline{\mathbb{E}}[cf] \leq c\overline{\mathbb{E}}[f]$. (Positive Homogeneity)*

UE3. *If $f_1, f_2 \in C(\mathcal{Y})$ such that $f_1 \leq f_2$, then $\overline{\mathbb{E}}[f_1] \leq \overline{\mathbb{E}}[f_2]$. (Monotonicity)*

UE4. *If $c \in \mathbb{R}$, then $\overline{\mathbb{E}}[c] = c$. (Translation Equivariance)*

UE5. *For every $\alpha \in \mathbb{R}$ the set $\{f \in C(\mathcal{Y}) \colon \overline{\mathbb{E}}[f] \leq \alpha\}$ is $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$-closed. (Lower Semicontinuity)*

[23] There are subtle differences between our notion of "Prohibiting Sure Gains" and the related notion of "avoiding sure loss" in imprecise probability, e.g., [Walley, 1991, Section 3.7.3]. Mainly, our definition contains a sign flip.

[24] The reason why we restrict ourselves to a smaller function space of gambles is that otherwise the representability of upper expectations (Definition 11.19) as support functions over closed, convex sets of probability distributions requires more technical tools which are beyond the scope of this paper, cf. [Delbaen, 2002, Section 5].

We first show that $\overline{\mathbb{E}}_{\mathcal{P}}[g] \coloneqq \sup_{\phi\in\mathcal{P}} \mathbb{E}_\phi[g]$ for any $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ is an upper expectation. Then, the set of all gambles with non-positive upper expectation form an offer.

Proof. Let $\overline{\mathbb{E}}_{\mathcal{P}}[g] \coloneqq \sup_{\phi\in\mathcal{P}} \mathbb{E}_\phi[g]$. We observe that [Aliprantis and Border, 2006, p. 251]

$$\overline{\mathbb{E}}_{\mathcal{P}}[g] \coloneqq \sup_{\phi\in\mathcal{P}} \mathbb{E}_\phi[g] = \sup_{\phi\in\overline{\text{co}}\mathcal{P}} \mathbb{E}_\phi[g],$$

where $\overline{\text{co}}$ denotes the $\sigma(\text{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closure of the convex hull of $\mathcal{P}$.

Hence, we can apply the fundamental representation result for support functions and closed convex sets [Aliprantis and Border, 2006, Theorem 7.51]. It guarantees that $\overline{\mathbb{E}}_{\mathcal{P}}$ satisfies axioms UE1,UE2 and UE5. Since, $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ it follows $\overline{\text{co}}\mathcal{P} \subseteq \Delta(\mathcal{Y})$ because $\Delta(\mathcal{Y})$ is $\sigma(\text{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed [Aliprantis and Border, 2006, Theorem 15.11] and convex. Translation equivariance UE4 follows because, for all $c \in \mathbb{R}_{\geq 0}$,

$$\sup_{\phi\in\mathcal{P}} \mathbb{E}_\phi[c] \overset{UE2}{=} c \sup_{\phi\in\mathcal{P}} \mathbb{E}_\phi[1] = c,$$

because $\mathcal{P} \subseteq \Delta(\mathcal{Y})$, respectively, for all $c \in \mathbb{R}_{<0}$,

$$\sup_{\phi\in\mathcal{P}} \mathbb{E}_\phi[c] \overset{UE2}{=} c \inf_{\phi\in\mathcal{P}} \mathbb{E}_\phi[1] = c,$$

because $\mathcal{P} \subseteq \Delta(\mathcal{Y})$. Monotonicity UE3 follows because $\mathcal{P} \subseteq \text{ca}(\mathcal{Y})_{\geq 0}$.

It remains to show that the set of gambles for which the upper expectation is non-positive form an offer. Let $\mathcal{G}_{\mathcal{P}} \coloneqq \{g \in C(\mathcal{Y}) \colon \overline{\mathbb{E}}_{\mathcal{P}}[g] \leq 0\}$. We step-by-step prove all axioms that $\mathcal{G}_{\mathcal{P}}$ has to fulfill to be an offer. Axiom O1 follows from axiom UE1. Axiom O2 follows from axiom UE2. Axiom O3 is given, because if $\inf g > 0$, then $\overline{\mathbb{E}}[g] \geq \overline{\mathbb{E}}[\inf g] > 0$ by axiom UE3 and axiom UE4, from which follows that $g \notin \{g \in C(\mathcal{Y}) \colon \overline{\mathbb{E}}[g] \leq 0\}$. Axiom O4 follows from axiom UE3 and axiom UE4. Let $g \in C(\mathcal{Y})$ with $\sup g \leq 0$. Then $\overline{\mathbb{E}}[g] \leq \overline{\mathbb{E}}[\sup g] \leq 0$ implies $g \in \mathcal{G}_{\overline{\mathbb{E}}}$. Axiom O5 directly follows from axiom UE5. □

The set of all available gambles is an offer. This reformulation, in fact, is not a one-way street. Given an arbitrary offer we can provide a forecasting set whose set of available gambles is exactly the offer (Theorem 11.22). This forecasting set is convex and closed. We call such forecasting sets *credal*.

Definition 11.20 (Credal Set)**.** *We call a non-empty set $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ a* credal set *if all of the following conditions are fulfilled*

CS1. *Let $p_1, \ldots, p_n \in \mathcal{P}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}_{\geq 0}$ such $\sum_{i=1}^n \alpha_i = 1$, then $\sum_{i=1}^n \alpha_i p_i \in \mathcal{P}$. (Convexity)*

CS2. *The set $\mathcal{P}$ is closed with respect to the $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-topology.* *(Closure)*

The term "credal set" is borrowed from the literature on imprecise probability [Augustin et al., 2014]. We emphasize that a credal set is not necessarily linked to any kind of a belief of Forecaster. We use the term here to describe a forecasting set with a particular structure. Note that $\overline{\mathrm{co}}\mathcal{P}$ is a credal set for every forecasting set $\mathcal{P} \subseteq \Delta(\mathcal{Y})$.

To formally state Theorem 11.22 we introduce polar sets.

Definition 11.21 (Polar Set)**.** *We define the polar set for non-empty $W \subseteq C(\mathcal{Y})$ and non-empty $V \subseteq \mathrm{ca}(\mathcal{Y})$ as*

$$W^\circ := \{\phi \in \mathrm{ca}(\mathcal{Y}) \colon \langle \phi, g \rangle \leq 1, \forall g \in W\},$$

respectively

$$V^\circ := \{g \in C(\mathcal{Y}) \colon \langle \phi, g \rangle \leq 1, \forall \phi \in V\}.$$

Finally, we are able to spell out a one-to-one correspondence between credal sets and offers.

Theorem 11.22 (Representation: Credal Sets – Offers)**.** *Let $\mathcal{G} \subseteq C(\mathcal{Y})$ be an offer. The set*

$$\mathcal{P}_\mathcal{G} := \mathcal{G}^\circ \cap \Delta(\mathcal{Y}) \subseteq \mathrm{ca}(\mathcal{Y})$$

is a credal set, such that

$$\mathcal{G} = \left\{ g \in C(\mathcal{Y}) \colon \sup_{\phi \in \mathcal{P}_\mathcal{G}} \mathbb{E}_\phi[g] \leq 0 \right\} = (\mathbb{R}_{\geq 0}\mathcal{P}_\mathcal{G})^\circ.$$

Conversely, let $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ be a credal set. The set

$$\mathcal{G}_\mathcal{P} := (\mathbb{R}_{\geq 0}\mathcal{P})^\circ \subseteq C(\mathcal{Y})$$

is an offer, such that

$$\mathcal{P} := \mathcal{G}_\mathcal{P}^\circ \cap \Delta(\mathcal{Y}).$$

Thus, every offer $\mathcal{G}$ is in one-to-one correspondence to a credal set $\mathcal{P}$.

In order to show Theorem 11.22, we use several helpful properties of polar sets.

Proposition 11.23 (Properties of Polar Set)**.** *For non-empty $W_1, W_2 \subseteq C(\mathcal{Y})$ we have:*

P1. *If $W_1 \subseteq W_2$, then $W_2^\circ \subseteq W_1^\circ$.*

P2. *The set W_1° is non-empty, convex, $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed and contains the zero element.*

P3. *If W_1 is non-empty, convex, $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$-closed and contains the zero element, then $W_1 = (W_1^\circ)^\circ$.*

P4. *If W_1 is a convex cone, then $W_1^\circ = \{\phi \in \mathrm{ca}(\mathcal{Y})\colon \langle \phi, g\rangle \leq 0, \forall g \in W_1\}$, which is a convex cone.*

P5. $(C(\mathcal{Y})_{\leq 0})^\circ = \mathrm{ca}(\mathcal{Y})_{\geq 0}$.

Analogous results hold for $V_1, V_2 \subseteq \mathrm{ca}(\mathcal{Y})$.

Proof. Statement P1 and P2 follow from Lemma 5.102 in [Aliprantis and Border, 2006]. Statement P3 is the famous Bipolar Theorem [Aliprantis and Border, 2006, Theorem 5.103]. The Statement P4 can be found on p. 215 in the same book. It remains to show Statement P5.

Since $C(\mathcal{Y})_{\leq 0}$ is a convex cone, we leverage Property P4 to get:

$$(C(\mathcal{Y})_{\leq 0})^\circ = \{\phi \in \mathrm{ca}(\mathcal{Y})\colon \langle \phi, g\rangle \leq 0, \forall g \in C(\mathcal{Y})_{\leq 0}\}.$$

We can easily see that for every $\phi \in \mathrm{ca}(\mathcal{Y})_{\geq 0}$, $\langle \phi, g\rangle \leq 0$ for all $g \in C(\mathcal{Y})_{\leq 0}$. For the reverse direction, consider any $\phi \in \mathrm{ca}(\mathcal{Y})$ such that ϕ is negative on a measurable set A. We show that such $\phi \notin (C(\mathcal{Y})_{\leq 0})^\circ$. Let $g(y) \coloneqq -[\![y \in A]\!]$[25]. Obviously, $g \in C(\mathcal{Y})_{\leq 0}$. Furthermore,

$$\langle \phi, g\rangle = \int_{\mathcal{Y}} g d\phi = \int_A g d\phi + \int_{\mathcal{Y}\setminus A} g d\phi = -\int_A d\phi > 0.$$

It follows that $\phi \notin (C(\mathcal{Y})_{\leq 0})^\circ$. Hence, $(C(\mathcal{Y})_{\leq 0})^\circ \subseteq \mathrm{ca}(\mathcal{Y})_{\geq 0}$. □

Furthermore, we guarantee that the polar of an offer is non-trivial.

Lemma 11.24 (Polar of Offer is Non-Trivial)**.** *Let $\mathcal{G} \subseteq C(\mathcal{Y})$ be an offer. Then, $\mathcal{G}^\circ \neq \{0\}$.*

Proof. If $\mathcal{G}^\circ = \{0\}$, then $\mathcal{G} = \mathcal{G}^{\circ\circ} = \{0\}^\circ = C(\mathcal{Y})$ (Proposition 11.23). But this offer $\mathcal{G}$ is not legitimate, since it violates Axiom O3. □

The proof of Theorem 11.22 is given in five steps. First, we show that $\mathcal{P}_\mathcal{G}$ is a credal set by going through the axioms. Then, we argue that $\mathcal{G} = \left\{g \in C(\mathcal{Y})\colon \sup_{\phi\in\mathcal{P}_\mathcal{G}} \mathbb{E}_\phi[g] \leq 0\right\} =$

[25] We denote the Iverson-bracket as $[\![\cdot]\!]$, i.e. $[\![S]\!] = 1$, if S is true, 0 otherwise.

$(\mathbb{R}_{\geq 0}\mathcal{P}_{\mathcal{G}})^\circ$. Thirdly, $\mathcal{G}_{\mathcal{P}}$ is shown to be an offer. After that, we prove $\mathcal{P} \coloneqq \mathcal{G}_{\mathcal{P}}^\circ \cap \Delta(\mathcal{Y})$. Finally, we argue that the mapping between offers and credal sets is bijective.

Proof. **of Theorem 11.22**

1. Let $\mathcal{G} \subseteq C(\mathcal{Y})$ be an offer. We show that $\mathcal{G}^\circ \cap \Delta(\mathcal{Y}) \subseteq \mathrm{ca}(\mathcal{Y})$ is a credal set. First, $C(\mathcal{Y})_{\leq 0} \subseteq \mathcal{G}$ (Condition O4). Thus, $\mathcal{G}^\circ \subseteq \mathrm{ca}(\mathcal{Y})_{\geq 0}$ by Proposition 11.23, Statement P1 and Statement P5. Furthermore, $\mathcal{G}^\circ$ is $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed and convex (Proposition 11.23, Statement P2). In particular, $\mathcal{G}^\circ$ contains at least one element $\phi \in \mathrm{ca}(\mathcal{Y})_{\geq 0}$, such that $\|\phi\|_q = 1$. We can easily see this fact, because if $\phi \in \mathcal{G}^\circ$ is not equal to zero, then $\phi' \coloneqq \frac{\phi}{\|\phi\|_q} \in \mathcal{G}^\circ$ (Proposition 11.23, Statement P4) and $\|\phi'\|_q = 1$. Importantly, such a non-zero $\phi \in \mathcal{G}^\circ$ exists (Lemma 11.24). From all these considerations it follows that the intersection $\mathcal{G}^\circ \cap \Delta(\mathcal{Y})$ is non-empty. Furthermore, it is $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed and convex, because both $\mathcal{G}^\circ$ and $\Delta(\mathcal{Y})$ fulfill those intersection-stable properties.

2. We have,

$$
\begin{aligned}
\left\{ g \in C(\mathcal{Y}) \colon \sup_{\phi \in \mathcal{P}_{\mathcal{G}}} \mathbb{E}_\phi[g] \leq 0 \right\} &= \{ g \in C(\mathcal{Y}) \colon \langle \phi, g \rangle \leq 0, \forall \phi \in \mathcal{P}_{\mathcal{G}} \} \\
&= \{ g \in C(\mathcal{Y}) \colon \langle r\phi, g \rangle \leq 0, \forall \phi \in \mathcal{P}_{\mathcal{G}}, \forall r \in \mathbb{R}_{\geq 0} \} \\
&= \{ g \in C(\mathcal{Y}) \colon \langle t, g \rangle \leq 0, \forall t \in \mathbb{R}_{\geq 0}\mathcal{P}_{\mathcal{G}} \} \\
&\overset{P4}{=} (\mathbb{R}_{\geq 0}\mathcal{P}_{\mathcal{G}})^\circ \\
&= (\mathbb{R}_{\geq 0}(\mathcal{G}^\circ \cap \Delta(\mathcal{Y})))^\circ \\
&= (\{ r\phi \colon r \in \mathbb{R}_{\geq 0}, \phi \in \mathcal{G}^\circ, \phi \in \Delta(\mathcal{Y}) \})^\circ \\
&= (\{ r\phi \colon r \in \mathbb{R}_{\geq 0}, \phi \in \mathcal{G}^\circ \} \cap \{ r\phi \colon r \in \mathbb{R}_{\geq 0}, \phi \in \Delta(\mathcal{Y}) \})^\circ \\
&= (\mathbb{R}_{\geq 0}\mathcal{G}^\circ \cap \mathbb{R}_{\geq 0}\Delta(\mathcal{Y}))^\circ \\
&\overset{P4}{=} (\mathcal{G}^\circ \cap \mathbb{R}_{\geq 0}\Delta(\mathcal{Y}))^\circ \\
&= (\mathcal{G}^\circ \cap \mathrm{ca}(\mathcal{Y})_{\geq 0})^\circ \\
&\overset{O4}{=} (\mathcal{G}^\circ)^\circ \\
&\overset{P3}{=} \mathcal{G}.
\end{aligned}
$$

3. Let $\mathcal{P} \subseteq \mathrm{ca}(\mathcal{Y})$ be a credal set. We show that $\mathcal{G}_{\mathcal{P}} = (\mathbb{R}_{\geq 0}\mathcal{P})^\circ$ is an offer. First, $\mathcal{G}_{\mathcal{P}}$ is $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$-closed (Condition O5), a convex cone (Condition O2 and O1) that contains the zero element (Proposition 11.23, Statement P2 and P4). Furthermore, $\mathbb{R}_{\geq 0}\mathcal{P} \subseteq \mathrm{ca}(\mathcal{Y})_{\geq 0}$ implies $(\mathbb{R}_{\geq 0}\mathcal{P})^\circ \supseteq C(\mathcal{Y})_{\leq 0}$ which is Condition O4 (Proposition 11.23, Statement P1 and P5). For any $g \in (\mathbb{R}_{\geq 0}\mathcal{P})^\circ$ we have that $\inf g \leq 0$,

otherwise $c \coloneqq \inf g > 0$, hence,

$$\langle \phi, g \rangle = \int_{\mathcal{Y}} g d\phi \geq \int_{\mathcal{Y}} c d\phi = c \int_{\mathcal{Y}} d\phi > 0,$$

for some $\phi \in \mathcal{P}$ (Axiom O3).

4. We have,

$$\mathcal{P}_{\mathcal{G}} = \mathcal{G}^\circ \cap \Delta(\mathcal{Y}) = ((\mathbb{R}_{\geq 0}\mathcal{P})^\circ)^\circ \cap \Delta(\mathcal{Y}) = \mathcal{P}.$$

5. For the bijectivity of the mapping, we have shown $\tilde{\mathcal{G}} = \mathcal{G}_{\mathcal{P}_{\tilde{\mathcal{G}}}}$ for any offer $\tilde{\mathcal{G}}$ and $\tilde{\mathcal{P}} = \mathcal{P}_{\mathcal{G}_{\tilde{\mathcal{P}}}}$ for any credal set $\tilde{\mathcal{P}}$. Hence, the mapping between offers and credal sets has a left and a right inverse, hence is a bijection.

□

In conclusion, every forecasting set provides a set of available gambles which fulfills the axioms of an offer. An offer in turn is in one-to-one correspondence with respect to credal sets, forecasting sets which are closed and convex. We will make use of the obtained statements in the following two sections in which we characterize the set of available gambles if the forecasting sets are not obtained through arbitrary properties, but if the properties are elicitable (respectively identifiable).

11.4.7 Proof of Theorem 11.13: Characterizing Available Gambles of Forecasts of Elicitable Properties

Via the introduced notion of an offer (Definition 11.17) and the fundamental correspondence Theorem 11.22, we can show that the set of available gambles for a forecast aligned to an elicitable property is defined through the set of scaled regret gambles.

Proof. First, we show that $\mathcal{H}_{\ell_\gamma}$ is an offer following Definition 11.17:

- O1: Let $g, f \in \mathcal{H}_{\ell_\gamma}$, then,

$$\begin{aligned} f + g &\leq \alpha_f(\ell_\gamma - \ell_{c_f}) + \alpha_g(\ell_\gamma - \ell_{c_g}) \\ &= (\alpha_f + \alpha_g)\left(\ell_\gamma - \left(\frac{\alpha_f}{\alpha_f + \alpha_g}\ell_{c_f} + \frac{\alpha_g}{\alpha_f + \alpha_g}\ell_{c_g}\right)\right) \\ &\leq (\alpha_f + \alpha_g)\left(\ell_\gamma - \ell_{c'}\right) \end{aligned}$$

  with $(\alpha_f + \alpha_g) \in \mathbb{R}_{\geq 0}$ and $c' \in \mathcal{V}$. Such a c' exists because $\ell_{c_f}, \ell_{c_g} \in \text{spr}(\ell)$ and the superprediction set $\text{spr}(\ell)$ is convex by assumption.

- O2: Let $g \in \mathcal{H}_{\ell_\gamma}$ and $r \geq 0$. Then $rg \leq r\alpha(\ell_\gamma - \ell_c)$ with $r\alpha \in \mathbb{R}_{\geq 0}$.
- O3: Let $g \in \mathcal{H}_{\ell_\gamma}$, then there is $\alpha \in \mathbb{R}_{\geq 0}$ and $c \in \mathbb{R}$ such that $g \leq \alpha(\ell_\gamma - \ell_c)$. In particular, $\inf g \leq \inf \alpha(\ell_\gamma - \ell_c)$. Hence, we require that $\inf \alpha(\ell_\gamma - \ell_c) \leq 0$ for every $\alpha \geq 0$. This is guaranteed because there exist $\phi \in \Gamma^{-1}(\gamma) \neq \emptyset$ such that $\langle \phi, \ell_\gamma - \ell_c \rangle \leq 0$ by consistency of the scoring function. Thus, Lemma 11.54 applies.
- O4: Let $g \in C(\mathcal{Y})$ with $\sup g \leq 0$, then $g \leq \alpha(\ell_\gamma - \ell_c)$ for $\alpha = 0$. Thus, $g \in \mathcal{H}_{\ell_\gamma}$.
- O5: By definition of $\mathcal{H}_{\ell_\gamma}$ (11.6).

Then, we compute

$$
\begin{aligned}
&\mathcal{H}_{\ell_\gamma}^\circ \cap \Delta(\mathcal{Y}) \\
&\overset{(a)}{=} \{g \in C(\mathcal{Y}) \colon g \leq \alpha(\ell_\gamma - \ell_c) \text{ for some } \alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\}^\circ \cap \Delta(\mathcal{Y}) \\
&\overset{(b)}{=} \{\phi \in \operatorname{ca}(\mathcal{Y}) \colon \langle \phi, g \rangle \leq 0, \forall g \in C(\mathcal{Y}) \text{ s.t. } g \leq \alpha(\ell_\gamma - \ell_c) \text{ for some } \alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\} \cap \Delta(\mathcal{Y}) \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \langle \phi, g \rangle \leq 0, \forall g \in C(\mathcal{Y}) \text{ s.t. } g \leq \alpha(\ell_\gamma - \ell_c) \text{ for some } \alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\} \\
&\overset{(c)}{=} \{\phi \in \Delta(\mathcal{Y}) \colon \langle \phi, \alpha(\ell_\gamma - \ell_c) \rangle \leq 0, \forall \alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\} \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \alpha\left(\langle \phi, \ell_\gamma \rangle - \langle \phi, \ell_c \rangle\right) \leq 0, \forall \alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\} \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \langle \phi, \ell_\gamma \rangle \leq \langle \phi, \ell_c \rangle, \forall c \in \mathcal{V}\} \\
&= \left\{\phi \in \Delta(\mathcal{Y}) \colon \arg\min_{c \in \mathcal{V}} \langle \phi, \ell_c \rangle = \gamma\right\} \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \Gamma(\phi) = \gamma\} \\
&= \Gamma^{-1}(\gamma).
\end{aligned}
$$

(a) For convex sets containing zero, such as,

$$A \coloneqq \{g \in C(\mathcal{Y}) \colon g \leq \beta(\ell_\gamma - \ell_c) \text{ for some } \beta \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\},$$

we have $\operatorname{cl} A = A^{\circ\circ}$, which implies $A^{\circ\circ\circ} = A^\circ = (\operatorname{cl} A)^\circ$ (Proposition 11.23).

(b) We have shown that $\{g \in C(\mathcal{Y}) \colon g \leq \alpha(\ell_\gamma - \ell_c) \text{ for some } \alpha \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\}$ is a cone in the first part of the proof. Thus, Proposition 11.23 Statement P4 applies.

(c) We know that for every $g \in C(\mathcal{Y})$ such that $g \leq \alpha(\ell_\gamma - \ell_c)$ for some $\alpha \in \mathbb{R}_{\geq 0}$ and $c \in \mathcal{V}$,

$$
\begin{aligned}
\langle \phi, g \rangle &= \langle \phi, \alpha(\ell_\gamma - \ell_c) + g - \alpha(\ell_\gamma - \ell_c) \rangle \\
&= \langle \phi, \alpha(\ell_\gamma - \ell_c) \rangle + \langle \phi, g - \alpha(\ell_\gamma - \ell_c) \rangle \\
&\leq \langle \phi, \alpha(\ell_\gamma - \ell_c) \rangle,
\end{aligned}
$$

because $\phi \geq 0$ and $g - \alpha(\ell_\gamma - \ell_c) \leq 0$.

We have already proven that $\Gamma^{-1}(\gamma)$ is a credal set for all γ (Proposition 11.9). Furthermore, credal sets and offers are in one-to-one correspondence (Theorem 11.22). It follows that $\mathcal{G}_{\Gamma^{-1}(\gamma)}$ is equal to $\mathcal{H}_{\ell_\gamma}$. □

11.4.8 Proof of Theorem 11.15: Characterizing Available Gambles of Forecasts of Identifiable Properties

The characterization of available gambles for forecasts aligned to identifiable properties follows analogously.

Proof. First, we show that $\mathcal{H}_{\nu_\gamma}$ is an offer following Definition 11.17:

- O1: Let $g, f \in \mathcal{H}_{\nu_\gamma}$, then $f + g \leq \alpha_f \nu_\gamma + \alpha_g \nu_\gamma = (\alpha_f + \alpha_g)\nu_\gamma$, with $(\alpha_f + \alpha_g) \in \mathbb{R}$.
- O2: Let $g \in \mathcal{H}_{\nu_\gamma}$ and $c \geq 0$. Then $cg \leq c\alpha\nu_\gamma$ with $c\alpha \in \mathbb{R}$.
- O3: Let $g \in \mathcal{H}_{\nu_\gamma}$, then there is $\alpha \in \mathbb{R}$ such that $g \leq \alpha\nu_\gamma$. In particular, $\inf g \leq \inf \alpha\nu_\gamma$. Hence, we require that $\inf \alpha\nu_\gamma \leq 0$ for every $\alpha \in \mathbb{R}$. To see this we remind the reader that $\Gamma^{-1}(\gamma)$ is non-empty (by assumption), i.e., there exist $\phi \in \Delta(\mathcal{Y})$ such that $\langle \phi, \nu_\gamma \rangle = 0$. Thus, Lemma 11.54 applies for ν_γ and $-\nu_\gamma$, which gives the desired result.
- O4: Let $g \in C(\mathcal{Y})$ with $\sup g \leq 0$, then $g \leq \alpha\nu_\gamma$ for $\alpha = 0$. Thus, $g \in \mathcal{H}_{\nu_\gamma}$.
- O5: By definition of $\mathcal{H}_{\nu_\gamma}$ (11.8).

Then, we compute

$$
\begin{aligned}
\mathcal{H}_{\nu_\gamma}^\circ \cap \Delta(\mathcal{Y}) &\overset{(a)}{=} \{g \in C(\mathcal{Y}) \colon g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\}^\circ \cap \Delta(\mathcal{Y}) \\
&\overset{(b)}{=} \{\phi \in \mathrm{ca}(\mathcal{Y}) \colon \langle \phi, g \rangle \leq 0, \forall g \in C(\mathcal{Y}) \text{ such that } g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\} \cap \Delta(\mathcal{Y}) \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \langle \phi, g \rangle \leq 0, \forall g \in C(\mathcal{Y}) \text{ such that } g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\} \\
&\overset{(c)}{=} \{\phi \in \Delta(\mathcal{Y}) \colon \langle \phi, \alpha\nu_\gamma \rangle \leq 0, \forall \alpha \in \mathbb{R}\} \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \alpha\langle \phi, \nu_\gamma \rangle \leq 0, \forall \alpha \in \mathbb{R}\} \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \langle \phi, \nu_\gamma \rangle = 0\} \\
&= \{\phi \in \Delta(\mathcal{Y}) \colon \Gamma(\phi) = \gamma\} \\
&= \Gamma^{-1}(\gamma).
\end{aligned}
$$

(a) For convex sets containing zero, such as $A \coloneqq \{g \in C(\mathcal{Y}) \colon g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\}$, we have $\mathrm{cl}\, A = A^{\circ\circ}$, which implies $A^{\circ\circ\circ} = A^\circ = (\mathrm{cl}\, A)^\circ$ (Proposition 11.23).

(b) We have shown that $\{g \in C(\mathcal{Y})\colon g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\}$ is a cone in the first part of the proof. Thus, Proposition 11.23 Statement P4 applies.

(c) The top to bottom inclusion is trivial. Since we consider all $g \in C(\mathcal{Y})$ such that $g \leq \alpha\nu_\gamma$, we in particular consider all $\alpha\nu_\gamma$ for $\alpha \in \mathbb{R}$. For the reverse set inclusion, we know that for every $g \in C(\mathcal{Y})$ there is $\alpha \in \mathbb{R}$ such that $g \leq \alpha\nu_\gamma$. Thus,

$$\begin{aligned}\langle \phi, g\rangle &= \langle \phi, \alpha\nu_\gamma + g - \alpha\nu_\gamma\rangle \\ &= \langle \phi, \alpha\nu_\gamma\rangle + \langle \phi, g - \alpha\nu_\gamma\rangle \\ &\leq \langle \phi, \alpha\nu_\gamma\rangle,\end{aligned}$$

because $\phi \geq 0$ and $g - \nu_\gamma \leq 0$. Hence, we can simply focus on all $\alpha\nu_\gamma$ functions, instead of the previously specified g.

Recall that we have already proven that $\Gamma^{-1}(\gamma)$ is a credal set for all γ (Proposition 11.9). Furthermore, credal sets and offers are in one-to-one correspondence (Theorem 11.22). It follows that $\mathcal{G}_{\Gamma^{-1}(\gamma)}$ is equal to $\mathcal{H}_{\nu_\gamma}$. □

The proof of Corollary 11.16 then simply requires to show that the set of gambles upper bounded by calibration gambles is closed for finite $\mathcal{Y}$.

Proof. **of Corollary 11.16** We have to show that,

$$\mathcal{F}_{\nu_\gamma} \coloneqq \{g \in C(\mathcal{Y})\colon g \leq \alpha\nu_\gamma \text{ for some } \alpha \in \mathbb{R}\},$$

is $\sigma(\text{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed in a finite dimensional $\text{ca}(\mathcal{Y})$-space, i.e., $\mathcal{F}_{\nu_\gamma} = \mathcal{H}_{\nu_\gamma}$

Given that $\mathcal{Y}$ is finite, $C(\mathcal{Y})$ (respectively $\text{ca}(\mathcal{Y})$) reduces to $\mathbb{R}^d$ for $d = |\mathcal{Y}|$. The topology induced by the supremum norm is equivalent to the Euclidean topology [Schechter, 1997, p. 580] and the equivalent to the $\sigma(C(\mathcal{Y}), \text{ca}(\mathcal{Y}))$-topology [Schechter, 1997, 28.17 (e)]. In short, we can use the standard topology on finite dimensional Euclidean spaces to express the next results.

First, we observe that we can write $\mathcal{F}_{\nu_\gamma}$ in terms of a Minkowski-sum

$$\begin{aligned}\mathcal{F}_{\nu_\gamma} &= \{g \in C(\mathcal{Y})\colon g - \alpha\nu_\gamma \leq 0 \text{ for some } \alpha \in \mathbb{R}\} \\ &= \{\alpha\nu_\gamma\colon \alpha \in \mathbb{R}\} + \{g \in C(\mathcal{Y})\colon g \leq 0\}.\end{aligned}$$

Let us introduce the shorthands $V_\gamma \coloneqq \{\alpha\nu_\gamma\colon \alpha \in \mathbb{R}\}$ for the line and $C(\mathcal{Y})_{\leq 0} \coloneqq \{g \in C(\mathcal{Y})\colon g \leq 0\}$ for the negative orthant.

We distinguish between two simple cases:

Case 1 Assume that $\nu_\gamma \geq 0$ or $-\nu_\gamma \geq 0$. Axiom O3 guarantees that there exist at least one $y \in \mathcal{Y}$ such that $\pm\nu_\gamma(y) = 0$. We denote the subset of those y's in which ν_γ is zero $\mathcal{Y}_0 \subseteq \mathcal{Y}$. It is relatively easy to see that

$$\mathcal{F}_{\nu_\gamma} = \{g \in C(\mathcal{Y}) \colon g(y) \leq 0, \forall y \in \mathcal{Y}_0\}.$$

The left to right set inclusion is clear by definition of $\mathcal{F}_{\nu_\gamma}$. For the right to left set inclusion choose $\alpha_g \coloneqq \max_{y \in \mathcal{Y} \setminus \mathcal{Y}_0} \left(\frac{g(y)}{\nu_\gamma(y)}\right)$ for arbitrary $g \in C(\mathcal{Y})$ such that $g(y) \leq 0$ for all $y \in \mathcal{Y}_0$. Then, $g \leq \alpha_g \nu_\gamma$. Concluding, the obtained set is closed in the Euclidean topology.

Case 2 In the other case, we have to leverage a result on the closedness of Minkowski-sums going back to Debreu [1959]. A Minkowski-sum is closed if V_γ and $C(\mathcal{Y})_{\leq 0}$ are closed and the asymptotic cones of V_γ and $C(\mathcal{Y})_{\leq 0}$ are positively semi-independent [Border, 2019-2020, Theorem 20.2.3] or [Debreu, 1959, p. 23 (9)].

Asymptotic cones, roughly stated, form the set of directions in which a set is unbounded. Positively semi-independent requires that for $v \in V_\gamma$ and $o \in C(\mathcal{Y})_{\leq 0}$, $v + o = 0$ implies $v = o = 0$. In our case, V_γ and $C(\mathcal{Y})_{\leq 0}$ are closed cones, i.e., for all $\alpha \in \mathbb{R}$ and elements $v \in V_\gamma$ and $o \in C(\mathcal{Y})_{\leq 0}$ it holds $\alpha v \in V_\gamma$ respectively $\alpha o \in C(\mathcal{Y})_{\leq 0}$. Hence, the asymptotic cones of V_γ and $C(\mathcal{Y})_{\leq 0}$ are the sets themselves [Border, 2019-2020, Theorem 20.2.2 f) i)].

It remains to show that the two sets are positively semi-independent, which follows from the following observation. If $a + o = 0$ for $a \in C(\mathcal{Y})$ and $o \in C(\mathcal{Y})_{\leq 0}$, then $a \in C(\mathcal{Y})_{\geq 0}$ where $C(\mathcal{Y})_{\geq 0} \coloneqq \{g \in C(\mathcal{Y}) \colon g \geq 0\}$. Now, $V_\gamma \cap C(\mathcal{Y})_{\geq 0} = \{0\}$ because there exist $y, y' \in \mathcal{Y}$ such that $\nu_\gamma(y) < 0$ and $\nu_\gamma(y') > 0$. Hence, $a = 0$, which implies $o = 0$.

□

11.4.9 On the Relationship of Unscaled Regret and Calibration Gambles

In Section 11.4.2 we have argued that (unscaled) regret gambles and calibration gambles are both available. The characterization theorems (Theorem 11.13 and Theorem 11.15) complete the picture. *Scaled* regret gambles and calibration gambles are, up to topological approximation, dominating all available gambles.

To remind the reader, if a gamble g dominates another gamble h, then playing gamble g is favorable to the gambler. No matter which outcome $y \in \mathcal{Y}$ realizes, the value $g(y) \geq h(y)$. In other words, scaled regret gambles (respectively calibration gambles) are among the best gambles a gambler can play to disprove Forecaster to provide adequate forecasts. In

principle, a gambler has no need to explore any other gambles than the scaled regret gambles or the calibration gambles. However, practically relevant evaluation metrics due to further restrictions put on the gambler embrace a broader scope of available gambles (Section 11.5).

Nevertheless, the characterization statements offer two further interesting consequences. First, in the case that a property is elicited by several distinct loss functions[26], the set of available gambles does not change. Hence, the set of of scaled regret gambles still (approximately) dominate all those available gambles. But, the sets of *un*scaled regret gambles for difference loss functions are not necessarily equal.

Second, in the case that a property is identifiable and elicitable[27], then *scaled* regret gambles are equally describing the set of all available gambles as calibration gambles do.

Corollary 11.25 (Duality of Calibration and Regret)**.** *Let* $\Gamma\colon Q \to 2^{\mathcal{V}}$ *be an identifiable and elicitable property with identification function* $\nu\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ *and scoring function* $\ell\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ *which has a convex superprediction set. Let* $\gamma \in \mathbb{R}$ *such that* $\Gamma^{-1}(\gamma) \neq \emptyset$*. Then,*

$$\mathcal{H}_{\ell_\gamma} = \mathcal{H}_{\nu_\gamma}.$$

Proof. The statement is a simple consequence of Theorem 11.13 and Theorem 11.15. □

One may interpret this statement as that, in principle, scaled regret is as powerful as calibration in proving Forecaster's forecast wrong. With this statement we provide another facet of a decades-old discussion on the relation of loss (or regret) and calibration (Section 11.6.2).

Scaled regret gambles, to the best of our knowledge, have not been used to construct an evaluation metric,[28], but, calibration gambles have been (Section 11.5.3). The equivalence in Corollary 11.25 states that optimizing for an evaluation metric expressed through calibration gambles is equivalent to optimizing for an evaluation metric expressed through according scaled regret gambles. Hence, scaled regret gambles have implicitly been used already. However, unscaled regret gambles seem to be of more relevance to current machine learning as we observe in the following section. The relationship of regret-type evaluation metrics and calibration-type evaluation metrics is thus more subtle than Corollary 11.25 suggests.

[26] For instance, the mean is elicited by all Bregman-divergences [Gneiting, 2011].

[27] Continuous, real-valued, elicitable properties are, under mild technical conditions, identifiable and vice-versa [Lambert et al., 2008, Steinwart et al., 2014].

[28] However, activation functions, such as introduced in [Cesa-Bianchi and Lugosi, 2006, § 4.6], allow for restricted scaling by 0 or 1.

11.5 Fair, Restricted and Rational Gamblers for Recovery of Evaluation Metrics

We already observed that a small capital of a fair gambler does not necessarily guarantee "good" forecasts. (a) A gambler playing constant negative gambles (which are available by default) does not prove anything about the forecasts. (b) A gambler who has access to arbitrarily scaleable gambles can always upscale a single gamble such that the resulting capital is essentially dominated by the result of this single gamble (Example 11.55).

Towards more reasonable evaluation metrics, we introduce two further conditions on the gambler, which turn out to be sufficient to elicit several known evaluation metrics. We demand the gambler to have a *restricted* set from which its gambles can be chosen. One might interpret this as *budget constraint*. This way gambles cannot be arbitrarily scaled. Furthermore, we equip the gambler with a belief about the actual distribution from which the next outcome is realized. The gambler behaves *rationally*, i.e., as an expected utility maximizer, with respect to its belief. One might interpret the belief as *alternative hypothesis* about the distribution of the outcome. This way the gambler will not play unfavourable gambles such as constant negative ones. In conclusion, we obtain a two-dimensional structure of evaluation metrics based on the type of forecast, choice of belief and the restriction summarized in Table 11.2.

For instance, the capital of a gambler which is restricted to play regret gambles and has a (true) belief about the average outcome is equal to external regret (Proposition 11.31. Or, the capital of gamblers with access to all gambles with a bounded norm and a (true) belief about the average outcome on the subsets on which a certain value has been predicted is aggregated to calibration scores (Proposition 11.38). Roughly speaking, the belief expresses a certain knowledge which gets more and more refined in Table 11.2 from top to bottom (Proposition 11.45), while the restriction gets less stringent in Table 11.2 from left to right (Proposition 11.42). Besides that the recovery results justify the usefulness of the evaluation protocol in understanding current evaluation metrics, the above statements shed light on the debated duality of calibration scores and loss scores Section 11.6.2.

Let us formally introduce restriction and rationality for the gamblers in an evaluation protocol.

Definition 11.26 (Fair, Restricted and Rational Gambler)**.** *Let* $(\mathcal{Y}, \Gamma, T, v, G^I, y)$ *be an evaluation protocol. Fix an index* $i \in I$*. Let* $R \subseteq C(\mathcal{Y})$ *such that* $0 \in R$*. Let* $b_t \in \Delta(\mathcal{Y}) \cup \{\square\}$[29] *for every* $t \in T$*. The gambler* G^i *in the evaluation protocol is called* fair, restricted and rational *if for all* $t \in T$*, all of the following conditions are fulfilled:*

[29] We use $\square$ as a placeholder for a fully non-informative belief about the next outcome.

(a) $G_t^i \in \mathcal{G}_{\Gamma^{-1}(v_t)}$. *(Availability, Fairness)*

(b) $G_t^i \in R$. *(Restrictiveness)*

(c) If $b_t \in \Delta(\mathcal{Y})$, *then*

$$\mathbb{E}_{b_t}[G_t^i] \geq \mathbb{E}_{b_t}[g],$$

for all $g \in \mathcal{G}_{\Gamma^{-1}(v_t)} \cap R$. *(Rationality)*

(d) If $b_t = \square$, *then* $G_t^i = 0$. *(Opt-Out)*

We shortly discuss three key aspects of the newly introduced conditions on Gambler: (a) the option to withdraw from gambling, i.e., there is no obligation for Gambler to gamble. (b) the relationship of rationality and optimality with respect to an alternative hypothesis and (c) the content of the belief.

(A) No Obligation to Gamble. The belief b_t for some $t \in T$ is not necessarily a distribution, but as well can be equal to $\square$. If $b_t = \square$, then Gambler is not willing to commit to any precise belief. We use $\square$ as a placeholder for a fully non-informative belief about the next outcome. Assuming maxmin-rationality following [Gilboa and Schmeidler, 1989], we can recover the opt-out rule. The following proposition simply shows that the gamble $g^* = 0$ is an optimal action facing the full ambiguity of the own belief.

Proposition 11.27 (Opt-Out is Maxmin-Rationality wrt. Vacuous Belief)**.** *Let* $R \subseteq C(\mathcal{Y})$ *such that* $0 \in R$ *and* $\square = \Delta(\mathcal{Y})$. *If* $g^* = 0$, *then*

$$\inf_{\phi \in \square} \mathbb{E}_\phi[g^*] \geq \inf_{\phi \in \square} \mathbb{E}_\phi[g],$$

for all $g \in \mathcal{G}_\mathcal{P} \cap R$.

Proof. By definition of g^*, $\inf_{\phi \in \square} \mathbb{E}_\phi[g^*] = 0$. Let us assume that,

$$\inf_{\phi \in \square} \mathbb{E}_\phi[g] > 0.$$

It follows that $\inf_{y \in \mathcal{Y}} g(y) > 0$, because for every $y \in \mathcal{Y}$, there exists $\delta(y) \in \square$. This is a contradiction, as $g \in \mathcal{G}_\mathcal{P}$ and $\mathcal{G}_\mathcal{P}$ is an offer (Theorem 11.18) with condition O3. $\square$

(B) Rationality is Type-II Error Minimization. The belief of Gambler is nothing else than an alternative hypothesis suggested against the "null hypothesis" made by Forecaster. The introduction of an alternative hypothesis for testing probabilistic statements is

not new and at least goes back to the seminal work by Neyman and Pearson [1933]. Essentially, alternative hypotheses try to control the Type-II error of a probabilistic statement. We can provide analogous interpretations to the conditions on Gambler: Availability roughly guarantees control over Type-I error, the false rejection of true forecasts. Rationality minimizes the Type-II error, the false acceptance of a wrong forecast. This correspondence is not surprising for readers familiar to the literature on e-values. The availability criterion is captured in the definition of e-value itself, while rationality is expressed through growth rate optimality [Grünwald et al., 2020]. There are some central differences, though. For a detailed discussion see Section 11.6.1.

(c) Belief can be True. As it turns out in Section 11.5.1, many existing evaluation metrics summarize the capital of the gamblers which play "as if" they would know parts of the outcomes of Nature. Usually, alternative hypothesis express external knowledge about a certain phenomenon. We will leverage our a posteriori perspective to align belief to truth. If the beliefs of an agent are aligned to truth, the beliefs can be called *knowledge*.

Definition 11.28 (Alignment to Truth)**.** *Let $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ be an evaluation protocol and denote by $b_t \in \Delta(\mathcal{Y}) \cup \{\square\}$ for every $t \in T$ the belief of G. We define $T_\beta \coloneqq \{t \in T \colon b_t = \beta\}$. The beliefs $(b_t)_{t\in T}$ are* aligned to truth*, if for all $\beta \in \{b_t \colon t \in T\}$, $\beta = \square$ or $\beta = \frac{1}{|T_\beta|} \sum_{t\in T_\beta} \delta(y_t) \in \Delta(\mathcal{Y})$.*

We illustrate the three conditions on the gambler in Figure 11.4.

Finally, we provide a simple derivative of a fair, restricted and rational gambler, which is only approximately rational. In fact, this definition let's us ignore the $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$-closure of sets in later results (e.g., Proposition 11.39).

Definition 11.29 (Fair, Restricted and Approximately Rational Gambler)**.** *Let $(\mathcal{Y}, \Gamma, T, v, G^I, y)$ be an evaluation protocol. Fix an index $i \in I$. Let $R \subseteq C(\mathcal{Y})$ such that $0 \in R$. Let $b_t \in \Delta(\mathcal{Y}) \cup \{\square\}$ for every $t \in T$. The ith gambler G^i in the evaluation protocol is called* fair, restricted and approximately rational *if for all $t \in T$, all of the following conditions are fulfilled:*

(a) For all $t \in T$, $G^i_t \in \mathcal{G}_{\Gamma^{-1}(v_t)}$. (Availability, Fairness)

(b) For all $t \in T$, $G^i_t \in R$. (Restrictiveness)

(c) For all $t \in T$ with $b_t \in \Delta(\mathcal{Y})$, then

$$\mathbb{E}_{b_t}[G^i_t] \geq \mathbb{E}_{b_t}[g] - \epsilon,$$

for all $\epsilon > 0$ and $g \in \mathcal{G}_{\Gamma^{-1}(v_t)} \cap R$. (Approximate Rationality)

(d) For all $t \in T$ with $b_t = \square$, then $G_t^i = 0$. (Opt-Out)

11.5.1 Recovery of Existing Evaluation Metrics

Machine learning scholars have introduced many notions which try to capture the quality of forecasts [Cesa-Bianchi and Lugosi, 2006, Williamson and Cranko, 2023, Derr et al., 2026]. We suggest that our approach via evaluation protocols (Definition 11.2) and fair, restricted and rational gamblers (Definition 11.26) provides an insightful framework to recover and structure several of those notions.

In the following, we provide a series of results, which follow a simple scheme summarized in Proposition 11.30.

Proposition 11.30 (Schematic Recovery Proposition)**.** *Set up the evaluation protocol, in particular, the outcome space $\mathcal{Y}$ and the predicted property Γ. Define several gamblers by:*

(a) Defining Gambler's beliefs.

(b) Restricting the allowed gambles.

The aggregated capital among the gamblers is equal to some known metric of evaluation.

11.5.2 From Gambles to Regret

Let us start with a simple finger exercise. We show that within the evaluation protocol with a certain choice of beliefs and restrictions, we can recover external regret. The proof is relatively direct and largely builds upon a specially structured restriction, which bounds the gambler to (unscaled) regret gambles.

Proposition 11.31 (Recovery of External Regret)**.** *Let $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ be an evaluation protocol with $\Gamma \colon Q \to 2^{\mathcal{V}}$ being an elicitable property with consistent scoring function ℓ. We define a fair, restricted and rational gambler G such that:*

(a) On every instance $t \in T$ the gambler G has a (true) belief b_t about the average outcome distribution,

$$b_t = \hat{D}_Y \coloneqq \frac{1}{|T|} \sum_{t' \in T} \delta(y_{t'}) \in Q.$$

(b) On every instance $t \in T$ the gambler is restricted to $\mathcal{L}_{v_t} \coloneqq \{\ell_{v_t} - \ell_c \colon c \in \mathcal{V}\}$.

Then, the reweighted capital of the gambler G is equal to,

$$|T|K(G) = \sum_{t \in T} \ell(y_t, v_t) - \min_{c \in \mathcal{V}} \ell(y_t, c),$$

which is external regret.

Proof. Let us consider a fixed instance $t \in T$. Given the gambler's belief, the fair, restricted and rational gambler plays (Lemma 11.32),

$$G_t \colon y \mapsto \ell(y, v_t) - \ell(y, \Gamma(\hat{D}_Y)).$$

In total, we obtain the following aggregated capital for Gambler G,

$$K(G) = \frac{1}{|T|} \sum_{t \in T} G_t(y_t) = \frac{1}{|T|} \sum_{t \in T} \ell(y_t, v_t) - \ell(y_t, \Gamma(\hat{D}_Y)).$$

The statement follows. □

Lemma 11.32 (Gamble for Rational Gambler Restricted to Regret Gambles)**.** *Let $(\mathcal{Y}, \Gamma, T, v, G^I, y)$ be an evaluation protocol with $\Gamma \colon Q \to 2^{\mathcal{V}}$ being an elicitable property with consistent scoring function ℓ. Fix $t \in T$ and consider a fair, restricted and rational gambler G^i with belief $b_t \in Q$ and restriction $\mathcal{L}_{v_t} \coloneqq \{\ell_{v_t} - \ell_c \colon c \in \mathcal{V}\}$. Then, in this round $t \in T$, the fair, restricted and rational gambler G^i plays the gamble,*

$$G_t^i \colon y \mapsto \ell(y, v_t) - \ell(y, \Gamma(b_t)).$$

Proof. The statement follows because,

$$\begin{aligned}
\mathbb{E}_{b_t}[G_t^i] &= \mathbb{E}_{b_t}[\ell_{v_t} - \ell_{\Gamma(b_t)}] \\
&\overset{D11.7}{\geq} \max_{c \in \mathcal{V}} \mathbb{E}_{b_t}[\ell_{v_t} - \ell_c] \\
&= \max_{g \in \mathcal{L}_{v_t}} \mathbb{E}_{b_t}[g] \\
&\overset{P11.10}{=} \max_{g \in \mathcal{G}_{\Gamma^{-1}(v_t)} \cap \mathcal{L}_{v_t}} \mathbb{E}_{b_t}[g].
\end{aligned}$$

□

Arguably, the knowledge of Gambler is "too" coarse to provide a hard benchmark. Hence, we suggest to make the beliefs more "fine-grained" (cf. Definition 11.44). In Proposition 11.33 we equip a set of gamblers with a true belief about the average outcome on all instances on

which the forecaster forecasted a certain fixed valued, which lets us recover swap regret (cf. Example 4.4 in [Cesa-Bianchi and Lugosi, 2006]).

Proposition 11.33 (Recovery of Swap Regret)**.** *Let* $v_T \coloneqq \{v_t \colon t \in T\}$ *and* $\Gamma \colon Q \to 2^{\mathcal{V}}$ *be an elicitable property with consistent scoring function* ℓ*. Let* $(\mathcal{Y}, \Gamma, T, v, G^{v_T}, y)$ *be an evaluation protocol. For every* $\gamma \in v_T$*, we define a fair, restricted and rational gambler* G^γ *such that:*

(a) On all instances $T_\gamma \coloneqq \{t \in T \colon v_t = \gamma\}$*, the gambler* G^γ *has a (true) belief* b_t^γ *about the average outcome distribution,*

$$b_t^\gamma = \hat{D}_{Y|\gamma} \coloneqq \frac{1}{|T_\gamma|} \sum_{t' \in T_\gamma} \delta(y_{t'}) \in Q.$$

On all other instances $T \setminus T_\gamma$*, the gambler has belief* $b_t^\gamma = \square$*.*

(b) The gambler is restricted to play gambles in $\mathcal{L}_\gamma \coloneqq \{\ell_\gamma - \ell_c \colon c \in \mathcal{V}\}$*.*

Then, the re-weighted (by $|T|$*) and summed up capitals of all gamblers* G^γ *is equal to,*

$$|T| \sum_{\gamma \in v_T} K(G^\gamma) = \max_{\sigma \in \{\mathcal{V} \to \mathcal{V}\}} \sum_{t \in T} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)),$$

which is the definition of swap regret following [Cesa-Bianchi and Lugosi, 2006, Example 4.4].

Proof. Let us fix a gambler G^γ for some $\gamma \in v_T$. Consider a fixed instance $t \in T_\gamma$. Given the gambler's belief, the fair, restricted and rational gambler plays (Lemma 11.32)„

$$G_t^\gamma \colon y \mapsto \ell(y, \gamma) - \ell(y, \Gamma(\hat{D}_{Y|\gamma})).$$

For any fixed instance $t \in T \setminus T_\gamma$, the gambler's belief $b_t^\gamma = \square$ is vacuous, hence by Definition 11.26, $G_t^\gamma = 0$. In total, we obtain the following aggregated capital for a gambler G^γ,

$$K(G^\gamma) = \frac{1}{|T|} \sum_{t \in T_\gamma} G_t^\gamma(y_t) = \frac{1}{|T|} \sum_{t \in T_\gamma} \ell(y_t, v_t) - \ell(y_t, \Gamma(\hat{D}_{Y|\gamma}).$$

The statement follows by aggregating the capitals of the different gamblers,

$$\begin{aligned}
|T| \sum_{\gamma \in v_T} K(G^\gamma) &= \sum_{\gamma \in v_T} \sum_{t \in T_\gamma} \ell(y_t, v_t) - \ell(y_t, \Gamma(\hat{D}_{Y|\gamma}) \\
&= \sum_{\gamma \in v_T} \max_{c \in \mathcal{V}} \sum_{t \in T_\gamma} \ell(y_t, v_t) - \ell(y_t, c) \\
&= \max_{\sigma \in \{\mathcal{V} \to \mathcal{V}\}} \sum_{\gamma \in v_T} \sum_{t \in T_\gamma} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)) \\
&= \max_{\sigma \in \{\mathcal{V} \to \mathcal{V}\}} \sum_{t \in T} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)).
\end{aligned}$$

□

By refining the belief to group-wise knowledge about the average outcome given a prediction we obtain a notion which we term *group-wise swap regret score*. This notion is analogous to multicalibration (Section 11.5.4).

Proposition 11.34 (Recovery of Group-Wise Swap Regret Score)**.** *Let $\mathcal{S} \subseteq 2^T$ be a set of groups and $v_T \coloneqq \{v_t \colon t \in T\}$. Let $(\mathcal{Y}, \Gamma, T, v, G^{v_T \times \mathcal{S}}, y)$ be an evaluation protocol with $\Gamma \colon Q \to 2^{\mathcal{V}}$ being an elicitable property with consistent scoring function ℓ. For every $\gamma \in v_T$ and $S \in \mathcal{S}$ we define a fair, restricted and approximate rational gambler $G^{\gamma,S}$ such that:*

(a) On all instances $T_{\gamma,S} \coloneqq \{t \in T \colon v_t = \gamma\} \cap S$, the gambler $G^{\gamma,S}$ has a (true) belief $b_t^{\gamma,S}$ about the average outcome distribution,

$$b_t^{\gamma,S} = \hat{D}_{Y|\gamma,S} \coloneqq \frac{1}{|T_{\gamma,S}|} \sum_{t' \in T_{\gamma,S}} \delta(y_{t'}) \in Q.$$

On all other instances $T \setminus T_{\gamma,S}$, the gambler has no belief.

(b) The gambler is restricted to play gambles in $\mathcal{L}_\gamma \coloneqq \{\ell_\gamma - \ell_c \colon c \in \mathcal{V}\}$.

Then, the reweighted and aggregated capital of the gamblers $G^{\gamma,S}$ is equal to,

$$\sup_{S \in \mathcal{S}} \sum_{\gamma \in v_T} |K(G^{\gamma,S})| = \frac{1}{|T|} \sup_{S \in \mathcal{S}} \max_{\sigma \in \{\mathcal{V} \to \mathcal{V}\}} \sum_{t \in S} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)),$$

what we call group-wise swap regret score.

Proof. Fix $\gamma \in v_T$ and $S \in \mathcal{S}$. Let us consider a fixed instance $t \in T_{\gamma,S}$. Given the gambler's belief, the fair, restricted and rational gambler plays (Lemma 11.32),

$$G_t^{\gamma,S} \colon y \mapsto \ell(y, \gamma) - \ell(y, \Gamma(\hat{D}_{Y|\gamma,S})).$$

For any fixed instance $t \in T \setminus T_{\gamma,S}$, the gambler's belief $b_t^{\gamma,S} = \square$ is vacuous, hence by Definition 11.26, $G_t^{\gamma,S} = 0$. In total, we obtain the following aggregated capital for the gambler $G^{\gamma,S}$,

$$K(G^{\gamma,S}) = \frac{1}{|T|} \sum_{t \in T_{\gamma,S}} G_t^{\gamma}(y_t) = \frac{1}{|T|} \sum_{t \in T_{\gamma,S}} \ell(y_t, \gamma) - \ell(y_t, \Gamma(\hat{D}_{Y|\gamma,S})).$$

The statement follows,

$$\begin{aligned}
\sup_{S \in \mathcal{S}} \sum_{\gamma \in v_T} \left| K(G^{\gamma,S}) \right| &= \sup_{S \in \mathcal{S}} \sum_{\gamma \in v_T} \left| \frac{1}{|T|} \sum_{t \in T_{\gamma,S}} \ell(y_t, \gamma) - \ell(y_t, \Gamma(\hat{D}_{Y|\gamma,S})) \right| \\
&= \sup_{S \in \mathcal{S}} \sum_{\gamma \in v_T} \frac{1}{|T|} \max_{c \in \mathcal{V}} \sum_{t \in T_{\gamma,S}} \ell(y_t, v_t) - \ell(y_t, c) \\
&= \frac{1}{|T|} \sup_{S \in \mathcal{S}} \max_{\sigma \in \{\mathcal{V} \to \mathcal{V}\}} \sum_{\gamma \in v_T} \sum_{t \in T_{\gamma,S}} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)) \\
&= \frac{1}{|T|} \sup_{S \in \mathcal{S}} \max_{\sigma \in \{\mathcal{V} \to \mathcal{V}\}} \sum_{t \in S} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)).
\end{aligned}$$

□

Finally, we push the knowledge of a gambler to the extreme and assume it is "clairvoyant". This way we recover standard, but zeroed, loss scores. Arguably, that is the most widely used evaluation metric in machine learning and often summarized in benchmark tables, e.g., accuracy, false positive rate (FPR), false negative rate (FNR), mean squared error (MSE), Brier score, cross entropy loss. The zeroed form of the loss scores might seem strange on the first sight. It is the average loss of the forecaster subtracted by the loss a forecaster which has access to the next outcome would have. For most loss functions, if the forecaster knows the outcome, then the loss is equal to 0. Formally, $\ell(y, \Gamma(\delta(y)) = 0$ for all $y \in \mathcal{Y}$. This is true for instance for the Brier score, accuracy or cross entropy loss. Hence, we call the scores "zeroed" as the perfect forecaster would achieve loss score 0.

Proposition 11.35 (Recovery of Zeroed Loss Scores)**.** *Let $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ be an evaluation protocol with $\Gamma \colon Q \to 2^{\mathcal{V}}$ being an elicitable property with consistent scoring function ℓ and fair, restricted and rational gambler G such that:*

(a) On every instance $t \in T$ the gambler G is "clairvoyant", i.e., has a (true) belief b_t about the outcome revealed by Nature, i.e., $b_t = \delta(y_t) \in Q$.

(b) The gambler is restricted to play gambles in $\mathcal{L}_{v_t} \coloneqq \{\ell_{v_t} - \ell_c \colon c \in \mathcal{V}\}$.

Then, the capital of the gambler G is equal to,

$$K(G) = \frac{1}{|T|}\sum_{t\in T}\ell(y_t, v_t) - \frac{1}{|T|}\sum_{t\in T}\ell(y_t, \Gamma(\delta(y_t))).$$

Proof. Let us consider a fixed instance $t \in T$. Given the gambler's belief, the fair, restricted and rational gambler plays (Lemma 11.32),

$$G_t \colon y \mapsto \ell(y, v_t) - \ell(y, \Gamma(\delta(y_t))).$$

In total, we obtain the following aggregated capital for Gambler G,

$$K(G) = \frac{1}{|T|}\sum_{t\in T} G_t(y_t) = \frac{1}{|T|}\sum_{t\in T}\ell(y_t, v_t) - \ell(y_t, \Gamma(\delta(y_t))).$$

□

In all of the above propositions, the reader might remain disappointed in that the restriction to $\mathcal{L}_{v_t}$ (respectively $\mathcal{L}_\gamma$) artificially forces the gambler to recover certain types of regret. Clearly, regret gambles are available. However, they form only one particular choice for evaluating a forecast within the framework of a fair evaluation protocol. Hence, we are not exploiting the full freedom of available gambles.

In the next section, the restriction of the gambles will appear more natural. In particular, the results will require a full characterization of available gambles.[30]

11.5.3 From Gambles to Calibration

The following recovery results share with the above the order in which the belief of the gambler(s) is refined step-by-step. However, we consider another restricted set of gambles from which Gambler can choose which is defined through the scaled norm ball $\mathbb{B}_\alpha \coloneqq \{g \in C(\mathcal{Y}) \colon \|g\| \leq \alpha\}$ for some norm $\|\cdot\|$ on $C(\mathcal{Y})$ and we assume that the predicted property Γ is identifiable. In particular, the first two statements consider mean predictors in the binary setting. The reason for this is that most of the existing notions focus on this limited setup.

First, we recover bias in the large (for binary outcomes) following [Murphy and Epstein, 1967] (for more modern formulations see [Vovk et al., 2005] or the generalization to subgroups called "accuracy in expectation" [Hébert-Johnson et al., 2018][31]).

[30] The use of the full characterization is marked by reference to Theorem 11.15 (respectively Corollary 11.16).

[31] We can recover the exact formulation of "accuracy in expectation" for a subgroup by simply assuming that the gambler in Proposition 11.36 has a true belief about the average outcome on the subgroup, and a vacuous belief, i.e., □, outside the subgroup (cf. Proposition 11.39)

Proposition 11.36 (Recovery of Bias in the Large)**.** *Let $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ be an evaluation protocol with $\mathcal{Y} = \{0, 1\}$ and Γ being the mean. We define a fair, restricted and rational gambler G such that:*

(a) On every instance $t \in T$, the gambler G has a (true) belief b_t about the average outcome distribution,

$$b_t = \hat{D}_Y \coloneqq \frac{1}{|T|} \sum_{t' \in T} \delta(y_{t'}) \in \Delta(\mathcal{Y}).$$

(b) The gambler is restricted to play gambles in $\mathbb{B}_1 \coloneqq \{g \in C(\mathcal{Y}) \colon \|g\|_1 \leq 1\}$ where $\|\cdot\|_1$ is the l_1-norm.[32]

Then, the capital of the gambler G is equal to,

$$K(G) = \left| \frac{1}{|T|} \sum_{t \in T} (y_t - v_t) \right|,$$

which is the definition of bias in the large for binary outcomes following [Murphy and Epstein, 1967].

Proof. Let us consider a fixed instance $t \in T_S$. Given the gambler's belief, the fair, restricted and rational gambler plays (Lemma 11.37),

$$G_t \colon \begin{cases} y \mapsto y - v_t \text{ if } \mathbb{E}_{\hat{D}_Y}[Y] > v_t \\ y \mapsto v_t - y \text{ otherwise.} \end{cases}$$

In total, we obtain the following aggregated capital for the gambler G,

$$\begin{aligned} K(G) &= \frac{1}{|T|} \sum_{t \in T} \begin{cases} y_t - v_t \text{ if } \mathbb{E}_{\hat{D}_Y}[Y] > \gamma \\ v_t - y_t \text{ otherwise.} \end{cases} \\ &= \frac{1}{|T|} \left| \sum_{t \in T} (y_t - v_t) \right|. \end{aligned}$$

□

Lemma 11.37 (Gamble for Rational Gambler Restricted to Norm Ball - Binary $\mathcal{Y}$)**.** *Let $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ be an evaluation protocol with $\mathcal{Y} = \{0, 1\}$ and Γ being the mean. Fix $t \in T$ and consider a fair, restricted and rational gambler G with belief $b_t \in \Delta(\mathcal{Y})$ and*

[32] The l_1-norm is well-defined as $C(\mathcal{Y}) = \mathbb{R}^2$ here.

restriction $\mathbb{B}_1 \coloneqq \{g \in C(\mathcal{Y})\colon \|g\|_1 \leq 1\}$. *Then, in this round* $t \in T$, *the fair, restricted and rational gambler* G *plays the gamble,*

$$G_t\colon \begin{cases} y \mapsto y - v_t \text{ if } \mathbb{E}_{b_t}[Y] > v_t \\ y \mapsto v_t - y \text{ otherwise.} \end{cases}.$$

Proof. The statement follows because,

$$\begin{aligned}
\mathbb{E}_{b_t}[G_t] &= |\mathbb{E}_{b_t}[Y] - v_t| \\
&= \max_{\alpha \in [-1,1]} \alpha \left(\mathbb{E}_{b_t}[Y] - v_t\right) \\
&= \max_{\alpha \in [-1,1]} \mathbb{E}_{b_t}[\alpha \nu_{v_t}] \\
&\overset{(*)}{=} \max_{\alpha \in \mathbb{R} \text{ and } \|\alpha\nu_{v_t}\|_1 \leq 1} \mathbb{E}_{b_t}[\alpha \nu_{v_t}] \\
&\geq \max_{g \in \{g \in C(\mathcal{Y})\colon g \leq \alpha\nu_{v_t}, \alpha \in \mathbb{R}\} \cap \mathbb{B}_1} \mathbb{E}_{b_t}[g] \\
&\overset{(C11.16)}{=} \max_{g \in \mathcal{G}_{\Gamma^{-1}(\gamma)} \cap \mathbb{B}_1} \mathbb{E}_{b_t}[g].
\end{aligned}$$

Equation $(*)$ is true because $\|\alpha\nu_\gamma\|_1 = |\alpha|\|\nu_\gamma\|_1 = |\alpha|$ for all $\gamma \in \mathcal{V}$. □

It is not a far way to the recovery of bias in the small [Murphy and Epstein, 1967], better known as standard calibration. The beliefs' of the gambler in the following proposition are equal to the beliefs' of the gamblers in Proposition 11.33.

Proposition 11.38 (Recovery of Calibration Score)**.** *Let* $v_T \coloneqq \{v_t\colon t \in T\}$, $\mathcal{Y} = \{0,1\}$ *and* Γ *being the mean. Let* $(\mathcal{Y}, \Gamma, T, v, G^{v_T}, y)$ *be an evaluation protocol. For every* $\gamma \in v_T$, *we define a fair, restricted and rational gambler* G^γ *such that:*

(a) On all instances $T_\gamma \coloneqq \{t \in T\colon v_t = \gamma\}$, *the gambler* G^γ *has a (true) belief* b_t^γ *about the average outcome distribution,*

$$b_t^\gamma = \hat{D}_{Y|\gamma} \coloneqq \frac{1}{|T_\gamma|} \sum_{t' \in T_\gamma} \delta(y_{t'}) \in \Delta(\mathcal{Y}).$$

On all other instances $T \setminus T_\gamma$, *the gambler has belief* $b_t^\gamma = \square$.

(b) The gambler is restricted to play gambles in $\mathbb{B}_1 \coloneqq \{g \in C(\mathcal{Y})\colon \|g\|_1 \leq 1\}$ *where* $\|\cdot\|_1$ *is the* l_1*-norm.*[33]

[33] The l_1-norm is well-defined as $C(\mathcal{Y}) = \mathbb{R}^2$ here.

Then, the re-weighted (by $\frac{|T|}{|T_\gamma|}$) and squared capitals of all gamblers G^γ is equal to,

$$\sum_{\gamma \in v_T} \frac{|T|}{|T_\gamma|} K(G^\gamma)^2 = \sum_{\gamma \in v_T} \frac{|T_\gamma|}{|T|} \left(\frac{1}{|T_\gamma|} \sum_{t \in T_\gamma} y_t - \gamma \right)^2, \tag{11.9}$$

which is the definition of calibration score following [Foster and Vohra, 1998]. Furthermore, the accumulated, absolute capitals of all gamblers G^γ indexed by γ is equal to,

$$\sum_{\gamma \in v_T} |K(G^\gamma)| = \sum_{\gamma \in v_T} \frac{|T_\gamma|}{|T|} \left| \frac{1}{|T_\gamma|} \sum_{t \in T_\gamma} y_t - \gamma \right|, \tag{11.10}$$

which is the definition of expected calibration score (ECE), e.g., following [Gopalan et al., 2022a, Definition 3.1].

Proof. Let us fix a gambler G^γ for some $\gamma \in v_T$. Let us consider a fixed instance $t \in T_\gamma$. Given the gambler's belief, the fair, restricted and rational gambler plays (Lemma 11.37),

$$G_t^\gamma \colon \begin{cases} y \mapsto y - \gamma \text{ if } \mathbb{E}_{\hat{D}_{Y|\gamma}}[Y] > \gamma \\ y \mapsto \gamma - y \text{ otherwise,} \end{cases}$$

For any fixed instance $t \in T \setminus T_\gamma$, the gambler's belief $b_t^\gamma = \square$ is vacuous, hence by Definition 11.26, $G_t^\gamma = 0$. In total, we obtain the following aggregated capital for the gambler G^γ,

$$\begin{aligned} K(G^\gamma) &= \frac{1}{|T|} \sum_{t \in T_\gamma} G_t^\gamma(y_t) \\ &= \frac{1}{|T|} \sum_{t \in T_\gamma} \begin{cases} y_t - \gamma \text{ if } \mathbb{E}_{\hat{D}_{Y|\gamma}}[Y] > \gamma \\ \gamma - y_t \text{ otherwise.} \end{cases} \\ &= \frac{1}{|T|} \left| \sum_{t \in T_\gamma} y_t - |T_\gamma| \gamma \right|. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{\gamma \in v_T} \frac{|T|}{|T_\gamma|} K(G^\gamma)^2 &= \sum_{\gamma \in v_T} \frac{|T|}{|T_\gamma|} \left(\frac{|T_\gamma|}{|T|} \left| \frac{1}{|T_\gamma|} \sum_{t \in T_\gamma} y_t - \gamma \right| \right)^2 \\ &= \sum_{\gamma \in v_T} \frac{|T_\gamma|}{|T|} \left(\frac{1}{|T_\gamma|} \sum_{t \in T_\gamma} y_t - \gamma \right)^2, \end{aligned}$$

and

$$\sum_{\gamma \in v_T} |K(G^\gamma)| = \sum_{\gamma \in v_T} \frac{|T_\gamma|}{|T|} \left| \frac{1}{|T|} \sum_{t \in T_\gamma} y_t - \gamma \right|.$$

□

The previous two recovery results might seem slightly restrictive as the predictive scenarios were simply estimating the mean of a binary outcome. However, this restriction is not necessary as the following recovery result shows. In particular, we refine the beliefs of the gamblers by adding group membership information. The recovered notion of multicalibration generalizes parts of Proposition 11.38.

Proposition 11.39 (Recovery of (Generalized) Multicalibration Score)**.** *Let $\mathcal{S} \subseteq 2^T$ be a set of groups and $v_T \coloneqq \{v_t \colon t \in T\}$. Let $(\mathcal{Y}, \Gamma, T, v, G^{v_T \times \mathcal{S}}, y)$ be an evaluation protocol with $\mathcal{Y} \subseteq \mathbb{R}$ and $\Gamma \colon Q \to 2^{\mathcal{V}}$ being an identifiable property with identification function ν. Furthermore, we assume there is no $\gamma \in \mathcal{V}$ such that $\Gamma^{-1}(\gamma) = Q$. For every $\gamma \in \{v_t \colon t \in T\}$ and $S \in \mathcal{S}$ we define a fair, restricted and approximate rational gambler $G^{\gamma,S}$ such that:*

(a) On all instances $T_{\gamma,S} \coloneqq \{t \in T \colon v_t = \gamma\} \cap S$, the gambler $G^{\gamma,S}$ has a (true) belief $b_t^{\gamma,S}$ about the average outcome distribution,

$$b_t^{\gamma,S} = \hat{D}_{Y|\gamma,S} \coloneqq \frac{1}{|T_{\gamma,S}|} \sum_{t' \in T_{\gamma,S}} \delta(y_{t'}) \in Q.$$

On all other instances $T \setminus T_{\gamma,S}$, the gambler has no belief.

(b) The gambler is restricted to play gambles in $\mathbb{B}_{\alpha_\gamma} \coloneqq \{g \in C(\mathcal{Y}) \colon \|g\| \leq \alpha_\gamma\}$ where $\alpha_\gamma = \|\nu_\gamma\|$ for some norm $\|\cdot\|$.

Then, the reweighted and aggregated capital of the gamblers $G^{\gamma,S}$ indexed by γ and S is equal to,

$$\sup_{S \in \mathcal{S}} \sum_{\gamma \in v_T} \frac{|T|}{|T_{\gamma,S}|} K(G^{\gamma,S})^2 = \sup_{S \in \mathcal{S}} \sum_{\gamma \in v_T} \frac{1}{|T||T_{\gamma,S}|} \left(\sum_{t \in T_{\gamma,S}} \nu(y_t, \gamma) \right)^2,$$

which is approximate $(\mathcal{S}, \nu)$-multicalibration [Noarov and Roth, 2023, Definition 4.2][34].

[34] This is a generalization of L_2-multicalibration following Definition 2.3 in [Garg et al., 2024].

Furthermore, the aggregated capital of the gamblers $G^{\gamma,S}$ indexed by γ and S is equal to,

$$\sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \left|K(G^{\gamma,S})\right| = \sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \frac{|T_{\gamma,S}|}{|T|} \left| \frac{1}{|T_{\gamma,S}|} \sum_{t\in T_{\gamma,S}} \nu(y_t,\gamma) \right|,$$

which is a generalization of L_1-multicalibration [Garg et al., 2024, Definition 2.3].

Proof. Fix $\gamma \in v_T$ and $S \in \mathcal{S}$. Define $\alpha_\gamma = \frac{1}{\|\nu_\gamma\|_1}$. Note that $\|\nu_\gamma\|_1 > 0$, because otherwise $\nu_\gamma = 0$ which implies that $\Gamma(\phi) = \gamma$ for all $\phi \in \Delta(\mathcal{Y})$, which we ruled out by assumption.

Let us consider a fixed instance $t \in T_{\gamma,S}$. Given the gambler's belief, the fair, restricted and approximate rational gambler plays (Lemma 11.40),

$$G_t^{\gamma,S} \colon \begin{cases} y \mapsto \nu(y,\gamma) \text{ if } \mathbb{E}_{\hat{D}_{Y|\gamma,S}}[\nu_\gamma] > 0 \\ y \mapsto -\nu(y,\gamma) \text{ otherwise,} \end{cases}$$

For any fixed instance $t \in T \setminus T_{\gamma,S}$, the gambler's belief $b_t^{\gamma,S} = \square$ is vacuous, hence by Definition 11.29, $G_t^{\gamma,S} = 0$. In total, we obtain the aggregated capital for the gambler $G^{\gamma,S}$,

$$\begin{aligned} K(G^{\gamma,S}) &= \frac{1}{|T|} \sum_{t\in T_{\gamma,S}} G_t^{\gamma,S}(y_t) \\ &= \frac{1}{|T|} \sum_{t\in T_{\gamma,S}} \begin{cases} \nu(y_t,\gamma) \text{ if } \mathbb{E}_{\hat{D}_{Y|\gamma,S}}[\nu_\gamma] > 0 \\ -\nu(y_t,\gamma) \text{ otherwise.} \end{cases} \\ &= \frac{1}{|T|} \left| \sum_{t\in T_{\gamma,S}} \nu(y_t,\gamma) \right|. \end{aligned}$$

Thus,

$$\begin{aligned} \sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \frac{|T|}{|T_{\gamma,S}|} K(G^{\gamma,S})^2 &= \sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \frac{|T|}{|T_{\gamma,S}|} \frac{1}{|T|^2} \left(\sum_{t\in T_{\gamma,S}} \nu(Y,\gamma) \right)^2 \\ &= \sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \frac{1}{|T||T_{\gamma,S}|} \left(\sum_{t\in T_{\gamma,S}} \nu(Y,\gamma) \right)^2, \end{aligned}$$

and

$$\sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \left|K(G^{\gamma,S})\right| = \sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \frac{1}{|T|} \left| \sum_{t\in T_{\gamma,S}} \nu(y_t,\gamma) \right|$$
$$= \sup_{S\in\mathcal{S}} \sum_{\gamma\in v_T} \frac{|T_{\gamma,S}|}{|T|} \left| \frac{1}{|T_{\gamma,S}|} \sum_{t\in T_{\gamma,S}} \nu(y_t,\gamma) \right|.$$

□

Lemma 11.40 (Gamble for Rational Gambler Restricted to Norm Ball - $\mathcal{Y} \subseteq \mathbb{R}$)**.** *Let $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ be an evaluation protocol with $\mathcal{Y} \subseteq \mathbb{R}$ and $\Gamma\colon Q \to 2^{\mathcal{V}}$ being an identifiable property with identification function ν. Furthermore, assume there is no $\gamma \in \mathcal{V}$ such that $\Gamma^{-1}(\gamma) = \Delta(\mathcal{Y})$. Fix $t \in T$ and consider a fair, restricted and approximately rational gambler G with belief $b_t \in Q$ and restriction $\mathbb{B}_{\alpha_{v_t}} \coloneqq \{g \in C(\mathcal{Y})\colon \|g\| \leq \alpha_t\}$ where $\alpha_t = \|\nu_{v_t}\|$ for some norm $\|\cdot\|$.*[35] *Then, in this round $t \in T$, the fair, restricted and approximately rational gambler G plays the gamble,*

$$G_t\colon \begin{cases} y \mapsto \nu(y, v_t) \text{ if } \mathbb{E}_{b_t}[\nu_{v_t}] > 0 \\ y \mapsto -\nu(y, v_t) \text{ otherwise,} \end{cases}$$

Proof. The statement follows since for every $\epsilon > 0$,

$$\begin{aligned} \mathbb{E}_{b_t}[G_t] &= |\mathbb{E}_{b_t}[\nu_{v_t}]| \\ &= \max_{\alpha\in[-1,1]} \mathbb{E}_{b_t}[\alpha\nu_{v_t}] \\ &\overset{(i)}{=} \max_{\alpha\in\mathbb{R} \text{ and } \|\alpha\nu_{v_t}\|_1\leq\alpha_t} \mathbb{E}_{b_t}[\alpha\nu_{v_t}] \\ &\overset{(ii)}{\geq} \max_{g\in\mathcal{F}_{\nu_{v_t}}\cap\mathbb{B}_{\alpha_t}} \mathbb{E}_{b_t}[g] \\ &\overset{(iii)}{\geq} \max_{g\in\mathcal{H}_{\nu_{v_t}}\cap\mathbb{B}_{\alpha_t}} \mathbb{E}_{b_t}[g] - \epsilon \\ &\overset{(iv)}{=} \max_{g\in\mathcal{G}_{\Gamma^{-1}(v_t)}\cap\mathbb{B}_{\alpha_t}} \mathbb{E}_{b_t}[g] - \epsilon, \end{aligned}$$

because,

(i) It holds $\|\alpha\nu_{v_t}\| = |\alpha|\|\nu_{v_t}\| = |\alpha|\alpha_t$.

(ii) By defining $\mathcal{F}_{\nu_{v_t}} \coloneqq \{g \in C(\mathcal{Y})\colon g \leq \alpha\nu_{v_t} \text{ for some } \alpha \in \mathbb{R}\}$.

[35] Note that $\|\nu_{v_t}\| > 0$, because otherwise $\nu_{v_t} = 0$ which implies that $\Gamma(\phi) = v_t$ for all $\phi \in \Delta(\mathcal{Y})$, which we ruled out by assumption.

(iii) Equation 11.8 and for every $\epsilon > 0$ and every $h \in \mathcal{H}_{\nu_{v_t}}$ there exists an $f \in \mathcal{F}_{\nu_{v_t}}$ such that $|\mathbb{E}_{\hat{D}_{Y|S}}[h] - \mathbb{E}_{\hat{D}_{Y|S}}[f]| < \epsilon$ by definition of $\sigma(C(\mathcal{Y}), \mathrm{ca}(\mathcal{Y}))$-closure of $\mathcal{F}_{\nu_{v_t}}$.

(iv) Theorem 11.15.

□

Finally, we equip the gambler with "clairvoyance", analogous to Proposition 11.35. We recover a notion of individual calibration inspired by Equation (4) in [Luo et al., 2022] (cf. [Höltgen and Williamson, 2023] for the notion of calibration on individuals).

Proposition 11.41 (Recovery of Individual Calibration Score)**.** *Let $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ be an evaluation protocol with $\mathcal{Y} \subseteq \mathbb{R}$ and $\Gamma\colon Q \to 2^{\mathcal{V}}$ being an identifiable property with identification function ν. Furthermore, we assume there is no $\gamma \in \mathcal{V}$ such that $\Gamma^{-1}(\gamma) = \Delta(\mathcal{Y})$. We define a fair, restricted and approximate rational gambler G such that:*

(a) On every instance $t \in T$ the gambler G is "clairvoyant", i.e., has a (true) belief b_t about the outcome sampled from a distribution chosen by Nature, $b_t = \delta(y_t) \in Q$.

(b) The gambler is restricted to play gambles in $\mathbb{B}_{\alpha_\gamma} \coloneqq \{g \in C(\mathcal{Y})\colon \|g\| \leq \alpha_\gamma\}$ where $\alpha_\gamma = \|\nu_\gamma\|$ for some norm $\|\cdot\|$.

Then, the capital of the gambler G is equal to,

$$K(G) = \frac{1}{|T|} \sum_{t \in T} |\nu(y_t, v_t)|.$$

Proof. Define $\alpha_t = \frac{1}{\|\nu_{v_t}\|_1}$. Note that $\|\nu_{v_t}\|_1 > 0$, because otherwise $\nu_{v_t} = 0$ which implies that $\Gamma(\phi) = v_t$ for all $\phi \in \Delta(\mathcal{Y})$, which we ruled out by assumption.

Given the gambler's belief, the fair, restricted and approximate rational gambler plays (Lemma 11.40),

$$G_t\colon \begin{cases} y \mapsto \nu(y, v_t) \text{ if } \nu(y_t, v_t) > 0 \\ y \mapsto -\nu(y, v_t) \text{ otherwise.} \end{cases}$$

In total, we obtain the following aggregated capital for the gambler G,

$$K(G) = \frac{1}{|T|} \sum_{t \in T} G_t(y_t) = \frac{1}{|T|} \sum_{t \in T} |\nu(y_t, v_t)|.$$

□

The given recoveries only cover a small fraction of all existing notions. However, as the breadth of provided recoveries suggests, more notions can be recovered by adopting the evaluation protocol, the restriction and the belief on a fair, restricted and (approximately) rational gambler. As the proofs share a lot of their structure, we do not provide more examples in this work.

11.5.4 On the Relationship between Regret and Calibration: Second Act

Besides that the recovery results justify the usefulness of the evaluation protocol in understanding current evaluation metrics, the above statements shed light on the debated duality of calibration scores and loss scores (cf. Section 11.6.2). See Table 11.2 for a summary.

(Zeroed) Loss scores are the result of a clairvoyant gambler with a restriction to regret gambles, while calibration scores are the result of forecast-wise informed gamblers with a restriction to the unit ball. Hence, calibration scores are in two dimensions, belief and restriction, different to (zeroed) loss scores. That makes any comparison between the scores misled.[36] In contrast to the observation that in general *scaled* regret gambles are to a certain extent equivalent to calibration gambles (Section 11.4.9).

11.5.5 Bounds on Evaluation Scores by Hierarchy on Beliefs and Restrictions

What is possible is the comparison along a single dimension. For instance, what is the relationship between two notions when the belief is fixed but the restriction is loosened? Or what is the relationship between two notions when restriction is fixed, but the beliefs of one gambler are more informative about the true outcomes than the beliefs of another gambler? We provide capital bounds along both dimensions. On the one hand, a set-theoretic order on the restrictions provides a hierarchy on the capital of gamblers with fixed belief. On the other hand, a refinement order (Definition 11.44) of the beliefs provides a hierarchy on the capitals of gamblers with fixed restriction.

Let us first take a look on the dimension of restrictions. Intuitively speaking, a gambler which has access to more gambles can choose "better" gambles which make the gambler richer than one which has access to less gambles. Interestingly, the following more general proposition can be used to recover a known result on a bound of swap regret by calibration scores (Corollary 11.43).

[36] One might be inclined to state that this comparison would be a comparison between "apples and oranges". This can be a fitting allegory, as on an abstract level the comparison between spherical objects seems implementable. In our case, the comparability of sets of available gambles for regret and calibration as in Section 11.4.9 forms exactly this abstraction.

Proposition 11.42 (Hierarchy of Capital by Order on Restrictions)**.** *Consider an evaluation protocol* $(\mathcal{Y}, \Gamma, T, v, \{G, G'\}, y)$*. Let* G *and* G' *be fair, restricted and rational gamblers with, for every* $t \in T$*, equal beliefs* $b_t \in \Delta(\mathcal{Y}) \cup \{\square\}$ *which are aligned to truth, but different restricting sets* $R_t \subseteq R'_t \subseteq C(\mathcal{Y})$ *such that* $0 \in R_t$*. Suppose whenever* $b_t = b_{t'}$ *for* $t, t' \in T$*, then* $v_t = v_{t'}$*,* $R_t = R_{t'}$ *and* $R'_t = R'_{t'}$*. Then, the accumulated capital of the gambler* G *is smaller than the accumulated capital of the gambler* G'*,*

$$K(G) \leq K(G').$$

Proof. First, we rewrite,

$$K(G) = \frac{1}{|T|} \sum_{t \in T} G_t(y_t) = \frac{1}{|T|} \sum_{\beta \in b_T} \sum_{t \in T_\beta} G_t(y_t) = \sum_{\beta \in b_T} \frac{|T_\beta|}{|T|} \frac{1}{|T_\beta|} \sum_{t \in T_\beta} G_t(y_t), \tag{11.11}$$

where $T_\beta \coloneqq \{t \in T \colon b_t = \beta\}$ and $b_T \coloneqq \{b_t \colon t \in T\}$. Let us focus on a single $\beta \in b_T$. If $\beta = \square$, then $G_t = G'_t = 0$ for $t \in T_\beta$ and we are done. Hence, suppose that $\beta \neq \square$.

By assumption, for every $t \in T_\beta$, v_t, G_t, G'_t are constant, hence we write without loss of generality, $v_\beta, G_\beta, G'_\beta$. By rationality of the gamblers, for every $t \in T_\beta$,

$$\mathbb{E}_\beta[G_t] \geq \mathbb{E}_\beta[g]$$

for all $g \in G_\beta \cap \mathcal{G}_{\Gamma^{-1}(v_\beta)}$, and, for every $t \in T_\beta$,

$$\mathbb{E}_\beta[G'_t] \geq \mathbb{E}_\beta[g]$$

for all $g \in G'_\beta \cap \mathcal{G}_{\Gamma^{-1}(v_\beta)}$. It follows, for all $t \in T_\beta$,

$$\mathbb{E}_\beta[G_t] \leq \mathbb{E}_\beta[G'_t],$$

because $G_\beta \subseteq G'_\beta$. Furthermore, for all $t, t' \in T_\beta$, $G_t = G_{t'}$ (as well as $G'_t = G'_{t'}$), i.e., we can write G_β (respectively G'_β instead).

By alignment to truth (Definition 11.28),

$$\frac{1}{|T_\beta|} \sum_{t \in T_\beta} G_t(y_t) = \mathbb{E}_\beta[G_\beta].$$

With the last two observations we can continue rewriting Equation (11.11),

$$\sum_{\beta \in b_T} \frac{|T_\beta|}{|T|} \mathbb{E}_\beta[G_\beta] \leq \sum_{\beta \in b_T} \frac{|T_\beta|}{|T|} \mathbb{E}_\beta[G'_\beta] = \frac{1}{|T|} \sum_{t \in T} G'_t(y_t) = K(G').$$

□

For Proposition 11.42 two hidden assumptions are of crucial importance. (a) The beliefs have to be aligned to truth. This guarantees that the expected outcome of the gamble aligns to the true (average) outcome. (b) The forecasts and restrictions have to be constant for instances for which the belief is constant. This guarantees that the set of gambles from which both gamblers choose remains constant. Otherwise, it is hard to show the superiority of the choice of one gambler over the choice of another gambler.

We can directly apply Proposition 11.42 to better understand the relationship between calibration scores and swap regret. We recover Theorem 12 in [Kleinberg et al., 2023], which exists in different formulations and flavors at several places of the literature (cf. Section 11.6.2)

Corollary 11.43 (Recovery of Theorem 12 in [Kleinberg et al., 2023])**.** *Let* $\mathcal{Y} = \{0,1\}$ *and* $v_T \coloneqq \{v_t \colon t \in T\}$*. Let Forecaster* $v_t \in [0,1]$ *predict the mean, denoted as* Γ*, for all* $t \in T$*. Let* $\ell \colon \{0,1\} \times [0,1] \to [-1,1]$ *be a consistent scoring function for* Γ*. For a fixed tuple of outcomes* $(y_t)_{t\in T}$*, the swap regret with respect to* ℓ *of the forecaster* v *is upper bounded by the re-scaled expected calibration error,*

$$\max_{\sigma\in\{\mathcal{V}\to\mathcal{V}\}} \sum_{t\in T} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)) \leq 4 \sum_{\gamma\in v_T} |T_\gamma| \left| \frac{1}{|T_\gamma|} \sum_{t\in T_\gamma} y_t - \gamma \right|.$$

Proof. First, we define $\ell' \coloneqq \frac{1}{4}\ell$ and an evaluation protocol $(\mathcal{Y}, \Gamma, T, v, \{G, \tilde{G}\}, y)$. For every $\gamma \in \{v_t \colon t \in T\}$ let us introduce two fair, restricted and rational gamblers: G^γ and $\tilde{G}^\gamma$. Both share their beliefs $b_t^\gamma = \hat{D}_{Y|\gamma} \subseteq \Delta(\mathcal{Y})$ for all $t \in T_\gamma \coloneqq \{t \in T \colon v_t = \gamma\}$ and $b_t^\gamma = \square$ for all $t \in T \setminus T_\gamma$. The beliefs are aligned to the truth (Definition 11.28). But the gamblers are differently restricted, $G_t^\gamma \in \mathcal{L}'_\gamma$ and $\tilde{G}_t^\gamma \in \mathbb{B}_1$. Where $\mathcal{L}'_\gamma \coloneqq \{\ell'_\gamma - \ell'_c \colon c \in [0,1]\}$. Note that $\mathcal{L}'_\gamma \subseteq \mathbb{B}_1$, as for every $g \in \mathcal{L}'_\gamma$,

$$\|g\|_1 = \|\ell_\gamma - \ell_c\|_1 = |\ell(1,\gamma) - \ell(1,c)| + |\ell(0,\gamma) - \ell(0,c)| \leq 1,$$

because $\ell' \colon \{0,1\} \times [0,1] \to [-\frac{1}{4}, \frac{1}{4}]$.

By Proposition 11.38 Equation 11.10 the right hand side of the above inequality can be written as,

$$|T| \sum_{\gamma\in v_T} K(\tilde{G}), \quad \text{and} \quad |T| \sum_{\gamma\in v_T} K(G).$$

respectively for the left hand side (Proposition 11.33). The inequality then follows by Proposition 11.42 and rescaling the loss from ℓ' to ℓ. □

Along the dimension of beliefs, we define a hierarchy on the beliefs by their fine-grainedness. Essentially, the more fine-grained a belief is, the more informed the gambler with the belief.

Definition 11.44 (Refinement of Belief)**.** *Let $T \coloneqq \{1, \dots, n\}$ and $b_t, b'_t \in \Delta(\mathcal{Y}) \cup \{\square\}$ for every $t \in T$. We define $T_\beta \coloneqq \{t \in T \colon b_t = \beta\}$ and $b_T \coloneqq \{b_t \colon t \in T\}$. The beliefs $(b_t)_{t\in T}$ are* refined by *$(b'_t)_{t\in T}$, if for all $\beta \in b_T$, $\beta = \square$ or $b'_t \in \Delta(\mathcal{Y})$ for all $t \in T_\beta$ and $\beta = \frac{1}{|T_\beta|}\sum_{t\in T_\beta} b'_t$.*

Refinement of belief does not require any belief to be aligned to truth (Definition 11.28). However, by comparing the Definition 11.44 to Definition 11.28, we observe that, beliefs $(b_t)_{t\in T}$ are aligned to truth, if and only if, $(\delta(y_t))_{t\in T}$ refines (b_t) under an appropriate choice of evaluation protocol.

The more informed a gambler is about the true outcomes the better it can optimize its capital. That is what we formalize in the following.

Proposition 11.45 (Hierarchy on Capital by Refinements on Belief)**.** *Consider an evaluation protocol $(\mathcal{Y}, \Gamma, T, v, \{G, G'\}, y)$. Let G and G' be fair, restricted and rational gamblers with, for every $t \in T$, equal restriction $G_t \subseteq C(\mathcal{Y})$ such that $0 \in R$, but different beliefs $b_t, b'_t \in \Delta(\mathcal{Y}) \cup \{\square\}$ for every $t \in T$. Suppose whenever $b_t = b_{t'}$ and $b'_t = b'_{t'}$ for $t, t' \in T$, then $v_t = v_{t'}$ and $G_t = G_{t'}$. If the beliefs $(b'_t)_{t\in T}$ are aligned to truth and $(b'_t)_{t\in T}$ refines $(b_t)_{t\in T}$, then the accumulated capital of the gambler G is smaller than the accumulated capital of the gambler G',*

$$K(G) \leq K(G').$$

Proof. First, we rewrite, analogous to the proof of Proposition 11.42,

$$\begin{aligned} K(G) = \frac{1}{|T|}\sum_{t\in T} G_t(y_t) &= \frac{1}{|T|}\sum_{\beta\in b_T}\sum_{t\in T_\beta} G_t(y_t) && (11.12)\\ &= \sum_{\beta\in b_T}\frac{|T_\beta|}{|T|}\frac{1}{|T_\beta|}\sum_{t\in T_\beta} G_t(y_t) && (11.13)\\ &= \sum_{\beta\in b_T}\frac{|T_\beta|}{|T|}\sum_{t\in T_\beta}\frac{|T_{\beta'}|}{|T_\beta|}\sum_{t\in T_{\beta'}}\frac{1}{|T_{\beta'}|}G_t(y_t), && (11.14) \end{aligned}$$

where $T_\beta \coloneqq \{t \in T \colon b_t = \beta\}$, $b_T \coloneqq \{b_t \colon t \in T\}$ and $T_{\beta'} = T_\beta \cap \{t \in T \colon b'_t = \beta'\}$ for $\beta' \in \{b'_t \colon t \in T_\beta\}$.

Let us focus on the last sum over $t \in T_{\beta'}$. Clearly, b_t and b'_t are constant for all $t \in T_{\beta'}$. Hence, by assumption, for all $t \in T_{\beta'}$ v_t and G_t are constant. We rewrite without loss of generality $v_{\beta'}$ and $G_{\beta'}$. Furthermore, by fairness, restrictiveness and rationality of the gamblers, for all $t, t' \in T_{\beta'}$, $G_t = G_{t'} \in G_{\beta'} \cap \mathcal{G}_{\Gamma^{-1}(v_{\beta'})}$ (respectively $G'_t = G'_{t'} \in G_{\beta'} \cap \mathcal{G}_{\Gamma^{-1}(v_{\beta'})}$), i.e., we can write $G_{\beta'}$ (respectively $G'_{\beta'}$ instead). We open two cases:

Suppose $\beta' = \square$**.** Then, $\beta = \square$ by refinement of belief (Definition 11.44), hence $G_{\beta'} = G'_{\beta'} = 0$. Thus,

$$\sum_{t \in T_{\beta'}} \frac{1}{|T_{\beta'}|} G_t(y_t) = 0 = \sum_{t \in T_{\beta'}} \frac{1}{|T_{\beta'}|} G'_t(y_t).$$

Suppose $\beta' \in \Delta(\mathcal{Y})$**.** Focusing on the gambler G', for every $t \in T_{\beta'}$,

$$\mathbb{E}_{\beta'}[G'_{\beta'}] \geq \mathbb{E}_{\beta'}[g]$$

for all $g \in G_{\beta'} \cap \mathcal{G}_{\Gamma^{-1}(v_{\beta'})}$, in particular,

$$\mathbb{E}_{\beta'}[G'_{\beta'}] \geq \mathbb{E}_{\beta'}[[G_{\beta'}].$$

Since $(b'_t)_{t \in T}$ is aligned to truth,

$$\sum_{t \in T_{\beta'}} \frac{1}{|T_{\beta'}|} G_t(y_t) = \mathbb{E}_{\beta'}[G_{\beta'}] \leq \mathbb{E}_{\beta'}[G'_{\beta'}] = \sum_{t \in T_{\beta'}} \frac{1}{|T_{\beta'}|} G'_t(y_t).$$

Combining both of the cases and Equation (11.12) gives the result. □

As in Proposition 11.42, we need consistency assumptions on the restrictions and forecasts when the beliefs don't change. Again, this guarantees that the more informed gambler actually chooses among the same set of gambles the superior one (in terms of final capital) in comparison to the less informed gambler. Otherwise the statement does not hold in general as the following examples shows.

Example 11.46. *Proposition 11.45 does* not *hold if the restriction G varies arbitrarily over time. For instance, let $\mathcal{Y} = \{0, 1\}$, Γ be the mean, $T = \{t_1, t_2, t_3\}$, $v = (\frac{5}{12}, \frac{5}{12}, \frac{5}{12})$, $y = (0, 1, 0)$ and G, G' be two fair, restricted and rational gamblers with beliefs $b_t = \frac{1}{3}$ for all $t \in T$ respectively $(b'_t)_{t \in T} = (\frac{1}{2}, \frac{1}{2}, 0)$ and restrictions $(G_t)_{t \in T} = (\mathbb{B}_1, C(\mathcal{Y})_{\leq 0}, C(\mathcal{Y})_{\leq 0})$. We observe the following:*

1. *Both gamblers have beliefs which are aligned to truth. The beliefs of G' refine the beliefs of G.*
2. *Both gamblers, as being rational and restricted with $C(\mathcal{Y})_{\leq 0}$, play $G_t = G'_t = 0$ for $t \in \{t_2, t_3\}$.*
3. *For $t = t_1$, the gambler G plays $G_{t_1} : y \mapsto \frac{5}{12} - y$, while the gambler G' plays $G'_{t_1} : y \mapsto y - \frac{5}{12}$.*

As a result, the capitals compare as follows,

$$K(G) = \frac{1}{3}\left(0 + 0 + \frac{5}{12}\right) \geq \frac{1}{3}\left(0 + 0 - \frac{5}{12}\right) = K(G').$$

The attentive reader might have noticed that in Table 11.2 we essentially provided a hierarchy of refining beliefs, from $\hat{D}_Y$ over $\hat{D}_{Y|\gamma}$ to $\delta(y_t)$. This suggests the two trivial recovery corollaries below. External regret is upper bounded by swap regret, which is upper bounded by (zeroed) loss score. Bias in the large is upper bounded by calibration score (ECE), which is upper bounded by individual calibration score. We omit the detailed proof as the statements follow almost immediately. The ideas are simple. From external to swap regret/bias in the large to calibration score: For every gambler with constant belief $\hat{D}_{Y|\gamma}$ which we introduce in Proposition 11.33, we introduce a gambler with constant belief $\hat{D}_Y$. Then, apply Proposition 11.45. From swap regret to zeroed loss scores/calibration score to individual calibration score: Directly apply Proposition 11.45.

Corollary 11.47 (Recovery of Known Order of Regret Scores)**.** *Let $\mathcal{Y} \subseteq \mathbb{R}^d$, $|T| < \infty$, $y_t \in \mathcal{Y}$ for all $t \in T$, $v_t \in \mathcal{V}$ for all $t \in T$ and $\Gamma\colon Q \to 2^{\mathcal{V}}$ an elicitable property with consistent scoring function ℓ. It holds,*

$$\sum_{t\in T} \ell(y_t, v_t) - \min_{c\in\mathcal{V}} \ell(y_t, c) \leq \max_{\sigma\in\{\mathcal{V}\to\mathcal{V}\}} \sum_{t\in T} \ell(y_t, v_t) - \ell(y_t, \sigma(v_t)) \leq \sum_{t\in T} \ell(y_t, v_t) - \ell(y_t, \Gamma(\delta(y_t)).$$

Corollary 11.48 (Recovery of Known Order of Calibration Scores)**.** *Let $\mathcal{Y} = \{0, 1\}$, $|T| < \infty$, $y_t \in \mathcal{Y}$ for all $t \in T$, $v_t \in [0, 1]$ for all $t \in T$ and $\Gamma\colon Q \to [0, 1]$ be the mean with identification function $\nu(y, \gamma) = y - \gamma$. It holds,*

$$\left|\frac{1}{|T|}\sum_{t\in T}(y_t - v_t)\right| \leq \sum_{\gamma\in v_T} \frac{|T_\gamma|}{|T|}\left|\frac{1}{|T_\gamma|}\sum_{t\in T_\gamma} y_t - \gamma\right| \leq \frac{1}{|T|}\sum_{t\in T}|y_t - v_t|.$$

11.6 Related Work

The content of this paper has been motivated by and built upon several strands of work. We contextualize our findings in two sections. One section focuses on the testing of forecasts. The other section concentrates on the duality of regret and calibration.

11.6.1 Testing Forecasts

Our suggested evaluation protocol for the recovery of existing notions was motivated by existing literature.

Meteorology and historical probabilistic forecast evaluation. Historically, the question of how to evaluate (probabilistic) forecasts has been a major concern in meteorology (e.g., [Brier, 1950, Murphy and Epstein, 1967]). With the advent of powerful machine learning systems which produce forecasts of arbitrary kinds, we can observe the renaissance of old debates on what constitutes a "good" (weather) forecast. In [Murphy and Epstein, 1967], the authors suggest a separation of operational evaluation, what is the value of the forecast to the user, and empirical evaluation, how well do forecasts and observations align. Even though, we don't fully agree to the meaningfulness of this separation we do want to highlight in this respect that our work is focusing on "statistical adequateness" criteria alone. We are convinced that for a "successful" forecast more adequateness criteria have to be met (cf. Section 11.7.1).

Manipulability and the limit of sequential tests. The testing of (sequential) forecasts has received interest in the economics literature. In particular, this strand of literature has been concerned with the manipulability of tests, i.e., whether Forecaster can fake knowing the truth and still pass a test. Key contributions included proofs which show that if a test is complete, i.e., a test accepts the truth (which is type-I error freeness), then the test is manipulable [Foster and Vohra, 1998, Lehrer, 2001, Sandroni, 2003, Olszewski and Sandroni, 2011]. Central to these results is the option for Forecaster to randomize over possible forecasts [Vovk et al., 2005, Vovk, 2005, 2007, Olszewski and Sandroni, 2011]. We showed in Proposition 11.6 that in our evaluation protocol completeness is equivalent to a fair gambler, i.e., a gambler which only plays available gambles. In follow-up works, scholars argued about the existence of non-manipulable but complete tests, e.g., in [Fortnow and Vohra, 2009] the authors show that the faking forecaster is computationally "complex",in [Shmaya, 2008] the author suggests a Borel-set construction of a non-manipulable test via the axiom of choice, and in [Olszewski and Sandroni, 2009] the authors equip the test to ask for counterfactual queries which makes it resilient to manipulating forecasts.

Forecast evaluation as test - From classical tests to e-variables and non-negative supermartingales. Finally, the evaluation of a forecast can equivalently be understood as a classical statistical test (e.g., [Neyman and Pearson, 1933]) whether the (composite) null hypothesis, i.e., the forecasting set defined through the forecast and the forecasted property, fits to the data observed, i.e., what Nature reveals. Recently, e-values, a replacement for p-values, have gained increasing attention [Grünwald et al., 2020, Vovk and Wang, 2021]. Essentially, e-variables, i.e., unrealized e-values, is a gambling approach to testing statistical hypothesis testing. The studied questions and the setup in this work has been motivated by the surrounding field of game-theoretic probability and statistics [Shafer and Vovk, 2019] (going back to [Ville, 1939]).

E-variables are, in our language, gambles, for which the availability constraint (Definition 11.5) is replaced by an expectation being equal or less than 1 and the restriction (Definition 11.26) is replaced by non-negativity. The analogue of rationality (Definition 11.26) is growth rate optimality (GRO). An e-variable is growth rate optimal, if its expected logarithmic value is maximal under the alternative hypothesis. In our case, a gambler is rational, if the expected values of its gambles is maximal under its belief, which is the alternative hypothesis. Hence, our rationality demand is comparable to GRO up to the concave monotone transformation of the "units" of the gamble's outcome.

In the realm of game-theoretic statistics, the closest work to ours is [Casgrain et al., 2022]. Casgrain et al. [2022] propose test supermartingales (non-negative supermartingales with initial expected value less or equal to 1) introduced in [Shafer et al., 2011] to test a composite null hypothesis which is equivalent to the level sets which we consider for elicitable (respectively identifiable) properties (cf. Definition 11.3 and Proposition 11.9). Test supermartingales are, (a) a special case of an e-process [Ramdas et al., 2022], which is the sequential generalization of an e-value [Ramdas et al., 2023] and (b) a sum of available gambles plus an initial value in $[0, 1]$.

A central difference between ours and their work is that in [Casgrain et al., 2022] the authors consider a stationary forecaster. In our setting, Forecaster can provide a new forecast for every instance. That is why we concentrate on identifying properties of single-instance based gambles, while [Casgrain et al., 2022] consider sequential tests. Nevertheless, they provide answers to related questions. The authors identify a subset of the possible supermartingales which tests the level set null hypothesis. In comparison, we provide full characterizations of the set of available gambles, i.e., "instance-wise tests" (Theorem 11.13 and Theorem 11.15). The authors maximize the power of their test, i.e., minimize the type-II error, based on GRO against an alternative hypothesis which is the empirical distribution of the seen instances. We consider rationality, the analogue of GRO as laid out above, against an alternative hypothesis which can include the actual outcomes of unseen realizations. This is obviously irrealistic, hence, we emphasis that our (recovery) results can only be interpreted as "as if"-statements. Finally, they state that the test supermartingales which they suggest for level sets of identifiable properties are more powerful, than for elicitable properties [Casgrain et al., 2022, p. 9 & 10]. We observe an analogous hierarchy in Section 11.4.9.

Because of the multiplicative structure of the test supermartingales in [Casgrain et al., 2022] and the logarithm in the GRO criterion [Grünwald et al., 2020], we summarize on an abstract level, that our theory is providing an "additive analogue" to the more "multiplicative" theory of game-theoretic statistics.

11.6.2 Duality of Regret-Type and Calibration-Type Consistency

In Section 11.4.9 and Section 11.5.4 we stumble upon a fundamental duality of consistency criteria of probability assignments: regret-type and calibration-type consistency criteria.

Internal consistency. In understanding "consistency" as *internal consistency*, i.e., the probability assignments on a single happening are not contradicting, the loss-type and calibration-type duality dates back at least to [De Finetti, 1974/2017]. De Finetti suggested two different definitions of internal consistency, what he calls *coherence* [Schervish et al., 2009]. One definition, the regret-type, states that a probabilistic assignment is coherent if it minimizes a scoring rule. The other definition, the calibration-type, requires that a gambler with access to available, scaled calibration gambles (cf. Theorem 11.4.4) does not get rich by gambling across events. If one type of coherence is fulfilled then the probabilities follow the standard rules of probability, e.g., sum to one, disjoint additivity.[37]

External consistency. On the other hand, we can understand "consistency" as *external consistency*, i.e., the probability assignments match empirical observations. Analogous to internal consistency there exist equivalence statements between regret-type and calibration-type definitions. First, to our knowledge, were DeGroot and Fienberg [1983, Equation 4.1 & 4.2] who provide a direct translation of calibration scores to swap regret scores for the Brier scoring function. Similar bridges have been established in Foster and Vohra [1998] and [Foster and Vohra, 1999, Section 2.3]. Independent of these works, Dawid [1985a, Theorem 8.1] showed that a (computably) calibrated (computable) forecast has lower proper scoring rule than any compared computable forecast for infinite time horizon.

More recently, equivalences between regret and calibration criteria of forecast regained interest. The reason for this is that multicalibration, a generalization of calibration, was introduced as a possible candidate to guarantee fair forecasts [Chouldechova, 2017, Hébert-Johnson et al., 2018]. For instance, Globus-Harris et al. [2023, Theorem 3.2] provide a characterization for multicalibration as kind of a "swap regret". Closely related Gopalan et al. [2024b] showed the equivalence of "swap regret" omnipredictor and multicalibrated predictor. The equivalence relationship between "swap regret" and calibration has been generalized beyond the mean in [Derr et al., 2026]. Finally, Kleinberg et al. [2023, Theorem 12] and Noarov et al. [2023] give results which imply low swap regret for (multi-)calibrated predictors for relatively general loss functions.[38] In particular, we recover [Kleinberg et al.,

[37] Schervish et al. [2009] trace back accounts for coherence of probabilistic statements and show limits when it comes to their generalization to imprecise probabilities. This has been done for the approach via calibration gambles in the seminal works of Walley [1991] and Williams [2007]. An analogous generalized theory for losses is still under development [Konek, 2023].

[38] Surprisingly, already DeGroot and Fienberg [1983, Theorem 4] made some unnoticed progress in that

2023, Theorem 12] in Corollary 11.43 using the more general Proposition 11.42.

The close relationship between regret-type external consistency and calibration-type external consistency becomes apparent time and again. Nevertheless, cases are made advocating calibration over regret [Foster and Vohra, 1998, Zhao et al., 2021, Kleinberg et al., 2023] other cases are made arguing for the advantages of regret over calibration [Schervish, 1985, Seidenfeld, 1985, Dawid, 1985a]. Our Section 11.4.9 and Section 11.5.4 highlight why the debate "calibration versus regret" needs to be done at the right level of comparison: along the level of available gambles (Section 11.4.9) or along the level of scores (Section 11.5.4).

Elicitation and identification of properties. Finally, the strand of statistical literature on property elicitation mentioned already in Section 11.3.1 provides further relationships between the regret and the calibration world. Because of the centrality of scoring functions, regret, on an abstract level, is linked to elicitability (Definition 11.7). To remind the reader, a property is elicitable if it minimizes a consistent scoring function. The fundamental result by Steinwart et al. [2014] generalizes the equivalence of elicitable and identifiable properties, if the property is real-valued and continuous given in [Lambert et al., 2008]. In particular, a sufficiently differentiable scoring function which elicits a property and an identification function which identifies the same property are linked via the gradients of the scoring function [Fissler and Ziegel, 2016, Theorem 3.2]. Then, in [Noarov and Roth, 2023] and [Gneiting and Resin, 2023] calibration got understood as approximating the identification criterion of a forecast. Similar to [Gopalan et al., 2022c, Deng et al., 2023] they observe that standard calibration, e.g., Proposition 11.38, largely is based on a specific instance-wise comparison between observation and forecast, namely $y - v$. This comparison is an identification function of the mean. Hence, [Noarov and Roth, 2023] and [Gneiting and Resin, 2023] generalized this instance-wise comparison to arbitrary identification functions. In particular, Noarov and Roth [2023] identified the set of calibratable properties by the set of elicitable respectively identifiable properties. Concluding, calibration, on an abstract level, is linked to identifiability. If elicitablity and identifiability both are given, then equivalence of the two worlds calibration and regret is, at least to a certain extent modulo scaling, given (Corollary 11.25).

11.7 Conclusion

We have argued that most used evaluation metrics used in machine learning can be viewed as the result of a fair gambling protocol between a forecaster and a gambler with certain restrictions and beliefs. Provocatively written, **all machine learning evaluation is gam-**

regard. They argued that swap regret is controlled by the calibration of Forecaster. The authors didn't use the term "swap regret". We refer to Equation (5.5) in their work as swap regret.

bling. We are convinced this insight adds to the debate on the value of current evaluation practice in machine learning when facing (social) reality, e.g., [Liu et al., 2022, Wang et al., 2024]. For instance, we can now ask whether we want to let the success of gamblers decide which methods to favor when doing breast cancer prediction, e.g., [Chen et al., 2025]?

Furthermore, our work sheds light on the relationship of regret-type and calibration-type evaluation metrics. Both those types of metrics are fair. The difference of the types are different restrictions on the set of gambles which the gambler is allowed to play beyond the available ones. Then, a meaningful comparison between the obtained scores is possible when the belief is fixed, e.g., individual calibration scores with loss scores, standard calibration scores with swap regret or multicalibration scores with group-wise swap regret. Nevertheless, scaled regret-type evaluation metrics can, in principle, make up for the difference between the set restrictions. Regret and calibration are two flavors of evaluation metrics, theoretically equivalent in their power, practically not comparable.

11.7.1 Future Work

Several interesting, open questions remain to be answered in future work. For instance, can we characterize the set of available gambles for a forecasting set induced by an elicitable or identifiable property by a criterion beyond calibration or regret? Related to this, how can it be guaranteed that a forecasting set of an arbitrary property, which we only demand to be credal, adequately describes Nature's outcome (cf. Appendix 11.8.4)? On the other hand, what are more meta-criteria for reasonable evaluation metrics beyond "fairness"? Finally, how does this evaluation protocol interact with evaluation methodologies, e.g., benchmarking [Hardt, 2025], which provide more semantics to the obtained scores?

Acknowledgements

The authors are very grateful to Christian Fröhlich, Benedikt Höltgen, Aaron Roth, Giacomo Molinari and Min Jae Song for insightful discussions. Thanks as well for the feedback of the audience attending the talk "Four Facets of Evaluating Predictions" at Harvard University, USA, November 2023, and the talk "Four Facets of Forecast Felicity" during the Workshop on "Learning under Weakly Structured Information" at the University of Tübingen, Germany, April 2025. Thanks to the International Max Planck Research School for Intelligent System (IMPRS-IS) for supporting Rabanus Derr.

This work was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645; it was also supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center.

11.8 Appendix

11.8.1 Representability of Evaluation Metrics

Proposition 11.49 (Characterization of Representability)**.** *Let $T \coloneqq \{1, \dots, n\}$, $\mathcal{Y}$ be an outcome set and $\mathcal{V}$ a set of property values. An evaluation metric $m \colon \mathcal{Y}^T \times \mathcal{V}^T \to \mathbb{R}$ is* representable in an evaluation protocol*, if and only if, for all $y \in \mathcal{Y}^T$ and $v \in \mathcal{V}^T$, there exists a set of functions $f_t^i \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ indexed by $t \in T$ and $i \in I$ for some finite I and an aggregation $\mathbf{A} \colon \mathbb{R}^{|I|} \to \mathbb{R}$, such that,*

$$m((y_t)_{t\in T}, (v_t)_{t\in T}) = \mathbf{A}\left[\left(\frac{1}{|T|}\sum_{t\in T} f_t^i(y_t, v_t)\right)_{i\in I}\right].$$

Proof. Given T, v, y and $\mathcal{V}$, let $\Gamma \colon Q \to 2^{\mathcal{V}}$ be arbitrary and define the evaluation protocol $(\mathcal{Y}, \Gamma, T, v, G^I, y)$, where for every $i \in I$, we define a gambler G^i such that $G_t^i \colon y \mapsto f_t^i(y, v_t)$. Then, the capital of each such gambler is,

$$K(G^i) = \frac{1}{|T|}\sum_{t\in T} f_t^i(y_t, v_t).$$

By choosing the appropriate aggregation function $\mathbf{A}$ we obtain,

$$\mathbf{A}[(K(G^i))_{i\in I}] = \mathbf{A}\left[\left(\frac{1}{|T|}\sum_{t\in T} f_t^i(y_t, v_t)\right)_{i\in I}\right].$$

The reverse direction follows analogously, by demanding that for every $i \in I$ and $t \in T$ there is a function $f_t^{G^i}(\cdot, v_t) \colon y \mapsto G_t^i(y)$. □

In particular, every single-instance based evaluation metric is representable.

Proposition 11.50 (Single-Instance Based Evaluation Metrics are Representable)**.** *Let $T \coloneqq \{1, \dots, n\}$, $\mathcal{Y}$ be an outcome set and $\mathcal{V}$ a set of property values. The single-instance based evaluation metric $m \colon \mathcal{Y}^T \times \mathcal{V}^T \to \mathbb{R}$,*

$$m((y_t)_{t\in T}, (v_t)_{t\in T}) = \mathbf{A}[(f_t(y_t, v_t))_{t\in T}],$$

for some aggregation function $\mathbf{A} \colon \mathbb{R}^T \to \mathbb{R}$, and some $f_t \colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ for each $t \in T$ is representable in an evaluation protocol.

Proof. Given T, v, y and $\mathcal{V}$, let $\Gamma \colon Q \to 2^{\mathcal{V}}$ be arbitrary and define the evaluation protocol $(\mathcal{Y}, \Gamma, T, v, G^I, y)$, where for every $\tilde{t} \in T$, we define a gambler $G^{\tilde{t}}$ such that $G_t^{\tilde{t}} = 0$ if $t \neq \tilde{t}$

and $G^{\tilde{t}}_{\tilde{t}}\colon y \mapsto f_{\tilde{t}}(y_{\tilde{t}}, y)$. Then, the averaged capital of each such gambler is,

$$K(G^{\tilde{t}}) = \frac{1}{|T|} f_{\tilde{t}}(y_{\tilde{t}}, v_{\tilde{t}}).$$

By choosing the appropriate aggregation function $\mathbf{A}'$, which is $\mathbf{A}'[(r_t)_{t\in T}] = \mathbf{A}[|T|(r_t)_{t\in T}]$ for $(r_t)_{t\in T} \in \mathbb{R}^{|T|}$, we obtain,

$$\mathbf{A}'[(K(G^t))_{t\in T}] = \mathbf{A}[(f_t(y_t, v_t))_{t\in T}].$$

□

11.8.2 Elicitable or Identifiable Properties Have Convex and Closed Level Sets

Proposition 11.51 (Convex and Closed Level Sets)**.** *For all $\gamma \in \mathcal{V}$ the level set $\Gamma^{-1}(\gamma) \subseteq \Delta(\mathcal{Y})$ of an elicitable or identifiable property Γ is credal, i.e., convex and $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed.*

Proof. Let Γ be elicitable with scoring function ℓ. By Definition 11.7, for all $\gamma \in \mathcal{V}$, $\ell_\gamma \in C(\mathcal{Y})$. Hence, all integrals are well-defined. Fix any $\gamma \in \mathcal{V}$. By assumption in Definition 11.3, $\Gamma^{-1}(\gamma) \neq \emptyset$. The convexity of the level sets follows by a straightforward argument comparable to [Steinwart et al., 2014, Appendix B Theorem 13]). Let $\phi_1, \phi_2 \in \Gamma^{-1}(\gamma)$ and $\alpha \in [0,1]$, for all $\gamma' \in \mathcal{V}$,

$$\begin{aligned}\langle \alpha\phi_1 + (1-\alpha)\phi_2, \ell_{\gamma'}\rangle &= \alpha\langle\phi_1, \ell_{\gamma'}\rangle + (1-\alpha)\langle\phi_2, \ell_{\gamma'}\rangle \\ &\geq \alpha\langle\phi_1, \ell_{\gamma}\rangle + (1-\alpha)\langle\phi_2, \ell_{\gamma}\rangle \\ &= \langle \alpha\phi_1 + (1-\alpha)\phi_2, \ell_{\gamma}\rangle,\end{aligned}$$

hence, $\alpha\phi_1 + (1-\alpha)\phi_2 \in \Gamma^{-1}(\gamma)$ by convexity of Q. For the closure, observe that

$$\begin{aligned}\Gamma^{-1}(\gamma) &\coloneqq \{\phi \in Q\colon \gamma \in \Gamma(\phi)\} \\ &= \{\phi \in Q\colon \mathbb{E}_\phi[\ell(Y,\gamma)] \leq \mathbb{E}_\phi[\ell(Y,c)], \forall c \in \mathcal{V}\} \\ &= \{\phi \in Q\colon \langle \phi, \ell_\gamma - \ell_c\rangle \leq 0, \forall c \in \mathcal{V}\} \\ &= \{\phi \in \mathrm{ca}(\mathcal{Y})\colon \langle \phi, \ell_\gamma - \ell_c\rangle \leq 0, \forall c \in \mathcal{V}\} \cap Q \\ &= \bigcap_{c\in\mathcal{V}} \{\phi \in \mathrm{ca}(\mathcal{Y})\colon \langle \phi, \ell_\gamma - \ell_c\rangle \leq 0\} \cap Q,\end{aligned}$$

is $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed since $\{\phi \in \mathrm{ca}(\mathcal{Y})\colon \langle \phi, \ell_\gamma - \ell_c\rangle \leq 0\}$ is closed by the definition of $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-topology and any intersection of closed sets is closed.

Let Γ be identifiable with identification function ν. By Definition 11.8, for all $\gamma \in \mathcal{V}$, $\nu_\gamma \in C(\mathcal{Y})$. Hence, all integrals are well-defined. Fix any $\gamma \in \mathcal{V}$. By assumption in Definition 11.3, $\Gamma^{-1}(\gamma) \neq \emptyset$. The convexity of the level sets follows directly by the definition of identification function. For the closure, observe that

$$\begin{aligned}\Gamma^{-1}(\gamma) &\coloneqq \{\phi \in Q \colon \gamma \in \Gamma(\phi)\} \\ &= \{\phi \in Q \colon \langle \phi, \nu_\gamma \rangle = 0\} \\ &= \{\phi \in \mathrm{ca}(\mathcal{Y}) \colon \langle \phi, \nu_\gamma \rangle = 0\} \cap Q \\ &= \{\phi \in \mathrm{ca}(\mathcal{Y}) \colon \langle \phi, \nu_\gamma \rangle \leq 0\} \cap \{\phi \in \mathrm{ca}(\mathcal{Y}) \colon \langle \phi, \nu_\gamma \rangle \geq 0\} \cap Q,\end{aligned}$$

is $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed. □

11.8.3 Additional Lemmas and Examples

Lemma 11.52 (Convex Cone)**.** *If $A \subseteq C(\mathcal{Y})$ or $A \subseteq \mathrm{ca}(\mathcal{Y})$ is convex, then $\mathbb{R}_{\geq 0}A$ is a convex cone.*

Proof. Note, $\mathbb{R}_{\geq 0}A$ is closed under positive scalar multiplication. It remains to show that for $r_1 a_1, r_2 a_2 \in \mathbb{R}_{\geq 0}A$ the sum $r_1 a_1 + r_2 a_2 \in \mathbb{R}_{\geq 0}A$. To this end, we choose $r' \coloneqq r_1 + r_2 \in \mathbb{R}_{\geq 0}$ and $a' \coloneqq \frac{r_1}{r_1+r_2} a_1 + \frac{r_2}{r_1+r_2} a_2 \in A$ by convexity. We note $r'a' \in \mathbb{R}_{\geq 0}A$ and $r'p' = r_1 a_1 + r_2 a_2$. □

Lemma 11.53 (Vacuous Forecasts Only Make Non-Positive Gambles Available)**.** *Let $\mathcal{P} = \Delta(\mathcal{Y})$. Then,*

$$\{g \in C(\mathcal{Y}) \colon \sup_{\phi \in \mathcal{P}} \mathbb{E}_\phi[g] \leq 0\} = C(\mathcal{Y})_{\leq 0}.$$

Proof. Note that $\mathcal{P}$ is credal, i.e., $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed and convex. Theorem 11.22 and Proposition P5 give

$$\{g \in C(\mathcal{Y}) \colon \sup_{\phi \in \mathcal{P}} \mathbb{E}_\phi[g] \leq 0\} = \mathcal{G}_{\mathcal{P}} = (\mathbb{R}_{\geq 0}\mathcal{P})^\circ = (\mathrm{ca}(\mathcal{Y})_{\geq 0})^\circ = C(\mathcal{Y})_{\leq 0}.$$

□

Lemma 11.54. *Let $f \in C(\mathcal{Y})$. If there exists $\phi \in \Delta(\mathcal{Y})$ such that $\langle \phi, f \rangle \leq 0$, then* $\inf f \leq 0$.

Proof. Let us give a proof by contraposition. Suppose that $\inf f > 0$. Then, for all $\phi \in \Delta(\mathcal{Y})$, $\langle \phi, \alpha f \rangle \leq 0$, because $\phi(A) \geq 0$ for all $A \in \Sigma(\mathcal{Y})$. □

Example 11.55 (Dominance of Single Gamble by Scaleability)**.** *Consider an evaluation protocol* $(\mathcal{Y}, \Gamma, T, v, \{G\}, y)$ *with* $K(G) = \frac{1}{|T|}\sum_{t\in T} G_t(y_t)$*. Pick an arbitrary* $\tilde{t} \in T$ *and change the gamble* $G'_{\tilde{t}} = \alpha G_{\tilde{t}}$ *for* $\alpha \in [0, \infty)$*. The updated capital is,*

$$K(G') = \frac{1}{|T|}\sum_{t\in T} G_t(y_t) + \frac{1}{|T|}(\alpha - 1)G_{\tilde{t}}(y_{\tilde{t}}),$$

hence,

$$\frac{K(G')}{\alpha} = \frac{1}{|T|\alpha}\sum_{t\in T} G_t(y_t) + \frac{\alpha - 1}{|T|\alpha}G_{\tilde{t}}(y_{\tilde{t}}) \overset{\alpha\to\infty}{\longrightarrow} G_{\tilde{t}}(y_{\tilde{t}}).$$

11.8.4 Availability Criterion for Imprecise Forecasts

The problem of "What is a good *precise* forecast?" has been part of debates for decades [Schervish, 1989, Schervish et al., 2009, Gneiting and Raftery, 2007, Gneiting, 2011, Zhao and Ermon, 2021, Zhao et al., 2021]. However, the more general, "what are "good" *imprecise* forecasts?", regains focus just recently [Zhao and Ermon, 2021, Gupta and Ramdas, 2022, Konek, 2023, Verma et al., 2025, Fröhlich and Williamson, 2024, Singh et al., 2025][39]. The question's answer has not been settled yet. We hope our work can contribute by clarifying the problem setup.

One of the central problems in evaluating imprecise forecasts is the fine balance of (de-)incentivizing imprecision. A vacuous forecast, i.e., the totality of all probability distributions, is, for example, a forecast in which a fair player cannot increase his capital beyond 0. In this paper, we circumvent the "problem of imprecision" by requiring the forecaster to make property-based forecasts. More specifically, in our approach, we focus on the evaluation of imprecise forecasts which are level sets of elicitable (respectively identifiable) properties.

[39] Arguably, Schervish et al. [2009] started this project during the search for a generalization of de Finetti's coherence type II. The first generalization of de Finetti's coherence type I already let to the development of the field of imprecise probability [Walley, 1991, Williams, 2007].

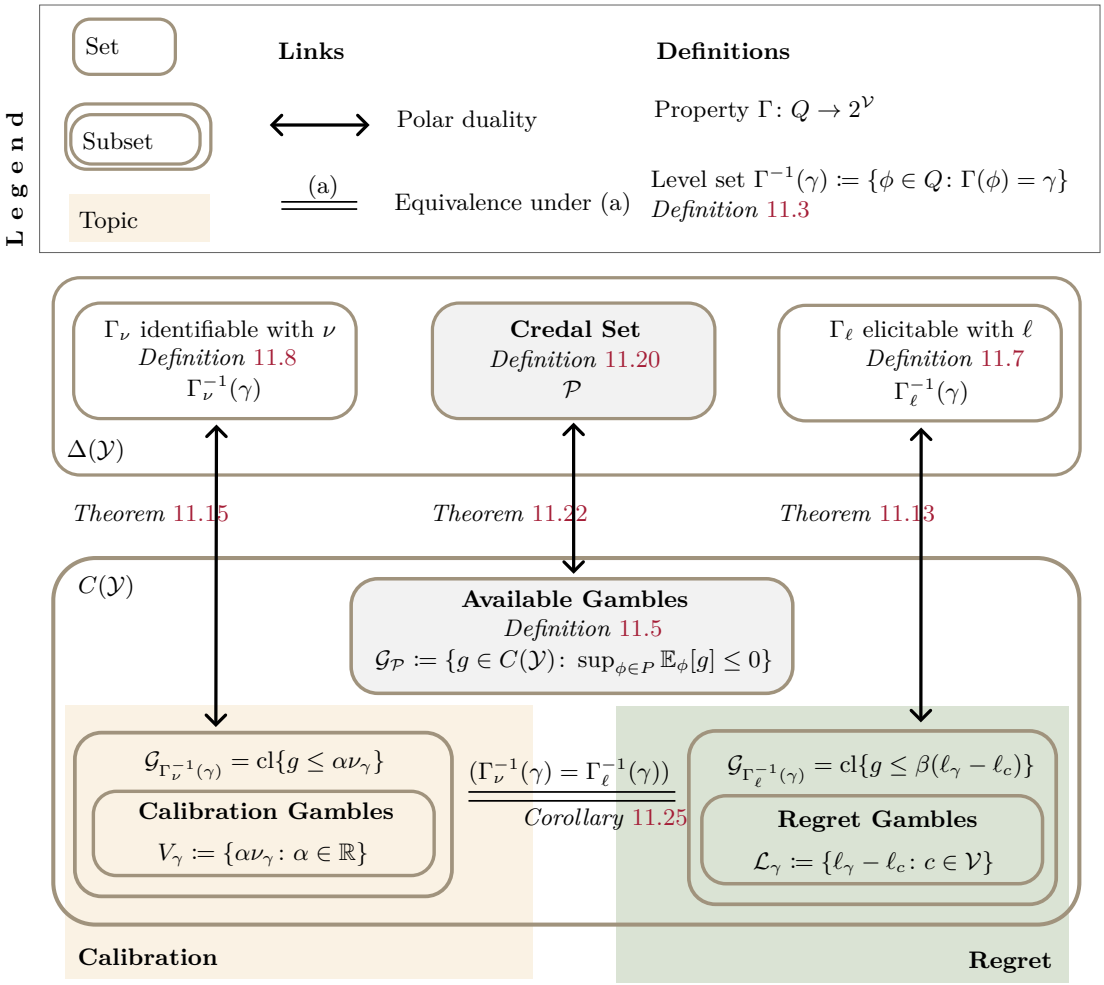


Figure 11.1: *Graphical summary of the results in Section 11.3.4.* The mathematical notations are explained in Table 11.3. The central element is the polar duality between closed, convex sets of probability distributions, i.e., credal sets, and convex cones of gambles, i.e., available gambles (highlighted in gray). We show that the set of available gambles, i.e., unfavorable gambles for a gambler, with respect to a forecasting set induced by an identifiable (respectively elicitable) property is equal to the set of gambles bounded above by calibration gambles (respectively scaled regret gambles). For example, the forecast of the set of probability distributions which share the mean $M \in \mathbb{R}$, are tested by the set of gambles which are upper bounded by $\alpha(y - M)$ for some $\alpha \in \mathbb{R}$, respectively upper bounded by $\beta(y - M)^2 - \beta(y - C)^2$ for some $\beta \geq 0$ and $C \in \mathbb{R}$. Roughly speaking, calibration and scaled regret are the most stringent forms of testing a forecast.

Set of probability distributions on $\mathcal{Y}$	$\Delta(\mathcal{Y}) \coloneqq \{\phi \in \mathrm{ca}(\mathcal{Y})\colon \phi \geq 0, \phi(\mathcal{Y}) = 1\}$
Set of continuous real-valued functions on $\mathcal{Y}$	$C(\mathcal{Y})$
Forecasting set	$\mathcal{P} \subseteq \Delta(\mathcal{Y})$
Domain of property	$Q \subseteq \Delta(\mathcal{Y})$, $\sigma(\mathrm{ca}(\mathcal{Y}), C(\mathcal{Y}))$-closed and convex
Property of distribution	$\Gamma\colon Q \to 2^{\mathcal{V}}$
Level set of property	$\Gamma^{-1}(\gamma) \coloneqq \{\phi \in Q\colon \Gamma(\phi) = \gamma\}$, Definition 11.3
Consistent scoring function for property Γ	$\ell\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ s.t. $\Gamma(\phi) = \arg\min_{c \in \mathcal{V}} \mathbb{E}_\phi[\ell_c]$, short $\ell_\gamma\colon y \mapsto \ell(\gamma, y)$, Definition 11.7
Identification function for property Γ	$\nu\colon \mathcal{Y} \times \mathcal{V} \to \mathbb{R}$ s.t. $\mathbb{E}_\phi[\nu_\gamma] = 0 \Leftrightarrow \Gamma(\phi) = \gamma$, short $\nu_\gamma\colon y \mapsto \nu(\gamma, y)$, Definition 11.8
Credal set	closed and convex $\mathcal{P} \subseteq \Delta(\mathcal{Y})$, Definition 11.20
Forecaster aligned to property Γ	v, for every $t \in T \coloneqq \{1, \ldots, n\}$, $v_t \in \mathcal{V}$, Definition 11.2
Set of gamblers with index set I	$G^I \coloneqq \{G^i\colon i \in I\}$, Definition 11.2
ith gambler	G^i for $i \in I$
Gamble at instance t of ith gambler	$G^i_t \in C(\mathcal{Y})$
Generic gamble	$g \in C(\mathcal{Y})$
Nature	y, for every $t \in T \coloneqq \{1, \ldots, n\}$, $y_t \in \mathcal{Y}$, Definition 11.2
Capital of ith gambler G^i	$K(G^i)$, Definition 11.2
Available gamble	$g \in C(\mathcal{Y})$ such that $\sup_{\phi \in \Gamma^{-1}(v)} \mathbb{E}_\phi[g] \leq 0$, Definition 11.5
Set of available gambles	$\mathcal{G}_{\Gamma^{-1}(v)}$, Definition 11.5 and $\mathcal{G}_{\mathcal{P}}$, Theorem 11.18
Offer	$\mathcal{G} \subseteq C(\mathcal{Y})$ additive, positive homogeneous coherent, default available and closed, Definition 11.17
(Unscaled) regret gambles	$\mathcal{L}_\gamma \coloneqq \{\ell_\gamma - \ell_c\colon c \in \mathcal{V}\}$, Equation (11.3)
Scaled regret gambles	$\mathcal{S}_\gamma \coloneqq \{\beta(\ell_\gamma - \ell_c)\colon \beta \in \mathbb{R}_{\geq 0}, c \in \mathcal{V}\}$, Equation (11.5)
Calibration gambles	$V_\gamma \coloneqq \{\alpha\nu_\gamma\colon \alpha \in \mathbb{R}\}$, Equation (11.4)
Superprediction set for scoring function ℓ	$\mathrm{spr}(\ell) \coloneqq \{g \in C(\mathcal{Y})\colon \exists c \in \mathcal{V}, \ell_c \leq g\}$, Definition 11.12

Table 11.3: Summary of important definitions.

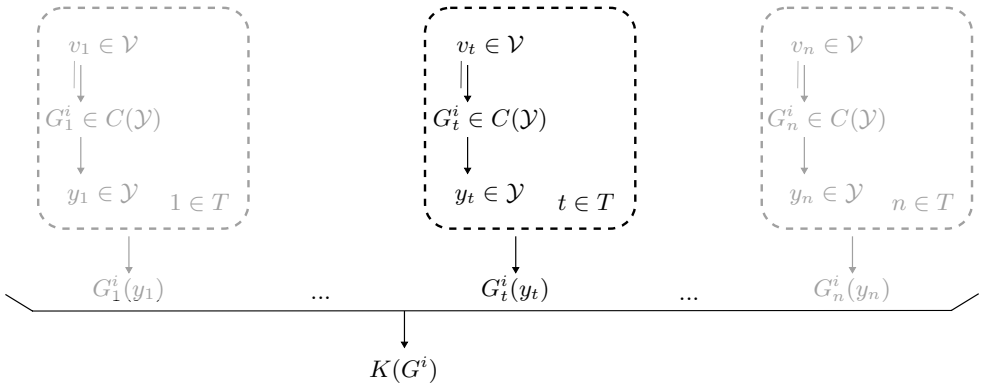


Figure 11.2: Evaluation Protocol following Definition 11.2 – The Game is played in (potentially concurrent) rounds, for every $t \in T$. In a single round, first Forecaster provides a forecast aligned to a property Γ, then the ith gambler reacts with a gamble, and finally Nature reveals an outcome. The realized values of the gambles are aggregated along the axis of all played rounds.

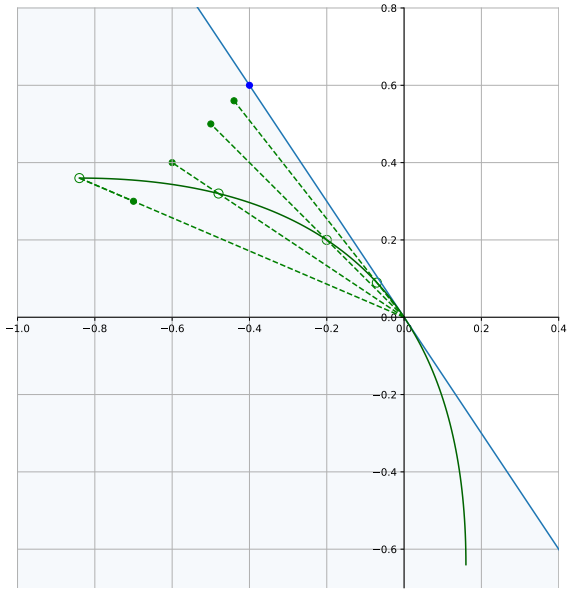


Figure 11.3: Available Gambles in the setup as described in Example 11.14 – We fix Forecaster to $v = 0.4$. The set of available gambles is shaded in light blue. The set of calibration gambles is given by the blue line. An exemplary available gamble $G = (-0.4, 0.6)$ is marked as a blue dot. The set of regret gambles for fixed $\beta = 1$ is drawn in green. An approximation of G via rescaled regret gambles ℓ_ϵ as in Equation (11.7) is shown in green.

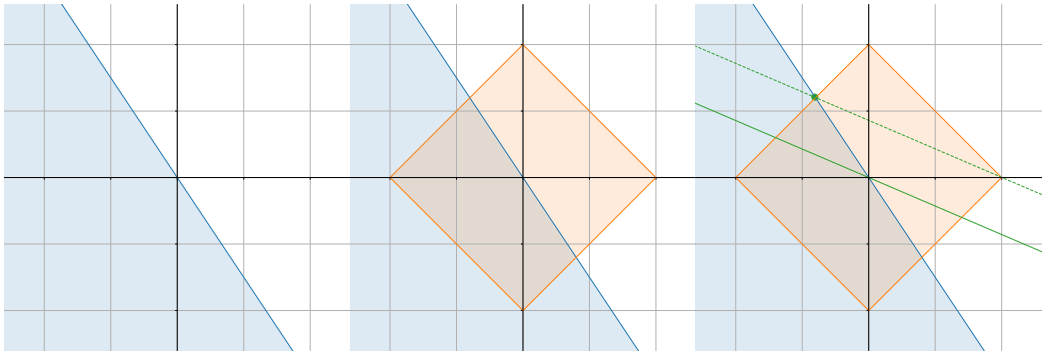


Figure 11.4: Illustration of a fair, restricted and rational gambler. – (a) Exemplary set of available gambles for a two-element outcome set (blue). (b) Restriction to norm ball $\mathbb{B}_1$ (orange). (c) Optimal choice of gamble (green dot) with respect to a belief (green line).

Part IV

More Facets

12

Individual, Randomness and Predictiveness

I elaborated on four facets of *statistical* forecast felicity in the previous chapters. It is not the scope, and I believe it cannot be the scope, of this work to fully flesh out an exhaustive list of statistical felicity conditions, their relationships and their intrinsic relative and contextual choices.

Nevertheless, I want to further expand from the four facets to two more. These additional facets are central for felicitous, statistical forecasting but have not received the attention they deserve. Both this and the following chapter can be regarded as summarizing open arguments or thoughts. They outline the path for potential future work.

12.1 Choice of Individual – Unit in Aggregates

The *aggregate* played a major role in the choices made explicit so far (Chapter 6 and Chapter 9). In the context of a *statistical forecast*, I defined an aggregate as the collection of *individuals* (Section 3.2). Hence, another more atomic question is: what (respectively who) is the *individual* about whose attributes the forecast is made?

An individual, as emphasized, is not necessarily to be understood as an individual person. In the rain forecasting example, the individual was one day. It would also be reasonable to consider the individual to be one hour, so "Chance of rain between 10 am and 11 am: 60%", or one week "Chance of rain this week: 60%". Again, we seek for felicity of this choice. An hourly rain forecast is helpful for a cyclist to know when to leave home and when to return. A weekly rain forecast is helpful for a farmer to know when to mow the lawn. The universally right individual does not exist.

An illustrative real-world forecasting scenario towards the felicitous choice of an individual is [Perdomo et al., 2025]. In this study the authors investigate the policy success of drop-out

early warning systems for public school students. The predictive systems provide drop-out risks on the level of students. The authors observe that a predictive system which provide overall drop-out risks on the level of schools would likely be sufficient to achieve comparable policy success.

The Individualization Narrative The narrative of "individualized forecasts" is part of the success story of machine learning. Apparently, machine learning methods enable to provide forecasts which are tailored to single individuals most often persons (cf. [Suriyakumar et al., 2023]). That promise is tightly related to personalized insurance [Fröhlich and Williamson, 2024]. In this section, I question the depth of the individualization.

First, "individualized forecasts" presume the existence of an ultimate individual. If, as in a social forecasting scenarios, we understand an individual as an individual person, then it might seem natural to assume that there exist an atomic individual, the person, which cannot be dissected further. However, it seems reasonable to consider the individual "You in the morning" and "You in the evening" to be two different individuals for whom different forecasts are desired. Hence, I doubt that we can reasonably presume the existence of universal atoms in forecasting scenarios.

Second, what is the usefulness of an "individualized forecast" when the forecast itself is statistical, which arguably most machine learning forecasts are. A statistical forecasts necessarily needs to refer to an aggregate. An individualized Type A2I forecast aims for an unreachable ideal, the singleton aggregate. Type I2A forecasts allow to distinguish between individuals even within groups, hence to be *individualized* [Dawid, 2017]. But this individualization is a hollow promise, as its justification, the evaluation, demands an aggregate again [Wang et al., 2024, Recht, 2025].

The individual statistical forecast is justified because of the adequate performance of the forecaster on an overarching aggregate. However, whether the individual profits from the forecast quality on the aggregate depends on the context. In [Suriyakumar et al., 2023] the authors observe that the individualization of forecasts can lead to worse forecasting performance on certain subgroups. Their findings can be understood as a type of Simpson's paradox on the level of forecasting performance. Their results challenge the narrative "individualized forecasts". Concluding, it remains the difficulty of a felicitous choice of individual.

Aggregate and Individual The choice of individual strongly interacts with the choice of aggregate. What is an aggregate when defining an individual in one scenario could be an individual in a different scenario [Latour, 2012]. The relationship of the individual and aggregate within forecast felicity remains underexplored in this thesis, but is of significant

interest [Tarde, 2012]. A more detailed elaboration of the relationship between the choice of individual, the choice of aggregate and the statisticalness of the forecasts are left for future work. The operationalization of individual as *individualization*, along the lines of Gilbert Simondon, e.g., [Simondon, 1992, Voss, 2020], could provide a helpful starter.

12.2 Randomness and Predictiveness – Two Sides of the Same Coin

In Chapter 7, I already guided us through a definition of randomness and its relationship to forecast felicity. In this section, I'll argue that the evaluation of forecasts implicitly defines a notion of randomness. Hence, Type I2A forecasts implicitly guarantee the randomness of the forecasted attributes, which then resolves the demand for a felicitous choice of randomness in the felicitous choice of an evaluation. Parts of this section have been included in a first version of Chapter 11.

A Hand-Wavy Intuition Let us start with a hand-wavy intuition why "good" forecasts on a aggregate of realized attributes is tantamount to "random" attributes with respect to an aggregate of forecasts. Caution, I'll throw around not-further-discussed and vague concepts. A forecasting method which "performs well" essentially provides "all the information" about the yet to be realized attributes. Hence, any more detailed view on the attributes will simply result in observing "noise". The "signal" is captured by the forecast. The attributes seem to be "random" with respect to the forecasts.

The duality switch I turn is, instead of asking whether the forecasts match the realized attributes, I ask whether the realized attributes match the forecasts. At first sight this perspective seems odd, but in fact it is intrinsic to an entire field of research: algorithmic randomness. Realized attributes which match the forecasts are random with respect to the forecasts. This observation is not new, [Vovk and Shen, 2010, Vovk, 2020, Dwork et al., 2023], but receives little attention from the machine learning community.

Short History of Algorithmic Randomness Let's start from the beginning. Mathematical literature on randomness started with the groundbreaking, frequential axiomatization of probability by von Mises [Von Mises, 1919]. Instead of using randomness as an ascription to a source of outcomes, he tried to capture whether a realized sequence of outcomes is random or not. This distinction is sometimes called "process" versus "product" randomness [Eagle, 2021].

In a nutshell, von Mises' random sequences (called "collective" by Von Mises [1919]) are defined as sequences for which the "law of the excluded gambling strategy" holds [Von Mises, 1919]. Remember that a gambling strategy is defined as a selection rule by von Mises.

If every selection rule strategy would be excluded, then no non-trivial sequence would be random (Section 7), thus the amount of selection rules, i.e., the abilities of the gambler, have to be restricted. The arguably standard approach faces this definitional question via tools of computation. For instance, the gambler has access to all "computable" gambling strategies, i.e., all computable selection rules [Church, 1940]. Hence, randomness and computability are opposed. This idea culminated in the theory of *algorithmic randomness*, e.g., [Martin-Löf, 1966, Uspenskii et al., 1990, Bienvenu et al., 2009, Eagle, 2021].

For the sake of simplicity, scholarship usually analyzes the randomness of infinite sequences of 0's and 1's, i.e., idealized infinite aggregates with realized binary attributes.[1] Further, the problem setup asks for a *forecasting system* with respect to which the sequence is called random or not.[2] A forecasting systems is a method which sequentially produces forecasts for every next individual.

Definitions of *algorithmic randomness* try to capture which sequences under the forecasting systems are considered random or not. Roughly stated, algorithmic randomness sorts out specific "computably structured" sequences such as e.g., 01010101010 . . .. The arguably standard notion of randomness in this literature is the so-called Martin-Löf randomness [Uspenskii et al., 1990]. This notion, introduced by Per Martin-Löf, defines a sequence to be random if it passes all computable, statistical tests with respect to a forecasting system [Martin-Löf, 1966]. Interestingly, for our purposes, this notion can be expressed equivalently in terms of gambling strategies and evaluation criteria [Vovk and Watkins, 1998, Vovk and Shen, 2010, De Cooman and De Bock, 2021].

Three Notions of Randomness with Evaluations Let me roughly introduce three notions of randomness which share that they are expressed in terms of evaluation criteria for forecasts.

In [Nobel, 2004] Andrew Nobel introduced *memoryless* sequences.[3] For a chosen loss function, a sequence is *memoryless* if no finite-horizon Markov predictor, i.e., a function which is allowed to look back on finitely many elements of the past sequence, performs better than a constant forecasting system.[4] In other words, if the asymptotic regret of any finite-Markov predictor against the best constant forecasting system is larger or equal to zero, then the sequence is memoryless. Hence, the definition of memoryless is built upon the evaluation a constant forecasting system in comparison to finite-Markov predictors. If a

[1] More general treatments can be found for instance in [Gács, 2005] or [De Cooman and De Bock, 2021].

[2] In fact, most often the definitions need a probability model. Every probability model can, under some restrictions, be understood as a forecasting system, cf. [Dawid and Vovk, 1999, §3.1.1], [Lehrer, 2001], [Shafer and Vovk, 2019, p. 193] or Shafer and Vovk [2019, p. 177]).

[3] The work got extended in [Frongillo and Nobel, 2017] and [Frongillo and Nobel, 2020].

[4] Different to the randomness definitions below, the forecasting system is required to be constant for this definition.

constant forecasting system performs "well", then the realized attributes are *memoryless*.

The definition of *game-randomness* [Vovk and Shen, 2010] requires the mathematical tools of prequential probability [Dawid, 1984, Dawid and Vovk, 1999, Shafer, 2021], lower semicomputability [Martin-Löf, 1966, Gács, 2005, Vovk and Shen, 2010] and superfarthingales[5]. In the author's terms, a sequence is *game-random*, if every lower semicomputable superfarthingale is bounded on the sequence.In simpler words and neglecting the technicalities, a sequence is *game-random*, if every *computable* and *fair* gambler (a) gambles on every next element of the sequence, (b) adds up the realized values of the gambles and (c) does not get infinitely rich. Importantly, *computability* is meant to capture the notion of lower semicomputability. *Fairness* is defined with respect to the forecasting system as in the protocol of Section 11. Hence, the forecasts are evaluated by the gambles. Only if the forecasts are "good", the sequence, i.e., the realized attributes, is "random".

A formally related notion is *predictive-complexity-randomness*. In a series of work starting with [Vovk and Watkins, 1998], Vladimir Vovk and collaborators elaborated a definition of predictive complexity which captures the unpredictability of a sequence.[6] Predictive complexity is roughly the accumulated loss the best computable expert would incur on a sequence. A *predictive-complexity-random* sequence with respect to a forecasting system is a sequence on which the forecasts perform at least as good as the best computable expert. Hence, as for memoryless sequences, randomness is captured by a regret-type definition. "Good" forecasts make the realized attributes "random".

Note that predictive complexity with respect to the logarithmic loss, sometimes called cross-entropy-loss, is equivalent to Levin's semimeasures [Zvonkin and Levin, 1970] and hence its predictive-complexity-randomness is equivalent to Martin-Löf-randomness Vovk and Watkins [1998]. Furthermore, game-randomness is under mild assumptions as well equivalent to Martin-Löf randomness [Vovk and Shen, 2010, Corollary 1].

In conclusion, randomness in the given examples is expressed through the evaluation of a forecasting system on an infinite sequence of realized attributes. All notions share that if the forecasting system is evaluated to perform well, the sequence is called random. This justifies our earlier statement: it is mathematically equivalent to state that forecasts fit realized attributes or realized attributes fit forecasts, but the semantic changes. Forecasts which match realized attributes are "good". Realized attributes which match forecasts are "random" with respect to those forecasts.

A critical reader might question the implicit stretch from infinite sequences to finite aggregates in the argument above. Practical scenarios are bound to observe only finitely many forecasts and finitely many realized attributes.

[5] Superfarthingales are the prequential analogue of supermartingales.

[6] For a nice summary see the PhD-thesis by Yuri Kalnishkan [Kalnishkan, 2002].

There are at least three responses to sweep this concern off the table. (a) There exist attempts to formalize randomness on finite aggregates [Putnam, 1990/2011] and [Kolmogorov, 1965]. However, their relationship to evaluation is yet to be explored. (b) One can interpret the actual realized attributes and forecasts as the start of an idealized sequence. Then, predictiveness on a finite aggregate would by assumption imply randomness on an infinite sequence. (c) I am convinced that the conceptual idea of randomness is not bound to infinite sequences. It is an artifact of the formalization that randomness is defined by a subset of the uncountable set of sequences. In principle, other kinds of aggregates, like functions on the real numbers [Kac, 1982, Leśkow and Napolitano, 2006, Napolitano and Gardner, 2022] or nets [Ivanenko, 2010] would be imaginable to be the subject of a definition of randomness. Analogously, a randomness definition for finite aggregates would then be a rule which excludes certain aggregates to be non-random. It is not a far stretch to simply define random aggregates as the ones on which a forecasting system performs well.

Accepting the argument, it follows that the predictiveness of a forecaster implies the randomness of the realized attributes. Hence, a felicitous choice of evaluation would imply a definition of randomness. This randomness, in turn, suffices as argument for a felicitous, statistical forecast. The choice of randomness serves as inherently felicitous modeling choice. It matches the felicitous evaluation. Different to forecasts of Type A2I, Type I2A forecasts seem to not require an elaborate choice of randomness.

13

Beyond Felicity Conditions about the Statistical Nature

The exemplary list of felicity conditions presented in Paragraph 4.2 included conditions reaching beyond the statistical nature of forecasts. This is a natural consequence of understanding a forecast as a speech act. Austin's felicity conditions, e.g., [Austin, 1975, page 15 and 18], for speech acts share this generality.

Hitherto, I focused on statistical felicity conditions. In this section, I want to risk a glimpse beyond statistical aspects of forecasts. I picked two topics, communication and performativity. Both are central to felicitous forecasting. Both have partially been treated in a formal, mathematical way. And, both are full of unanswered, but interesting questions.

13.1 Communication – Forecaster and Receiver

Forecasts as speech acts are uttered in a context of receivers. Forecasts hence are communicated with at least two parties involved, the forecast and the receiver.

Understandability and Understanding One of the infelicities Austin is alluding to are "misunderstandings" [Austin, 1975, page 22]. Hence, a prerequisite for a felicitous forecasts is that it is understandable to the receivers and understood by them in accordance with the forecaster's aims and purposes. Otherwise the communication of the forecast fails.

I have not paid attention to this relatively fundamental aspect, which is practically important as the study by [Gigerenzer et al., 2005] showed. A forecast like "Chance of rain today: 60%"[1] is understood in a variety of ways, sometimes matching the forecaster's meaning

[1] In fact, they studied the question of what people understood under "30% chance of rain tomorrow".

sometimes not.

A reason for these misunderstandings was and is the multitude of meanings of such rain forecasts given by the forecasting agencies (cf. Section 5.1 and [Stewart et al., 2016]). As a worrying or amusing side note, even experts of the professional atmospheric science community hold different understandings of the probabilistic rain forecasts, and they do so with great conviction [Stewart et al., 2016].

The understandability and the understanding of a forecast with respect to the intended receivers of the forecast is part of the success of it. It drives the receivers to change beliefs, take actions, or react more generally.

Decision Schemes by Receivers A standard way to avoid the understanding problem is to assume a certain decision scheme by the receiver. The seminal work by von Neumann and Morgenstern [1953] laid the foundation of decision theory and provided a formal treatment of such decision makers. The expected utility maximizer is an idealized agent which is governed by a utility function which specifies the utility for a yet unobserved attribute and a possible decision. Such an agent reacts to a probabilistic forecast as if it were the true probability distribution by taking the decision which maximizes the expected utility. Von Neumann and Morgenstern furthermore showed that agents which have a structured preference relationship on different probability distributions over the outcomes actually behave as if they would be governed by such a utility function [von Neumann and Morgenstern, 1953].

This model of response to a forecast essentially underpins most of the evaluation structures which we have seen before. Loss functions, necessary for notions of regret, are negative utilities which determine the optimal forecasting model. Calibration is (partially) rationalized by the argument that any decision maker profits, in terms of no swap regret, from calibrated forecasts (cf. Section 10). Note that the link between decision theory and statistics has a history dating back far beyond machine learning forecasting systems, e.g., [Wald, 1950].

An interesting tweak to the assumption of a certain response behavior is the introduction of competing interests between forecaster and receiver. [Jain and Perchet, 2024, Tang et al., 2025] and [Feng and Tang, 2025] are formal studies of how a receiver can be nudged to certain decisions while the forecasts still fulfills a certain evaluation criterion. In particular, the response behavior is rationalized as it is, under certain conditions, for the receiver's best, i.e., minimizing a certain swap regret (Definition 10.26). This information design following the idea of Bayes persuasion [Kamenica and Gentzkow, 2011] exemplarily shows that truth is not the ultimate goal of the forecast here, but the felicity through the eyes of the forecaster.

From Decision-Support to Discourse-Support In a forthcoming work by Sebastian Zezulka and Konstantin Genin [Zezulka and Genin, 2025], they argue that we should not take for granted certain response behavior. Forecasts need not only support the decisions to take by the receivers, but support the discourse among the receivers to improve, in their case, policy decisions. They define discourse-supportiveness based on Habermas' "communicative rationality" [Habermas, 1998].

Their argument relies on the assumption that the forecasts are *performative* in the sense that they change the subject of the forecast (Section 13.2). In a nutshell, suppose that the receivers react in some general, but pre-specified, way to a forecast by a decision. Suppose that the receiver is subject of the forecast. Then, optimal forecasting, i.e., decision support, intertwines epistemic and pragmatic interests. They suggest to disentangle them via counterfactual, causal estimation, which they claim to be discourse-supportive.

Interaction between Forecaster and Receiver Finally, the communication of forecasts is not bound to be a one-way street. In fact, often forecaster and receiver interact in more complicated ways, e.g., to form a forecast [Charusaie et al., 2022, Charusaie and Samadi, 2024, Collina et al., 2025b,a].

Open Questions In all those aspects of communicating forecasts there wait several interesting question yet to be answered:

Receiver Models for Statistical Forecasts What are other models of receivers' reactions to forecasts which respect the statisticalness of the forecast not through uncertainty expressed by a probability distribution but because of their relationship to an aggregate? What is the role of individual and aggregate for the reactions in those models?

Formalization of Discourse-Supportiveness How can discourse-supportiveness be formalized beyond counterfactual forecasting?

Competing Interests To which extent can forecasts simultaneously serve competing interests of forecasters and receivers?

Agreement and Understandability How does "agreement" in the interaction between forecaster and receiver relate to the understanding and understandability of forecasts?

13.2 Performativity – Beyond the Concept in Machine Learning

Let me know dive a deeper into the scenario sketched before. Machine learning applied in social context, e.g., traffic prediction, credit default risk, leads to scenarios in which the

subjects of forecasts react to the deployment of the method, hence change the forecasting problem and in particular the data distribution on which the machine learning model is trained on.

Performativity in Machine Learning This phenomenon has been studied as *performativity* in machine learning [Perdomo et al., 2020].[2]

A special subtype of performativity, in the sense of machine learning, is the response of the subjects of forecasts to the forecasts [Kim and Perdomo, 2023]. This is a restricted setting. The more general model assumes that machine learning deployment changes the entire distribution of covariates and predictands. A good example of such restricted setting is the forecast about the utilization of trains such as implemented by the Deutsche Bahn [Deutsche Bahn AG, 2025]. The forecast directly influences the behavior of potential travelers and thus utilization. In contrast, the much-discussed rain forecast does not change the rain.[3]

Performativity in Speech Act Theory When earlier (Section 4.1) referring to "performatives", was I referring to the machine learning understanding? No, performativity is a natural and rich concept in speech act theory. To shortly remind the reader, Austin [1975] suggests to look at three subacts of a speech act, the locutionary, illocutionary and perlocutionary act. Interestingly, both the illocutionary and the perlocutionary act are tightly related to performativity.

The perlocutionary act is the (intentional) consequence or effect of the forecast. Performativity in machine learning is most often part of this act. The reaction of the subject of a forecast to change its attributes, e.g., "take the train", is a consequence of the forecast.

The illocutionary act of forecasting is *in* the forecasting itself. The forecast *does* warn, inform, promise, describe, approve, hence, shapes reality. This performativity is more subtle but constructs our (social) reality. To illustrate, a warning rain forecast brings to existence the warning. The world is different after the warning.

Different to the perlocutionary act, the illocutionary act in forecasting has, to the best of my knowledge, eluded mathematical formalization so far. In machine learning we are in the situation of separating between constantives and performatives. One of Austin's main goals in [Austin, 1975] was to replace this separation by a unified treatment as speech acts with all their sub-acts included. Hence, I believe that a treatment of the illocutionary act in the formal words and language of machine learning would indeed be fruitful.

[2] Note that for research on economics or sociology, *performativity* is not a new concept, e.g., [Luhmann, 1984]. For instance, MacKenzie and Millo [2003] observed that the Black-Scholes-Merton option pricing theory [Black and Scholes, 1973, Merton, 1973] indeed reified.

[3] At least if we disregard the fact that people's behavior can, in principle, change the weather conditions.

Furthermore, an extension of the project "forecasts as speech acts" can be "machine learning as act". This leads to the rich and developing question: how does machine learning make (social) reality? For instance, through construction of evaluation schemes [Wang et al., 2024, Recht, 2025], through definitions of variables [Hu and Kohler-Hausmann, 2020, Fröhlich and Williamson, 2024], through use in everyday life [Bareither, 2024], and many more ways not mentioned. I leave this to future work.

Part V

Epilogue

14

A Practical Conclusion

This thesis introduces *forecast felicity* and focuses on *statistical* forecasts. In particular, I attempt to resolve four problems, what does the forecast mean, how to determine whether the forecast is "good", what is the argument that the forecast is a quantity, what is the argument that the forecast is statistical. I mainly focus on four facets captured by the four projects which form the core of this thesis. The facets do not map one-to-one to the problems. In their entirety they sketch a picture of statistical forecast felicity which I want to summarize and translate into the following "4F-Questionnaire" (Table 14.1).

The questionnaire serves as a practical tool that results from the questions examined. It makes explicit internal choices required for statistical forecasting. Furthermore, the questionnaire asks for an argument why the choice is felicitous. Hence, it supports agencies who provide forecasts to check the relevant aspects of their forecasts. A priori, it should help to produce felicitous forecast i.e., forecasts which meet their aim and purpose. A posteriori, I should facilitate the analysis why a forecast has not been felicitous.

To exemplify, consider, for a last time, the forecast "Chance of rain today: 60%" and let me iterate through the questionnaire. I presume that the forecast is used by farmers to decide on when to mow their lawns (cf. [Deutscher Wetterdienst (DWD), 2025]). The forecast *supports* the mowing decision. The choices made and the arguments presented are made up. The arguments might require further extension or corroboration. Note that the scope of the argument is again a choice depending on the context, e.g., what are the stakes involved?

I would be "felicitous" to see weather agencies providing *felicitous* explanations of their precipitation forecasts. Forecast felicity is more than a theoretical concept. It is a practical tool to understand, construct, analysis (un)successful forecasts. Forecast felicity is everywhere.

	What is the choice?	What is the argument for the felicity of the choice?
Statistical Forecast		
Individual	*Day.*	*Decision on mowing is made for a day.*
Aggregate	*Days with similar forecasted weather, average temperature within* $\pm 2°C$ *and pressure within* $\pm 50, \mathrm{hPa}$*, over the past* 10 *years within a* $10, \mathrm{km}$ *radius.*	*Due to climate change, older data is not reliable for future predictions. Similarity measure has shown relevance in tests for precipitation forecasting.*
Type	*Forecast Type A2I but additionally evaluated as Type I2A forecasts.*	*Additional evaluation supports conclusiveness of forecasts.*
Forecast of Type A2I		
Measurability	*In the range of* $-2°C$ *to* $31°C$ *for the forecasted weather situation, measurability is given. Otherwise not.*	*Sufficient past reference data exists for statistical claims.*
Measure	*Relative number of days, rounded to* 5%*.*	*Simple measure and rounding facilitates understanding. Mowing decision is supported.*
Randomness	*Aggregate of days is random with respect to relative humidity and/or maximum temperature.*	*Similarity defined through relative humidity and/or temperature does not change relative count.*
Forecast of Type I2A		
Internal Consistency	*Given through relative count.*	*Mowing decision is supported as this is usually taken on a subjective and varying threshold on the chance of rain.*
Evaluation metric	*Binary distribution calibration.*	*Calibration guarantees that consequences of mowing decision can be estimated.*
Randomness.	*By calibration and see above.*	*Resolved because predictions are calibrated and randomness with respect to certain additional information is guaranteed above.*

Table 14.1: The 4F-Questionnairem makes internal choices required for statistical forecasting explicit. *Exemplary forecast, "Chance of rain today:* 60%*".*

References

Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Conference on Learning Theory*, pages 27–46. PMLR, 2011.

Kenneth S. Abraham. Efficiency and fairness in insurance risk classification. *Virginia Law Review*, 71(3):403–451, 1985.

Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory*, pages 2704–2729. PMLR, 2022.

Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 3rd edition, 2006.

Andrew Altman. Discrimination. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, (Summer 2020) edition, 2020.

Klaus Ambos-Spies, Elvira Mayordomo, Yongge Wang, and Xizhong Zheng. Resource-bounded balanced genericity, stochasticity and weak randomness. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 61–74. Springer, 1996.

Philip Amortila, Dylan J. Foster, Nan Jiang, Akshay Krishnamurthy, and Zakaria Mhammedi. Reinforcement learning under latent dynamics: toward statistical and algorithmic modularity. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 133007–133091, 2024.

Anonymous. Hypergraph. *Wikipedia*, 2023a. URL `https://en.wikipedia.org/wiki/Hypergraph`.

Anonymous. Intersectionality. *Wikipedia*, 2023b. URL `https://en.wikipedia.org/wiki/Intersectionality`.

Anonymous author. St. Petersburg paradox. `https://en.wikipedia.org/wiki/St._Petersburg_paradox`, 2025. Wikipedia, accessed 27 August 2025.

Kwame Anthony Appiah. *As if: Idealization and ideals*. Harvard University Press, 2017.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI magazine*, 36(1):15–24, 2015.

Kenneth J. Arrow. Uncertainty and the welfare economics of medical care. In *Uncertainty in economics*, pages 345–375. Elsevier, 1978.

Thomas Augustin, Frank P.A. Coolen, Gert De Cooman, and Matthias C.M. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

John Langshaw Austin. *How to do things with words*. Oxford university press, 1975.

Maria Baghramian and J. Adam Carter. Relativism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2025 edition, 2025.

Prasanta S. Bandyopadhyay and Malcolm R. Forster. Philosophy of statistics: An introduction. In Prasanta S. Bandyopadhyay and Malcolm R. Forster, editors, *Philosophy of Statistics*, volume 7. Elsevier, 2011.

Christoph Bareither. Cultures of artificial intelligence. AI assemblages and the transformations of everyday life. *Zeitschrift für Empirische Kulturwissenschaft*, 120:I–XIX, 2024.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Mark Battersby. The rhetoric of numbers: Statistical inference as argumentation. In *OSSA Conference Archive*, 2003.

BBC. About BBC Weather. `https://www.bbc.com/weather/about/17185651`. Accessed: 9 August 2025.

Michael Ian Beardmore. *Ancient weather signs: texts, science and tradition*. PhD thesis, The University of St. Andrews, 2013.

Alessio Benavoli, Alessandro Facchini, and Marco Zaffalon. Quantum mechanics: The Bayesian theory generalized to the space of Hermitian matrices. *Physical Review A*, 94: 042106, 2016.

Alessio Benavoli, Alessandro Facchini, Marco Zaffalon, and José Vicente-Pérez. A polarity theory for sets of desirable gambles. In *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pages 37–48. PMLR, 2017.

Deborah Bennett. Defining randomness. In Prasanta S. Bandyopadhyay and Malcolm R. Forster, editors, *Philosophy of Statistics*, volume 7. Elsevier, 2011.

Claude Berge. *Hypergraphs: combinatorics of finite sets*. North-Holland, 1989.

Joseph Berkovitz, Roman Frigg, and Fred Kronz. The ergodic hierarchy, randomness and Hamiltonian chaos. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 37(4):661–691, 2006.

Dan Biddle. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower Publishing, Ltd., 2006.

Laurent Bienvenu, Glenn Shafer, and Alexander Shen. On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics*, 5(1): 1–40, 2009.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.

Garrett Birkhoff. *Lattice theory*. American Mathematical Society, 1940.

Garrett Birkhoff and John von Neumann. The logic of quantum mechanics. *Annals of Mathematics*, 37:1–26, 1936.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pages 1006–1014. PMLR, 2015.

Norbert Blum. *Einführung in Formale Sprachen, Berechenbarkeit, Informations-und Lerntheorie*. Oldenbourg Wissenschaftsverlag, 2009.

Kim C. Border. Lecture notes on convex analysis and economic theory, topic 20: When are sums closed?, 2019-2020. Division of the Humanities and Social Sciences, California Institute of Technology.

Jiří Brabec. Compatibility in orthomodular posets. *Časopis pro pěstování matematiky*, 104 (2):149–153, 1979.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Percy W. Bridgman. Remarks on the present state of operationalism. *The Scientific Monthly*, 79(4):224–226, 1954.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

John Broome. Selecting people randomly. *Ethics*, 95(1):38–55, 1984.

Costantino Budroni, Adán Cabello, Otfried Gühne, Matthias Kleinmann, and Jan-Åke Larsson. Kochen-Specker contextuality. *Reviews of Modern Physics*, 94(4):045007, 2022.

John Burden, Marko Tešić, Lorenzo Pacchiardi, and José Hernández-Orallo. Paradigms of ai evaluation: Mapping goals, methodologies and culture. *arXiv preprint arXiv:2502.15620*, 2025.

Bureau of Meteorology (Australia). Right as rain: How to interpret the daily rainfall forecast. https://media.bom.gov.au/social/blog/209/right-as-rain-how-to-interpret-the-daily-rainfall-forecast/, 2018. Accessed: 9 September 2025.

Bureau of Meteorology (Australia). Forecast accuracy. https://beta.bom.gov.au/about-the-bureau/plans-performance-and-accountability/forecast-accuracy, 2025. Accessed: 12 September 2025.

Paul Busch, Teiko Heinosaari, Jussi Schultz, and Neil Stevens. Comparing the degrees of incompatibility inherent in probabilistic physical theories. *Europhysics Letters*, 103, 2012.

Sam Buss and Mia Minnes. Probabilistic algorithmic randomness. *The Journal of Symbolic Logic*, 78(2):579–601, 2013.

Jarosław Błasiok, Parikshit Gopalan, Linlin Hu, and Venkatesan Guruswami. A unifying theory of distance from calibration. *Symposium on the Theory of Computing*, 2023.

Federico Cabitza, Andrea Campagner, and Lorenzo Famiglini. Global interpretable calibration index, a new metric to estimate machine learning models' calibration. In Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 82–99. Springer International Publishing, 2022.

Armando J. Cabrera Pacheco, Rabanus Derr, and Robert C. Williamson. Aggregating algorithm and axiomatic loss aggregation. *Preprint*, 2024.

Chengyi Cai, Zesheng Ye, Lei Feng, Jianzhong Qi, and Feng Liu. Bayesian-guided label mapping for visual reprogramming. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 17656–17695, 2024.

Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Donald T. Campbell. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, 3(1):14–25, 1958.

Rudolf Carnap. *Einführung in die Philosophie der Naturwissenschaften, Hrsg. Martin Gardner*. Ullstein Materialien Corporation, 1986. Translated from the original: "Philosophical Foundations of Physics".

Thyago P. Carvalho, Fabrízzio A. A. M. N. Soares, Roberto Vita, Roberto da P. Francisco, João P. Basto, and Symone G. S. Alcalá. A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137:106024, 2019.

Arianna Casanova, Juerg Kohlas, and Marco Zaffalon. Information algebras in the theory of imprecise probabilities, an extension. *International Journal of Approximate Reasoning*, 150:311–336, 2022.

George Casella and Roger L. Berger. *Statistical Inference: Second Edition*. Duxbury Advanced Series, 2002.

Philippe Casgrain, Martin Larsson, and Johanna Ziegel. Anytime-valid sequential testing for elicitable functionals via supermartingales. *arXiv preprint arXiv:2204.05680*, 2022.

Emanuele Castano, Vincent Yzerbyt, Maria-Paolo Paladino, and Simona Sacchi. I belong, therefore, I exist: Ingroup identification, ingroup entitativity, and ingroup bias. *Personality and Social Psychology Bulletin*, 28(2):135–143, 2002.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Gregory J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the ACM (JACM)*, 13(4):547–569, 1966.

Anjan Chakravartty. Scientific Realism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2017 edition, 2017.

Hasok Chang. Operationalism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Mohammad-Amin Charusaie and Samira Samadi. A unifying post-processing framework for multi-objective learn-to-defer problems. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 23705–23755, 2024.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, pages 2972–3005. PMLR, 2022.

Junwei Chen, Teng Pan, Zhengjie Zhu, Lijue Liu, Ning Zhao, Xue Feng, Weilong Zhang, Yuesong Wu, Cuidan Cai, Xiaojin Luo, et al. A deep learning-based multimodal medical imaging model for breast cancer screening. *Scientific Reports*, 15(1):14696, 2025.

Graciela Chichilnisky. The foundations of probability with black swans. *Journal of Probability and Statistics*, 2010:1–11, 2010.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer, 2nd edition, 1988.

Alonzo Church. On the concept of a random sequence. *Bulletin of the American Mathematical Society*, 46(2):130–135, 1940.

Elizabeth R. Cole. Intersectionality and research in psychology. *American Psychologist*, 64 (3):170, 2009.

Natalie Collina, Rabanus Derr, and Aaron Roth. The value of ambiguous commitments in multi-follower games. *arXiv preprint arXiv:2409.05608*, 2024.

Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. Collaborative prediction: Tractable information aggregation via agreement. *arXiv preprint arXiv:2504.06075*, 2025a.

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1532–1543, 2025b.

Donald C. Cook. The concept of independence in accounting. In *Federal Securities Law and Accounting 1933–1970: Selected Addresses*, pages 198–222. Routledge, 2020.

Roger Cooke. The anatomy of the squizzel: the role of operational definitions in representing uncertainty. *Reliability Engineering & System Safety*, 85(1-3):313–319, 2004.

Core Writing Team. IPCC, 2023: Summary for Policymakers. In *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1–34. IPCC, Geneva, Switzerland, 2023.

D.R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.

Patrick Cronin. The authorship and sources of the peri semeion ascribed to theophrastus. *Fortenbaugh, WW & Gutas, D.(edd.) Theophrastus: His Psychological, Doxographical and Scientific Writings. London*, pages 307–345, 1992.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

Carlos María Cuadras, Josep Fortiana, and José A. Rodriguez-Lallena. *Distributions with given marginals and statistical modelling*. Springer, 2002.

Mark H. A. Davis. Verification of internal risk measure estimates. *Statistics & Risk Modeling*, 33(3-4):67–93, 2016.

A. Philip Dawid. The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

A. Philip Dawid. Statistical theory: The prequential approach (with discussion and rejoinder). *Journal of the Royal Statistical Society Ser. A*, 147:278–292, 1984.

A. Philip Dawid. Calibration-based empirical probability. *The Annals of Statistics*, 13(4): 1251–1274, 1985a.

A. Philip Dawid. Probability, symmetry and frequency. *The British journal for the philosophy of science*, 36(2):107–128, 1985b.

A. Philip Dawid. Probability, causality and the empirical world: A Bayes–de Finetti–Popper–Borel synthesis. *Statistical Science*, 19(1):44–57, 2004.

A. Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

A. Philip Dawid and Vladimir G. Vovk. Prequential probability: principles and properties. *Bernoulli*, pages 125–162, 1999.

Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017.

Luciano I. De Castro and Alain Chateauneuf. Ambiguity aversion and trade. *Economic Theory*, 48:243–273, 2011.

Gert De Cooman and Jasper De Bock. Randomness and imprecision: a discussion of recent results. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 110–121. PMLR, 2021.

Gert De Cooman and Jasper De Bock. Randomness is inherently imprecise. *International Journal of Approximate Reasoning*, 141:28–68, 2022.

Bruno De Finetti. Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 179–190, 1929.

Bruno De Finetti. *Theory of probability: A critical introductory treatment*. John Wiley & Sons, 1974/2017.

Abraham De Moivre. *The Doctrine of Chances: A Method of Calculating the Probabilities of Events in Play*. Frank Cass and Company Limited, 2nd edition, 1738/1967.

Anna De Simone and Pavel Pták. Measures on circle coarse-grained systems of sets. *Positivity*, 14(2):247–256, 2010.

Anna De Simone, Mirko Navara, and Pavel Pták. Extending states on finite concrete logics. *International Journal of Theoretical Physics*, 46(8):2046–2052, 2007.

Gerard Debreu. *Theory of value: An axiomatic analysis of economic equilibrium*. Yale University Press, 1959.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1/2):12–22, 1983.

Freddy Delbaen. Coherent risk measures on general probability spaces. In Klaus Sandmann and Philipp J. Schönbucher, editors, *Advances in finance and stochastics*, pages 1–37. Springer, 2002.

Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325 – 339, 1967.

Zhun Deng, Cynthia Dwork, and Linjun Zhang. HappyMap : A Generalized Multicalibration Method. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 41:1–41:23, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Dieter Denneberg. *Non-additive measure and integral*. Springer Science & Business Media, 1994.

Rabanus Derr and Robert C. Williamson. The set structure of precision. In *International Symposium on Imprecise Probability: Theories and Applications*, 2023a.

Rabanus Derr and Robert C. Williamson. Systems of precision: coherent probabilities on pre-Dynkin systems and coherent previsions on linear subspaces. *Entropy*, 25(9):1283, 2023b.

Rabanus Derr and Robert C. Williamson. Fairness and randomness in machine learning: Statistical independence and relativization. *The New England Journal of Statistics in Data Science*, 3:55–72, 2024.

Rabanus Derr and Robert C. Williamson. Evaluation metrics as averaged outcomes of fair gambles. *arXiv preprint arXiv:2401.14483*, 2025.

Rabanus Derr, Jessie Finocchiaro, and Robert C. Williamson. Three types of calibration using properties and their semantic and formal relationships. *Journal of Machine Learning Research*, 27(110):1–52, 2026.

David Deutsch. *The beginning of infinity: Explanations that transform the world*. Penguin, 2011.

Deutsche Bahn AG. DB Navigator. https://www.bahn.de/service/mobile/db-navigator, 2025. Accessed: 4 August 2025.

Deutscher Wetterdienst. Niederschlagswahrscheinlichkeit. https://www.dwd.de/DE/service/lexikon/Functions/glossar.html?lv2=101812&lv3=101912. Accessed: 9 August 2025.

Deutscher Wetterdienst. Aktuelle Warnungen und Prognosen auf Gemeindeebene (WarnWetter). https://www.dwd.de/DE/wetter/warnungen_gemeinden/warnWetter_node.html, 2025. Accessed: 6 August 2025.

Deutscher Wetterdienst (DWD). agrowetter Prognose: agrarmeteorologische Produkte. https://www.dwd.de/DE/leistungen/agrowetter_prognose/agroprog.html, 2025. Accessed: 22 September 2025.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 1996.

Joseph Diestel and John J. Uhl. *Vector Measures*. American Mathematical Society, 1977.

Miriam Doh, Benedikt Höltgen, Piera Riccio, and Nuria M. Oliver. Position: The categorization of race in ML is a flawed premise. In *International Conference on Machine Learning – Position Paper Track*, 2025.

Rodney G. Downey and Denis R. Hirschfeldt. *Algorithmic randomness and complexity*. Springer Science & Business Media, 2010.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference - ITCS 2012*, pages 214–226, 2012.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

Cynthia Dwork, Daniel Lee, Huijia Lin, and Pranay Tankala. From pseudorandomness to multi-group fairness and back. In *Conference on Learning Theory*, pages 3566–3614. PMLR, 2023.

Antony Eagle. Chance versus Randomness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

Elad Eban, Elad Mezuman, and Amir Globerson. Discrete Chebyshev classifiers. In *International Conference on Machine Learning*, pages 1233–1241. PMLR, 2014.

Jürgen Elstrodt. *Maß-und Integrationstheorie*. Springer, 2018.

Paul Embrechts, Tiantian Mao, Qiuqi Wang, and Ruodu Wang. Bayes risk, elicitability, and the expected shortfall. *Mathematical Finance*, 31(4):1190–1217, 2021.

Larry G. Epstein and Jiankang Zhang. Subjective probabilities on subjectively unambiguous events. *Econometrica*, 69:265–306, 2001.

European Central Bank. ECB Survey of professional forecasters for the third quarter of 2025, 2025.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689, 12 July 2024, 2024. Entered into force on 1 August 2024; progressive applicability through August 2027.

Brian S. Everitt. *The Cambridge dictionary of statistics*. Cambridge, 2nd edition, 2002.

Matthias Faes and David Moens. Recent trends in the modeling and quantification of non-probabilistic uncertainty. *Archives of Computational Methods in Engineering*, 27(3): 633–671, 2020.

Yiding Feng and Wei Tang. Persuasive calibration. *arXiv preprint arXiv:2504.03211*, 2025.

James H. Fetzer. Reichenbach, reference classes, and single case 'probabilities'. In *Hans Reichenbach: Logical Empiricist*, pages 187–219. Springer, 1979.

Terrence L. Fine. *Theories of probability: An examination of foundations*. Academic Press, 1973.

James Gordon Finlayson and Dafydd Huw Rees. Jürgen Habermas. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

Jessica Finocchiaro and Rafael Frongillo. Convex Elicitation of Continuous Properties. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Unifying lower bounds on prediction dimension of convex surrogates. In *Advances in Neural Information Processing Systems*, volume 34, pages 22046–22057, 2021.

Jessie Finocchiaro, Rafael M. Frongillo, and Bo Waggoner. An Embedding Framework for the Design and Analysis of Consistent Polyhedral Surrogates. *Journal of Machine Learning Research*, 25(63):1–60, 2024.

Tobias Fissler and Johanna F. Ziegel. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

Hans Föllmer and Alexander Schied. *Stochastic finance: an introduction in discrete time*. Walter de Gruyter, 2011.

Lance Fortnow and Rakesh V. Vohra. The complexity of forecast testing. *Econometrica*, 77(1):93–105, 2009.

Dean P. Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and Nash dynamics. *Games and Economic Behavior*, 109:271–293, 2018.

Dean P. Foster and Sergiu Hart. Forecast hedging and calibration. *Journal of Political Economy*, 129(12):3447–3490, 2021.

Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(589):40–55, 1997.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Dean P. Foster and Rakesh V. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7–35, 1999.

Paul L. Franco. Speech act theory and the multiple aims of science. *Philosophy of Science*, 86(5):1005–1015, 2019.

Christian Fröhlich and Robert C. Williamson. Risk measures and upper probabilities: Coherence and stratification. *Journal of Machine Learning Research*, 25:1–100, 2024.

Christian Fröhlich and Robert C. Williamson. Insights from insurance for fair machine learning. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, page 407–421. Association for Computing Machinery (ACM), 2024.

Christian Fröhlich and Robert C. Williamson. Scoring rules and calibration for imprecise probabilities. *arXiv preprint arXiv:2410.23001*, 2024.

Christian Fröhlich, Rabanus Derr, and Robert C. Williamson. Towards a strictly frequentist theory of imprecise probability. In *International Symposium on Imprecise Probability: Theories and Applications*, volume 215, pages 230–240, 2023.

Christian Fröhlich, Rabanus Derr, and Robert C. Williamson. Strictly frequentist imprecise probability. *International Journal of Approximate Reasoning*, 168:109148, 2024.

Rafael Frongillo and Ian A. Kash. Vector-valued property elicitation. In *Conference on Learning Theory*, pages 710–727. PMLR, 2015.

Rafael Frongillo and Ian A. Kash. Elicitation complexity of statistical properties. *Biometrika*, 108(4):857–879, December 2021.

Rafael Frongillo and Andrew Nobel. Memoryless sequences for differentiable losses. In *Conference on Learning Theory*, pages 925–939. PMLR, 2017.

Rafael Frongillo and Andrew Nobel. Memoryless sequences for general losses. *Journal of Machine Learning Research*, 21:80–1, 2020.

Peter Gács. Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science*, 341(1-3):91–137, 2005.

Walter Bryce Gallie. Essentially contested concepts. In *Proceedings of the Aristotelian society*, volume 56, pages 167–198, 1955.

Alexander Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163, 2007.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2725–2792. SIAM, 2024.

Gerd Gigerenzer, Ralph Hertwig, Eva Van Den Broek, Barbara Fasolo, and Konstantinos V. Katsikopoulos. "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis: An International Journal*, 25(3):623–629, 2005.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153, 1989.

Donald Gillies. *An Objective Theory of Probability*. Routledge, 2010. (first published 1973).

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In *International Conference on Machine Learning*, pages 11459–11492. PMLR, 2023.

Clark Glymour. Instrumental probability. *The Monist*, 84(2):284–300, 2001.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting and Johannes Resin. Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2):3226–3286, 2023.

Radosław Godowski. Varieties of orthomodular lattices with a strongly full set of states. *Demonstratio Mathematica*, 14(3):725–734, 1981.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Parikshit Gopalan and Lunjia Hu. Calibration through the lens of indistinguishability. *arXiv preprint arXiv:2509.02279*, 2025.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint arXiv:2210.08649*, 2022a.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:21, Dagstuhl, Germany, 2022b. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Parikshit Gopalan, Michael P. Kim, Mihir A. Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022c.

Parikshit Gopalan, Lunjia Hu, and G. Rothblum. On Computationally Efficient Multi-Class Calibration. *arXiv.org*, 2024a. doi: 10.48550/arxiv.2402.07821.

Parikshit Gopalan, Michael P. Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2024b.

Igor I. Gorban. *Randomness and Hyper-randomness*. Springer, 2018.

Erich Grädel and Jouko Väänänen. Dependence and independence. *Studia Logica*, 101(2): 399–410, 2013.

Matejka Grgic and Igor Z. Žagar. *How to do things with tense and aspect: Performativity before Austin*. Cambridge Scholars Publishing, 2020.

Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367 – 1433, 2004.

Peter D. Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

Stanley P. Gudder. Quantum probability spaces. *Proceedings of the American Mathematical Society*, 21(2):296–302, 1969.

Stanley P. Gudder. Generalized measure theory. *Foundations of Physics*, 3(3):399–411, 1973.

Stanley P. Gudder. *Stochastic methods in quantum mechanics*. Elsevier North Holland, 1979.

Stanley P. Gudder. An extension of classical measure theory. *SIAM Review*, 26(1):71–89, 1984.

Stanley P. Gudder and Julia E. Zerbe. Generalized monotone convergence and Radon–Nikodym theorems. *Journal of Mathematical Physics*, 22(11):2553–2561, 1981.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *International Conference on Machine Learning*, 2017.

Siyuan Guo, Chi Zhang, Karthika Mohan, Ferenc Huszár, and Bernhard Schölkopf. Do Finetti: on causal effects for exchangeable data. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 127317–127345, 2024.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. *International Conference on Learning Representations*, 2021.

Chirag Gupta and Aaditya Ramdas. Faster online calibration without randomization: interval forecasts and the power of two choices. In *Conference on Learning Theory*, pages 4283–4309. PMLR, 2022.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online Multivalid Learning: Means, Moments, and Prediction Intervals. *LIPIcs, Volume 215, ITCS 2022*, 215:82:1–82:24, 2022.

Jürgen Habermas. *On the pragmatics of communication*. MIT Press, 1998.

Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. Truthfulness of calibration measures. *Advances in Neural Information Processing Systems*, 37:117237–117290, 2024.

Alan Hájek. "Mises redux"-Redux: Fifteen arguments against finite frequentism. *Erkenntnis*, 45(2):209–227, 1996.

Alan Hájek. The reference class problem is your problem too. *Synthese*, 156(3):563–585, 2007.

Alan Hájek. Fifteen arguments against hypothetical frequentism. *Erkenntnis*, 70(2):211–235, 2009.

Alan Hájek. Interpretations of Probability. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

Ange-Marie Hancock. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on politics*, 5(1):63–79, 2007.

Ange-Marie Hancock. Empirical intersectionality: a tale of two approaches. *UC Irvine Law Review*, 3(2):259–296, 2013.

Bernard E. Harcourt. *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press, 2007.

Moritz Hardt. The emerging science of machine learning benchmarks. Online at `https://mlbenchmarks.org`, 2025. Manuscript.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Philipp Hennig, Michael A. Osborne, and Hans P. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

Corinna Hertweck, Christoph Heitz, and Michele Loi. On the moral justification of statistical parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 747–757, 2021.

David Hilbert. Mathematical problems. lecture delivered before the international congress of mathematicians at Paris in 1900. *Bulletin of the American Mathematical Society*, 8: 437–479, 1902. Translated for the Bulletin, with the author's permission, by Dr.Mary Winston Newson. The original appeared in the Göttinger Nachrichten, 1900, pp. 253-297, and in the Archiv der Mathematik und Physik, 3rd series, vol. 1 (1901), pp. 44-63 and 213-237.

Teophil H. Hildebrandt. On bounded linear functional operations. *Transactions of the American Mathematical Society*, 36(4):868–875, 1934.

Michael A. Hogg and Dominic Abrams. *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. Routledge, 1998.

Benedikt Höltgen. Practical foundations for probability: Prediction methods and calibration. *preprint*, 2024.

Alfred Horn and Alfred Tarski. Measures in Boolean algebras. *Transactions of the American Mathematical Society*, 64(3):467–497, 1948.

Lily Hu. Does calibration mean what they say it means; or, the reference class problem rises again. *Philosophical Studies*, pages 1–27, 2025.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, 2020.

Lunjia Hu and Yifan Wu. Predict to Minimize Swap Regret for All Payoff-Bounded Tasks, April 2024. URL `http://arxiv.org/abs/2404.13503`. arXiv:2404.13503 [cs, stat].

Peter J. Huber. *Robust statistics*. John Wiley & Sons, 1981.

Paul W. Humphreys. Randomness, independence, and hypotheses. *Synthese*, pages 415–426, 1977.

Kodi Husimi. Studies on the foundation of quantum mechanics. I. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, 19:766–789, 1937.

Benedikt Höltgen and Robert C. Williamson. On the richness of calibration. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23. ACM, 2023.

Katerina Ierodiakonou. Theophrastus. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.

International Institute of Forecasters. International journal of forecasting. https://forecasters.org/ijf/, 2025. Accessed: 4 August 2025.

Yoaav Isaacs, Alan Hájek, and John Hawthorne. Non-measurability, imprecise credences, and imprecise chances. *Mind*, 131(523):892–916, 2022.

Victor I. Ivanenko. *Decision systems and nonstochastic randomness*. Springer, 2010.

Atulya Jain and Vianney Perchet. Calibrated forecasting and persuasion. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 489–489, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Sung Jae Jun, Joris Pinkse, and Haiqing Xu. Tighter bounds in triangular systems. *Journal of Econometrics*, 161(2):122–128, 2011.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh V. Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023.

Mark Kac. The search for the meaning of independence. In *The Making of Statisticians*, pages 61–72. Springer, 1982.

Joseph B. Kadane, Mark J. Schervish, Teddy Seidenfeld, P. K. Goel, and A. Zellner. Statistical implications of finitely additive probability. *Bayesian Inference and Decision Techniques*, pages 69–76, 1986. reprinted in Rethinking the Foundations of Statistics, pages 211–232, Elsevier Science Publisher, 1999.

Sham M. Kakade and Dean P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pages 160–171, 2024.

Yuri Kalnishkan. *The Aggregating Algorithm and Predictive Complexity*. PhD thesis, Department of Computer Science, Royal Holloway, University of London, Egham, 2002.

Yuri Kalnishkan and Michael V. Vyugin. Mixability and the existence of weak complexities. In *International Conference on Computational Learning Theory*, pages 105–120. Springer, 2002.

Yuri Kalnishkan, Vladimir Vovk, and Michael V. Vyugin. A criterion for the existence of predictive complexity for binary games. In *Algorithmic Learning Theory: 15th International Conference, ALT 2004, Padova, Italy, October 2-5, 2004. Proceedings*, pages 249–263. Springer, 2004.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7524 LNAI(PART 2):35–50, 2012.

Tibor Katriňák and Tibor Neubrunn. On certain generalized probability domains. *Matematickỳ časopis*, 23(3):209–215, 1973.

Hans G. Kellerer. Verteilungsfunktionen mit gegebenen Marginalverteilungen. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 3(3):247–270, 1964.

A. Ia. Khinchin. 'Mises' theory of probability and the principles of statistical physics', 2015. Translated by Olga Kondrikova and Lukas M. Verburgt.

Andrei Khrennikov. External observer reflections on qbism, its possible modifications, and novel applications. *Quantum Foundations, Probability and Information*, pages 93–118, 2018.

Andrei Y. Khrennikov. *Contextual approach to quantum formalism*. Springer Science & Business Media, 2009a.

Andrei Y. Khrennikov. *Interpretations of probability – Second revised and extended edition*. Walter de Gruyter, 2009b.

Andrei Y. Khrennikov. Bell could become the Copernicus of probability. *Open Systems & Information Dynamics*, 23(02):1650008, 2016.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, 2017.

Michael P. Kim and Juan C. Perdomo. Making Decisions Under Outcome Performativity. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 79:1–79:15, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Michael Kirchhof, Mark Collier, Seong Joon Oh, and Enkelejda Kasneci. Pretrained visual uncertainties. *arXiv preprint arXiv:2402.16569*, 2024.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs*, 67:1–22, 2017.

Robert Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. *arXiv preprint, arXiv 2307.00168*, 2023. Extended abstract published in Proceedings of Thirty Sixth Conference on Learning Theory.

Marko Kolanovic and Rajesh T. Krishnamachari. Big data and AI strategies: Machine learning and alternative data approach to investing. *JP Morgan Global Quantitative & Derivatives Strategy Report*, 25, 2017.

Andreĭ Nikolaevich Kolmogorov. The general theory of measure and probability calculus. *Collected Works of the Mathematical Section, Communist Academy, Section for Natural and Exact Sciences*, 1:8–21, 1927/1929. In Russian. Translated to English in A.N. Shiryayev (Editor), *Selected Works of A.N. Kolmogorov, Volume II Probability and Mathematical Statistics*, pages 48–59, Springer 1992.

Andreĭ Nikolaevich Kolmogorov. *Grundbegriffe de Wahrscheinlichkeitsrechnung*. Julius Springer, 1933.

Andreĭ Nikolaevich Kolmogorov. *Foundations of the theory of probability: Second English Edition*. Chelsea Publishing Company, 1956.

Andreĭ Nikolaevich Kolmogorov. Three approaches to the definition of the concept "quantity of information". *Problemy peredachi informatsii*, 1(1):3–11, 1965.

Jason Konek. Evaluating imprecise forecasts. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 270–279. PMLR, 2023.

Jason Konek. Epistemic conservativity and imprecise credence. *Philosophy and Phenomenological Research*, forthcoming.

Wouter M. Koolen, Muriel Felipe Pérez-Ortiz, and Tyron Lardy. A generalisation of ville's inequality to monotonic lower bounds and thresholds. *arXiv preprint arXiv:2502.16019*, 2025.

Volodymyr Kuleshov and Shachi Deshpande. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, pages 11683–11693. PMLR, 2022.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, volume 32, 2019.

John Lafferty and Larry Wasserman. Challenges in statistical machine learning. *Statistica Sinica*, 16(2):307, 2006.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander

Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Nicolas S. Lambert. Elicitation and Evaluation of Statistical Forecasts. *preprint*, 2019.

Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM conference on Electronic commerce*, EC '08, pages 129–138, New York, NY, USA, 2008.

Bruno Latour. Gabriel Tarde and the end of the social. In *The social in question*, pages 117–132. Routledge, 2012.

Ehud Lehrer. Any inspection is manipulable. *Econometrica*, 69(5):1333–1347, 2001.

Ehud Lehrer. Coherent risk measures induced by partially specified probabilities. Technical report, Tel Aviv University, 2007.

Ehud Lehrer. Partially specified probabilities: decisions and games. *American Economic Journal: Microeconomics*, 4(1):70–100, 2012.

Jacek Leśkow and Antonio Napolitano. Foundations of the functional approach for signal analysis. *Signal processing*, 86:3796–3825, 2006.

Leonid A. Levin. The concept of random sequence. *Soviet Mathemathics Doklady 14*, pages 1413–1416, 1973.

Leonid A. Levin. A concept of independence with applications in various fields of mathematics. Technical Report MIT/LCS/TR-235, MIT, Laboratory for Computer Science, April 1980.

Haonan Lin, Wenbin An, Jiahao Wang, Yan Chen, Feng Tian, Mengmeng Wang, Guang Dai, Qianying Wang, and Jingdong Wang. Flipped classroom: aligning teacher attention with student in generalized category discovery. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 60897–60935, 2024.

Kasper Lippert-Rasmussen. The badness of discrimination. *Ethical Theory and Moral Practice*, 9(2):167–185, 2006.

Vitali Liskevich. *Analysis 1 Lecture Notes 2013/2014*. University of Bristol, 2014.

Lydia T. Liu, Serena Wang, Tolani Britton, and Rediet Abebe. Lost in translation: Reimagining the machine learning life cycle in education. *arXiv preprint arXiv:2209.03929*, 2022.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR, 2020.

Michele Loi, Anders Herlitz, and Hoda Heidari. A philosophical theory of fairness for prediction-based decisions. *SSRN Electronic Journal*, 2019.

Jiuyao Lu, Aaron Roth, and Mirah Shi. Sample efficient omniprediction and downstream swap regret for non-linear losses. *arXiv preprint arXiv:2502.12564*, 2025.

Niklas Luhmann. *Soziale Systeme*, volume 478. Suhrkamp Frankfurt am Main, 1984.

Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local calibration: metrics and recalibration. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1286–1295. PMLR, 01–05 Aug 2022.

David J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

Donald MacKenzie and Yuval Millo. Constructing a market, performing theory: The historical sociology of a financial derivatives exchange. *American journal of sociology*, 109 (1):107–145, 2003.

Dorothy Maharam. Consistent extension of linear functionals and of probability measures. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California*, 6:127–147, 1972.

Ryan Martin and Chuanhai Liu. Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501):301–313, 2013.

Per Martin-Löf. The definition of random sequences. *Information and control*, 9(6):602–619, 1966.

Craig McGarty. *Categorization in Social Psychology*. Sage Publications, 1999.

Xiao-Li Meng. A Heisenberg-esque uncertainty principle for simultaneous (machine) learning and error assessment? *Statistics and Applications, Special Issue in Memory of Prof. C. R. Rao*, 22(3):667–693, 2024.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

Robert C. Merton. Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4:141–183, 1973.

MeteoSwiss. Probability in forecasts. `https://www.meteoswiss.admin.ch/weather/weather-and-climate-from-a-to-z/probability-in-forecasts.html`. Accessed: 9 August 2025.

Enrique Miranda and Marco Zaffalon. Compatibility, coherence and the RIP. In *International Conference on Soft Methods in Probability and Statistics*, 2018.

Giacomo Molinari. Trust the evidence: two deference principles for imprecise probabilities. In *International Symposium on Imprecise Probability: Theories and Applications*, pages 356–366. PMLR, 2023.

Emanuel D. Moss. *The Objective Function: Science and Society in the Age of Machine Intelligence*. Phd thesis, City University of New York, New York, NY, United States, 2021.

Andrei A. Muchnik, Alexei L. Semenov, and Vladimir A. Uspensky. Mathematical metaphysics of randomness. *Theoretical Computer Science*, 207(2):263–317, 1998.

Kevin Mulligan. Promisings and other social acts: Their constituents and structure. In *Speech act and Sachverhalt: Reinach and the foundations of realist phenomenology*, pages 29–90. Springer, 1987.

James Munkres. *Topology*. Pearson Education Limited, 2nd edition, 2014.

Brian C. Murchison. The concept of independence in public law. *Emory Law Journal*, 41: 961, 1992.

Allan H. Murphy and Barbara G. Brown. A comparative evaluation of objective and subjective weather forecasts in the united states. *Journal of Forecasting*, 3(4):369–393, 1984.

Allan H. Murphy and Edward S. Epstein. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755, 1967.

Allan H. Murphy and Robert L. Winkler. Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *Applied Statistics*, 26(1):41, 1977.

Antonio Napolitano and William A. Gardner. Fraction-of-time probability: Advancing beyond the need for stationarity and ergodicity assumptions. *IEEE Access*, 10:34591–34612, 2022.

Louis Narens. An introduction to lattice based probability theories. *Journal of Mathematical Psychology*, 74:66–81, 2016.

Melvyn B Nathanson. *Elementary methods in number theory*, volume 195. Springer Science & Business Media, 2008.

National Weather Service. FAQ – What is the meaning of PoP. `https://preview.weather.gov/ffc/pop`. Accessed: 9 August 2025.

Mirko Navara and Pavel Pták. Considering uncertainty and dependence in Boolean, quantum and fuzzy logics. *Kybernetika*, 34(1):121–134, 1998.

Arch W. Naylor. On decomposition theory: generalized dependence. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(10):699–713, 1981.

Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

Georgy Noarov and Aaron Roth. The statistical scope of multicalibration. In *International Conference on Machine Learning*, pages 26283–26310. PMLR, 2023.

Georgy Noarov and Aaron Roth. Calibration for decision making: A principled approach to trustworthy ML, 2024. URL `https://www.let-all.com/blog/2024/03/13/calibration-for-decision-making-a-principled-approach-to-trustworthy-ml/`. Accessed: 11 February, 2025.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional unbiased prediction for sequential decision making. In *OPT 2023: Optimization for Machine Learning*, 2023.

Andrew B. Nobel. Some stochastic properties of memoryless individual sequences. *IEEE Transactions on Information Theory*, 50(7):1497–1505, 2004.

Wojciech Olszewski and Alvaro Sandroni. A nonmanipulable test. *Annals of Statistics*, 37 (2):1013–1039, 2009.

Wojciech Olszewski and Alvaro Sandroni. Falsifiability. *American Economic Review*, 101 (2):788–818, 2011.

OpenAI (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774v6*, 2024.

Donald S. Ornstein. Ergodic theory, randomness, and "chaos". *Science*, 243(4888):182–187, 1989.

Kent Harold Osband. *Providing Incentives for Better Cost Forecasting (Prediction, Uncertainty Elicitation)*. PhD Thesis, University of California, Berkeley, 1985.

Peter G. Ovchinnikov. Measures on finite concrete logics. *Proceedings of the American Mathematical Society*, 127(7):1957–1966, 1999.

Athanasios Papoulis. *Random variables and stochastic processes*. McGraw Hill, 3rd edition, 1965.

Joel Matthew Parker. *Randomness and legitimacy in selecting democratic representatives*. The University of Texas at Austin, 2011.

Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 39–48, 2019.

Renato Pelessoni and Paolo Vicig. Imprecise previsions for risk measurement. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 11(04):393–412, 2003.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Juan Carlos Perdomo and Benjamin Recht. In defense of defensive forecasting. *arXiv preprint arXiv:2506.11848*, 2025.

Juan Carlos Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult lessons on social prediction from Wisconsin public schools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, page 2682–2704. Association for Computing Machinery, 2025.

Fotios Petropoulos, Daniele Apiletti, Vassilios Assimakopoulos, Mohamed Zied Babai, Devon K. Barrow, Souhaib Ben Taieb, Christoph Bergmeir, Ricardo J. Bessa, Jakub Bijak, John E. Boylan, et al. Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871, 2022.

Karl Popper. *The logic of scientific discovery*. Routledge, 2010.

Christopher P. Porter. *Mathematical and philosophical perspectives on algorithmic randomness*. University of Notre Dame, 2012.

Pavel Pták. Some nearly Boolean orthomodular posets. *Proceedings of the American Mathematical Society*, 126(7):2039–2046, 1998.

Pavel Pták. Concrete quantum logics. *International Journal of Theoretical Physics*, 39(3): 827–837, 2000.

Hilary Putnam. *The Meaning of the Concept of Probability in Application to Finite Sequences (Routledge Revivals)*. Routledge, 1990/2011. First published in 1990 by Garland Publishing.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M. Koolen. Testing exchangeability: Fork-convexity, supermartingales and E-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.

Aaditya Ramdas, Peter D. Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

K.P.S. Bhaskara Rao and M. Bhaskara Rao. *Theory of charges: a study of finitely additive measures*. Academic Press, 1983.

John Rawls. *A theory of justice*. The Belknap Press of Harvard University Press, 1971.

Tim Räz. Group fairness: Independence revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 129–137, 2021.

Benjamin Recht. The actuary's final word on algorithmic decision making. *arXiv preprint arXiv:2509.04546*, 2025.

Hans Reichenbach. *The theory of probability*. University of California Press, 1949.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(3):731–817, 2011.

Anka Reuel-Lamparth, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 21763–21813, 2024.

Ángel Rivas. On the role of joint probability distributions of incompatible observables in Bell and Kochen–Specker theorems. *Annals of Physics*, 411:167939, 2019.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Science & Business Media, 2009.

Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? In *The Thirty Eighth Annual Conference on Learning Theory*, pages 4937–4972. PMLR, 2025.

Gian-Carlo Rota. Twelve problems in probability no one likes to bring up. In *Algebraic combinatorics and computer science*, pages 57–93. Springer, 2001.

Aaron Roth and Mirah Shi. Forecasting for Swap Regret for All Downstream Agents. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pages 466–488, New York, NY, USA, December 2024. Association for Computing Machinery.

Ben-Zion A. Rubshtein, Genady Ya. Grabarnik, Mustafa A. Muratov, and Yulia S. Pashkova. *Foundations of symmetric spaces of measurable functions*. Springer, 2016.

Andrew L. Rukhin. Loss functions for loss estimation. *The Annals of Statistics*, 16(3): 1262–1269, 1988.

Kornelius Räth and Nicole Ludwig. Marginal calibration in regression does not guarantee useful uncertainty estimates. *Forthcoming*, 2025.

Alvaro Sandroni. The reproducible properties of correct forecasts. *International Journal of Game Theory*, 32(1):151–159, 2003.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Teresa Scantamburlo. Non-empirical problems in fair machine learning. *Ethics and Information Technology*, 23(4):703–712, 2021.

Eric Schechter. *Handbook of analysis and its foundations*. Academic Press, 1997.

Henry Scheffe. A statistical theory of calibration. *The Annals of Statistics*, 1(1):1–37, 1973.

Mark J. Schervish. Self-calibrating priors do not exist: Comment. *Journal of the American Statistical Association*, 80(390):341–342, 1985.

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989.

Mark J. Schervish, Teddy Seidenfeld, and Joseph B. Kadane. Proper scoring rules, dominated forecasts, and coherence. *Decision Analysis*, 6(4):202–221, 2009.

Claus-Peter Schnorr. *Zufälligkeit und Wahrscheinlichkeit: eine algorithmische Begründung der Wahrscheinlichkeitstheorie*, volume 218 of *Lecture Notes in Mathematics*. Springer-Verlag, 1971.

Claus-Peter Schnorr. The process complexity and effective random tests. In *Proceedings of the fourth annual ACM symposium on Theory of computing*, pages 168–176, 1972.

Claus-Peter Schnorr. A survey of the theory of random sequences. In *Basic Problems in Methodology and Linguistics*, pages 193–211. Springer, 1977.

Miriam Schoenfield. Chilling out on epistemic rationality: A defense of imprecise credences. *Philosophical Studies*, 158(2):197–219, 2012.

Gerhard Schurz. No free lunch theorem, inductive skepticism, and the optimality of meta-induction. *Philosophy of Science*, 84(5):825–839, 2017.

Gerhard Schurz and Hannes Leitgeb. Finitistic and frequentistic approximation of probability measures with or without σ-additivity. *Studia Logica*, 89(2):257–283, 2008.

John R. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, 1969.

Teddy Seidenfeld. Calibration, coherence, and scoring rules. *Philosophy of Science*, 52(2): 274–294, 1985.

Teddy Seidenfeld. personal communication at ISIPTA 2023, 2023.

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Glenn Shafer. Discussion on Hedging Predictions in Machine Learning by A. Gammerman and V. Vovk . *The Computer Journal*, 50(2):164–172, 02 2007.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

Glenn Shafer. How the game-theoretic foundation for probability resolves the Bayesian vs. frequentist standoff. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*, pages 264–275. Chapman and Hall/CRC, 2024.

Glenn Shafer and Vladimir Vovk. *Probability and finance: it's only a game!*, volume 491. John Wiley & Sons, 2005.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*, volume 455. John Wiley & Sons, 2019.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Stephanie A. Shields. Gender: an intersectionality perspective. *Sex Roles*, 59:309–311, 2008.

Eran Shmaya. Many inspections are manipulable. *Theoretical Economics*, 3(3):367–382, 2008.

Gilbert Simondon. The genesis of the individual. In *Incorporations*, pages 296–319. Zone Books, 1992.

Anna De Simone and Pavel Pták. Extending coarse-grained measures. *Bulletin of The Polish Academy of Sciences. Mathematics*, 54:1–11, 2006.

John Simpson and Edmund Weine. "independence, n.". In *Oxford English Dictionary - Second Edition*. Oxford University Press, 1989.

Anurag Singh, Siu Lun Chau, and Krikamol Muandet. Truthful elicitation of imprecise forecasts. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, number 169 in UAI '25, 2025.

Ján Šipoš. Subalgebras and sublogics of σ-logics. *Mathematica Slovaca*, 28(1):3–9, 1978.

Barry Smith. Towards a history of speech act theory. *Speech acts, meanings and intentions. Critical approaches to the philosophy of John R. Searle*, pages 29–61, 1990.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265, 2015.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019.

Wolfgang Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical logic*, 9(1):73–99, 1980.

Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Conference on Learning Theory*, pages 482–526. PMLR, 2014.

Alan E. Stewart, Castle A. Williams, Minh D. Phan, Alexandra L. Horst, Evan D. Knox, and John A. Knox. Through the eyes of the experts: Meteorologists' perceptions of the probability of precipitation. *Weather and Forecasting*, 31(1):5–17, 2016.

Marshall H. Stone. The theory of representations for Boolean algebras. *Transactions of the American Mathematical Society*, 40:37–111, 1936.

Peter Stone. *The Luck of the Draw: The Role of Lotteries in Decision-Making*. Oxford University Press, 2011.

Patrick Suppes. The probabilistic argument for a non-classical logic of quantum mechanics. *Philosophy of Science*, 33(1):14–21, 1966.

Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms performance: reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*, pages 33209–33228. PMLR, 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Kohtaro Tadaki. An operational characterization of the notion of probability by algorithmic randomness. In *Proceedings of the 37th Symposium on Information Theory and its Applications (SITA2014)*, volume 5, pages 389–394, 2014.

Henri Tajfel. Social identity and intergroup behaviour. *Social Science Information*, 13(2): 65–93, 1974.

Eran Tal. Measurement in Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.

Jingwu Tang, Jiahao Zhang, Fei Fang, and Zhiwei Steven Wu. Persuasive prediction via decision calibration. *arXiv preprint arXiv:2505.16141*, 2025.

Terence Tao. 275a, notes 2: Product measures and independence. `https://terrytao.wordpress.com/2015/10/12/275a-notes-2-product-measures-and-independence/`, October 2015. Accessed: 12 October 2015.

Gabriel Tarde. *Monadology and sociology*. re.press, 2012. Translation of the 1895 edition of "Monadologie et Sociologie" by Theo Lorenc.

Matthias C.M. Troffaes and Gert De Cooman. *Lower Previsions*. John Wiley & Sons, Ltd., 2014.

Jennifer S. Trueblood and Jerome R. Busemeyer. A quantum probability account of order effects in inference. *Cognitive science*, 35(8):1518–1552, 2011.

U.S. Department of Education. Issue Brief: Early Warning Systems. Technical Report ED571990, U.S. Department of Education, 2016. Accessed: 6 August 2025.

Vladimir A. Uspenskii, Alexei L. Semenov, and A Kh. Shen. Can an individual sequence of zeros and ones be random? *Russian Mathematical Surveys*, 45(1):121, 1990.

Jóakim v. Kistowski, Jeremy A. Arnold, Karl Huppler, Klaus-Dieter Lange, John L. Henning, and Paul Cao. How to build a benchmark. In *Proceedings of the 6th ACM/SPEC international conference on performance engineering*, pages 333–336, 2015.

Hans Vaihinger. *Die Philosophie des Als Ob. System der theoretischen, prak- tischen und religiösen Fiktionen der Menschheit auf Grund eines idealistischen Positivismus. Mit einem Anhang über Kant und Nietzsche. Herausgegeben von H. Vaihinger.* Verlag von Reuther und Reichard, 1911. Translated in Hans Vaihinger and Michael A. Rosenthal, The philosophy of 'as if', 2021, Routledge.

Sergey S. Vallander. The structure of separable Dynkin algebras. *Vestnik St. Petersburg University: Mathematics*, 49(3):219–223, 2016.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Ben Van Calster, David J. McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: the Achilles heel of predictive analytics. *BMC medicine*, 17(1): 230, 2019.

Noah van Dongen, Riet van Bork, Denny Borsboom, Markus Eronen, and Jan-Willem Romeijn. Rethinking psychology's foundations: A reflection on five philosophical and methodological commitments. *PsyarXiv*, 2025.

C. Van Fraassen. Belief and the will. *The Journal of Philosophy*, 81(5):235–256, 1984.

Michiel Van Lambalgen. Von Mises' definition of random sequences reconsidered. *The Journal of Symbolic Logic*, 52(3):725–755, 1987.

Michiel Van Lambalgen. The axiomatization of randomness. *The Journal of Symbolic Logic*, 55(3):1143–1167, 1990.

J. H. P. Van Wel, J. Kinyua, A. L. N. van Nuijs, S. Salvatore, Jørgen G. Bramness, A. Covaci, and G. Van Hal. A comparison between wastewater-based drug data and an illicit drug use survey in a selected community. *International Journal of Drug Policy*, 34:20–26, 2016.

John Venn. *The Logic of Chance: An essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its applications to moral and social science, and to statistics*. Macmillan and Company, 1876.

Rajeev Verma, Volker Fischer, and Eric Nalisnick. On calibration in multi-distribution learning. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 938–950, New York, NY, USA, 2025. Association for Computing Machinery.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939.

Giuseppe Vitali. *Sul problema della misura dei Gruppi di punti di una retta: Nota*. Tipografia Gamberini e Parmeggiani, 1905.

Sean Vitousek, Patrick L. Barnard, Patrick Limber, Li Erikson, and Blake Cole. A model integrating longshore and cross-shore processes for predicting long-term shoreline response to climate change. *Journal of Geophysical Research: Earth Surface*, 122(4):782–806, 2017.

Sérgio B. Volchan. What is a random sequence? *The American mathematical monthly*, 109 (1):46–63, 2002.

Elart Von Collani. A note on the concept of independence. *Economic Quality Control*, 21 (1):155–164, 2006.

Richard Von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(1):52–99, 1919.

Richard Von Mises. *Probability, statistics, and truth*. Dover Publications, Inc., 1981.

Richard Von Mises and Hilda Geiringer. *Mathematical theory of probability and statistics*. Academic press, 1964.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 3rd edition, 1953.

Nikolai Nikolaevich Vorob'ev. Consistent families of measures and their extensions. *Theory of Probability & Its Applications*, 7(2):147–163, 1962.

Daniela Voss. Simondon on the notion of the problem: A genetic schema of individuation. In *Problems in Twentieth Century French Philosophy*, pages 94–112. Routledge, 2020.

Vladimir Vovk. Defensive prediction with expert advice. In *International Conference on Algorithmic Learning Theory*, pages 444–458. Springer, 2005.

Vladimir Vovk. Continuous and randomized defensive forecasting: unified view. *arXiv preprint arXiv:0708.2353*, 2007.

Vladimir Vovk. Non-algorithmic theory of randomness. In *Fields of Logic and Computation III: Essays Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday*, pages 323–340. Springer, 2020.

Vladimir Vovk and Alexander Shen. Prequential randomness and probability. *Theoretical Computer Science*, 411(29-30):2632–2646, 2010.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting. In *International Workshop on Artificial Intelligence and Statistics*, pages 365–372. PMLR, 2005.

Vladimir G. Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal statistical society: series B (Methodological)*, 55(2):317–341, 1993.

Volodya Vovk and Chris Watkins. Universal portfolio selection. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 12–23, 1998.

Kartik Waghmare and Johanna Ziegel. Proper scoring rules for estimation and forecast evaluation. *arXiv preprint arXiv:2504.01781*, 2025.

Abraham Wald. Die Widerspruchsfreiheit des Kollektivbegriffs. *Ergebnisse eines mathematischen Kolloquiums*, 8:38–72, 1937.

Abraham Wald. *Statistical decision functions*. John Wiley & Sons, 1950.

Peter Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, 1991.

Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2-3):125–148, 2000.

Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing*, 1(1):1–45, 2024.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

S. Laurel Weldon. Intersectionality. In Gary Goertz and Amy G. Mazur, editors, *Politics, Gender, and Concepts: Theory and Methodology*, pages 193–218. Cambridge University Press, 2008.

Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In *International Conference on Machine Learning*, volume 32. PMLR, 2020.

David Williams. *Probability with martingales*. Cambridge University Press, 1991.

Peter M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383, 2007. revised version of: Notes on conditional previsions. Research Report, School of Math. and Phys. Science, University of Sussex, 1975.

Robert C. Williamson. The rhetoric of machine learning, 2024. talk presented at "Persuasive Algorithms? A symposium on the rhetoric of generative AI".

Robert C. Williamson and Zac Cranko. The geometry and calculus of losses. *Journal of Machine Learning Research*, 24(342):1–72, 2023.

Robert C. Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797. PMLR, 2019.

Robert C. Williamson, Elodie Vernet, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(222):1–52, 2016.

William C. Wimsatt. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press, 2007.

Julia L. Wirch and Mary R. Hardy. Distortion risk measures: coherence and stochastic dominance. In *International congress on insurance: Mathematics and economics*, pages 15–17, 2001.

P. Wojtaszczyk. *Banach spaces for analysts*. Cambridge University Press, 1991.

Feng Xie, Zhen Yao, Lin Xie, Yan Zeng, and Zhi Geng. Identification and estimation of the bi-directional mr with some invalid instruments. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 48256–48291, 2024.

Zhen-Peng Xu and Adán Cabello. Necessary and sufficient condition for contextuality from incompatibility. *Physical Review A*, 99(2):020103, 2018.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019.

Sebastian Zezulka and Konstantin Genin. Prediction, performativity, and potential outcomes: Communicative rationality in prediction-allocation problems. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 299–299, 2025.

Jiankang Zhang. Subjective ambiguity, expected utility and Choquet expected utility. *Economic Theory*, 20(1):159–181, 2002.

Shengjia Zhao and Stefano Ermon. Right decisions from wrong predictions: A mechanism design alternative to individual calibration. In *International Conference on Artificial Intelligence and Statistics*, pages 2683–2691. PMLR, 2021.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pages 11387–11397. PMLR, 2020.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 34, pages 22313–22324, 2021.

Alexander K. Zvonkin and Leonid A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83, 1970.