Transformations, Equivalences and Comparisons of Learning Problems

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Laura Iacovissi, M.Sc.

aus Frosinone/Italien

Tübingen

2026

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

Tag der mündlichen Qualifikation:	28.07.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Robert C. Williamson
2. Berichterstatter:	Prof. Dr. Ingo Steinwart

Disclaimer. I acknowledge the use of generative AI tools to assist with the editing of the present thesis. Furthermore, I would like to thank Diego Baptista Theuerkauf, Jack Brady and Raphael Sayer for their help to improve the text.

Preface

Earlier this past year, I was in a tiny little countryside village close to my tiny little hometown in Italy. Friends of friends were there, people I never met before, and who live lives that are far outside of my current academic bubble (a kind of life that I have lived as well, as a first generation academic). They started talking about AI-powered chatbots, and how they seemed to be more or less of an expert in their field; they worked as doctors, developers, teachers and more. Some of them loved it, some of them found it stupid, the majority only liked it for certain tasks. All of them were extremely familiar with it.

What I brought back to the tiny little German town where I now live was a sense of astonishment at how quickly things have evolved during my PhD years. Even if I had been well-aware that large language models were becoming widespread in our current society, this conversation made the awareness very concrete. What struck me was how natural and comfortable everyone seemed to be in using and talking about this tool, which I have for a long time only associated with data science labs and academia. I imagine that hearing the idiom "I googled that" entering everyday speech must have felt somewhat similar.

Machine learning has been in a frenetic state for many years, long before diffusion and GPT models. It was in this climate that I completed my studies, chose to pursue a PhD, and had to decide which line of research to follow. The pace has not slowed since. Fortunately, messy empirical fields are a great opportunity for researchers with a taste for theory to try to bring some order.

Still, it would be untrue, and perhaps even presumptuous, to say that I was drawn to study what is in these pages purely out of perceived necessity. In fact, past and current observations of how the world has reacted to the introduction of AI to the public have led me to ask different, more socio-technical questions. While I aim to return to these questions in my future work, I look back at my research efforts in a more fundamental and theoretical area as steps needed to build my own scientific foundations.

However, just as much, if not more, I arrived at these topics because of my obsessive and stubborn nature. An obsession with order and structure originally pulled me toward mathematics; stubbornness kept me in the uncertain and competitive academic world even when my research did not seem to catch the attention of the less theoretically inclined machine learning scientists. And, of course, finding a community of empathetic and curious people made a journey with many downs and rare ups far more enjoyable.

I had the fortune to start looking for a PhD position right when Bob Williamson moved to the University of Tübingen. Bob, I am grateful for your consistent support every time I asked for it, and sometimes even when I did not. I deeply appreciate how your supervision helped me grow as a scientist and as a person, and how you have been available for feedback even through hard times. As you once very humbly stated at a group retreat, you also did a great job hiring people for the FMLS group. Thanks to Armando, Ben, Charlotte, Chris, Mina, Nan, Rabanus, Raphael and Wenkai, who made every group activity (research- or beer-related) fun and engaging, and the work environment feel like a safe space.

I would also like to thank Dan Zhang for supporting me during the first year of my PhD and for always being concerned about my well-being, as well as Caterina De Bacco and the whole PIO group, who continued to offer scientific exchange, mentorship, and very competitive Tischfußball tournaments even after my internship ended. I am grateful to the International Max Planck Research School for Intelligent Systems for their support, and to Ulrike von Luxburg and Ingo Steinwart for serving on my Thesis Advisory Committee.

I thank all the authors in the reference list for writing interesting work, and, equally importantly, for bringing me close to achieving my goal of having every letter of the alphabet represented in the bibliography. Special thanks to Osterreicher and Oyen for representing O, absent for a long time, and Bob and Raphael for trying to help. Sadly, it seems to be very hard to find first authors whose surnames start with U and who have done work I can cite in this thesis. In fact, it seems difficult to find U surnames at all (see Fig. 1). In case you are embarking on a similar quest: pick your research field carefully.

Lastly, I would like to thank all of my friends, who over the years have been a great source of support, regardless of where they were or what life brought them. I feel very fortunate to have so many deep and meaningful connections that withstand the constant flow of complaints that being a friend of mine entails. While most of them will be addressed personally, I would like to briefly mention some here. Jack and Nico, thank you for being such a great refuge: I really cannot overstate how important you have been to me. Thanks also to the amazing KS45 community, made up of the most beautiful constellation of people a WG could wish for: Anne, Caio, Kibidi, Lisa, Marti, Nath, Rebbi, and Tobi. Last but not least, thanks to my parents and siblings: Grazie per tutto il sostegno e per la cieca fiducia, nonostante spesso sia stata lontana, confusa, ipersensibile e complessivamente difficile da capire.

Laura Iacovissi
Tübingen, April 1st 2026

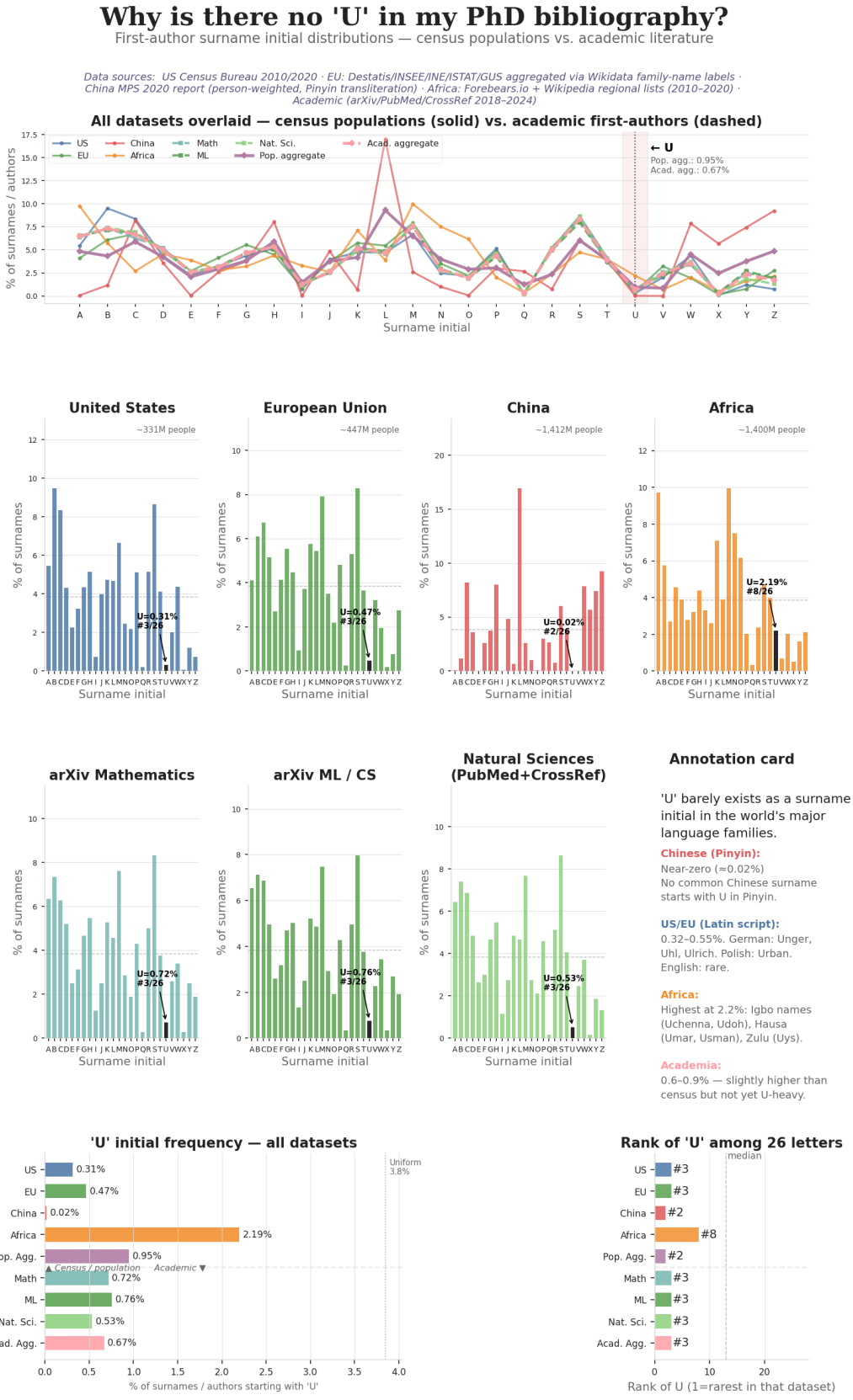


Figure 1: The data were collected and visualized with the assistance of Claude Sonnet 4.6, GPT-5.3, and GPT-5-mini. This figure has no scientific validity, as it has not been created for scientific purposes.

Abstract

While artificial intelligence (AI) has achieved striking practical success across diverse applications, the theoretical foundations of many modern methods remain incomplete. In particular, a rigorous account of how the individual components of an AI system influence the outcome of its (machine) learning process is yet to be established. These components are numerous and heterogeneous, contributing to the high complexity of AI models; consequently, there is no reason to expect a single approach to fully resolve this gap. Thus, the present thesis abstracts away from the particular data points, optimization algorithms, and other technical aspects of AI systems. Paying a price in specificity, we gain a significant advantage: a framework for theoretically characterizing the interplay between the core components of a machine learning problem — loss function, model class, and data-generating probability distribution. In particular, we pay specific attention to the latter, as it is an often overlooked component in many modern works.

We prove two main types of results, namely Bayes risk equalities and inequalities, obtaining *equivalences* and *comparisons* of learning problems, respectively. As a preliminary analysis, we provide an exhaustive taxonomy of stochastic transformations of a problem's data distribution by leveraging properties of Markov kernels, i.e., the objects used to model the transformations. Studying equalities, we then learn how different Markov kernel types lead to distinct consequences. For example, applying label noise to the data distribution yields a learning problem *equivalent* to one that retains the original data-generating probability but with a modified loss function. By contrast, attribute noise will jointly affect both loss and model class in the equivalent problem. These findings suggest that classical loss correction techniques for noisy data are only effective in the case of label noise, not for general attribute noise. We then compute what a generalized loss correction approach would entail for attribute noise. However, focusing solely on equivalences only partially explains how tweaking one component affects the learning outcome. Studying the inequalities, we explore when one learning problem can be considered superior to another. This challenge can be addressed in multiple ways, as problems can be *compared* by varying the probability distribution, model class, or loss. Previous work explored comparing conditional probabilities for every loss and model class, arriving at interesting characterization results nowadays grouped under the Blackwell-Sherman-Stein theorem. This work has found application in economics as well as learning theory, and nicely relates to classical results in information theory. Our contribution frames the existing results into the larger perspective

of comparing learning problems instead of conditional probabilities. We then focus on comparing couples of model classes and losses for every probability, introducing a theory complementary to Blackwell's work, revolving around establishing when a decision maker is better than another. We study this question by means of Bayes risk orderings, which are proved to be equivalent to the inclusion ordering on certain sets, called superprediction sets. These sets are defined starting from fixing a loss and model class, and therefore provide geometrical understanding to the question of comparing decision makers. We additionally prove sufficient conditions for the ordering to hold on certain types of loss and model classes, which are decided by the type of noisy data considered. Hence, these results additionally show the central role of knowing what are the data at hand so to be able to understand the related learning problem.

This work advances theoretical understanding of machine learning systems as highly entangled entities. Its first set of results underscores the critical importance of choosing your data carefully, as they demonstrate that changing the distribution is equivalent to changing the loss and model class. This can cause strong mismatches between what one expects the machine to learn and what the machine actually learns. In particular, the equalities proved offer guidance for designing corrections that aim to achieve accurate learning in the presence of noisy data. The second set of contributions introduces a novel theory for comparison of models and losses, and in particular for comparing them before and after a transformation is applied to the learning problem. This is relevant in the field of model complexity, as it introduces a decision-theoretic inspired framework to assess whether a certain decision maker is more reliable (in terms of risk) than another under a certain type of noise.

Contents

Glossary

$\mathcal{Z}, \mathcal{X}, \mathcal{A}$	Sets, usually in calligraphic
$\Sigma(\mathcal{Z})$	Borel σ-algebra on $\mathcal{Z}$
$(\mathcal{Z}, \Sigma(\mathcal{Z}))$	Standard Borel measurable space
$(\mathcal{X}, \Sigma(\mathcal{X}))$	Attribute space
$(\mathcal{Y}, \Sigma(\mathcal{Y}))$	Label space
μ and ν	(Signed) measures
$\mathrm{ba}(\mathcal{Z})$	Set of finitely additive, signed measures with bounded total variation on $(\mathcal{Z}, \Sigma(\mathcal{Z}))$
ϕ and ψ	Probability measures, or distributions
$\Delta_{\mathrm{ca}}(\mathcal{Z})$	Set of countably additive probabilities on $(\mathcal{Z}, \Sigma(\mathcal{Z}))$
$\Delta_{\mathrm{ba}}(\mathcal{Z})$	Set of finitely additive probabilities on $(\mathcal{Z}, \Sigma(\mathcal{Z}))$
$\mathcal{B}_b(\mathcal{Z})$	Set of bounded, Borel-measurable functions $f\colon \mathcal{Z} \to \mathbb{R}$
$\sigma(\mathrm{ba}, \mathcal{B}_b)(\mathcal{Z}),\ \sigma(\mathcal{B}_b, \mathrm{ba})(\mathcal{Z})$	Weak σ-algebra on $\mathrm{ba}(\mathcal{Z})$ and $\mathcal{B}_b(\mathcal{Z})$, resp.
$\langle \cdot, \cdot \rangle\colon \mathcal{Z} \times \mathcal{W} \to \mathbb{R}$	Dual pairing relating $\mathcal{Z}$ and $\mathcal{W}$
$\rho_{\mathcal{A}}\colon \mathcal{W} \to \mathbb{R}$	Concave support function of a set $\mathcal{A} \subset \mathcal{Z}$ w.r.t. a dual pairing with $\mathcal{W}$
X, Y, Z	Random variables

$\kappa\colon \mathcal{Z} \rightsquigarrow \mathcal{W}$	Markov kernel from $(\mathcal{Z}, \Sigma(\mathcal{Z}))$ to $(\mathcal{W}, \Sigma(\mathcal{W}))$
$\dot{\kappa}\colon \mathcal{Z} \to \Delta_{\mathrm{ca}}(\mathcal{W})$	Markov transition of κ
$\hat{\kappa}\colon \mathcal{B}_b(\mathcal{W}) \to \mathcal{B}_b(\mathcal{Z})$	Markov transition operator of κ
$\check{\kappa}\colon \mathrm{ba}(\mathcal{Z}) \to \mathrm{ba}(\mathcal{W})$	Markov transition adjoint operator of κ
$\kappa^{\dagger}\colon \mathcal{W} \rightsquigarrow \mathcal{Z}$	Bayesian Inversion of Markov kernel κ
$\mathcal{M}(\mathcal{Z}, \mathcal{W})$	Set of Markov kernels from $(\mathcal{Z}, \Sigma(\mathcal{Z}))$ to $(\mathcal{W}, \Sigma(\mathcal{W}))$
$D(\kappa)$	Domain of a Markov kernel
$I(\kappa)$	Image of a Markov kernel
τ s.t. $I(\tau) = \mathcal{X}$	Attribute corruption kernel
λ s.t. $I(\lambda) = \mathcal{Y}$	Label corruption kernel
$\kappa_\nu \equiv \nu, \nu \in \Delta_{\mathrm{ca}}(\mathcal{Z})$	Degenerate kernel, or totally uninformative
$\kappa_f\colon \mathcal{Z} \rightsquigarrow \mathcal{W}$	Deterministic kernel induced by $f\colon \mathcal{Z} \to \mathcal{W}$
$\delta\colon \mathcal{Z} \rightsquigarrow \mathcal{Z}$	Identity kernel, or Dirac delta kernel
$\#$	Push forward operation
$\circ$	Kernel chain composition
$\times$	Kernel product composition
$\otimes$	Kernel superposition
$\circ_{\mathcal{Z}}$	Kernel partial chain composition w.r.t. $\mathcal{Z}$
π	Prior probability measure
$E\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$	Experiment
$F = E^{\dagger}\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$	Posterior kernel

$\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$	Loss function
$\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$	Model class
$\ell \circ \mathcal{H}$	Attainable predictions
$(\ell, \mathcal{H}, \phi)$	Learning problem
$\tilde{z}$, $\tilde{\mathcal{Z}}$, and $\tilde{\phi}$	Corrupted object, set, and probability
R	Risk
CBR	Conditional Bayes risk
BR	Bayes risk
$\mathrm{spr}(\ell)$	Superprediction set of ℓ
$\mathrm{spr}(\ell \circ \mathcal{H})$	Constrained superprediction set of $\ell \circ \mathcal{H}$
$\equiv_{\mathrm{BR}}$	Equivalence relation on the set of learning problems w.r.t. Bayes risk
$\succeq_{\mathrm{BR}}$	Partial ordering on the set of attainable predictions $\ell \circ \mathcal{H}$ w.r.t. Bayes risk
$\mathcal{A}^{\diamond}$	Antipolar of a set $\mathcal{A}$
$\rho^{\diamond}_{\mathcal{A}}\colon \mathcal{Z} \to \overline{\mathbb{R}}_{\geq 0}$	Antipolar of the support function ρ, for $\mathcal{A} \in \mathcal{W}$
$\alpha_{\ell \circ \mathcal{H}}(\kappa)$	Inflation coefficient of κ w.r.t. $\ell \circ \mathcal{H}$

Chapter 1

Introduction

> *Mathematics may be compared to a mill of exquisite workmanship, which grinds you stuff of any degree of fineness; but, nevertheless, what you get out depends on what you put in; and as the grandest mill in the world will not extract wheat-flour from peas-cods, so pages of formulae will not get a definite result out of loose data.*
>
> *Huxley (1869)*

Learning something new requires gathering examples. This principle is applicable to any context, and machine learning is no exception. The initial step in the development of a model, regardless of its complexity, is to establish a goal and gather the relevant data to achieve it. Hence, data constitute the primary interface between an observed phenomenon and an algorithm: the decisions taken during the data collection process will influence the learned model of such a phenomenon. In practice, however, data often occupy a secondary role in published work. With the exception of certain subfields, emphasis is heavily placed on improving model performance through architectural innovation or, more recently, through advances in training and evaluation techniques. Even when data curation is indeed crucial, how the training *corpora* are collected and processed is frequently undisclosed, proprietary, or so under-documented that it effectively becomes part of the black box.

That machine learning evolved such a secretive approach to data is rather surprising. In statistical modeling, from which many of the current practices originate, a data-centric perspective is far more common. In fact, as articulated by Craiu et al. (2023), one of the defining abilities of a statistician is to "see through" the given observations and extract "enlightenment from dark data". Hence, the emphasis lies not only on what is observed, but also on what is hidden (*dark*), and on the mechanisms that produced the observations (*data*) in the first place. It follows that a better understanding of what the possible forms and causes of dark data are can be beneficial to learning, as the learning objective is often to find

a function matching as closely as possible the generating process of the observations.

We review in the following what is currently understood as dark data and what could be added to their definition. In the same paper (Craiu et al., 2023), the authors describe dark data as "any kind of data that are lost or distorted before analysis, as well as unobserved or unobservable data [...] that are constructed or conceptualized because they serve modeling and computational purposes". Examples of the former include missing or incomplete data and noisy measurements; examples of the latter include latent variables, auxiliary variables, and hidden states. However, they do not attempt to describe nor resolve all possible sources and manifestations of dark data. Hand (2020)'s book instead provides a long, and yet non-exhaustive description of many of the forms dark data can take, as well as their causes. While appreciating the contribution, we note that the author only focuses on statistical learning, and does not take into account the environment in which a model is deployed. To adapt the notion of dark data to modern machine learning, we add to the above list also intentionally manipulated data, such as adversarial noise (Goodfellow et al., 2015) and misreporting in survey statistics (Bound et al., 2001), as well as data modified to ensure privacy (Li and Unger, 2012; Kobsa et al., 2016; Zhou and Wu, 2025) or as an attempt of a user to "domesticate" an algorithm and obtain what they want out of it (Siles et al., 2019; Simpson et al., 2022). In other words, we also want to account for an *interaction* between the machine learning model and the data generating process. Formulations for this interaction have been proposed by allowing the data model to explicitly react to a deployed algorithm, as in strategic classification or performative effects (Hardt et al., 2016; Perdomo et al., 2020; Hardt et al., 2023; Zezulka and Genin, 2025). Importantly, we emphasize that such modifications are not necessarily malicious. In our examples, data may be altered to satisfy ethical, legal, or cooperative objectives, including privacy protection or fairness constraints. The term "dark" should therefore not be read as inherently negative: the consequences of hidden data depend on the goals and assumptions under which learning takes place. They are only hard to see, hidden in the darkness. These phenomena still fit within the broader notion of dark data, as they concern observations that are shaped, filtered, or altered, but it is hard to properly define what "before analysis" means when in the presence of a feedback loop.

How can dark data be accounted for in machine learning? Williamson (2020) proposed starting by questioning the view of data as static facts, and instead regarding them as a dynamic element of a learning task. Beyond the traditional focus on prediction models and loss functions, attention can be directed toward the data dynamics themselves: that is, how different processes lead to the observation of particular data, and how these processes subsequently affect the learning procedure. This concept can be mathematically formalized, following Csiszár (1972), by means of *observation channels*. The naming manifests the author's intention of translating into statistical theory what Shannon (1948) referred to as communication channels, which is the (possibly probabilistic) filter through which we get to observe certain phenomena. Talking about observation channels presupposes the existence of a "true" state of such phenomena, which decision theorists often call Nature, perhaps

showing a rather spiritual view of randomness.[1] In formulas, this is a stochastic matrix, or conditional probability of observing a certain object given some unknown parameters or states. We adopt the formalization, and use it in this thesis to propose a mathematical framework for understanding dark data in the context of machine learning problems.

We refer to the process transforming a learning problem into a different one as *corruption*, borrowing the term from the computer science literature. However, our conceptualization of corruption extends beyond the creation of dark data starting from some "ideal" state (i.e., the action of some observation channel). A practical understanding of corruption and its consequences requires considering also the goals and constraints inherent to the learning task at hand. These are encoded in the chosen loss function, which formalizes preferences over what is considered an acceptable mistake, and in the effective model class, which captures the set of all admissible ways to process data, e.g., optimization procedures, data handling, model initialization, and of course what structures can be learned from data, that is the architecture.

Hence corruption encompasses any modification of a learning problem: changes to the loss function, the model class, or the underlying probability distribution from which data are drawn. Like dark data, corruption is not inherently pejorative; it is a *transformation process* whose effects depend on context. We try to answer two fundamental questions:

1. When does a corruption make one learning problem *equivalent* to another?

2. When does a corruption make one learning problem *worse* than another?

In other words, we want to find structure in the space of learning problems. Since we can do it in two different ways – with equivalences and comparisons – we organize our contributions by theme, splitting them into two main parts, as well as reserve an entire part for setting up the machine learning problem abstraction.

Thesis Outline and Contributions

This thesis is divided into three parts: Abstracting a Learning Problem, Transformations and Equivalences of Problems, and Comparison of Problems via Inequalities. Each part contains topic-specific introductions and reviews of related work. Here, we introduce the main questions and summarize the overall contributions.

[1] Meng (2022) referred to this conceptualization of the data-generating process as the "divine probability" construct. We do not philosophically defend this decision in the thesis, but we do adopt such a conceptualization, as it proves itself to be convenient for the purpose of studying transformations of learning problems. Whether this is the "right" way of modeling a random process, or a data-generating process, is not debated here. The reader can for instance refer to (Derr and Williamson, 2024; Fröhlich and Williamson, 2024; Höltgen and Williamson, 2024) for additional pointers and discussions of how such modeling choices can influence machine learning.

Part 1: Abstracting a Learning Problem The first part establishes the formal framework in which learning is defined, which forms the backbone of this thesis. Learning refers to the problem of selecting adequate actions based on observations and a specified criterion; actions are considered adequate when optimal with respect to a specified objective. To formalize this notion, a number of mathematical tools are required. The most central is the concept of a Markov kernel, introduced in § 2.3. Using this formalism, a learning problem is decomposed into its fundamental components: a data-generating process for the observations, a loss function encoding rewards and penalties for actions and thereby specifying the objective, and a model class describing the admissible mappings from observations to actions (§ 3.1). The learning criterion is treated as an explicit component of the problem, since multiple principles can be used to combine these elements. The analysis is however restricted to risk minimization, which fixes the criterion for the remainder of the thesis. An alternative formulation of the learning optimization problem, based on specific sets, is also developed. Such a formulation allows one to view the minimized risk as known geometrical objects, and enables understanding learning through its lenses.

Part 2: Transformations and Equivalences of Problems The second part develops the view that the building blocks of learning problems (loss function, model class, and data distribution) are intertwined, as applying a transformation to one is equivalent to transforming one or all others. The motivation for investigating this topic initially originated in survey statistics, where selection bias has been observed to have disproportionate effects when the selection variable is correlated with the data. In fact, Meng (2018) has shown that even correlations typically regarded as small can drastically reduce the effective sample size in mean estimation problems. Hence, he proved an equivalence between two different problems, related by a transformation induced by selection bias, and that has rather surprising consequences. This study prompted the question of whether analogous phenomena arise in more general learning settings, a direction later explored in (Meng, 2022) but not generally solved.

In other words, we aim to investigate what are the consequences of applying a specific type of noise to data, and to understand whether certain forms of noise are universally more detrimental than others. Given the absolute nature of such a claim, a prior step is to characterize the space of noise mechanisms of interest, that we call Markovian corruptions. This led to the search for a formulation that is *exhaustive* (capturing all corruptions), *hierarchical* (in terms of corruption complexity), and *compositional* (allowing complex mechanisms to be represented as compositions of simpler ones). The result of this investigation is a taxonomy of "dark data", modeled as Markov kernels. We give a detailed explanation of how much our taxonomy fits the three desiderata in Sections 4.1.2, 4.1.4 and 4.2. Markov kernels, however, provide an exhaustive representation of stochastic noise mechanisms, while alternative, non-Markovian representations are also possible. Building on this taxonomy, the analysis concentrates on modifications of the underlying probability distribution, emphasizing data as a process rather than a static object. Variations in problem components are formalized probabilistically via Markov

kernels, that is, conditional distributions. The effect of such modifications is studied through equivalences between learning problems, characterized in terms of their Bayes risk. Comparing the equivalence classes generated by different corruption types allows *qualitative* comparison of noise types, which partially answers our initial question on whether certain corruptions can be considered as more detrimental than others.

Part 3: Comparison of Problems via Inequalities The third part focuses on finding ways to *quantitatively* compare learning problems. We identify here classes of corruptions that are detrimental when performance is measured via generalized entropy, itself induced by a chosen loss and model class (§ 7.2). We arrive at these conclusions using the set-based view of learning problems, and building on top of information-theoretical and statistical results. Mathematically, the results are in the form of inequalities, which allow to give a relative answer to the question of when does a corruption worsen a learning problem: it depends on the loss and model chosen. Specific families of learning problems will be negatively affected by specific classes of kernels, i.e., types of dark data. Again, we find interesting relations between some characteristics of the learning problem and the data at hand. Moreover, beyond focusing on data, we remark that the second part of the thesis contributes inequalities that enable direct comparison of loss functions and model classes, independent of any specific dataset or corruption mechanism. These comparisons take the form of Bayesian orderings, extending concepts from the theory of comparison of experiments to the present learning framework.

The results in this thesis have been developed in collaboration with Robert C. Williamson, my supervisor, as well as Dr. Nan Lu for Part II (Iacovissi et al., 2026b) and Rabanus Derr for § 3.3 and § 7 (Iacovissi et al., 2026a). All results are original unless stated otherwise; key concepts are emphasized with a gray box.

Part I

Abstracting a Learning Problem

Chapter 2

Mathematical Tools

This chapter fixes notation and introduces existing tools needed for the development of the thesis. First, we give a refresher of measure and probability theory, but only for the purpose of serving the later chapters. We assume that the reader is already familiar with these concepts, as they will not be analyzed in depth. Later, we shift our focus to presenting key definitions and properties about Markov kernels, which are included in the second section. These compose the main building block for our theory, therefore we review them more carefully and highlight important statements with a gray box. For a complete presentation of both topics, refer to Aliprantis and Border (2006) and Çinlar (2011).

2.1 Measure and Probability Theory

Sets and Topology

Definition 2.1 (Topology)**.** *Let $\mathcal{Z}$ be a set. A* topology $\mathfrak{T}$ *on $\mathcal{Z}$ is a collection of subsets of $\mathcal{Z}$ (called* open sets*) such that:*

1. $\emptyset \in \mathfrak{T}$ *and* $\mathcal{Z} \in \mathfrak{T}$,
2. *Any union of sets in $\mathfrak{T}$ is also in $\mathfrak{T}$: if $\{\mathcal{A}_i\}_{i\in I} \subseteq \mathfrak{T}$, then $\bigcup_{i\in I} \mathcal{A}_i \in \mathfrak{T}$,*
3. *Any finite intersection of sets in $\mathfrak{T}$ is also in $\mathfrak{T}$: if $\mathcal{A}_1, \dots, \mathcal{A}_n \in \mathfrak{T}$, then $\bigcap_{i=1}^n \mathcal{A}_i \in \mathfrak{T}$.*

A *topological space* is a pair $(\mathcal{Z}, \mathfrak{T})$ of a set and a topology on it. A function $f\colon \mathcal{Z}_1 \to \mathcal{Z}_2$ between two topological spaces $(\mathcal{Z}_1, \mathfrak{T}_1)$, $(\mathcal{Z}_2, \mathfrak{T}_2)$ is *continuous* if, for every open set $\mathcal{A} \in \mathfrak{T}_2$, the inverse image

$$f^{-1}(\mathcal{A}) = \{z \in \mathcal{Z}_1 \mid f(z) \in \mathcal{A}\}$$

is an open set in $\mathfrak{T}_1$.

A *topological vector space* is a vector space $\mathcal{Z}$ over $\mathbb{R}$, equipped with a topology $\mathfrak{T}$ that makes the vector addition and scalar multiplication maps continuous. In this thesis, we simply

refer to it as a vector space and specify the topology when not apparent.

A subset $\mathcal{X} \subseteq \mathcal{Z}$ is called *convex* if for every pair of points $y, w \in \mathcal{X}$ and every $\lambda \in [0,1]$, the point $\lambda y + (1-\lambda)w$ also belongs to $\mathcal{X}$.

The *convex hull* of a set $\mathcal{Z} \subseteq \mathbb{R}^d$, denoted $\text{co}(\mathcal{Z})$, is the smallest convex set containing $\mathcal{Z}$. Equivalently, it consists of all finite convex combinations of points in $\mathcal{Z}$:

$$\text{co}(\mathcal{Z}) = \left\{ \sum_{i=1}^{n} \lambda_i z_i : n \in \mathbb{N},\ z_i \in \mathcal{Z},\ \lambda_i \geq 0,\ \sum_{i=1}^{n} \lambda_i = 1 \right\}.$$

The *closure* of a set $\mathcal{Z}$ with respect to the topology $\mathfrak{T}$ is the intersection of all sets in $\mathfrak{T}$ that contain $\mathcal{Z}$, i.e., $\overline{\mathcal{Z}} \coloneqq \bigcap\{\mathcal{A} \in \mathfrak{T} \mid \mathcal{Z} \subseteq \mathcal{A}\}$. A set is closed if it is equal to its closure.

A subset $\mathcal{X} \subseteq \mathcal{Z}$ is called *compact* with respect to the topology $\mathfrak{T}$ if for every collection $\{\mathcal{A}_i\}_{i\in I} \subseteq \mathfrak{T}$ such that $\mathcal{X} \subseteq \bigcup_{i\in I} \mathcal{A}_i$, there exists a finite subcollection $\{i_1, \ldots, i_n\} \subset I$ such that $\mathcal{X} \subseteq \mathcal{A}_{i_1} \cup \cdots \cup \mathcal{A}_{i_n}$.

We denote the *Minkowski sum* as $\oplus$, defined as: let $\mathcal{Z}, \mathcal{Z}'$, then $\mathcal{Z} \oplus \mathcal{Z}' \coloneqq \{z + z' \mid z \in \mathcal{Z}, z' \in \mathcal{Z}'\}$.

We use the notation $\alpha \cdot \mathcal{Z}$ or $\alpha\mathcal{Z}$ to indicate the scaled set $\{\alpha z \mid z \in \mathcal{Z}\}$. When $\alpha = -1$, we simply write $-\mathcal{Z}$.

Finally, a point $z \in \mathcal{Z}$ is called an *extreme point* of $\mathcal{Z}$ if it cannot be expressed as a nontrivial convex combination of other points in $\mathcal{Z}$. Formally, z is extreme if for any $y, w \in \mathcal{Z}$ and any $\lambda \in (0,1)$, $z = \lambda y + (1-\lambda)w \implies y = w = z$. The set of all extreme points of $\mathcal{Z}$ is denoted $\text{ext}(\mathcal{Z}) \coloneqq \{z \in \mathcal{Z} : z \text{ is an extreme point of } \mathcal{Z}\}$.

Probability Spaces

Definition 2.2 (Algebra)**.** *Given a set $\mathcal{Z}$, an* algebra $\mathfrak{A}$ *is a collection of subsets of $\mathcal{Z}$ such that*

1. $\mathcal{Z} \in \mathfrak{A}$ *and* $\emptyset \in \mathfrak{A}$*;*
2. *If $\mathcal{X} \in \mathfrak{A}$, then $\mathcal{X}^{\text{c}} \coloneqq \mathcal{Z} \setminus \mathcal{X} \in \mathfrak{A}$;*
3. *For any finite collection of sets $\{\mathcal{Z}_i \in \mathfrak{A}\}_{i\in I\subsetneq\mathbb{N}}$, it holds that $\bigcup_{i\in I} \mathcal{Z}_i \in \mathfrak{A}$.*

Definition 2.3 (σ-algebra)**.** *Given a set $\mathcal{Z}$, a* σ-algebra Σ *is a collection of subsets of $\mathcal{Z}$ such that*

1. Σ *is an algebra;*
2. *For any countable collection of sets $\{\mathcal{Z}_i \in \Sigma\}_{i\in I\subseteq\mathbb{N}}$, it holds that $\bigcup_{i\in I} \mathcal{Z}_i \in \Sigma$.*

We refer to the smallest σ-algebra including all sets in a family of sets Ω as $\Sigma(\Omega)$, while to

the smallest σ-algebra containing a set $\mathcal{Z}$ is referred to as $\Sigma(\mathcal{Z})$. If $\mathcal{Z}$ is a topological space with open sets $\Omega_{\mathcal{Z}}$, then $\Sigma(\Omega_{\mathcal{Z}})$ is the *Borel σ-algebra* on $\mathcal{Z}$. In the following, we will only use the notation $\Sigma(\mathcal{Z})$ for the Borel σ-algebra, as we will mostly use it in our setting.

An ordered pair $(\mathcal{Z}, \Sigma)$, consisting of a set $\mathcal{Z}$ and a σ-algebra Σ on it, is called *measurable space*. When $\mathcal{Z}$ is a separable, completely metrizable topological space, it is said to be a *Polish space*; if $\mathcal{Z}$ is a Polish space and $\Sigma(\mathcal{Z})$ the associated Borel σ-algebra generated by the topological space $(\mathcal{Z}, \mathfrak{T})$, the measurable space $(\mathcal{Z}, \Sigma(\mathcal{Z}))$ is said to be a *standard Borel measurable space*.

Let $(\mathcal{Z}, \Sigma)$ and $(\mathcal{Z}', \Sigma')$ be measurable spaces. The *product σ-algebra* on $\mathcal{Z} \times \mathcal{Z}'$ is the σ-algebra $\Sigma \otimes \Sigma' := \Sigma\big(\{\mathcal{A} \times \mathcal{B} \mid \mathcal{A} \in \Sigma, \mathcal{B} \in \Sigma'\}\big)$, i.e., the smallest σ-algebra on $\mathcal{Z} \times \mathcal{Z}'$ that contains all measurable rectangles $\mathcal{A} \times \mathcal{B}$.

When $(\mathcal{Z}, \Sigma)$ and $(\mathcal{Z}', \Sigma')$ are standard Borel spaces, the product measurable space $(\mathcal{Z} \times \mathcal{Z}',\ \Sigma \otimes \Sigma')$ is again a standard Borel space (Srivastava, 1998, Proposition 3.1.23).

Definition 2.4 (Σ-Measurable Function)**.** *Let $(\mathcal{Z}, \Sigma)$ and $(\mathcal{Z}', \Sigma')$ be measurable spaces. A function $f : \mathcal{Z} \to \mathcal{Z}'$ is said to be measurable if for every $\mathcal{A} \in \Sigma'$ the pre-image of $\mathcal{A}$ under f is in Σ; that is, for all $\mathcal{A} \in \Sigma'$,*

$$f^{-1}(\mathcal{A}) := \{z \in \mathcal{Z} \mid f(x) \in \mathcal{A}\} \in \Sigma.$$

Definition 2.5 (Finitely Additive Measures)**.** *Let $(\mathcal{Z}, \mathfrak{A})$ be an ordered pair of a set and an algebra defined on it.*

- *Consider the map $\mu : \mathfrak{A} \to \mathbb{R}$ such that $\mu(\emptyset) = 0$. If given a finite collection of pairwise disjoint sets $\{\mathcal{Z}_i \in \Sigma,\ i \in I \subsetneq \mathbb{N} \mid \mathcal{Z}_i \cap \mathcal{Z}_j = \emptyset\ \forall i \neq j\}$ it holds that $\mu\big(\bigcup_{i\in I} \mathcal{Z}_i\big) = \sum_{i\in I} \mu(\mathcal{Z}_i)$, we call μ a* finitely additive signed measure*;*

- *If μ is a finitely additive signed measure such that $\mu : \mathfrak{A} \to \mathbb{R}_{\geq 0}$, then we call it a* finitely additive measure*;*

- *If ϕ is a finitely additive measure such that $\phi : \mathfrak{A} \to [0, 1]$ and $\phi(\mathcal{Z}) = 1$, then we call it a* finitely additive probability measure (or distribution)*.*

For a complete treatment of finitely additive measures, we refer the reader to (Rao and Rao, 1983).

Definition 2.6 (Countably Additive Measures)**.** *Let $(\mathcal{Z}, \Sigma)$ be a measurable space.*

- *Consider the map $\mu : \Sigma \to \mathbb{R}$ such that $\mu(\emptyset) = 0$. If given a countable collection of pairwise disjoint sets $\{\mathcal{Z}_i \in \Sigma,\ i \in I \subset \mathbb{N} \mid \mathcal{Z}_i \cap \mathcal{Z}_j = \emptyset\ \forall i \neq j\}$ it holds that $\mu\big(\bigcup_{i\in I} \mathcal{Z}_i\big) = \sum_{i\in I} \mu(\mathcal{Z}_i)$, we call μ a* countably additive probability measure*;*

- *If μ is a countably additive signed measure such that $\mu\colon \Sigma \to \mathbb{R}_{\geq 0}$, then we call it a* countably additive measure*;*

- *If ϕ is a countably additive measure such that $\phi\colon \Sigma \to [0,1]$ and $\phi(\mathcal{Z}) = 1$, then we call it a* countably additive probability measure (or distribution).

The set of all countably additive probability measures is denoted $\Delta_{\text{ca}}(\mathcal{Z})$, and the set of finitely additive probabilities is $\Delta_{\text{ba}}(\mathcal{Z})$; we omit the ($\sigma$-)algebra as it is assumed to be the Borel σ-algebra in both cases, with open sets identified by an order-induced topology when the set $\mathcal{Z}$ is made of (vectors of) real numbers.[1] Trivially, we get that $\Delta_{\text{ba}}(\mathcal{Z}) \supset \Delta_{\text{ca}}(\mathcal{Z})$.

Definition 2.7 (Absolute Continuity)**.** *Let $(\mathcal{Z}, \Sigma)$ be a measurable space, and let $\mu, \nu\colon \Sigma \to \mathbb{R}$ be countably additive signed measures. We say that μ is* absolutely continuous *with respect to ν, and write $\mu \ll \nu$ when $\nu(\mathcal{A}) = 0 \implies \mu(\mathcal{A}) = 0 \quad \forall \mathcal{A} \in \Sigma$.*

Theorem 2.8 (Radon–Nikodym Theorem for Signed Measures)**.** *Let $(\mathcal{Z}, \Sigma)$ be a measurable space. Let $\mu, \nu\colon \Sigma \to \mathbb{R}$ be countably additive signed measures such that ν is σ-finite and $\mu \ll \nu$. Then there exists a Σ-measurable function $f\colon \mathcal{Z} \to \mathbb{R}$ such that for all $\mathcal{A} \in \Sigma$,*

$$\mu(\mathcal{A}) = \int_{\mathcal{A}} f \, d\nu.$$

The function f is unique ν-almost everywhere, and is called the Radon–Nikodym derivative *of μ with respect to ν, denoted $f = \frac{d\mu}{d\nu}$.*

A *probability space* is a triple $(\mathcal{Z}, \mathfrak{A}, \phi)$; commonly, $\phi\colon \mathfrak{A} \to [0,1]$ is assumed to be a countably additive probability measure and therefore $\mathfrak{A}$ is required to be a σ-algebra. When ϕ is only finitely additive, we call the space a *finitely additive probability space*.

Let $(\mathcal{Z}, \Sigma, \phi)$ be a probability space and $(\mathcal{E}, \Sigma')$ a measurable space. A mapping $Z\colon \mathcal{Z} \to \mathcal{E}$ is called a *random variable* taking values in $(\mathcal{E}, \Sigma')$, provided that it is measurable relative to Σ and Σ', i.e., for all $\mathcal{A} \in \Sigma'$, $Z^{-1}(\mathcal{A}) \coloneqq \{Z \in \mathcal{A}\} = \{z \in \mathcal{Z} \mid Z(z) \in \mathcal{A}\}$ is an element in the σ-algebra Σ. If Σ' is understood from context, then we merely say that Z takes values in, or is defined on, $\mathcal{E}$.

Definition 2.9 (Coupling of Probability Spaces)**.** *Consider two probability spaces $(\mathcal{Z}_1, \Sigma(\mathcal{Z}_1), \phi_1)$ and $(\mathcal{Z}_2, \Sigma(\mathcal{Z}_2), \phi_2)$, s.t. $\phi_i \in \Delta_{\text{ca}}(\mathcal{Z}_i)$, $i \in \{1,2\}$. A coupling is a joint probability space $(\mathcal{Z}_1 \times \mathcal{Z}_2, \Sigma(\mathcal{Z}_1) \times \Sigma(\mathcal{Z}_2), \phi)$, such that, for all $\mathcal{A} \in \Sigma(\mathcal{Z}_2)$ and $\mathcal{B} \in \Sigma(\mathcal{Z}_1)$,*

$$\phi_2(\mathcal{A}) = \int_{\mathcal{A}} \int_{\mathcal{Z}_1} \phi(dz_1, dz_2);$$
$$\phi_1(\mathcal{B}) = \int_{\mathcal{B}} \int_{\mathcal{Z}_2} \phi(dz_1, dz_2).$$

[1] In the case of vector spaces, we consider the product topology of the order topology on $\mathbb{R}$. That being, the Euclidean topology.

2.2 Duality Between Functions and Measures

Definition 2.10 (Dual pair, Definition 5.90 in Aliprantis and Border (2006))**.** *A dual pair is a pair of vector spaces $\mathcal{Z}, \mathcal{Z}'$ together with a bilinear functional $(z, z') \to \langle z, z' \rangle$ from $\mathcal{Z} \times \mathcal{Z}'$ to $\mathbb{R}$ such that:*

1. *$z \to \langle z, z' \rangle$ is linear for each $z' \in \mathcal{Z}'$;*
2. *$z' \to \langle z, z' \rangle$ is linear for each $z \in \mathcal{Z}$;*
3. *If $\langle z, z' \rangle = 0 \ \forall z' \in \mathcal{Z}'$, then $z = 0$;*
4. *If $\langle z, z' \rangle = 0 \ \forall z \in \mathcal{Z}$, then $z' = 0$.*

Each component space involved in a dual pair can be interpreted as a set of linear functionals on the other. Conditions 1 and 2 are the ones required for the definition of a *bilinear functional*. The bilinear functional $(z, z') \mapsto \langle z, z' \rangle$ is also called the *duality* of the dual pair.

An example of dual pair in a finite dimensional setting is a vector space $\mathcal{X}$ and its algebraic dual $\mathcal{X}^*$, and $\langle x, x^* \rangle = x^*(x)$. In the infinite dimensional setting, the algebraic dual and the topological dual (the one we call here simply as *dual*) do not generally coincide.

The set of function $f : \mathcal{Z} \to \mathbb{R}$ that are *bounded*, i.e., $\exists C \in \mathbb{R}_{\geq 0} \mid |f(z)| \leq C, \forall z \in \mathcal{Z}$, and $\Sigma(\mathcal{Z})$-*measurable* is denoted as $\mathcal{B}_b(\mathcal{Z})$. This set together with the supremum norm,

$$\|f\|_\infty = \sup_{z \in \mathcal{Z}} |f(z)|$$

for $f \in \mathcal{B}_b(\mathcal{Z})$ forms a Banach space (Schechter, 1997, page 804).

The norm dual, i.e., the *set of all continuous, linear functionals* on $\mathcal{B}_b(\mathcal{Z})$ coincides with $\text{ba}(\mathcal{Z})$, the space of all finitely additive, signed measures $\mu : (\mathcal{Z}, \Sigma(\mathcal{Z})) \to \mathbb{R}$ with bounded total variation $\|\mu\|_\infty$ (Aliprantis and Border, 2006, Theorem 14.4). That is defined as

$$\|\mu\|_\infty := \sup_{\mathcal{A} \in \Sigma(\mathcal{Z})} |\mu(\mathcal{A})|,$$

equivalently to the supremum norm. Notice that the space $\text{ba}(\mathcal{Z})$ normed with $\|\mu\|_\infty$ forms a Banach space (Schechter, 1997, 29.6.c & 29.29.f).

Proposition 2.11 (Theorem 14.4 in Aliprantis and Border (2006))**.** *Every continuous linear functional on $F : \mathcal{B}_b(\mathcal{Z}) \to \mathbb{R}$ can be written as $F : f \mapsto \int f \, d\mu$ for a unique $\mu \in \text{ba}(\mathcal{Z})$.*

Proposition 2.12 (Theorem 11.8 in Aliprantis and Border (2006) and page 804 of Schechter (1997))**.** *The integral $\int f \, d\mu$ is well-defined, furthermore, it defines a continuous bilinear functional*

of the sets $\mathcal{B}_b(\mathcal{Z}) \times \mathrm{ba}(\mathcal{Z}) \to \mathbb{R}$, *as*

$$\langle f, \mu \rangle \coloneqq \int f d\mu \,.$$

It follows that $\langle \mathcal{B}_b(\mathcal{Z}), \mathrm{ba}(\mathcal{Z}) \rangle$ forms a *dual pair* by (Schechter, 1997, 28.4 HB22), which applies here as the two spaces $\mathcal{B}_b(\mathcal{Z}), \mathrm{ba}(\mathcal{Z})$ are Banach spaces and therefore fulfill the hypothesis by (Schechter, 1997, 16.1 & 26.1).

Relevant Topologies and Sets

When working with dual pairs, it is standard and helpful to consider a weak re-topologization of the sets $\mathcal{B}_b$ and ba (Aliprantis and Border, 2006, Chapter 15). We define $\sigma(\mathcal{B}_b, \mathrm{ba})$ as the topology on $\mathcal{B}_b(\mathcal{Z})$ which makes every functional $f \mapsto \langle f, \mu \rangle$ (for some $\mu \in \mathrm{ba}(\mathcal{Z})$) continuous. Analogously, $\sigma(\mathrm{ba}, \mathcal{B}_b)$ is the topology on $\mathrm{ba}(\mathcal{Z})$ which makes every functional $\mu \mapsto \langle f, \mu \rangle$ (for some $f \in \mathcal{B}_b(\mathcal{Z})$) continuous. The topologies are called weak as they are contained within the norm-topologies of the respective spaces (Schechter, 1997, 28.13.b & 28.22.a).

Every closed convex subset of $\mathcal{B}_b$ in the norm topology is $\sigma(\mathcal{B}_b, \mathrm{ba})$-closed, and vice-versa (Schechter, 1997, 28.14 HB24).

We introduce the notation $\mathrm{ca}(\mathcal{Z})$ to denote the set of all countably additive, signed measures with bounded total variation. Notice that we can give an alternative definition of the set of probability measures as,

$$\Delta_{\mathrm{ca}}(\mathcal{Z}) \coloneqq \{\phi \in \mathrm{ca}(\mathcal{Z}) \colon \phi \geq 0, |\phi(\mathcal{Z})| = 1\},$$

and similarly for finitely additive probability measures,

$$\Delta_{\mathrm{ba}}(\mathcal{Z}) \coloneqq \{\phi \in \mathrm{ba}(\mathcal{Z}) \colon \phi \geq 0, |\phi(\mathcal{Z})| = 1\}.$$

For $\phi \in \Delta_{\mathrm{ba}}(\mathcal{Z})$ we denote the expectation of a function $f \in \mathcal{B}_b(\mathcal{Z})$ as,

$$\mathbb{E}_\phi[f] \coloneqq \langle f, \phi \rangle.$$

The following property seems to be a known fact, as it is referenced, without proof, in Walley (1991, Appendix D) and Duanmu and Weiss (2018). As we have not found any proof, we provide one here.[2]

Lemma 2.13 (Closure of Countably Additive Probabilities are Finitely Additive Probabilities)**.** *Let* $\mathcal{Z}$ *be a Polish space. Then, with respect to the* $\sigma(\mathcal{B}_b, \mathrm{ba})$*-topology,* $\overline{\Delta_{\mathrm{ca}}(\mathcal{Z})} = \Delta_{\mathrm{ba}}(\mathcal{Z})$.

Proof. Let $\mathrm{ext}\,\mathcal{Z}$ be the set of extreme points of a set. The set $\Delta_{\mathrm{ba}}(\mathcal{Z})$ is compact as it is a closed subset (Walley, 1991, Appendix D4) of the compact norm-ball. The norm is the

[2] We thank Rabanus Derr for the proof.

supremum norm and the ball is compact by Banach-Alaoglu-Bourbaki-Theorem (Schechter, 1997, 28.29.UF28). By the Krein-Milman-Theorem (Aliprantis and Border, 2006, 7.68) and (Schechter, 1997, 16.1 & 26.1) we know that $\Delta_{\text{ba}}(\mathcal{Z}) = \overline{\text{co}}\,\text{ext}\,\Delta_{\text{ba}}(\mathcal{Z})$. The extreme points of the $\Delta_{\text{ba}}(\mathcal{Z})$, i.e., $\text{ext}\,\Delta_{\text{ba}}(\mathcal{Z})$, is equal to the set of all Dirac-measures (Aliprantis and Border, 2006, Theorem 15.9). The convex closure of those extreme points $\text{co}\,\text{ext}\,\Delta_{\text{ba}}(\mathcal{Z})$ is a subset of the countably additive probability measures $\Delta_{\text{ca}}(\mathcal{Z})$, as all Dirac measures are countably additive, and the set $\Delta_{\text{ca}}(\mathcal{Z})$ is convex. Thus, $\overline{\Delta_{\text{ca}}(\mathcal{Z})} = \Delta_{\text{ba}}(\mathcal{Z})$. □

2.3 Markov Kernels

Markov kernels provide a formal and general framework for modeling probabilistic transformations between measurable spaces. They extend the notion of stochastic maps beyond finite or discrete settings, enabling a consistent and rigorous treatment of conditional probability in continuous or uncountable spaces.

From a probabilistic viewpoint, a Markov kernel represents a family of probability measures, parameterized by the elements of a source space. Each input determines a distribution over the target space, formalizing the concept of a conditional law. In this sense and under certain conditions (which are met in our framework), they coincide with regular conditional distributions, providing the standard measure-theoretic framework for conditional probability (Çinlar, 2011). Beyond abstract probability, Markov kernels underpin the formal representation of information-processing and decision-making mechanisms (Shannon, 1948; Blackwell, 1953; Csiszár, 1972). In information and control theory, they are used to model *observation channels*; in statistical decision theory, they formalize *statistical experiments*. In both cases, they define a way of generating data conditioned on underlying states or signals. We formally talk about these concepts in § 3.

Definition 2.14 (Transition and Markov Kernels)**.** *Let* $(\mathcal{Z}_1, \Sigma_1)$ *and* $(\mathcal{Z}_2, \Sigma_2)$ *be standard Borel. Let* $\kappa\colon \mathcal{Z}_1 \times \Sigma_2 \to [0, +\infty)$*. Then,* κ *is called a* (transition) kernel *from* $(\mathcal{Z}_1, \Sigma_1)$ *to* $(\mathcal{Z}_2, \Sigma_2)$ *if*

1. *the mapping* $z_1 \mapsto \kappa(z_1, B)$ *is* Σ_1*-measurable for every set* $B \in \Sigma_2$*;*
2. *the mapping* $B \mapsto \kappa(z_1, B)$ *is a measure on* $(\mathcal{Z}_2, \Sigma_2)$ *for every* $z_1 \in \mathcal{Z}_1$*.*

When the second mapping induces a probability distribution, then κ *is a* Markov kernel*.*

We refer to the set of transition kernels as $\mathcal{T}(\mathcal{Z}_1, \mathcal{Z}_2)$, while $\mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ is its subset of Markov kernels. In addition, we denote the *domain* and the *image* spaces of such a kernel, respectively as

$$D(\kappa) = \mathcal{Z}_1 \quad \text{and} \quad I(\kappa) = \mathcal{Z}_2.$$

Recalling how we introduced the concept of Markov kernel, one can see it as a parameterized

family $\kappa(z_1, \cdot)\,, z_1 \in \mathcal{Z}_1$ of probability measures on the space $(\mathcal{Z}_2, \Sigma(\mathcal{Z}_2))$. In this context, a Markov kernel serves as a detailed probabilistic description of the generative process leading from a "hidden value" $\mathcal{Z}_1$ to *observed* distribution on $\mathcal{Z}_2$.[3] As such, for finite spaces they can be represented as stochastic matrices.

Example 2.15. *Consider a set $\mathcal{Z} = \{0, 1\}$, a kernel $\kappa \in \mathcal{M}(\mathcal{Z}, \mathcal{Z})$, and random variables $Z, \tilde{Z}$ on the probability space $(\mathcal{Z}, \Sigma(\mathcal{Z}), \phi)$. We can conveniently represent the kernel with the a matrix whose entries are the associated conditional density ψ's values:*

$$\kappa = \begin{bmatrix} \psi(\tilde{Z} = 1 \,|\, Z = 1) & \psi(\tilde{Z} = 1 \,|\, Z = 0) \\ \psi(\tilde{Z} = 0 \,|\, Z = 1) & \psi(\tilde{Z} = 0 \,|\, Z = 0) \end{bmatrix} = \begin{bmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{bmatrix}.$$

Associated with a Markov kernel are the functionals called *Markov transition*, *Markov operator* and *Markov operator adjoint*. We define them and provide basic properties in the following.

Proposition 2.16 (Markov Transition Function, Theorem 19.13, Aliprantis and Border (2006))**.** *Let κ be a Markov kernel from $(\mathcal{Z}_1, \Sigma_1)$ to $(\mathcal{Z}_2, \Sigma_2)$. The function*

$$\dot{\kappa} \colon \mathcal{Z}_1 \to \Delta_{\text{ca}}(\mathcal{Z}_2), \qquad z_1 \mapsto \dot{\kappa}(z_1) \coloneqq \kappa(z_1, \cdot),$$

is known as Markov transition*, and it is measurable with respect to $\Sigma(\Delta_{\text{ca}}(\mathcal{Z}_2))$, the Borel-$\sigma$-algebra induced by the $\sigma(\text{ba}, \mathcal{B}_b)$-topology on $\Delta_{\text{ca}}(\mathcal{Z}_2)$.*

Remark 2.17 *An alternative and compact notation for a Markov kernel $\kappa \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$, reminding us of its role in defining a Markov transition functional, is*

$$\kappa \colon \mathcal{Z}_1 \rightsquigarrow \mathcal{Z}_2.$$

Once we establish the concept of Markov transition, we can define the related operator and adjoint operator. These functionals are sometimes also referred to as Markov kernel *actions*, as they allow the object to be combined with functions and probabilities.

Proposition 2.18 (Kernel Actions)**.** *Let $\dot{\kappa} \colon \mathcal{Z}_1 \to \Delta_{\text{ca}}(\mathcal{Z}_2)$ be a Markov transition corresponding to a Markov kernel κ.*

1. *The* Markov transition operator *$\hat{\kappa} \colon \mathcal{B}_b(\mathcal{Z}_2) \to \mathcal{B}_b(\mathcal{Z}_1)$ is the mapping*

$$f \mapsto \hat{\kappa} f \quad \text{s.t.} \quad z_1 \in \mathcal{Z}_1 \mapsto (\hat{\kappa} f)(z_1) \coloneqq \int_{\mathcal{Z}_2} f(z_2)\, \kappa(z_1, dz_2);$$

[3] In fact, under our assumptions they are the *regular conditional probabilities* associated to the coupling of the spaces $(\mathcal{Z}_1, \Sigma(\mathcal{Z}_1))$ and $(\mathcal{Z}_2, \Sigma(\mathcal{Z}_2))$ (see Çinlar, 2011).

2. *The norm adjoint of $\hat{\kappa}$, denoted $\check{\kappa}\colon \mathrm{ba}(\mathcal{Z}_1) \to \mathrm{ba}(\mathcal{Z}_2)$,*

$$\mu \mapsto \check{\kappa}\mu \quad s.t. \quad B \in \Sigma_2 \mapsto (\check{\kappa}\mu)(B) \coloneqq \int_{\mathcal{Z}_1} \kappa(z_1, B) d\mu.$$

In particular, if $\mu \in \mathrm{ca}(\mathcal{Z}_1)$, then $\check{\kappa}\mu \in \mathrm{ca}(\mathcal{Z}_2)$. If $\mu \in \Delta_{\mathrm{ca}}(\mathcal{Z}_1)$, then $\check{\kappa}\mu \in \Delta_{\mathrm{ca}}(\mathcal{Z}_2)$.

Proof. The first statement is proven in Theorem 19.7 in (Aliprantis and Border, 2006). The second statement is proven in Theorem 19.9, point 2 in (Aliprantis and Border, 2006). □

Remark 2.19 *Proposition 2.18 states and proves that $\check{\kappa}$ is the norm adjoint of $\hat{\kappa}$, for which the defining property is that*

$$\langle f, \check{\kappa}\mu \rangle_{\mathcal{Z}_2} = \langle \hat{\kappa} f, \mu \rangle_{\mathcal{Z}_1} \quad \forall f \in \mathcal{B}_b(\mathcal{Z}_2), \mu \in \mathrm{ba}(\mathcal{Z}_1).$$

This observation will be particularly useful in the context of learning theory.

Notice that we can rewrite the definition Markov transition operator and its adjoint using the dual pairing, i.e.,

$$z_1 \in \mathcal{Z}_1 \mapsto (\hat{\kappa} f)(z_1) = \int_{\mathcal{Z}_2} f(z_2)\, \kappa(z_1, dz_2) = \langle f, \dot{\kappa}(z_1) \rangle_{\mathcal{Z}_2},$$
$$\mathcal{B} \in \Sigma_2 \mapsto (\check{\kappa}\mu)(\mathcal{B}) = \int_{\mathcal{Z}_1} \kappa(z_1, \mathcal{B}) d\mu = \langle \kappa(z_1, \mathcal{B}), \mu \rangle_{\mathcal{Z}_1} = \langle \dot{\kappa}(z_1)(\mathcal{B}), \mu \rangle_{\mathcal{Z}_1}.$$

Plugging this in Remark 2.19, we can write for all $f \in \mathcal{B}_b(\mathcal{Z}_2)$ and $\mu \in \mathrm{ba}(\mathcal{Z}_1)$

$$\langle \langle f, \dot{\kappa}(z_1) \rangle_{\mathcal{Z}_2}, \mu \rangle_{\mathcal{Z}_1} = \langle f, \langle \dot{\kappa}(z_1), \mu \rangle_{\mathcal{Z}_1} \rangle_{\mathcal{Z}_2}.$$

Lastly, we give here some definitions, mostly of notational relevance, that will come in hand in future derivations.

Definition 2.20 (Kernel Action on a Set of Functions)**.** *Let $\hat{\kappa}$ be a Markov transition operator associated to the Markov transition $\dot{\kappa}\colon \mathcal{Z}_1 \to \Delta_{\mathrm{ca}}(\mathcal{Z}_2)$. We define the action of the kernel on a set as,*

$$\hat{\kappa}\mathcal{F} \coloneqq \{\hat{\kappa} f,\ f \in \mathcal{F}\}.$$

Definition 2.21 (Deterministic Markov Kernel)**.** *A* deterministic *Markov kernel $\kappa_f\colon \mathcal{Z}_1 \times \Sigma_2 \to [0,1]$ induced by some measurable function $f\colon \mathcal{Z}_1 \to \mathcal{Z}_2$, is defined as*

$$(z_1, \mathcal{A}) \mapsto \kappa_f(z_1, \mathcal{A}) \coloneqq \delta_{f(z_1)}(\mathcal{A}),$$

where δ denotes the Dirac distribution on x, *i.e.*,

$$\mathcal{A} \mapsto \delta_x(\mathcal{A}) := \begin{cases} 1 & \text{if } x \in \mathcal{A}, \\ 0 & \text{if } x \notin \mathcal{A}. \end{cases}$$

As a consequence, every measurable function $f : \mathcal{Z}_1 \to \mathcal{Z}_2$ can be represented by a kernel.

Proposition 2.22. *Let $\kappa_f \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ for some $f : \mathcal{Z}_1 \to \mathcal{Z}_2$, and $\kappa \in \mathcal{M}(\mathcal{Z}_2, \mathcal{Z}_3)$. The Markov transition associated to the kernel $\kappa' := \hat{\kappa}_f \circ \kappa$, is*

$$\dot{\kappa'} = \dot{\kappa} \circ f : \mathcal{Z}_1 \to \Delta_{\text{ca}}(\mathcal{Z}_3)\,.$$

Definition 2.23 (Degenerate Markov Kernel)**.** *Let $\nu \in \Delta_{\text{ca}}(\mathcal{Z}_2)$. We define the* degenerate *or* non-informative *kernel $\kappa_\nu \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ as the one with associate operator*

$$\dot{\kappa}_\nu : \mathcal{Z}_1 \to \Delta_{\text{ca}}(\mathcal{Z}_2)\,, \quad \dot{\kappa}_\nu(z_1) = \nu \ \forall z_1 \in \mathcal{Z}_1.$$

Operations between Markov Kernels

Kernels can be combined through different operations. We introduce them here briefly, mainly inspired by (Kallenberg, 2017; Johnston, 2023), with the aim of self-containment. We remark that specifying the kernel action operator κf for all measurable f effectively defines a kernel as $\kappa(z_1, \mathcal{B}) := (\hat{\kappa}\chi_{\mathcal{B}})(z_1)$ (Çinlar, 2011, Remark 6.4), where $\chi_{\mathcal{B}}(z_2)$ is the indicator function of $\mathcal{B} \in \Sigma(\mathcal{Z}_2)$ for $z_2 \in \mathcal{Z}_2$; for this reason, we state the following definitions using the action on a general function.

The first set of operations defined here can be referred to as *in-series operations*, given that the involved kernels are required to satisfy specific conditions on the spaces for which they are defined. These operations impose a more stringent set of feasibility conditions.

Definition 2.24 (Chain Composition)**.** *Given $\tau \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ and $\lambda \in (\mathcal{Z}_2, \mathcal{Z}_3)$, their* chain composition *$\tau \circ \lambda$ is uniquely determined by the following kernel action:*

$$\hat{\kappa} f(z_1) := (\hat{\tau} \circ \hat{\lambda}) f(z_1) := \int_{\mathcal{Z}_2} \tau(z_1, dz_2) \int_{\mathcal{Z}_3} \lambda(z_2, dz_3) f(z_3)\,,$$

for every $f : \mathcal{Z}_3 \to \mathbb{R}$ positive and $\Sigma(\mathcal{Z}_3)$-measurable.

Definition 2.25 (Product Composition)**.** *Given $\tau \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ and $\lambda \in (\mathcal{Z}_1 \times \mathcal{Z}_2, \mathcal{Z}_3)$, their* product composition *$\tau \times \lambda$ is uniquely determined by the following kernel action:*

$$\hat{\kappa} f(z_1) := (\hat{\tau} \times \hat{\lambda}) f(z_1) := \int_{\mathcal{Z}_2} \tau(z_1, dz_2) \int_{\mathcal{Z}_3} \lambda((z_1, z_2), dz_3)\, f(z_2, z_3)\,,$$

for every $f : \mathcal{Z}_2 \times \mathcal{Z}_3 \to \mathbb{R}$ *positive and* $\Sigma(\mathcal{Z}_2 \times \mathcal{Z}_3)$*-measurable.*

The operations defined above can be naturally understood using well-known probability theory results.

Consider the trivial Markov kernel τ_ν for some $\nu \in \Delta_{\text{ca}}(\mathcal{Z}_1)$. In this setting, the operations Def. 2.24 and Def. 2.25 apply to τ_ν when composed with a kernel $\lambda_1 \colon \mathcal{Z}_1 \rightsquigarrow \mathcal{Z}_2$ (for the chain composition) and $\lambda_2 \colon \{*\} \times \mathcal{Z}_1 \rightsquigarrow \mathcal{Z}_2$ (for the product composition). This allows us to write distribution-kernel combinations using the same notation as kernel-kernel ones, i.e. $\tau_\nu \circ \lambda_1$ and $\tau_\nu \times \lambda_2$.[4] Both of these constructions result in *new probability measures*:

- The composition
$$\nu \circ \lambda_1 \coloneqq \tau_\nu \circ \lambda_1 \ \in \Delta_{\text{ca}}(\mathcal{Z}_2)$$
is equivalent to the kernel action on probabilities $\check{\lambda}_1 \nu$, and corresponds to the Law of Total Probability.

- The product
$$\nu \times \lambda_2 \coloneqq \tau_\nu \times \lambda_2 \ \in \Delta_{\text{ca}}(\mathcal{Z}_1 \times \mathcal{Z}_2)$$
corresponds to the Bayesian decomposition of a joint probability into a marginal ν and a conditional probability λ_2.

Notice that λ_2 is essentially of the same type as λ_1, apart from a dummy variable over the irrelevant set $\{*\}$—it only needs to be the same as the input set of τ_ν. In what follows, we will overload the notation and write $\nu \times \lambda_1 \in \Delta_{\text{ca}}(\mathcal{Z}_1 \times \mathcal{Z}_2)$ for the product operation, and $\nu \circ \lambda_2 \in \Delta_{\text{ca}}(\mathcal{Z}_2)$ for the action of a kernel on a probability.

The second set of operations defined here can be referred to as *parallel operations*. Compared to in-series operations as in Def. 2.24 and Def. 2.25, it allows for more flexible combinations of kernels.

Definition 2.26 (Superposition Composition)**.** *Given* $\tau \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ *and* $\lambda \in \mathcal{M} \in (\mathcal{Z}_3, \mathcal{Z}_4)$*, their* superposition $\tau \otimes \lambda$ *is uniquely determined by the following kernel action:*

$$\hat{\kappa} f(z_1, z_3) \coloneqq (\hat{\tau} \otimes \hat{\lambda}) f(z_1, z_3) \coloneqq \int_{\mathcal{Z}_2} \tau(z_1, dz_2) \int_{\mathcal{Z}_4} \lambda(z_3, dz_4)\, f(z_2, z_4)\ ,$$

for every $f : \mathcal{Z}_2 \times \mathcal{Z}_4 \to \mathbb{R}$ *positive and* $\Sigma(\mathcal{Z}_2 \times \mathcal{Z}_4)$*-measurable.*

Remark 2.27 *Observe that no restriction is imposed on the parameter spaces to be equal, e.g.,* $\mathcal{Z}_1 = \mathcal{Z}_3$*, or Cartesian products with some space in common, e.g.,* $\mathcal{Z}_1 = \mathcal{Y}_1 \times \mathcal{Y}_2, \mathcal{Z}_3 = \mathcal{Y}_1 \times \mathcal{Y}_3$*. When this happens, the actions of the two kernels "superpose" on the same space. It is possible for*

[4] Strictly speaking, $\tau_\nu \circ \lambda_1 \in \mathcal{M}(\{*\}, \mathcal{Z}_2)$, while $\check{\lambda}_1 \nu \in \Delta_{\text{ca}}(\mathcal{Z}_2)$. So this identification holds up to a suitable "projection". However, we avoid this level of technicality to keep the presentation clear and simple.

the superposition operation to produce a kernel such that

$$(\hat{\tau} \otimes \hat{\lambda}) f(z_1) \coloneqq \int_{\mathcal{Z}_2} \kappa(z_1, dz_2) \int_{\mathcal{Z}_3} \lambda(z_1, dz_3)\, f(z_2, z_3),$$

with $\tau \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ *and* $\lambda \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_3)$*, so that* $\tau \otimes \lambda \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2 \times \mathcal{Z}_3)$*.*

Another possible case is $\kappa' \colon \mathcal{Z}_1 \rightsquigarrow \mathcal{Z}_2$ *and* $\lambda' \colon \mathcal{Z}_1 \times \mathcal{Z}_2 \rightsquigarrow \mathcal{Z}_3$*, leading to*

$$\kappa' \otimes \lambda' = \kappa' \times \lambda',$$

which makes the superposition a generalization of the product operation. In this case, we will use the $\times$ *symbol.*

However, in case we have more than one measure acting on the same space, the superposition integral would be ill-defined, making some combinations unfeasible.

Because of the above properties, we say that Def. 2.26 is the operation with the *weakest feasibility conditions*, the set of rules to fulfill for a well-defined operation.

The last set of operations, introduced by us, can be described as a mid-way between the chain composition Def. 2.24 and the superposition Def. 2.26.

Definition 2.28 (Partial Chain Composition)**.** *Given* $\tau \in \mathcal{M}(\mathcal{Z}_1 \times \mathcal{Z}_2, \mathcal{Z}_3)$ *and* $\lambda \in \mathcal{M}(\mathcal{Z}_1 \times \mathcal{Z}_3, \mathcal{Z}_4)$*, their* partial chain composition *is uniquely determined by the following kernel action:*

$$\hat{\kappa} f(z_1, z_2) \coloneqq (\hat{\tau} \circ_{\mathcal{Z}_3} \hat{\lambda}) f(z_1, z_2) \coloneqq \int_{\mathcal{Z}_3} \tau((z_1, z_2), dz_3) \int_{\mathcal{Z}_4} \lambda((z_1, z_3), dz_4)\, f(z_4)\ ,$$

for all $f \colon \mathcal{Z}_4 \to \mathbb{R}$ *positive and* $\Sigma(\mathcal{Z}_4)$*-measurable function.*

Essentially, this operation only chains the kernels on the specified space, here $\mathcal{Z}_3$, while superposing them on the common parameter, $\mathcal{Z}_1$.

Extreme Points of the Set of Markov kernels

In the finite case, we can prove some useful properties from Markov kernels, i.e., for rectangular column stochastic matrices. A class of related matrices are the *rectangular permutation matrices*, defined as binary matrices with exactly one entry per row allowed to be 1. We can easily obtain a generalization of Birkhoff-von Neumann theorem for Markov kernels, referring to (Cao et al., 2022) for a full, simple proof.

Proposition 2.29. *The extreme points of the set of rectangular column stochastic matrices are the transposed rectangular permutation matrices.*

Proof. It is known, e.g., (Cao et al., 2022, Theorem 1), that the extreme points of the set of row stochastic matrices are rectangular permutation matrices. In symbols, when A is a

rectangular row stochastic matrix, there exist a finite collection of rectangular permutation matrices B_i, and related coefficients $\alpha_i \in \mathbb{R}_{\geq 0}$ summing to one, such that

$$A = \sum_i \alpha_i B_i.$$

Taking the transpose of the above proves the thesis. □

Taking again the Markov kernel point of view, we note that transposed rectangular permutation matrices are in fact deterministic kernels induced by some function f, as we have the condition of exactly one 1 in each column. However, this function is not required to be injective nor surjective, as we instead would have observed with permutation matrices. For this reason, the term "permutation matrix" paired with the adjective "rectangular" may be misleading, and within this thesis we will only refer to these objects as deterministic Markov kernels.

In more general spaces, the question of finding the set of extreme points of the set of Markov kernels becomes harder to answer. However, our standard Borel setting allows for us to apply the following result.

Proposition 2.30. *The extreme points of the set of Markov kernels are all deterministic kernels, i.e.,*

$$\operatorname{ext} \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2) = \{\kappa_f \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2) \mid f : \mathcal{Z}_1 \to \mathcal{Z}_2 \text{ measurable}\}.$$

Proof. The statement is a direct consequence of (González Hernández and Hernández Lerma, 2005, Lemma 3.3) considering their unconstrained setting. □

Chapter 3

Learning Problems via Kernels and Sets

After establishing the notation for working with kernels, we are now ready to deploy the framework in the learning context. The content presented here is connected to the literature on *statistical experiments* and *decision theory* (Torgersen, 1991; Shiryaev and Spokoiny, 2000), as well as related to *machine learning theory* work (Reid and Williamson, 2011; Williamson and Cranko, 2024). We summarize the key concepts crucial to our analysis and direct readers to relevant books for a more comprehensive perspective.

3.1 The General Learning Problem

In statistical decision theory, a general decision problem can be viewed as a two-player game between *Nature* and a *decision-maker*. Here, Nature represents an unknown process that generates the observed phenomena; the decision-maker observes the said phenomena and seeks to find the optimal action for each observation within the context of a given task. Slightly more formally, Nature here stands for the (stochastic) force choosing an *observation* $o \in \mathcal{O}$ given some hidden *state* $\theta \in \Theta$. The stochastic process generating o given θ is referred to as the *experiment* E.

Definition 3.1 (Statistical Experiment)**.** *An experiment $E \in \mathcal{M}(\Theta, \mathcal{O})$ is a Markov kernel from the hidden state space $(\Theta, \Sigma(\theta))$ to the observation space $(\mathcal{O}, \Sigma(\mathcal{O}))$.*

The parameter space Θ and the observation space $\mathcal{O}$ are fixed by the setting of the decision problem. We need to specify an additional set, the decision space $\mathcal{A}$, to introduce the modeling of the decision-maker. Having observed the phenomena, the decision-maker aims to construct a *decision rule* D mapping from the observation space $\mathcal{O}$ to the action space $\mathcal{A}$. The decision-making task can be represented by the transition diagram $\Theta \overset{E}{\underset{\text{experiment}}{\rightsquigarrow}} \mathcal{O} \overset{D}{\underset{\text{decision rule}}{\rightsquigarrow}} \mathcal{A}$, where the decision rule is also modeled by a Markov kernel. Hence, D is interpreted as a *stochastic rule*, fixing a probability on the action space $\mathcal{A}$, instead of the classical deterministic view. In order to evaluate the performance of the

decision maker with respect to the optimal decision, one introduces the concept of a loss function, which induces a learning problem in the form of an optimization.

Definition 3.2 (General Learning Problem)**.** *Consider the standard Borel product space* $(\Theta \times O \times \mathcal{A},\ \Sigma(\Theta \times O \times \mathcal{A}))$*, where* $(\Theta, \Sigma(\Theta))$*,* $(O, \Sigma(O))$*, and* $(\mathcal{A}, \Sigma(\mathcal{A}))$ *denote the measurable space for parameters, observations and decisions respectively. We refer to the space* $(\Theta \times O,\ \Sigma(\Theta \times O))$ *as the* data space *on which we aim to learn. A* general learning problem *on such a product space is a pair* $(\mathfrak{L}, \mathfrak{C})$*, where* $\mathfrak{L}$ *denotes the* learning context *and* $\mathfrak{C}$ *specifies the* learning criterion*. Specifically, the learning context is defined as* $\mathfrak{L}(\Theta \times O \times \mathcal{A})$*, and it is composed by the elements* $(\ell, \mathcal{H}, \phi)$*, where:*

- $\ell \colon \Delta_{\text{ca}}(\mathcal{A}) \times \Theta \to \mathbb{R}$ *is a* loss function *on* $\Delta_{\text{ca}}(\mathcal{A}) \times \Theta$*, measurable in both its arguments,*
- $\mathcal{H} \subseteq \mathcal{M}(O, \mathcal{A})$ *is a decision class, or* model class*,*
- *and* $\phi := \pi_\theta \times E$ *is the joint* data-generating distribution*, with associated* experiment $E \in \mathcal{M}(\Theta, O)$ *modeling the stochastic observations, and* prior *distribution* $\pi_\theta \in \Delta_{\text{ca}}(\Theta)$ *modeling the likelihood of the parameters to manifest.*

The learning criterion $\mathfrak{C}$ *is chosen by the decision maker to evaluate their overall performance against Nature.*

Remark 3.3 *The correct notation for a general learning context is* $\mathfrak{L}(\Theta \times O \times \mathcal{A})$*. However, when specifying the action space is unnecessary or redundant (as it will be for supervised learning problems defined in the next section), we use the more compact notation* $\mathfrak{L}(\Theta \times O)$*.*

We highlight that many different choices of $\mathfrak{C}$ are available, some of them better studied than others. Popular examples include expected risk and minmax risk. We refrain from stating a preference in this section, and defer it to the next one.

3.2 Supervised Learning with Experiments

In the specific setup of *supervised learning*, the observation space O is the attribute space $\mathcal{X} \subseteq \mathbb{R}^d$, $d \geq 1$, while both states Θ and actions $\mathcal{A}$ correspond to the label space $\mathcal{Y}$. Then, the experiment $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ identifies a probability associated with the attribute $\mathcal{X}$ given the state $\mathcal{Y}$. Here we focus on the *classification task*, that assumes the label space to be *finite*. We define here the formal framework for this setting, first introducing some useful additional notions, and then formally defining Bayes risk and a supervised learning problem.

Definition 3.4 (Posterior Markov Kernel)**.** *A* posterior Markov kernel *for a data space* $(\mathcal{X} \times \mathcal{Y}, \Sigma(\mathcal{X} \times \mathcal{Y}))$ *is a Markov kernel* $F \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ *from the label space* $(\mathcal{Y}, \Sigma(\mathcal{Y}))$ *to the attribute space* $(\mathcal{X}, \Sigma(\mathcal{X}))$*.*

Observe that, as we noted in the previous chapter (§ 2.3), the product operation for Markov kernels can be used to write the Bayes decomposition theorem for a joint probability

$\phi \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y})$. Hence, given $F \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ and $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$, we can write

$$\phi = \pi_X \times F = \pi_Y \times E,$$

for some suitable marginal probabilities $\pi_X \in \Delta_{\text{ca}}(\mathcal{X})$ and $\pi_Y \in \Delta_{\text{ca}}(\mathcal{Y})$.[1]

Proposition 3.5. *Let $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ be a model class and $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ a bounded loss function, then for all $h \in \mathcal{H}$ with corresponding Markov transition $\dot{h} \colon \mathcal{X} \to \Delta_{\text{ca}}(\mathcal{Y})$,*

$$\ell \circ h \colon (x, y) \mapsto \ell(\dot{h}(x), y) \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}).$$

Proof. As $\dot{h}$ is measurable by construction (Proposition 2.16) and ℓ is measurable by definition, the function $(x, y) \mapsto \ell(\dot{h}(x), y)$ is measurable. Furthermore, it is bounded, as ℓ is bounded. □

The set of functions obtained through a composition of a loss and a model is written as $\ell \circ \mathcal{H} := \{\ell \circ \dot{h} \colon h \in \mathcal{H}\}$, and called the set of *attainable prediction losses*, or just *attainable predictions*.

Definition 3.6 (Risk and Bayes Risk)**.** *Consider a loss $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, a model class $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ and a joint probability distribution $\phi := \pi_Y \times E \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y})$, with $\pi_{\mathcal{Y}} \in \Delta_{\text{ca}}(\mathcal{Y})$ and $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$. Then, the* Bayes risk (**BR**) *is defined as*

$$\begin{aligned} \mathbf{BR}_{\ell,\mathcal{H}}(\pi_Y \times E) &:= \inf_{h \in \mathcal{H}} \mathbf{R}_{\pi_Y \times E}(\ell \circ h)\,, \\ \mathbf{R}_{\pi_Y \times E}(\ell \circ h) &:= \mathbb{E}_{Y \sim \pi_Y} \mathbb{E}_{X \sim \dot{E}(Y)} \ell(h(X), Y)\,. \end{aligned}$$

Here, **R** *is known as the* risk*.*

We remark that, when $\mathcal{H} = \mathcal{M}(\mathcal{X}, \mathcal{Y})$, we talk about unconstrained Bayes risk and indicate this with the simpler notation $\mathbf{BR}_\ell(\phi)$.

Definition 3.7 (Conditional Bayes Risk)**.** *Given a measurable loss $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and a label distribution $\phi \in \Delta_{\text{ca}}(\mathcal{Y})$, the conditional Bayes risk is the quantity*

$$\mathbf{CBR}_\ell(\phi) := \inf_{\psi \in \Delta_{\text{ca}}(\mathcal{Y})} \mathbb{E}_{Y \sim \phi}[\ell(\psi, Y)].$$

In the literature, this function is called "conditional" or "prior" risk depending on which probability it takes in input – a conditional or a prior one. Here we abuse the conditional Bayes risk notation to indicate both functionals, as they are clearly distinguishable when the domain is specified.

Remark 3.8 (Relation between unconstrained **BR** and **CBR**) *Let $\mathcal{Y}$ be finite and $\mathcal{H} = \mathcal{M}(\mathcal{X}, \mathcal{Y})$.*

[1] To be precise, $\pi_X \times F \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y})$ and $\pi_Y \times E \in \Delta_{\text{ca}}(\mathcal{Y} \times \mathcal{X})$ by Def. 2.25. The equality of the two decomposition holds up to a permutation. However, we avoid this level of technicality to keep the presentation clear and simple.

Consider a loss function ℓ and a Markov kernel $F \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$. We have,

$$\begin{aligned}\mathbf{BR}_\ell(\phi) &= \inf_{h \in \mathcal{M}(\mathcal{X},\mathcal{Y})} \mathbb{E}_{(X,Y)\sim\phi}[\ell \circ h] \\ &= \inf_{h \in \mathcal{M}(\mathcal{X},\mathcal{Y})} \mathbb{E}_{X\sim\pi_X}[\mathbb{E}_{Y\sim\dot{F}(X)}[\ell \circ h]] \\ &\overset{(\star)}{=} \mathbb{E}_{\pi_X}[\inf_{\psi \in \Delta_{\mathrm{ca}}(\mathcal{Y})} \mathbb{E}_{\dot{F}(X)}[\ell(\psi, Y)]] \\ &\overset{D3.7}{=} \mathbb{E}_{\pi_X}[\mathbf{CBR}_\ell(\dot{F}(X))]\,,\end{aligned}$$

where $\pi_X \in \Delta_{\mathrm{ca}}(\mathcal{X})$ and s.t. the data-generating distribution fulfills the condition $\phi = \pi_X \times F$. The equality $(\star)$ holds by (Rockafellar and Wets, 1998, Theorem 14.60). For constrained model classes, the theorem—and therefore this remark—may not hold; they should satisfy a specific property called decomposability *(Rockafellar and Wets, 1998, Definition 14.59).*[2]

We can now introduce our notion of supervised learning problem in an abstract way, but such that it resembles what machine learning theory research studies.

Definition 3.9 (Supervised Learning Problem)**.** *Let $(\mathcal{X} \times \mathcal{Y}, \Sigma(\mathcal{X} \times \mathcal{Y}))$ be a standard Borel data space, consisting of labels in $\mathcal{Y}$ and attributes in $\mathcal{X}$. A* supervised learning problem *on it is a general learning problem $(\mathfrak{L}, \mathfrak{C})$, where:*

- *the loss is $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, measurable and bounded,*
- *the model class is $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$,*
- *and the data-generating distribution is $\phi := \pi_Y \times E \in \Delta_{\mathrm{ca}}(\mathcal{X} \times \mathcal{Y})$ with prior $\pi_Y \in \Delta_{\mathrm{ca}}(\mathcal{Y})$ and experiment $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$.*

In this setting, the learning criterion is by default risk minimization, i.e., finding the optimal action $h \in \mathcal{H}$ that achieves the associated Bayes risk*. Thus, we refer to a supervised learning problem using the compact notation $\mathfrak{L}(\mathcal{X} \times \mathcal{Y})$, with $\mathfrak{C}$ implicitly given by risk minimization.*

The definition above fits in the general decision problem framework by considering the specific diagram $\mathcal{Y} \overset{E}{\rightsquigarrow} \mathcal{X} \overset{h}{\rightsquigarrow} \mathcal{Y}$, where h is a decision rule chosen in $\mathcal{M}(\mathcal{X}, \mathcal{Y})$, therefore choosing a probability on $\mathcal{Y}$ associated to a point in $\mathcal{X}$.[3]

For some cases, the formulation of a learning problem in terms of the experiment, or loss and model class, can be restrictive; for this reason, we introduce an alternative way of writing $\mathfrak{L}$ starting from the *posterior kernel* $F \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$. This can be naturally justified by considering the following simple proposition.

[2] We do not explore this here, as we do not use this result in the constrained case. The interested reader can refer to (Giner, 2009; Chancelier et al., 2022; Bùi and Combettes, 2024) for examples of conditions s.t. this property holds.

[3] That is, considering the hypothesis, or decision, as stochastic. Several techniques exist to obtain a deterministic label from a stochastic decision rule, with different consequences (Cotter et al., 2019).

Proposition 3.10. *A supervised learning problem $\mathfrak{L} = (\ell, \mathcal{H}, \phi = \pi_Y \times E)$ on the standard Borel space $(\mathcal{X} \times \mathcal{Y}, \Sigma(\mathcal{X} \times \mathcal{Y}))$ can be equivalently expressed*

1. *using the attainable predictions $\ell \circ \mathcal{H}$, i.e., as a pair $\mathfrak{L} = (\mathcal{F}, \phi)$ with $\mathcal{F} \coloneqq \ell \circ \mathcal{H}$;*
2. *using the posterior kernel, i.e., as $\phi = \pi_X \times F$ for some prior $\pi_X \in \Delta_{\mathrm{ca}}(\mathcal{X})$ on the attribute space. We will then refer to it as $\mathfrak{L} = (\ell, \mathcal{H}, \phi = \pi_X \times F)$ on the measurable space $(\mathcal{X} \times \mathcal{Y}, \Sigma(\mathcal{X} \times \mathcal{Y}))$;*
3. *using the joint distribution ϕ,* agnostic *regarding its factorization. We will then refer to it as $\mathfrak{L} = (\ell, \mathcal{H}, \phi)$ on the measurable space $(\mathcal{Z}, \Sigma(\mathcal{Z}))$.*

We refer to an $\mathfrak{L} = (\ell, \mathcal{H}, \phi = \pi_Y \times E)$ as generative, *while an $\mathfrak{L} = (\ell, \mathcal{H}, \phi = \pi_X \times F)$ is* discriminative.

Since our focus in the following will exclusively be on supervised learning problems, we will simply term them *learning problems*.

Remarks for Finitely Additive Probability Measures

For a complete treatment of finitely additive measures, we refer the reader to (Rao and Rao, 1983). In general, many nice results that hold for countably additive measures are not proved (at least, not in their exact form) for the finitely additive counterpart. Notable examples are the Radon-Nikodym theorem (Rao and Rao, 1983, Chapter 6) and the disintegration theorem. We report here some properties useful in the context of learning problems.

Definition 3.11 (Risk and Bayes Risk Operators for Finitely Additive Measures)**.** *Consider a function class $\mathcal{F} \in \mathcal{B}_b(\mathcal{Z})_{\geq 0}$, a function $f \in \mathcal{F}$, and a probability distribution $\phi \in \Delta_{\mathrm{ba}}(\mathcal{Z})$. Then, the* Bayes risk *and* risk operators *are defined as*

$$\mathbf{BR}_{\mathcal{F}}(\phi) \coloneqq \inf_{f \in \mathcal{F}} \mathbf{R}_\phi(f)\,,$$
$$\mathbf{R}_\phi(f) \coloneqq \int f \, d\phi = \langle f, \phi \rangle\,.$$

Proposition 3.12. *Def. 3.11 is well-defined and extends Def. 3.6.*

Proof. The well-posedness of risk and its infimum directly follows from the well-posedness of the integral w.r.t. a finitely additive signed measure, proved by Proposition 2.12. In particular, as $\ell \circ \mathcal{H} \subseteq \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$ and $\Delta_{\mathrm{ca}}(\mathcal{X} \times \mathcal{Y}) \subset \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$, the extension also follows. □

In this extended definition of risk, we allow the generating probability measure to be a finitely additive one. However, we keep the standard definition of loss, and therefore only admit a model class that is a set of Markov kernels, by definition countably additive. In this

way, we can make use of the dual pairing described in § 2.2, while still being allowed to use a standard Bayesian framework for inference.

Remark 3.13 *Consider a Markov kernel $\kappa \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ and its associated operator as per Proposition 2.18. The operator $\check{\kappa}$ is defined via the dual pairing relating $\mathcal{B}_b$ and* ba*. For this reason, we have*

$$\check{\kappa} \colon \mathrm{ba}(\mathcal{Z}_1) \to \mathrm{ba}(\mathcal{Z}_2)$$

and that $\check{\kappa}\,\mathrm{ba}(\mathcal{Z}_1) \subset \mathrm{ba}(\mathcal{Z}_2)$. It follows that $\check{\kappa}\Delta_{\mathrm{ba}}(\mathcal{Z}_1) \subset \Delta_{\mathrm{ba}}(\mathcal{Z}_2)$. This is a weaker statement than the one holding for countably additive probabilities, which enjoy the disintegration theorem and for which it can therefore be proved that $\check{\kappa}\Delta_{\mathrm{ca}}(\mathcal{Z}_1) = \Delta_{\mathrm{ca}}(\mathcal{Z}_2)$. We will formally prove and use this property in § 4.1.4 (when presenting properties of our taxonomy) and give additional details in Appendix C.2.1.

3.3 Supervised Learning with Superprediction Set

In the following sections, we will introduce the classical notion of superprediction set, and then present our generalization with its related properties.

The notion of superprediction has been originally introduced in the context of online learning. Its definition goes back to Vovk (1990), where the author introduced the Aggregating Algorithm for learning with expert advice. This algorithm enjoys strong theoretical guarantees for optimality; key to this result is the introduction of the notion of *superprediction set* of a loss function. In later works, people resorted again to using the superprediction set in the context of machine learning, e.g., for generalizing Vovk's Aggregating Algorithm (Cabrera Pacheco et al., 2025) and for geometrically characterizing loss functions (Williamson et al., 2016; Williamson and Cranko, 2023; Cabrera Pacheco and Williamson, 2023). Cranko (2021) further generalized the theory of the unconstrained superprediction set to a set $\mathcal{Y}$ of infinite dimension.

Definition 3.14 (Unconstrained Superprediction Set)**.** *The unconstrained superprediction set for a bounded loss function $\ell : \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is defined as*

$$\mathrm{spr}\ell := \bigcup_{\phi \in \Delta_{\mathrm{ca}}(\mathcal{Y})} \left\{ u \in \mathcal{B}_b(\mathcal{Y}) \mid \ell(\phi, y) \leq u(y) \quad \forall y \in \mathcal{Y} \right\}.$$

Remark 3.15 *Notice that our definition slightly differs from Williamson and Cranko (2023)'s, as they deal with possibly unbounded functions and therefore consider their restriction to the set*

$$\mathrm{ri}\big(\Delta_{\mathrm{ca}}(\mathcal{Y})\big) := \left\{ \phi \in \Delta_{\mathrm{ca}}(\mathcal{Y}) \mid \lambda x + (1-\lambda) y \in \Delta_{\mathrm{ca}}(\mathcal{Y}) \text{ for some } \lambda > 1 \right\},$$

called the relative interior of $\Delta_{\mathrm{ca}}(\mathcal{Y})$, for well-posedness. For instance, the logarithmic loss $\ell_{\log}(\phi, y) := -\log(\phi_y)$ is not a valid choice for us as is, but only if restricted to (a subset of) $\mathrm{ri}\big(\Delta_{\mathrm{ca}}(\mathcal{Y})\big)$ or modified adequately.

As defined above, the superprediction set is isomorphic to a set in $\mathbb{R}^{|\mathcal{Y}|}$, because every

function $u \in \mathcal{B}_b(\mathcal{Y})$ can be represented as a vector $u \coloneqq \big(u(1), \dots, u(|\mathcal{Y}|)\big)^\top$. In particular, we can write $\ell(\phi) \in \mathbb{R}^{|\mathcal{Y}|}$. To better visualize the set, one can use Equation 29 from Williamson and Cranko (2023), which states that the unconstrained superprediction set for a loss function can be written as

$$\operatorname{spr}\ell = \ell\big(\Delta_{\text{ca}}(\mathcal{Y})\big) \oplus \mathbb{R}^{|\mathcal{Y}|}_{\geq 0}. \tag{3.1}$$

In other words, the superprediction set has the loss values on its boundary, and vectors with all entries larger or equal than all loss values (aka, the superpredictions) in its interior. We give a visualization of this object in Fig. 3.1.

Williamson and Cranko (2023) also prove the connection between the superprediction set and the conditional Bayes risk (Def. 3.7), later extended to a connection with a generalized notion of information (Williamson and Cranko, 2024).[4] We state here the result in a simplified fashion.

Proposition 3.16 (Theorem 15(1), Williamson and Cranko (2023))**.** *Let ℓ be a loss function, defined as $\ell\colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Then, it holds that*

$$\mathbf{CBR}_\ell(\phi) = \inf_{l \in \overline{\text{co}}\operatorname{spr}\ell} \langle l, \phi \rangle.$$

Proof.

$$\begin{aligned}\mathbf{CBR}_\ell(\phi) &= \inf_{\psi \in \Delta_{\text{ca}}(\mathcal{Y})} \sum_y \ell(\psi, y)\, \phi_y =: \inf_{\psi \in \Delta_{\text{ca}}(\mathcal{Y})} \langle \ell(\psi), \phi \rangle \\ &= \inf_{l \in \ell\left(\Delta_{\text{ca}}(\mathcal{Y})\right)} \langle l, \phi \rangle \overset{(\star)}{=} \inf_{l \in \operatorname{spr}\ell} \langle l, \phi \rangle = \inf_{l \in \overline{\text{co}}\operatorname{spr}\ell} \langle l, \phi \rangle.\end{aligned}$$

where $\langle l, \phi \rangle$ here means the dot product in $\mathbb{R}^{|\mathcal{Y}|}$. The last equality holds as the minimized function is linear; $(\star)$ uses Eq. (3.1) together with the positiveness of ϕ and increasing monotonicity of $\langle \cdot, \phi \rangle$. □

There are different limitations that come with studying a learning problem through its conditional risk, and therefore the associated unconstrained superprediction set (or, superprediction set of the loss). The most apparent is that this formulation does not take into account the full data space $\mathcal{X} \times \mathcal{Y}$. By definition, the conditional Bayes risk only accounts for label distributions, and its relationship with the full Bayes risk is well-posed only for the unconstrained case (see the Remark 3.8). Hence, the unconstrained superprediction set is not a viable option to study the full Bayes risk or to characterize constrained supervised learning problems.

[4] Note that the unconstrained superprediction set in Williamson and Cranko (2023) is defined for the *relative interior of* $\Delta_{\text{ca}}(\mathcal{Y})$, while here we are using the full space.

3.3.1 Constrained Superprediction Set

We now focus on generalizing the superprediction set to include the constrained model class, as well as to properly connect it with Bayes risk and supervised learning problems. This generalization is non-trivial, as we in general consider a possibly uncountable attribute set $\mathcal{X} \subseteq \mathbb{R}^d$ and a model class constituted of Markov kernels, i.e., measurable functions mapping points in $\mathcal{X} \times \Sigma(\mathcal{Y})$ to a probability value in $[0, 1]$. This makes it impossible to have the constrained superprediction set to live in the finite-dimensional euclidean space $\mathbb{R}^{|\mathcal{Y}|}$, as we instead noted for the unconstrained case.

Definition 3.17 (Constrained Superprediction Set with Respect to Model Class)**.** *Let $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ be a model class and $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function. We define the* constrained superprediction set of $\ell \circ \mathcal{H}$ *as*

$$\operatorname{spr}(\ell \circ \mathcal{H}) \coloneqq \bigcup_{h \in \mathcal{H}} \left\{ f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}) \mid (\ell \circ h)(x, y) \leq f(x, y) \ \ \forall (x, y) \in \mathcal{X} \times \mathcal{Y} \right\}.$$

We can also define the superprediction set in a more general fashion, not necessarily assuming a loss and model class but functions in $\mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$. This definition will be useful later, when studying how the superprediction set is modified by the action of a Markov kernel.

Definition 3.18 (Superprediction Set Operator)**.** *Let $\mathcal{F} \subseteq \mathcal{B}_b(\mathcal{Z})_{\geq 0}$ non-empty. The* superprediction set of $\mathcal{F}$ *is defined as,*

$$\operatorname{spr}(\mathcal{F}) \coloneqq \bigcup_{a \in \mathcal{F}} \left\{ f \in \mathcal{B}_b(\mathcal{Z}) \mid a(z) \leq f(z) \ \ \forall z \in \mathcal{Z} \right\}.$$

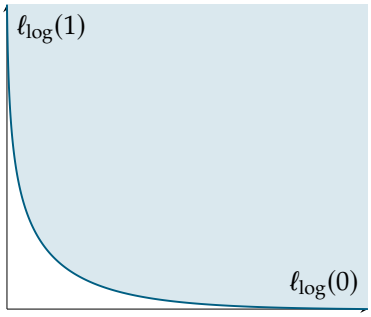
One characterizing property of the superprediction set is that it extends a set of positive functions only in the positive orthant; this result extends Eq. (3.1) to the constrained learning case.

Lemma 3.19. *Let $\mathcal{F} \subseteq \mathcal{B}_b(\mathcal{Z})_{\geq 0}$. The associated superprediction set can be written as*

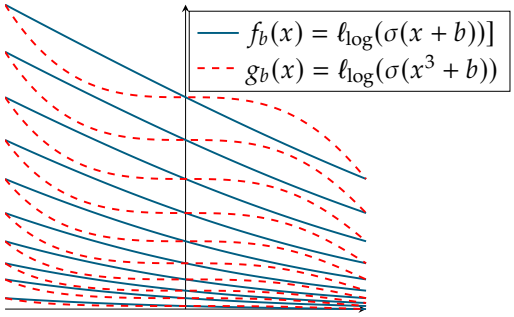
$$\operatorname{spr}(\mathcal{F}) = \mathcal{F} \oplus \mathcal{B}_b(\mathcal{Z})_{\geq 0}.$$

Proof. We first show the set inclusion from left to right. Let $f \in \operatorname{spr}(\mathcal{F})$, i.e., there exists $a \in \mathcal{F}$ such that $a \leq f$. Define $g \coloneqq f - a \in \mathcal{B}_b(\mathcal{Z})_{\geq 0}$. Hence, clearly $f = a + g \in \mathcal{F} \oplus \mathcal{B}_b(\mathcal{Z})_{\geq 0}$. For the reverse direction, note that any $f \in \mathcal{F} \oplus \mathcal{B}_b(\mathcal{Z})_{\geq 0}$ can be written as $f = a + g$ for some $a \in \mathcal{F}$ and $g \in \mathcal{B}_b(\mathcal{Z})_{\geq 0}$, hence $f \geq a$. □

Remark 3.20 *One could imagine a different construction for a superprediction set for a loss function ℓ and a model class $\mathcal{H}$ which goes as follows. First, note that we can define for every $x \in \mathcal{X}$, $\mathcal{H}_x \coloneqq \{\dot{h}(x) \in \Delta_{\text{ca}}(\mathcal{Y}) \mid h \in \mathcal{H}_x\}$. Then, we essentially restrict the unconstrained superprediction*



(a) Set spr $\ell_{\log}(\Delta_{\text{ca}}(\mathcal{Y}))$ for the log loss, with curve $\{(-\log(p), -\log(1-p)), p \in \Delta_{\text{ca}}(\mathcal{Y})\}$.



(b) Values for $x \in [-1,1] \mapsto \ell_{\log}(h_{y=0}(x), 0)$ for linear and cubic models, varying $b \in [-2,3]$.

Figure 3.1: Panel (a) shows that, no matter what the model class is, the superprediction set of the loss (in this case for logarithmic loss $\ell_{\log}$) will look the same as long as the model class covers the whole simplex, i.e., $\Delta_{\text{ca}}(\mathcal{Y}) = \bigcup_{x\in\mathcal{X}}\{\dot{h}(x), h \in \mathcal{H}\}$. However, in panel (b) we see that the single components of the elements of the constrained superprediction set, i.e., $\left[\ell_{\log}(h_y(x), y)\right]_{x,y}$, take different values for different models $h \in \mathcal{H}_{\text{lin}} = \left\{\dot{h}(x) := \left[-\log\left(\sigma(x+b)\right), -\log\left(1-\sigma(x+b)\right)\right]\right\}$ and $h \in \mathcal{H}_{\text{cub}} = \left\{\dot{h}(x) := \left[-\log\left(\sigma(x^3+b)\right), -\log\left(1-\sigma(x^3+b)\right)\right]\right\}$. Hence, the corresponding constrained superprediction set will look differently for each of these model classes.

set to,

$$\text{spr}(\ell \circ \mathcal{H}_x) := \bigcup_{\phi\in\mathcal{H}_x} \left\{u \in \mathcal{B}_b(\mathcal{Y}) \,|\, \ell(\phi) \leq u\right\} \subseteq \text{spr}(\ell)\,.$$

Naively, one might then take the product over $x \in \mathcal{X}$. However, in general,

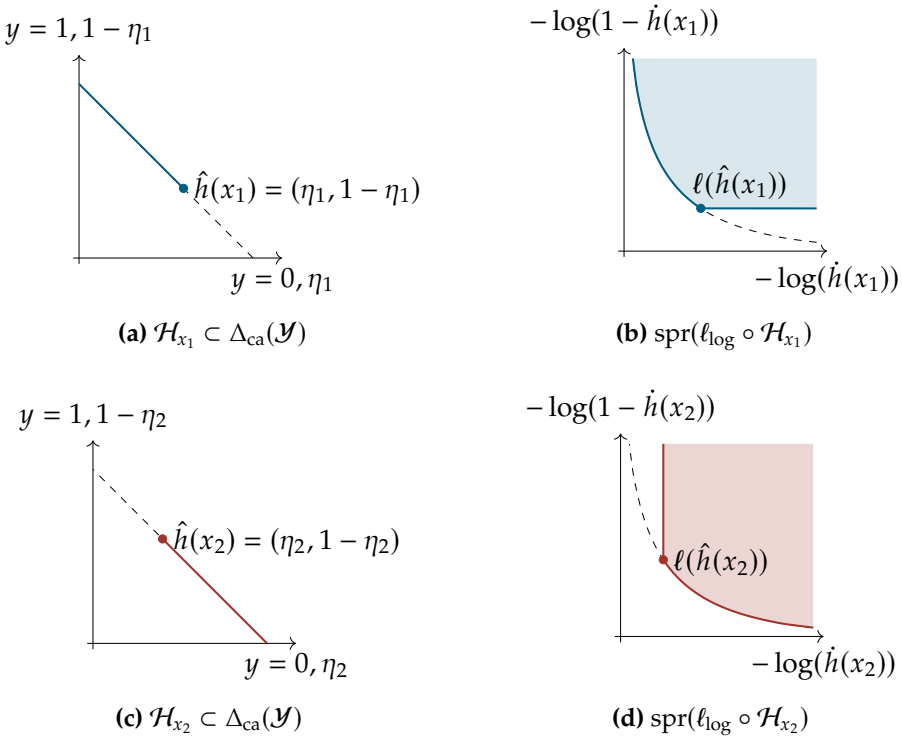
$$\text{spr}(\ell \circ \mathcal{H}) \subset \bigotimes_{x\in\mathcal{X}} \text{spr}(\ell \circ \mathcal{H}_x)\,,$$

where the subsethood can be strict ($\subsetneq$). This is the case whenever the model class is not expressive enough to independently assign predictions to different $x \in \mathcal{X}$. For instance, when a model class is such that $\dot{h}(x_1) = \phi \in \Delta_{\text{ca}}(\mathcal{Y})$ implies that $\dot{h}(x_2) = \psi(\mathcal{Y})$ for every $h \in \mathcal{H}$ and $\phi \neq \psi$, then the prediction assignment to x_1 interacts with the prediction assignment to x_2.

3.3.2 Superprediction Set and Bayes Risk

Definition 3.21 (Support Functions). *Let $\mathcal{Z}$ be a Polish space. Let $\mathcal{A} \subseteq \mathcal{B}_b(\mathcal{Z})$ and $\mathcal{B} \subseteq \text{ba}(\mathcal{Z})$ be non-empty, the functions*

$$\sigma_{\mathcal{A}}(\mu) := \sup_{f\in\mathcal{A}} \langle f, \mu\rangle, \qquad \sigma_{\mathcal{B}}(f) := \sup_{\mu\in\mathcal{B}} \langle f, \mu\rangle,$$



(a) $\mathcal{H}_{x_1} \subset \Delta_{\text{ca}}(\mathcal{Y})$ **(b)** $\text{spr}(\ell_{\log} \circ \mathcal{H}_{x_1})$

(c) $\mathcal{H}_{x_2} \subset \Delta_{\text{ca}}(\mathcal{Y})$ **(d)** $\text{spr}(\ell_{\log} \circ \mathcal{H}_{x_2})$

Figure 3.2: Visualization of the set $\text{spr}(\ell_{\log} \circ \mathcal{H}_x)$ for a model class that does not cover the whole simplex. We fix $\eta_1 = 0.6$ and $\eta_2 = 0.4$.

are respectively called convex support function *of* $\mathcal{A}$ *and* $\mathcal{B}$*. The functions*

$$\rho_{\mathcal{A}}(\mu) \coloneqq \inf_{f \in \mathcal{A}} \langle f, \mu \rangle, \qquad \rho_{\mathcal{B}}(f) \coloneqq \inf_{\mu \in \mathcal{B}} \langle f, \mu \rangle,$$

are called concave support functions.

Lemma 3.22 (Properties of Concave Support Functions)**.** *Let* $\mathcal{Z}$ *be a Polish space and* $\mathcal{A}, \mathcal{B} \subseteq \mathcal{B}_b(\mathcal{Z})$ *be non-empty. The following statements hold.*

(P1) $\sigma_{\mathcal{A}} = -\rho_{-\mathcal{A}}$.

(P2) ρ *is* upper semi-continuous *with respect to the* $\sigma(\mathcal{B}_b, \text{ba})$*-topology.*

(P3) ρ *is* sublinear*, i.e.,* $\rho_{\mathcal{A}}(\alpha\mu) = \rho_{\mathcal{A}}(\mu)$ (positive homogeneity) *and* $\rho_{\mathcal{A}}(\mu + \nu) \geq \rho_{\mathcal{A}}(\mu) + \rho_{\mathcal{A}}(\nu)$ (superadditivity) *for all* $\alpha \in \mathbb{R}_{\geq 0}$ *and all* $\mu, \nu \in \text{ba}(\mathcal{Z})$.

(P4) $\rho_{\mathcal{A}} = \rho_{\overline{\text{co}}(\mathcal{A})}$.

(P5) $\rho_{\mathcal{A} \oplus \mathcal{B}} = \rho_{\mathcal{A}} + \rho_{\mathcal{B}}$.

Analogous properties hold for non-empty sets $\mathcal{C}, \mathcal{D} \subseteq \text{ba}(\mathcal{Z})$.

Proof.

1. Trivially follows by the linearity of the dual pairing and properties of sup and inf functions.

2. The pointwise infimum over a family of $\sigma(\mathrm{ba}, \mathcal{B}_b)$-continuous functions is upper semicontinuous (Aliprantis and Border, 2006, Lemma 2.41).

3. Positive homogeneity is a direct consequence of the definition and the linearity of the bilinear mapping. Furthermore, the pointwise infimum over linear functions is concave ((P1) and Aliprantis and Border (2006, Lemma 5.40)) and positive homogeneity with concavity implies superadditivity ((P1) and Aliprantis and Border (2006, Definition 5.45)).

4. Follows from (P2) and (P3) together with Aliprantis and Border (2006, Theorem 7.51).

5. Follows from (P1) and Lemma 7.54 (2) in Aliprantis and Border (2006).

□

In the rest of this work, we will mostly make use of the concave support function, and will refer to it simply as *support function* unless ambiguous.

Proposition 3.23 (Support Function of Superprediction Set is Bayes Risk)**.** *Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. It holds,*

$$\rho_{\mathrm{spr}(\ell \circ \mathcal{H})}(\phi) = \mathbf{BR}_{\ell \circ \mathcal{H}}(\phi)\,, \quad \phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})\,.$$

Proof. Let $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$,

$$\begin{aligned}
\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) &= \inf_{h \in \mathcal{H}} \mathbb{E}_{\phi}[\ell \circ h] \\
&= \inf_{h \in \mathcal{H}} \langle \ell \circ h, \phi \rangle = \inf_{f \in \ell \circ \mathcal{H}} \langle f, \phi \rangle \\
&\overset{(\star)}{=} \inf_{f \in \ell \circ \mathcal{H}} \langle f, \phi \rangle + \inf_{f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}} \langle f, \phi \rangle \\
&= \inf_{f \in (\ell \circ \mathcal{H}) \oplus \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}} \langle f, \phi \rangle \\
&\overset{L.3.19}{=} \inf_{f \in \mathrm{spr}(\ell \circ \mathcal{H})} \langle f, \phi \rangle = \rho_{\mathrm{spr}(\ell \circ \mathcal{H})}(\phi)\,.
\end{aligned}$$

Where the equality $(\star)$ is true as, for $f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$ and $\phi \in \Delta_{\mathrm{ba}}(\mathcal{Y})$, we have $\langle f, \phi \rangle \geq 0$. Hence, $\inf_{f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}} \langle f, \phi \rangle = \langle 0, \phi \rangle = 0$. □

Proposition 3.23 generalizes the correspondence of the unconstrained superprediction set with the conditional Bayes risk stated in (Williamson and Cranko, 2024, Remark 18) and (Williamson and Cranko, 2023, Theorem 15). Given Remark 3.8, we can also recover the connection with conditional Bayes risk for $\phi \in \Delta_{\mathrm{ca}}(\mathcal{X} \times \mathcal{Y})$.[5]

As a last observation, we notice that a direct consequence of Lemma 3.22 (P4) is that we can, without changing the support function, take the convex closure of the superprediction set.

Corollary 3.24. *Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. It holds,*

$$\rho_{\mathrm{spr}(\ell \circ \mathcal{H})}(\mu) = \rho_{\overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H})}(\mu)\,, \quad \mu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y})\,.$$

Remark 3.25 *For the topologically inclined readers, we note that Proposition 3.23 suggests that the Bayes risk functional is* not *continuous with respect to the $\sigma(\mathrm{ba}, \mathcal{B}_b)$-topology. That seems counterintuitive, as all linear functionals of the type $\phi \mapsto \langle f, \phi \rangle$ are $\sigma(\mathrm{ba}, \mathcal{B}_b)$-continuous by definition of $\sigma(\mathrm{ba}, \mathcal{B}_b)$. The infimum over those functions is, however, not necessarily continuous, but only upper semicontinuous (Aliprantis and Border, 2006, Lemma 2.41). Hence, it follows that even though $\overline{\Delta_{\mathrm{ca}}} = \Delta_{\mathrm{ba}}$ (Lemma 2.13), the Bayes risk of a finitely additive probability distribution might not be approximated arbitrarily well by the Bayes risk of a countably additive probability distribution.*

[5] Note that we need to assume that $\phi \in \Delta_{\mathrm{ca}}(\mathcal{X} \times \mathcal{Y})$, as it is not sufficient to assume $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$ to ensure the existence and (almost sure) uniqueness of the associated conditional expectation—in other words, the validity of the disintegration theorem.

Part II

Transformations and Equivalences of Learning Problems

Chapter 4

An Exhaustive Taxonomy of Markovian Corruptions in Supervised Learning

The most widespread conception of data defines them as atomic facts, perfectly describing some reality of interest (Poovey, 1998). In learning theories, this is reflected by the often-used assumption that training and test data are drawn identically and independently from some fixed probability distribution. The goal of learning is then construed as identifying and synthesizing patterns based on the knowledge, or information, embedded in these data. In practice, however, *corruption* regularly occurs in data collection. This creates a mismatch between training and test distributions, forcing us to learn from imperfect facts. Upon this observation, machine learning researchers should go beyond the traditional emphasis on prediction models and loss functions, and focus also on the data at hand. As is already common practice in statistics, researchers and practitioners need to understand how different processes may have led us to the observation of particular data, to understand how they can subsequently impact the learning process. While the necessity of investigating this topic is recognized both at a practical (World Economic Forum Global Future Council on Human Rights 2016-2018, 2018; Malinin et al., 2021; Koh et al., 2021) and a theoretical (Meng, 2021; Rostamzadeh et al., 2021) level, no standardized way to model and analyze the dynamic generative process of data has been so far created.

In view of understanding corruption and producing models that are "robust",[1] there has been a sustained surge of research on specific corruption models, predominantly characterized through changes or invariances in selected probability distributions. In Tab. 4.1, we present a non-exhaustive list of such models, each of which can be interpreted as corruption in our sense. These approaches, however, do not permit a principled comparison between different corruption mechanisms. Each framework is developed from distinct premises and tailored to a particular empirical phenomenon. As a result, the

[1] Here, robustness refers to correctness under noisy or corrupted data, acknowledging that the term carries different meanings across the literature.

Models	**Descriptions**
Attribute noise	$\phi(X)$ is corrupted due to, e.g., additive attribute noise or missingness, while the labels remain untouched
Random classification noise	Considering $\phi(Y \mid X)\phi(X)$, $\phi(Y \mid X)$ is corrupted by flipping each label independently with a constant probability, while $\phi(X)$ remains invariant
Class-conditional noise	Considering $\phi(Y \mid X)\phi(X)$, $\phi(Y \mid X)$ is corrupted by flipping labels with a probability dependent on the label, while $\phi(X)$ remains invariant
Instance-dependent noise	Considering $\phi(Y \mid X)\phi(X)$, $\phi(Y \mid X)$ is corrupted by flipping labels with a probability dependent on the instance, while $\phi(X)$ remains invariant
Instance- & label-dependent noise	Considering $\phi(Y \mid X)\phi(X)$, $\phi(Y \mid X)$ is corrupted by flipping labels with an instance- & label-dependent probability, while $\phi(X)$ remains invariant
Mutually contaminated distributions	Considering $\phi(X \mid Y)\phi(Y)$, $\phi(X \mid Y)$ is corrupted by a mixture model, and $\phi(Y)$ can also be corrupted
Combined simple noise	Considering $\phi(Y \mid X)\phi(X)$, $\phi(X)$ is corrupted by additive noise, and $\phi(Y \mid X)$ is corrupted by flipping labels with a probability dependent on the label
Target shift	Considering $\phi(X \mid Y)\phi(Y)$, $\phi(Y)$ is corrupted while $\phi(X \mid Y)$ remains invariant
Covariate shift	Considering $\phi(Y \mid X)\phi(X)$, $\phi(X)$ is corrupted while $\phi(Y \mid X)$ remains invariant
Generalized target shift	Considering $\phi(X \mid Y)\phi(Y)$, $\phi(Y)$ and $\phi(X \mid Y)$ are corrupted, subject to specific invariance assumptions on conditional distributions in the latent space
Style transfer	To model it probabilistically, we express it as $\phi(X \mid Y)$ being changed given the designated style
Adversarial noise	To model it probabilistically, we express it as $\phi(X)$ being intentionally corrupted by an adversary to alter the correct prediction for each instance
Concept drift	$\phi(X, Y)$ changes over time
Concept shift	Considering $\phi(Y \mid X)\phi(X)$, $\phi(Y \mid X)$ changes over time
Sampling shift	Considering $\phi(Y \mid X)\phi(X)$, $\phi(X)$ changes over time, while $\phi(Y \mid X)$ is invariant
Selection bias	$\phi(X, Y)$ is corrupted to $\tilde{\phi}(X, Y)$ s.t. $\tilde{\phi} \ll \phi, \exists!\ \alpha = \frac{d\tilde{\phi}}{dP}$ & $\|\|\alpha\|\|_\infty < \infty$

Table 4.1: Examples of models proposed in the literature that capture data dynamics with probabilistic descriptions. Here, X represents the attribute, and Y represents the label. Details and references are included in Appendix A.

relationships between models remain opaque. The absence of a standardized nomenclature further compounds this fragmentation.

Rather than contributing another isolated model, we propose a taxonomy that systematically organizes the existing landscape. This taxonomy provides a comprehensive map of probabilistic corruption models, currently lacking in the literature. In contrast to much of the existing work, which introduces new corruption settings together with bespoke mitigation algorithms, our objective is not algorithmic novelty, but theoretical clarification. We analyze mitigation strategies within a unified perspective to better understand the structural principles underlying existing methods.

Although prior efforts have aimed to construct unifying frameworks, important limitations persist with respect to homogeneity and exhaustiveness. An early and influential attempt was by Quiñonero-Candela et al. (2008), which assembled a broad range of results on dataset shift without offering a fully integrated treatment. Subsequent works (Moreno-Torres et al., 2012; Kull and Flach, 2014; Sáez, 2022; Subbaswamy et al., 2022) pursued more homogeneous formulations, yet typically relied upon invariance assumptions concerning marginal or conditional distributions. Whether such frameworks exhaustively capture the full spectrum of corruption mechanisms is generally conjectured rather than established.

Therefore, the primary objective of this chapter is to improve the existing understanding of the effects of corruption. We aim to do so by systematically studying and comparing the possible types of corruption in supervised learning problems, providing a general framework for analyzing their mitigation as an initial step toward unraveling these fundamental questions.

4.1 Corruption Definition and Types

In this section, we formally define our conceptualization of corruption within the context of a learning problem, utilizing the mathematical tool of Markov kernels. Given the diverse forms of corruption that can arise, we categorize them through a novel taxonomy based on their input and output spaces–essentially classifying them by *type*. We demonstrate the exhaustiveness of this general framework, which facilitates the systematic study of the various corruption types and combinations. Finally, we analyze the relationships between our taxonomy of corruption and existing corruption models, elucidating novel insights generated by our framework.

As formulated in § 3.1, a learning problem comprises three key components: the loss function ℓ, the model class $\mathcal{H}$, and the probability distribution ϕ from which we draw the data. In the field of machine learning, considerable attention has been devoted by engineers and researchers to the task of designing suitable loss functions or model architectures; however, less effort has been put into data, given that they are often not responsible for collecting them but rather for processing them (Sambasivan et al., 2021).

In contrast to the traditional concept of corruption in machine learning, which only focuses

on data generation and is defined as *distribution shift*–an alteration of the probability distribution to deviate from its original test counterpart–we argue that corruption can occur in any of the components. In this broader sense, opting for *surrogate losses* can be regarded as a form of corruption to the original loss function. For instance, surrogate losses are often chosen in place of the 0-1 loss in the classification problems (Bartlett et al., 2006). Moreover, a *misspecified model*, such as when the model class of choice, e.g., linear functions, does not include the true model, e.g., a quadratic function, can also be considered a form of corruption. Therefore we define the general corruption as any alterations in $(\ell, \mathcal{H}, \phi)$ as follows.

Definition 4.1 (General Corruption)**.** *Let* $(\mathcal{Z}, \Sigma(\mathcal{Z}))$, $(\mathcal{Z}', \Sigma(\mathcal{Z}'))$ *be measurable spaces on which general learning problems can be defined, as prescribed by Def. 3.2. A general corruption is a mapping sending a general learning problem* $\mathfrak{L}(\mathcal{Z})$ *into another general learning problem* $\tilde{\mathfrak{L}}(\mathcal{Z}')$*.*

To initiate a comprehensive taxonomy of corruption, we begin by examining a specific case where corruption is defined as a Markov kernel with fixed input and output probability spaces.

This definition subsumes a significant portion of existing literature, including classical works on distribution shift and noisy data (van Rooyen and Williamson, 2018; Williamson and Cranko, 2024).[2] We therefore turn our attention to this subcase, formally defined below, with the aim of establishing connections between our types of corruption and the diverse corruption models laid out in previous studies.

Definition 4.2 (Markovian Corruption)**.** *A Markovian corruption is a general corruption that maps a general learning problem* $\mathfrak{L}(\mathcal{Z})$ *to another general learning problem* $\tilde{\mathfrak{L}}(\mathcal{Z}')$ *through the action of a Markov kernel* $\kappa \in \mathcal{M}(\mathcal{Z}, \mathcal{Z}')$*. That is, such that* $\tilde{\mathfrak{L}}(\mathcal{Z}) = (\ell, \mathcal{H}, \tilde{\phi} = \check{\kappa}\phi)$*. Two important subcases are:*

1. *A* ***joint Markovian corruption****, which has* $\mathcal{Z} = \mathcal{Z}' = \mathcal{X} \times \mathcal{Y}$, $\Sigma(\mathcal{Z}) = \Sigma(\mathcal{Z}') = \Sigma(\mathcal{X} \times \mathcal{Y})$*, and only maps supervised learning problems into supervised learning problems;*
2. *A* ***partial Markovian corruption****, that is such that* $\mathcal{Z}, \mathcal{Z}'$ *can differ, and may be* $\mathcal{X}, \mathcal{Y}$*, or* $\mathcal{X} \times \mathcal{Y}$*, with the associated* σ*-algebras varying accordingly.*

We remark that the definition above does not necessarily assume the Markov kernel κ to be known. We only require κ to exist, and for us to know the values assumed by the kernel action when evaluated on ϕ, i.e. $\tilde{\phi} = \check{\kappa}\phi$. The kernel is therefore not uniquely identified by the corruption definition, since multiple Markov kernels can generate $\tilde{\phi}$ from ϕ. However, for the analysis carried on in the rest of the chapter, we assume κ to be known.

The rationale behind this choice for modeling corruption lies in viewing a Markov kernel, or *observation channel* (Csiszár, 1972) in the context of information theory, as a point-wise

[2] While Markov kernels have been utilized in formalizing corruption, their primary foci were solely on label corruption, attribute corruption, or simple joint corruption.

description of the stochastic process that leads to an observed probability distribution. This process is determined by external conditions that, in some sense, limit our ability of "seeing" the truth (probabilistic world), consequently giving rise to corruptions (distorted data distribution). Given the probabilistic focus of this work, we will from now on refer to Markovian corruption simply as *corruption*. For formal statements, we abuse the kernel notation and refer to the corruption induced by a kernel as the kernel itself, i.e., $\kappa \in \mathcal{M}(\mathcal{Z}, \mathcal{Z}')$.

4.1.1 A New Taxonomy of Partial Corruptions

Partial corruptions can be classified in different ways based on the domain and image of their associated kernels. Starting from the most general corruption, i.e., the joint corruption on $\mathcal{X} \times \mathcal{Y}$ induced by $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$, when one space (either $\mathcal{X}$ or $\mathcal{Y}$ space) is absent in the image or domain of the corruption kernel, we obtain a partial corruption. In Fig. 4.1, we present all possible types of partial corruption, with the exception of those that are deterministic or degenerate kernels, as they can be seen as obvious subcases of other partial corruptions. By construction, we can state:

Proposition 4.3. *There are no partial Markovian corruptions outside of those listed in Fig. 4.1.*

In the same Fig. 4.1, we classify partial corruptions based on their *signature type*, that is, which sets of $\mathcal{X}$ and $\mathcal{Y}$ constitute the domain and image of the corruption kernel. Specifically, we employ the following nomenclature: *joint* corruption when $D(\kappa) = I(\kappa) = \mathcal{X} \times \mathcal{Y}$; *1-parameter joint* corruption when $D(\kappa) = \mathcal{X} \times \mathcal{Y}$ and $I(\kappa)$ is either $\mathcal{X}$ or $\mathcal{Y}$; *2-dependent* when $D(\kappa) = \mathcal{X} \times \mathcal{Y}$ and $I(\kappa)$ is either $\mathcal{X}$ or $\mathcal{Y}$; *1-dependent* when $D(\kappa) = \mathcal{X}$ and $I(\kappa) = \mathcal{Y}$ or the opposite; *simple* corruption when $D(\kappa) = I(\kappa) \neq \mathcal{X} \times \mathcal{Y}$, so when they are either equal to $\mathcal{X}$ or $\mathcal{Y}$.

4.1.2 Constructing Joint Corruption as a Combination of Partial Ones

We now aim to enumerate all possible ways of constructing joint corruptions, i.e., of the type $\kappa \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$, by combining the nodes in Fig. 4.1 through the superposition operation Def. 2.26. To this end, we introduce an additional condition to be imposed on the combinations of partial corruptions.

Definition 4.4 (One-Step Markovian Corruption)**.** *A Markovian corruption, induced by a kernel $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$, is said to be a* one-step *Markovian joint corruption if κ is formed as a superposition of two partial corruptions, i.e. $\tau \otimes \lambda$, such that neither τ nor λ can be further decomposed using the operations defined in Definitions 2.24–26 and 2.28.*

This definition is intentionally crafted to capture the most fundamental forms of composable corruption: those that occur in a single step, i.e., without further factorizing the kernels. By identifying and analyzing these atomic components, we establish a clear and comprehensive taxonomy of combined corruptions.

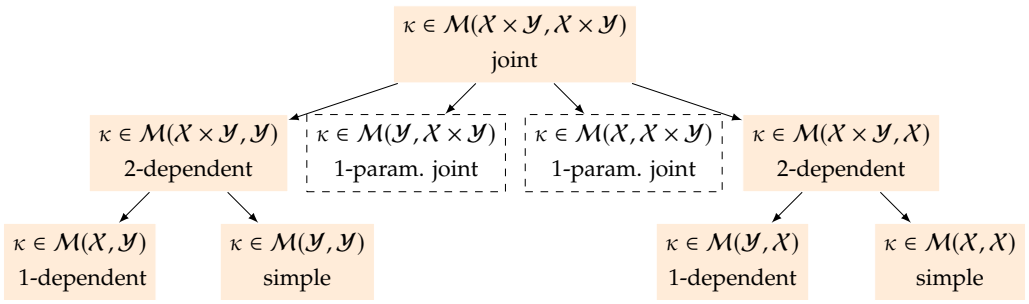


Figure 4.1: *Hierarchy of partial corruption types.* The partial corruption types are hierarchically organized based on their dependence on the instance $\mathcal{X}$ and label $\mathcal{Y}$ space, as depicted through a tree structure. At the root of the tree lies the most general form of corruption, where the domain and image spaces are the joint one $\mathcal{X} \times \mathcal{Y}$, i.e., $D(\kappa) = I(\kappa) = \mathcal{X} \times \mathcal{Y}$. The arrows signify that a child node has its domain or image constant w.r.t. exactly one of the variables in its parent. Therefore, the children nodes can be expressed as subcases of their parent, but the parents generally cannot be expressed by only one of their children. The partial corruption types that cannot be combined with others are shown in dotted boxes. Note that corner cases involving independence from all variables or identity kernels are excluded from this analysis.

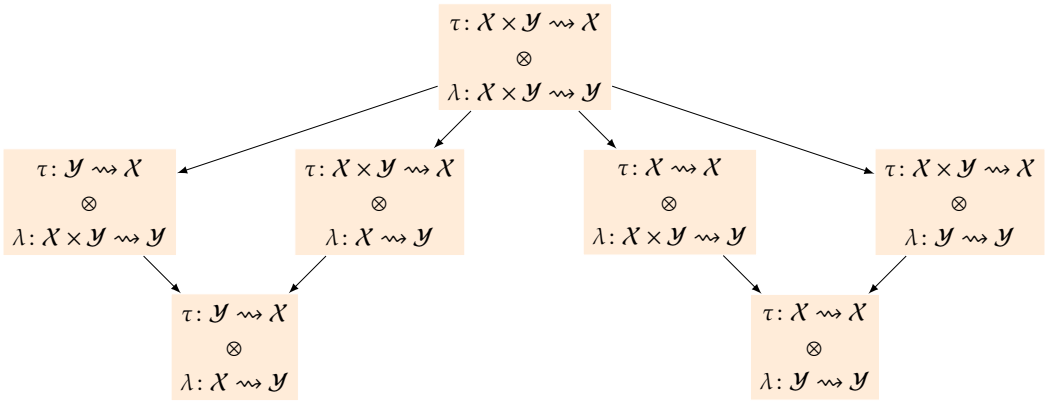


Figure 4.2: *Feasible combinations of partial corruptions.* Joint corruptions, i.e. of type $\kappa : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$, are obtained by combining two compatible partial corruptions in Fig. 4.1. The graph structure is induced by that of the partial corruption types. Notice that we can only combine a partial corruption with $I(\tau) = \mathcal{X}$ with another such that $I(\lambda) = \mathcal{Y}$, following Proposition 4.6. Therefore, the arrows signify that both τ and λ in a child node inherit their domains from the parent node with either τ or λ constant w.r.t. exactly one of their domain variables.

Remark 4.5 *It is also possible for a corruption kernel $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ to not respect the one-step condition in Def. 4.4. This case can occur in different ways: a first option is that τ or λ are obtained through combination of other kernels; or, it can happen that τ and λ are not combined via superposition; lastly, it can also happen that a factorization w.r.t. to Definitions 2.24–26 and 2.28 does not exist for κ. We use a single umbrella term for all such kernels, called* non-factorized joint corruptions*, and treat them as a distinct class in our analysis. In fact, what we are enforcing is a non-factorized representation of all non-one-step Markovian corruptions. Later in this section, we will also give examples of what can fall within this set of corruptions. While a full characterization of these more general scenarios is beyond the scope of this work, we will observe that studying one-step Markovian corruptions gives us many insights also on this distinct set of corruptions.*

Using the requirements introduced until now to shape the objects of interest, we prove the following statement.

Proposition 4.6. *The set of feasible one-step Markovian corruptions $\kappa = \tau \otimes \lambda$ is such that $I(\tau) = \mathcal{X}$ and $I(\lambda) = \mathcal{Y}$.*

Proof. According to Def. 4.2, we must map a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ into another joint one. Hence, we must exclude the combinations of a simple corruption with a 1-dependent corruption since such a pairing cannot generate a joint corruption. Additionally, combinations such as $\tau \otimes \lambda$ with $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X} \times \mathcal{Y}) \otimes \lambda \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ are not allowed because, according to Def. 2.26, the measure on the corrupted labels would be ill-defined. By taking λ and τ from the partial corruption in Fig. 4.1, enumerating all of their possible combinations, and checking which of them are feasible, we can see that only the ones with $I(\tau) = \mathcal{X}$ and $I(\lambda) = \mathcal{Y}$ or $I(\tau) = \mathcal{Y}$ and $I(\lambda) = \mathcal{X}$ respect the condition. Therefore, fixing the notation so that τ is an attribute corruption and λ a label corruption, we get the proposition. □

The set of feasible combinations is depicted and hierarchically organized in Fig. 4.2. Proposition 4.6 formalizes a desirable property of corruption, allowing it to change the distribution on attribute X and the distribution on label Y in a distinguishable way, either independently or dependently. Therefore, corruptions with indistinguishable effects on labels and attributes, such as 1-parameter joint ones, are incorporated in the class of joint non-factorized corruptions, i.e. $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$.[3] In the next section we will see this is an appropriate choice.

Lastly, we can state the following characterization result as a direct consequence of Proposition 4.6.

[3] Note that a 1-parameter joint corruption can be seen as a subcase of a joint one, i.e., as the mapping $\mathcal{A} \in \Sigma(\mathcal{X} \times \mathcal{Y}) \mapsto \kappa(x, y, \mathcal{A}) = \lambda(x, \mathcal{A})\mathbf{1}(y)$, where $\mathbf{1}(y)$ only trivially depends on y since it is the matrix with all entries equal to 1.

Corollary 4.7. *There are no one-step Markovian corruptions outside of those listed in Fig. 4.2, and therefore it constitutes a* complete *taxonomy.*

4.1.3 A Practical Example of Markovian Corruption

Here, we present an illustration of a Markov kernel (in the finite case) within a practical scenario to facilitate the reader's understanding of corruption through kernels. The provided example is adapted from (Fogliato et al., 2020) which considers the prediction of recidivism in the criminal justice system–predicting who goes on to commit future crimes.

Surveys have shown that "in the case of drug crimes, whites are at least as likely as blacks to sell or use drugs; yet blacks are more than twice as likely to be arrested for drug-related offenses" (Rothwell, 2014). Given this, we consider the observed outcome "rearrest", denoted as $\tilde{Y} \in \mathcal{Y} := \{+1, -1\}$, as a corrupted version of the true outcome "reoffense", denoted as $Y \in \mathcal{Y} := \{+1, -1\}$, depending on the attribute $X \in \mathcal{X} = \{b, w\}$. Specifically, the disparity between Y and $\tilde{Y}$ can be captured by a higher probability of flipping the reoffense label ($Y = +1$) to the no rearrest label ($\tilde{Y} = -1$) for white population ($X = w$) compared to the black population ($X = b$):

$$\alpha(w) > \alpha(b), \text{ where } \alpha(x) := \psi(\tilde{Y} = -1 \,|\, Y = +1, X = x),$$

where ψ here indicates the conditional probability density of the corrupted random variable $\tilde{Y}$ on $\mathcal{Y}$ given the clean variables X, Y on $\mathcal{X}$ and $\mathcal{Y}$ respectively.[4] Moreover, we employ the additional assumption that the corruption arises solely from the hidden recidivists, and not from erroneous arrests of individuals not committing the offense. In other words, the probability of flipping the no reoffense label ($Y = -1$) to the rearrest label ($\tilde{Y} = +1$) is zero for both the black and white populations:

$$\beta(w) = \beta(b) = 0, \text{ where } \beta(x) := \psi(\tilde{Y} = +1 \,|\, Y = -1, X = x).$$

A possible Markov kernel modeling the setting would be therefore of the type $\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$, exemplifying 2-dependent label corruption. More specifically, it can be written as a joint kernel $\delta \otimes \lambda$ where $\delta : \mathcal{X} \rightsquigarrow \mathcal{X}$. Being defined in discrete probability spaces, both δ and λ can be expressed as matrices with entries representing conditional probabilities. In particular, we rewrite for clarity λ as its parameterized version $\lambda\big|_{X=x} : \mathcal{Y} \rightsquigarrow \mathcal{Y}$ for $x \in \mathcal{X}$,

[4] In order to define this conditional, we have to consider clean and corrupted learning problems, and a coupling of the clean and corrupted data-generating distributions defining them. We omit it here for brevity.

obtaining:

$$\delta \coloneqq \begin{bmatrix} \psi(\tilde{X} = b \mid X = b) & \psi(\tilde{X} = b \mid X = w) \\ \psi(\tilde{X} = w \mid X = b) & \psi(\tilde{X} = w \mid X = w) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$\lambda\big|_{X=x} \coloneqq \begin{bmatrix} \psi(\tilde{Y} = +1 \mid Y = +1, X = x) & 0 \\ \psi(\tilde{Y} = -1 \mid Y = +1, X = x) & 1 \end{bmatrix} = \begin{bmatrix} 1 - \alpha(x) & 0 \\ \alpha(x) & 1 \end{bmatrix},$$

where the entries are determined according to the given problem setting. To illustrate, consider an example of $\alpha(b) = 1/10$ and $\alpha(w) = 1/5$, yielding the following expressions:

$$\lambda\big|_{X=b} = \begin{bmatrix} 1 - \alpha(b) & 0 \\ \alpha(b) & 1 \end{bmatrix} = \begin{bmatrix} 9/10 & 0 \\ 1/10 & 1 \end{bmatrix} \text{ and } \lambda\big|_{X=w} = \begin{bmatrix} 1 - \alpha(w) & 0 \\ \alpha(w) & 1 \end{bmatrix} = \begin{bmatrix} 4/5 & 0 \\ 1/5 & 1 \end{bmatrix}.$$

From Def. 4.2 we know that defining a Markovian corruption requires specifying a learning problem. In particular, it is necessary to fix the clean probability distribution for us to observe the effect of the corruption kernel on it. Therefore, we consider a joint density on $\mathcal{X} \times \mathcal{Y}$ in the form

$$\phi = [1/4, 1/4, 1/4, 1/4]^\top,$$

where the specific order is assumed to be

$$\phi \coloneqq [\, \phi(X = b, Y = +1),\ \phi(X = b, Y = -1),\ \phi(X = w, Y = +1),\ \phi(X = w, Y = -1) \,]^\top.$$

Note that in finite spaces, the superposition operation Def. 2.26 reduces to the Kronecker product, hence we write the joint corruption kernel $\delta \otimes \lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$ as

$$\delta \otimes \lambda = \left[\begin{array}{c|c} 1 \cdot \lambda\big|_{X=b} & 0 \cdot \lambda\big|_{X=w} \\ \hline 0 \cdot \lambda\big|_{X=b} & 1 \cdot \lambda\big|_{X=w} \end{array}\right] = \left[\begin{array}{c|c} \lambda\big|_{X=b} & \mathbf{0} \\ \hline \mathbf{0} & \lambda\big|_{X=w} \end{array}\right] = \begin{bmatrix} 9/10 & 0 & 0 & 0 \\ 1/10 & 1 & 0 & 0 \\ 0 & 0 & 4/5 & 0 \\ 0 & 0 & 1/5 & 1 \end{bmatrix},$$

which is a 4×4 block diagonal matrix. Written in its probabilistic form, the entries of the matrix representation would hence be the probability of $\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y} \mid X = x, Y = y$. Then we can obtain *a corrupted joint probability* $\tilde{\phi}$ in the following manner:

$$\tilde{\phi} = (\check{\delta} \otimes \check{\lambda})\phi = \begin{bmatrix} 9/10 & 0 & 0 & 0 \\ 1/10 & 1 & 0 & 0 \\ 0 & 0 & 4/5 & 0 \\ 0 & 0 & 1/5 & 1 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \frac{1}{40} \cdot \begin{bmatrix} 9 \\ 11 \\ 8 \\ 12 \end{bmatrix},$$

In this case, the chain composition operation Def. 2.24 is reduced to matrix multiplication. After applying the corruption kernel $\delta \otimes \lambda$, the original clean learning problem with joint probability ϕ, is transformed into a distinct problem governed by $\tilde{\phi}$.

4.1.4 On the Exhaustiveness of Markovian Corruptions

We now know from Corollary 4.7 that the taxonomy of one-step corruption is complete. However, we also noticed that one-step corruptions are not the only possible type of corruptions, as kernels can also come in different forms. We will now leverage a well-known property of Markov kernels—their bijection with coupling of probability spaces—to define an *exhaustive* taxonomy based on the one-step corruption one.

Definition 4.8 (Exhaustive Taxonomy of Corruptions)**.** *We say that a taxonomy of Markovian corruptions is* exhaustive *if, for every possible pair of distributions $\phi, \tilde{\phi} \in \Delta_{\text{ca}}(\mathcal{Z})$, there exists a Markovian corruption from $\mathfrak{L}(\mathcal{Z})$ to $\tilde{\mathfrak{L}}(\mathcal{Z})$ s.t. $\tilde{\phi} = \check{\kappa}\phi$ for all loss ℓ and model class $\mathcal{H}$.*

Proposition 4.9. *The set of feasible one-step Markovian joint corruptions, together with the set of non-factorized ones, constitutes an* exhaustive *taxonomy of Markovian corruptions.*

Proof. Recall from Def. 2.9 that a coupling is defined for two probability spaces $(\mathcal{Z}_1, \Sigma(\mathcal{Z}_1), \phi_1)$, $(\mathcal{Z}_2, \Sigma(\mathcal{Z}_2), \phi_2)$ as a probability space $(\mathcal{Z}_1 \times \mathcal{Z}_2, \Sigma(\mathcal{Z}_1) \times \Sigma(\mathcal{Z}_2), \phi)$, such that the marginal probabilities associated to ϕ w.r.t. $\mathcal{Z}_i$, $i \in \{1, 2\}$, are the respective ϕ_i. By construction, Markov kernels with fixed input and output probabilities $\phi, \tilde{\phi}$ are in bijection with all the possible couplings existing on $\mathcal{Z} \times \mathcal{Z}$ with two *fixed* probability measures; for us, $\phi, \tilde{\phi}$ (see details in Appendix C). This, by definition, proves the exhaustiveness of the taxonomy, as it implies that every corruption that is not one-step Markovian but still sends ϕ to $\tilde{\phi}$ has a non-factorized representation. □

4.2 Scope of the Taxonomy and Excluded Models

The proposed taxonomy enjoys an interesting property of Markov kernels, which is their ability to give an alternative (non-factorized or one-step) stochastic representation to changes of a probability distribution into another. It is important to keep in mind that these are only *possible* representations. For some pairs $(\phi, \tilde{\phi})$, there exist also representations not within our taxonomy, as shown by the examples of multi-step corruptions included in the next section (§ 4.2.1). They are also valid modeling choices for corruption, as they may be functional to analyses different from ours; here we ascribed non-one-step cases to the non-factorized joint corruption group, as our framework is enough for obtaining the results presented in the next chapter. Nevertheless, studying the properties and consequences of one-step corruptions gives hints on properties and consequences of what is included in the non-factorized class. Again, this will be further discussed in § 4.2.1.

Focusing on what is instead included in our taxonomy, we get a novel perspective on how to understand many existing works under a unified framework, hierarchically organized based on the notion of dependence on the instance and label space. We reorganize the corruption models outlined in Tab. 4.1 and illustrate them within our taxonomy, as depicted

Corruption name in literature	**Corruption type**	**Kernel representation**
Attribute noise	simple	$\tau : \mathcal{X} \rightsquigarrow \mathcal{X}$
Style transfer	1-dep.	$\tau : \mathcal{Y} \rightsquigarrow \mathcal{X}$
Adversarial noise	2-dep.	$\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}$
Random classification noise	simple	$\lambda : \{*\} \rightsquigarrow \mathcal{Y}$
Class-conditional noise	simple	$\lambda : \mathcal{Y} \rightsquigarrow \mathcal{Y}$
Instance-dependent noise	1-dep.	$\lambda : \mathcal{X} \rightsquigarrow \mathcal{Y}$
Instance- & label-dependent noise	2-dep.	$\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$
Combined simple noise	two simple combined	$(\tau : \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda : \mathcal{Y} \rightsquigarrow \mathcal{Y})$
Generalized target shift	two 2-dep. combined	$(\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$
Target shift	min. simple, max. 1-dep. & 2-dep. combined	$\lambda : \mathcal{Y} \rightsquigarrow \mathcal{Y}$, $(\tau : \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$
Concept shift	min. simple, max. two 2-dep. combined	$\lambda : \mathcal{Y} \rightsquigarrow \mathcal{Y}$, $(\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$
Covariate shift Sampling shift	min. simple, max. 2-dep. & 1-dep. combined	$\tau : \mathcal{X} \rightsquigarrow \mathcal{X}$, $(\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda : \mathcal{X} \rightsquigarrow \mathcal{Y})$
Concept drift	can be any type, including the non-factorized one	-
Mutually contaminated distributions Selection bias	non-Markovian corruption	-
Selection bias	non-Markovian corruption	-

Table 4.2: Illustration of the taxonomy with examples of existing corruption models. When only one kernel is indicated, missing variables remain unchanged. "dep." is short for "dependent". Details and references in Appendix A.

in Tab. 4.2.[5] Notably, Tab. 4.2 unveils several new insights that are not readily evident from Tab. 4.1.

Firstly, it enables a qualitative comparison of different corruption models by considering their domain and image spaces (refer to Fig. 4.1 and Fig. 4.2). If a kernel exhibits more complexities in its domain and image, it will induce a more intricate form of corruption. Conversely, corruptions demonstrating less complexity in the domain and image of the associated kernel can be regarded as subcases of those with greater complexity. For example, instance- & label-dependent noise characterized as a 2-dependent label corruption $\lambda\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$, is more intricate than class-conditional noise, a simple label corruption $\lambda\colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$, with more dependence on $\mathcal{X}$; the latter can be seen as a subcase of the former with its kernel being constant w.r.t. $\mathcal{X}$.

Secondly, it reveals that some corruption models in Tab. 4.1 correspond to multiple corruption types or combinations of partial corruptions in our taxonomy. This is because their definitions in Tab. 4.1 often consider the corruption of either the probability in the $\mathcal{X}$ space or the $\mathcal{Y}$ space, leaving freedom for corrupting the other. In extreme cases like concept drift, which assumes corruption only in the joint distribution, it can be of any corruption type and may not even be factorized to partial corruptions.

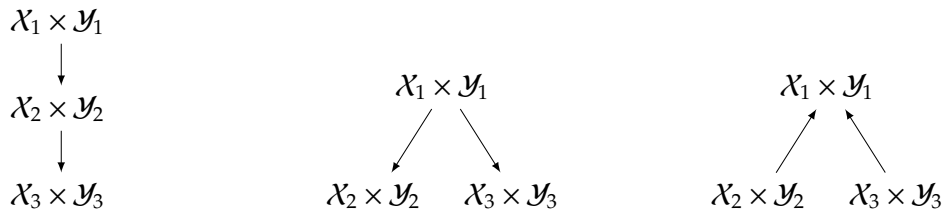
Thirdly, it provides new insights into some corruptions that have been taken for granted to share the same probabilistic nature as others, but turn out to be non-Markovian corruptions. Examples include mutually contaminated distributions and selection bias, topics that will be further elaborated in the following.

4.2.1 Multi-Step Corruption

In machine learning research, the corruption process typically involves two environments—training and test time—but other settings are also possible. For example, scenarios involving more than two spaces arise when learning from multiple domains (Ben-David et al., 2010), or in the presence of concept drift over time (Widmer and Kubat, 1996; Gama et al., 2014; Lu et al., 2019). By relaxing certain assumptions, the applicability of our framework can be extended into such cases by combining one-step Markovian corruptions. In these cases, kernels act in a "sequential" (Definitions 2.24 and 2.25) or "parallel" (Def. 2.26) manner, enabling the modeling of more complex patterns of corruption.

Multi-Step Markovian Corruption To illustrate how one-step corruptions give insights about non-one-step ones, consider a scenario with three environments $\mathcal{X}_i \times \mathcal{Y}_i$, $i = 1, 2, 3$. We aim to model Markovian corruptions occurring among these spaces and represent them using directed acyclic graphs, as usually done in causality literature (Pearl et al., 2016). In this representation, an arrow indicates the presence of a non-trivial Markov kernel between two spaces.

[5] Details about the relationships with these corruption instances are given in Appendix A, and discussions on the relationships with other data shift taxonomies are given in Appendix B.



(a) Chain structure. **(b)** Fork structure. **(c)** Collider structure.

Figure 4.3: Possible non-degenerate relations among three probability spaces. Arrows represent non-trivial Markov kernel $\kappa \in \mathcal{M}(\mathcal{X}_i \times \mathcal{Y}_i, \mathcal{X}_j \times \mathcal{Y}_j)$.

We focus on three corruption configurations depicted in Fig. 4.3, excluding triangular structures with three arrows, as they lack a clear distinction between input and output spaces, making the corruption flow not interpretable.

The first configuration is the *chain* structure shown in Fig. 4.3(b), where the spaces influence each other sequentially. A concrete example is when $\kappa_1 \colon \mathcal{X} \rightsquigarrow \mathcal{W}$ models a feature extractor for the attribute space $\mathcal{X}$, and $\kappa_2 \colon \mathcal{W} \rightsquigarrow \mathcal{X}$ represents a corruption depending on the latent features only. Although this structure does not strictly satisfy our one-step corruption definition as per Def. 4.4, it can still be represented in our framework by composing kernels as $\kappa := \kappa_1 \circ \kappa_2$, with each κ_i being a one-step Markovian corruption. Similar structures also occur in scenarios involving concept drift or online learning with corruption (Widmer and Kubat, 1996; Cesa-Bianchi et al., 2010; Lu et al., 2019).

A second option is that the spaces relate according to the *triangular* structures shown in Fig. 4.3(a) and (c). In particular, case (a) reflects assumptions made in settings combining data from different domains (Ben-David et al., 2010; van Rooyen and Williamson, 2018; Redko et al., 2022), where distinct observed distributions are assumed to arise from a common underlying clean distribution but corrupted differently depending on their environment. We can model this as a single Markov kernel obtained via superposition of two others, i.e., $\kappa := \kappa_1 \otimes \kappa_2$, with $\kappa_1 \in \mathcal{M}(\mathcal{X}_1 \times \mathcal{Y}_1, \mathcal{X}_2 \times \mathcal{Y}_2)$ and $\kappa_2 \in \mathcal{M}(\mathcal{X}_1 \times \mathcal{Y}_1, \mathcal{X}_3 \times \mathcal{Y}_3)$. As for case (c), it can be used to model scenarios with merged (noisy, or clean and noisy) datasets, which is a fairly common practice in robustness research, e.g. Veit et al. (2017); Fatras et al. (2022), as well as causality, e.g. (Gresele et al., 2022; Garrido Mejia et al., 2024). The corruption would then be represented with a $\kappa \in \mathcal{M}(\mathcal{X}_2 \times \mathcal{Y}_2 \times \mathcal{X}_3 \times \mathcal{Y}_3, \mathcal{X}_1 \times \mathcal{Y}_1)$, with a decoupled joint probability as input, e.g., $\phi = \phi_2 \times \kappa_{\phi_3}$ where κ_{ϕ_3} is a degenerate kernel and ϕ_2, ϕ_3 are the clean underlying probabilities.

These three basic structures can be themselves combined to form more complex graphical models, capturing relationships among n environments. Since these are built from Markovian one-step corruptions or partial Markovian corruptions, our framework still offers insights into these more complex combinations, which motivates our focus on one-step corruptions.

Multi-Step Non-Markovian Corruption. Consider two Markov kernels $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y} \times \mathcal{X}, \mathcal{Y})$ and $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$. Let $\mathcal{A}$ be an element of the Borel sigma algebra on $\mathcal{X} \times \mathcal{Y}$. We can obtain a corrupted measure $\tilde{\phi} \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y})$ as

$$\begin{aligned}\tilde{\phi}(\mathcal{A}) = (\tilde{F} \times \tilde{\pi})(\mathcal{A}) &\coloneqq [(\phi \circ_{\mathcal{X}\times\mathcal{Y}} \lambda) \times \check{\tau}\pi](\mathcal{A}) \\ &= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} [\phi \circ_{\mathcal{X}\times\mathcal{Y}} \lambda](\tilde{x}, d\tilde{y}) \cdot \check{\tau}\pi(d\tilde{x}) \\ &= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \lambda(x, y, \tilde{x}, d\tilde{y})\, \phi(dx, dy)\right] \cdot \left[\int_{x'\in\mathcal{X}} \tau(x', d\tilde{x})\, \pi(dx')\right],\end{aligned}$$

where $\phi \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y})$ is the joint clean probability and $\pi \in \Delta_{\text{ca}}(\mathcal{X})$ the associated clean $\mathcal{X}$-marginal. It is apparent that the domain space of λ does not respect the one-step assumption: It assumes a coupling defining joint probability space on $(\mathcal{X} \times \mathcal{Y} \times \mathcal{X})$, where the latter $\mathcal{X}$ is meant to be equipped with the *corrupted* marginal probability $\pi \circ \tau$, and $\mathcal{X} \times \mathcal{Y}$ has marginal ϕ. This modeling choice amounts to a corruption that does not act in a "parallel" fashion on $\mathcal{X}$ and $\mathcal{Y}$, i.e., through $\otimes$. The corruption of the label space happens on a second time step, depending on the outcome of an initial corruption phase carried out by τ and only involving $\mathcal{X}$.

Notice that, more in general, we cannot write the combined action of λ and τ as a unique kernel κ acting on ϕ; the available operations do not allow it. Consequently, it cannot be expressed as a single graphical model, although each individual step can. However, we can write $\tilde{\phi}(\mathcal{A}) = [\tilde{F} \times \tilde{\pi}](\mathcal{A})$, with $\tilde{\pi} \coloneqq \check{\tau}\pi$ and $\tilde{F} \coloneqq \phi \circ_{\mathcal{X}\times\mathcal{Y}} \lambda$. This shows that the resulting corruption is *non-Markovian* in how it acts on the clean probability ϕ, *but has (almost) Markovian components*.[6]

4.2.2 Mutually Contaminated Distributions

While being a term with less widespread recognition, *mutually contaminated distributions* (MCD) (Blanchard and Scott, 2014; Menon et al., 2015; Blanchard et al., 2016; Katz-Samuels et al., 2019) is a popular corruption model that has been studied in the literature under more familiar names, for example, *learning from positive and unlabeled data* (Elkan and Noto, 2008; Ward et al., 2009; Du Plessis et al., 2014, 2015; Kiryo et al., 2017) in the binary class case, and *learning from label proportions* (Quadrianto et al., 2008; Yu et al., 2014; Liu et al., 2019; Scott and Zhang, 2020; Tang et al., 2023). Despite the popularity, less is understood about how MCD relates to other corruption models. Our framework offers new insights into such relationships and demonstrates how MCD extends beyond Markovian corruptions.

To initiate this analysis, we first formally define MCD in the sense of Tab. 4.1. Fix a measurable instance space $\mathcal{X}$, and denote by ϕ a distribution over $\mathcal{X} \times [K]$ for $[K] \coloneqq \{1, 2, \cdots, K\}$ with random variables $(X, Y) \sim \phi$.

Definition 4.10 (Katz-Samuels et al. (2019))**.** *Let $\psi(X = x \mid Y = k)$ be class-conditional*

[6] The "almost" refers to λ acting on ϕ via partial chaining instead of standard chaining. This is very similar to our Markovian corruption definition, but it still is a slight relaxation.

*distribution and π_k be the base rate, i.e., the probability of $Y = k$, $k \in [K]$, s.t. $\phi = \pi \times \psi$. Consider some mixing probabilities $\{\pi_{m,k}\}_{m\in[M],k\in[K]}$, with $\pi_{m,k} \geq 0$ and $\sum_m \pi_{m,k} = 1$. Then, **the** MCD **corruption model** assumes that there is a general corruption from $(\ell, \mathcal{H}, \phi)$ to $(\ell, \mathcal{H}, \tilde{\phi})$, such that the corrupted class-conditional distributions are of the form*

$$\tilde{\psi}(X = x \mid Y = m) \coloneqq \sum_{k=1}^{K} \pi_{m,k}\, \psi(X = x \mid Y = k) \quad \forall x \in \mathcal{X}\,,$$

where $m \in [M]$ denotes the corrupted class.

This definition has some clear differences with our definition of corruption. First, Def. 4.10 uses the class conditional probabilities $\psi(X = x \mid Y = k)$ instead of the joint probability. In our language, this means expressing corruption via the experiment E. Secondly, their mixing probabilities defined a mixing matrix $\mathbf{\Pi} \coloneqq (\pi_{m,k})_{m\in[M],k\in[K]}$ that is *row-stochastic* instead of our column stochastic kernels. We can therefore translate the MCD into our notation as in the following:

$$\tilde{\phi}(\mathcal{A}) = \sum_{\mathcal{Y}} \int_{\mathcal{A}} \int_{\mathcal{X}} \delta(x, d\tilde{x})\, \kappa_M(\tilde{y}, dy)\, E(y, dx)\, \tilde{\pi}(d\tilde{y}) \tag{4.1}$$

$$= \int_{\mathcal{A}} \left[\left(\kappa_M \circ (E \circ \delta)\right) \times \tilde{\pi} \right] (d\tilde{x}, d\tilde{y}) \neq \int_{\mathcal{A}} \left[\left(\delta \otimes \kappa_M\right) \circ (\pi \times E) \right] (d\tilde{x}, d\tilde{y})\,, \tag{4.2}$$

where now $\tilde{\phi}$ and ϕ are joint probabilities, $\kappa_M(\tilde{y}, dy) = \mathbf{\Pi}_{\tilde{y},y} \in \mathcal{M}(\mathcal{Y}, \mathcal{Y})$, and $\mathcal{Y} \coloneqq [\max(K, M)]$ to get a square matrix. In particular, we underline that $\tilde{\pi}(d\tilde{y})$ is a marginal probability on the corrupted space, while $\pi(dy)$ is on the clean one. It is not specified by Katz-Samuels et al. (2019) how the corrupted marginal probability is obtained, nor whether it is the same one given in input as the clean one. Generally, we can always write the following relationship:

$$\tilde{\pi}(d\tilde{y}) = \int_{\mathcal{Y}} \lambda_M(\hat{y}, d\tilde{y})\, \pi(d\hat{y})\,, \tag{4.3}$$

where $\lambda_M \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$, so the variable $\hat{y}$ is defined on the clean probability space $(\mathcal{Y}, \Sigma(\mathcal{Y}), \pi)$.

Mutually Contaminated Distributions Model is Non-Markovian. The formula derived above classifies MCD as multi-step, because plugging Eq. (4.3) in Eq. (4.1) violates Def. 4.4. This gets even clearer when looking at Eq. (4.2): the right-hand side would imply that there exists a single kernel in $(\delta \otimes \kappa_M)(x, \tilde{y}, d\tilde{x}, dy) \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ representing the MCD corruption scheme, but such representation is not possible because of how the MCD kernel acts on E by definition. In addition, we underline that the existence of $(\delta \otimes \kappa_M)(x, \tilde{y}, d\tilde{x}, dy)$ would still *not make a viable Markovian corruption* in the sense of Def. 4.2 because of the variables not being compatible with the probability ϕ for generating $\tilde{\phi}(d\tilde{x}, d\tilde{y}) = \big((\check{\delta} \otimes \check{\kappa}_M)\phi\big)(d\tilde{x}, d\tilde{y}) = \int_{\mathcal{X}\times\mathcal{Y}} \phi(dx, dy) \left(\delta \otimes \kappa_M\right)(x, \tilde{y}, d\tilde{x}, dy)$, the latter being an ill-posed integral since we have two measures on $y \in \mathcal{Y}$.

Menon et al. (2015) assume it plausible to have a corrupted label marginal totally unrelated to the original clean one; we model this case as a degenerate kernel constantly equal to the output probability, i.e. $\lambda_M(\hat{y}, d\tilde{y}) = \lambda_{\tilde{\pi}}(\hat{y}, d\tilde{y}) = \tilde{\pi}(d\tilde{y})$. The other extreme case is for the corrupted and clean marginals to not differ, and in such a case we are still in the presence of a corruption that is not one-step. This is because, having $\lambda_M(\hat{y}, d\tilde{y}) = \delta(\hat{y}, d\tilde{y})$ we write the marginal

$$\pi(d\tilde{y}) = \int_{\mathcal{Y}} \lambda_M(\hat{y}, d\tilde{y})\, \pi(d\hat{y}) = \int_{\mathcal{Y}} \delta(\hat{y}, d\tilde{y})\, \pi(d\hat{y})\,.$$

Comparison with Class Conditional Noise. We can lastly look at the comparison of MCD with *class-conditional noise* (CCN) to understand more in depth its non-Markovian nature. Clearly we cannot reduce MCD to CCN, as we have already shown above. However, Menon et al. (2015, Section 2.3) prove that CCN can be mapped to the MCD model in the binary case, and claim that CCN is a special case of MCD. Their argument can be trivially extended to the multi-class setting by taking

$$[\kappa_M]_{ij} \coloneqq \frac{[\check{\lambda}_C]_{ij}\tilde{\pi}_j}{\sum_{j=1}^{|\mathcal{Y}|}[\check{\lambda}_C]_{ij}\tilde{\pi}_j}\,,$$

where λ_C is the Markov kernel associated to CCN, and $\delta \otimes \lambda_C$ would be its joint form. In plain words, the usual definition of CCN via λ_C can be manipulated such that the κ_M will subsume the label corruption, and the marginal corruption of the MCD is assumed to be a delta, i.e. $\lambda_M = \delta$. However, it would still act on the clean probability as a non-Markovian corruption, as we have proved above. We therefore gain a new insight on MCD: It cannot be thought as a generalization of CCN; the two models can be equivalent in certain regimes of their parameters, but they are in general non-comparable noise models when written in their respective original definition.

4.2.3 Selection Bias

Another corruption model that has been widely studied in the literature is selection bias. Different definitions have been proposed over the years, as we briefly discuss in Appendix A in comparison with covariate shift.

We show that, under its classical formulation based on the Radon–Nikodym derivative, selection bias cannot be captured within the Markovian corruption framework. In contrast, under a probabilistic formulation, it does fall within this family. These two definitions are therefore incompatible and cannot be jointly assumed in a single theoretical analysis. This illustrates the utility of our kernel-based taxonomy, which classifies corruptions by their probabilistic nature and provides a unified basis for understanding them.

Definition 4.11 (Chapter 3.2, Quiñonero-Candela et al. (2008))**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ and the Borel σ-algebra $\Sigma(\mathcal{Z})$ on $\mathcal{Z}$ form a measurable space. Consider a clean probability space $(\mathcal{Z}, \Sigma(\mathcal{Z}), \phi)$ and a corrupted one $(\mathcal{Z}, \Sigma(\mathcal{Z}), \tilde{\phi})$, from which we aim to learn. We define selection bias as a general corruption such that $\mathfrak{L} = (\ell, \mathcal{H}, \phi)$ and $\tilde{\mathfrak{L}} = (\ell, \mathcal{H}, \tilde{\phi})$, and that fulfills the following conditions:*

1. *Support condition, or absolute continuity of the measures:* $\exists\ \alpha \in L^1(\mathcal{Z}, \Sigma(\mathcal{Z}), \phi)$ *s.t.* $\tilde{\phi}(\mathcal{A}) = \int_{\mathcal{A}} \alpha(z)\phi(dz)\ \ \forall \mathcal{A} \in \Sigma(\mathcal{Z})$*, where* α *is the (almost surely unique) Radon–Nikodym derivative;*

2. *Selection condition:* $\sup_{z\in\mathcal{Z}} \alpha(z) < +\infty$.

Non-Markovian Definition of Selection Bias. Clearly, selection bias can in principle include different instances of our taxonomy, since its type is not specified by its characterizing conditions. We now try to understand if it meets the requirement for being Markovian in the first place. Comparing it with the action of a general $\kappa \in \mathcal{M}(\mathcal{Z}, \mathcal{Z})$ on the input probability ϕ, we get the condition

$$\int_{\mathcal{A}} \int_{\mathcal{Z}} \kappa(z, d\tilde{z})\phi(dz) = \int_{\mathcal{A}} \alpha(z)\phi(dz) \quad \forall \mathcal{A} \in \Sigma(\mathcal{Z})\,.$$

It is easy to check that the kernel satisfying the condition is $\kappa(z, d\tilde{z}) \coloneqq \delta_z(d\tilde{z})\alpha(z)$, which respects the definition of kernel, but does not fulfill the Markov property unless $\alpha(z) = 1\ \forall z \in \mathcal{Z}$. This kernel is defined such that ϕ is corrupted into $\tilde{\phi}$, but it does not preserve mass for every input probability measure, therefore it is not what we are looking for to say that selection bias is a Markovian corruption. Is this κ the only possible guess?

Consider a general $\kappa \in \mathcal{M}(\mathcal{Z}, \mathcal{Z})$. It can be rewritten through its density w.r.t. a suitable measure, i.e.,

$$\tilde{\phi}(\mathcal{A}) = \int_{\mathcal{A}} \int_{\mathcal{Z}} \kappa(z, d\tilde{z})\, \phi(dz) = \int_{\mathcal{A}} \int_{\mathcal{Z}} k(z, \tilde{z})\, \nu(d\tilde{z})\, \phi(dz) \quad \forall \mathcal{A} \in \Sigma(\mathcal{Z})\,,$$

and defining $\beta(\tilde{z}) \coloneqq \int_{\mathcal{Z}} k(z, \tilde{z})\, \phi(dz)$, we obtain $\tilde{\phi}(\mathcal{A}) = \int_{\mathcal{A}} \beta(\tilde{z})\, \nu(d\tilde{z}) \quad \forall \mathcal{A} \in \Sigma(\mathcal{Z})$. Imposing κ to act as selection bias, we get $\beta(z) = \alpha(z)\ \forall z \in \mathcal{Z}$ and $\nu = \phi \in \Delta_{\text{ca}}(Z)$, $\mu \coloneqq \frac{\nu+\phi}{2}$ -a.e. On the other hand, the Markov condition asks

$$\int_{\mathcal{Z}} k(z, \tilde{z})\, \nu(d\tilde{z}) = \int_{\mathcal{Z}} k(z, \tilde{z})\, \phi(d\tilde{z}) = 1 \quad \forall z \in \mathcal{Z} \quad \Rightarrow \quad \beta(z) = \alpha(z) = 1 \quad \forall z \in \mathcal{Z}\,.$$

Hence, we reached a contradiction and proved that selection bias cannot be directly represented as a Markov kernel if we impose it to be acting on probabilities *exactly* as the Radon–Nikodym derivative α. Obviously, there exists a Markovian corruption relating ϕ and $\tilde{\phi}$, since they are probability measures and our exhaustiveness argument holds. Therefore, we can represent selection bias via a one-step Markovian corruption. However, that would not reflect the "natural" definition of selection bias that acts through the weighting function α.

Markovian Definition of Selection Bias. Interestingly, in the same Chapter 3.2 of Quiñonero-Candela et al. (2008), the authors also make use of the probabilistic definition of selection bias, as originally introduced by Rubin (1976). More precisely, they define the probability $\phi_{tr}(\tilde{X} = x, \tilde{Y} = y \mid S = 1)$ as the one generating the training samples, where

S is some Bernoulli selection variable that determines whether a sample is included (or corrupted). The test data are instead drawn from the uncorrupted distribution $\phi \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y})$. In other words, they assume two joint probability spaces on $(\mathcal{X} \times \mathcal{Y} \times \{0,1\})$: one clean, where S and the data are independent, and the marginal data distribution is ϕ; one with dependence, and the data distribution is some $\tilde{\phi}$. It follows that these objects exist:

1. $\pi \in \Delta_{\text{ca}}(\{0,1\})$, the marginal of S w.r.t. the clean joint probability on $(\mathcal{X} \times \mathcal{Y} \times \{0,1\})$;

2. $\kappa_\pi \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \{0,1\}$, a degenerate kernel that assumes the value π regardless of the point in $\mathcal{X} \times \mathcal{Y}$ it is evaluated on.[7] This is the kernel description of the random sampling case, where S is independent of (X, Y);

3. $\kappa_{tr} \colon \{0,1\} \rightsquigarrow \mathcal{X} \times \mathcal{Y}$, the kernel representation of the biased training probability ϕ_{tr}, i.e.,

$$\int_{\mathcal{A} \subseteq \mathcal{X} \times \mathcal{Y}} \kappa_{tr}(s = 1, d\tilde{x}, d\tilde{y}) = \phi_{tr}\Big((\tilde{X}, \tilde{Y}) \in \mathcal{A} \mid S = 1\Big)$$

Thus, we can write

$$\tilde{\phi}(\mathcal{A}) \coloneqq (\check{\kappa}_{tr}\check{\kappa}_\pi\phi)(\mathcal{A}) =: (\check{\kappa}\phi)(\mathcal{A}),$$

for $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{Y}$, while $\tilde{\phi}$ is the training distribution and ϕ the test one. This is therefore a Markovian representation of selection bias, and it is used as definition of the involved probability as if it was interchangeable with Def. 4.11. Having proved that Def. 4.11 leads to a non-Markovian corruption, we know the two definitions of selection bias considered here are at least not using same "representation" for the corruption process. In addition, a Markovian corruption is generally not guaranteed to generate a corrupted distribution respecting the absolute continuity condition. This particular kernel $\kappa = \kappa_\pi \circ \kappa_{tr}$ is defined through the degenerate kernel κ_π, and therefore (the support of) $\tilde{\phi}$ does not at all depend on (the support of) ϕ. The corruption process is only dependent on S and its relationship with $(\tilde{X}, \tilde{Y})$ in a latent variable model. We conclude that the two definitions are not equivalent, and should be used in conjunction only upon careful examination.

[7] This is equivalent to the definition of degenerate kernel we gave in Def. 2.23.

Chapter 5

Equivalent Learning Problems under Corruption

Having identified all the types of one-step Markovian corruptions, a natural subsequent question is how to systematically understand their effects on learning problems. We build upon work from Williamson and Cranko (2024) and use equalities involving Bayes risk (Data Processing Equality) to formalize such effects, and then leverage them to discuss how different corruption types require different mitigation strategies. The mitigation we investigate is loss correction.

5.1 Data Processing Equalities

Recently, in (Williamson and Cranko, 2024), Data Processing Equality results have been studied within the supervised learning framework and from an information-theoretic point of view. They have been inspired by the theory of comparison of statistical experiments (Blackwell, 1951; Torgersen, 1991) and the information-theoretic Data Processing Inequality (Polyanskiy and Wu, 2025), which relates Bayes risk after and before corruption. Their work adapts the theory to machine learning, including more realistic assumptions such as a restricted model class $\mathcal{H}$, a fixed loss of interest ℓ, and a fixed prior distribution. Together with an experiment E, the fixed prior uniquely identifies the joint distribution $\phi \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y})$. More formally, their equalities are of the form

$$\mathbf{BR}_{\ell \circ \mathcal{H}}[\pi \times \tilde{E}] = \mathbf{BR}_{\widetilde{\ell \circ \mathcal{H}}}[\pi \times E] \,,$$

where the notation $\widetilde{(\cdot)}$ indicates the action of some Markov kernel changing an object. In their framework, the quantity $\tilde{E}$ is a corrupted experiment in $\mathcal{M}(\mathcal{Y}, \mathcal{X})$, which is computed either as $E \circ \tau \,, \tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ or as $\lambda \circ E \,, \lambda \in \mathcal{M}(\mathcal{Y}, \mathcal{Y})$. The attainable prediction set is defined as $\ell \circ \mathcal{H} \coloneqq \{ (x, y) \mapsto \ell(\dot{h}(x), y) \,|\, h \in \mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y}) \}$, and its corrupted counterpart $\widetilde{\ell \circ \mathcal{H}}$ is obtained as the action of the kernel τ or λ on the set $\ell \circ \mathcal{H}$—the same kernel acting on E and determining $\tilde{E}$. We will formally give this statement later in Proposition 5.3.

The equality trivially induces an equivalence relation on the space of all possible learning problems, given the bijection between Markov kernels and couplings described in the previous section. In its general form, we write it as

$$(\ell \circ \mathcal{H}, \pi \times \tilde{E}) \equiv_{\text{BR}} (\widetilde{\ell \circ \mathcal{H}}, \pi \times E) \,.$$

Williamson and Cranko (2024) only considered corruption acting on the sole experiment by composition; specifically, they use what is referred to by us as simple $\mathcal{X}$ and $\mathcal{Y}$ corruption. In our contributions we also adopt the equality approach, but relate the clean and corrupted *learning problems* through Bayes risk. We prove equivalences that formally characterize how problems are affected by different kinds of joint corruption; the kinds are identified by our taxonomy.

This class of results cannot be directly considered as part of the comparison of experiments theory, as we are not providing any inequality results and are considering a different setting. Rather, we complement that line of work by identifying when problems are equivalent in terms of appropriate measures of risk, and by analyzing how the structure of these equivalences changes depending on the type of corruption applied.

While the equality of Bayes risks is a direct consequence of formalizing corruptions via Markov kernels, we give precise formulas describing how the corruptions act on prior and posterior distributions, as well as on the attainable prediction set $\ell \circ \mathcal{H}$. Accordingly, the main goal of this section is to present qualitative results in terms of conserved "entropy"[1] between corrupted and clean learning problems, and establish a bridge between the problems themselves. These results also lay the basis for the loss-correction framework in Section § 5.2, as they enable a precise differentiation of the loss correction techniques based on the corruption type.

5.1.1 Preliminary Properties

When introducing Markov kernels in Def. 2.14, we allowed it to be defined on different input and output spaces. However, we also align with a more classical view of kernels as related to Markov chains, considering them a modification of the same set of objects, while rearranging the probability measure defined on it. Hence, a corruption from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{Y}$ has to be considered as a *parameterized version* of the corruption on $\mathcal{Y}$, where the parameter is x. For this reason, we also introduced the partial chain composition operation Def. 2.28, which allows chaining while keeping a specified free parameter. A degenerate sub-case takes place when we deal with a kernel from $\mathcal{X}$ to $\mathcal{Y}$.

Markov kernels prove themselves as a useful modeling choice not only because of their interpretation and flexibility; they also have the property of preserving expectations under certain modifications. For this result to hold, one should assume a certain setting for the considered learning problems. A possibility is to take a more geometrical approach, e.g.

[1] For readers interested in how exactly Bayes risk relates to entropy, we refer to Grünwald and Dawid (2004); Williamson and Cranko (2024); Polyanskiy and Wu (2025).

the one described in Section 19.2 of Aliprantis and Border (2006). Here we choose a less involved one to present—but very similar in the imposed restrictions–and introduce one key assumption on the loss function: we require it to be *bounded*.

This requirement restricts the definition of loss function we gave in § 3; positivity and boundedness together ensure that the dual pairing can be used safely in the following proof. In addition, we know that in this setting the Fubini–Tonelli theorem still holds, which is key to many of the statements provided in this chapter.[2] We can now give a formal statement of the aforementioned property of kernels.

Lemma 5.1. *Consider a bounded loss ℓ, a model $h \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, and a probability distribution ϕ on $(\mathcal{X}\times\mathcal{Y}, \mathcal{X}\times\mathcal{Y})$ standard Borel. Let κ be a joint corruption kernel in $\mathcal{M}(\mathcal{X}\times\mathcal{Y}, \mathcal{X}\times\mathcal{Y})$. Then,*

$$\mathbf{R}_{\phi}[\hat{\kappa}(\ell \circ h)] = \mathbf{R}_{\check{\kappa}\phi}[\ell \circ h].$$

Proof. We know that $\mathbf{R}_{\phi}[f] \coloneqq \int f \, d\phi$, which is the definition of the dual pairing used to connect functions and probability measures. Hence, the lemma holds by applying the adjoint property in Remark 2.19. □

This result is central to enabling a general version of the Data Processing Equality, as it will trivially hold for the unconstrained Bayes risk. What remains unknown is how the *form* of the equality changes, specifically in terms of the clean distribution ϕ and the attainable predictions $\ell \circ \mathcal{H}$, under different types of corruption.

Lastly, we specify that in all the following statements the joint corruption action on the learning problem is written as the superposition $\tau \otimes \lambda$, where $I(\tau) = \mathcal{X}$ and $I(\lambda) = \mathcal{Y}$. Their full signature will be provided in each theorem.

Remark 5.2 *The following section will focus on Bayes risk because of its connection with information theory: The* **BR** *equalities (or, Data Processing Equalities) prove changes in the entropy measure induced by some "learnable information", i.e., the maximum amount of information contained in some distribution w.r.t. a learning problem. We will elaborate more on this point in the upcoming sections. However, Lemma 5.1 shows that the results are valid also for a sub-optimal hypothesis in $\mathcal{H}$. The equivalences proved in the following sections can also be seen as risk induced ones (opposed to Bayes risk ones).*

5.1.2 Existing Result: Data Processing Equalities for Simple Noise

As a first step, we can show that our framework subsumes the existing result proved by Williamson and Cranko (2024). We give our own proof in Appendix D. In our taxonomy, their "combined noise" corresponds to our superposition of simple label and attribute noise, called "combined simple noise".

[2] Note that this is not the most general setting for this theorem to apply, but it is a valid setting.

Proposition 5.3 (Combined simple noise, Williamson and Cranko (2024))**.** *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, \phi)$, $E\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y, and $F\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ its associated posterior such that $\phi = \pi_X \times F$ for a suitable π_X. Let $(\tau\colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on this problem. Then, the kernel action on ϕ can be rewritten as*

$$(\check{\tau} \otimes \check{\lambda})\, \phi = \pi_Y \circ \left[(E \circ \tau) \otimes \lambda\right]$$

or, equivalently,

$$(\check{\tau} \otimes \check{\lambda})\, \phi = \pi_X \circ \left[\tau \otimes (F \circ \lambda)\right]. \tag{5.1}$$

Then, the **BR** *Data Processing Equality,*

$$\left(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\lambda})\, \phi\right) \equiv_{\mathbf{BR}} \left(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \phi\right),$$

holds such that the functions contained in the new attainable prediction set satisfy

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) = \{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_y \circ h)](x) \mid h \in \mathcal{H}\}.$$

A Fundamental Difference Between Label and Attribute Noise. Proposition 5.3 allows us to observe some properties of corruption that will be conserved also in the subsequent, more complex cases. Consider its hypotheses, i.e., $\tau \otimes \lambda = \delta \otimes \lambda$ being a corruption acting only on labels. The formula in Eq. (5.1) therefore tells:

$$\mathbf{BR}_{\ell \circ \mathcal{H}}\left[\pi_X \circ \left(\delta \otimes (F \circ \lambda)\right)\right] = \mathbf{BR}_{(\hat{\lambda}\ell \circ \mathcal{H})}\left[\pi_X \times F\right].$$

When looking at the right-hand side, we see that the λ component only modifies the loss function, and leaves the model class untouched, i.e.,

$$\hat{\lambda}(\ell \circ \mathcal{H}) = \hat{\lambda}\ell \circ \mathcal{H}.$$

That means, *simple label (Markovian) corruptions are equivalent in Bayes risk to loss corruptions*, which is non-Markovian in the sense of Def. 4.2, but induced by a Markov kernel. On the other hand, when considering $\tau \otimes \lambda = \tau \otimes \delta$, we obtain

$$\mathbf{BR}_{\ell \circ \mathcal{H}}\left[\pi_Y \circ \left((E \circ \tau) \otimes \delta\right)\right] = \mathbf{BR}_{\hat{\tau}(\ell \circ \mathcal{H})}\left[\pi_Y \times E\right],$$

and in this case notice that the action of $\kappa = \tau \otimes \lambda$ affects the whole attainable prediction set when considering the Bayes risk on the clean distribution.

5.1.3 Novel Data Processing Equalities for Other Corruptions

We now present the results for each of the remaining corruption combinations in Fig. 4.2. All proofs for this section are in Appendix D. From now on, we will refer to the **BR** equality, defined as

$$\Big(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\lambda})\, \phi\Big) \equiv_{\text{BR}} \Big(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \phi\Big). \tag{5.2}$$

The following two results are a strict generalization of Proposition 5.3.

Theorem 5.4 (2-dependent τ, simple λ)**.** *Let ℓ be a bounded loss function. Consider the learning problem $(\ell, \mathcal{H}, \phi)$ and suppose $E\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ is its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y. Let $(\tau\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on this problem. Then,*

$$(\check{\tau} \otimes \check{\lambda})\, \phi = (\check{\tau} \otimes \check{\lambda})(\pi_Y \times E) = \pi_Y \circ \left[(E \circ_{\mathcal{X}} \tau) \otimes \lambda\right].$$

and the **BR** *Data Processing Equality in Eq. (5.2) holds such that the functions contained in the new attainable prediction set satisfy*

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) = \left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_y \circ h)](x, y) \mid h \in \mathcal{H}\right\}.$$

Here in Theorem 5.4 we have shown the **BR** equality for the experiment E. However, for some corruptions the equalities are better formulated using the posterior kernel F, i.e., the discriminative formulation of a learning problem; this change allows us to gain more insights about the form of the corrupted attainable prediction sets.

Theorem 5.5 (simple τ, 2-dependent λ)**.** *Let ℓ be a bounded loss function. Consider the learning problem $(\ell, \mathcal{H}, \phi)$ and suppose $F\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ is its associated posterior such that $\phi = \pi_X \times F$ for a suitable π_X. Let $(\tau\colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on this problem. Then, the kernel action on ϕ can be written as*

$$(\check{\tau} \otimes \check{\lambda})\, \phi = (\check{\tau} \otimes \check{\lambda})(\pi_X \times F) = \pi_X \circ \left[\tau \otimes (F \circ_{\mathcal{Y}} \lambda)\right],$$

and the **BR** *equivalence Eq. (5.2) holds such that the functions contained in the new attainable prediction set satisfy*

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) = \left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](x) \mid h \in \mathcal{H}\right\}.$$

We can notice, thanks to Theorems 5.4 and 5.5, that when corruption involves dependent structures in the factorization, the loss function or the whole attainable prediction set are modified in a parameterized, *dependent* way. Consider, for instance, the action of $\lambda\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$ on the attainable prediction set, when $\tau = \delta$. By definition, it generates the

measurable functions

$$\hat{\lambda}\ell \circ \mathcal{H} = \left\{(x,y) \mapsto \hat{\lambda}\ell(\dot{h}(x),x,y)\,, h \in \mathcal{H}\right\} = \left\{(x,y) \mapsto (\hat{\lambda}\ell_{(x,y)} \circ h)(x)\right\}\,,$$

which is a strong change in the definition of the loss function considered, although still a valid choice. We additionally underline here that *corruptions on $\mathcal{Y}$ only affect the loss function and do not touch the model class, even in the dependent case.*

The next theorems cover the factorizations involving 1-dependent corruptions. In the first case, we are again forced to use either E or F, depending on the involved factors. We group the two results into one theorem for brevity.

Theorem 5.6 (1-dependent, 2-dependent)**.** *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, \phi)$, suppose $E\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ is its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y, and $F\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ its associated posterior such that $\phi = \pi_X \times F$ for a suitable π_X.*

1. *Let $(\tau\colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on the problem. Then,*

$$(\check{\tau} \otimes \check{\lambda})\,\phi = (\check{\tau} \otimes \check{\lambda})(\pi_Y \times E) = \pi_Y \circ \left[\tau \otimes (E \circ_{\mathcal{X}} \lambda)\right]. \tag{5.3}$$

   The **BR** *equivalence Eq. (5.2) holds such that the functions contained in the new attainable prediction set satisfy*

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) = \left\{(x,y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](y) \mid h \in \mathcal{H}\right\}.$$

2. *Let $(\tau\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \rightsquigarrow \mathcal{Y})$ be a corruption acting on the problem. Then,*

$$(\check{\tau} \otimes \check{\lambda})\,\phi = (\check{\tau} \otimes \check{\lambda})(\pi_X \times F)\pi_X \circ \left[(F \circ_{\mathcal{Y}} \tau) \otimes \lambda\right].$$

   The **BR** *equivalence Eq. (5.2) holds such that the functions contained in the new attainable prediction set satisfy*

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) = \left\{(x,y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_x \circ h)](x,y) \mid h \in \mathcal{H}\right\}.$$

Since the 1-dependent κ and λ combination is a subcase of both previous corruptions, we can prove the result as a simple corollary. Notice that this implies both E and F formulations hold.

Corollary 5.7 (1-dependent τ, general λ)**.** *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, \phi)$, $E\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y, and $F\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ its associated posterior such that $\phi = \pi_X \times F$ for a suitable*

π_X. *Let* $(\tau\colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \rightsquigarrow \mathcal{Y})$ *be a corruption acting on the problem. Then,*

$$(\check{\tau} \otimes \check{\lambda})\,\phi = (\check{\tau} \otimes \check{\lambda})(\pi_Y \times E) = \pi_Y \circ \big[\tau \otimes (E \circ \lambda)\big].$$

or, equivalently,

$$(\check{\tau} \otimes \check{\lambda})\,\phi = (\check{\tau} \otimes \check{\lambda})(\pi_X \times F) = \pi_X \circ \big[(F \circ \tau) \otimes \lambda\big]. \tag{5.4}$$

The **BR** *equivalence Eq. (5.2) holds such that the functions contained in the new attainable prediction set satisfy*

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) = \left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_x \circ h)](y) \mid h \in \mathcal{H}\right\}.$$

In all the theorems involving a 1-dependent corruption, the attainable prediction set is heavily modified. To better understand how, we take a closer look at the functions contained in the clean and corrupted minimization sets. To see it in detail, we first need to slightly rework the notation for the attainable prediction set. Consider the loss function $\ell(\cdot, y)$ as a parameterized one, i.e., $\ell_y(\cdot)\colon \Delta_{\text{ca}}(\mathcal{Y}) \to \mathbb{R}_{\geq 0}$; then, the set

$$\ell \circ \mathcal{H} \coloneqq \left\{ (x, y) \mapsto \ell(\dot{h}(x), y) \,|\, h \in \mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y}) \right\}$$

can be equivalently rewritten as

$$\left\{ (x, y) \mapsto (\ell_y \circ \dot{h})(x) \,|\, h \in \mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y}) \right\}.$$

In Eq. (5.3), we have again the kernel $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$ acting on the loss; hence, we obtain

$$\tilde{\ell}_{(x,y)} = \tilde{\ell}(\cdot, x, y) \coloneqq \hat{\lambda}\ell(\cdot, x, y).$$

Therefore, we are inducing a more general notion of loss, namely

$$\tilde{\ell}\colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}.$$

Additionally, the whole composition with the model h, i.e. $\big(\tilde{\ell}_{(x,y)} \circ h\big)(\tilde{x})$, is modified by the action of $\tau \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$, which "swaps" the input $\tilde{x} \in \mathcal{X}$ with $y \in \mathcal{Y}$ in addition to modifying the function itself. Combining them together, we get the new attainable prediction set containing functions of the form

$$f(x, y) = \big[\hat{\tau}\big(\tilde{\ell}_{(x,y)} \circ h\big)\big](y),$$

which is not anymore comparable to the initial form $(\ell_{\tilde{y}} \circ \dot{h})(\tilde{x})$, nor interpretable as a performance evaluation for the model h.

A similar strong modification is observed for the attainable prediction set in Eq. (5.4), which contains functions of the form $f(x, y) = \big[\hat{\tau}\big(\tilde{\ell}_x \circ h\big)\big](y) \coloneqq \big[\hat{\tau}\big(\hat{\lambda}\ell_x \circ h\big)\big](y)$. That is

caused both by the action of $\lambda \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ on $\ell(\cdot, y)$, which results in a new loss function $\hat{\lambda}\ell_x(\cdot) \coloneqq \hat{\lambda}\ell(\cdot, x)$, as well as the action of τ on $\ell \circ h$.

The final result of the factorization, involving the most complex corruption case.

Theorem 5.8 (2-dependent κ, general λ)**.** *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, \phi)$, and let $(\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on the problem. Then:*

1. *the action of such corruption on the joint probability ϕ is equivalent to the one of the non-factorized joint corruption;*
2. *the* **BR** *Data Processing Equality in Eq. (5.2) holds;*
3. *the functions contained in the new attainable prediction set satisfy*

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) = \left\{(\mathcal{X}, \mathcal{Y}) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](x, y) \mid h \in \mathcal{H}\right\}.$$

This result is due to the full dependence on the joint space $\mathcal{X} \times \mathcal{Y}$ for both τ and λ, making it impossible in general to derive a simplifying decomposition of the action on ϕ via Definitions 2.24, 2.25 and 2.28. However, we can still distinguish the effect of λ on the loss, as achieved in all previous cases, and of τ on the full attainable prediction set. For a detailed analysis and proof, see Appendix D.

5.2 Loss Correction Mitigations from Equalities

We now leverage our corruption framework and the derived Data Processing Equalities to reason about the fundamental question:

Can corruption be mitigated so as to guarantee accurate learning from corrupted data?

In our chosen formalization, the training data are drawn from a corrupted distribution $\tilde{\phi} \coloneqq \check{\kappa}\phi$, while evaluation is performed on data drawn from the original clean distribution ϕ. We define *accurate learning* as the property that an optimal model, obtained via risk minimization over a constrained model class using a corrupted training distribution, coincides with the model trained on the clean distribution with the same model class.

While there is prior work on corruption mitigation, often approached by constructing corrected loss functions tailored to specific types of corruption, such methods are typically limited to a specific noise model. To overcome this limitation, we introduce the notion of *Bayesian inverse* of a corruption kernel, which enables a principled and generalized form of loss correction that, in theory, applies systematically to all one-step Markovian corruptions captured in our taxonomy.

5.2.1 Existing Work on Corruption-Corrected Learning

A vast amount of theoretical research on corruption-corrected learning has been conducted in the fields of *learning with noisy labels* and *learning under distribution shift*. Their goal is to achieve *unbiased learning* from biased data,[3] where "biased data" means corrupted data in our context, while "unbiased" refers to the use of a corrected loss $\tilde{\ell}$ that yields:

$$\mathbf{R}_{\tilde{\phi}}[\tilde{\ell} \circ h] = \mathbf{R}_{\phi}[\ell \circ h], \forall h \tag{5.5}$$

with $\mathbf{R}_P[f] \coloneqq \int f dP$. This equality can be obtained as a direct consequence of Lemma 5.1 when an appropriate loss correction is applied. When the Bayes risk is attained for some $h^* \in \mathcal{H}$, such unbiasedness directly implies accurate learning, formally written as

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathbf{R}_{\tilde{\phi}}[\tilde{\ell} \circ h] \quad \text{and} \quad h^* \in \arg\min_{h \in \mathcal{H}} \mathbf{R}_{\phi}[\ell \circ h], \tag{5.6}$$

and therefore constitutes a stronger requirement. Given that Eq. (5.6) is not always a well-defined condition in our setting, we prefer using Eq. (5.5). We refer to the family of approaches for achieving such a goal collectively as *corruption-corrected learning* (CL).

Here, we examine two well-established approaches to unbiased learning through such loss correction, which serve as comparisons to our framework. We then highlight their limitations and propose a generalized correction scheme grounded in Data Processing Equalities to address them.

Reconstruction-based Method

The first one studies loss correction by means of a *reconstruction matrix* (van Rooyen and Williamson, 2018; Patrini et al., 2017; Natarajan et al., 2013), which is derived from a Markovian corruption $\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$ assumed to be reconstructible. In the finite space case, reconstructibility means that the corruption kernel matrix admits a left inverse λ^* (so $\lambda^*\lambda = I$ on functions over Y). The loss correction is then defined as

$$\tilde{\ell}(\dot{h}(x), \tilde{y}) \coloneqq (\lambda^*\ell)(\dot{h}(x), \tilde{y}) \coloneqq \sum_y \lambda^*_{y\tilde{y}} \ell(\dot{h}(x), y), \quad \lambda \in \mathcal{M}(\mathcal{Y}, \mathcal{Y}), \tag{5.7}$$

where λ^* is the reconstruction matrix, and $\lambda^*\ell$ the matrix acting on the loss viewed as a function of the label. This yields unbiased learning as per Eq. (5.5), since an analogue of van Rooyen and Williamson (2018)'s Theorem 5 holds. In our notation, for all $x \in \mathcal{X}$ and $h \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$,

$$\begin{aligned}\mathbb{E}_{\tilde{Y} \sim F \circ \lambda}[\tilde{\ell}(\dot{h}(x), \tilde{Y})] &\overset{L5.1}{=} \mathbb{E}_{Y \sim F}[\lambda(\tilde{\ell}(\dot{h}(x), Y)] \\ &\overset{(5.7)}{=} \mathbb{E}_{Y \sim F}[(\lambda^*\lambda)(\ell(\dot{h}(x), Y)] = \mathbb{E}_{Y \sim F}[\ell(\dot{h}(x), Y)].\end{aligned}$$

[3] Existing literature defines unbiased learning via unbiasedness of the empirical risk estimator. This notion is trivially implied by our Eq. (5.5).

This method is called *backward correction* by Patrini et al. (2017), and the *method of unbiased estimators* by Natarajan et al. (2013).

Such a reconstruction matrix λ^* is in general not a Markov kernel and may contain negative entries, which can make the corrected loss negative and cause problems for optimization. Moreover, although the underlying framework in these works is similar to ours, they can only handle simple Y corruption, i.e., $\delta \otimes \lambda$, $\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$, $\delta \colon \mathcal{X} \rightsquigarrow \mathcal{X}$, and do not generalize to more complex joint corruption cases.

Importance-weighting-based Method

A second line of work corrects loss functions through *importance weighting* (IW) (Shimodaira, 2000; Cortes et al., 2010; Sugiyama and Kawanabe, 2012), originally developed for *covariate shift*, where the input distribution is corrupted by τ such that $I(\tau) = \mathcal{X}$ while the conditional distribution F remains unchanged. Under model misspecification (White, 1981),[4] IW provides a principled correction by requiring the clean data distribution to be absolutely continuous w.r.t. the corrupted one, i.e., $\phi \ll \tilde{\phi}$. The corrected loss takes the form

$$\tilde{\ell}(\dot{h}(x), \tilde{y}) \coloneqq w(x)\, \ell(\dot{h}(x), \tilde{y}), \;\; w(x) \coloneqq \frac{d\phi}{d\tilde{\phi}}(x), \tag{5.8}$$

where $w(x)$ is called the *importance weight*, typically expressed via densities w.r.t. the Lebesgue measure. This guarantees the unbiased learning goal in Eq. (5.5).

The key limitations of IW are that it applies directly only to covariate shift, and that it depends critically on the assumptions of absolute continuity as well as model misspecification. Although recent work extends IW to joint corruptions by weighting with $w(x, y) \coloneqq \frac{dP}{d\tilde{\phi}}(x, y)$ over the joint distribution (Liu and Tao, 2015; Fang et al., 2020, 2023), these methods remain restricted by the absolute continuity requirement (at least on the overlapping support of ϕ and $\tilde{\phi}$), which is essential for the importance weights to be well-defined.

At first glance, the IW correction formula Eq. (5.8) looks similar to the ones we derive later, however, it is not a subcase of our kernel-based loss correction. IW operates by reweighting losses through the Radon-Nikodym derivatives $\frac{dP}{d\tilde{\phi}}$, whereas our framework performs correction through kernel inversion. The two mechanisms are fundamentally different, as Radon–Nikodym derivatives exist only under absolute continuity assumptions, while kernel-based loss corrections are free from such restrictions and can systematically handle general corruptions beyond that. In addition, Radon–Nikodym derivatives cannot be represented as Markov kernels, as we have seen in § 4.2.3.

5.2.2 A Generalized Framework for Corruption-Corrected Learning

In CL, existing approaches such as reconstruction-based or importance-weighting based methods are either designed for specific types of corruption, or rely on restrictive assump-

[4] By contrast, when the model is well specified, standard empirical risk minimization remains consistent under covariate shift and IW provides no benefit.

tions, and therefore cannot resolve the question posed above in a systematic or general way. To this end, we introduce the concept of the Bayesian inverse of a Markov kernel, which enables a systematic analysis of CL across the wide range of corruptions in our taxonomy.

Bayesian Inverse of a Markov Kernel

To study the CL problem within our framework, we first define a principled way to reverse the corruption process. We introduce here the Bayesian inverse of a Markov kernel (Dahlqvist et al., 2016; Cho and Jacobs, 2019), which preserves the Markov property and yields a mathematically well-defined mechanism for inverting corruptions.

Definition 5.9 (Bayesian Inverse). *Let $\phi \in \Delta_{\text{ca}}(\mathcal{Z}_1)$ be a probability distribution, and $\kappa \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ be a Markov kernel with the property $\check{\kappa}\phi = \tilde{\phi}$ for some $\tilde{\phi} \in \Delta_{\text{ca}}(\mathcal{Z}_2)$. The* Bayesian inverse *of κ is defined as the Markov kernel $\kappa^\dagger \in \mathcal{M}(\mathcal{Z}_2, \mathcal{Z}_1)$ such that it induces together with $\tilde{\phi}$ the same coupling induced by κ and ϕ, i.e.,*

$$(\phi \times \kappa)(\mathcal{A} \times \mathcal{B}) = (\tilde{\phi} \times \kappa^\dagger)(\mathcal{A} \times \mathcal{B}), \ \forall\, \mathcal{A} \times \mathcal{B} \in \Sigma(\mathcal{Z}_1) \times \Sigma(\mathcal{Z}_2).$$

By taking $\mathcal{A}$ or $\mathcal{B}$ equal to the $\mathcal{Z}_2$ or $\mathcal{Z}_1$, we respectively get the property of $\kappa^\dagger$ and κ being a "weak" inverse of each other.

Proposition 5.10. *Let $\kappa\colon \mathcal{Z}_1 \rightsquigarrow \mathcal{Z}_2$ be a Markov kernel with the property $\tilde{\phi} = \check{\kappa}\phi$ for $\phi \in \Delta_{\text{ca}}(\mathcal{Z}_1)$, $\tilde{\phi} \in \Delta_{\text{ca}}(\mathcal{Z}_2)$, and $\kappa^\dagger$ its Bayesian inverse. Then, $\kappa^\dagger$ reverses the action of κ for the fixed input and output probabilities, i.e.,*

$$\phi(\mathcal{A}) = (\check{\kappa}^\dagger \tilde{\phi})(\mathcal{A}), \ \forall\, \mathcal{A} \in \Sigma(\mathcal{Z}_2).$$

Remark 5.11 *Recalling how we have defined the posterior kernel $F \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ in Def. 3.4,*

$$\pi_X \times F = \pi_Y \times E = \phi \in \Delta_{\text{ca}}(\mathcal{X} \times \mathcal{Y}),$$

we notice that $F = E^\dagger$. We will use this notation in the rest of this thesis.

We will refer to the Bayesian inverse of the corruption kernel as the *cleaning kernel*. In general, the Bayesian inverse is not unique, since it corresponds to a class of equivalence induced by the probability measures on $\mathcal{Z}_1$ and $\mathcal{Z}_2$. However, we are always sure it exists given the assumption of using standard Borel measure spaces when defining Markov kernels (more details in Appendix C). This is a weaker existence condition w.r.t. the one given for the reconstruction matrix.[5]

[5] Such a weaker condition comes at the price of the Bayesian inverse kernel being a *typed inverse*, meaning that to compute $\kappa^\dagger$, we need not only the kernel κ but also the initial probability ϕ; having a different ϕ induces a different Bayesian inverse. This point becomes clearer when considering the discrete case, as explained in Remark 5.12.

Remark 5.12 *In the discrete case, the Bayesian inverse always exists and is defined by Bayes rule. Furthermore, it is uniquely defined $\tilde{\phi}$-a.s. This is easy to see by unfolding Def. 5.9 into:*

$$\int_B \kappa(z_1, A)\, \phi(dz_2) = \int_A \kappa^\dagger(z_2, B)\, \tilde{\phi}(dz_2) \quad \forall A \in \Sigma(\mathcal{Z}_2),\ \forall B \in \Sigma(\mathcal{Z}_1).$$

This formulation extends the discrete Bayes' rule $\phi(z_2 \mid z_1)\phi(z_1) = \phi(z_1 \mid z_2)\phi(z_2)\ \forall z_1, z_2$. Indeed, in the discrete case, the Bayesian inverse always exists and can be expressed as

$$\kappa^\dagger(z_1 \mid z_2) \coloneqq \frac{\phi(z_1)\kappa(z_2 \mid z_1)}{\tilde{\phi}(z_2)} \qquad \forall z_1, z_2 \ \ s.t. \ \ \tilde{\phi}(z_2) \neq 0.$$

This formula ensures the uniqueness of $\kappa^\dagger$ within the support of $\tilde{\phi}$, as all components are unique. However, outside the support when $\tilde{\phi}$ is zero, the uniqueness may not hold.

The Bayesian inversion operation has the following desirable property of preserving the expectations.

Proposition 5.13 (Inverse Data Processing Equality)**.** *Consider a learning problem $(\ell, \mathcal{H}, \phi)$ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$, a corruption $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$, and a $f \in \ell \circ \mathcal{M}(\mathcal{X}, \mathcal{Y})$. Let ℓ be a bounded loss function. Then,*

$$\mathbf{R}_\phi[f(Z)] = \mathbf{R}_{\check{\kappa}\phi}[\hat{\kappa}^\dagger f(\tilde{Z})], \tag{5.9}$$

and

$$\mathbf{BR}_{\ell \circ \mathcal{H}}\big(\phi\big) = \mathbf{BR}_{\hat{\kappa}^\dagger(\ell \circ \mathcal{H})}\big(\check{\kappa}\phi\big),$$

where $\kappa^\dagger$ is the cleaning kernel and $(\hat{\kappa}^\dagger(\ell \circ \mathcal{H}), \check{\kappa}\phi)$ the (corruption) corrected problem.

Proof. The first part of the proof follows by applying Lemma 5.1 for $\check{\kappa}^\dagger \check{\kappa}\phi = \phi$, where the equality holds because of Proposition 5.10. The second is proved by taking the infimum over $\ell \circ \mathcal{H}$ of both sides of Eq. (5.9). □

Remark 5.14 *Notice that, even when a risk minimizer exists, Proposition 5.13 does not imply $\ell \circ h^* = \hat{\kappa}^\dagger(\ell \circ h^*)$, but only that their risks coincide, i.e., $\mathbb{E}_\phi(\ell \circ h^*) = \mathbb{E}_{\check{\kappa}\phi}[\hat{\kappa}^\dagger(\ell \circ h^*)]$ as functions of $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Under certain conditions on the learning problem, some minimizers $h^*, \tilde{h}^* \in \mathcal{H}$ exist such that $\ell \circ h^* = \hat{\kappa}^\dagger(\ell \circ \tilde{h}^*)$, but it is not necessarily the case that $\tilde{h}^* = h^*$.*

In the following, we will make use of these facts assuming $\mathcal{Z}_1 = \mathcal{Z}_2 = \mathcal{X} \times \mathcal{Y} =: \mathcal{Z}$, to match the setting of our taxonomy of corruption.

Label Corruption Correction with Bayesian Inverse

Similarly to the reconstruction-matrix and importance-weighting approaches, we enforce the condition in Eq. (5.5) and derive a principled loss correction method based on the Bayesian inverse $\kappa^\dagger$ of the corruption kernel κ.

We first illustrate this in the case of simple label corruption $\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$. Consider its associated cleaning kernel sending $\tilde{\mathfrak{L}} = (\tilde{\ell}, \mathcal{H}, \tilde{\phi})$ to the clean one $\mathfrak{L} = (\ell, \mathcal{H}, \phi)$, i.e., the

Bayesian inverse $\lambda^\dagger \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$. By considering a corruption $\delta \otimes \lambda$, $\delta \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ and its inverse $\delta \otimes \lambda^\dagger$, Proposition 5.13 allows us to write

$$\mathbf{R}_{\tilde{\phi}}(\hat{\lambda}^\dagger \ell \circ h) = \mathbf{R}_{\phi}(\ell \circ h), \quad \text{where } \tilde{\phi} \coloneqq (\check{\delta}_X \otimes \check{\lambda})\phi, \quad \forall\, h \in \mathcal{H}\,.$$

This directly yields the loss correction formula:

$$\tilde{\ell}(\dot{h}(x), \tilde{y}) \coloneqq (\hat{\lambda}^\dagger \ell)(\dot{h}(x), \tilde{y}), \quad \lambda^\dagger \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}. \tag{5.10}$$

The corrected loss satisfies the unbiased learning criterion in Eq. (5.5), and hence also achieves the accurate learning goal in Eq. (5.6) when the minimizer exists.

Notably, in this simple label corruption case, our kernel-based correction in Eq. (5.10) resembles the reconstruction-matrix correction in Eq. (5.7). However, the two differ fundamentally, as illustrated below.

Example 5.15. *Following van Rooyen and Williamson (2018, Sec. 4.1.2), consider symmetric label noise in binary classification setting. The clean distribution is given by $\phi = (p_1, p_2)^\top$, while the corruption kernel is*

$$\lambda = \begin{bmatrix} \sigma & 1-\sigma \\ 1-\sigma & \sigma \end{bmatrix}.$$

Its associated reconstruction matrix exists, and it amounts to

$$\lambda^* = \frac{1}{1-2\sigma} \cdot \begin{bmatrix} 1-\sigma & -\sigma \\ -\sigma & 1-\sigma \end{bmatrix},$$

while the Bayesian inverse takes the form

$$\lambda^\dagger = \begin{bmatrix} \frac{p_1(1-\sigma)}{p_1(1-\sigma)+p_2\sigma} & \frac{p_1\sigma}{p_2(1-\sigma)+p_1\sigma} \\ \frac{p_2\sigma}{p_1(1-\sigma)+p_2\sigma} & \frac{p_2(1-\sigma)}{p_2(1-\sigma)+p_1\sigma} \end{bmatrix}.$$

The two objects are clearly different, with the Bayesian inverse approach requiring the knowledge of the clean probability values in order to compute the cleaning matrix.

We note that in simple corruption settings, such as $\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$ as discussed above, the Bayesian inverse naturally takes the form $\lambda^\dagger \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$. However, for a joint corruption kernel of the form $\kappa = \tau \otimes \lambda$, the Bayesian inverse does not, in general, distribute over $\otimes$; that is, one typically has $\kappa^\dagger \neq \tau^\dagger \otimes \lambda^\dagger$. Consequently, in what follows, we assume that $\kappa^\dagger$ is *one-step Markovian* and therefore admits the form $\tau \otimes \lambda$.

We now extend this paradigm to the $\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$ case, as from Theorem 5.5 we see that also in the dependent case, label corruption only affects the loss function.

Theorem 5.16. *Let $(\ell, \mathcal{H}, \phi)$ be a clean learning problem with ℓ being a bounded loss function, and κ a Markovian corruption on it. Let $\kappa^\dagger = \delta \otimes \lambda$, $\delta \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ be the cleaning kernel reversing κ, such that $(\hat{\kappa}^\dagger(\ell \circ \mathcal{H}), \check{\kappa}\phi)$ is its associated corrected problem. When $\lambda \in \mathcal{M}(\mathcal{Y}, \mathcal{Y})$, we have*

$$\tilde{\ell}(\dot{h}(\tilde{x}), \tilde{y}) \coloneqq (\hat{\lambda}\ell)(\dot{h}(\tilde{x}), \tilde{y}) \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y} ,$$

while, when $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$ we have a more general notion of loss, i.e.,

$$\tilde{\ell}(\dot{h}(\tilde{x}), \tilde{x}, \tilde{y}) \coloneqq (\hat{\lambda}\ell)(\dot{h}(\tilde{x}), \tilde{x}, \tilde{y}) \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y} ,$$

with $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Similarly, when $\lambda \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$,

$$\tilde{\ell}(\dot{h}(\tilde{x}), \tilde{x}) \coloneqq (\hat{\lambda}\ell)(\dot{h}(\tilde{x}), \tilde{x}) \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y} ,$$

with $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{X} \to \mathbb{R}_{\geq 0}$.

Proof. Let us consider a general function in the set $\hat{\kappa}^\dagger(\ell \circ \mathcal{H})$. Assuming $\kappa^\dagger = \delta \otimes \lambda$, $\delta(\tilde{x}, dx) \in \mathcal{M}(\mathcal{X}, \mathcal{X})$, it will take the form

$$\hat{\kappa}^\dagger(\ell \circ h)(\tilde{x}, \tilde{y}) = (\hat{\delta} \otimes \hat{\lambda})(\ell \circ h)(\tilde{x}, \tilde{y}) \coloneqq \sum_{y \in Y} \int_{x \in X} \ell(\dot{h}(x), y)\, \delta(\tilde{x}, dx)\, \lambda(\tilde{x}, \tilde{y}, dy) .$$

Now let λ be a 2-dependent label corruption, i.e., in $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$. We can define the loss correction $\tilde{\ell}$ as

$$\begin{aligned}(\dot{h}(\tilde{x}), \tilde{x}, \tilde{y}) \mapsto \tilde{\ell}(\dot{h}(\tilde{x}), \tilde{x}, \tilde{y}) &\coloneqq \sum_{y \in Y} \int_{x \in X} \ell(\dot{h}(x), y)\, \delta(\tilde{x}, dx)\, \lambda(\tilde{x}, \tilde{y}, dy) \\ &= \int_{x \in X} (\hat{\lambda}\ell)(\dot{h}(x), \tilde{x}, \tilde{y})\, \delta(\tilde{x}, dx) = (\hat{\lambda}\ell)(\dot{h}(\tilde{x}), \tilde{x}, \tilde{y}),\end{aligned}$$

where $(\dot{h}(\tilde{x}), \tilde{x}, \tilde{y}) \in \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{X} \times \mathcal{Y}$. Hence, the case $\lambda(\tilde{x}, \tilde{y}, dy)$ and its subcases $\lambda(\tilde{x}, dy)$, $\lambda(\tilde{y}, dy)$ combined with an identity kernel on $\mathcal{X}$ do not change the hypothesis function. Lastly, by definition of Markov transition, we have that when $\lambda \in \mathcal{M}(\mathcal{Y}, \mathcal{Y})$

$$\tilde{\ell}(\dot{h}(\tilde{x}), \tilde{y}) \coloneqq \sum_{y \in Y} \int_{x \in X} \ell(\dot{h}(x), y)\, \delta(\tilde{x}, dx)\, \lambda(\tilde{y}, dy) = (\hat{\lambda}\ell)(\dot{h}(\tilde{x}), \tilde{y}),$$

which respects the classical definition of loss function, as $(\dot{h}(\tilde{x}), \tilde{y}) \in \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y}$. This proves the thesis, as by construction $\tilde{\ell}(\dot{h}(\tilde{x}), \tilde{x}, \tilde{y}) \in \hat{\kappa}^\dagger(\ell \circ \mathcal{H})$. Proposition 5.13 implies that $\hat{\kappa}^\dagger(\ell \circ \mathcal{H})$ makes Eq. (5.5) hold. □

For the dependent label noise cases, we need to relax the definition of loss function, and allow it to take x as an additional input. This generalization is not unprecedented; for example, it has been employed in Steinwart and Christmann (2008).

Generalized Corruption-Corrected Learning with Bayesian Inverse

For corruptions that involve more than just label corruption, we cannot obtain loss corrections using the same proof technique as Theorem 5.16. This is because, when applying the risk conservation formula in Proposition 5.13, the model class itself is also affected by the corruption kernel. Formally, assume $\kappa^\dagger = \tau \otimes \lambda$ with $I(\tau) = \mathcal{X}$ and $I(\lambda) = \mathcal{Y}$. Then, Eq. (5.5) becomes

$$\begin{aligned}\mathbf{R}_\phi(\ell \circ h) &= \mathbf{R}_{\check{\kappa}\phi}\left[\hat{\kappa}^\dagger(\ell \circ h)\right] \\ &= \mathbf{R}_{\check{\kappa}\phi}\left[(\hat{\tau} \otimes \hat{\lambda})(\ell \circ h)\right], \quad \forall h \in \mathcal{M}(\mathcal{X}, \mathcal{Y}),\end{aligned} \tag{5.11}$$

which does not necessarily imply Eq. (5.6). In Eq. (5.11), the corruption effect on loss and model class is in general *indistinguishable*; we are not able to rewrite the set $(\hat{\tau} \otimes \hat{\lambda})(\ell \circ \mathcal{H})$ as the composition $\tilde{\ell} \circ \tilde{\mathcal{H}}$, let alone $\tilde{\ell} \circ \mathcal{H}$ as per the CL case. Nevertheless, our framework still allows for some new understanding on how attribute corruption may be corrected.

To this end, we formalize a new version of the CL paradigm, requiring to find a loss correction formula $\tilde{\ell}$ that depends on ℓ, h and $\kappa^\dagger$. That is the *generalized corruption-corrected learning* (GCL) paradigm, defined as the condition

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathbf{R}_{\tilde{\phi}}[\tilde{\ell} \circ h] \quad \text{and} \quad h^* \in \arg\min_{h \in \mathcal{H}} \mathbf{R}_\phi[\ell \circ h], \tag{5.12}$$

with the additional *factorization requirement* for $\hat{\kappa}^\dagger(\ell \circ h)$, i.e.,

$$\tilde{\ell}(h, x, y) = \hat{\kappa}^\dagger(\ell \circ h)(x, y), \quad \forall h \in \mathcal{M}(\mathcal{X}, \mathcal{Y}), x \in \mathcal{X}, y \in \mathcal{Y}, \tag{5.13}$$

where $\tilde{\ell}$ is a *generalized loss* $\tilde{\ell} \colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.[6]

A key assumption made in the above paradigm is the existence of a minimizer for the problems $(\ell, \mathcal{H}, \phi)$ and $(\tilde{\ell}, \mathcal{H}, \phi)$; hence, for it to be used, a practitioner would need to impose additional conditions. We will discuss the feasibility of the paradigm after presenting the loss correction formulas. For now, we impose an additional assumption and show that it is enough to imply the well-posedness of GCL.

A1 Given the clean learning problem $(\ell, \mathcal{H}, \phi)$, there exists a minimizer $h^* \in \mathcal{H}$ that attains the Bayes risk value associated to it, i.e., $\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) = \mathbf{R}_\phi[\ell \circ h^*]$.

Lemma 5.17. *Let $(\ell, \mathcal{H}, \phi)$ be a clean learning problem respecting A1, and $\kappa = \tau \otimes \lambda$ a Markovian corruption kernel in $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ on it. Denote the risk minimizer for the clean problem as $h^* \in \mathcal{H}$. Then, if Eq. (5.13) is fulfilled by some $\tilde{\ell}$, h^* is also the minimizer of the GCL-corrected problem $(\tilde{\ell}, \mathcal{H}, \check{\kappa}\phi)$, as per Eq. (5.12).*

Proof. Consider the problem $(\hat{\kappa}^\dagger(\ell \circ \mathcal{H}), \check{\kappa}\phi)$ and assume that there is some (possibly

[6] As for the requirements of the factorization on the whole set of Markov kernels, it can be weakened for instance to $\mathcal{H}$ if we know that the minimizer is contained in the set.

generalized) loss function $\tilde{\ell}$ such that Eq. (5.13). By Proposition 5.13 and Eq. (5.13),

$$\mathbf{R}_{\phi}(\ell \circ h) \overset{P5.13}{=} \mathbf{R}_{\check{\kappa}\phi}\left[(\hat{\tau} \otimes \hat{\lambda})(\ell \circ h)\right] \overset{(5.13)}{=} \mathbf{R}_{\check{\kappa}\phi}(\tilde{\ell} \circ h),$$

which, taking infimum on the right- and left-most equations, implies the thesis. □

Hence, finding loss correction formulas respecting Eq. (5.13) and A1 allows us to use the GCL correction paradigm.

Loss Correction Formulas. We now give the correction results for all the corruption case lying within the GCL paradigm, while deferring their computation to Appendix E. Recall that the notation $f\#\mu$ stands for the push-forward probability measure of the distribution μ through the function f, defined as $(f\#\mu)(\mathcal{A}) \coloneqq \mu(f^{-1}(\mathcal{A}))$ for a suitable set $\mathcal{A}$. By definition of kernel, τ is a measure when fixing $\tilde{x} \in \mathcal{X}$, and by Proposition 2.16 $\dot{h} \colon \mathcal{X} \to \Delta_{\mathrm{ca}}(\mathcal{Y})$ is a measurable function; hence, we can write

$$(\dot{h}\#\tau)(\tilde{x})(\mathcal{A}) \coloneqq \tau(\tilde{x}, \dot{h}^{-1}(\mathcal{A})),\ \mathcal{A} \subseteq \Delta_{\mathrm{ca}}(\mathcal{Y}),$$

that is a family of distributions defined on a set of probability measures on $\mathcal{Y}$, evaluated on $\mathcal{A}$ and indexed by $\tilde{x}$. Since it is indexed and induced by Markov kernels, we can see it as a posterior probability on the set $\Delta_{\mathrm{ca}}(\mathcal{Y})$, given $\tilde{x}$.

Theorem 5.18. *Let $(\ell, \mathcal{H}, \phi)$ be a clean learning problem with $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ being a bounded loss function and such that A1 holds. Let $\kappa^{\dagger} = \tau \otimes \lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ be the Markovian cleaning kernel reversing κ, such that $(\hat{\kappa}^{\dagger}(\ell \circ \mathcal{H}), \check{\kappa}\phi)$ is its associated corrected problem. Then, we can find a generalized loss $\tilde{\ell} \colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ respecting the GCL paradigm. In particular:*

1. *When $\kappa^{\dagger}$ is of the form $(\tau \colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$, or $(\tau \colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$, or $(\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$, we have*

   $$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{x})}[\hat{\lambda}\ell(u, \tilde{y})] \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\,.$$

   When both corruptions are simple, the $\hat{\lambda}\ell$ formula remains unchanged. When λ is 2-dependent, it induces $\hat{\lambda}\ell(u, \tilde{x}, \tilde{y})$. Lastly, we get $(\dot{h}\#\tau)(\tilde{x}) = (\dot{h}\#\tau)(\tilde{x}, \tilde{y})$ when τ is 2-dependent.
2. *When $\kappa^{\dagger}$ is of the form $(\tau \colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \rightsquigarrow \mathcal{Y})$, we have*

   $$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{y})}[\hat{\lambda}\ell(u, \tilde{x})] \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\,.$$

3. *When $\kappa^{\dagger}$ is of the form $(\tau \colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$, or $(\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes$*

$(\lambda\colon \mathcal{X} \rightsquigarrow \mathcal{Y})$, *we respectively have*

$$\tilde{\ell}(h,\tilde{x},\tilde{y}) \coloneqq \mathbb{E}_{u\sim(\dot{h}\#\tau)(\tilde{y})}[\hat{\lambda}\ell\,(u,\tilde{x},\tilde{y})] \quad \forall\,(\tilde{x},\tilde{y}) \in \mathcal{X}\times\mathcal{Y}\,;$$
$$\tilde{\ell}(h,\tilde{x},\tilde{y}) \coloneqq \mathbb{E}_{u\sim(\dot{h}\#\tau)(\tilde{x},\tilde{y})}[\hat{\lambda}\ell\,(u,\tilde{x})] \quad \forall\,(\tilde{x},\tilde{y}) \in \mathcal{X}\times\mathcal{Y}\,.$$

4. *When* $\kappa^{\dagger}$ *is of the form* $(\tau\colon \mathcal{X}\times\mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X}\times\mathcal{Y} \rightsquigarrow \mathcal{Y})$, *we have*

$$\tilde{\ell}(h,\tilde{x},\tilde{y}) \coloneqq \mathbb{E}_{u\sim(\dot{h}\#\tau)(\tilde{x},\tilde{y})}[\hat{\lambda}\ell\,(u,\tilde{x},\tilde{y})] \quad \forall\,(\tilde{x},\tilde{y}) \in \mathcal{X}\times\mathcal{Y}\,.$$

Discussion of the GCL Paradigm and Its Correction Formulas. Theorem 5.18 provides a constructive proof for the existence of GCL corrected losses, i.e., that respect both Eq. (5.12) and Eq. (5.13), granted that A1 holds. What is therefore shown is that the GCL paradigm is closely related to the CL one, but it is only defined for certain classes of learning problem, as it imposes the existence of a minimizer s.t. the condition in Eq. (5.12) is well-posed. Within our setup, laid out in § 2.2, we could fulfill the assumption by requiring the loss ℓ not only to be positive, measurable and bounded but also *continuous* on $\Delta_{\mathrm{ca}}(\mathcal{Y})$ in the Euclidean topology, and $\dot{\mathcal{H}} \coloneqq \{\dot{h} \mid h \in \mathcal{H}\} \subsetneq \mathcal{B}_b(\mathcal{X})$ to be *compact* in the norm topology. This would imply continuity of the functional $\dot{h} \in \dot{\mathcal{H}} \mapsto \mathbf{R}_{\phi}[\ell(\dot{h}(x),y)]$, and therefore the existence of a minimizer on the compact set $\dot{\mathcal{H}}$.

While GCL is on one hand adding more restrictive conditions, it is also weakening some of the CL requirements: it asks the corrected loss to be only a generalized loss, which is then used in the factorization condition Eq. (5.13). Having a generalized loss is necessary for the attribute corruption, as standard losses may fail to mitigate this case. The following example showcases one of such scenarios.

Example 5.19. *Consider a simple neural network with one hidden layer of two ReLU neurons and a softmax output,*

$$\dot{h}\colon \mathcal{X} \to \Delta_{\mathrm{ca}}(\mathcal{Y}), \qquad \dot{h}(\boldsymbol{x}) \coloneqq \mathrm{softmax}\big(W_2\,\sigma_{ReLU}(W_1\boldsymbol{x}+b_1)+b_2\big),$$

where $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^2$, $y \in \mathcal{Y} = \{0,1\}$, $W_1 \in \mathbb{R}^{2\times 2}$, $b_1 \in \mathbb{R}^2$, $W_2 \in \mathbb{R}^{2\times 2}$, $b_2 \in \mathbb{R}^2$, *and* $\sigma_{ReLU}(z) = \max(0,z)$. *Writing the hidden activations* n_j *and logits explicitly, we get*

$$n_j(\boldsymbol{x}) = \max\{0, \langle [W_1]_j, \boldsymbol{x}\rangle + b_{1,j}\}, \quad j=1,2,$$
$$z_i(\boldsymbol{x}) = \langle [W_2]_i, \boldsymbol{n}(\boldsymbol{x})\rangle + b_{2,i},$$
$$\dot{h}(\boldsymbol{x})_i = \frac{e^{z_i(\boldsymbol{x})}}{\sum_{j=1}^{2} e^{z_j(\boldsymbol{x})}}.$$

Here $[W_1]_j$ *denotes the* j*-th row of* W_1 *and* $[W_2]_i$ *denotes the* i*-th row of* W_2. *For simplicity and without loss of generality, we can choose* $b_{2,i} = 0$ *and the weights* W_1 *to be always positive.*

For our argument, we want to consider the points in $\mathcal{X} \subset \mathbb{R}^2$ *for which the ReLU activation makes the model* $\dot{h}(x)$ non injective*: that happens if both pre-activations* $\langle [W_1]_j, \boldsymbol{x}\rangle + b_{1,j}$ *are negative,*

i.e., for the set of inputs

$$\mathcal{S}_h := \left\{ \boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2 \;\middle|\; x_2 < \min\left(\frac{-b_{1,1} - [W_1]_{1,1} x_1}{[W_1]_{1,2}}, \;\; \frac{-b_{1,2} - [W_1]_{2,1} x_1}{[W_1]_{2,2}} \right) \right\}.$$

Hence, we get that $n_1(\boldsymbol{x}) = n_2(\boldsymbol{x}) = 0$*, and therefore* $\dot{h}(\boldsymbol{x}) = (0.5, 0.5)$*. This implies that*

$$\ell(\dot{h}(\boldsymbol{x}_1), y) = \ell(\dot{h}(\boldsymbol{x}_2), y), \quad \forall y \in \mathcal{Y}, \; \forall \boldsymbol{x}_i \in \mathcal{S}_h.$$

However, if we apply a deterministic attribute corruption kernel, i.e., $\kappa^\dagger := \delta_{f(\boldsymbol{x})} \otimes \delta$ *with* $f : \mathcal{X} \to \mathcal{X}$ *measurable and* $\delta \in \mathcal{M}(\mathcal{Y}, \mathcal{Y})$*, we have that*

$$\hat{\kappa}^\dagger(\ell \circ \mathcal{H}) = \Big\{ \ell(\dot{h}(f(\boldsymbol{x})), y) \mid h \in \mathcal{H} \Big\}.$$

Imposing $\tilde{\ell}(\dot{h}(\boldsymbol{x}), y) = \ell(\dot{h}(f(\boldsymbol{x})), y)$ *as per the* CL *definition (i.e., stretching Eq. (5.10) to the attribute corruption case), with* $\tilde{\ell} : \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$*, implies that*

$$\ell\Big(\dot{h}(f(\boldsymbol{x}_2)), y\Big) = \tilde{\ell}\Big(\dot{h}(\boldsymbol{x}_2), y\Big) = \tilde{\ell}\Big(\dot{h}(\boldsymbol{x}_1), y\Big) = \ell\Big(\dot{h}(f(\boldsymbol{x}_1)), y\Big),$$

for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{S}_h$ *and for all* $y \in \mathcal{Y}$*. This yields a contradiction. Indeed, since* $\dot{h}(x)$ *is constant on* $\mathcal{S}_h$*, any standard corrected loss of the form* $\tilde{\ell}(\dot{h}(x), y)$ *must assign the same value to all* $x \in \mathcal{S}_h$*. However, one can construct a measurable transformation* f *(for instance, a suitable translation depending on* h*) such that* $\dot{h}(f(\boldsymbol{x}_1)) \neq \dot{h}(f(\boldsymbol{x}_2))$*, which implies*

$$\ell(\dot{h}(f(\boldsymbol{x}_1)), y) \neq \ell(\dot{h}(f(\boldsymbol{x}_2)), y).$$

This contradicts the requirement that the corrected loss depends only on $\dot{h}(x)$*. Therefore, in this case a standard corrected loss* $\tilde{\ell} \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ *cannot mitigate effect of this attribute corruption effectively. The correction necessarily depends on* h*, which justifies the need for generalized losses in the* GCL *framework.*

The factorization requirement in Eq. (5.13) is the key property that defines GCL, as it implies

$$\mathbf{R}_{\check{\kappa}\phi}(\tilde{\ell} \circ h^*) = \mathbf{R}_{\check{\kappa}\phi}\left[\hat{\kappa}^\dagger(\ell \circ h^*)\right],$$

and by Proposition 5.13,

$$\mathbf{R}_{\check{\kappa}\phi}(\tilde{\ell} \circ h^*) = \mathbf{R}_{\phi}(\ell \circ h^*).$$

Since the factorization holds for every $h \in \mathcal{H}$, it follows that h^* is also a minimizer for the factorized corrected problem $(\tilde{\ell}, \mathcal{H}, \check{\kappa}\phi)$. Investigating which conditions on the learning problem ensure the factorization property is beyond the scope of this analysis. The conclusion of our loss correction formulas is not to provide a new practical tool, but to showcase how even under optimistic assumptions, loss correction techniques are too complex and involved when dealing with attribute corruption kernels.

More in general, all the novel correction formulas found in this chapter are more complex

than the ones defined in previous work (van Rooyen and Williamson, 2018; Patrini et al., 2017), which only considers a label noise scenario similar to our CL for simple label corruption. Our version, Theorem 5.16, further extends the setting by including the dependent label corruption. The second set of results, included in Theorem 5.18, are instead fulfilling the weaker conditions imposed by the GCL framework, but only for a subset of learning problems for which an optimal model exists. Our loss correction formulas offer us an important insight: under a Bayes risk point of view, *there is another fundamental difference between label and attribute corruption*. They induce distinct corrupted learning settings, and traditional loss correction does not ensure unbiased learning in the sense of CL in the presence of the latter.

5.3 A Taxonomy of Corruption Consequences and Mitigations

In this chapter, we have leveraged our complete taxonomy of one-step Markovian corruptions, based on Markov kernels, to understand how distributional changes of this sort influence a learning problem. As a result, we identified an induced *taxonomy of corruption consequences and mitigations*, reported in Tab. 5.1.

The consequences are here studied in the form of a qualitative comparison of the effects of the corruption kernel on the data distribution ϕ and the attainable prediction set $\ell \circ \mathcal{H}$. Formally, these are the data processing equalities and associated equivalences that we proved in § 5.1. We underlined that, when the degree of corruption complexity we consider is higher (identified by the partial hierarchy in Fig. 4.2), also the effects on the attainable predictions $\ell \circ \mathcal{H}$ have higher complexity. These results complement, rather than belong to, the classical comparison-of-experiments tradition a là Blackwell (1951) and Torgersen (1991), and expand previous work from Williamson and Cranko (2024). However, in this framework, talking about "easier" or "harder" corruptions is not really adequate, as we do not show that the structural complexity of corruptions (or equivalently, the corrupted minimization sets) implies a harder optimization problem in terms of higher error rates. We can instead conclude from our theorems that there is a fundamental difference between the two corruption types: attribute noise alters the minimization set itself, making the induced change to the model class inseparable from the change to the loss. Nevertheless, we are not claiming that one cannot define a comparison of problems framework when using Markov kernels. We point for instance to (Oyen et al., 2022) for an empirical study, and to our own framework for quantitative comparison via inequalities in Part III of this thesis.

The corruption mitigation techniques studied in § 5.2 fall within the framework of loss correction methods. This is not the only possible choice of mitigation techniques one can develop, but it is one that is rather inexpensive (or at least, it is in the label noise case) as it does not involve processing the dataset at hand or change the whole architecture; it instead only targets a single function, the loss function, and re-weights it appropriately depending on the corruption. Following existing literature, we name this type of mitigations (generalized) corruption-corrected learning and identified instances in the literature that

belong to this class. When minimized, the corrected losses will, by construction, give back the optimal hypothesis h^* for the clean problem, and achieve accurate learning in the sense of matching loss scores and in the distributional sense. However, we do not intend to propose the formulas proved in Theorems 5.16 and 5.18 as new tools for defining robust losses. These are complex and, in the generalized case, require possibly restrictive assumptions. Instead, Theorem 5.18 is suggesting a *negative result*: even allowing a factorization $\tilde{\ell} \circ h^*$ to exist by weakening the loss requirements, classical loss correction formulas do not seem to be enough for learning the correct model in a corruption setting that involves an attribute corruption. One example of what kind of problems can arise is given in Example 5.19. Instead, one should also account for the set of posterior probabilities $\dot{h}^*\#\tau$ and average on it, instead of only "reweighting" the loss as done under the CL paradigm. This conclusion is, furthermore, perfectly in line with the fundamental differences between label and attribute corruption already observed in § 5.1.

Table 5.1: Corruption types, Bayes risk equalities, and loss correction formulas. Each corruption kernel κ is factorized as τ (acting on inputs $\mathcal{X}$) and λ (acting on labels $\mathcal{Y}$), modifying both the data distribution ϕ and the attainable prediction set $\ell \circ \mathcal{H}$. The corresponding Data Processing Equality for Bayes risk is $\left(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\lambda})\phi\right) \equiv_{\text{BR}} \left(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \phi\right)$ (§ 5.1.3), and the resulting loss correction formulas (§ 5.2) are summarized. All operators are expressed using the kernel operations defined in § 2.3: Def. 2.24 chain composition ($\circ$), Def. 2.25 product composition ($\times$), Def. 2.26 superposition ($\otimes$), and Def. 2.28 partial chain composition ($\circ_{(\cdot)}$).

Corruption type $\kappa \coloneqq \tau \otimes \lambda$	Corruptions action on ϕ, $\tilde{\phi} \coloneqq (\check{\tau} \otimes \check{\lambda})\phi$	Corruption action $\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H})$ on the minimization set $\ell \circ \mathcal{H} \coloneqq \{(x, y) \mapsto \ell(h_x, y) \mid h \in \mathcal{H}\}$	Loss correction formula $\tilde{\ell}(h, \tilde{x}, \tilde{y})$, $\forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}$
$(\delta \colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes$ $(\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$	$\pi_Y \circ \left[(E \circ \check{\delta}) \otimes \check{\lambda}\right]$ $\pi_X \circ \left[\check{\delta} \otimes (E^\dagger \circ \check{\lambda})\right]$	$\left\{(x, y) \mapsto \hat{\lambda}\ell(\dot{h}(x), y), h \in \mathcal{H}\right\}$	$\hat{\lambda}\ell\,(\dot{h}(x), \tilde{y})$
$(\tau \colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes$ $(\lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$	$\pi_Y \circ \left[(E \circ \check{\tau}) \otimes \check{\lambda}\right]$ $\pi_X \circ \left[\check{\tau} \otimes (E^\dagger \circ \check{\lambda})\right]$	$\left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_y \circ h)](x), h \in \mathcal{H}\right\}$	$\mathbb{E}_{u \sim (\dot{h}\#\hat{\tau})(\tilde{x})}[\hat{\lambda}\ell\,(u, \tilde{y})]$
$(\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes$ $(\hat{\lambda} \colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$	$\pi_Y \circ \left[(E \circ_{\mathcal{X}} \check{\tau}) \otimes \check{\lambda}\right]$	$\left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_y \circ h)](x, y), h \in \mathcal{H}\right\}$	$\mathbb{E}_{u \sim (\dot{h}\#\hat{\tau})(\tilde{x}, \tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{y})]$
$(\tau \colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes$ $(\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$	$\pi_X \circ \left[\check{\tau} \otimes (E^\dagger \circ_{\mathcal{Y}} \check{\lambda})\right]$	$\left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](x), h \in \mathcal{H}\right\}$	$\mathbb{E}_{u \sim (\dot{h}\#\hat{\tau})(\tilde{x})}[\hat{\lambda}\ell\,(u, \tilde{x}, \tilde{y})]$
$(\tau \colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes$ $(\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$	$\pi_Y \circ \left[\check{\tau} \otimes (E \circ_{\mathcal{X}} \check{\lambda})\right]$	$\left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](y), h \in \mathcal{H}\right\}$	$\mathbb{E}_{u \sim (\dot{h}\#\hat{\tau})(\tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{x}, \tilde{y})]$
$(\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes$ $(\lambda \colon \mathcal{X} \rightsquigarrow \mathcal{Y})$	$\pi_X \circ \left[(E^\dagger \circ_{\mathcal{Y}} \check{\tau}) \otimes \check{\lambda}\right]$	$\left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_x \circ h)](x, y), h \in \mathcal{H}\right\}$	$\mathbb{E}_{u \sim (\dot{h}\#\hat{\tau})(\tilde{x}, \tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{x})]$
$(\tau \colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes$ $(\lambda \colon \mathcal{X} \rightsquigarrow \mathcal{Y})$	$\pi_Y \circ \left[\check{\tau} \otimes (E \circ \check{\lambda})\right]$ $\pi_X \circ \left[(E^\dagger \circ \check{\tau}) \otimes \check{\lambda}\right]$	$\left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_x \circ h)](y), h \in \mathcal{H}\right\}$	$\mathbb{E}_{u \sim (\dot{h}\#\hat{\tau})(\tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{x})]$
$(\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes$ $(\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$	$(\check{\tau} \otimes \check{\lambda})\phi$	$\left\{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](x, y), h \in \mathcal{H}\right\}$	$\mathbb{E}_{u \sim (\dot{h}\#\hat{\tau})(\tilde{x}, \tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{x}, \tilde{y})]$

Part III

Comparison of Learning Problems via Inequalities

Chapter 6

Entropy and Information for Unconstrained Learning Problems

To give a proper introduction and motivation to the third part of this thesis, we need to focus on some information-theoretic concepts, analyze how they have been treated in the more classical literature, and relate them to machine learning problems as we have abstracted them in Part I. This chapter therefore covers the well-known connection between *entropy* (Shannon, 1948) and the *uncertainty* of statistical decision problems (Blackwell, 1951; DeGroot, 1962). Such a relationship has many interesting implications, the most central arguably being that a form of data processing inequality (DPI) holds for decision problems; this result is the basis for the comparison of experiments theory (Torgersen, 1991). This is relevant to us as machine learning theorists because a statistical decision problem is what we know as an unconstrained learning problem. Recalling Def. 3.2 and switching to the machine learning setting, these problems consist of a data space $\mathcal{X} \times \mathcal{Y}$, a learning context $(\ell, \mathcal{M}(\mathcal{X}, \mathcal{Y}), \phi)$ and a learning criterion $\mathbf{BR}_{\ell \circ \mathcal{M}(\mathcal{X},\mathcal{Y})} = \mathbf{BR}_{\ell}$. Hence, we talk about unconstrained supervised learning problems, which are the only type of learning settings we will discuss in the following sections; their less-explored constrained counterpart will be studied in § 7.

6.1 From Shannon to DeGroot

The term entropy was first introduced by Clausius (1850) with the aim of explaining the second law of thermodynamics. The law states that:

> "There does not exist a machine that operates in a cycle (i.e., returns to its initial state), produces useful work and whose only other effect on the outside world is drawing heat from a warm body."

That is, every such machine must also expend some amount of heat to a cold body. Clausius defined entropy as an attribute of the cyclic machine, a *transformative content*, and proposed that it must have the property of returning to its original value when the machine goes

back to its initial state. Boltzmann (1877) later provided a microscopic interpretation of entropy, connecting macroscopic disorder to the number of compatible microstates, and Gibbs (1902) generalized this viewpoint to probability distributions over phase space. At this stage, entropy was formally described as we still do it nowadays, i.e., as a functional of a probability distribution. Within this new statistical physics framework, entropy quantifies uncertainty about the microscopic configuration of a system consistent with macroscopic constraints. One could then interpret entropy as measuring some *lack of information*.

Later, Shannon (1948) introduced an analogous functional in the context of communication systems. Such a system roughly amounts to a transmitter, a communication channel and a receiver; what is transmitted is *a message containing some information*. Key questions in communication systems are how to make the messages arrive "intact", and how much of the message may be lost during transmission. Shannon worked on how to answer these questions using entropy to measure the information content of messages. In this setting, entropy and mutual information were defined operationally, without physical interpretation, as quantities governing coding rates and channel capacity. However, the connection with statistical mechanics was apparent, and was emphasized by Shannon himself.[1] In particular, he justified his modeling choice of using the logarithm with three, rather pragmatic, arguments: first, it had been observed to be "practically more useful" in many engineering applications, as many important quantities "such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities"; second, "it is nearer to our intuitive feeling as to the proper measure", because properties of the logarithm reflect certain semantics of the problem that Shannon wanted to preserve in his formalization; last, "it is mathematically more suitable", which is an elegant way of saying that the logarithm allowed for simple computations.[2]

This historical note already highlights many key concepts that are central to our perspective as well. We can start by noticing that, while entropy originated as a state function with a (vague) informational interpretation, the quantity is then used to relate two objects, such as states of some machines, collection of particles, or transmitted messages. Citing Brillouin (2013)'s informal definition of information given in his Introduction,

> "We consider a problem involving a certain number of possible answers, if we have no special information on the actual situation. When we happen to be in possession of some information on the problem, the number of possible answers is reduced, and complete information may even leave us with only one possible answer. Information is a function of the ratio of the number of possible answers before and after, and we choose a logarithmic law in order to ensure additivity of the information contained in independent situations."

Entropy is measured in a setting that will undergo or has witnessed a *transformation* from one state to another. Accordingly, throughout this work we use the term information to refer to quantities that are measured in the presence of a transformation, and interpreted

[1] For instance, see discussion under (Shannon, 1948, Theorem 2).
[2] See (Shannon, 1948, Introduction) for his full argumentation.

as the difference between two beliefs about the state of a system. From this standpoint, information can be regarded as the quantification of the strength of a relationship between states, measured with respect to a chosen measure, unit, or task. The beliefs on the states are formalized by entropy, which is compatible with entropy's standard interpretations and uses, but in a way that does not necessarily require the logarithmic form. Rather, the choice of the functional defining entropy is understood as a design choice tied to the unit, task, or notion of comparison one wishes to impose. This is still compatible with Shannon (1948)'s pragmatic discussion of the logarithmic definition of entropy and mutual information, since all we will be doing in the next sections is generalizing his concepts to a larger set of problems. Such different problems, i.e. statistical learning problems, will specify different needs and therefore require different design choices; to address this, we first follow the steps of DeGroot (1962) and then generalize further.

Our standpoint avoids conflating the two distinct roles of information and entropy (dynamic for the former, static for the latter) in measuring quantities related to a system. However, we also acknowledge their conceptual and mathematical similarity, which justifies the practice of regarding entropy as itself a measure of information. Indeed, our semantical distinction of entropy and information is not universally adopted.

6.1.1 Mutual Information and Entropy

We now delve into the mathematical definitions of the objects described so far: entropy and information. The simplest notion is entropy in the discrete case, as Shannon (1948) introduced it.

Definition 6.1 (Shannon Entropy)**.** *Let $\mathcal{Z}$ be a countable space and $\phi \in \Delta_{\text{ca}}(\mathcal{Z})$ a discrete probability distribution. We define the* entropy *of ϕ as*

$$H(\phi) \coloneqq \sum_{z \in \mathcal{Z}} -\phi_z \log(\phi_z).$$

A generalization of entropy to continuous variables is trickier to obtain. One widely accepted attempt is differential entropy, which mimics the original definition but does not preserve all the properties of its discrete counterpart H (see Theorem 2.7, and the warning before it, in Polyanskiy and Wu (2025)). In particular, we remark that differential entropy cannot be obtained as the limit of Shannon entropy (Polyanskiy and Wu, 2025, Eq. (2.21)), as one would instead expect when switching from a discrete to an uncountable setting.

Definition 6.2 (Differential Entropy)**.** *Let $(\mathcal{Z}, \Sigma(\mathcal{Z}), \mu)$ be a standard Borel measure space, and ϕ a probability distribution such that $\phi \ll \mu$. Then, we define the* differential entropy *of $\phi \in \Delta_{\text{ca}}(\mathcal{Z})$ w.r.t. μ as*

$$H_\mu(\phi) \coloneqq \mathbb{E}_\phi \left[-\log \left(\frac{d\phi}{d\mu} \right) \right].$$

We follow Grünwald and Dawid (2004, Eq. (21)) in giving the above definition for a

general measure, while noting that Polyanskiy and Wu (2025) gives it using the Lebesgue measure.

Definition 6.3 (Conditional Entropy)**.** *Consider the standard Borel measurable spaces* $(\mathcal{Z}_i, \Sigma(\mathcal{Z}_i))$ *for* $i \in \{1, 2\}$*, with* $\pi_i \in \Delta_{\text{ca}}(\mathcal{Z}_i)$ *being their associated probabilities, and consider their coupling generated by the Markov kernel* $\kappa \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$*. Let* μ *be a reference measure on the space* $(\mathcal{Z}_2, \Sigma(\mathcal{Z}_2))$ *such that* $\dot{\kappa}(z_1) \ll \mu$ *for all* $z_1 \in \mathcal{Z}_1$*. We define the* conditional entropy *of* κ *with respect to* π_1 *as*

$$H_\mu(\kappa, \pi_1) \coloneqq \mathbb{E}_{\pi_1} H_\mu(\dot{\kappa}(z_1)) = \int \left[\int -\log \left(\frac{d\dot{\kappa}(z_1)}{d\mu}(z_2) \right) \kappa(z_1, dz_2) \right] \pi_1(dz_1).$$

When all the involved sets are countable, and μ *is the counting measure, the definition simplifies to*

$$H(\kappa, \pi) \coloneqq \mathbb{E}_\pi H(\dot{\kappa}(z_1)) = \sum_{z_1, z_2} \left[-\log(\kappa_{z_2, z_1}) \, \pi_{z_1} \right],$$

with $[\kappa_{z_2,z_1}]_{z_2,z_1}$ *being the matrix representation of the Markov kernel* κ *and* $\pi = \pi_1$*.*

We choose the unusual notation taking as input both the regular conditional probability κ and the prior π_1 to underline that the choice of the prior is central for fully specifying the conditional entropy; other random-variable-based notations only emphasize the dependence on the conditional distribution, which may be misleading.

Definition 6.4 (Kullback-Leibler Divergence)**.** *Let* $(\mathcal{Z}, \Sigma(\mathcal{Z}))$ *be a standard Borel measurable space. The functional* $D_{KL} \colon \Delta_{\text{ca}}(\mathcal{Z}) \times \Delta_{\text{ca}}(\mathcal{Z}) \to [0, +\infty]$*, called the* Kullback-Leibler *(KL)* divergence*, is defined as*

$$D_{KL}(\phi \,||\, \psi) \coloneqq \int \log \left(\frac{d\phi}{d\psi} \right) d\phi.$$

When the absolute continuity condition $\phi \ll \psi$ *does not hold, we assume the divergence to be* $+\infty$*.*

If we relax the definition and allow ψ to be a positive measure, the connection between divergence and entropy becomes apparent:

$$H_\mu(\phi) = -D_{KL}(\phi \,||\, \mu).$$

Definition 6.5 (Mutual Information)**.** *Consider the standard Borel measurable spaces* $(\mathcal{Z}_i, \Sigma(\mathcal{Z}_i))$ *for* $i \in \{1, 2\}$*, with* $\pi_i \in \Delta_{\text{ca}}(\mathcal{Z}_i)$ *being their associated probabilities, and consider their coupling generated by the Markov kernel* $\kappa \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ *such that* $\pi_1 \times \kappa \in \Delta_{\text{ca}}(\mathcal{Z}_1 \times \mathcal{Z}_2)$ *is the resulting joint distribution. We define the* mutual information *generated by* κ *w.r.t.* π_1 *as*

$$I(\kappa, \pi_1) \coloneqq D_{KL}(\pi_1 \times \kappa \,||\, \pi_1 \times \kappa_{\pi_2}),$$

recalling that κ_{π_2} *is a degenerate kernel constantly equal to* π_2*, and* π_2 *is identified by marginalizing the joint distribution* $\pi_1 \times \kappa$ *on* $\mathcal{Z}_2$*.*

Our kernel notation is quite different from the standard one, which usually relies on random variables and writes mutual information as a function of two of them. In the setting used by the definition, one would normally write mutual information as $I(Z_1, Z_2)$. Our notation emphasizes that the core element of the definition is κ: this measure of information aims to quantify the effect of the transformation induced by the kernel, comparing it to the state before its application. In addition, as we also noted for entropy, this quantity is only fully identified if one of the two marginal distributions involved is fixed. These concepts become even clearer when expressing mutual information through entropy (Polyanskiy and Wu, 2025, Theorem 3.4).

Proposition 6.6 (Mutual Information as Difference of Entropies)**.** *Consider the standard Borel measurable spaces $(\mathcal{Z}_i, \Sigma(\mathcal{Z}_i))$ for $i \in \{1, 2\}$, $\pi \in \Delta_{\text{ca}}(\mathcal{Z}_1)$ an associated marginal probability, and a Markov kernel $\kappa \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ such that $\pi_1 \times \kappa \in \Delta_{\text{ca}}(\mathcal{Z}_1 \times \mathcal{Z}_2)$ is the resulting joint distribution on the product space and π_2 is identified by marginalizing the joint distribution $\pi_1 \times \kappa$ on $\mathcal{Z}_2$. Then, if $\mathcal{Z}_1$ is discrete,*

$$I(\kappa, \pi_1) = H(\pi_2) - H(\kappa, \pi_1),$$

provided $H(\kappa, \pi_1) < \infty$. When $\mathcal{Z}_2$ is uncountable, and assuming that μ dominates both the marginal π_2 and all Markov transition values $\dot{\kappa}(z_1)$, we have

$$I(\kappa, \pi_1) = H_\mu(\pi_2) - H_\mu(\kappa, \pi_1),$$

provided $H_\mu(\kappa, \pi_1) < \infty$.

Remark 6.7 *A consequence of the above proposition is that* the gap between entropy and conditional entropy is invariant under the choice of dominating reference measure *used to define the differential entropies. While the individual entropies $H_\mu(\pi_2)$ and $H_\mu(\kappa, \pi_1)$ depend on μ, their dependence cancels in the difference due to the definition via Radon–Nikodym derivative of $\pi_1 \otimes \kappa$, i.e.,*

$$I(\kappa, \pi_1) = \int \log\left(\frac{d(\pi_1 \otimes \kappa)}{d(\pi_1 \otimes \kappa_{\pi_2})}\right) d(\pi_1 \otimes \kappa).$$

Hence mutual information is an intrinsic quantity associated with the joint distribution, consistent with its representation as a Kullback-Leibler divergence. In particular, changes of dominating measure modify entropy terms by additive expectation shifts that cancel exactly in the conditional decomposition, so that only the joint law matters.

Hence we have shown that we can define mutual information in two different yet equivalent ways: using Kullback-Leibler divergence or as difference of entropies.

6.1.2 Ways of Defining Generalized Information

As we already discussed at the beginning of this chapter, entropy and mutual information were originally defined using the logarithmic function because of their original connection

to communication systems; Shannon (1948) gives the list of his own intuitions for which the logarithm seemed to be the most sensible choice. All of them are rather pragmatic, appealing to practical observations as well as mathematical convenience. However, practical arguments hold only as long as the field of application is not changed; looking at learning problems, we are putting ourselves in a rather different field of application, and Shannon's information may become obsolete, or at least not the only suitable choice. As DeGroot (1962) observes, "in any particular situation, the selection of an appropriate uncertainty [i.e., generalized entropy] function would typically be based on the use to which the experimenter's knowledge about [the parameter of interest] is to be put." In other words: the tasks of interest should decide the measure of information (and entropy) used. For this reason, we explore here other ways of defining entropy and information. We start doing so by abstracting away from the quantities introduced in the previous section, starting from considering f-divergences instead of the Kullback-Leibler divergence (Polyanskiy and Wu, 2025, Definition 7.1).

Definition 6.8 (f-divergence)**.** *Let $f : (0, +\infty) \to \mathbb{R}$ be convex, and s.t. $f(0) \coloneqq \lim_{z\to 0+} f(z)$ and $f(1) = 0$. Let ϕ, ψ be two probability distributions on $(\mathcal{Z}, \Sigma(\mathcal{Z}))$. Assuming $\phi \ll \psi$, we define f-*divergence *as*

$$D_f(\phi \,||\, \psi) \coloneqq \int f\left(\frac{d\phi}{d\psi}\right) d\psi.$$

Notice that, for $f(t) \coloneqq t \log t$, we recover the KL divergence. In analogy with Shannon's setting, we define a more general notion of mutual information.

Definition 6.9 (f-Information)**.** *Consider the standard Borel measurable spaces $(\mathcal{Z}_i, \Sigma(\mathcal{Z}_i))$ for $i \in \{1, 2\}$, with $\pi_i \in \Delta_{\text{ca}}(\mathcal{Z}_i)$ being their associated probabilities, and consider their coupling generated by the Markov kernel $\kappa \in \mathcal{M}(\mathcal{Z}_1, \mathcal{Z}_2)$ such that $\pi_1 \times \kappa \in \Delta_{\text{ca}}(\mathcal{Z}_1 \times \mathcal{Z}_2)$ is the resulting joint distribution. Assuming $\pi_1 \times \kappa \ll \pi_1 \times \kappa_{\pi_2}$, we define the f-*information *generated by κ with respect to π_1 as*

$$I_f(\kappa, \pi_1) \coloneqq D_f(\pi_1 \times \kappa \,||\, \pi_1 \times \kappa_{\pi_2})$$

with $f : (0, +\infty) \to \mathbb{R}$ convex, and s.t. $f(0) \coloneqq \lim_{z\to 0+} f(z)$, $f(1) = 0$.

Now that we have generalized mutual information, what is its associated generalized entropy? Following what we have done before, we *could* define it as

$$H_{\mu,f}(\phi) \coloneqq -D_f(\phi \,||\, \mu),$$

and hope that the difference of this entropy and its conditional counterpart would be equal to the f-information. Unfortunately, f-information does not behave as nicely as classical mutual information. Notice that a function that makes f-information equal to the difference of f-entropies is such that $f(xy) = f(x) + f(y)$; if we want f continuous, the only choice is $f(x) = c \log x, c < 0$. We therefore have to abandon this attempt and look

elsewhere to define generalized entropy. Another line of work, tightly related to statistical decision problems, investigated ways to define such an object: we review some of the main contributions in the following.

According to Csiszár (1969), it was Perez (1957) who first introduced the term "generalized entropy" in the literature, even though for a slightly different object than what we use here. DeGroot (1962) later described it as a notion of uncertainty, represented through a function of a probability distribution; he then proceeds to characterize these functions, and gives an example derived from statistical decision problems—the *risk* of the optimal decision (Eq. 2.14). Grünwald and Dawid (2004) treated generalized entropy within a more modern framework, also considering general spaces instead of finite-dimensional, and translated the maximum generalized entropy principle into robust Bayes decision making. In all these works, the used notion of information is different from the one arising from f-divergences: they (implicitly or explicitly) consider information as derived from entropy. Therefore, let us now go back to the realm of learning problems, and define their associated entropy and information following DeGroot (1962)'s work.

Definition 6.10 (Statistical Information and Generalized Entropy)**.** *Let $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss, $(\mathcal{X} \times \mathcal{Y}, \Sigma(\mathcal{X}) \times \Sigma(\mathcal{Y}), \phi)$ a standard Borel space with $\phi \in \Delta_{\mathrm{ca}}(\mathcal{X} \times \mathcal{Y})$. Let $\pi_Y \in \Delta_{\mathrm{ca}}(\mathcal{Y}), \pi_X \in \Delta_{\mathrm{ca}}(\mathcal{X})$ be their associated marginals, and $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ the experiment matching the coupling, i.e., $\phi = \pi_Y \times E$. Then, we define the* ℓ-statistical information *as*

$$I_\ell(E, \pi_Y) \coloneqq H_\ell(\pi_Y) - H_\ell(E, \pi_Y),$$

where the generalized entropy $H_\ell(\pi_Y)$ and its conditional version $H_\ell(E, \pi_Y)$ are respectively defined as

$$H_\ell(\pi_Y) \coloneqq \inf_{\psi \in \Delta_{\mathrm{ca}}(\mathcal{Y})} \mathbb{E}_{Y \sim \pi_Y}[\ell(\psi, Y)] = \mathbf{CBR}_\ell(\pi_Y)$$
$$H_\ell(E, \pi_Y) \coloneqq \mathbb{E}_{Y \sim \pi_Y} \mathbf{CBR}_\ell(\dot{E}(Y)) = \mathbf{BR}_\ell(\phi)$$

where the last equality in the definition of the generalized conditional entropy $H_\ell(E, \pi_Y)$ holds because of Remark 3.8.

A natural question arises: are the two definitions of information and entropy related? A first step towards answering this question can be made by considering Shannon entropy.

Proposition 6.11. *Let $\ell_{\log}(p) = -\log(p)\colon \mathrm{ri}\big(\Delta_{\mathrm{ca}}(\mathcal{Y})\big) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be the logarithmic loss restricted to the relative interior of the simplex (as defined and discussed in Remark 3.15). Let $(\mathcal{X} \times \mathcal{Y}, \Sigma(\mathcal{X}) \times \Sigma(\mathcal{Y}), \phi)$ be a standard Borel space with $\phi \in \Delta_{\mathrm{ca}}(\mathcal{X} \times \mathcal{Y})$, $\pi_Y \in \Delta_{\mathrm{ca}}(\mathcal{Y}))$ the associated marginal, and $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ the experiment s.t. $\phi = \pi_Y \times E$. Then, it holds that the statistical information is equal to the f-information for $f(t) = t \log t$, i.e.,*

$$I_{\ell_{\log}}(E, \pi_Y) = I_{t \mapsto t \log t}(E^\dagger, \pi_X)$$

Proof. We prove the statement by noticing that, since $\ell_{\log}$ is a proper loss, i.e., a loss s.t. the true generating probability $E^\dagger$, defined such that $\pi_Y \times E = \pi_X \times E^\dagger$ (see Def. 5.9), attains the value $\mathbf{BR}_{\ell_{\log}}(\phi)$, then

$$H_{\ell_{\log}}(E, \pi_Y) = \mathbf{BR}_{\ell_{\log}}(\pi_X \times E^\dagger) = H(E^\dagger, \pi_X).$$

Similarly, for the prior entropy we notice that

$$\mathbf{CBR}_{\ell_{\log}}(\pi_Y) \coloneqq \inf_{\psi \in \Delta_{\text{ca}}(\mathcal{Y})} \sum_{y \in \mathcal{Y}} -\log([\psi_Y]_y)\,[\pi_Y]_y \;= \mathbb{E}_{Y \sim \pi_Y}\left[-\log(\pi_Y)\right] = H(\pi_Y).$$

□

Unlike their logarithmic counterparts, it is not immediately obvious how to determine whether an equivalence result holds for statistical and f-information, and the answer depends on whether we are in a binary or multiclass learning setting. A first relation has been proved in Theorems 1 and 2 of Osterreicher and Vajda (1993) for binary problems, showing that every f-divergence is a ℓ-statistical information and vice versa. A later refinement, but still in the binary case, appeared in Reid and Williamson (2011) by adding the relationship with Bregman divergence to the picture (Theorems 9 and 10).

Proposition 6.12 (Bridge Between ℓ-Statistical Information and f-Information)**.** *Let $\mathcal{Y} = \{0, 1\}$ and $(\mathcal{X}, \Sigma(\mathcal{X}))$ a standard Borel space. Consider an experiment $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ and a discrete $\pi_Y \in \Delta_{\text{ca}}(\mathcal{Y})$. For a suitable pair (f, ℓ),*[3]

$$I_f(E_0 \,||\, E_1) = I_\ell(E, \pi_Y) = \mathbf{CBR}_\ell(\pi_Y) - \mathbf{BR}_\ell(\pi_Y \times E).$$

The f-information computed above has a nice interpretation in terms of experiments: since $E_0 \coloneqq E(y = 0, dx)$ and $E_1 \coloneqq E(y = 1, dx)$ are the two possible realizations of a binary statistical experiment, I_ℓ is measuring information by quantifying the "dissimilarity" (in the sense of f-divergence) between the two.

Modern results extending this proposition to multiclass cases are included in García García and Williamson (2012) and Williamson and Cranko (2024). Theorem 23 from Williamson and Cranko (2024) finally generalized the Osterreicher and Vajda (1993) result to multiclass problems by means of a new notion of information, the D-information. The authors show that their notion of information[4] is a multi-distribution generalization of the f-divergence (Williamson and Cranko, 2024, Section 3.2), and then prove it to be equal to Bayes risk, or

[3] Reid and Williamson (2011) provide examples of such pairs (Table 2), as well as a mapping from f to ℓ and vice versa.

[4] Williamson and Cranko (2024)'s notion of D-information cannot be trivially included in our presentation, as it does not respect the classical definition of information as a comparison of entropies before and after a transformation. It in fact subsumes f-divergences, and therefore is a more general object than entropy and information; it can in fact be equal to both, depending on the input. Looking at the information-risk bridge result, the authors seem to align with the school of thought regarding (generalized) entropy as a notion of information itself (indeed, in some literature, our Def. 6.1 is called Shannon information). This view is a valid stance, although different from ours.

generalized conditional entropy as we define it here.

Proposition 6.13 (D-Information-Risk Bridge)**.** *Suppose $(\mathcal{X}\times\mathcal{Y}, \Sigma(\mathcal{X})\times\Sigma(\mathcal{Y}), \phi)$ is a standard Borel and $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\geq 0}$ be a bounded loss. Let $\pi\in\Delta_{\mathrm{ca}}(\mathcal{Y})$ be the associated marginal, and $E\in\mathcal{M}(\mathcal{Y},\mathcal{X})$ and experiment. Consider the set $\pi\cdot\mathrm{spr}\ell := \{(\pi_y f_y)_{y\in\mathcal{Y}} \mid f\in\mathrm{spr}\ell\}$. Then,*

$$\mathbf{BR}_\ell(\pi\times E) = -I_{-\pi\cdot\mathrm{spr}\ell}(E),$$

where I is meant as the D-information, defined as

$$I_D(E) := \int \sigma_D\left(\frac{dE}{d\mu}\right)d\mu = \int -\rho_{-D}\left(\frac{dE}{d\mu}\right)d\mu,$$

with μ being a reference measure on $(\mathcal{X},\Sigma(\mathcal{X}))$, $\rho_D = -\sigma_{-D}$ being the concave support function (Def. 3.21), and

$$\frac{dE}{d\mu} := \left(x\mapsto \frac{d\dot{E}(y)}{d\mu}(x)\right)_{y\in\mathcal{Y}}$$

a vector of real functions on $\mathcal{X}$.

6.2 Data Processing Inequalities for Entropy

One key property of mutual information is that it is *always a non-negative quantity*. This property is proved by observing that $H(\pi_2)\geq H(\kappa,\pi_1)$ (Polyanskiy and Wu, 2025, Corollary 3.5), which formalized the idea that the uncertainty about a certain variable Z_2 *decreases* once we have additional observations Z_1 giving us "information" on Z_2.

Generalized entropy also respects such a property, as proved in Theorem 2.1 (DeGroot, 1962) by additionally observing that conditional Bayes risk is concave (e.g., Wijsman, 1970, Theorem 2). We give here a statement and a proof to introduce the reader to the technical details of the result.

Proposition 6.14. *Consider a joint probability space $(\mathcal{X}\times\mathcal{Y}, \Sigma(\mathcal{X})\times\Sigma(\mathcal{Y}), \phi)$, and its decompositions $\phi = E\times\pi_Y$ and $\phi = E^\dagger\times\pi_X$, where $E\in\mathcal{M}(\mathcal{Y},\mathcal{X})$ is an experiment and $\pi_Y\in\Delta_{\mathrm{ca}}(\mathcal{Y}), \pi_X\in\Delta_{\mathrm{ca}}(\mathcal{X})$ some compatible prior probabilities, i.e., $\pi_Y = \pi_X\circ E^\dagger$. Let $\ell\colon\Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\geq 0}$ be a bounded loss function. Then, it holds that*

$$\mathbf{CBR}_\ell(\pi_Y)\geq\mathbf{BR}_\ell(\phi).$$

Proof. Let us remind the reader that conditional risk is a concave functional, and that

Remark 3.8 holds in the unconstrained setting. Therefore, we can write that

$$\begin{aligned}\mathbf{BR}_\ell(\phi) &= \mathbb{E}_{X\sim\pi_X}\mathbf{CBR}_\ell[\dot{E}^\dagger(X)]\\ &\overset{(\star)}{\leq} \mathbf{CBR}_\ell[\mathbb{E}_{X\sim\pi_X}\dot{E}^\dagger(X)]\\ &= \mathbf{CBR}_\ell[\pi_X \circ E^\dagger]\\ &= \inf_{\psi\in\Delta_{\text{ca}}(\mathcal{Y})} \mathbb{E}_{Y\sim\pi_X\circ E^\dagger}\ell(\psi, Y)\\ &= \inf_{\psi\in\Delta_{\text{ca}}(\mathcal{Y})} \mathbb{E}_{Y\sim\pi_Y}\ell(\psi, Y)\end{aligned}$$

where $(\star)$ holds by Jensen's inequality. This proves the thesis, as the last line of the equation is the prior Bayes risk. □

In other words, we proved that additional transformation of our prior induced by an experiment reduces entropy. This can also be interpreted in statistical terms as, observing an additional variable X related to Y decreases uncertainty. Recalling the interpretation of information as the comparison of entropy (or, uncertainty) before and after some transformation, we can derive a simple corollary from the above proposition, drawing a connection with a core class of results of information theory: *data processing inequalities*.

Corollary 6.15 (Data Processing Inequality for a Non-Informative Experiment)**.** *Consider a joint probability space $(\mathcal{X}\times\mathcal{Y}, \Sigma(\mathcal{X})\times\Sigma(\mathcal{Y}), \phi)$, and a disintegration of the measure as $\phi = E\times\pi_Y$, where $E \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ is an experiment and $\pi_X \in \Delta_{\text{ca}}(\mathcal{X})$, $\pi_Y \in \Delta_{\text{ca}}(\mathcal{Y})$ compatible prior probability, i.e., $\pi_Y = \pi_X \circ E^\dagger$. Let $\ell\colon \Delta_{\text{ca}}(\mathcal{Y})\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $E_{\pi_X} \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ a non-informative experiment, i.e., $\dot{E}(y) = \pi_X \forall y$ (Def. 2.23). Then, it holds that*

$$\mathbf{BR}_\ell(E_{\pi_X}\times\pi_Y) \geq \mathbf{BR}_\ell(\phi).$$

Proof. Observe that we can write

$$\begin{aligned}\mathbf{CBR}_\ell(\pi_Y) &:= \inf_{\psi\in\Delta_{\text{ca}}(\mathcal{Y})} \mathbb{E}_{Y\sim\pi_Y}\ell(\psi, Y)\\ &\overset{(\star)}{=} \inf_{\psi\in\Delta_{\text{ca}}(\mathcal{Y})} \mathbb{E}_{Y\sim\pi_Y}\int_X \ell(\psi, Y)d\pi_X\\ &\overset{(\star)}{R3.8} \inf_{\psi\in\Delta_{\text{ca}}(\mathcal{Y})} \mathbb{E}_{Y\sim\pi_Y}\mathbb{E}_{X\sim\pi_X}\ell(\psi, Y)\\ &=: \mathbf{BR}_\ell(E_{\pi_X}\times\pi_Y),\end{aligned}$$

where $(\star)$ holds because $\ell(\psi, Y)$ does not depend on X. □

Hence, the uncertainty of a joint probability with a non trivial coupling of the marginals is always lower than the one of the decoupled marginals. However, the more interesting form of data processing inequality includes far more complex cases than decoupled marginals. We will explore here the existing results.

6.2.1 Blackwell's Comparison of Experiments

In his paper from 1951, Blackwell introduces a framework for *comparing experiments*.[5] Key to this framework are two concepts: sufficiency and informativity of one experiment with respect to another. The former appears in the literature under different names: famous instances are Le Cam (1964)'s and Torgersen (1991)'s *deficiency* and Shiryaev and Spokoiny (2000)'s *randomization*; different literature strains also refer to it as *dominated experiments* (Strasser, 1985). The latter has been less investigated as the field progressed, and the word "informativity" became synonym of sufficiency after an equivalence result was proved as a joint effort of Blackwell, Sherman and Stein.[6] We report here in brief the key concepts that lead to proving a general data processing inequality for Bayes risk, as well as the equivalence result.

Definition 6.16 (Sufficiency)**.** *Consider two arbitrary measurable spaces $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$. An experiment $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ is* sufficient *for another experiment $\tilde{E} \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ if there is a Markov kernel $\kappa \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ such that $\tilde{E} = E \circ \kappa$.*

Remark 6.17 *Degenerate Markov kernels $\kappa_\mu \in \mathcal{M}(\mathcal{X}, \mathcal{X})$, i.e., kernels that ignore their input and output a fixed probability measure μ on $\mathcal{X}$, produce experiments $\tilde{E} = E \circ \kappa_\mu$ that are* totally non-informative *about their parameter $\mathcal{Y}$. That is because such experiments are independent of the original signal carried by E and therefore represent the minimal (worst) elements in the partial order induced by sufficiency on $\mathcal{M}(\mathcal{Y}, \mathcal{X})$. In fact, as a consequence Corollary 6.15, they will also have the maximal Bayes risk.*

Now consider a probability measure π on $\mathcal{X}$: it can also be viewed as a degenerate experiment E_π that is constant in $\mathcal{Y}$. Following the same reasoning as above, any degenerate experiment is sufficient for another degenerate experiment. Applying Corollary 6.15 to all ordered pairs (E_{π_1}, E_{π_2}) combined with some $\pi_Y \in \Delta_{\mathrm{ca}}(\mathcal{Y})$, we get that all degenerate experiments have the same maximal Bayes risk, and therefore minimal information content once a prior is fixed.

Definition 6.18 (Informativity)**.** *Consider two arbitrary measurable spaces $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$. Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $E \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ is an experiment. Let $\mathcal{A}_{\ell,\mathcal{D}}(E)$ the set of attainable (average) losses, i.e.,*

$$\mathcal{A}_{\ell,\mathcal{D}}(E) \coloneqq \left\{ \left(\mathbb{E}_{X \sim E_1} \ell(\dot{h}(X), 1), \dots, \mathbb{E}_{X \sim E_{|\mathcal{Y}|}} \ell(\dot{h}(X), |\mathcal{Y}|) \right) \mid h \in \mathcal{H}_{\mathcal{D}} \right\} \subseteq \mathbb{R}_{\geq 0}^{|\mathcal{Y}|},$$

where $\mathcal{H}_{\mathcal{D}}$ contains all kernels such that their associated transitions $\dot{h}$ are measurable functions mapping $\mathcal{X}$ to the decision space $\mathcal{D} \subseteq \Delta_{\mathrm{ca}}(\mathcal{Y})$. An experiment E is more informative than

[5] He specifies that he did not come up with the concept entirely by himself, and refers to private communication with Bohnenblust, Shapley and Sherman. These results are only later made public, in (Bohnenblust et al., 1949). We decide to still attribute the definition to Blackwell, as he has the merit of properly formalizing and analyzing the concept.

[6] A first proof was given in (Blackwell, 1953), but only in finite spaces; Boll (1955) extended the equivalence to Polish state spaces. See (Khan et al., 2024) for a literature review and complete presentation of the equivalent statements to Blackwell-Sherman-Stein theorem.

another *experiment* $\tilde{E} \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ *if*

$$\overline{\text{co}}\, \mathcal{A}_{\ell,\mathcal{D}}(E) \supseteq \overline{\text{co}}\, \mathcal{A}_{\ell,\mathcal{D}}(\tilde{E}) \quad \forall \ell, \forall \mathcal{D}.$$

We use the definition of informativity initially suggested by Blackwell (1951), with the convex closure and possible equality of the sets. As shown by Khan et al. (2025) in Example 1, this is the necessary definition for properly arriving to a data processing result.[7]

Indeed, an inequality was proved for DeGroot (1962)'s notion of conditional entropy, called "uncertainty functions". Key to this theorem is the equivalence between sufficiency and informativeness, first proved for finite $\mathcal{Y}$ and bounded $\mathcal{X}$ in (Blackwell, 1953). A modern proof and formulation of this result is given by Khan et al. (2025) in their Theorem 1.

Theorem 6.19 (Blackwell-Sherman-Stein Theorem)**.** *For any two experiments $E, \tilde{E} \in \mathcal{M}(\mathcal{Y}, \mathcal{X})$ on arbitrary spaces $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$, sufficiency implies informativity. If there exists a measure μ on $(\mathcal{X}, \Sigma(\mathcal{X}))$ standard Borel such that $\dot{E}(y) \ll \mu$ for all $y \in \mathcal{Y}$, sufficiency and informativity are equivalent.*

Remark 6.20 *Notice that when $(\mathcal{X}, \Sigma(\mathcal{X}))$ is standard Borel and $\mathcal{Y}$ is finite, we can pick $\mu := \frac{1}{|\mathcal{Y}|} \sum_y \dot{E}(y)$ as the measure dominating all conditioned experiments $\dot{E}(y)$. Hence, the Blackwell-Sherman-Stein Theorem holds as an equivalence.*

6.2.2 DeGroot's Uncertainty Functions

Informativeness as-it-is does not trivially induce a relationship between uncertainty functions. The only result that can be immediately proved is included in (Blackwell, 1951, Theorem 2), which is very close to a data processing inequality result, but it only applies to finite sets $\mathcal{X}$. Such a result later inspired De Groot's work on measures of information and entropy functions, which provides the missing result needed to understand the relationship between informativity, sufficiency and Bayes risk inequality: he established in (DeGroot, 1962, Theorem 4.3) that sufficiency implies the data processing inequality, without constraints on the reference measure and using general *uncertainty functions*.

Theorem 6.21 (Data Processing Inequality for Uncertainty Functions)**.** *Consider an experiment E sufficient for another experiment $\tilde{E}$, i.e., $\tilde{E} = E \circ \kappa$ for some $\kappa \in \mathcal{M}(\mathcal{X}, \mathcal{X})$. Let $U : \Delta_{\text{ca}}(\mathcal{Y}) \to \mathbb{R}_{\geq 0}$ be an uncertainty function, assumed to be concave and $\Sigma(\Delta_{\text{ca}}(\mathcal{Y}))$-*

[7] Related work from the same authors (Khan et al., 2024) studies other notions of informativity and sufficiency, that are more sophisticated and capture additional interesting properties. We do not review them here, as they lead to a data processing inequality for randomized strategies, i.e., defining risk by averaging on a distribution on $\mathcal{D}$. This has also been the focus of van Rooyen (2015). What we study here is instead a data processing inequality for Bayes risk, that only deals with "pure strategies", as the authors of (Khan et al., 2024, 2025) name them.

measurable, and V the conditional uncertainty induced by it defined as

$$V_{\pi_Y}(E) \coloneqq \int_{\mathcal{X}} U(\dot{E}^{\dagger}(X)) \left[\sum_y \pi_y \frac{d\dot{E}(y)}{d\mu}(x) \right] d\mu,$$

where μ is a reference measure such that $\dot{E}(y) \ll \mu \; \forall \, y \in \mathcal{Y}$. Hence, it holds that

$$V_{\pi_Y}(E) \leq V_{\pi_Y}(\tilde{E}) \quad \forall \pi_Y, \forall U.$$

As remarked by DeGroot himself, one well-known uncertainty function is prior Bayes risk. Hence we can state our data processing inequality of interest as a corollary.

Corollary 6.22 (Data Processing Inequality for Bayes Risk)**.** *Consider an experiment E sufficient for another experiment $\tilde{E}$, i.e., $\tilde{E} = E \circ \kappa$ for some $\kappa \in \mathcal{M}(\mathcal{X}, \mathcal{X})$. Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function. Then, it holds that*

$$\mathbf{BR}_\ell(\pi_Y \times E) \leq \mathbf{BR}_\ell(\pi_Y \times \tilde{E}) \quad \forall \pi_Y, \forall \ell.$$

Proof. We define the uncertainty function $U \coloneqq \mathbf{CBR}_\ell$ for some bounded loss function ℓ. Applying Remark 3.8 proves the statement for Bayes risk. □

As a subsequent immediate result, we also obtain the inequality for statistical information.

Corollary 6.23 (Data Processing Inequality for Statistical Information)**.** *Consider an experiment E sufficient for another experiment $\tilde{E}$, i.e., $\tilde{E} = E \circ \kappa$ for some $\kappa \in \mathcal{M}(\mathcal{X}, \mathcal{X})$. Let $\pi_Y \in \Delta_{\mathrm{ca}}(\mathcal{Y})$ have full support and ℓ be a loss function. Then, it holds that*

$$I_\ell(E, \pi_Y) \geq I_\ell(\tilde{E}, \pi_Y) \quad \forall \ell.$$

Proof. The inequality trivially follows from Corollary 6.22 after noticing that $\mathbf{CBR}(\pi_Y)$ risk is constant in the experiment, and therefore is not affected by modifying E into $\tilde{E} = \kappa \circ E$. □

6.2.3 Uses and Structural Limitations of the Data Processing Inequality for Learning Problems

The data processing result for Bayes risk and statistical information have interesting consequences for learning problems. The interpretation of the inequality is that any stochastic (and therefore, also deterministic) transformation of the data distribution cannot decrease the best-case error, meaning that no additional useful (statistical) information for the task can be created beyond what is already present in the original data distribution. This principle has been widely used across several areas of machine learning, including representation learning through the information bottleneck framework (Tishby et al., 1999), generalization theory via information-theoretic bounds (Russo and Zou, 2016; Xu and

Raginsky, 2017), and the study of invariances and sufficient representations in supervised and unsupervised learning (Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017; Saxe et al., 2019). However, it mostly uses mutual information, and therefore assumes the logarithmic loss when measuring entropy. As we have seen here, a learning problem with a loss that is not the logarithmic one will instead have the goal of minimizing a different entropy. What is the exact relationship between different kinds of entropy (and information) based on the chosen loss is unclear. We do know that there exist bounds between the Kullback–Leibler divergence and other f-divergences (Polyanskiy and Wu, 2025, Chapter 7), which can help in understanding such a relationship for specific pairs of problems by using the entropy-divergence relationships from Propositions 6.12 and 6.13.

Blackwell's work and his legacy not only differ from existing machine learning literature because of the different learning problems considered. Most machine learning applications use a specific architecture, optimization algorithm, data pipeline and other technical solutions when training their models, and these factors contribute explicitly or implicitly in constraining the model class $\mathcal{H}$. Such constraints cannot be guaranteed to be included in what is used in Blackwell (1951) definition of informativity (Def. 6.18). In addition, one learning problem's data distribution can be affected not only by attribute noise, but also label or combined noise (see Part II). The tools we have described in this chapter cannot account for these additional important aspects affecting a learning problem. In the next chapter, we give an overview of existing work that already challenged these limitations, and propose a novel way of dealing with them.

Chapter 7

Data Processing for Constrained Learning Problems

In this chapter, we further generalize and analyze the inequality used in Corollary 6.22, with the aim of addressing the limitations arising in machine learning when we want to compare learning problems using the data processing inequality (see § 6.2.3). The transition from unconstrained to constrained problems already introduces substantial complications, discussed in detail in the next section. In particular, the result of DeGroot (1962) does not extend to constrained Bayes risk. Moreover, as shown in Part II, the class of admissible corruptions is much broader than attribute corruptions alone, and different corruptions act differently depending on their signature. The classical formulation of the data processing inequality excludes important cases such as label corruption and, more generally, combined corruption, and it does not account for changes in the joint distribution $\phi = \pi_Y \times E$. Consequently, the DPI must be modified to account for these cases.

The following sections therefore develop a generalized DPI based on the functional

$$\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y}) \mapsto \mathbf{BR}_{\ell \circ \mathcal{H}}(\phi),$$

rather than restricting attention to comparisons between experiments. We also extend the underlying class of probabilities from countably additive to finitely additive measures. This extension is natural in view of the dual pairing between bounded measurable functions and finitely additive measures (see § 2). Framed more broadly as a comparison of learning problems, the generalized DPI admits several variants depending on which components are held fixed and which are allowed to vary. In this way, the inequality can be used to compare probabilities, losses, models, or combinations thereof. Note that the term 'generalized' is well justified: the classical DPI result will be recovered within our framework, which additionally offers an alternative proof of the theorem.

Some of these alternative inequalities have been considered in the literature. For example, Goel and DeGroot (1979, Section 5) studies the data processing inequality result assuming fixed marginal probabilities. Feldman (1972) instead allows comparisons of experiments

for a fixed model class and loss, while still quantifying over all priors. To our knowledge, however, there is no prior work on data processing inequalities aimed at comparing model class performance itself, that is, the composite set $\ell \circ \mathcal{H}$. Motivated by this gap, we characterize a generalized DPI for fixed loss and model class while allowing ϕ to vary. This perspective is not only novel, but also deliberately removes the least accessible component of a learning problem—the data-generating distribution—and centers the analysis on the loss and model class, which are specified by the practitioner and encode their domain knowledge and needs.

Superprediction sets are useful objects for building a tool to answer our question, as we can prove that such a Bayes risk inequality (or, Bayes risk ordering) is in fact equivalent to set ordering in the space of superprediction sets. We then leverage this characterization to prove sufficient conditions for the Bayes risk orderings to hold, and then proceed to build an antipolar version of the ordering which is related to the strong version of the data processing result. Two main differences distinguish our contributions from Blackwell's and DeGroot's work:

1. The comparison of problems flavor we choose shifts the focus from the probability distribution to the decision space (our model class $\mathcal{H}$) and the evaluator (our loss function ℓ). We do so by studying the couple $(\ell, \mathcal{H})$, by combining them in the set $\ell \circ \mathcal{H}$ and taking its superprediction set;

2. We consider the *actual* attainable prediction losses by evaluating all models in $\mathcal{H}$ through ℓ; Blackwell's attainable losses are in fact *average* losses, as we can see in Def. 6.18. For the sake of simplicity, and to allude to the superprediction set, we call the set $\ell \circ \mathcal{H}$ the set of *attainable predictions*.

We now proceed to introduce the generalization of the data processing inequality, first to constrained sets and then to joint distributions.

7.1 Comparison of Experiments in Constrained Case

A straightforward way to generalize the DPI to the constrained setting is to consider the inequality

$$\mathbf{BR}_{\ell \circ \mathcal{H}}[\pi_Y \times E] \leq \mathbf{BR}_{\ell \circ \mathcal{H}}[\pi_Y \times (E \circ \kappa)]\,, \tag{7.1}$$

where $\kappa \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ is simple attribute corruption. A natural question is whether the classical DPI result extends to this setting, that is, *does Eq. (7.1) hold?*

When the model is well-specified for both data distributions, the answer is yes, for all π_Y and ℓ, as both constrained Bayes risks assume the same value of the unconstrained versions. The situation changes when the model class is constrained and not necessarily well-specified: in this case there is no general relationship between conditional Bayes risk

and full Bayes risk, as discussed in Remark 3.8, and Theorem 6.21 cannot be applied. Upon further investigation, we discover that in this setting we are less fortunate than in DeGroot (1962)'s framework: the inequality in Eq. (7.1) can fail to hold, and a simple counterexample can be constructed.

Example 7.1 (Counterexample for Constrained DPI)**.** *Let $\mathcal{X} \coloneqq \{0,1\}$ and $\mathcal{Y} \coloneqq \{0,1\}$. We consider the experiment E and the joint data distribution ϕ in their matrix representation, with columns being the $\mathcal{Y}$-values and rows being the $\mathcal{X}$-values:*

$$E \coloneqq \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \phi \coloneqq \pi_Y \times E = \begin{pmatrix} 1-p & 0 \\ 0 & p \end{pmatrix},$$

where $\pi_Y = [1-p, p]$, $p \in (0,1)$ is some general label prior with full support.

Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function such that $\ell(y, h_y(x)) = 0$ if and only if $h_y(x) = 1$, i.e., the model assigns probability one to the correct label. This implies that for $h_y(x) < 1$ the loss function will be strictly greater than zero. Let $\mathcal{H} \coloneqq \{h \colon \mathcal{X} \to \Delta(\mathcal{Y}) \mid h_y(x) = 1 - \delta_{x,y}\}$ be a model class with only one element in it. Written in matrix form, that element is equal to:

$$h \coloneqq \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Evaluating the Bayes risk gives

$$\min_{h \in \mathcal{H}} \mathbb{E}_\phi[(\ell \circ h)(X,Y)] = \ell(h_0(0),0) * (1-p) + \ell(h_1(1),1) * p > 0\,.$$

If we consider the Markov kernel which puts full corruption on $\mathcal{X}$, we obtain the corrupted distributions,

$$\tilde{E} \coloneqq \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \quad \tilde{\phi} \coloneqq \begin{pmatrix} \frac{(1-p)}{2} & \frac{p}{2} \\ \frac{(1-p)}{2} & \frac{p}{2} \end{pmatrix}.$$

This gives the Bayes risk,

$$\min_{h \in \mathcal{H}} \mathbb{E}_{\tilde{\phi}}[\ell(\dot{h}(X),Y)] = 0.5 * \left[\, \ell(\dot{h}_0(0),0) * (1-p) + \ell(h_1(1),1) * p \,\right]$$

which is half of the Bayes risk of the original task.

The example shows that the data processing inequality stated in Eq. (7.1) does not generally hold. Although the construction uses extreme cases involving singleton model classes, the same reasoning can be extended to settings with larger $\mathcal{H}$. In the following section we will focus only on generalized entropy, and try to find properties that can help prove data processing inequality results. We do not focus on any related notion of information at this stage, and leave it for future work. However, it is interesting to notice that one can still meaningfully define a notion of constrained statistical information, which already appeared in a simpler form in the literature.

7.1.1 Predictive Information

Statistical information (Def. 6.10) is defined as the difference between some notion of entropy and some notion of conditional entropy. While full Bayes risk can be used as a conditional entropy, we cannot use $\mathbf{CBR}_{\ell\circ\mathcal{H}}(\pi_Y)$ as entropy and be sure to enjoy the non-negativity property of their difference. Many different alternative measures of information can be defined in a way that circumvents this issue, notably (García García and Williamson, 2012; Williamson and Cranko, 2024; Finzi et al., 2026). We define here an immediate one, following Xu et al. (2020) in introducing the notion of predictive families of models, and define a possible notion of "constrained statistical information" with a generalized version of their predictive information.

Definition 7.2 (Predictive Family)**.** *A model class $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X},\mathcal{Y})$ is a* predictive family *if, for all probabilities that are in the image of $\mathcal{H}$ through $\mathcal{X}$, the model class contains its associated degenerate kernel h_π. In formulas, we ask that*

$$\exists\, h_\pi \in \mathcal{M}(\mathcal{X},\mathcal{Y}) \mid \dot{h}_\pi(x) \equiv_x \pi \quad \forall\, \pi \in \bigcup_{h\in\mathcal{H}} \dot{h}(\mathcal{X}).$$

We denote the subset of these constant predictors as $\mathcal{H}_c \subseteq \mathcal{H}$.

Definition 7.3 (Predictive Information)**.** *Let ℓ be a loss function, $\mathcal{H}$ a predictive family and $\phi \in \Delta_{\mathrm{ca}}(\mathcal{X}\times\mathcal{Y})$. Then, we call*

$$I_{\ell\circ\mathcal{H}}(\phi) \coloneqq \mathbf{BR}_{\ell\circ\mathcal{H}_c}(\phi) - \mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)\,,$$

the predictive $\ell\circ\mathcal{H}$-information *of ϕ.*

Notice that our notion of information is a direct generalization of Xu et al. (2020)'s constrained entropy and information.

Lemma 7.4 (Recovery of $\mathcal{H}$-Information)**.** *Let $\mathcal{H} \subset \mathcal{M}(\mathcal{X},\mathcal{Y})$ be a predictive family. Let $\phi \in \Delta_{\mathrm{ca}}(\mathcal{X}\times\mathcal{Y})$. If $(p,y) \in \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y} \mapsto \ell_{\log}(p,y) = -\log(p_y)$ is the logarithmic loss, and $H_{\mathcal{H}}$ is the predictive (conditional) $\mathcal{H}$-entropy (as per Xu et al., 2020, Definition 2), we have that*

$$\mathbf{BR}_{\ell_{\log}\circ\mathcal{H}_c}(\phi) = H_{\mathcal{H}}(Y) \quad \text{and} \quad \mathbf{BR}_{\ell_{\log}\circ\mathcal{H}}(\phi) = H_{\mathcal{H}}(Y|X).$$

In addition, we recover their definition of predictive $\mathcal{H}$-information, as

$$I_{\mathcal{H}}(X,Y) = I_{\ell_{\log}\circ\mathcal{H}}(\phi).$$

Proof. Let $\mathcal{H}(\mathcal{X}) \coloneqq \bigcup_{h\in\mathcal{H}} \dot{h}(\mathcal{X}) \subseteq \Delta_{\mathrm{ca}}(\mathcal{Y})$. The statement is proved by noticing that

$$H_{\mathcal{H}}(Y) \coloneqq \inf_{\pi'\in\mathcal{H}(\mathcal{X})} \mathbb{E}_\pi[-\log(\pi'_Y)] \overset{(\star)}{=} \inf_{h\in\mathcal{H}_c} \mathbb{E}_{Y\sim\pi}\mathbb{E}_{X\sim\dot{E}(Y)}[-\log([\dot{h}(X)]_Y)] = \mathbf{BR}_{\ell_{\log}\circ\mathcal{H}_c}(\phi)\,,$$
$$H_{\mathcal{H}}(Y|X) \coloneqq \inf_{h\in\mathcal{H}} \mathbb{E}_\phi[-\log([\dot{h}(X)]_Y)] =: \mathbf{BR}_{\ell_{\log}\circ\mathcal{H}}(\phi),$$

for some experiment $E \in \mathcal{M}(\mathcal{X},\mathcal{Y})$ such that $\pi\times E = \phi$, and for $[\dot{h}(X)]_y$ being the y-th

entry of the vector $\dot{h}(X) \in \Delta_{\text{ca}}(\mathcal{Y})$. The equality in ($\star$) holds for this specific model class, since $\mathbb{E}_{X\sim\dot{E}(Y)}[-\log([\dot{h}(X)]_Y)]$ is in fact independent from X. □

Analogously to Proposition 2 in (Xu et al., 2020) we can show some basic properties of the constrained information.

Proposition 7.5 (Properties of Predictive Information)**.** *Let $\mathcal{H}$ and $\mathcal{H}'$ be predictive families such that $\mathcal{H} \subseteq \mathcal{H}'$. For any distribution $\phi \in \Delta_{\text{ba}}(\mathcal{X}\times\mathcal{Y})$:*

1. ***Monotonicity:*** $\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi) \geq \mathbf{BR}_{\ell\circ\mathcal{H}'}(\phi)$.
2. ***Non-Negativity:*** $I_{\ell\circ\mathcal{H}}(\phi) \geq 0$.
3. ***Respecting Independence:*** *If X is independent of Y in ϕ, then $I_{\ell\circ\mathcal{H}}(\phi) = 0$.*

Proof.

1. Follows by definition of Bayes risk.
2. Since $\mathcal{H}_c \subseteq \mathcal{H}$ by definition, the statement holds because of point 1.
3. Let $\mathcal{H}(\mathcal{X}) \coloneqq \bigcup_{h\in\mathcal{H}} \dot{h}(\mathcal{X}) \subseteq \Delta_{\text{ca}}(\mathcal{Y})$ and $\phi = \pi_X \times \pi_Y$:

$$\begin{aligned}\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi) &= \inf_{h\in\mathcal{H}} \mathbb{E}_{X\sim\pi_X}[\mathbb{E}_{Y\sim\pi_Y}[(\ell\circ h)(X,Y)]] \geq \mathbb{E}_{X\sim\pi_X}[\inf_{h\in\mathcal{H}} \mathbb{E}_{Y\sim\pi_Y}[(\ell\circ h)(X,Y)]] \\ &= \mathbb{E}_{X\sim\pi_X}[\inf_{h\in\mathcal{H}_c} \mathbb{E}_{Y\sim\pi_Y}[(\ell\circ h)(X,Y)]] = \inf_{\psi\in\mathcal{H}(\mathcal{X})} \mathbb{E}_{Y\sim\pi_Y}[\ell(\psi,Y)] = \mathbf{BR}_{\ell\circ\mathcal{H}_c}(\phi),\end{aligned}$$

   which makes their difference zero.

□

Finally, we can immediately transfer the estimation results in (Xu et al., 2020) to our information measure.

Definition 7.6 (Empirical Predictive Information)**.** *Let $\phi \in \Delta_{\text{ca}}(\mathcal{X}\times\mathcal{Y})$. Let $\mathcal{D}$ be a finite, independently and identically distributed sample from ϕ. Let $\mathcal{H}$ be a predictive family and $\ell\colon \Delta_{\text{ca}}(\mathcal{Y})\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function. The* empirical predictive information *w.r.t. $\ell\circ\mathcal{H}$ is given by,*

$$\hat{I}_{\ell\circ\mathcal{H}}(\mathcal{D}) \coloneqq \inf_{h\in\mathcal{H}_c}\left[\frac{1}{|\mathcal{D}|}\sum_{y_i\in\mathcal{D}}(\ell\circ h)(y_i)\right] - \left[\inf_{h\in\mathcal{H}}\frac{1}{|\mathcal{D}|}\sum_{(x_i,y_i)\in\mathcal{D}}(\ell\circ h)(x_i,y_i)\right].$$

Proposition 7.7 (Estimation of Predictive Information)**.** *Let $\phi \in \Delta_{\text{ca}}(\mathcal{X}\times\mathcal{Y})$. Let $\mathcal{D}$ be a finite, independently and identically distributed sample from ϕ. Let $\mathcal{H}$ be a predictive family and*

$\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\geq 0}$ be a loss function, bounded on the set $\mathcal{H}(\mathcal{X})\times\mathcal{Y} := \bigcup_{h\in\mathcal{H}} \dot{h}(\mathcal{X})\times\mathcal{Y}$ by the constant $\mathcal{B}\geq 0$. Then for any $\delta\in(0,0.5)$, with probability at least $1-2\delta$,

$$|I_{\ell\circ\mathcal{H}}(\phi) - \hat{I}_{\ell\circ\mathcal{H}}(\mathcal{D})| \leq 4\mathfrak{R}_{|\mathcal{D}|}(\ell\circ\mathcal{H}) + 2B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}},$$

where $\mathfrak{R}_{|\mathcal{D}|}(\ell\circ\mathcal{H})$ denotes the Rademacher complexity of $\ell\circ\mathcal{H}$ with sample number $|\mathcal{D}|$, defined as

$$\mathrm{Rad}_{|\mathcal{D}|}(\mathcal{F}) = \frac{1}{|\mathcal{D}|}\mathbb{E}_R\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{|\mathcal{D}|}(1-2R_i)f(z_i)\right|\right], \tag{7.2}$$

for a sample $\mathcal{D}=\{z_i\}_i$, a function class $\mathcal{F}=\{z\mapsto f(z)\in\mathbb{R}_{\geq 0}\}$ and $R=(R_i)_i$ with R_i i.i.d. and Bernoulli distributed with parameter $\frac{1}{2}$.

Proof. The argument laid out in Theorem 1 in (Xu et al., 2020) holds for general, bounded ℓ. □

For possible applications of this information estimator, see follow-up works on Xu et al. (2020)'s predictive information (Ethayarajh et al., 2022; Lu et al., 2023; Fel et al., 2024).

7.2 Comparison of Attainable Predictions

We have discussed above why the classical data processing inequality result does not hold in the constrained case. In fact, a similar counterexample to the one given in Example 7.1 can be built for label corruption and posterior kernel $F=E^\dagger$. With the aim of finding a different data processing inequality that encompasses all meaningful transformations of learning problems (attribute and label ones), and that we can hope to study with tools different from Blackwell (1951)'s and DeGroot (1962)'s, we consider a generalized version of the DPI (GDPI).

Definition 7.8 (Generalized Data Processing Inequality for Bayes Risk)**.** *Given a loss function $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\geq 0}$ and a model class $\mathcal{H}\subseteq\mathcal{M}(\mathcal{X},\mathcal{Y})$, we write the generalized data processing inequality for Bayes risk as*

$$\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell\circ\mathcal{H}}(\check{\kappa}\phi)\,, \quad \phi\in\Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})\,, \quad \kappa\in\mathcal{M}(\mathcal{X}\times\mathcal{Y},\mathcal{X}\times\mathcal{Y}).$$

In words, we are asking the joint corruption κ to be bad for a fixed whole learning problem instead of solely for the experiment. Note that this version of the data processing inequality is clearly more general than the one used in the statement of Corollary 6.22, as choosing $\mathcal{H}=\mathcal{M}(\mathcal{X},\mathcal{Y})$ and $\kappa\in\mathcal{M}(\mathcal{X},\mathcal{X})$ we obtain the classical data processing inequality statement.

Given the nature of the more general inequality, following Blackwell's path is not feasible: a notion of sufficiency for joint probabilities is not meaningful, as two probabilities are always

related by at least a degenerate kernel constantly equal to one of the two. We therefore turn our attention to studying the validity of Def. 7.8 for a fixed set of attainable predictions while varying the probability distribution. As we have seen in Proposition 3.23, one can relate the Bayes risk associated to a probability distribution ϕ, a loss ℓ, and a model class $\mathcal{H}$ to the support function of the set $\operatorname{spr}(\ell \circ \mathcal{H})$ evaluated on ϕ. Therefore, we can characterize the Bayes risk inequality by means of geometrical properties. This is similar in spirit but not equivalent to the work from Blackwell (1951) and DeGroot (1962), that we have analyzed in the previous chapter.

To distinguish our contributions from previous work on informativity and sufficiency, we call *orderings for attainable predictions* our flavor of the Bayes risk inequalities and their characterizing geometrical properties (proved in the following). Recall that, in general, an ordering on a set is defined as a binary relation between two elements of the said set, e.g.,

$$\geq : (a, b) \mapsto \{a \geq b, b > a, \text{'not comparable'}\},$$

such that it is reflexive ($a \geq a$), antisymmetric (if $a \geq b$ and $b \geq a$ then $a = b$) and transitive (if $a \geq b$ and $b \geq c$ then $a \geq c$); an ordering is a total ordering if all pairs are comparable, i.e., $\geq : (a, b) \mapsto \{a \geq b, b > a\}$. An ordering always induces an *inverse* ordering, e.g., $\leq : (a, b) \mapsto \{a \leq b, b < a, \text{'not comparable'}\}$. Classical orderings are the inclusion relation on the powerset $2^{\mathcal{Z}}$ of some $\mathcal{Z}$, and the point-wise function comparison $f \geq g$ iff $f(z) \geq g(z)$ for all points $z \in \mathcal{Z}$ and functions $f, g : \mathcal{Z} \to \mathbb{R}$.

A useful yet less-known ordering is introduced in (Khan et al., 2024, Definition 4), which compares experiments looking at their Bayes risk value. For a fixed prior probability, when the Bayes risk of one experiment is consistently better over losses and action sets than the risk of another, we say that the said experiment is *Bayesian-better than* the other. We introduce a similar ordering for our attainable prediction sets.

Definition 7.9 (Bayesian-better-than Ordering for Attainable Predictions)**.** *Given two sets of attainable predictions $\mathcal{F} := \ell \circ \mathcal{H}$ and $\mathcal{G} := \ell' \circ \mathcal{H}'$, with $\ell, \ell' : \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ bounded losses and $\mathcal{H}, \mathcal{H}' \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$,*

$$\mathcal{F} \preceq_{\mathrm{BR}} \mathcal{G} \quad \Leftrightarrow \quad \mathbf{BR}_{\mathcal{F}}(\phi) \leq \mathbf{BR}_{\mathcal{G}}(\phi) \quad \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y}) .$$

Note that $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y}) \mapsto \mathbf{BR}_{\mathcal{F}}(\phi)$ is a well-defined functional, even if the loss function is defined on $\Delta_{\mathrm{ca}}(\mathcal{Y})$. That is because the finitely additive probabilities only appear in the computation of the expectation, and do not affect the loss input.

As a first step in the direction of studying orderings and their relations, we recall that the support function is in a natural correspondence with convex sets. Such natural correspondence is formalized in the following lemma.

Lemma 7.10 (Duality between Closed, Convex Sets and Support Functions)**.** *Let $\mathcal{Z}$ be a*

Polish space and $\mathcal{A} \subseteq \mathcal{B}_b(\mathcal{Z})$ be closed convex and non-empty. Then,

$$\mathcal{A} = \{f \in \mathcal{B}_b(\mathcal{Z}) \colon \rho_{\mathcal{A}}(\mu) \leq \langle f, \mu \rangle,\ \mu \in \mathrm{ba}(\mathcal{Z})\}.$$

Proof. This lemma is a consequence of Property P1 in Lemma 3.22 and Theorem 7.51 in (Aliprantis and Border, 2006). □

Hence, a first connection between orderings of support functions and sets can be immediately proved.

Lemma 7.11 (Equivalence between Set and Support Function Orderings)**.** *Consider two closed convex and non-empty sets $\mathcal{F}, \mathcal{G} \subseteq \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})$. Then,*

$$\mathcal{G} \subseteq \mathcal{F} \quad \Leftrightarrow \quad \rho_{\mathcal{F}}(\mu) \leq \rho_{\mathcal{G}}(\mu) \quad \forall \mu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y}),$$

assuming that $-\infty \leq -\infty$.

Proof. That set containment implies the inequality follows directly. The reverse direction is a consequence of Lemma 7.10. □

Once the technical setup has been laid out, one can also recall another property of the support function: the invariance to convex closure of the set, such that $\rho_{\mathrm{spr}\,\mathcal{F}} = \rho_{\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}}$ (Lemma 3.22 (P4)). In addition, we also aim to ultimately characterize the behavior of learning problems, which involve probability measures Δ_{ba} instead of general, signed measures ba. For this reason, it makes sense to prove the following key result in terms of convex closures of sets and the set of finitely additive probability measures. What it proves is that we can express the correspondence between support functions *of superprediction sets* and closed convex sets in terms of a restricted class of measures, the finitely additive probability distributions.

Lemma 7.12. *Let $\mathcal{F} \subseteq \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$ non-empty. Then, it holds,*

$$\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} = \{\, g \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}) \mid \rho_{\mathrm{spr}\,\mathcal{F}}(\phi) \leq \langle g, \phi \rangle,\ \ \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y}) \,\}.$$

Proof. We start by considering the well-known relationship between sets and support functions formulated in Lemma 7.10, which for $\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}$ amounts to

$$\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} = \{g \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}) \mid \rho_{\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}}(\mu) \leq \langle g, \mu \rangle,\ \forall \mu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y})\}.$$

Now notice that every bounded signed measure $\mu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y})$ can be written as $\alpha\mu' = \mu$ for some $\mu' \in \mathbb{B}_1 := \{\nu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y}) \colon \|\nu\|_\infty = 1\}$ and $\alpha \in \mathbb{R}_{\geq 0}$, e.g. if $\mu \neq 0$, by choosing

$\alpha := \|\mu\|_\infty$ and $\mu' := \frac{\mu}{\alpha}$. Thus, we can rewrite the set as

$$\begin{aligned}\overline{\text{co}}\,\text{spr}\,\mathcal{F} &= \{\, g \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \mid \rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\mu) \le \langle g, \mu\rangle,\ \ \forall \mu \in \text{ba}(\mathcal{X}\times\mathcal{Y})\,\}\\ &= \{\, g \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \mid \rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\alpha\mu) \le \langle g, \alpha\mu\rangle,\ \ \forall \mu \in \mathbb{B}_1, \alpha \in \mathbb{R}_{\ge 0}\,\}\\ &= \{\, g \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \mid \rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\mu) \le \langle g, \mu\rangle,\ \ \forall \mu \in \mathbb{B}_1\,\},\end{aligned}$$

where the last step uses the linearity of the bilinear mapping and Lemma 3.22 (P4) of support functions. Let us introduce a partition of $\mathbb{B}_1$ into two sets of positive and signed finitely additive measures with norm 1, such that,

$$\begin{aligned}\mathcal{P}_+ &:= \mathbb{B}_1 \cap \text{ba}(\mathcal{X}\times\mathcal{Y})_{\ge 0},\\ \mathcal{P}_- &:= \mathbb{B}_1 \setminus \text{ba}(\mathcal{X}\times\mathcal{Y})_{\ge 0},\\ \mathcal{P}_- \cup \mathcal{P}_+ &= \mathbb{B}_1, \mathcal{P}_- \cap \mathcal{P}_+ = \emptyset.\end{aligned}$$

Hence it holds that

$$\overline{\text{co}}\,\text{spr}\,\mathcal{F} = \{\, g \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \mid \rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\mu) \le \langle g, \mu\rangle,\ \ \forall \mu \in \mathcal{P}_- \cup \mathcal{P}_+\,\}.$$

We show now that for every $g \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})$, the condition $\rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\mu) \le \langle g, \mu\rangle$ is satisfied for all $\mu \in \mathcal{P}_-$, and thus we can neglect this set inside the condition. To this end, let us rewrite the support function as

$$\rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\mu) = \inf_{f\in\mathcal{F}} \langle f, \mu\rangle + \inf_{f\in\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\ge 0}} \langle f, \mu\rangle,$$

by Lemma 3.19 and Lemma 3.22 (P5). Observe that for every measure $\mu \in \mathcal{P}_-$ there exists $\mathcal{A} \in \Sigma(\mathcal{X}\times\mathcal{Y})$ such that $-1 \le \mu(\mathcal{A}) < 0$.[1] We can define a subset of functions in $\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\ge 0}$ given by the scaled indicator function of this set, i.e., $\{\alpha\chi_{\mathcal{A}} \mid \alpha \ge 0\} \subset \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\ge 0}$. Hence, we have that

$$\inf_{f\in\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\ge 0}} \langle f, \mu\rangle \le \inf_{\alpha\ge 0} \langle \alpha\chi_{\mathcal{A}}, \mu\rangle = \alpha\mu(\mathcal{A}) \ \ \Rightarrow \inf_{f\in\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\ge 0}} \langle f, \mu\rangle = -\infty,$$

as α gets arbitrarily large while μ is bounded from below. Since $\inf_{f\in\mathcal{F}}\langle f, \mu\rangle$ is finite because $\mathcal{F} \neq \emptyset$, it follows that $\rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\mu) = -\infty$. Hence, we can indeed neglect $\mathcal{P}_-$ in the condition, and write

$$\begin{aligned}\overline{\text{co}}\,\text{spr}\,\mathcal{F} &= \{\, g \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \mid \rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\mu) \le \langle g, \mu\rangle,\ \ \forall \mu \in \mathbb{B}_1\,\}\\ &= \{\, g \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \mid \rho_{\overline{\text{co}}\,\text{spr}\,\mathcal{F}}(\phi) \le \langle g, \phi\rangle,\ \ \forall \phi \in \mathcal{P}_+\,\}.\end{aligned}$$

We conclude the proof observing that, by monotonicity of positive measures, $\mathcal{P}_+ = \Delta_{\text{ba}}(\mathcal{X}\times\mathcal{Y})$. □

[1] Notice that this is not the negative set of $\mathcal{X}\times\mathcal{Y}$ w.r.t. μ, as we are considering the family of finitely additive measures. The Hahn decomposition theorem may not hold (Rao and Rao, 1983)[Chapter 2.6].

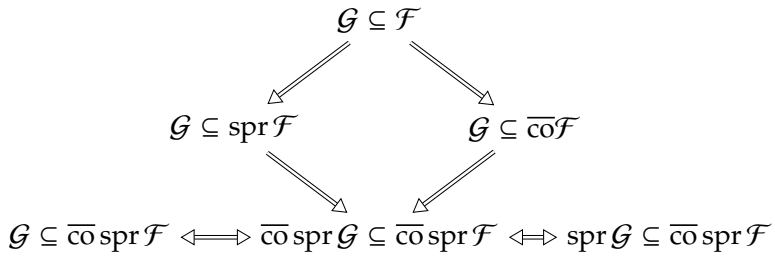


Figure 7.1: Hierarchy of set containment conditions, where the condition in Theorem 7.13 (bottom) and its equivalents are the weakest and the inclusion of the simple sets is the strongest (top). The bottom level equivalences hold by definition of convex hull and superprediction set.

Theorem 7.13 (Equivalence between Set and Support Functions of the Superprediction Orderings)**.** *Consider $\mathcal{F}, \mathcal{G} \subseteq \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$ non-empty. Then, the condition*

$$\overline{\operatorname{co}}\operatorname{spr}\mathcal{F} \supseteq \overline{\operatorname{co}}\operatorname{spr}\mathcal{G} \quad \Leftrightarrow \quad \rho_{\mathcal{F}}(\phi) \leq \rho_{\mathcal{G}}(\phi) \quad \forall \phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})\,.$$

Proof. That set containment implies the inequality follows directly from the fact that ρ is the concave support function defined in Def. 3.21 as an infimum, and that it is immune to convexification and closure of the set. The reverse direction is a consequence of Lemma 7.12. □

The theorem above is a characterization of some generalized version of the data processing inequality for Bayes risk. Key to this interpretation is Lemma 7.12, as it restricts the set of measures on which we have to verify the inequality to the positive and normalized ones. Further specializations of the function sets $\mathcal{F}$ and $\mathcal{G}$ allow for using this result for learning problems. In addition, we can make the set containment weaker or find equivalent conditions by simply using properties of convex hulls. A complete hierarchy of the set containment conditions is illustrated in Fig. 7.1.

Corollary 7.14. *Let $\ell, \ell' \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be bounded loss functions and $\mathcal{H}, \mathcal{H}' \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ model classes. The condition*

$$\overline{\operatorname{co}}\operatorname{spr}(\ell \circ \mathcal{H}) \supseteq \overline{\operatorname{co}}\operatorname{spr}(\ell' \circ \mathcal{H}')$$

holds, if and only if

$$\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell' \circ \mathcal{H}'}(\phi) \quad \forall \phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})\,.$$

Proof. The statement follows directly from Proposition 3.23 and Theorem 7.13. □

We can rephrase the above theorem as the following statement:

> *"An ordering for attainable predictions, $\ell \circ \mathcal{H} \preceq_{\text{BR}} \ell' \circ \mathcal{H}'$, induced by the data processing inequality, is equivalent to the inverse inclusion order on the convex hull of the associated superprediction sets, $\overline{\text{co}}\,\text{spr}\,(\ell \circ \mathcal{H}) \supseteq \overline{\text{co}}\,\text{spr}(\ell' \circ \mathcal{H}')$."*

Observe that this property is quite similar to the equivalence of the Bayesian-better-than and informativeness orderings, as illustrated in Khan et al. (2024). However, our "informativeness" is not Blackwell's analogous for the $\ell \circ \mathcal{H}$ comparison, as it is instead only defined as a property of the attainable predictions instead of attainable (average) losses.

Remark 7.15 (Regularization) *The subsethood condition characterizing the Bayes risk inequality above can be achieved by $\mathcal{H}' \subset \mathcal{H}$ and $\ell = \ell'$. In other words, the Bayes risk of a smaller model class evaluated with the same loss cannot decrease, and therefore we observe a best-case performance degradation. This means that the benefits of regularizing the model class cannot be seen at the Bayes risk level.*

7.2.1 Orderings under Corruption

We can provide another interpretation of Theorem 7.13 by stating a corollary focusing on the comparison of Bayes risk computed for the same loss function and model class, but on a corrupted distribution. In fact, thanks to Lemma 5.1, we know that this is equivalent to comparing $\ell \circ \mathcal{H}$ and $\hat{\kappa}(\ell \circ \mathcal{H})$ through the Bayes risk ordering. A visualization of this property is given in Fig. 7.2.

Corollary 7.16 (Orderings under Corruption)**.** *Let $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. Consider a Markov kernel $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$. Then, the condition*

$$\overline{\text{co}}\,\text{spr}(\ell \circ \mathcal{H}) \supseteq \overline{\text{co}}\,\text{spr}(\hat{\kappa}(\ell \circ \mathcal{H}))\,,$$

holds, if and only if

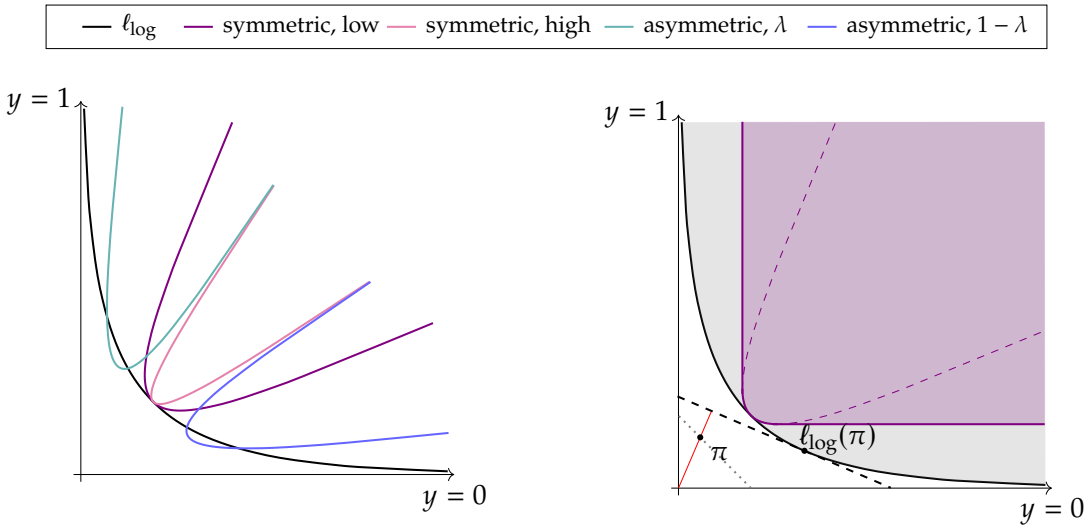
$$\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\kappa}\phi)\,, \quad \forall \phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})\,.$$

Proof. Define $\mathcal{F} \coloneqq \ell \circ \mathcal{H}$ and $\mathcal{G} \coloneqq \hat{\kappa}(\ell \circ \mathcal{H})$ The statement follows directly from Theorem 7.13 together with Proposition 3.23 and Lemma 5.1. □

We now focus on defining and analyzing the set of kernels ensuring comparability of a clean and corrupted attainable prediction set.

Definition 7.17. *Let $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class, we define the set of Markov kernels for which the Bayes risk inequality holds as*

$$\mathbf{K}(\ell \circ \mathcal{H}) \coloneqq \{\, \kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y}) \mid \mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\kappa}\phi) \quad \forall \phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y}) \,\}.$$



(a) Effects of different label corruption kernels on $\ell_{\log}$. Symmetric noise is s.t. the matrix representation's diagonal values are α, the anti-diagonal $1-\alpha$. "Low" and "high" refer to α being respectively lower or higher than $1-\alpha$.

(b) Two $\operatorname{spr}(\ell \circ \mathcal{H})$ projected on the $\mathcal{Y}$ axes, for $\ell_{\log}$ (gray area) and $\hat{\lambda}\ell_{\log}$ (violet area). The black dashed line is the supporting hyperplane at $\ell_{\log}(\pi)$, π in the simplex $\Delta_{\text{ca}}(\mathcal{Y})$ (gray, dotted).

Figure 7.2: We consider a learning problem composed of: a model class $\mathcal{H}$ including all Markov kernels whose transitions h are all measurable functions mapping to the relative interior of $\Delta_{\text{ca}}(\mathcal{Y})$; the logarithmic loss $\ell_{\log}$; $|\mathcal{X}| = 1$; $\mathcal{Y} = \{0, 1\}$. The corruption applied is a label corruption, i.e., $\kappa = \delta_{\mathcal{X}} \otimes \lambda$, only acting on the loss as depicted in panel (a). In the *asymmetric* cases depicted here, we can graphically see that the generalized data processing inequality *does not hold*, as the set containment condition is not fulfilled.
Panel (b) represents the clean and corrupted superprediction sets. The latter is such that some parts of the loss function lie inside the set (drawn as a violet dashed curve). Here we instead show that, for this learning problem and associated *symmetric* corruption, the generalized data processing inequality *holds* because of Corollary 7.16, since $\overline{\text{co}}\operatorname{spr}(\hat{\lambda}\ell_{\log} \circ \mathcal{H}) \subset \overline{\text{co}}\operatorname{spr}(\ell_{\log} \circ \mathcal{H})$. In the next section (Proposition 7.27), we will formally prove that this holds for a generalization of symmetric losses. To better understand visually why set containment implies the Bayes ordering, notice that the Bayes risk associated to the clean problem and $\pi = (0.3, 0.7)$ is the length of the red line, connecting the origin to the hyperplane (dashed line) and perpendicular to it; a similar observation holds for the corrupted superprediction set.

Note that the identity kernel $\delta_{\mathcal{X}} \times \mathcal{Y} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ always belongs to this set, hence it is ensured to be non-empty. Using Corollary 7.16, we can alternatively write,

$$\mathbf{K}(\ell \circ \mathcal{H}) = \{\, \kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y}) \mid \overline{\mathrm{co}}\, \mathrm{spr}(\hat{\kappa}(\ell \circ \mathcal{H})) \subseteq \overline{\mathrm{co}}\, \mathrm{spr}(\ell \circ \mathcal{H}) \,\}.$$

We now prove some useful properties of this set.

Lemma 7.18. *Let $\kappa_1, \kappa_2 \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ and $\alpha \in (0,1)$. Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. For $\kappa \coloneqq \alpha\kappa_1 + (1-\alpha)\kappa_2$,*

$$\mathrm{spr}(\hat{\kappa}(\ell \circ \mathcal{H})) \subseteq \alpha \mathrm{spr}(\hat{\kappa}_1(\ell \circ \mathcal{H})) + (1-\alpha)\mathrm{spr}(\hat{\kappa}_2(\ell \circ \mathcal{H})).$$

Proof. Observe that, by definition, $\mathrm{spr}(\hat{\kappa}(\ell \circ \mathcal{H})) = \mathrm{spr}(\alpha\hat{\kappa}_1 + (1-\alpha)\hat{\kappa}_2(\ell \circ \mathcal{H}))$. Let $u \in \mathrm{spr}(\alpha\hat{\kappa}_1 + (1-\alpha)\hat{\kappa}_1(\ell \circ \mathcal{H}))$, then there exists $h_u \in \mathcal{H}$ such that

$$(\alpha\hat{\kappa}_1 + (1-\alpha)\hat{\kappa}_2)(\ell \circ h_u) = \alpha\hat{\kappa}_1(\ell \circ h_u) + (1-\alpha)\hat{\kappa}_2(\ell \circ h_u) \leq u,$$

where the equality holds by definition of superprediction set Def. 3.17. Let $x_u \coloneqq \alpha\hat{\kappa}_1(\ell \circ h_u) + (1-\alpha)\hat{\kappa}_2(\ell \circ h_u)$ be the point bounding u from below. We can rewrite, for $\alpha \in (0,1]$,

$$u = \alpha\hat{\kappa}_1(\ell \circ h_u) + (u - x_u) + (1-\alpha)\hat{\kappa}_2(\ell \circ h_u) = \alpha\left(\hat{\kappa}_1(\ell \circ h_u) + \frac{1}{\alpha}(u - x_u)\right) + (1-\alpha)\hat{\kappa}_2(\ell \circ h_u),$$

where $\hat{\kappa}_1(\ell \circ h_u) + \frac{1}{\alpha}(u - x_u)$ is a point of the set $\mathrm{spr}(\hat{\kappa}_1(\ell \circ \mathcal{H}))$, because $u - x_u \geq 0$ and $\alpha \in (0,1)$. Since also $\hat{\kappa}_1(\ell \circ h_u) \in \mathrm{spr}(\hat{\kappa}_2(\ell \circ \mathcal{H}))$, we get $u \in \alpha\, \mathrm{spr}(\hat{\kappa}_1(\ell \circ \mathcal{H})) + (1-\alpha)\, \mathrm{spr}(\hat{\kappa}_1(\ell \circ \mathcal{H}))$. □

Proposition 7.19 (**K** is a Convex Set)**.** *Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. The associated set $\mathbf{K}(\ell \circ \mathcal{H})$ is convex.*

Proof. Let $\kappa_1, \kappa_2 \in \mathbf{K}(\ell \circ \mathcal{H})$ and $\alpha \in [0,1]$. We show that $\kappa \coloneqq \alpha\kappa_1 + (1-\alpha)\kappa_2 \in \mathbf{K}$, as convexity then follows by a trivial inductive argument. By definition of $\mathbf{K}(\ell \circ \mathcal{H})$, it holds that

$$\hat{\kappa}_i(\ell \circ \mathcal{H}) \subseteq \overline{\mathrm{co}}\, \mathrm{spr}(\ell \circ \mathcal{H}), \quad i \in \{1,2\}. \tag{7.3}$$

Then, considering the corrupted superprediction set, we can write

$$\mathrm{spr}(\hat{\kappa}(\ell \circ \mathcal{H})) \overset{L.7.18}{\subseteq} \alpha\, \mathrm{spr}(\hat{\kappa}_1(\ell \circ \mathcal{H})) + (1-\alpha)\, \mathrm{spr}(\hat{\kappa}_2(\ell \circ \mathcal{H})) \overset{(\star)}{\subseteq} \overline{\mathrm{co}}\, \mathrm{spr}(\ell \circ \mathcal{H})$$

where $(\star)$ holds by convexity and the set containment in Eq. (7.3). Hence, it follows that $\mathrm{spr}(\hat{\kappa}(\ell \circ \mathcal{H})) \subseteq \overline{\mathrm{co}}\, \mathrm{spr}(\ell \circ \mathcal{H})$, which implies the thesis. □

It is then a matter of simple computations to show a weaker type of Bayes risk inequality based upon the definition of $\mathbf{K}(\ell \circ \mathcal{H})$ and Corollary 7.16. In words, we prove that for a given distribution $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$, a loss function ℓ and a model class $\mathcal{H}$, we can construct

a set of distributions $\psi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})$ for which the Bayes risk is lower than for the reference distribution ϕ.

Corollary 7.20. *Let $\phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})$, $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. We can define the set of probabilities that can reach ϕ through some $\kappa \in \mathbf{K}(\ell \circ \mathcal{H})$ as*

$$\mathbf{K}(\ell \circ \mathcal{H})^{\dagger}_{\phi} := \{\check{\kappa}^{\dagger}_{\phi}\phi \mid \kappa \in \mathbf{K}(\ell \circ \mathcal{H})\} \subseteq \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y}),$$

where $\check{\kappa}^{\dagger}_{\phi}$ denotes the adjoint Markov kernel operator of the Bayesian inverse $\kappa^{\dagger}_{\phi}$ of κ. Hence, the data processing inequality holds on it, i.e.,

$$\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\psi) \quad \forall \psi \in \mathbf{K}(\ell \circ \mathcal{H})^{\dagger}_{\phi}\,.$$

Proof. For every $\psi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})$ such that there exists $\kappa \in \mathbf{K}(\ell \circ \mathcal{H})$ with $\check{\kappa}\psi = \phi$, it holds by definition of $\mathbf{K}(\ell \circ \mathcal{H})$ that,

$$\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\psi)\,.$$

Under our assumption of standard Borel measure spaces for $\mathcal{X}, \mathcal{Y}$, we can characterize the set of ψ s.t. the inequality holds using the Bayesian inverse of the Markov kernel $\kappa \in \mathbf{K}(\ell \circ \mathcal{H})$ w.r.t. the chosen ϕ (see Def. 5.9). Hence, we can write $\psi := \check{\kappa}^{\dagger}_{\phi}\phi$. We therefore obtain the thesis. □

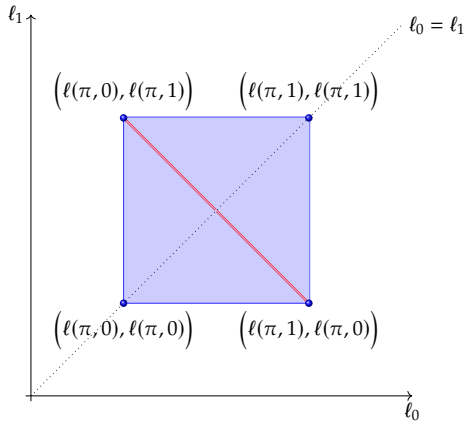
7.2.2 Convex Combination of Kernels on Superprediction Set

Corollary 7.16 teaches us that, for a general $\phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})$, it is equivalent to understanding and the superprediction set of $\hat{\kappa}(\ell \circ \mathcal{H})$, to show a Bayes risk inequality. Understanding what the effect of the kernel κ is on $\ell \circ \mathcal{H}$, however, is non-trivial.

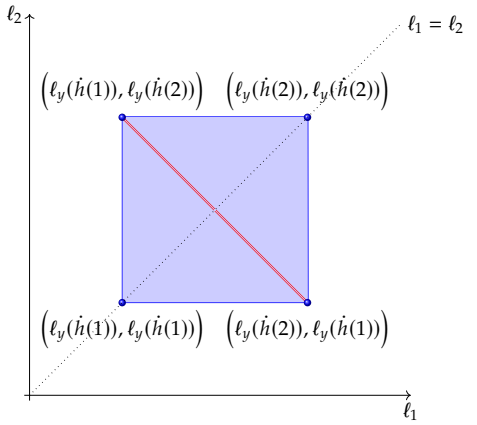
A first observation comes from the following: because of its normalization property ($\kappa(\mathcal{A}, z) = 1\ \forall z \in \mathcal{Z}$, for some $\mathcal{A} \in \mathcal{Z}'$), and because a kernel $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ action on a function is defined as,

$$\hat{\kappa}(\ell(\dot{h}(x), y)) := \int_{\mathcal{X} \times \mathcal{Y}} \kappa(x, y, d(\tilde{x}, \tilde{y}))\, \ell(\dot{h}(\tilde{x}), \tilde{y}) \qquad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\,,$$

we know that the kernel convexly combines the values of a mapping $(x, y) \mapsto \ell(\dot{h}(x), y)$ for different (x, y)-pairs. For illustration, fix $x \in \mathcal{X}$ and suppose that $\mathcal{Y} = \{0, 1\}$ is binary. Choose a loss ℓ, a model $h \in \mathcal{H}$ with $\dot{h}(x) = \pi$ and a label corruption $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$. We are interested in the point $q_1 = \big(\ell(\pi, 0), \ell(\pi, 1)\big)$ and its element-wise convex combinations $(\lambda\ell)_x(\pi, \tilde{y}) \in \text{co}(\{\ell(\pi, y)\}_{y \in \mathcal{Y}})$ for a fixed x. The effect of the kernel is the point $q_\lambda = \big(\lambda\ell(\pi, 0), \lambda\ell(\pi, 1)\big)$. Note that because λ is label corruption it does not have an effect on the $x \in \mathcal{X}$, but it only depends on it. An illustration of all the possible outcomes for all possible λ kernels is given by Fig. 7.3a, i.e., the blue area. We will use the intuition gained through this example to build conditions for the data processing inequality to hold.



(a) Visualization of the effect of label corruption $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$ on a loss function ℓ evaluated on a prediction $\pi := \dot{h}(x) \in \Delta_{\text{ba}}(\mathcal{Y})$ for some fixed $x \in \mathcal{X}$, assuming $\mathcal{Y} = \{0, 1\}$.



(b) Visualization of the effect of attribute corruption $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ on a loss function ℓ, evaluated on a fixed model $h \in \mathcal{H}$ and some fixed $y \in \mathcal{Y}$, assuming $\mathcal{X} := \{1, 2\}$.

Figure 7.3: Visualization of a point $\hat{\kappa}(\ell \circ h)$ given a point $\ell \circ h \in \text{spr}(\ell \circ \mathcal{H})$, by corruption type and properties. Blue area: corruption outcome for all possible kernels; Red line: bistochastic corruption; Blue marks: points generated by a degenerate binary kernel (ones only in one row, zeros otherwise).

Additional insights on possible strategies to prove a data processing result can come from Proposition 7.19. The proposition shows that also the subset of kernels fulfilling the data processing inequality for a given choice of loss and model class is convex. In general, this gives us no knowledge on the extreme points of such a set knowing the ones of the whole set of Markov kernels. Nevertheless, if we assume that we know which extreme points generate such subset of kernels, we may get interesting results. We prove a first preliminary result needed to proceed in this direction.

Proposition 7.21 (Generalized Data Processing Inequality with Convex Combination of Kernels)**.** *Let $\kappa_1, \kappa_2 \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$. Let $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class such that $\hat{\kappa}_i(\ell \circ h) \in \overline{\text{co}}\,\text{spr}(\ell \circ \mathcal{H})$ for all $h \in \mathcal{H}$ and $\kappa_i \in \{\kappa_1, \kappa_2\}$. Then, for all $\kappa := \alpha\kappa_1 + (1-\alpha)\kappa_2$ with $\alpha \in [0, 1]$,*

$$\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\kappa}\phi)\,, \quad \forall \phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})\,.$$

Proof. The statement follows immediately from Proposition 7.19 and Corollary 7.16. □

We will now take a closer look at the label and attribute corruption, for which this result will have different consequences.

7.3 Sufficient Conditions for the Bayes Risk Ordering

In the following, we leverage our characterization result (Corollary 7.16) to derive sufficient conditions for the Bayes risk inequality in the case of dependent corruptions, while we

find a new characterization for certain types of simple corruptions. These conditions apply to specific families of learning problems and corruptions, where learning problems are specified through assumptions on the loss function and the model class, but not on the underlying probability distribution. Such distributional generality is required to obtain a data processing result in the sense of Def. 7.8. Our objective is to identify learning problems for which certain classes of Markov kernels are included in $\mathbf{K}(\ell \circ \mathcal{H})$. Since $\mathbf{K}(\ell \circ \mathcal{H})$ is convex, it suffices to study its extreme points, which considerably simplifies the analysis.

In the binary setting depicted in Fig. 7.3, all relevant kernel actions can be expressed as convex combinations of deterministic kernels induced by constant functions (degenerate binary kernels) or certain bijective functions (symmetric bistochastic kernels). Applying Proposition 7.21, it is therefore sufficient to impose conditions on $\ell \circ \mathcal{H}$ that guarantee a decrease in Bayes risk under both degenerate and bistochastic kernels: under such conditions, the generalized data processing inequality holds for all kernels. For larger domains, the class of relevant kernels is substantially richer. Nevertheless, we keep in mind these two fundamental cases as examples, while we establish sufficient conditions for the generalized DPI with families of kernels that are generated by certain deterministic kernels. As the extreme points of the set of Markov kernels are deterministic kernels (Proposition 2.30), these sets will share a portion of their boundary with the set of all Markov kernels with the same signature. We imagine that showing results for differently constructed sets of kernels may require more sophisticated arguments; investigating the GDPI for such cases is left as future work.

In the next section, we begin by showing that when the loss function and the model class satisfy a suitable closure property with respect to label transformations, then optimal performance as measured by the Bayes risk can only deteriorate under label corruptions induced by such transformations, and vice versa. By strengthening this invariance assumption, we extend the result to label corruptions that depend on the attribute $x \in \mathcal{X}$ as a sufficient condition. An important special case is that of bistochastic label corruption, for which we derive sufficient conditions on ℓ and $\mathcal{H}$ ensuring that the Bayes risk inequality holds.

We then establish analogous results for attribute corruption and illustrate the corresponding statements using commonly encountered and practically relevant transformations. Finally, we present examples demonstrating that all conditions introduced in this section are sufficient but not necessary for the Bayes risk inequality to hold.

Some Corruption Kernels of Interest

We give here definitions for different types of corruption kernels, which have a relevant structure for our data processing results. They complement the taxonomy illustrated in § 4.

Definition 7.22 (Structured Label Corruptions)**.** *Let $\mathcal{X}$ be an input space and $\mathcal{Y}$ be an output space.*

1. *A Markov kernel $\lambda\colon \mathcal{X}\times\mathcal{Y}\times\Sigma(\mathcal{Y})\to[0,1]$ is called* $\mathcal{F}$-label corruption*, if for every $x\in\mathcal{X}$ there exists a finite tuple $(\alpha_i^x)_{i\in I}\in[0,1]^{|I|}$ such that $\sum_{i\in I}\alpha_i^x=1$ and a finite set of measurable mappings $f_i^x\in\mathcal{F}\subseteq\{\mathcal{Y}\to\mathcal{Y}\}$ for every $i\in I$, such that,*

$$\lambda_x\colon \mathcal{Y}\times\Sigma(\mathcal{Y})\to[0,1];\quad (y,\mathcal{B})\mapsto\sum_{i\in I}\alpha_i^x\kappa_{f_i^x}(y,\mathcal{B})\ \ \forall y\in\mathcal{Y},\mathcal{B}\in\Sigma(\mathcal{Y}),$$

   where the deterministic kernels $\kappa_{f_i^x}$ are defined as in Def. 2.21.

2. *A Markov kernel $\lambda\colon\mathcal{Y}\times\Sigma(\mathcal{Y})\to[0,1]$ is called* bistochastic label corruption*, if it is a $\mathcal{F}$-label corruption where $\mathcal{F}$ is the set of all measurable, bijective mappings.*

Analogously, we can also define the attribute corruption counterpart to the previous definition.

Definition 7.23 (Structured Attribute Corruptions)**.** *Let $\mathcal{X}$ be an input space and $\mathcal{Y}$ be an output space.*

1. *A Markov kernel $\tau\colon\mathcal{X}\times\mathcal{Y}\times\Sigma(\mathcal{X})\to[0,1]$ is called* $\mathcal{G}$-attribute corruption*, if for every $y\in\mathcal{Y}$ there exists a finite tuple $(\alpha_i^y)_{i\in I}\in[0,1]^{|I|}$ such that $\sum_{i\in I}\alpha_i^y=1$ and a finite set of measurable mappings $g_i^y\in\mathcal{G}\subseteq\{\mathcal{X}\to\mathcal{X}\}$ for every $i\in I$, such that,*

$$\tau_y\colon\mathcal{X}\times\Sigma(\mathcal{X})\to[0,1];\quad (x,\mathcal{A})\mapsto\sum_{i\in I}\alpha_i^y\kappa_{g_i^y}(y,\mathcal{B})\ \ \forall x\in\mathcal{X},\mathcal{A}\in\Sigma(\mathcal{X}),$$

   where the deterministic kernels $\kappa_{g_i^y}$ are defined as in Def. 2.21.

2. *A Markov kernel $\tau\colon\mathcal{X}\times\mathcal{Y}\times\Sigma(\mathcal{Y})\to[0,1]$ is called* bistochastic attribute corruption*, if it is a $\mathcal{G}$-label corruption where $\mathcal{G}$ is the set of all measurable, bijective mappings.*

The illustrations in Fig. 7.3 show how finite bistochastic kernels act on a point of the superprediction set, and how this differs from a general corruption kernel. It suggests that the GDPI may hold under assumptions on the superprediction set determined by the corruption kernel structure.

Remark 7.24 *On finite spaces, bistochastic corruptions are represented by doubly stochastic matrices, as we know that the Birkhoff-von Neumann theorem holds (Birkhoff, 1946). The same statement does not hold for general continuous spaces, which is a relevant observation to quantum mechanics. We do not investigate here possible generalizations of the theorem, as it is an open problem outside the scope of this manuscript, but we point the interested reader to some recent developments in this direction, e.g., (Safarov, 2005; Păunescu and Rădulescu, 2017).*

7.3.1 Label Corruption

We start from the simplest form of label corruption in our taxonomy (§ 4): simple label corruption generated by deterministic kernels.

Proposition 7.25. *Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{Y} \to \mathcal{Y}$. Let us define,*

$$\mathcal{L} \coloneqq \mathrm{co}\{\lambda_f \in \mathcal{M}(\mathcal{Y}, \mathcal{Y}) \mid f \in \mathcal{F}\},$$

where λ_f is a deterministic kernel, equal to $\delta_{f(y)}$. Let $\ell : \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. The set containment condition

$$\bigcup_{f \in \mathcal{F}} \{\hat{\lambda}_f \ell \circ h \mid h \in \mathcal{H}\} \subseteq \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H}), \tag{7.4}$$

holds if and only if the generalized data processing inequality $\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\lambda}\phi)$ holds for all $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y}), \lambda \in \mathcal{L}$.

Proof. By definition of superprediction set (Def. 3.17) the superprediction set containment $\overline{\mathrm{co}}\,\mathrm{spr}\mathcal{G} \subseteq \overline{\mathrm{co}}\,\mathrm{spr}\mathcal{F}$ holds if and only if $\mathcal{G} \subseteq \overline{\mathrm{co}}\,\mathrm{spr}\mathcal{F}$. Choose $\mathcal{F} = \ell \circ \mathcal{H}$ and $\mathcal{G} = \mathrm{co}\left(\bigcup_{f \in \mathcal{F}} \{\hat{\lambda}_f \ell \circ h \mid h \in \mathcal{H}\}\right)$. We know that $\mathcal{G} = \bigcup_{\lambda \in \mathcal{L}} \{\hat{\lambda} \ell \circ h \mid h \in \mathcal{H}\}$, as

$$(\lambda \ell \circ h)(x, y) \coloneqq \left[\left(\sum_{i=1}^{n} \alpha_i \lambda_{f_i} \ell\right) \circ h\right](x, y) = \left[\sum_{i=1}^{n} \alpha_i (\lambda_{f_i} \ell \circ h)\right](x, y).$$

Now noticing that $\rho_{\mathcal{A} \cup \mathcal{B}}(\phi) = \min\{\rho_{\mathcal{A}}(\phi), \rho_{\mathcal{B}}(\phi)\} \leq \rho_*(\phi)$ for $(*) \in \{\mathcal{A}, \mathcal{B}\}$ because of properties of the infimum, we can apply the support function characterization result from Theorem 7.13 and the equality between Bayes risk and support function from Proposition 3.23 to obtain the thesis. □

Example 7.26 (Degenerate Simple Label Corruption)**.** *Let $\mathcal{F}$ be the class of constant functions, i.e., $f \in \mathcal{F}$ is such that $f(y) = y_c$ for some $y_c \in \mathcal{Y}$ and all $y \in \mathcal{Y}$. Because of the behavior of the constant functions ignoring the input, we name the kernels in the induced $\mathcal{L}$ as "degenerate corruptions", in line with the caption in Fig. 7.3. For these kernels, the condition in Eq. (7.4) translates to requiring a specific set of constant functions to be included in the superprediction set, namely,*

$$\{(x, y) \mapsto \tilde{\ell}(\dot{h}(x), y) \coloneqq \min_{y' \in \mathcal{Y}} \ell(\dot{h}(x), y') \mid h \in \mathcal{H}\} \subseteq \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H}). \tag{7.5}$$

In other words, it is sufficient to take a loss function ℓ constant in y for the condition to be met. However, the condition is not particularly relevant when studying learning problems. A more constructive condition is found noticing that the condition in Eq. (7.5) can be further rewritten as

$$\{(x, y) \mapsto \inf_{h \in \mathcal{H}} \min_{y' \in \mathcal{Y}} \ell(\dot{h}(x), y') = \min_{y' \in \mathcal{Y}} \inf_{h \in \mathcal{H}} \ell(\dot{h}(x), y') \mid h \in \mathcal{H}\} \subseteq \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H}),$$

which is asking for the vector of label-wise minimum losses to be included in the superprediction set. To build sets of functions with such a property, we first pick the best performing h^* for each attribute x, which is possibly not in $\mathcal{H}$; then, we select the y^* for which such model achieves the lowest loss; lastly, we impose that the loss function is constant only on certain probability values, i.e.,

$$\ell(p, y) := \begin{cases} \ell(p, y^*) & \text{if } \exists x \in \mathcal{X} \mid p = h^*(x) \\ \ell(p, y) & \text{if } p \neq h^*(x)\ \forall x \in \mathcal{X} \end{cases}$$

and additionally impose $h^* \in \mathcal{H}$. However, if the model h^* covers the whole simplex, we will be forced to require the loss to be always constant.

We now show that we can go beyond simple label corruption when strengthening the assumptions on the ℓ and $\mathcal{H}$ combination.

Proposition 7.27 (Generalized Data Processing Inequality with $\mathcal{F}$-Label Corruption)**.** *Let $\mathcal{Y}$ be finite and $\mathcal{F}$ a class of functions $\mathcal{Y} \to \mathcal{Y}$. Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. Consider the following conditions:*

1. *The loss function ℓ is $\mathcal{F}$-*symmetric*, i.e., $\ell(\pi, y) = \ell([\pi_{f(y)}]_y, f(y))$, for all $f \in \mathcal{F}$, where $[\pi_{f(y)}]_y$ denotes the corresponding coordinate-wise rearrangement of the probability vector $\pi \in \Delta_{\mathrm{ba}}(\mathcal{Y})$ induced by f;*
2. *The model class is $\mathcal{F}$-*label-invariant*, i.e., for every $h \in \mathcal{H}$, and every $f \in \mathcal{F}$, there exists $h_f \in \mathcal{H}$ such that $\dot{h}_f(x) = [\dot{h}(x)_{f(y)}]_y$, for all $x \in \mathcal{X}$;*
3. *The kernel is a $\mathcal{F}$-label corruption $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$.*

If all of these conditions are fulfilled, then $\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\lambda}\phi)$ for all $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$.

Proof. We show that the application of the kernel λ on the attainable predictions $\ell \circ \mathcal{H}$ leads to the subsethood condition $\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\lambda}(\ell \circ \mathcal{H})) \subseteq \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H})$. Then, we apply Corollary 7.16. We can write a generic element of $\hat{\lambda}\ell \circ \mathcal{H}$ as

$$\begin{aligned} \hat{\lambda}((\ell \circ h)(x, y)) &= \int_{\mathcal{Y}} \ell(\dot{h}(x), \tilde{y})\, \lambda(x, y, d(\tilde{y})) = \sum_{i \in I} \alpha_i^x \int_{\mathcal{Y}} \ell(\dot{h}(x), \tilde{y})\, \kappa_{f_i^x}(y, d(\tilde{y})) \\ &= \sum_{i \in I} \alpha_i^x \ell(\dot{h}(x), f_i^x(y)) = \sum_{i \in I} \alpha_i^x \ell([\dot{h}(x)_{f_i^x(y)}]_y, y). \end{aligned}$$

By $\mathcal{F}$-label invariance, there exists $\dot{h}_{f_i^x}(x) = [\dot{h}(x)_{f_i^x(y)}]_y$ s.t. $h_{f_i^x} \in \mathcal{H}$, hence $(x, y) \mapsto \ell([\dot{h}(x)_{f_i^x(y)}]_y, y) \in \ell \circ \mathcal{H}$ by definition. It follows that

$$\sum_{i \in I} \alpha_i^x \ell([\dot{h}(x)_{f_i^x(y)}]_y, y) \in \mathrm{co}(\ell \circ \mathcal{H}) \subseteq \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H}).$$

□

Remark 7.28 *In Proposition 7.25 we make use of the left implication chain in Fig. 7.1 while in the*

proof of Proposition 7.27 we make use of the right implication chain in Fig. 7.1. In particular, the assumptions $\mathcal{F}$-symmetry and $\mathcal{F}$-label-invariance, essentially guarantee that the strong sufficient condition $\hat{\lambda}(\ell \circ \mathcal{H}) \subseteq \ell \circ \mathcal{H}$ holds for λ being a $\mathcal{F}$-label corruption. This assumption is not enough to prove the necessity of the condition, differently from the approach used in Proposition 7.25.

Corollary 7.29 (Generalized Data Processing Inequality with Bistochastic Label Corruption)**.** *Let $\mathcal{Y}$ be finite and $\mathcal{F}$ a class of functions $\mathcal{Y} \to \mathcal{Y}$. Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a bounded loss function and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. Consider the following conditions:*

1. *The loss function ℓ is* symmetric*, i.e., $\ell(\pi, y) = \ell([\pi_{\sigma(y)}]_y, \sigma(y))$, for all bijective $\sigma \colon \mathcal{Y} \to \mathcal{Y}$, where $[\pi_{\sigma(y)}]_y$ denotes the corresponding coordinate-wise rearrangement of the probability vector $\pi \in \Delta_{\mathrm{ba}}(\mathcal{Y})$ induced by σ;*
2. *The model class is $\mathcal{Y}$-*permutation invariant*, i.e., for every $h \in \mathcal{H}$, and every bijective $\sigma \colon \mathcal{Y} \to \mathcal{Y}$, there exists $h_\sigma \in \mathcal{H}$ such that $\dot{h}_\sigma(x) = [\dot{h}(x)_{\sigma(y)}]_y$, for all $x \in \mathcal{X}$;*
3. *The kernel is a bistochastic label corruption $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$.*

If all of them are fulfilled, then $\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\lambda}\phi)$ holds for all $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$.

Proof. We apply Proposition 7.27 with its $\mathcal{F}$ the set of all bijective functions from $\mathcal{Y}$ to $\mathcal{Y}$. □

Remark 7.30 *The conditions in Corollary 7.29 agree with previous findings regarding robust losses. In particular, Ghosh et al. (2017, Theorem 1) proves that symmetric losses are robust to low-rate symmetric label corruption, as well other non-symmetric and attribute-dependent forms of labels corruption (Theorems 2 and 3). Both cases are subsumed by our result, which includes all noise rates, but in the form of an inequality.*

Example 7.31. *Let $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$, $\mathcal{Y} = \{0, 1\}$. The loss function $\ell(\pi, y) = (y - \pi)^2$ is symmetric. Any model class $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ such that for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$, there is $h' \in \mathcal{H}$ such that $1 - \dot{h}(x) = \dot{h}'(x)$, is $\mathcal{Y}$-permutation invariant. Hence, by Corollary 7.29, label corruption does not degrade the performance of this learning problem.*

7.3.2 Attribute Corruption

Analogously to Proposition 7.25 we can provide a sufficient condition for the Bayes risk inequality under attribute transformations.

Proposition 7.32. *Let $\mathcal{F}$ be a class of measurable functions $f \colon \mathcal{X} \to \mathcal{X}$. Let us define,*

$$\mathcal{T} \coloneqq \mathrm{co}\left\{\tau_f \in \mathcal{M}(\mathcal{X}, \mathcal{X}) \,|\, f \in \mathcal{F}\right\}.$$

where τ_f is a deterministic kernel, equal to $\delta_{f(x)}$. Let $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ bounded and

$\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$. *The set containment condition*

$$\bigcup_{f \in \mathcal{F}} \left\{ \hat{\tau}_f(\ell \circ h) \,|\, h \in \mathcal{H} \right\} \subseteq \overline{\text{co}}\, \text{spr}(\ell \circ \mathcal{H}), \tag{7.6}$$

holds if and only if the generalized data processing inequality $\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\tau}\phi)$ *holds for all* $\phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y}), \tau \in \mathcal{T}$.

Proof. Choose $\mathcal{F} = \ell \circ \mathcal{H}$ and $\mathcal{G} = \text{co}\left(\bigcup_{f \in \mathcal{F}} \left\{ (x, y) \to \hat{\tau}_f(\ell \circ h)(x, y) = \ell((\dot{h} \circ f)(x), y) \,|\, h \in \mathcal{H} \right\} \right)$. We know that $\mathcal{G} = \bigcup_{\tau \in \mathcal{T}} \left\{ \hat{\tau}(\ell \circ h) \,|\, h \in \mathcal{H} \right\}$, because

$$[\tau(\ell \circ h)](x, y) \coloneqq \left[\left(\sum_{i=1}^{n} \alpha_i \tau_{f_i} \right) (\ell \circ h) \right](x, y) = \left[\sum_{i=1}^{n} \alpha_i \tau_{f_i} (\ell \circ h) \right](x, y),$$

as we can exchange the finite sum with the action of a kernel on some functions, as it is defined via an integral of integrable functions. The rest of the proof is analogous to the one of Proposition 7.25. □

Remark 7.33 (Recovery of Classical Data Processing Inequality for Bayes Risk) *The standard data processing result for unconstrained settings can be easily proved using a technique similar to the proposition above. Let* $\ell \colon \Delta_{\text{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ *be a bounded loss function,* $\mathcal{H} = \mathcal{M}(\mathcal{X}, \mathcal{Y})$ *and* $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$. *Then, for all* $h \in \mathcal{H}$,

$$\hat{\tau}(\ell \circ h) \coloneqq \int_{\tilde{x} \in \mathcal{X}} \ell(\dot{h}(\tilde{x}), y)\, \tau(x, \tilde{x})\,, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\,.$$

We know that Proposition 2.30 holds. It follows that we can write τ *as the convex combination of some deterministic Markov kernels. Hence, by definition of extreme points of a set, there exists a finite set of measurable functions* $\mathcal{G} \subseteq \{g \colon \mathcal{X} \to \mathcal{X}\}$ *and* $\alpha_g \in [0, 1]$ *with* $\sum_{g \in \mathcal{G}} \alpha_g = 1$, *such that,*

$$\tau = \sum_{g \in \mathcal{G}} \alpha_g \tau_g\,,$$

where τ_g *denotes the deterministic kernel corresponding to the function* $g \in \mathcal{G}$. *Using the notation* $\tau_g(\tilde{x}, x)$ *for the matrix entry corresponding to row* $\tilde{x}$, *column* x, *we can write*

$$\begin{aligned} \hat{\tau}(\ell \circ h) &= \int_{\tilde{x} \in \mathcal{X}} \ell(\dot{h}(\tilde{x}), y) \sum_{g \in \mathcal{G}} \alpha_g \tau_g(x, d\tilde{x})\,, \\ &= \sum_{g \in \mathcal{G}} \alpha_g \int_{\tilde{x} \in \mathcal{X}} \tau_g(x, d\tilde{x})\, \ell(\dot{h}(\tilde{x}), y) = \sum_{g \in \mathcal{G}} \alpha_g\, \ell(\dot{h}_g(x), y)\,, \end{aligned}$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, *and where* $\dot{h}_g \coloneqq \dot{h} \circ g$ *is a transition of some kernel* h_g *in* $\mathcal{M}(\mathcal{X}, \mathcal{Y})$. *By*

definition of all sets involved,

$$\sum_{g\in\mathcal{G}} \alpha_g \ell(\dot{h}_g(x), y) \in \operatorname{co}(\ell \circ \mathcal{H}) \subseteq \overline{\operatorname{co}}\operatorname{spr}(\ell \circ \mathcal{H})\,.$$

The Bayes risk inequality follows by Corollary 7.16. Since the statements are true for a general loss ℓ and joint probability ϕ, we can pick $\phi = \pi_Y \times E$ and obtain that the inequality holds for all ℓ, π_Y. We therefore recover the classical DPI, as stated in Corollary 6.22.

As done for the label corruption case, we can go beyond the simple $\mathcal{G}$-attribute corruption.

Proposition 7.34 (Generalized Data Processing Inequality with $\mathcal{F}$-Attribute Corruption)**.** *Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathcal{X}$. Let $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. Consider the following conditions:*

1. *The model class is $\mathcal{F}$-*attribute-invariant*, i.e., for every $h \in \mathcal{H}$, and every $f \in \mathcal{F}$, there exists $\dot{h}_f \in \mathcal{H}$ such that $\dot{h}_f(x) = \dot{h}(f(x))$, for all $x \in \mathcal{X}$;*
2. *The kernel is a $\mathcal{F}$-attribute corruption $\tau \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X})$.*

If both these conditions are fulfilled, then $\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell\circ\mathcal{H}}(\check{\tau}\phi)$ holds for all $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$.

Proof. We show that the application of the kernel τ on the set $\ell \circ \mathcal{H}$ leads to the subsethood condition $\overline{\operatorname{co}}\operatorname{spr}(\hat{\tau}(\ell \circ \mathcal{H})) \subseteq \overline{\operatorname{co}}\operatorname{spr}(\ell \circ \mathcal{H})$. Then, we apply Corollary 7.16. Observe that we can write

$$\begin{aligned}\hat{\tau}((\ell \circ h)(x, y)) &= \int_{\mathcal{X}} \ell(\dot{h}(\tilde{x}), y)\tau(x, y, d(\tilde{x})) \\ &= \sum_{i\in I} \alpha_i^y \int_{\mathcal{X}} \ell(\dot{h}(\tilde{x}), y)\, \kappa_{f_i^y}(x, d(\tilde{x})) = \sum_{i\in I} \alpha_i^y \ell(\dot{h}(f_i^y(x)), y)\,.\end{aligned}$$

By $\mathcal{G}$-attribute invariance, there exists $\dot{h}_{f_i^y}(x) = \dot{h}(f_i^y(x)) \in \mathcal{H}$. It follows that

$$\sum_{i\in I} \alpha_i^y \ell(\dot{h}(f_i^y(x)), y) \in \operatorname{co}(\ell \circ \mathcal{H}) \subseteq \overline{\operatorname{co}}\operatorname{spr}(\ell \circ \mathcal{H}).$$

□

Remark 7.35 *An analogous statement to Remark 7.28 holds.*

Corollary 7.36 (GDPI with Bistochastic Attribute Corruption)**.** *Let $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function and $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ a model class. Consider the following conditions:*

1. *The model class is $\mathcal{X}$-*permutation invariant*, i.e., given $h \in \mathcal{H}$, for every $\sigma\colon \mathcal{X} \to \mathcal{X}$ bijective and measurable, there exists h_σ such that $\dot{h}(\sigma(x)) = h_\sigma(x)$;*

2. *The kernel is a bistochastic attribute corruption* $\tau \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X})$.

If both these conditions are fulfilled, then $\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\tau}\phi)$ *holds for all* $\phi \in \Delta_{\text{ba}}(\mathcal{X} \times \mathcal{Y})$.

Proof. Proposition 7.34 with $\mathcal{F}$ the set of all bijective functions from $\mathcal{X}$ to $\mathcal{X}$. □

7.3.3 Sufficient but not Necessary Conditions

The set containment condition in Eq. (7.4) (respectively Eq. (7.6)) is necessary and sufficient for all simple label corruption $\lambda \in \mathcal{L}$ (respectively $\tau \in \mathcal{T}$) to fulfill Corollary 7.16 only when we impose the set containment condition for the closed and convexified super prediction set. Considering the simple superprediction set would instead lead the condition to be only sufficient, as it can be deduced from Fig. 7.1. We can construct a simple example to better understand this phenomenon.

Example 7.37. *Let* $\mathcal{X} = \{x\}$ *and* $\mathcal{Y} = \{0, 1\}$. *A predictor is a distribution on* $\mathcal{Y}$ *which is defined through the probability on* $y = 1$, *for which we simply write* $p \in [0, 1]$. *Let* $\mathcal{H} = [0, \gamma]$ *for some* $\gamma \in (0, 0.5)$. *Let* ℓ *be a cost-weighted loss function, i.e.,* $\ell(y, p) = \mathbf{1}[p < \gamma, y = 1] + \mathbf{1}[p > \gamma, y = 0] + \gamma \mathbf{1}[p = \gamma]$. *If* $\gamma = 0.5$, *and neglecting the last tie-breaking summand, then* ℓ *is the* 0-1*-loss. The restrictions put to* ℓ *and* $\mathcal{H}$ *lead to,*

$$\ell \circ \mathcal{H} = \left\{ \big((x, 0) \mapsto 0, (x, 1) \mapsto 1\big), \big((x, 0) \mapsto \gamma, (x, 1) \mapsto \gamma\big) \right\}.$$

Let λ *be uniform bistochastic simple label corruption. Hence, it is the convex combination of bistochastic simple label noises which are defined through the two permutations on* $\mathcal{Y}$ *by Birkhoff-von Neumann theorem (Birkhoff, 1946). It is now a matter of simple computations to show that the Bayes risk inequality is preserved for the uniform bistochastic simple label corruption, but not for both simple permutation label noises.*

As we mentioned earlier, Proposition 7.27 (respectively Proposition 7.34) are both built upon sufficient conditions at the strongest level, namely $\kappa(\ell \circ \mathcal{H}) \subseteq \ell \circ \mathcal{H}$ (cf. the top node in Fig. 7.1). The sufficiency of this condition can be illustrated in multiple ways. A first example is given by constructing an open model class that violate $\mathcal{X}$-permutation invariance without invalidating the Bayes risk inequality for bistochastic attribute corruption. An analogous construction applies to bistochastic label corruption.

Example 7.38. *Let* $\mathcal{X} = \{a, b\}$ *and* $\mathcal{Y} = \{0, 1\}$. *We consider binary probabilistic predictors which are defined through points in* $\Delta_{\text{ca}}(\mathcal{Y})^{\mathcal{X}}$. *The model class* $\mathcal{H}_{(0,1]} := \{a \mapsto (p, 1 - p), b \mapsto (1 - p, p) \colon p \in (0, 1]\} \subseteq \Delta_{\text{ba}}(\mathcal{Y})^{\mathcal{X}}$ *is not* $\mathcal{X}$*-permutation invariant. However, for any bistochastic attribute corruption* $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$,

$$\mathbf{BR}_{\ell \circ \mathcal{H}_{(0,1]}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}_{(0,1]}}(\check{\tau}\phi),$$

for all $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$ *and all* ℓ*. The reasoning is simple,*

$$\begin{aligned}\mathbf{BR}_{\ell \circ \mathcal{H}_{(0,1]}}(\check{\tau}\phi) = \mathbf{BR}_{\hat{\tau}(\ell \circ \mathcal{H}_{(0,1]})}(\phi) &= \inf_{f \in \hat{\tau}(\ell \circ \mathcal{H}_{(0,1]})} \langle f, \phi \rangle \\ &= \inf_{f \in \ell \circ \mathcal{H}_{[0,1]}} \langle f, \phi \rangle \geq \inf_{f \in \ell \circ \mathcal{H}_{(0,1]}} \langle f, \phi \rangle = \mathbf{BR}_{\ell \circ \mathcal{H}_{(0,1]}}(\phi)\,.\end{aligned}$$

Even without invoking closure conditions, which are irrelevant for infimum operators appearing in the Bayes risk, we can show that $\mathcal{X}$-permutation invariance is not a necessary condition for the Bayes risk inequality under attribute corruption.

Example 7.39. *Let* $\mathcal{X} = \{a, b\}$ *and* $\mathcal{Y} = \{0, 1\}$*. Let* $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ *be the* 0-1*-loss function. We consider binary probabilistic predictors which are uniquely identified through points in* $[0,1]^{|\mathcal{X}|}$ *because of the finiteness of all sets involved. The model class*

$$\begin{aligned}\mathcal{M}(\mathcal{X}, \mathcal{Y}) \supset \mathcal{H} \coloneqq \{\, &(a \mapsto (0,1), b \mapsto (0.9, 0.1)), \\ &(a \mapsto (1,0), b \mapsto (0.25, 0.75)), \\ &(a \mapsto (0.2, 0.8), b \mapsto (0.15, 0.85)), \\ &(a \mapsto (0.6, 0.4), b \mapsto (0.7, 0.3))\,\},\end{aligned}$$

here expressed in its Markov transition form for convenience, is not $\mathcal{X}$*-permutation invariant. However, for any (including bistochastic) attribute corruption* $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$*,*

$$\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \leq \mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\tau}\phi),$$

for all $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$*, as* $\hat{\tau}(\ell \circ \mathcal{H}) = \ell \circ \mathcal{H}$*. This is true because the loss function only distinguishes between predictors using the threshold value* 0.5*. In other words, the class* $\mathcal{H}$ *is, through the lens of* ℓ*, as powerful as* $\mathcal{M}(\mathcal{X}, \mathcal{Y})$*.*

7.4 Antipolar Sets and Inequalities

We now proceed to develop a framework based on antipolar sets and functions, which expands what done until now for Bayes risk orderings. Leveraging these objects and their properties, we introduce an antipolar data processing inequality and establish some characterization results, both at the level of antipolar sets and at the level of the associated antipolar support functions.

These results are then used in two ways. First, we obtain a characterization of the Strong Bayes Risk Inequality in terms of an upper bound expressed via an antipolar support function. Second, we show that the Strong Bayes Risk Inequality coefficient itself admits an exact representation as the value of an antipolar support function evaluated at a certain function.

7.4.1 Antipolar Duality

Key objects for the following sections are antipolar sets and functions. They are the concave counterparts of the more common polar and gauge functions, widely studied in e.g. Rockafellar (1970); Hiriart-Urruty and Lemaréchal (2004); Chancelier and De Lara (2025). We prove here relevant results for our concave setting.

Definition 7.40 (Antipolar Set)**.** *Consider the dual pairing* $\langle \cdot, \cdot \rangle \colon \mathcal{B}_b(\mathcal{Z}) \times \mathrm{ba}(\mathcal{Z}) \to \mathbb{R}$ *for some Polish space* $\mathcal{Z}$*. Let* $\mathcal{A} \subseteq \mathcal{B}_b(\mathcal{Z})$ *and* $\mathcal{B} \subseteq \mathrm{ba}(\mathcal{Z})$*. The associated antipolar sets are defined as,*

$$\mathcal{A}^\diamond := \{\mu \in \mathrm{ba}(\mathcal{Z}) \,|\, \forall f \in \mathcal{A}, \langle f, \mu \rangle \geq 1\}$$
$$\mathcal{B}^\diamond := \{f \in \mathcal{B}_b(\mathcal{Z}) \,|\, \forall \mu \in \mathcal{B}, \langle f, \mu \rangle \geq 1\}\,.$$

Definition 7.41 (Co-star-shaped and Shady Set)**.** *A non-empty subset* $\mathcal{A} \subset \mathcal{B}_b(\mathcal{Z})$ *is said to be* shady *if and only if it is convex,* $0 \notin \mathcal{A}$ *and such that,*

$$[1, \infty)\, \mathcal{A} \subseteq \mathcal{A}\,.$$

The above property alone identifies co-star-shaped *sets. We can give an equivalent definition for sets* $\mathcal{B} \subset \mathrm{ba}(\mathcal{Z})$*.*

Proposition 7.42 (Properties of Antipolar Sets)**.** *Given* $\mathcal{A} \subseteq \mathcal{B}_b(\mathcal{Z})$*:*

(P1) $\mathcal{A}^\diamond = level_{\geq 1}(\rho_{\mathcal{A}}) := \{\mu \in \mathrm{ba}(\mathcal{Z}) \,|\, \rho_{\mathcal{A}}(\mu) \geq 1\}$*;*

(P2) $\mathcal{A}^\diamond = \overline{\mathrm{co}}\Big([1, \infty)\mathcal{A}\Big)^\diamond$*;*

(P3) The mapping $\mathcal{A} \mapsto \mathcal{A}^\diamond$ *takes closed shady sets into closed shady sets;*

(P4) $\mathcal{A}^{\diamond\diamond} = \overline{\mathrm{co}}\Big([1, \infty)\mathcal{A}\Big)$*;*

(P5) $\mathcal{G} \subseteq \mathcal{A} \Rightarrow \overline{\mathrm{co}}\Big([1, \infty)\mathcal{G}\Big) \subseteq \overline{\mathrm{co}}\Big([1, \infty)\mathcal{A}\Big) \Leftrightarrow \mathcal{A}^\diamond \subseteq \mathcal{G}^\diamond$*;*

(P6) $(\lambda\mathcal{A})^\diamond = \lambda\mathcal{A}^\diamond$ *for* $\lambda \neq 0$*,* $(\lambda\mathcal{A})^\diamond = \emptyset$ *for* $\lambda = 0$*;*

(P7) Given a collection of sets $\mathcal{C}$ *in* $\mathcal{B}_b(\mathcal{X} \times \mathcal{Y})$*, it holds that* $\big(\bigcup_{\mathcal{A} \in \mathcal{C}} \mathcal{A}\big)^\diamond = \bigcap_{\mathcal{A} \in \mathcal{C}} \mathcal{A}^\diamond$*.*

Analogous properties hold for $\mathcal{B} \subseteq \mathrm{ba}(\mathcal{Z})$*.*

Proof.

(P1) True by definition.

(P2) (Penot and Zalinescu, 2000, Lemma 4.2)

(P3) (Penot and Zalinescu, 2000, Lemma 4.2)

(P4) Follows from (P1).

(P5) The first $\Rightarrow$ follows immediately. For the second $\Rightarrow$: pick any $f' \in \mathcal{A}^\diamond$, by definition $\langle f, f' \rangle \geq 1$ for all $f \in \mathcal{A}$, hence in particular for all $f \in \mathcal{G}$, from which follows $f' \in \mathcal{A}^\diamond$. The direction $\Leftarrow$ follows from (P4).

(P6) By definition, we can write

$$\begin{aligned}(\lambda\mathcal{A})^\diamond &:= \{\mu \in \mathrm{ba}(\mathcal{Z}) \,|\, \langle f, \mu \rangle \geq 1 \ \forall f \in \lambda\mathcal{A}\} \\ &= \{\mu \in \mathrm{ba}(\mathcal{Z}) \,|\, \langle \lambda f', \mu \rangle \geq 1 \ \forall f' \in \mathcal{A}\} \\ &= \{\mu \in \mathrm{ba}(\mathcal{Z}) \,|\, \langle f', \lambda\mu \rangle \geq 1 \ \forall f' \in \mathcal{A}\} \\ &= \{\mu' \in \lambda\,\mathrm{ba}(\mathcal{Z}) \,|\, \langle f', \mu' \rangle \geq 1 \ \forall f \in \mathcal{A}\}.\end{aligned}$$

It follows that, for $\lambda \neq 0$, $(\lambda\mathcal{A})^\diamond = \lambda\mathcal{A}^\diamond$. For $\lambda = 0$, we have

$$(\lambda\mathcal{A})^\diamond := \{\mu \in \mathrm{ba}(\mathcal{Z}) \,|\, \langle 0, \mu \rangle \geq 1\} = \emptyset.$$

(P7) "$\subseteq$" Let $\mu \in \left(\bigcup_{\mathcal{A}\in\mathcal{C}} \mathcal{A}\right)^\diamond$. By definition, for every $f \in \bigcup_{\mathcal{A}\in\mathcal{C}} \mathcal{A}$ we have $\langle f, \mu \rangle \geq 1$. In particular, for any fixed $\mathcal{A} \in \mathcal{C}$, since $\mathcal{A} \subseteq \bigcup_{\mathcal{A}\in\mathcal{C}} \mathcal{A}$, it follows that $\langle f, \mu \rangle \geq 1$ for all $f \in \mathcal{A}$, i.e. $\mu \in \mathcal{A}^\diamond$. As this holds for every $\mathcal{A} \in \mathcal{C}$, we can conclude that $\mu \in \bigcap_{\mathcal{A}\in\mathcal{C}} \mathcal{A}^\diamond$.

"$\supseteq$" Let $\mu \in \bigcap_{\mathcal{A}\in\mathcal{C}} \mathcal{A}^\diamond$. Then for every $\mathcal{A} \in \mathcal{C}$ we have $\langle f, \mu \rangle \geq 1$ for all $f \in \mathcal{A}$. If $f \in \bigcup_{\mathcal{A}\in\mathcal{C}} \mathcal{A}$, then $f \in \hat{\mathcal{A}}$ for some $\hat{\mathcal{A}} \in \mathcal{C}$, and hence $\langle f, \mu \rangle \geq 1$. Therefore $\mu \in \left(\bigcup_{\mathcal{A}\in\mathcal{C}} \mathcal{A}\right)^\diamond$.

□

Definition 7.43 (Antipolar Support Function)**.** *Let $\mathcal{A}$ be a set in $\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})$, its associated* antipolar support function $\rho^\diamond\colon \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \to \overline{\mathbb{R}}_{\geq 0}$ *is defined as*

$$\rho^\diamond_{\mathcal{A}}(f) := \sup\{\lambda \geq 0 \mid \langle f, \mu \rangle \geq \lambda \rho_{\mathcal{A}}(\mu) \quad \forall \mu \in \mathrm{ba}(\mathcal{X}\times\mathcal{Y})\}\,.$$

Notice that, since both the bilinear form $\langle f, \mu \rangle$ and the support function are 1-homogeneous function, an equivalent definition of the antipolar support function only involves signed measures of norm one, i.e.,

$$\rho^\diamond_{\mathcal{A}}(f) := \sup\{\lambda \geq 0 \mid \langle f, \mu \rangle \geq \lambda \rho_{\mathcal{A}}(\mu) \ , \ \forall \mu \in \mathbb{B}_1 \cup \{0\}\}\,.$$

Definition 7.44 (Antigauge Function)**.** *Let $\mathcal{A}$ be a set in $\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})$, its associated* antigauge

function $\beta\colon \mathcal{B}_b(\mathcal{X}\times\mathcal{Y}) \to \overline{\mathbb{R}}_{\geq 0}$ *is defined as*

$$\beta_{\mathcal{A}}(f) \coloneqq \sup\{\lambda \geq 0 \mid f \in \lambda\mathcal{A}\}\,.$$

One can prove that this formulation of the antigauge function is upper semi-continuous when $0 \notin \overline{\mathcal{A}}$ (Penot and Zalinescu, 2000, Proposition 2.4 (c)), unlike other available definitions in the literature. It also holds that, if $0 \notin \overline{\mathcal{A}}$, the antigauge function does not take the value $+\infty$. For a detailed discussion and proof, see from Example 2.1 on in Penot and Zalinescu (2000).

Proposition 7.45 (Properties of Support Function and Antipolar)**.** *Given a closed shady set* $\mathcal{A} \subseteq \mathcal{B}_b(\mathcal{Z})$*:*

(P1) $\rho_{\mathcal{A}}(\mu) = \rho_{\mathcal{A}^{\circ\circ}}(\mu) = \rho_{[1,+\infty)\mathcal{A}}(\mu)$*;*

(P2) $\rho_{\mathcal{A}^\circ}(f) = \beta_{\mathcal{A}}(f)$ *and* $\rho_{\mathcal{A}}(f) = \beta_{\mathcal{A}^\circ}(f)$*;*

(P3) $\rho^\diamond_{\mathcal{A}}(f) = \rho_{\mathcal{A}^\circ}(f)$*;*

(P4) For a general set $B \subseteq \mathcal{B}_b(\mathcal{Z})$*,* $\rho^\diamond_B(f) = \rho^\diamond_B(f)$*.*

Proof.

1. Notice that the support function is invariant to convex closure and that $\mathcal{A}$ is closed shady, hence the bipolar theorem holds (Proposition 7.42 (P4)) and proves the statement.

2. We define the indicator function, which takes values

$$\iota_{\mathcal{A}}(x) \coloneqq \begin{cases} 0 & \text{if } x \in \mathcal{A}, \\ +\infty & \text{otherwise.} \end{cases} \tag{7.7}$$

   We observe it has the following connection with the support function,

$$-\rho_{\mathcal{A}^\circ}(f) \overset{(L.3.22)}{=} \sigma_{-\mathcal{A}^\circ}(f) = \sup_{\mu\in -\mathcal{A}^\circ} \langle f,\mu\rangle = \sup_{\mu\in \mathrm{ba}(\mathcal{Z})} \langle f,\mu\rangle - \iota_{-\mathcal{A}^\circ}(\mu) =: \iota^*_{-\mathcal{A}^\circ}(f)\,.$$

   The right hand side is known as the Legendre-Fenchel conjugate of a function,

$$F^*(f) \coloneqq \sup_{\mu\in \mathrm{ba}(\mathcal{Z})} \langle f,\mu\rangle - F(\mu),$$

   computed for $F = \iota_A$. Hence, we can apply Penot and Zalinescu (2000, Lemma 4.1)

and get that

$$[(-\beta_{\mathcal{A}}(f))^*]^* = \iota^*_{-\mathcal{A}^\diamond}(f) = -\rho_{\mathcal{A}^\diamond}(f)\,.$$

As a consequence of the bipolar theorem (Proposition 7.42 (P4)) the closed and shady set $\mathcal{B} \coloneqq \mathcal{A}^\diamond$ is s.t. $B^\diamond = \mathcal{A}$. Thus,

$$[(-\beta_{\mathcal{B}^\diamond}(\mu))^*]^* = \iota^*_{-\mathcal{B}}(\mu) = -\rho_{\mathcal{B}}(\mu)\,.$$

To conclude the argument, we notice that under the assumption of $\mathcal{A}$ being closed and shady, its β is a concave, upper semi-continuous function, therefore the biconjugate theorem (Zălinescu, 2002, Theorem 2.3.3) holds for $-\beta_{\mathcal{A}}$ and $-\beta_{\mathcal{A}^\diamond}$.

3. Consider the definition of antipolar of the support function:

$$\begin{aligned}
\rho^\diamond_{\mathcal{A}}(f) &\coloneqq \sup\{\lambda \geq 0 \mid \langle f, \mu\rangle \geq \lambda \rho_{\mathcal{A}}(\mu) \;\; \forall \mu \in \mathrm{ba}(\mathcal{Z})\} \\
&= \sup\{\lambda \geq 0 \mid \inf_{\mu} \Big(\langle f, \mu\rangle - \lambda \rho_{\mathcal{A}}(\mu)\Big) \geq 0\} \\
&= \sup\{\lambda \geq 0 \mid -\sup_{\mu} \Big(\langle -f, \mu\rangle + \lambda \rho_{\mathcal{A}}(\mu)\Big) \geq 0\} \\
&\overset{(P2)}{=} \sup\{\lambda \geq 0 \mid -\sup_{\mu} \Big(\langle -f, \mu\rangle + \lambda \beta_{\mathcal{A}^\diamond}(\mu)\Big) \geq 0\} \\
&\overset{(\star)}{=} \sup\{\lambda \geq 0 \mid -\iota_{-(\lambda\mathcal{A})^{\diamond\diamond}}(-f) \geq 0\} \overset{(7.7)}{=} \sup\{\lambda \geq 0 \mid f \in (\lambda\mathcal{A})^{\diamond\diamond}\} \\
&\overset{(P6)}{=} \sup\{\lambda \geq 0 \mid f \in \lambda\mathcal{A}^{\diamond\diamond}\} = \sup\{\lambda \geq 0 \mid f \in \lambda\mathcal{A}\} = \beta_{\mathcal{A}}(f),
\end{aligned}$$

where $(\star)$ is Lemma 4.1 in Penot and Zalinescu (2000). Notice that we are allowed to use P6 here, since the case $\lambda = 0$ cannot occur in the supremum. That is because, it would cause the condition to become "$f \in (0\mathcal{A})^{\diamond\diamond} = \emptyset$", which cannot be met. Using P2 again, we obtain the thesis.

4. As a consequence of Lemma 3.22, we get

$$\begin{aligned}
\rho^\diamond_{\mathcal{A}}(f) &\coloneqq \sup\{\lambda \geq 0 \mid \langle f, \mu\rangle \geq \lambda \rho_{\mathcal{A}}(\mu) \;\; \forall \mu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y})\} \\
&= \sup\{\lambda \geq 0 \mid \langle f, \mu\rangle \geq \lambda \rho_{\overline{\mathrm{co}}\mathcal{A}}(\mu) \;\; \forall \mu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y})\} = \rho^\diamond_{\overline{\mathrm{co}}\mathcal{A}}(f)\,.
\end{aligned}$$

□

7.4.2 Antipolar Bayes Risk Inequality

Inspired by the analysis of the antipolars of superprediction set, loss and conditional Bayes risk carried out in Williamson and Cranko (2023), we aim to study the antipolar version of the Bayes Risk Inequality. First, we notice that a characterization result of the Bayes Risk Inequality through subsethood holds also in the antipolar setting, for general shady sets.

Lemma 7.46. *Given two shady sets $\mathcal{F}, \mathcal{G} \subseteq \mathcal{B}_b(\mathcal{Z})$ and a dual pairing $\langle \cdot, \cdot \rangle \colon \mathcal{B}_b(\mathcal{Z}) \times \mathrm{ba}(\mathcal{Z}) \to \mathbb{R}$, the following statements are equivalent:*

1. $\rho_{\mathcal{G}}(\mu) \leq \rho_{\mathcal{F}}(\mu) \quad \forall \mu \in \mathrm{ba}(\mathcal{Z})$;
2. $\rho_{\mathcal{F}^\diamond}(f) \leq \rho_{\mathcal{G}^\diamond}(f) \quad \forall f \in \mathcal{B}_b(\mathcal{Z})$;
3. $\rho^\diamond_{\mathcal{F}}(f) \leq \rho^\diamond_{\mathcal{G}}(f) \quad \forall f \in \mathcal{B}_b(\mathcal{Z})$;
4. $\mathcal{G} \subseteq \mathcal{F}$;
5. $\mathcal{F}^\diamond \subseteq \mathcal{G}^\diamond$.

Proof. We know already that 1 $\Leftrightarrow$ 4 from the proof of Lemma 7.11 and that 4 $\Leftrightarrow$ 5 because of P5 of Proposition 7.42. Because of the bijection between support functions and closed convex sets (Lemma 7.10), we know that also 2 $\Leftrightarrow$ 5 holds. The statements in 2 and 3 are equivalent because of Proposition 7.45 (P3). This concludes the proof. □

To be sure that the result above can be applied to the set $\overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H})$, we need to prove the following lemmas.

Lemma 7.47. *Consider a set $\mathcal{F} \subseteq \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$. Then, it holds that*

$$\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} \oplus \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0} \subseteq \overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}\,.$$

Proof. We show that $\overline{\mathrm{co}}\,\mathrm{spr}\left(\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} \oplus \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}\right) = \overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}$. We use Lemma 7.12, hence we have to show that,

$$\rho_{\overline{\mathrm{co}}\,\mathrm{spr}\left(\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} \oplus \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0}\right)}(\phi) = \rho_{\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}}(\phi), \ \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})\,.$$

For this, note that, for all $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$,

$$\begin{aligned}\rho_{\overline{\mathrm{co}}\,\mathrm{spr}(\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} \oplus \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0})}(\phi) &\overset{(L3.19)}{=} \rho_{\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} \oplus \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0} \oplus \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0}}(\phi) \\ &\overset{(P5)}{=} \rho_{\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}}(\phi) + \rho_{\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0}}(\phi) + \rho_{\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0}}(\phi) = \rho_{\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}}(\phi)\,,\end{aligned}$$

because $\rho_{\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0}}(\phi) = 0$ as $0 \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$. □

Lemma 7.48. *Let $\mathcal{F} \subseteq \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$. If $0 \notin \overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}$, the set $\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}$ is shady.*

Proof. Then, we use the logic given in (Williamson and Cranko, 2023, Lemma 18). For every $b \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$ and all $\alpha \in [0, \infty)$, $\alpha b \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$, hence,

$$\overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F} \oplus \alpha b \subseteq \overline{\mathrm{co}}\,\mathrm{spr}\,\mathcal{F}\,,$$

by Lemma 7.47. Since $\overline{\text{co}}\,\text{spr}\,\mathcal{F} \subseteq \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0}$, we can assume that $b \in \overline{\text{co}}\,\text{spr}\,\mathcal{F}$, hence,

$$\overline{\text{co}}\,\text{spr}\,\mathcal{F} \oplus \alpha b = (1+\alpha)\,\overline{\text{co}}\,\text{spr}\,\mathcal{F} \subseteq \overline{\text{co}}\,\text{spr}\,\mathcal{F}\,,$$

which concludes the proof. □

Proposition 7.49 (Characterization of Antipolar Generalized Data Processing Inequality)**.** *Let $\ell\colon \Delta_{\text{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\geq 0}$ be a loss function and $\mathcal{H}\subseteq\mathcal{M}(\mathcal{X},\mathcal{Y})$ a model class. Let $\langle\cdot,\cdot\rangle$ be the canonical dual pairing relating $\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})$ and $\text{ba}(\mathcal{X}\times\mathcal{Y})$. Consider a Markov operator $\hat{\kappa}$ associated to a kernel $\kappa\in\mathcal{M}(\mathcal{X}\times\mathcal{Y},\mathcal{X}\times\mathcal{Y})$. Then, assuming that $0\notin\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})$, the condition*

$$\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})^{\diamond}\subseteq\overline{\text{co}}\,\text{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))^{\diamond}$$

is equivalent to the Bayes Risk Inequality and to its antipolar counterpart

$$\rho^{\diamond}_{\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})}(f)\leq\rho^{\diamond}_{\overline{\text{co}}\,\text{spr}\hat{\kappa}(\ell\circ\mathcal{H})}(f)\quad\forall f\in\mathcal{B}_b(\mathcal{Z})\,.$$

as well as to all the statements in Lemma 7.46.

Proof. Consider the antipolar of the convex closure of the constrained superprediction set,

$$\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})^{\diamond}=\{\phi\in\text{ba}(\mathcal{X}\times\mathcal{Y})\,|\,\forall f\in\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})\,,\,\langle f,\phi\rangle\geq 1\}\,.$$

Using Lemma 7.46,

$$\overline{\text{co}}\,\text{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))\subseteq\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})\quad\Leftrightarrow\quad\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})^{\diamond}\subseteq\overline{\text{co}}\,\text{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))^{\diamond}\,.$$

Given that Lemma 7.48 holds when $0\notin\ell\circ\mathcal{H}$, we have that the corrupted superprediction set is shady. Thus, there is a bijection between the antipolar functions and antipolar superprediction sets. Applying Lemma 7.11 and Proposition 7.45 (P3) to the RHS, we get

$$\rho^{\diamond}_{\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})}(f)\leq\rho^{\diamond}_{\overline{\text{co}}\,\text{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))}(f)\quad\forall f\in\mathcal{B}_b(\mathcal{Z})\,.$$

Then rest of the thesis is a direct consequence of Lemma 7.46 and Theorem 7.13. □

7.4.3 Strong Bayes Risk Inequality

We now reconnect with a more learning theoretic perspective and relate all the antipolar objects to Risk-related ones.

Definition 7.50 (Minimal Relative Risk)**.** *Given a set $\ell\circ\mathcal{H}\subseteq\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{\geq 0}$ such that $\overline{\text{co}}\,\text{spr}(\ell\circ\mathcal{H})$ is closed and shady, and a function $f\in\mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{>0}$, we define the quantity*

$$\mathbf{MR}_{\ell\circ\mathcal{H}}(f):=\inf_{\phi\in\Delta_{\text{ba}}(\mathcal{X}\times\mathcal{Y})}\frac{\langle f,\phi\rangle}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)}\in(0,+\infty).$$

as the minimal relative risk *of* f.

The quantity above is well defined: When $\overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H})$ is a closed shady set, we know that $0 \notin \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H})$. Hence, for every $\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$ and $f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{>0}$ we have $\langle f, \phi \rangle > 0$ and $\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) > 0$. In addition, as f is a bounded measurable function, and $\phi \in \Delta_{\mathrm{ba}}$, the numerator will not diverge.

Proposition 7.51. *Given a set* $\ell \circ \mathcal{H} \subseteq \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$, *it holds that*

$$\mathbf{MR}_{\ell \circ \mathcal{H}}(f) \geq \rho^{\diamond}_{\mathrm{spr}(\ell \circ \mathcal{H})}(f) \quad \forall f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{>0}.$$

Proof. By definition of Minimal relative risk and support function, we can write

$$\begin{aligned}\mathbf{MR}_{\ell \circ \mathcal{H}}(f) &\coloneqq \inf_{\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})} \frac{\langle f, \phi \rangle}{\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi)} \\ &= \inf_{\phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})} \langle f, \frac{\phi}{\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi)} \rangle =: \rho_{R_{\Delta_{\mathrm{ba}}}}(f) \quad \forall f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{>0},\end{aligned}$$

where $R_{\Delta_{\mathrm{ba}}} \coloneqq \bigcup_{\phi \in \Delta_{\mathrm{ba}}} \left\{ \frac{\phi}{\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi)} \right\} \subset \mathrm{ba}(\mathcal{X} \times \mathcal{Y})_{>0}$. Let us now consider the antipolar set of $R_{\Delta_{\mathrm{ba}}}$,

$$\begin{aligned}R^{\diamond}_{\Delta_{\mathrm{ba}}} &= \left(\bigcup_{\phi \in \Delta_{\mathrm{ba}}} \left\{ \frac{\phi}{\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi)} \right\} \right)^{\diamond} \overset{P.7.42}{=} \bigcap_{\phi \in \Delta_{\mathrm{ba}}} \left\{ \frac{\phi}{\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi)} \right\}^{\diamond} \\ &= \bigcap_{\phi \in \Delta_{\mathrm{ba}}} \left\{ f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}) \mid \langle f, \frac{\phi}{\mathbf{BR}_{\ell \circ \mathcal{H}}(\phi)} \rangle \geq 1 \right\} \\ &= \bigcap_{\phi \in \Delta_{\mathrm{ba}}} \left\{ f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}) \mid \langle f, \phi \rangle \geq \mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \right\} \\ &= \bigcap_{\phi \in \Delta_{\mathrm{ba}}} \left\{ f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y}) \mid \langle f, \phi \rangle \geq \rho_{\mathrm{spr}(\ell \circ \mathcal{H})}(\phi) \right\} \supseteq \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H}).\end{aligned}$$

Hence, by properties of the support function (Lemma 7.11),

$$\rho_{\overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H})^{\diamond}}(f) \leq \rho_{R^{\diamond\diamond}_{\Delta_{\mathrm{ba}}}}(f) = \rho_{\overline{\mathrm{co}}([1,+\infty) R_{\Delta_{\mathrm{ba}}})}(f) = \rho_{[1,+\infty) R_{\Delta_{\mathrm{ba}}}}(f).$$

As $R_{\Delta_{\mathrm{ba}}} \subseteq [1, +\infty) R_{\Delta_{\mathrm{ba}}}$, we have the thesis. □

Definition 7.52 (Strong Bayes Risk Inequality)**.** *Let* $\ell \colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ *be a loss function and* $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Y})$ *a model class. Let* $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ *be a Markov kernel. We refer as the Strong Bayes Risk Inequality to the following inequality,*

$$\mathbf{BR}_{\ell \circ \mathcal{H}}(\check{\kappa}\phi) \geq \alpha_{\ell \circ \mathcal{H}}(\kappa)\, \mathbf{BR}_{\ell \circ \mathcal{H}}(\phi) \quad \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y}),$$

where $\alpha_{\ell \circ \mathcal{H}}(\kappa)$ *is some suitable* inflation coefficient.

We call this equivalence Strong Bayes Risk Inequality, aligning it with its classical information

theoretic counterpart, called Strong Data Processing Inequality for the f-divergence,

$$D_f(\check{\kappa}\phi \,\|\, \check{\kappa}\psi) \le \eta_f(\kappa)\, D_f(\phi \,\|\, \psi)\,,$$
$$\eta_f(\kappa) \coloneqq \sup_{\phi,\psi\in\mathcal{P}} \frac{D_f(\check{\kappa}\phi\|\check{\kappa}\psi)}{D_f(\phi\|\psi)}\,,$$

where $\kappa \in \mathcal{M}(\mathcal{X},\mathcal{X})$, $\mathcal{P}$ is the set of probabilities for which D_f takes positive and finite values and $\eta_f(\kappa)$ is a called a *contraction coefficient*.[2] To be compatible with our setting while analogous to the above, the coefficient α must take a different form than η. In the following, we give a suitable definition and show its connection to the antipolar support function.

Definition 7.53 (Inflation coefficient)**.** *Let $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\ge 0}$ be a loss function and $\mathcal{H}\subseteq\mathcal{M}(\mathcal{X},\mathcal{Y})$ a model class inducing a shady $\overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})$. Consider a Markov kernel $\kappa\in\mathcal{M}(\mathcal{X}\times\mathcal{Y},\mathcal{X}\times\mathcal{Y})$. Then, we define*

$$\alpha_{\ell\circ\mathcal{H}}(\kappa) \coloneqq \inf_{\phi\in\Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})} \frac{\mathbf{BR}_{\ell\circ\mathcal{H}}(\check{\kappa}\phi)}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)}\,.$$

Lemma 7.54 (Infimum over Minimal Relative Risk is Inflation Coefficient)**.** *Let $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\ge 0}$ be a loss function and $\mathcal{H}\subseteq\mathcal{M}(\mathcal{X},\mathcal{Y})$ a model class s.t. $0\notin\overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})$. Furthermore, suppose that $\kappa\in\mathcal{M}(\mathcal{X}\times\mathcal{Y},\mathcal{X}\times\mathcal{Y})$ s.t. $0\notin\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))$. Then,*

$$\alpha_{\ell\circ\mathcal{H}}(\kappa) = \inf_{f\in\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))} \mathbf{MR}_{\ell\circ\mathcal{H}}(f)\,.$$

Proof. By swapping infima we obtain,

$$\begin{aligned}
\inf_{f\in\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))} \mathbf{MR}_{\ell\circ\mathcal{H}}(f) &\overset{D.7.50}{=} \inf_{f\in\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))} \inf_{\phi\in\Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})} \frac{\mathbb{E}_\phi(f)}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)} \\
&= \inf_{\phi\in\Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})} \frac{\inf_{f\in\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))}\mathbb{E}_\phi(f)}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)} \\
&= \inf_{\phi\in\Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})} \frac{\rho_{\hat{\kappa}(\ell\circ\mathcal{H}))}(\phi)}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)} \overset{L.5.1}{=} \alpha_{\ell\circ\mathcal{H}}(\kappa)\,.
\end{aligned}$$

□

Proposition 7.55. *Let $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y}\to\mathbb{R}_{\ge 0}$ be a loss function and $\mathcal{H}\subseteq\mathcal{M}(\mathcal{X},\mathcal{Y})$ a model class inducing a generalized superprediction set $\overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})$ s.t. $0\notin\overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})$. Consider a Markov kernel $\kappa\in\mathcal{M}(\mathcal{X}\times\mathcal{Y},\mathcal{X}\times\mathcal{Y})$ s.t. $0\notin\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))$. Let $\alpha_{\ell\circ\mathcal{H}}(\kappa)$ be its associated coefficient, defined as in Def. 7.53. Then, the following statements are equivalent:*

1. *the Bayes Risk Inequality holds;*

[2] More details at (Polyanskiy and Wu, 2025, Chapter 33.2).

2. *the Strong Bayes Risk Inequality holds with inflation coefficient* $\alpha_{\ell\circ\mathcal{H}}(\kappa) \geq 1$;
3. *it holds that* $\mathbf{MR}_{\ell\circ\mathcal{H}}(f) \geq 1\,,\ \ \forall f \in \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))$.

Proof.

(1) $\Rightarrow$ (2): This is an immediate consequence of the Bayes risk inequality.

(2) $\Rightarrow$ (3): Lemma 7.54.

(3) $\Rightarrow$ (1): The remaining implication is given by,

$$
\begin{aligned}
& \mathbf{MR}_{\ell\circ\mathcal{H}}(f) \geq 1\,,\ \ \forall f \in \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H})) \\
\Leftrightarrow\quad & \inf_{\phi\in\Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})} \frac{\mathbb{E}_\phi(f)}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)} \geq 1\,,\ \ \forall f \in \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H})) \\
\Rightarrow\quad & \frac{\mathbb{E}_\phi(f)}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)} \geq 1\,,\ \ \forall f \in \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H})), \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y}) \\
\Rightarrow\quad & \mathbb{E}_\phi(f) \geq \mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)\,,\ \ \forall f \in \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H})), \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y}) \\
\Rightarrow\quad & \inf_{f\in\overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))} \mathbb{E}_\phi(f) \geq \mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)\,,\ \ \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y}) \\
\overset{L5.1}{\Rightarrow}\quad & \mathbf{BR}_{\ell\circ\mathcal{H}}(\check{\kappa}\phi) \geq \mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)\,,\ \ \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})\,.
\end{aligned}
$$

□

Corollary 7.56 (Sufficient Condition for Strong Bayes Risk Inequality)**.** *Let* $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y})\times\mathcal{Y} \to \mathbb{R}_{\geq 0}$ *be a loss function and* $\mathcal{H} \subseteq \mathcal{M}(\mathcal{X},\mathcal{Y})$ *a model class inducing a generalized superprediction set* $\overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})$ *s.t.* $0 \notin \overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})$*. Consider a Markov kernel* $\kappa \in \mathcal{M}(\mathcal{X}\times\mathcal{Y}, \mathcal{X}\times\mathcal{Y})$ *s.t.* $0 \notin \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H}))$*. Then,*

$$
\begin{gathered}
\rho^{\diamond}_{\overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})}(f) \geq 1 \quad \forall f \in \mathcal{B}_b(\mathcal{X}\times\mathcal{Y})_{>0} \\
\Downarrow \\
\alpha_{\ell\circ\mathcal{H}}(\kappa) \geq 1 \\
\textit{and} \\
\mathbf{BR}_{\ell\circ\mathcal{H}}(\check{\kappa}\phi) \geq \alpha_{\ell\circ\mathcal{H}}(\kappa)\ \mathbf{BR}_{\ell\circ\mathcal{H}}(\phi) \quad \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y}).
\end{gathered}
$$

Proof. The result is given by

$$
\begin{aligned}
& \rho^{\diamond}_{\overline{\mathrm{co}}\,\mathrm{spr}(\ell\circ\mathcal{H})}(f) \geq 1\,,\ \ \forall f \in \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H})) \\
\overset{P7.51}{\Rightarrow}\quad & \inf_{\phi\in\Delta_{\mathrm{ba}}(\mathcal{X}\times\mathcal{Y})} \frac{\mathbb{E}_\phi(f)}{\mathbf{BR}_{\ell\circ\mathcal{H}}(\phi)} \geq 1\,,\ \ \forall f \in \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell\circ\mathcal{H})).
\end{aligned}
$$

Then, we apply Proposition 7.55. □

We conclude this section by remarking that, even if we have found new relationships between the SDPI coefficients and quantities related to the learning problem, i.e., minimal relative risk and antipolar support function, we are still far from proposing a formula for computing the coefficient itself. This problem is known to be complex to tackle. Contraction coefficients for f-divergence (see formula and short discussion under Def. 7.52) quantify how much an information measure shrinks under the action of a Markov kernel κ. As we have already seen, the SDPI constant η_f for a given divergence D_f is defined as the smallest value such that

$$D_f(\check{\kappa}\phi \,\|\, \check{\kappa}\psi) \leq \eta_f(\kappa)\, D_f(\phi \,\|\, \psi).$$

These coefficients depend intricately on both the choice of f-divergence and the structure of the kernel κ, and finding closed form expressions for them can be rather complex. Classical cases include (Makur and Zheng, 2020): the χ^2-divergence, where η_{χ^2} is proven to be equal to the squared maximal correlation, and can be used to bound coefficients for other choices of f; the Dobrushin coefficient for the total variation ($f = \frac{1}{2}|x-1|$); the Kullback–Leibler divergence coefficient, which is typically harder to characterize but admits closed-form expressions for certain families of distributions or kernels (see also Lee and Makur, 2024).

In general, determining contraction coefficients remains a technically demanding problem, with progress often limited to specific divergences or channel classes. This seems to be the case also for our inflation coefficients. Nevertheless, we contributed a new class of objects, namely the constrained version of the strong data processing inequality for Bayes risk and its associated coefficient, and a set of preliminary results relating them both to geometrical and learning theory quantities. We hope these can provide a new avenue for future research at the intersection of information theory and machine learning theory.

7.5 New Data Processing Inequalities and Partial Validity Results

In this chapter, the constrained superprediction set was used to characterize the generalized data processing inequality for Bayes risk in the constrained setting. This framework provides a systematic method to compare different choices of loss and model class through Bayesian orderings, which complements the existing literature on comparing experiments or joint probabilities. The table summarizes the equivalent formulations of the generalized data processing inequality we obtained, written in its most general setting which does not necessarily include Markov kernels. These equivalences clarify the geometric and functional structure underlying the inequality.

When specialized to the case of Markov kernels, both characterizing and sufficient conditions for the Bayesian orderings were derived. These conditions are motivated by the way kernels act on functions associated with the superprediction set, revealing how transformations

Objects	Formula	Reference
Set	$\mathcal{G} \subseteq \mathcal{F}$	-
Support Function	$\rho_{\mathcal{G}}(\mu) \geq \rho_{\mathcal{F}}(\mu)$ $\forall \mu \in \mathrm{ba}(\mathcal{X} \times \mathcal{Y})$	Lemma 7.11
Bayes Risk	$\mathbf{BR}_{\mathcal{G}}(\phi) \geq \mathbf{BR}_{\mathcal{F}}(\phi)$ $\forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$	Corollary 7.14
Antipolar Set	$\mathcal{G}^{\diamond} \supseteq \mathcal{F}^{\diamond}$	Lemma 7.46
Antipolar Support Function	$\rho^{\diamond}_{\mathcal{G}}(f) \leq \rho^{\diamond}_{\mathcal{F}}(f)$ $\forall f \in \mathcal{B}_b(\mathcal{X} \times \mathcal{Y})$	Proposition 7.49
Bayes Risk	$\mathbf{BR}_{\mathcal{G}}(\phi) = \mathbf{BR}_{\mathcal{F}}(\check{\kappa}\phi) \geq \alpha_{\mathcal{F}}(\kappa)\, \mathbf{BR}_{\mathcal{F}}(\phi)$ $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})\,,\ \forall \phi \in \Delta_{\mathrm{ba}}(\mathcal{X} \times \mathcal{Y})$	Proposition 7.55
Inflation Coefficient	$\alpha_{\mathcal{F}}(\kappa) \geq 1$	Proposition 7.55
Minimal Relative Risk	$\mathbf{MR}_{\mathcal{F}}(f) \geq 1$ $\forall f \in \mathcal{G}$	Proposition 7.55

Table 7.1: Equivalent statements proved in the manuscript, related to the Generalized Data Processing Inequality. The sets $\mathcal{F}, \mathcal{G}$ are assumed to be shady subsets of $\mathcal{B}_b(\mathcal{X} \times \mathcal{Y})_{\geq 0}$ and can be chosen such that $\mathcal{F} = \overline{\mathrm{co}}\,\mathrm{spr}(\ell \circ \mathcal{H})$ and $\mathcal{G} = \overline{\mathrm{co}}\,\mathrm{spr}(\hat{\kappa}(\ell \circ \mathcal{H}))$.

of experiments translate into transformations of Bayes risk. A fundamental distinction between label noise and attribute noise emerges also in the sufficient conditions derived for the generalized data processing inequality. This difference may explain the absence of a data processing result or Blackwell-Sherman-Stein-like theorem for label noise, even in the unconstrained case. The role of the loss function is again central in this discrepancy.

In the final section, the relationship between the strong data processing inequality (SDPI) for Bayes risk and the antipolar of the support function was examined. Both characterizing and sufficient conditions for the SDPI were established. In particular, the inflation coefficient can be simply related to the minimal relative risk. This yields an equivalent formulation of the SDPI in terms of the minimal relative risk. Moreover, the SDPI—and consequently the generalized data processing inequality—holds whenever the antipolar of the superprediction set exceeds one.

Chapter 8

Discussion and Future Work

This thesis advanced the understanding of how learning problems are formally related to one another by adopting a twofold perspective: establishing equivalences and deriving inequalities. Together with recent work by Mémoli et al. (2025), this introduces a substantial amount of structure into the space of learning problems, addressing a gap that had previously persisted in the literature. We summarize below the main insights gained and remaining open questions, grouped by theme.

A Taxonomy of Markovian Corruption In § 4 we introduced a taxonomy of Markovian corruptions, studied its structural properties, and analyzed its advantages and limitations. Its main strength lies in its exhaustiveness: it provides a full representation of all possible pairs of probability distributions (one clean, one corrupted) as a corruption mechanism. In addition, its formulation is sufficiently general to encompass many existing corruption models in the literature (as shown in Tab. 4.2 and later in Appendix A) and provides a solid theoretical foundation for systematic analysis of corruption effects and mitigations, as shown in § 5.1 and § 5.2 (summary in Tab. 5.1). However, it does not align directly with many existing taxonomies in the literature (Appendix B) and does not subsume all possible corruption *representations* (§ 4.2.2 and § 4.2.3). Moreover, we did not investigate whether our notion of hierarchical complexity corresponds to a form of learning complexity, i.e., whether higher degree of dependence of the corruption on the data leads to measurable degradation in performance at the level of Bayes risk. Interestingly, recent results on multi-class problems (Oyen et al., 2022) indicate that simple label corruption is significantly easier to mitigate than dependent label noise. This suggests that our structural notion of hierarchical complexity (based on dependence) may have operational consequences, but this remains to be formally established.

On the Use of the Taxonomy The taxonomy as-is should not be understood as a tool directly revealing new relationships between different types of corruption models, but as a framework. Such a kind of analysis can only be done by choosing two models of interest, manipulating them, and carefully comparing them, as we for instance do in the

sections discussing multi-step corruptions, selection bias, and mutually contaminated distributions, or in (Oyen et al., 2022) as explained above. Our work provides practitioners with a schematic representation of what forms a (one-step) corruption model can take, and therefore is an a priori step that enables the comparison of models. Consider a specific example: a corruption model acting on labels, only dependent on labels, and acting on a probability distribution by transforming it into another probability distribution. This corruption *must have* a representation as a one-step Markovian corruption, as ensured by our Proposition 4.9, since it cannot be non-factorizable—the type (image and domain being $\mathcal{X} \times \mathcal{Y}$) would not match this label noise setting. Note that the resulting Markov kernel representation may differ from the "natural" one with which the label corruption was originally defined. However, its existence allows models to be made directly comparable as Markov kernels. Here comparable can be understood both as qualitative (e.g., my corruption is dependent on $\mathcal{X}$, so its consequences are more intricate than the simple noise ones) or quantitative (e.g., comparing the values of the matrix representation of the kernel for two corruptions of the same type, comparing the error rates, comparing some other relevant metric). In Part II, we have only focused on a qualitative comparison, and did not explore if this has any quantitative consequence.

What is Non-Markovian? What is Multi-Step? In § 4.2 we discussed examples that fall outside the Markovian framework, as well as possible approaches and conceptual obstacles in extending the taxonomy to such cases. Additional examples appear in Mémoli et al. (2025). Future work could investigate extensions based on imprecise probability or more general theories of corruption that do not rely on a fully specified probabilistic structure. In addition, a systematic treatment of multi-step or non-Markovian corruptions remains open.

Data Processing Equalities In § 5.1 we extended the framework of Williamson and Cranko (2024) beyond simple corruptions and established equality results connecting the Bayes risks of clean and corrupted supervised learning problems for all corruptions in our taxonomy. These equalities demonstrate an equivalence between two learning problems: one corrupted in a *Markovian* manner, and another described through a *general corruption* that modifies the model class and loss function via an appropriate Markov kernel.

A central insight is that for corruptions acting solely on $\mathcal{Y}$, only the loss function is affected, while the model class remains unchanged. In contrast, for corruptions that also involve $\mathcal{X}$, both the loss function and the model class are altered through the corresponding factorized corruption kernel.

General Loss Correction Formulas Building on the Bayes risk equalities, in § 5.2 we derived corruption-corrected loss functions for all corruption types in our framework. This required extending the classical notion of accurate learning under corruption, which is no longer adequate once corruptions go beyond labels and affect attributes. Within this generalized formulation, the hierarchical structure of the taxonomy induces corresponding

changes in the optimization problem and determines how loss corrections can be computed in abstract terms. The analysis shows that increasing corruption complexity (in terms of the hierarchy in Fig. 4.2) leads to structurally more demanding correction schemes. In particular, classical loss correction is insufficient to fully mitigate corruptions involving attributes, revealing a further fundamental distinction between label and attribute corruption. Future work can investigate the consequence of such formulas in practical settings, and how they compare to existing and comparable loss correction techniques.

Characterization of Bayes Risk Orderings The results in § 7 employed the constrained superprediction set to characterize the generalized data processing inequality for Bayes risk in the constrained setting. The resulting framework provides a systematic way to compare losses and model classes via Bayesian orderings, extending classical comparisons of experiments or joint distributions. Tab. 7.1 summarizes the established equivalent formulations, highlighting the geometric and functional structure underlying the inequality. When restricted to Markovian corruptions acting on the learning problems, both characterizing and sufficient conditions for the Bayesian ordering were derived, based on how kernels act on functions generating the superprediction set. The analysis reveals a structural distinction between label noise and attribute noise, which may account for the absence of a Blackwell–Sherman–Stein-type result for label noise, even in the unconstrained case. The loss function plays a decisive role in this discrepancy. Finally, the strong data processing inequality (SDPI) for Bayes risk was linked to the antipolar of the support function. Characterizing and sufficient conditions were obtained also for this case, and the inflation coefficient was shown to coincide with the minimal relative risk. This yields an equivalent formulation of the SDPI in terms of minimal relative risk and implies that the SDPI (and hence the generalized data processing inequality) holds whenever the antipolar of the superprediction set exceeds one.

Can the Loss and Model Class be Treated as Distinct Entities? The reader may have observed that many of our results in fact characterize the set $\ell \circ \mathcal{H}$, and that many definitions and propositions are directly stated for a generic set $\mathcal{F}$ of positive and bounded measurable functions. This is not unseen: measures of model expressivity are in fact dependent on the loss chosen, e.g., the Rademacher complexity we briefly introduced in Eq. (7.2). This suggests that, although loss and model class are semantically different constraints on a learning problem (the former "softer" than the latter), they can be formally coupled.

Information Theory for Constrained Problems In § 7 we investigated the data processing inequalities, both in their standard (§ 7.2) and strong (§ 7.4) formulations, and provided many results for characterizing them, as well as proving sufficient conditions for the inequalities to hold. We have done so focusing primarily on generalized entropy rather than statistical information. The field of information theory, however, offers a broader range of tools that may be relevant to constrained learning problems if appropriately translated. In order to do that, we need alternative notions of information tailored to such a setting.

Fortunately, the framework of Williamson and Cranko (2024) already provides notions of unconstrained and constrained information that generalize f-divergences and are linked to generalized entropy as defined here. An immediate continuation of our work would be to exploit the characterization via superprediction sets within this context and to examine how classical information-theoretic results adapt to constrained model classes.

A Data Processing Inequality that Better Serves Machine Learning Our results initiate a study of data processing principles in constrained learning environments, building on the classical work of Blackwell (1951) and DeGroot (1962). Given the nature of our results, we can generally conclude that, in the Bayes optimal setting, the data processing equality is expected to hold as long as the model class is large enough to be "robust" to the corruption present in our data. This can be a useful sanity check for practitioners designing machine learning architectures and losses, and studying their behavior on different datasets.

At present, however, these results do not directly connect to commonly used architectures or settings. One possible bridge is the modeling of pre-trained representations on our theoretical framework. A representation can be formalized as a map $f : \mathcal{X} \to \mathcal{X}$, corresponding to a deterministic Markov kernel, or as a stochastic one if the model is inherently stochastic (e.g., variational autoencoders or diffusion models). Consider a pre-trained model $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ whose output is fed into a new linear classification layer, with only the final layer optimized on task-specific data. Formally, this induces the learning problem $(\ell, \hat{\tau}\mathcal{H}, \phi)$. In such a setting, it becomes meaningful to ask whether a data processing inequality for Bayes risk holds uniformly, or whether specific choices of ℓ and $\mathcal{H}$ systematically violate it. Investigating this question may contribute to explaining the empirical observation that large pre-trained models often outperform task-specific models trained from scratch. This direction provides a concrete avenue for further investigation.

Bibliography

R. Alaiz-Rodríguez and N. Japkowicz. Assessing the impact of changing environments on classifier performance. In *Advances in Artificial Intelligence: 21st Conference of the Canadian Society for Computational Studies of Intelligence*, pages 13–24. Springer, 2008.

C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, 3rd edition, 2006.

D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154, 1946.

D. Blackwell. Comparison of experiments. *Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102, 1951.

D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

G. Blanchard and C. Scott. Decontamination of mutually contaminated models. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1–9. PMLR, 2014.

G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10(2): 2780–2824, 2016.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory (COLT)*, pages 92–100, 1998.

H. F. Bohnenblust, L. S. Shapley, and S. Sherman. Reconnaissance in Game Theory. *Santa Monica, CA: RAND Corporation*, pages 1–18, 1949.

C. H. Boll. *Comparison of experiments in the infinite case and the use of invariance in establishing sufficiency*. PhD thesis, Stanford University, 1955.

L. Boltzmann. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung. *Wiener Berichte*, 1877.

J. Bound, C. Brown, and N. Mathiowetz. Measurement error in survey data. In *Handbook of econometrics*, volume 5, pages 3705–3843. Elsevier, 2001.

L. Brillouin. *Science and Information Theory*. Dover Publications, 2013.

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

M. N. Bùi and P. L. Combettes. Interchange rules for integral functions. *arXiv preprint arXiv:2305.04872*, 2024.

A. J. Cabrera Pacheco and R. C. Williamson. The Geometry of Mixability. *Transactions on Machine Learning Research*, 2023.

A. J. Cabrera Pacheco, R. Derr, and R. C. Williamson. Aggregating algorithm and axiomatic loss aggregation. *Transactions on Machine Learning Research*, 2025.

L. Cao, D. McLaren, and S. Plosker. Centrosymmetric stochastic matrices. *Linear and Multilinear Algebra*, 70(3):449–464, 2022.

E. Çinlar. *Probability and Stochastics*. Springer, 2011.

N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Online learning of noisy data with kernels. In *Conference on Computational Learning Theory (COLT)*, pages 218–230, 2010.

J.-P. Chancelier and M. De Lara. A unified view of polarity for functions. *Journal of Convex Analysis*, 32(1):227–262, 2025.

J.-P. Chancelier, M. De Lara, and B. Tran. Minimization interchange theorem on posets. *Journal of Mathematical Analysis and Applications*, 509(1):125927, 2022.

J. Cheng, T. Liu, K. Ramamohanarao, and D. Tao. Learning with bounded instance and label-dependent label noise. In *International conference on machine learning (ICML)*, pages 1789–1799. PMLR, 2020.

K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

R. Clausius. Ueber die bewegende Kraft der Wärme und die Gesetze, welche sich daraus für die Wärmelehre selbst ableiten lassen. *Annalen der Physik*, 155(4):500–524, 1850.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2010.

A. Cotter, M. Gupta, and H. Narasimhan. On making stochastic classifiers deterministic. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019.

R. V. Craiu, R. Gong, and X.-L. Meng. Six statistical senses. *Annual Review of Statistics and Its Application*, 10(1):699–725, 2023.

Z. Cranko. *An analytic approach to the structure and composition of General Learning Problems*. PhD thesis, The Australian National University, 2021.

I. Csiszár. On generalized entropy. *Studia Scientiarum Mathematicarum Hungarica*, 4:401–419, 1969.

I. Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4):191–213, 1972.

F. Dahlqvist, V. Danos, I. Garnier, and O. Kammar. Bayesian inversion by ω-complete cone duality. In *International Conference on Concurrency Theory*, 2016.

M. H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

R. Derr and R. C. Williamson. Fairness and randomness in machine learning: Statistical independence and relativization. *The New England Journal of Statistics in Data Science*, 3(1): 55–72, 2024.

J. Du and Z. Cai. Modelling class noise with symmetric and asymmetric distributions. In *AAAI Conference on Artificial Intelligence*, 2015.

M. Du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems (NeurIPS)*, 2014.

M. Du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning (ICML)*, pages 1386–1394. PMLR, 2015.

H. Duanmu and W. Weiss. Finitely-additive, countably-additive and internal probability measures. *Comment. Math. Univ. Carolin*, 59(4):467–485, 2018.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220, 2008.

K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding Dataset Difficulty with V–Usable Information. In *International conference on machine learning (ICML)*, 2022.

A. M. Faden. The existence of regular conditional probabilities: necessary and sufficient conditions. *The Annals of Probability*, 13(1):288–298, 1985.

T. Fang, N. Lu, G. Niu, and M. Sugiyama. Rethinking importance weighting for deep learning under distribution shift. In *Advances in neural information processing systems (NeurIPS)*, pages 11996–12007, 2020.

T. Fang, N. Lu, G. Niu, and M. Sugiyama. Generalizing importance weighting to a universal solver for distribution shift problems. In *Advances in neural information processing systems (NeurIPS)*, 2023.

K. Fatras, H. Naganuma, and I. Mitliagkas. Optimal Transport meets Noisy Label Robust Loss and MixUp Regularization for Domain Adaptation. In S. Chandar, R. Pascanu, and D. Precup, editors, *Proceedings of The 1st Conference on Lifelong Learning Agents*, pages 966–981. PMLR, 2022.

T. Fel, L. Béthune, A. K. Lampinen, T. Serre, and K. Hermann. Understanding visual feature reliance through the lens of complexity. In *Advances in neural information processing systems (NeurIPS)*, pages 69888–69924, 2024.

D. Feldman. Some Properties of Bayesian Orderings of Experiments. *The Annals of Mathematical Statistics*, 43(5):1428 – 1440, 1972.

M. Finzi, S. Qiu, Y. Jiang, P. Izmailov, J. Z. Kolter, and A. G. Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *arXiv preprint arXiv:2601.03220*, 2026.

R. Fogliato, A. Chouldechova, and M. G'Sell. Fairness evaluation in presence of biased noisy labels. In *Artificial intelligence and statistics (AISTATS)*, pages 2325–2336. PMLR, 2020.

C. Fröhlich and R. C. Williamson. Data models with two manifestations of imprecision. *arXiv e-prints*, pages arXiv–2404, 2024.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys*, 46(4):1–37, 2014.

D. García García and R. C. Williamson. Divergences and risks for multiclass experiments. In *Conference on Computational Learning Theory (COLT)*, pages 28.1–28.20. JMLR Workshop and Conference Proceedings, 2012.

S. H. Garrido Mejia, E. Kirschbaum, A. Kekić, and A. Mastakouri. Estimating joint interventional distributions from marginal interventional data. *arXiv preprint arXiv:2409.01794*, 2024.

L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

A. Ghosh, N. Manwani, and P. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

A. Ghosh, H. Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*, 2017.

J. W. Gibbs. *Elementary Principles in Statistical Mechanics*. Yale University Press, 1902.

E. Giner. Necessary and sufficient conditions for the interchange between infimum and the symbol of integration. *Set-Valued and Variational Analysis*, 17(4):321–357, 2009.

P. K. Goel and M. H. DeGroot. Comparison of Experiments and Information Measures. *The Annals of Statistics*, 7(5):1066 – 1077, 1979.

S. A. Goldman and R. H. Sloan. Can PAC learning algorithms tolerate random attribute noise? *Algorithmica*, 14(1):70–84, 1995.

M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning (ICML)*, pages 2839–2848. PMLR, 2016.

J. González Hernández and O. Hernández Lerma. Extreme points of sets of randomized strategies in constrained optimization and control problems. *SIAM Journal on Optimization*, 15(4):1085–1104, 2005.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015.

L. Gresele, J. V. Kügelgen, J. Kübler, E. Kirschbaum, B. Schölkopf, and D. Janzing. Causal inference through the structural causal marginal problem. In *International conference on machine learning (ICML)*. PMLR, 2022.

E. Grinstein, N. Q. Duong, A. Ozerov, and P. Pérez. Audio style transfer. In *IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 586–590, 2018.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.

D. J. Hand. *Dark data: why what you don't know matters*. Princeton University Press, 2020.

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

M. Hardt, E. Mazumdar, C. Mendler-Dünner, and T. Zrnic. Algorithmic collective action in machine learning. In *International Conference on Machine Learning*, pages 12570–12586. PMLR, 2023.

H. He and E. A. García. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

B. Höltgen and R. C. Williamson. We should avoid the assumption of data-generating probability distributions in social settings. *arXiv preprint arXiv:2407.17395*, 2024.

T. H. Huxley. Geological reform. *Quarterly Journal of the Geological Society of London*, 25: 38–53, 1869.

L. Iacovissi, R. Derr, and R. C. Williamson. Comparing Corrupted Constrained Learning Problems. *arXiv preprint arXiv:2608.25745*, 2026a.

L. Iacovissi, N. Lu, and R. C. Williamson. Corruptions of supervised learning problems: Typology and mitigations. *Journal of Machine Learning Research*, 27(72):1–73, 2026b.

T. Ishida, G. Niu, W. Hu, and M. Sugiyama. Learning from complementary labels. *Advances in neural information processing systems (NeurIPS)*, 2017.

T. Ishida, G. Niu, A. Menon, and M. Sugiyama. Complementary-label learning for arbitrary losses and models. In *International conference on machine learning (ICML)*, pages 2971–2980. PMLR, 2019.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, page 694, 2016.

D. Johnston. *Statistical Causal Modelling and Decision Theory*. PhD thesis, The Australian National University, 2023.

O. Kallenberg. *Random measures, theory and applications*. Springer, 2017.

J. Katz-Samuels, G. Blanchard, and C. Scott. Decontamination of mutual contamination models. *Journal of machine learning research*, 20(41):1–72, 2019.

M. A. Khan, H. Yu, and Z. Zhang. On comparisons of information structures with infinite states. *Journal of Economic Theory*, 218, 2024.

M. A. Khan, H. Yu, and Z. Zhang. A simple proof of Blackwell's theorem on the comparison of experiments for a general state space. *Economics Letters*, 247(C):111686, 2025.

R. Kiryo, G. Niu, M. C. Du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems (NeurIPS)*, 2017.

A. Kobsa, H. Cho, and B. P. Knijnenburg. The effect of personalization provider characteristics on privacy attitudes and behaviors: An Elaboration Likelihood Model approach. *Journal of the Association for Information Science and Technology*, 67(11):2587–2606, 2016.

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning (ICML)*, pages 5637–5664. PMLR, 2021.

M. Kull and P. Flach. Patterns of dataset shift. In *First International Workshop on Learning over Multiple Contexts (LMCE) at ECML-PKDD*, 2014.

A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

L. Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, pages 1419–1455, 1964.

D. Lee and A. Makur. Contraction coefficients of product symmetric channels. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1–8, 2024.

T. Li and T. Unger. Willing to pay for quality personalization? Trade-off between quality and privacy. *European Journal of Information Systems*, 21(6):621–642, 2012.

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning (ICML)*, pages 3122–3130. PMLR, 2018.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

J. Liu, B. Wang, Z. Qi, Y. Tian, and Y. Shi. Learning from label proportions with generative adversarial networks. *Advances in neural information processing systems (NeurIPS)*, 2019.

T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019.

S. Lu, S. Chen, Y. Li, D. Bitterman, G. K. Savova, and I. Gurevych. Measuring Pointwise V-Usable Information In-Context-ly. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

S. Mac Lane. *Categories for the working mathematician*. Springer Science & Business Media, 2013.

A. Makur and L. Zheng. Comparison of contraction coefficients for f-divergences. *Problems of Information Transmission*, 56(2):103–156, 2020.

A. Malinin, N. Band, Y. Gal, M. Gales, A. Ganshin, G. Chesnokov, A. Noskov, A. Ploskonosov, L. Prokhorenkova, I. Provilkov, V. Raina, V. Raina, D. Roginskiy, M. Shmatova, P. Tigas, and B. Yangel. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

F. Mémoli, B. Vose, and R. C. Williamson. Geometry and stability of supervised learning problems. *Journal of Machine Learning Research*, 26:1–99, 2025.

X.-L. Meng. Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2): 685–726, 2018.

X.-L. Meng. Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4): 1161–1175, 2021.

X.-L. Meng. Comments on "Statistical inference with non-probability survey samples"—Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology*, 48(2):339–360, 2022.

A. Menon, B. van Rooyen, C. S. Ong, and R. C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *International conference on machine learning (ICML)*, pages 125–134. PMLR, 2015.

A. K. Menon, B. van Rooyen, and N. Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, 2018.

J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. *Advances in neural information processing systems (NeurIPS)*, 2013.

F. Osterreicher and I. Vajda. Statistical information and discrimination. *IEEE Transactions on Information Theory*, 39(3):1036–1039, 1993.

D. Oyen, M. Kucer, N. Hengartner, and H. S. Singh. Robustness to label noise depends on the shape of the noise distribution in feature space. *arXiv preprint arXiv:2206.01106*, 2022.

N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1944–1952, 2017.

L. Păunescu and F. Rădulescu. A generalisation to Birkhoff–von Neumann theorem. *Advances in Mathematics*, 308:836–858, 2017.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, Chichester, UK, 2016.

J.-P. Penot and C. Zalinescu. Harmonic sum and duality. *Journal of Convex Analysis*, 7(1): 95–114, 2000.

J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, pages 7599–7609. PMLR, 2020.

A. Perez. Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales. *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 183–208, 1957.

Y. Polyanskiy and Y. Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

M. Poovey. *A history of the modern fact: Problems of knowledge in the sciences of wealth and society*. University of Chicago Press, 1998.

N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. In *International conference on machine learning (ICML)*, pages 776–783, 2008.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

K. B. Rao and M. B. Rao. *Theory of charges: A study of finitely additive measures*, volume 109. Academic Press, 1983.

I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2022.

M. Reid and R. Williamson. Information, divergence and risk for binary experiments. *Journal of machine learning research*, 12:731 – 817, 2011.

R. T. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer, 1998.

N. Rostamzadeh, B. Hutchinson, C. Greer, and V. Prabhakaran. Thinking beyond distributions in testing machine learned models. In *NeurIPS Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

J. Rothwell. How the war on drugs damages black social mobility. *The Brookings Institution*, 30, 2014.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1232–1240. PMLR, 2016.

J. A. Sáez. Noise models in classification: Unified nomenclature, extended taxonomy and pragmatic categorization. *Mathematics*, 10(20), 2022.

Y. Safarov. Birkhoff's theorem and multidimensional numerical range. *Journal of Functional Analysis*, 222(1):61–97, 2005.

M. Salganicoff. Tolerating concept and sampling shift in lazy learning using prediction error context switching. *Artificial Intelligence Review*, 11:133–155, 1997.

N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Conference on Human Factors in Computing Systems (CHI)*, pages 1–15, 2021.

A. M. Saxe, Y. Bansal, J. Dapello, M. S. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations (ICLR)*, 2019.

E. Schechter. *Handbook of Analysis and Its Foundations*. Academic Press, San Diego, 1997.

C. Scott and J. Zhang. Learning from label proportions: A mutual contamination framework. *Advances in neural information processing systems (NeurIPS)*, 33:22256–22267, 2020.

G. Shackelford and D. Volper. Learning k-DNF with noise in the attributes. In *First annual workshop on Computational Learning Theory (COLT)*, pages 97–103, 1988.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423 and 623–656, 1948.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

A. N. Shiryaev and V. G. Spokoiny. *Statistical Experiments And Decision, Asymptotic Theory*. World Scientific, 2000.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

I. Siles, J. Espinoza-Rojas, A. Naranjo, and M. F. Tristán. The Mutual Domestication of Users and Algorithmic Recommendations on Netflix. *Communication, Culture & Critique*, 12(4): 499–518, 2019.

E. Simpson, A. Hamann, and B. Semaan. How to tame "your" algorithm: LGBTQ+ users' domestication of TikTok. *Proceedings of the ACM on Human-computer Interaction*, 6(GROUP): 1–27, 2022.

S. M. Srivastava. *A Course on Borel Sets*. Springer, New York, 1998.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

H. Strasser. *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*. Walter de Gruyter, 1985.

A. Subbaswamy, B. Chen, and S. Saria. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, 10(1):64–89, 2022.

M. Sugiyama and M. Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

K. Tang, M. Tao, J. Qi, Z. Liu, and H. Zhang. Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision (ECCV)*, pages 709–726. Springer, 2022.

Y. Tang, N. Lu, T. Zhang, and M. Sugiyama. Multi-class classification from multiple unlabeled datasets with partial risk regularization. In *Asian Conference on Machine Learning (ACML)*, pages 990–1005. PMLR, 2023.

N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. *arXiv preprint arXiv:1503.02406*, 2015.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communications, Control and Computing*, pages 368–377, 1999.

E. Torgersen. *Comparison of statistical experiments*. Cambridge University Press, 1991.

A. Tsymbal. The problem of concept drift: Definitions and related work. Technical report, Department of Computer Science, The University of Dublin, Trinity College, Dublin, Ireland, 2004.

B. van Rooyen. *Machine learning via transitions*. PhD thesis, The Australian National University, 2015.

B. van Rooyen and R. C. Williamson. A theory of learning with corrupted labels. *Journal of machine learning research*, 18(228):1–50, 2018.

B. van Rooyen, A. Menon, and R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems (NeurIPS)*, 28, 2015.

A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 839–847, 2017.

P. Vorburger and A. Bernstein. Entropy-based concept shift detection. In *Sixth International Conference on Data Mining (ICDM)*, pages 1113–1118. IEEE, 2006.

V. Vovk. Aggregating strategies. In *Conference on Computational Learning Theory (COLT)*, pages 371–383, 1990.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Q. Wang, B. Han, T. Liu, G. Niu, J. Yang, and C. Gong. Tackling instance-dependent label noise via a universal probabilistic model. In *AAAI Conference on Artificial Intelligence*, pages 10183–10191, 2021.

G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick. Presence-only data and the EM algorithm. *Biometrics*, 65(2):554–563, 2009.

H. White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981.

G. Widmer and M. Kubat. Effective learning in dynamic environments by explicit context tracking. In *European Conference on Machine Learning (ECML)*, volume 6, pages 227–243, 1993.

G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996.

R. A. Wijsman. Continuity of the Bayes Risk. *The Annals of Mathematical Statistics*, 41(3): 1083–1085, 1970.

R. Williamson, E. Vernet, and M. Reid. Composite multiclass losses. *Journal of machine learning research*, 17(222):1–52, 2016.

R. C. Williamson. Process and Purpose, Not Thing and Technique: How to Pose Data Science Research Challenges. *Harvard Data Science Review*, 2(3), 2020.

R. C. Williamson and Z. Cranko. The geometry and calculus of losses. *Journal of machine learning research*, 24(342):1–72, 2023.

R. C. Williamson and Z. Cranko. Information processing equalities and the information–risk bridge. *Journal of machine learning research*, 25(103):1–53, 2024.

World Economic Forum Global Future Council on Human Rights 2016-2018. How to Prevent Discriminatory Outcomes in Machine Learning. White paper, World Economic Forum, 2018.

A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2521–2530, 2017.

Y. Xu, S. Zhao, J. Song, R. Stewart, and S. Ermon. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.

K. Yamazaki, M. Kawanabe, S. Watanabe, M. Sugiyama, and K.-R. Müller. Asymptotic bayesian generalization error when training and test distributions are different. In *International conference on machine learning (ICML)*, pages 1079–1086, 2007.

Y. Yao, T. Liu, M. Gong, B. Han, G. Niu, and K. Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4409–4420, 2021.

F. X. Yu, K. Choromanski, S. Kumar, T. Jebara, and S.-F. Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

X. Yu, T. Liu, M. Gong, K. Zhang, K. Batmanghelich, and D. Tao. Label-noise robust domain adaptation. In *International conference on machine learning (ICML)*, pages 10913–10924. PMLR, 2020.

S. Zezulka and K. 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.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning (ICML)*, pages 819–827. PMLR, 2013.

K. Zhang, M. Gong, P. Stojanov, B. Huang, Q. Liu, and C. Glymour. Domain adaptation as a problem of inference on graphical models. *Advances in neural information processing systems (NeurIPS)*, pages 4965–4976, 2020a.

T. Zhang, I. Yamane, N. Lu, and M. Sugiyama. A one-step approach to covariate shift adaptation. In *Asian Conference on Machine Learning (ACML)*, pages 65–80. PMLR, 2020b.

T. Zhou and X. Wu. Examining generative AI user disclosure intention: An ELM perspective. *Universal Access in the Information Society*, 24(2):1209–1220, 2025.

X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study. *The Artificial Intelligence Review*, 22(3):177, 2004.

C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, Singapore, 2002.

Appendix A

Existing Corruption Models

A Markov kernel-based taxonomy is substantially different from previous work. Therefore, in this section, we carefully examine how existing corruption models fit into our taxonomy. This involves reformulating them as specific instances of Markovian corruptions, thereby unveiling their relationships within the corruption hierarchy presented in Fig. 4.1a.

The primary challenge stems from the lack of consistency across the literature; different authors sometimes refer to the same corruption process with different names or use the same name to denote different settings. For instance, classical studies on concept drift (Widmer and Kubat, 1996; Lu et al., 2019) generally define it as a mismatch in the joint distributions between two different learning environments, e.g., training and test times. Meanwhile, works such as in Moreno-Torres et al. (2012) characterize it further by necessitating unchanged attribute or label priors.

We attempt a partial unification of the corruption models we are aware of by establishing connections as depicted in Tab. 4.2, while additional technical intricacies regarding correspondences and relationships are elucidated in subsequent sections.

A.1 Simple Corruptions

The most well-known and widely studied corruptions in the literature are the simple cases, where the corruption solely acts on the feature space $\mathcal{X}$ or the label space $\mathcal{Y}$. We discuss below various examples of simple corruptions, i.e. in the sets $\mathcal{M}(\mathcal{X}, \mathcal{X})$ and $\mathcal{M}(\mathcal{Y}, \mathcal{Y})$, as defined in Fig. 4.1.

Attribute Noise

The problem of attribute noise concerns errors that are introduced into the observations of attribute X, leaving the labels untouched (Shackelford and Volper, 1988; Goldman and Sloan, 1995; Zhu and Wu, 2004; Williamson and Cranko, 2024). Widely studied examples of such errors include erroneous attribute values and missing attribute values. Instead of observing (X, Y), in the first case, one can only observe a distorted version of X, e.g.,

$(X + N, Y)$ with some independent noise random variable $N \perp\!\!\!\perp X$; in the second case, one's observation of X contains missing values.

Let $\mathbf{X} = (x_{ij})_{1\le i\le n, 1\le j\le d}$ be the complete input matrix, with $|\mathcal{X}| = n$, and $\mathbf{M} = (m_{ij})_{1\le i\le n, 1\le j\le d}$ be the associated missingness indicator matrix such that $m_{ij} = 1$ if x_{ij} is observed and $m_{ij} = 0$ if x_{ij} is missing. Then the corresponding observed input matrix is $\mathbf{X}_o = \mathbf{X} \odot \mathbf{M}$ and its missing counterpart is $\mathbf{X}_m = \mathbf{X} - \mathbf{X}_o$, where $\odot$ denotes Hadamard product. The missing value mechanisms are further categorized into three types based on their dependencies (Rubin, 1976; Little and Rubin, 2019):[1]

- Missing completely at random (MCAR): the cause of missingness is entirely random, i.e., $p(M \mid X) = p(M)$ does not depend on X_o or X_m. This corresponds to having a trivial Markov kernel acting on the clean distribution, $\tau : \{*\} \rightsquigarrow \mathcal{X} \equiv \mu \in \Delta_{\text{ca}}(\mathcal{X})$.

- Missing not at random (MNAR): the cause of missingness depends on both observed variables and missing variables, i.e., $p(M \mid X) = p(M \mid X_o, X_m)$. This case corresponds to our non-trivial $\tau : \mathcal{X} \rightsquigarrow \mathcal{X}$.

- Missing at random (MAR): the cause of missingness depends on observed variables but not on missing variables, i.e., $p(M \mid X) = p(M \mid X_o)$. This case is a sub-case of the non-trivial $\tau : \mathcal{X} \rightsquigarrow \mathcal{X}$, which is not directly specifiable by our taxonomy because of the different premises it is built on.

Class-Conditional Noise (CCN)

The problem of CCN arises in situations where, instead of observing the clean labels, one can only observe corrupted labels that have been flipped with a label-dependent probability, while the marginal distribution of the instance remains unchanged (Natarajan et al., 2013; Patrini et al., 2017; van Rooyen and Williamson, 2018; Williamson and Cranko, 2024). CCN is an example of simple label corruption, $\mathcal{M}(\mathcal{Y}, \mathcal{Y})$, that can be formulated as a corrupted posterior. For classification tasks, $\mathcal{Y}$ is assumed to be a finite space. Therefore the corruption $\lambda : \mathcal{Y} \rightsquigarrow \mathcal{Y}$ can be represented by a column-stochastic matrix $\mathbf{T} = (\rho_{ij})_{1\le i\le |Y|, 1\le j\le |Y|}$ which specifies the probability of the clean label $Y = j$ being flipped to the corrupted label $\tilde{Y} = i$, i.e., $\forall i, j,\ \rho_{ij} = p(\tilde{Y} = i \mid Y = j)$. The corrupted joint distribution can be rewritten as $\tilde{P} = \sum_Y p(\tilde{Y} \mid Y)\, p(Y \mid X)\, p(X)$. In the literature, $\mathbf{T}$ is known as the noise transition matrix with its elements ρ_{ij} referred to as the noise rates, and is useful for designing loss correction approaches (our results in § 5.2 significantly generalize existing loss correction results in CCN to our broad class of simple, dependent and combined corruptions) (Patrini et al., 2017). Prior to the proposal of the CCN model, early studies primarily focused on a symmetric subcase of $\mathbf{T}$ in binary classification, known as random classification noise (RCN) (Angluin and Laird, 1988; Blum and Mitchell, 1998; van Rooyen et al., 2015). Note that in RCN, the output of the corruption $\lambda : \mathcal{Y} \rightsquigarrow \mathcal{Y}$ remains constant w.r.t. its parameters. Recently, some

[1] Assume the rows x_i, m_i are assigned a joint distribution, and X and M are treated as random variables.

variants of CCN have been further developed, for example, in Ishida et al. (2017, 2019), complementary labels are modeled via a symmetric $\mathbf{T}$ whose diagonal elements are all equal to zero.

A.2 Dependent Corruptions

Although simple corruptions have been well studied and understood, more complexity arises in dependent cases, yet they receive relatively less attention and understanding. We discuss below examples of dependent corruptions in the sets $\mathcal{M}(\mathcal{Y}, \mathcal{X})$, $\mathcal{M}(\mathcal{X}, \mathcal{Y})$, $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X})$ and $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$, as defined in Fig. 4.1a.

Style Transfer

Style transfer refers to the process of migrating the artistic style of a given image to the content of another image (Gatys et al., 2015; Johnson et al., 2016). The primary objective is to recreate the second image with the designated style of the first image. In recent developments, it has also been applied to audio signals (Grinstein et al., 2018). If we represent the style of the first image by Y, and the second image and the reconstructed image as X and $\tilde{X}$ respectively, style transfer serves as an illustrative example of $\tau \colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ "corruption". Note that the aim here is to *learn how to corrupt* instead of learning in the presence of corruption. We mention this connection because our framework can also be used with different purposes, but underline that our BR results are not applicable to this case. The process of style transfer can be formulated as a corrupted posterior.

Adversarial Noise

In contrast to additive random attribute noise, adversarial noise is specifically crafted by adversaries for each instance with the intent of changing the model's prediction of the correct label (Szegedy et al., 2013; Goodfellow et al., 2015; Papernot et al., 2016; Kurakin et al., 2018; Hendrycks et al., 2021). Such adversarial examples raise significant security concerns as they can be utilized to attack machine learning systems, even in scenarios where the adversary has no access to the underlying model. The adversarial noise is an example of $\tau \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X})$ corruption that can be formulated as a corrupted experiment.

Instance-Dependent Noise (IDN)

As a counterpart to CCN, the problem of IDN arises in situations where, instead of observing the clean labels, one can only observe corrupted labels that have been flipped with an instance-dependent (but not label-dependent) probability (Ghosh et al., 2015; Menon et al., 2018). It is a special case of the ILN noise model, which we will describe later. IDN is an example of $\lambda \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ corruption that can be formulated as a corrupted experiment.

Instance- and Label-Dependent Noise (ILN)

ILN is the most general label noise model, which arises in situations where, instead of observing clean labels, one can only observe corrupted labels that have been flipped with an instance- and label-dependent probability (Menon et al., 2018; Cheng et al., 2020; Yao et al., 2021; Wang et al., 2021). ILN is an example of $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$ corruption that can be formulated as a corrupted posterior. Compared to the matrix representation $\mathbf{T}$ of the CCN corruption $\kappa \in \mathcal{M}(\mathcal{Y}, \mathcal{Y})$, the ILN corruption $\kappa \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$ can be represented by a matrix-valued function of the instance $\mathbf{T}(x) = (\rho_{ij}(x))_{1 \leq i \leq |\tilde{Y}|, 1 \leq j \leq |Y|}$ which specifies the probability that the instance $X = x$ with the clean label $Y = j$ being flipped to the corrupted label $\tilde{Y} = i$, i.e., $\forall i, j,\ \rho_{ij}(x) = p(\tilde{Y} = i \mid Y = j, X = x)$. Some subcases of ILN have also been studied in the literature, for example, the boundary-consistent noise, which considers a label flip probability based on a score function of the instance and label. The score aligns with the underlying class-posterior probability function, resulting in instances closer to the optimal decision boundary having a higher chance of its label being flipped (Du and Cai, 2015).

A.3 Combined Corruptions

Given the simple and dependent corruptions, we can combine them to generate 2-parameter joint corruptions, i.e., $\mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$. Below, we discuss some examples of combined noise models illustrated in Fig. 4.1b.

Combined Simple Noise

The simplest combined corruption is the combined simple noise, where the observations of attribute X are subject to some errors and the observed labels Y are flipped with a label-dependent probability (Williamson and Cranko, 2024). Combined simple noise is an example of $\tau : \mathcal{X} \rightsquigarrow \mathcal{X} \otimes \lambda : \mathcal{Y} \rightsquigarrow \mathcal{Y}$ corruption that can be formulated as a corrupted experiment.

Target Shift

In the literature, target shift, also known as prior probability shift, refers to the situation where the prior probability $p(Y)$ is changed while the conditional distribution $p(X \mid Y)$ remains invariant across training and test domains (Japkowicz and Stephen, 2002; He and García, 2009; Buda et al., 2018; Lipton et al., 2018). The definition is established by assuming certain invariance from a generative perspective of the learning problem, that is, considering it as a corruption of the experiment according to $P = \pi_Y \times E$. However, when examining the learning problem from a discriminative perspective, the change in $p(Y)$ may cause changes in both $p(X)$ and $p(Y \mid X)$ due to the Bayes rule. Existing frameworks for the categorization of target shift do not capture these implications, as they are based on the notion of invariance from a single perspective of the E direction. In contrast, our framework categorizes corruptions based on their dependencies and therefore is advantageous by

offering dual perspectives from both the E and F directions. Specifically, target shift is a subcase of $\tau\colon \mathcal{Y} \rightsquigarrow \mathcal{X} \otimes \lambda\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$ corruption and can be formulated either as a corrupted experiment or as a corrupted posterior. The corrupted distribution is given by $\tilde{P} = (\pi_Y \times E) \circ (\tau \otimes \lambda)$ or $\tilde{P} = (\pi_X \times F) \circ (\tau \otimes \lambda)$.

Covariate Shift

In the literature, covariate shift refers to the situation where the marginal distribution $p(X)$ is changed while the class-posterior probability $p(Y \mid X)$ remains invariant across training and test domains (Shimodaira, 2000; Quiñonero-Candela et al., 2008; Sugiyama and Kawanabe, 2012; Zhang et al., 2020b). Similarly to target shift, the definition is based on assuming invariance from the discriminative perspective of the learning problem, treating it as a corruption of the posterior using $P = \pi_X \times F$. However, when viewed from a generative perspective, changes in $p(X)$ may lead to changes in $p(Y)$ and $p(X \mid Y)$ due to the Bayes rule. Covariate shift is a subcase of $\tau\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X} \otimes \lambda\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ corruption and can be formulated either as a corrupted posterior or as a corrupted experiment. The corrupted distribution is given by $\tilde{P} = (\pi_Y \times E) \circ (\tau \otimes \lambda)$ or $\tilde{P} = (\pi_X \times F) \circ (\tau \otimes \lambda)$.

It is important to clarify that while covariate shift is sometimes used interchangeably with sample selection bias in certain literature, the two are not synonymous. This point is also mentioned by the author of the original covariate shift paper (Shimodaira, 2000) in the book by Quiñonero-Candela et al. (2008, Chapter 11): they claim covariate shift to be a special form of selection bias when the latter is taken under assumption of missing at random, and in general, selection bias without such a structure is difficult. However, based on our definition of selection bias in Def. 4.11, it is not true that covariate shift is a special form of selection bias. Nonetheless, various definitions exist in the literature and they can relate in different ways.

We introduce here a classical definition of selection bias, which leads to the one we gave in the main text (see Quiñonero-Candela et al., 2008, Chapter 3.2). Let S be a binary selection variable deciding whether a datum is included in the training set ($S = 1$) or excluded from it ($S = 0$). The corrupted distribution by selection bias can be expressed as $\tilde{P}(X, Y) = P(X, Y \mid S = 1)$. By assuming the missing at random structure, where S is independent of Y given X: $P(S \mid X, Y) = P(S \mid X)$, we recover covariate shift where $P(X \mid S = 1) \neq P(X)$ and $P(Y \mid X, S) = P(Y \mid X)$.

Note that covariate shift is only harmful when the model class is misspecified (Shimodaira, 2000). This issue is typically addressed through importance-weighted empirical risk minimization–weighting the training losses according to the ratio of the test and training input densities (Sugiyama and Kawanabe, 2012; Fang et al., 2023). In such context, the additional assumption of $P \ll \tilde{P}$ is required so to obtain the weighted risk on the training set to be equal to the risk on the test set. This assumption is therefore in contrast with Def. 4.11, requiring for selection bias the support condition $\tilde{P} \ll P$.

More in general, selection bias necessitates both the support condition and the selection

condition with bounded $\frac{dP}{d\tilde{P}}(x_i, y_i)\, \forall i \in [n]$, which are stronger than the original definition of covariate shift assuming only the change of marginal distribution $p(X)$ and the invariance of the class-posterior probability $p(Y \mid X)$. As a result, there exist covariate shift scenarios that cannot be attributed to selection bias when $\tilde{P} \ll P$ is not the case.

Generalized Target Shift

In the literature, generalized target shift refers to the situation where the prior probability $p(Y)$ and the conditional distribution $p(X \mid Y)$ both change across training and test domains, however, with some invariance assumptions in the latent space (Zhang et al., 2013; Gong et al., 2016; Yu et al., 2020). Generalized target shift is a subcase of $\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow X \otimes \lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$ corruption that can be formulated as a corrupted experiment. Note that simplified sub-examples can also manifest as a generalized target shift; however, it is important to avoid degenerating into the basic $\tau \colon \mathcal{X} \rightsquigarrow \mathcal{X}$ corruption, as it would violate the requirement of corrupting the label distribution.

Concept Drift, Concept Shift, and Sampling Shift

Concept drift refers to the situation where data evolves over time, leading to different categorizations depending on the nature of the change. Typically, concept drift between time point t_0 and t_1 is characterized by $p_{t_0}(X, Y) \neq p_{t_1}(X, Y)$ (Widmer and Kubat, 1996; Gama et al., 2014; Lu et al., 2019). In our words, $p_{t_1}(X, Y)$ can be seen as a corrupted version of $p_{t_0}(X, Y)$. Given its generality, this case can be associated with every corruption in our framework; therefore, the most general correspondence is the $\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow X \otimes \lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$ joint Markov kernel.

There are two types of concept drifts popular in the literature:

- Concept shift (Vorburger and Bernstein, 2006; Widmer and Kubat, 1993; Salganicoff, 1997): in this case, $p(Y \mid X)$ changes over time, and such changes can occur with or without changes on $p(X)$, often referred to as concept shift; in our framework, this is a subcase of $\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow X \otimes \lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$ corruption. More details in Tab. B.1.

- Sampling shift (Tsymbal, 2004; Widmer and Kubat, 1993; Salganicoff, 1997): here, $p(X)$ changes over time while $p(Y \mid X)$ remains invariant, also known as virtual drift; in our framework, this is a subcase of $\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X} \otimes \lambda \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ corruption. More details in Tab. B.1.

However, in the literature, concept drift is also defined with more invariance assumptions. For example, in Moreno-Torres et al. (2012), they define concept drift as $p(Y \mid X)$ changing while $p(X)$ remains invariant or $p(X \mid Y)$ changing while $p(Y)$ remains invariant. Similar to instance- and label-dependent noise and covariate shift, they are examples of $\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$ corruption and $\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X} \otimes \lambda \colon \mathcal{Y} \rightsquigarrow \mathcal{Y}$ corruption that involve more corrupted spaces at different time points.

Appendix B

Comparison with Other Taxonomies

We notice that most of the taxonomies available in the literature are based on the notion of invariance, inducing taxonomies very different from ours. We here connect our work to other categorization paradigms for distribution shifts, although without claiming it to be a comprehensive review.

We divide taxonomies in two main groups: the traditional ones, focusing on identifying which probability in the set $\{\pi_Y, \pi_X, E, F\}$ is forced to be left invariant and which one is forced to be corrupted (Moreno-Torres et al., 2012), and the causal ones, where a causal graph structure is associated to the corruption process (Zhang et al., 2020a) and hidden structures are possibly involved so that some latent feature is left unchanged by the corruption (Kull and Flach, 2014; Subbaswamy et al., 2022). Notice that in none of the cited works the corrupted distribution is assumed to have a specific form or to be "close enough" to the clean one. We do not review these other cases, because too far from our point of view and objective.

Focusing on the first case, a complete *traditional taxonomy* has four types of possible corruptions. Taking into account which marginal or conditional probability is forced to be corrupted, we obtain a finite number of corruption subcases of these four macro-types. However, the different cases obtained may overlap, as it is schematically shown in Tab. B.1. The cases that have a clear correspondence with ours are the ones leaving invariant a marginal distribution, generating simple noises. All the other cases cannot be directly mapped into our taxonomy, so we explicitly write the range of corruption types covered by them.

As for *causal taxonomies*, based on causal graphs, they are more difficult to describe in a unified way since different applications lead to different notations. We then avoid doing so, and limit ourselves to qualitatively compare them with our work.

A common trend is to identify the current space we live in with a variable D, the *domain* or

Corrupted	Invariant	Name in (Moreno-Torres et al., 2012)	DAG in (Kull and Flach, 2014)	Ours
at least one among $\{\pi_X, F, E\}$, according to compatibility	π_Y	concept shift[1] when $Y \to X$	$D \swarrow$, $X \leftarrow Y$	subcase of $\kappa : X \rightsquigarrow X$
at least one among $\{\pi_Y, F, E\}$, according to compatibility	π_X	concept shift[2] when $X \to Y$	$D \searrow$, $X \rightarrow Y$	subcase of $\lambda : Y \rightsquigarrow Y$
at least π_Y, causing π_X or F to change	E	prior probability shift[3] when $Y \to X$	$D \searrow$, $X \leftarrow Y$	at least a $\lambda : Y \rightsquigarrow Y$ subcase, at most $\kappa : X \times Y \rightsquigarrow X$ $\otimes$ $\lambda : X \times Y \rightsquigarrow Y$
at least π_X, causing π_Y or E to change	F	covariate shift[4] when $X \to Y$	$D \swarrow$, $X \rightarrow Y$	at least a $\kappa : X \rightsquigarrow X$ subcase, at most $\kappa : X \times Y \rightsquigarrow X$ $\otimes$ $\lambda : X \times Y \rightsquigarrow Y$

Table B.1: Traditional taxonomies compared to ours.

environment, possibly taking values in $\mathbb{N}$. This variable is then included in the causal graph indicating on what it is acting, as done in the examples in Tab. B.1. In the case described by Def. 4.2 we restrict it to take values in $\{0, 1\}$, the clean and corrupted environments. This representation is again missing some of our corruptions, since it is only possible to encode X and Y changing across domains and not whether other environments influence the current one. The shifts in Kull and Flach (2014, Fig. 3) involving hidden variables (concept shift subcases) resemble our idea of a "latent process" influencing the current environment, but still fail to cover all the possible cross-domain influence in Fig. 4.1. An additional limitation of the causal approach lies in the causal assumption itself; we are forced, in this setting, to consider only one of E and F as a valid representation of the generative process, while in our framework we are not forced to make this choice. However, we still use it when it is available.

In both described classes of taxonomies, it is neither natural nor simple to define a hierarchy of corruptions. In particular, in the traditional taxonomy the specification of what is corrupted leaves room for other components to be forced to be influenced, creating overlaps between cases. As for composing them, a DAG representation of a corruption model can facilitate their chaining. Nevertheless, feasibility rules are rather complex and unclear, given the overlapping nature of the corruptions and identifiability problems for causal representations (Pearl, 2009).

[16] as sole attribute noise: (Shackelford and Volper, 1988; Goldman and Sloan, 1995; Zhu and Wu, 2004; Williamson and Cranko, 2024)

[17] as sole class-conditional noise: (Angluin and Laird, 1988; Blum and Mitchell, 1998; Natarajan et al., 2013; Patrini et al., 2017; van Rooyen and Williamson, 2018; Williamson and Cranko, 2024); in general: (Yamazaki et al., 2007; Alaiz-Rodríguez and Japkowicz, 2008)

[18] or label shift, or class imbalance: (Japkowicz and Stephen, 2002; He and García, 2009; Buda et al., 2018; Lipton et al., 2018; Tang et al., 2022)

[19] (Shimodaira, 2000; Quiñonero-Candela et al., 2008; Sugiyama and Kawanabe, 2012; Zhang et al., 2020b)

Appendix C

Bayesian Inversion in Category Theory

In this section, we provide a more formal definition of the Bayesian inverse of a Markov kernel, based on some existing results from category theory applied to Bayesian learning (Dahlqvist et al., 2016). In fact, Bayesian update is exactly kernel inversion. These results guarantee the valid and proper utilization of the inverse kernel in the current paper. Before delving into the details, we introduce relevant categorical concepts, establishing the necessary background to proceed. For a comprehensive overview of category theory, we recommend that interested readers refer to Mac Lane (2013).

C.1 Categorical Concepts

To begin, let `Mes` be the category of measurable spaces with measurable maps as morphisms, and `Pol` be the category of *Polish spaces*, i.e., separable metric spaces for which a complete metric exists, with continuous maps as morphisms. The functor $\mathcal{B} : \texttt{Pol} \to \texttt{Mes}$ associates any Polish space to the measurable space with the same underlying set equipped with the Borel σ-algebra, and interprets continuous maps as measurable ones. Measurable spaces in the range of $\mathcal{B}$ are *standard Borel spaces*, which are important because the *regular conditional probabilities* are known to exist in them, but not in general (Faden, 1985). Therefore, they will be used as the building block of the `Krn` category in the subsequent Bayesian inversion theorem.

The *Giry monad* is the monad on a category of suitable spaces which sends each suitable space X to the space of suitable probability measures on X. In this case, the set of suitable spaces is the one of the `Mes` category induced by the functor $\mathcal{B}$. To define it more formally, we now consider the triple $(\mathcal{P}, \mu, \delta)$:

- the functor $\mathcal{P}$ is such that we assign to every space X in `Mes` the set of all probability measures on X, $\mathcal{P}(X)$. This is equipped with the smallest σ-algebra that makes the

Categorical	Probabilistic
Kleisli category of Giry monad $\mathtt{G}$, $\mathcal{K}\ell$	measurable spaces as objects and Markov kernels as arrows
arrows in category $1 \downarrow \mathcal{K}\ell$	$\mathcal{M}(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ have marginals ϕ and ψ, respectively
arrows in category $1 \downarrow F$	subset of the above $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ with measure-preserving maps induced by identity kernels δ
Kleisli composition $\circ_{\mathtt{G}}$	chain composition $\circ$ in **P1** with transitional kernels
$\alpha^{\mathcal{X}}_{\mathcal{Y}} : \mathrm{Hom}_{\mathtt{Krn}}(\mathcal{X}, -) \to \Gamma(\mathcal{X}, -)$	product composition $\times$ in **P2** with a kernel and a probability

Table C.1: Comparison of categorical concepts in Dahlqvist et al. (2016) and probabilistic concepts in this paper.

evaluation function $ev_B : \mathcal{P}(\mathcal{X}) \to [0,1] = P \mapsto P(B)$ measurable, for B a measurable subset in $\mathcal{X}$;

- the multiplication of the monad, $\mu : \mathcal{P}^2 \Rightarrow \mathcal{P}$, is defined by

$$\mu_{\mathcal{X}}(Q)(B) = \int_{\psi \in \mathcal{P}(\mathcal{X})} ev_B(\psi) dQ \, ;$$

- the unit of the monad, $\delta : Id \Rightarrow \mathcal{P}$, sends a point $x \in \mathcal{X}$ to the Dirac measure at x.

This equips the endofunctor $\mathcal{P} : \mathtt{Mes} \to \mathtt{Mes}$ into a monad, that is, the Giry monad $\mathtt{G} := (\mathcal{P}, \mu, \delta)$ on measurable spaces.

The Kleisli category of $\mathtt{G}$, denoted by $\mathcal{K}\ell$, has the same objects as $\mathtt{Mes}$, and the morphism $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ in $\mathcal{K}\ell$ is a kernel $\kappa : X \to \mathcal{P}(\mathcal{Y})$ in $\mathtt{Mes}$. The Kleisli composition of kernel $\kappa : \mathcal{X} \rightsquigarrow \mathcal{Y}$ with $\lambda : \mathcal{Y} \rightsquigarrow \mathcal{Z}$ is given by $\lambda \circ_{\mathtt{G}} \kappa = \mu_Z \circ \mathcal{P}(\lambda) \circ \kappa$. The action of the functor $\mathcal{P}$ on a kernel results, by definition, in the push-forward operator $\mathcal{P}(\kappa)(\cdot) := (\cdot) \circ \kappa^{-1}$, defined on a suitable space of probabilities. Hence, $\circ_{\mathtt{G}}$ is the same as the chain composition we defined in Def. 2.24.

C.2 The Bayesian Inversion Theorem

Dahlqvist et al. (2016) investigates how and when the Bayesian inversion of the Markov kernel is defined, both directly on the category of measurable spaces, and indirectly by considering the associated linear operators (i.e., Markov transition, see Çinlar, 2011). Below, we only introduce the first result of the Bayesian inversion theorem, given the focus of Markov kernels we have in the current paper, and then describe the pseudo-inversion operation in Def. 5.9 in a more formal way.

The category of Markov kernels considered here is that of *typed kernels*. Their definition is tied to a fixed probability ϕ on $\mathcal{X}$ and a fixed probability ψ on $\mathcal{Y}$, so that $\kappa \circ_{\mathtt{G}} \phi = \psi$, instead of being characterized for every probability on $\mathcal{X}$ and every reachable output. In

general, one can define Markov kernels as operators on the space of probabilities; that is not our interest, as we tie the concept of corruption to a specific pair of clean and corrupted distributions. This remark is also crucial for understanding our notion of exhaustiveness in Appendix C.2.1.

The key object for building the inversion operation is the `Krn` category, similar to our notion of space of Markov kernels $\mathcal{M}(\mathcal{X}, \mathcal{Y})$, but with an equivalence relation acting on it. We describe its construction in the following steps.

1. Let $F\colon \texttt{Mes} \to \mathcal{K}\ell$ be the functor embedding `Mes` into $\mathcal{K}\ell$ which acts identically on spaces and maps measurable arrows $\kappa\colon \mathcal{X} \to \mathcal{Y}$ to Kleisli arrows $F(\kappa) = \delta \circ \kappa$. This means that $F(\kappa)$ only allows one possible jump at each x in $\mathcal{X}$, with δ an identical jump (i.e., a deterministic kernel on $\mathcal{Y}$).

2. It further induces the category $1 \downarrow F$ of probabilities $\phi\colon 1 \rightsquigarrow \mathcal{P}(\mathcal{X})$, denoted by $(\mathcal{X}, \phi)$, and morphisms $\kappa : (\mathcal{X}, \phi) \rightsquigarrow_\delta (\mathcal{Y}, \psi)$ as degenerate arrows $F(\kappa)\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ s.t. $\psi = F(\kappa) \circ_{\mathtt{G}} \phi = \mathtt{P}(\kappa)(\phi) = \phi \circ \kappa^{-1}$. In more familiar terms, this is saying that ψ is the push-forward of ϕ along κ. $1 \downarrow F$ includes all measure-preserving maps induced by degenerate arrows.

3. When the arrows are not degenerate, we obtain the supercategory $1 \downarrow \mathcal{K}\ell$ with the same objects. Specifically, in this category, an arrow from $(\mathcal{X}, \phi)$ to $(\mathcal{Y}, \psi)$ is any Kleisli arrow $\kappa\colon \mathcal{X} \rightsquigarrow Y$ s.t. $\psi = \kappa \circ_{\mathtt{G}} \phi$, and the arrows are what we denoted as $\mathcal{M}(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X}$ has marginal probability ϕ and $\mathcal{Y}$ has marginal probability ψ.

4. Markov kernels cannot be inverted as they are, because of their *non-singularity*. Lemma 3 in Dahlqvist et al. (2016) characterizes it by proving that for a kernel $\kappa\colon (\mathcal{X}, \phi) \rightsquigarrow (\mathcal{Y}, \psi)$ there are ϕ-negligibly many points jumping to ψ-negligible sets.

Once the non-singularity is understood, we can define an equivalence relation on $1 \downarrow \mathcal{K}\ell$ that allows a well-posed definition of the inverse kernel.

Definition C.1. *For all objects $(\mathcal{X}, \phi), (\mathcal{Y}, \psi)$, $R_{(\mathcal{X},\phi),(\mathcal{Y},\psi)}$ is the smallest equivalence relation on $Hom_{1\downarrow\mathcal{K}\ell}(X, Y)$ such that*

$$(\kappa, \kappa') \in R_{(\mathcal{X},\phi),(\mathcal{Y},\psi)} \quad \Leftrightarrow \quad \kappa = \kappa' \ \ \phi\text{-a.s.}$$

They prove R to be a congruence relation on $1 \downarrow \mathcal{K}\ell$ in their Lemma 4. This congruence relation allows us to define the quotient category, with proper morphisms.

Definition C.2. *The category* `Krn` *is the quotient category* $(1 \downarrow \mathcal{K}\ell)/R$.

Having defined the category, we have to build the functions that are going to constitute the Bayesian inversion operator, i.e., a bijection between $\mathrm{Hom}_{\mathtt{Krn}}((\mathcal{X}, \phi), (\mathcal{Y}, \psi))$ and $\mathrm{Hom}_{\mathtt{Krn}}((\mathcal{Y}, \psi), (\mathcal{X}, \phi))$. There are two mappings between the `Krn` category and the space

of couplings associated to $(\mathcal{X},\phi),(\mathcal{Y},\psi)$. The first is equivalent to the product composition we defined in Def. 2.25 applied to a kernel (i.e. conditional probability) and a probability, and is formally written as

$$\alpha^{\mathcal{X}}_{\mathcal{Y}}\colon \mathrm{Hom}_{\mathtt{Krn}}((\mathcal{X},\phi),(\mathcal{Y},\psi)) \to \Gamma((\mathcal{X},\phi),(\mathcal{Y},\psi)) \text{ s.t. } \alpha^{\mathcal{X}}_{\mathcal{Y}}(\kappa)(B_{\mathcal{X}}\times B_{\mathcal{Y}}) \coloneqq \int_{x\in B_{\mathcal{X}}} \kappa(x)(B_{\mathcal{Y}})dp\,,$$

with $\Gamma((\mathcal{X},\phi),(\mathcal{Y},\psi)) \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ the typed couplings associated to the marginals $(\mathcal{X},\phi),(\mathcal{Y},\psi)$. The second is defined as its inverse operation, and it is decomposing a joint probability along a fixed marginal distribution (aka, disintegrating it), i.e.,

$$\begin{aligned} &D^{\mathcal{X}}_{\mathcal{Y}}\colon \Gamma((\mathcal{X},\phi),(\mathcal{Y},\psi)) \to \mathrm{Hom}_{\mathtt{Krn}}((\mathcal{X},\phi),(\mathcal{Y},\psi))\\ &\text{s.t.}\quad D^{\mathcal{X}}_{\mathcal{Y}}(\gamma) \coloneqq \mathtt{P}(\pi_Y)\circ\pi_X^{\dagger},\ \gamma\in\Gamma((\mathcal{X},\phi),(\mathcal{Y},\psi)),\\ &\text{and}\quad \gamma(B_{\mathcal{X}}\times B_{\mathcal{Y}}) \coloneqq \int_{x\in B_{\mathcal{X}}} D^{\mathcal{X}}_{\mathcal{Y}}(\gamma)(x)(B_{\mathcal{Y}})dp\,, \end{aligned}$$

with $(\cdot)^{\dagger}$: adjoint operator. As one is the inverse of the other, they are both obviously bijective and the one-to-one correspondence between typed kernels and couplings is proved. Hence, we formally define the Bayesian inverse as in the following:

Definition C.3. *The Bayesian inverse of a typed kernel κ from $(\mathcal{X},\phi)$ to $(\mathcal{Y},\psi)$, is defined as*

$$(\cdot)^{\dagger}:\kappa\mapsto\kappa^{\dagger}\coloneqq\left(D^{\mathcal{Y}}_{\mathcal{X}}\circ\mathtt{P}(\pi_Y\times\pi_X)\circ\alpha^{\mathcal{X}}_{\mathcal{Y}}\right)(\kappa)\,,$$

with $\mathtt{P}(\pi_Y\times\pi_X):\Gamma((\mathcal{X},\phi),(\mathcal{Y},\psi))\to\Gamma((\mathcal{Y},\psi),(\mathcal{X},\phi))$ being the permutation map.

As the Bayesian inverse has been defined as a bijection between $\mathrm{Hom}_{\mathtt{Krn}}((\mathcal{X},\phi),(\mathcal{Y},\psi))$ and $\mathrm{Hom}_{\mathtt{Krn}}((\mathcal{Y},\psi),(\mathcal{X},\phi))$, it is always guaranteed to exist in this setting.

Proposition C.4 (Bayesian Inversion Theorem)**.** *The Bayesian inverse of a typed kernel κ from $(\mathcal{X},\phi)$ to $(\mathcal{Y},\psi)$ exists and is equivalently one of the following objects:*

1. $\kappa^{\dagger}:(\mathcal{Y},\psi)\to(\mathcal{X},\phi)\in\mathtt{Krn}$ *when κ is seen as element of* `Krn`*, such that $(\kappa^{\dagger}\circ_{\mathtt{G}}\kappa)\circ_{\mathtt{G}}\psi=\delta_Y\circ_{\mathtt{G}}\psi$ and $(\kappa\circ\kappa^{\dagger})\circ_{\mathtt{G}}\phi=\delta_X\circ_{\mathtt{G}}\phi$;*

2. $\kappa^{\dagger}:Y\rightsquigarrow X\in\mathcal{M}(\mathcal{Y},\mathcal{X})$ *when κ is seen as element of $\mathcal{M}(\mathcal{X},\mathcal{Y})$, such that $(\kappa^{\dagger}\circ_{\mathtt{G}}\kappa)\circ_{\mathtt{G}}\psi\equiv_R\delta_{\mathcal{Y}}\circ_{\mathtt{G}}\psi$ and $(\kappa\circ_{\mathtt{G}}\kappa^{\dagger})\circ_{\mathtt{G}}\phi\equiv_R\delta_{\mathcal{X}}\circ_{\mathtt{G}}\phi$.*

Here, $\delta_{(\cdot)}$ indicates the identity kernel on the set $(\cdot)$, induced by the Dirac delta distribution.

Proof. The statement in (1) is a direct consequence of Dahlqvist et al. (2016).As for (2), we are only using Def. C.1. □

Remark C.5 *We can understand the Bayesian inverse of a corruption kernel $\kappa\in\mathcal{M}(\mathcal{Z},\mathcal{Z})$ from $(\mathcal{Z},\Sigma(\mathcal{Z}),\phi)$ to $(\mathcal{Z},\Sigma(\mathcal{Z}),\tilde{\phi})$ that distorts $\tilde{P}(\mathcal{A})=\int_{\mathcal{A}}\int_{\mathcal{Z}}\kappa(z,d\tilde{z})\,\phi(dz)\ \ \forall A\in\mathcal{Z}$ as a Markov*

kernel $\kappa^\dagger \in \mathcal{M}(\mathcal{Z}, \mathcal{Z})$ *satisfying*

$$\int_{\mathcal{B}} \kappa(z, \mathcal{A})\, \phi(dz) = \int_{\mathcal{A}} \kappa^\dagger(\tilde{z}, \mathcal{B})\, \tilde{\phi}(d\tilde{z}) \quad \forall \mathcal{A} \in \Sigma(\mathcal{Z}),\, \mathcal{B} \in \Sigma(\mathcal{Z}).$$

This formulation extends the discrete Bayes' rule $P(\tilde{z} \mid z)\phi(z) = P(z \mid \tilde{z})P(\tilde{z})\, \forall z, \tilde{z} \in \mathcal{Z}$. *Hence, in the discrete case, the Bayesian inverse always exists and can be expressed as*

$$\kappa^\dagger(z \mid \tilde{z}) := \frac{\phi(z)\kappa(\tilde{z} \mid z)}{\tilde{\phi}(\tilde{z})} \text{ for } z, \tilde{z} \in \mathcal{Z} \text{ with } \tilde{\phi}(\tilde{z}) \neq 0.$$

This formula ensures the uniqueness of $\kappa^\dagger$ *within the support of* $\tilde{\phi}$*, as all components are unique. However, outside the support when* $\tilde{\phi}$ *is zero, the uniqueness may not hold, requiring a non-fixed value for* $\tilde{z} \in \mathcal{Z}$ *where* $\tilde{\phi}(\tilde{z}) = 0$.

In the continuous case, the Bayesian inverse may not exist. To ensure $\kappa^\dagger \in \mathcal{M}(\mathcal{Z}, \mathcal{Z})$ *is well-defined, it must satisfy the conditions of being a Markov kernel, as defined in Def. 2.14, where the mapping* $\tilde{z} \to \kappa^\dagger(\tilde{z}, \mathcal{B})$ *is* $\Sigma(\mathcal{Z})$*-measurable for every set* $\mathcal{B} \in \Sigma(\mathcal{Z})$*, and the mapping* $\mathcal{B} \to \kappa^\dagger(\tilde{z}, \mathcal{B})$ *is a probability measure on* $(\mathcal{Z}, \Sigma(\mathcal{Z}))$ *for every* $\tilde{z} \in \mathcal{Z}$, for the standard Borel space $(\mathcal{Z}, \Sigma(\mathcal{Z}))$. *Under this condition, the Bayesian inverse always exists, and it is uniquely defined within the support of* $\tilde{\phi}$*, where uniqueness is represented by an equivalence class of kernels that are* $\tilde{\phi}$*-a.s. equal.*

C.2.1 Exhaustiveness of Markovian Corruption

As we noticed in Appendix A, Markov kernels are not the only possibility for modeling corruption, but we proved that given a clean and corrupted space we can always find a Markov kernel that connects the two distributions. In particular, we define the operations $\alpha_{\mathbf{y}}^{\mathcal{X}}$ and $D_{\mathbf{y}}^{\mathcal{X}}$ for typed kernels, where one is the inverse of the other by construction (Appendix C.2). They are the operations representing the bijection between the space of Markov kernels typed for ϕ, ψ and the space of couplings with marginals ϕ, ψ. Hence, they are proving that *for each couple of probability spaces, there exists a Markov kernel sending one into the other corresponding to a possible associated coupling.*

Appendix D

Proofs for Data Processing Equalities

Recall that π_Y is a prior distribution on $\mathcal{Y}$, and the notation $\dot{\kappa}(X)$ stands for the Markov transition $\dot{\kappa}$ evaluated on the random variable X, e.g., $\dot{E}(Y)$, $\dot{E}^\dagger(X)$, while its deterministic counterpart is the evaluation $\dot{\kappa}(x)$. The kernel δ denotes a kernel equal to the Dirac delta from $(\mathcal{Z}, \Sigma(\mathcal{Z}))$ to $(\mathcal{Z}, \Sigma(\mathcal{Z}))$, i.e., the identity kernel.

In the proofs we will use a continuous notation for measures on $\mathcal{Y}$, for the sake of simplicity and homogeneity. However, notice that all the λ kernels are actually (parameterized) stochastic matrices, i.e., $\lambda = [\lambda_{\tilde{y}y}]$ where $\lambda_{\tilde{y}y} = p(\tilde{Y} = \tilde{y} \mid Y = y)$ for simple corruptions, and $\lambda_{\tilde{y}y}(x) = p(\tilde{Y} = \tilde{y} \mid Y = y, X = x)$ for dependent corruptions. Note that both y and $\tilde{y}$ range in $\mathcal{Y}$, and thus they are square matrices. In Theorem 5.5 and Lemma D.2, the kernel λ acting on the function $\ell \circ \mathcal{H}$ is actually the transpose of the stochastic matrix λ:

$$\begin{aligned}\hat{\lambda}(\ell_y \circ h)(x) &:= \sum_{\tilde{y} \in \mathcal{Y}} \dot{\lambda}(x, y)(d\tilde{y})\, \ell(\dot{h}(x), \tilde{y}) \\ &= \sum_{\tilde{y} \in \mathcal{Y}} \lambda^\top_{\tilde{y}y}(x)\, \ell_{\tilde{y}}(\dot{h}(x)) =: \lambda^\top(x) \cdot (\ell_y \circ h)(x)\,.\end{aligned}$$

Below, Theorems 5.4 and 5.5 are proved based on the Lemmas concerning **BR** changes under dependent $\mathcal{X}$ and $\mathcal{Y}$ corruptions, respectively.

Lemma D.1 (Attribute Corruption)**.** *Let ℓ be a bounded loss. Consider the learning problem $(\ell, \mathcal{H}, \phi)$ and $E \colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y. Let $\tau \otimes \delta$ be a corruption acting on this problem, with $\tau \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X})$ and $\delta \in \mathcal{M}(\mathcal{Y}, \mathcal{Y})$. Then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\delta})\phi\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big((E \circ_{\mathcal{X}} \tau) \otimes \delta\big)\Big) \equiv_{\mathbf{BR}} \Big(\hat{\tau}(\ell \circ \mathcal{H}), \pi_Y \times E\Big)\,.$$

Moreover, if $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$, we have

$$\Big(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\delta})\phi\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big((E \circ \tau) \otimes \delta\big)\Big) \equiv_{\text{BR}} \Big(\hat{\tau}(\ell \circ \mathcal{H}), \pi_Y \times E\Big)\,.$$

Proof. Let $\mathcal{A} \in \Sigma(\mathcal{X} \times \mathcal{Y})$, and π_y be the y-th entry of the π_Y probability vector. By definition of all the objects involved, the action of $\tau \otimes \delta$ on ϕ is

$$\begin{aligned}
\tilde{\phi}(\mathcal{A}) &= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \dot{\tau}(x,y)(d\tilde{x})\, \delta_y(d\tilde{y})\, \phi(dx,dy) \\
&= \int_{(\tilde{x},\mathbf{y})\in\mathcal{A}} \left[\sum_{y\in\mathcal{Y}} \left(\int_{x\in\mathcal{X}} \dot{\tau}(x,y)(d\tilde{x})\, \dot{E}(y)(dx)\right) \delta_y(d\tilde{y})\, \pi_y\right] \\
&= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\sum_{y\in\mathcal{Y}} (E \circ_{\mathcal{X}} \tau)_y(d\tilde{x})\, \delta_y(d\tilde{y})\, \pi_y\right] = \left[\pi_y \circ \big((E \circ_{\mathcal{X}} \tau) \otimes \delta\big)\right](\mathcal{A})\,.
\end{aligned}$$

We can hence rewrite the risk w.r.t. $\tilde{\phi} \coloneqq \pi_y \circ [(E \circ_{\mathcal{X}} \tau) \otimes \delta]$ as

$$\begin{aligned}
\mathbb{E}_{(\tilde{X},\tilde{Y})\sim\tilde{\phi}}\left[\ell(\dot{h}(\tilde{X}),\tilde{Y})\right] &= \sum_{\substack{y\in\mathcal{Y},\\ \tilde{y}\in\mathcal{Y}}} \left[\int_{\tilde{x}\in\mathcal{X}} \ell(\dot{h}(\tilde{x}),\tilde{y}) \left(\int_{x\in\mathcal{X}} \dot{\tau}(x,y)(d\tilde{x})\, \dot{E}(y)(dx)\right)\right] \delta_y(d\tilde{y})\, \pi_y \\
&= \sum_{y\in\mathcal{Y}} \left[\int_{x\in\mathcal{X}} \left(\int_{\tilde{x}\in\mathcal{X}} (\hat{\delta}\ell_y \circ h)(\tilde{x})\, \dot{\tau}(x,y)(d\tilde{x})\right) \dot{E}(y)(dx)\right] \pi_y \\
&= \sum_{y\in\mathcal{Y}} \left[\int_{x\in\mathcal{X}} [\hat{\tau}(\hat{\delta}\ell_y \circ h)](x,y)\, \dot{E}(y)(dx)\right] \pi_y \qquad \text{(D.1)} \\
&= \mathbb{E}_{(X,Y)\sim(\pi_Y\times E)}\left[\big(\hat{\tau}(\hat{\delta}\ell_Y \circ h)\big)(X,Y)\right]\,.
\end{aligned}$$

Let $\tilde{E}_y(d\tilde{x}) \coloneqq (E \circ_{\mathcal{X}} \tau)_y(d\tilde{x})$. We have that the associated **BR** is

$$\begin{aligned}
\min_{h\in\mathcal{H}} \mathbb{E}_{(\tilde{X},\tilde{Y})\sim\pi_Y\circ[(E\circ_{\mathcal{X}}\tau)\otimes\delta]}\left[\ell(\dot{h}(\tilde{X}),\tilde{Y})\right] &= \min_{h\in\mathcal{H}} \mathbb{E}_{\tilde{Y}\sim\pi_Y} \mathbb{E}_{\tilde{X}\sim\tilde{E}_{\tilde{Y}}}\left[\ell(\dot{h}(\tilde{X}),\tilde{Y})\right] \\
&= \mathbf{BR}_{\ell\circ\mathcal{H}}[\pi_Y \circ (\tilde{E} \otimes \delta)]\,,
\end{aligned}$$

$$\begin{aligned}
\min_{h\in\mathcal{H}} \mathbb{E}_{(X,Y)\sim(\pi_Y\times E)}\left[\big(\hat{\tau}(\hat{\delta}\ell_Y \circ h)\big)(X,Y)\right] &= \min_{f\in\hat{\tau}(\ell\circ\mathcal{H})} \mathbb{E}_{(X,Y)\sim(\pi_Y\times E)}\left[f(X,Y)\right] \\
&= \mathbf{BR}_{\hat{\tau}(\ell\circ\mathcal{H})}[\pi_Y \times E]\,, \qquad \text{(D.2)}
\end{aligned}$$

which are equal given the previous computations. We have defined and used in Eq. (D.2) that $f(x,y) \coloneqq [\hat{\tau}(\hat{\delta}\ell_y \circ h)](x,y), h \in \mathcal{H}$. Such functions are the ones populating the attainable prediction set $\hat{\tau}(\ell \circ \mathcal{H})$, denoting that τ acts on the composition of the loss and model class while δ only acts on ℓ and leaves it unchanged. If τ is simple, then the equations

from Eq. (D.1) lead to a slightly different model class:

$$\begin{aligned}\mathbf{BR}_{\ell\circ\mathcal{H}}\left[\pi_Y \circ \left((E\circ\tau)\otimes\delta\right)\right] &= \min_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} \left[\int_{x\in\mathcal{X}} [\hat{\tau}(\hat{\delta}\ell_y \circ h)](x)\, \dot{E}(y)(dx)\right] \pi_y \\ &= \min_{f\in\hat{\tau}(\ell\circ\mathcal{H})} \sum_{y\in\mathcal{Y}} \left[\int_{x\in\mathcal{X}} f(x,y)\, \dot{E}(y)(dx)\right] \pi_y = \mathbf{BR}_{\hat{\tau}(\ell\circ\mathcal{H})}(\pi_Y \times E)\,.\end{aligned}$$

□

Theorem (2-dependent τ, simple λ, Theorem 5.4)**.** *Let ℓ be a bounded loss. Consider the learning problem $(\ell,\mathcal{H},\phi)$ and $E\colon \mathcal{Y}\rightsquigarrow\mathcal{X}$ its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y. Let $(\tau\colon \mathcal{X}\times\mathcal{Y}\rightsquigarrow\mathcal{X})\otimes(\lambda\colon Y\rightsquigarrow\mathcal{Y})$ be a corruption acting on this problem, then, we obtain*

$$\left(\ell\circ\mathcal{H}, (\check{\tau}\otimes\check{\lambda})\phi\right) = \left(\ell\circ\mathcal{H}, \pi_Y\circ\left((E\circ_{\mathcal{X}}\tau)\otimes\lambda\right)\right) \equiv_{\text{BR}} \left(\hat{\tau}(\hat{\lambda}\ell\circ\mathcal{H}), \pi_Y\times E\right).$$

The functions contained in the new attainable prediction set are defined as

$$\hat{\tau}(\hat{\lambda}\ell\circ\mathcal{H}) \coloneqq \{(x,y)\mapsto[\hat{\tau}(\hat{\lambda}\ell_y\circ h)](x,y), h\in\mathcal{H}\}\,.$$

Proof. With this corruption formulation, we can replicate the proof of Lemma D.1 up to Eq. (D.1) by simply plugging in λ instead of δ. Therefore, we obtain the thesis. □

We remark that in this case $\tilde{\phi}\neq\pi_Y\times\tilde{E}$ with $\tilde{E}_y(d\tilde{x}) \coloneqq (E\circ_{\mathcal{X}}\tau)_y(d\tilde{x})$, i.e., the corrupted experiment is not given by the sole action of τ, but also by the influence of λ. That is clarified further by corruption formula $\tilde{\phi} = \pi_y\circ[(E\circ_{\mathcal{X}}\tau)\otimes\lambda]$. We conclude that, in this more general case, it does not make sense to distinguish the effect of corruption on E and π.

Lemma D.2 (Label Corruption)**.** *Let ℓ be a bounded loss. Consider the learning problem $(\ell,\mathcal{H},\phi)$ and $E^{\dagger}\colon\mathcal{X}\rightsquigarrow\mathcal{Y}$ its associated posterior such that $\phi = \pi_X\times E^{\dagger}$ for a suitable π_X. Let $\delta\otimes\lambda$ be a corruption acting on this problem, with $\lambda\in\mathcal{M}(\mathcal{X}\times\mathcal{Y},\mathcal{Y})$ and $\delta\in\mathcal{M}(\mathcal{X},\mathcal{X})$. Then, we obtain*

$$\left(\ell\circ\mathcal{H}, (\check{\delta}\otimes\check{\lambda})\phi\right) = \left(\ell\circ\mathcal{H}, \pi_X\circ\left(\delta\otimes(E^{\dagger}\circ_{\mathcal{Y}}\lambda)\right)\right) \equiv_{\text{BR}} \left(\lambda\ell\circ\mathcal{H}, \pi_X\times E^{\dagger}\right).$$

Moreover, if $\lambda\in\mathcal{M}(\mathcal{Y},\mathcal{Y})$, we have

$$\left(\ell\circ\mathcal{H}, (\check{\delta}\otimes\check{\lambda})\phi\right) = \left(\ell\circ\mathcal{H}, \pi_X\circ\left(\delta\otimes(E^{\dagger}\circ\lambda)\right)\right) \equiv_{\text{BR}} \left(\hat{\lambda}\ell\circ\mathcal{H}, \pi_X\times E^{\dagger}\right).$$

Proof. Let $\mathcal{A} \in \Sigma(\mathcal{X} \times \mathcal{Y})$. By definition of all the objects involved, the action of $\tau \otimes \delta$ on ϕ is

$$\begin{aligned}\tilde{\phi}(\mathcal{A}) &= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\int_{x\in\mathcal{X}} \left(\sum_{y\in\mathcal{Y}} \dot{\lambda}(x,y)(d\tilde{y}) \dot{E}^{\dagger}(x)(dy) \right) \delta_x(d\tilde{x})\, \pi_X(dx) \right] \\ &= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\int_{x\in\mathcal{X}} (E^{\dagger} \circ_{\mathcal{Y}} \lambda)_x(d\tilde{y})\, \delta_x(d\tilde{x})\, \pi_X(dx) \right] = \left[\pi_X \circ \left((E^{\dagger} \circ_{\mathcal{Y}} \lambda) \otimes \delta \right) \right](\mathcal{A}) \,.\end{aligned}$$

We can hence rewrite the risk w.r.t. $\tilde{\phi} \coloneqq \pi_X \circ [(E^{\dagger} \circ_{\mathcal{Y}} \lambda) \otimes \delta]$ as

$$\begin{aligned}\mathbb{E}_{(\tilde{X},\tilde{Y})\sim\tilde{\phi}} \left[\ell(\dot{h}(\tilde{X}), \tilde{Y}) \right] &= \int_{\substack{x\in\mathcal{X},\\ \tilde{x}\in\mathcal{X}}} \left[\sum_{\tilde{y}\in\mathcal{Y}} \ell(\dot{h}(\tilde{x}), \tilde{y}) \left(\sum_{y\in\mathcal{Y}} \lambda(x,y,d\tilde{y})\, \dot{E}^{\dagger}(x)(dy) \right) \right] \delta_x(d\tilde{x})\, \pi_X(dx) \\ &= \int_{\substack{x\in\mathcal{X},\\ \tilde{x}\in\mathcal{X}}} \left[\sum_{y\in\mathcal{Y}} \left(\sum_{\tilde{y}\in\mathcal{Y}} \ell(\dot{h}(\tilde{x}), \tilde{y})\, \lambda(x,y,d\tilde{y}) \right) \dot{E}^{\dagger}(x)(dy) \right] \delta_x(d\tilde{x})\, \pi_X(dx) \\ &= \int_{x\in\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} \left(\int_{\tilde{x}\in\mathcal{X}} (\hat{\lambda}\ell)(h_{\tilde{x}}, x, y) \delta_x(d\tilde{x}) \right) \dot{E}^{\dagger}(x)(dy) \right] \pi_X(dx) \\ &= \int_{x\in\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} \left((\hat{\lambda}\ell)(\dot{h}(x), x, y) \right) \dot{E}^{\dagger}(x)(dy) \right] \pi_X(dx) \qquad \text{(D.3)} \\ &= \mathbb{E}_{(X,Y)\sim(\pi_Y\times E)} \left[(\hat{\lambda}\ell)(h_X, X, Y) \right] = \mathbb{E}_{(X,Y)\sim(\pi_Y\times E)} \left[(\hat{\lambda}\ell_{(X,Y)} \circ h)(X) \right] \,.\end{aligned}$$

Similarly to the proof provided for Lemma D.1, we can switch to **BR** and obtain

$$\mathbf{BR}_{\ell\circ\mathcal{H}}[\pi_X \circ \left((E^{\dagger} \circ_{\mathcal{Y}} \lambda) \otimes \delta \right)] = \mathbf{BR}_{\hat{\lambda}\ell\circ\mathcal{H}}(\pi_X \times E^{\dagger}) \,,$$

with functions $\hat{\lambda}\ell(\dot{h}(x), x, y) = (\hat{\lambda}\ell_{(x,y)} \circ h)(x) \in \hat{\lambda}\ell \circ \mathcal{H}$. If λ is simple and $\delta \in \mathcal{M}(\mathcal{X}, \mathcal{X})$, then Eq. (D.3) leads to a simpler model class:

$$\begin{aligned}\mathbf{BR}_{\ell\circ\mathcal{H}}[\pi_X \circ \left((E^{\dagger} \circ \lambda) \otimes \delta \right)] &= \min_{h\in\mathcal{H}} \int_{x\in\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} (\hat{\lambda}\ell)(\dot{h}(x), y) \dot{E}^{\dagger}(x)(dy) \right] \pi_X(dx) \\ &= \min_{f\in\hat{\lambda}(\ell\circ\mathcal{H})} \int_{x\in\mathcal{X}} \left[\sum_{y\in\mathcal{Y}} f(x,y) \dot{E}^{\dagger}(x)(dy) \right] \pi_X(dx) \\ &= \mathbf{BR}_{\hat{\lambda}(\ell\circ\mathcal{H})}(\pi_X \times E^{\dagger}) \,.\end{aligned}$$

□

Theorem (simple τ, 2-dependent λ, Theorem 5.5)**.** *Let ℓ be a bounded loss. Consider the learning problem $(\ell, \mathcal{H}, \phi)$ and $E^{\dagger} \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ its associated posterior such that $\phi = \pi_X \times E^{\dagger}$ for a suitable π_X. Let $(\tau \colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on this problem, then,*

we obtain

$$\left(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\lambda})\phi\right) = \left(\ell \circ \mathcal{H}, \pi_X \circ \left(\tau \otimes (E^\dagger \circ_{\mathcal{Y}} \lambda)\right)\right) \equiv_{\text{BR}} \left(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \pi_X \times E^\dagger\right).$$

The functions contained in the new attainable prediction set are defined as

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) \coloneqq \{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](x), h \in \mathcal{H}\}.$$

Proof. With this corruption formulation, we can replicate the proof of Lemma D.2 up to Eq. (D.3) by simply plugging in τ instead of $\delta \in \mathcal{M}(\mathcal{X}, \mathcal{X})$. Therefore, we obtain the thesis. □

Theorem (1-dependent and a 2-dependent, Theorem 5.6)**.** *Let ℓ be a bounded loss. Consider the clean learning problem $(\ell, \mathcal{H}, \phi)$, $E \colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y, and $E^\dagger \colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ its associated posterior such that $\phi = \pi_X \times E^\dagger$ for a suitable π_X.*

1. *Let $(\tau \colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on the problem, then, we obtain*

$$\left(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\lambda})\phi\right) = \left(\ell \circ \mathcal{H}, \pi_Y \circ \left(\tau \otimes (E \circ_{\mathcal{X}} \lambda)\right)\right) \equiv_{\text{BR}} \left(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \pi_Y \times E\right).$$

   The functions contained in the new attainable prediction set are defined as

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H})\{(x, y) \mapsto \hat{\tau}[\hat{\lambda}\ell_{(x,y)} \circ h](y), h \in \mathcal{H}\}.$$

2. *Let $(\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \rightsquigarrow \mathcal{Y})$ be a corruption acting on the problem, then, we obtain*

$$\left(\ell \circ \mathcal{H}, \check{\tau}(\otimes\check{\lambda})\phi\right) = \left(\ell \circ \mathcal{H}, \pi_X \circ \left((E^\dagger \circ_{\mathcal{Y}} \tau) \otimes \lambda\right)\right) \equiv_{\text{BR}} \left(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \pi_X \times E^\dagger\right).$$

   The functions contained in the new attainable prediction set are defined as

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H})\{(x, y) \mapsto \hat{\tau}[\hat{\lambda}\ell_x \circ h](x, y), h \in \mathcal{H}\}.$$

Proof. Consider point 1 and let $A \in \mathcal{X} \times \mathcal{Y}$. By definition of all the objects involved, the action of $\tau \otimes \lambda$ on ϕ is

$$\begin{aligned}
\tilde{\phi}(\mathcal{A}) &= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\sum_{y\in\mathcal{Y}} \left(\int_{x\in\mathcal{X}} \dot{\lambda}(x, y)(d\tilde{y})\dot{E}(y)(dx)\right) \tau(y, d\tilde{x})\pi_y\right] \\
&= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\sum_{y\in\mathcal{Y}} (E \circ_{\mathcal{X}} \lambda)(y, d\tilde{y})\tau(y, d\tilde{x})\, \pi_y\right] \\
&= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\sum_{y\in\mathcal{Y}} \left(\tau \otimes (E \circ_{\mathcal{X}} \lambda)\right)(y, d\tilde{x}, d\tilde{y})\, \pi_y\right] = \left[\pi_Y \times \left(\tau \otimes (E \circ_{\mathcal{X}} \lambda)\right)\right](\mathcal{A}).
\end{aligned}$$

We can then write the associated risk w.r.t. $\tilde{\phi} \coloneqq \pi_Y \times \big(\tau \otimes (E \circ_{\mathcal{X}} \lambda)\big)$ as

$$\begin{aligned}
\mathbb{E}_{(\tilde{X},\tilde{Y})\sim\tilde{\phi}}[\ell(\dot{h}(\tilde{X}),\tilde{Y})] &= \sum_{\substack{y\in\mathcal{Y},\\ \tilde{y}\in\mathcal{Y}}} \left[\int_{\tilde{x}\in\mathcal{X}} \ell(\dot{h}(\tilde{x}),\tilde{y}) \left(\int_{x\in\mathcal{X}} \lambda(x,y,d\tilde{y})\, \dot{E}(y)(dx)\right)\right] \tau(y,d\tilde{x})\, \pi_y \\
&= \sum_{y\in\mathcal{Y}} \left[\int_{x\in\mathcal{X}} \left(\int_{\tilde{x}\in\mathcal{X}} \left(\sum_{\tilde{y}\in\mathcal{Y}} \lambda(x,y,d\tilde{y})\, \ell(\dot{h}(\tilde{x}),\tilde{y})\right) \tau(y,d\tilde{x})\right) \dot{E}(y)(dx)\right] \pi_y \\
&= \sum_{y\in\mathcal{Y}} \left[\int_{x\in\mathcal{X}} \left(\int_{\tilde{x}\in\mathcal{X}} (\hat{\lambda}\ell_{(x,y)} \circ h)(\tilde{x})\, \tau(y,d\tilde{x})\right) \dot{E}(y)(dx)\right] \pi_y \\
&= \sum_{y\in\mathcal{Y}} \left[\int_{x\in\mathcal{X}} [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](y)\, \dot{E}(y)(dx)\right] \pi_y \\
&= \mathbb{E}_{(X,Y)\sim(\pi_Y\times E)} \left[\big(\hat{\tau}(\hat{\lambda}\ell_{(X,Y)} \circ h)\big)(Y)\right] ,
\end{aligned}$$

which proves the thesis when minimizing over $h \in \mathcal{H}$. For proving point 2, we first rewrite the action of $\tau \otimes \lambda$ on ϕ as

$$\begin{aligned}
\tilde{\phi}(\mathcal{A}) &= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\int_{x\in\mathcal{X}} \lambda(x,d\tilde{y}) \left(\sum_{y\in\mathcal{Y}} \tau(x,y,d\tilde{x})\, E^{\dagger}(x,dy)\right) \pi_X(dx)\right] \\
&= \int_{(\tilde{x},\tilde{y})\in\mathcal{A}} \left[\int_{x\in\mathcal{X}} \lambda(x,d\tilde{y})\, (E^{\dagger} \circ_{\mathcal{Y}} \tau)(x,d\tilde{x})\, \pi_X(dx)\right] = \left[\pi_X \times \big(\lambda \otimes (E^{\dagger} \circ_{\mathcal{Y}} \tau)\big)\right](\mathcal{A}) ,
\end{aligned}$$

and repeat a similar argument but for the $E^{\dagger}$ kernel. We find a minimization space of functions $f(x,y) \coloneqq \tau[\lambda\ell_x \circ h](x,y)$. Thus, we obtain the thesis. □

Corollary (1-dependent τ and λ, Corollary 5.7)**.** *Let ℓ be a bounded loss. Consider the clean learning problem $(\ell,\mathcal{H},\phi)$, $E\colon \mathcal{Y} \rightsquigarrow \mathcal{X}$ its associated experiment such that $\phi = \pi_Y \times E$ for a suitable π_Y, and $E^{\dagger}\colon \mathcal{X} \rightsquigarrow \mathcal{Y}$ its associated posterior such that $\phi = \pi_X \times E^{\dagger}$ for a suitable π_X. Let $(\tau\colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \rightsquigarrow \mathcal{Y})$ be a corruption acting on the problem, then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\check{\tau} \otimes \check{\lambda})\phi\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big(\tau \otimes (E \circ \lambda)\big)\Big) \equiv_{\text{BR}} \Big(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \pi_Y \times E\Big) .$$

or, equivalently,

$$\Big(\ell \circ \mathcal{H}, (\pi_X \times E^{\dagger}) \circ (\tau \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_X \circ \big((E^{\dagger} \circ \tau) \otimes \lambda\big)\Big) \equiv_{\text{BR}} \Big(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \pi_X \times E^{\dagger}\Big) .$$

The functions contained in the new attainable prediction set are defined as

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) \coloneqq \{(x,y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_x \circ h)](y), h \in \mathcal{H}\} .$$

Proof. We can replicate the proof of Theorem 5.6 by simply substituting $\lambda(x,d\tilde{y})$ in place of $\lambda(x,y,d\tilde{y})$ in the first point, and $\tau(y,d\tilde{x})$ for $\tau(x,y,d\tilde{x})$ in the second point. We then in both cases obtain functions $f(x,y) \coloneqq \hat{\tau}[(\hat{\lambda}\ell)_x \circ h](y)$, i.e. comparing a point x with a kernel on $\Delta_{\text{ca}}(\mathcal{X})$ parameterized by y. Therefore, we obtain the thesis. □

Theorem (2-dependent κ and λ, Theorem 5.8)**.** *Let ℓ be a bounded loss. Consider the clean learning problem $(\ell, \mathcal{H}, \phi)$, and let $(\tau \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda \colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$ be a corruption acting on the problem. Then:*

1. *the action of such corruption on the joint probability ϕ is equivalent to the one of the non-decomposed joint corruption;*

2. *the action on the attainable prediction set $\ell \circ \mathcal{H}$ induces the following* **BR***-equivalence*

$$\left(\ell, \mathcal{H}, (\check{\tau} \otimes \check{\lambda})\phi\right) \equiv_{\text{BR}} \left(\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}), \phi\right);$$

3. *the functions contained in the new attainable prediction set are defined as*

$$\hat{\tau}(\hat{\lambda}\ell \circ \mathcal{H}) \coloneqq \{(x, y) \mapsto [\hat{\tau}(\hat{\lambda}\ell_{(x,y)} \circ h)](x, y), h \in \mathcal{H}\}.$$

Proof. Let $\mathcal{A} \in \Sigma(\mathcal{X} \times \mathcal{Y})$. By definition of all the objects involved, the action of $\tau \otimes \lambda$ on ϕ is

$$\begin{aligned}
\tilde{\phi}(\mathcal{A}) &= \int_{(\tilde{x},\tilde{y}) \in \mathcal{A}} \left[\sum_{y \in \mathcal{Y}} \left(\int_{x \in \mathcal{X}} \dot{\tau}(x, y)(d\tilde{x})\, \dot{\lambda}(x, y)(d\tilde{y})\, \dot{E}(y)(dx) \right) \pi_y \right] \\
&= \int_{(\tilde{x},\tilde{y}) \in \mathcal{A}} \left[\int_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}} \dot{\tau}(x, y)(d\tilde{x})\, \dot{\lambda}(x, y)(d\tilde{y})\, \dot{E}^{\dagger}(x)(dy) \right) \pi_X(dx) \right].
\end{aligned}$$

In both the formulations above, obtained by factorizing the joint probability ϕ in two different ways, we cannot isolate the action of one between λ and τ on $E^{\dagger}$ or E. That is, because of the dependence of λ and τ on the couple (x, y), and because the action of a kernel on a probability via a combination of Definitions 2.24, 2.25 and 2.28 requires sequential integration. This concludes point 1.

As for point 2, we now want to consider the action on functions. This uses integration w.r.t. the corrupted variables $(\tilde{x}, \tilde{y})$, and therefore allows sequential integration. We have that the associate risk is equal to

$$\begin{aligned}
\mathbb{E}_{(\tilde{X},\tilde{Y}) \sim \tilde{\phi}}[\ell(\dot{h}(\tilde{X}), \tilde{Y})] &= \int_{\tilde{x} \in \mathcal{X}, \tilde{y} \in \mathcal{Y}} \ell(\dot{h}(\tilde{x}), \tilde{y}) \left[\sum_{y \in \mathcal{Y}} \left(\int_{x \in \mathcal{X}} \tau(x, y, d\tilde{x})\, \lambda(x, y, d\tilde{y})\, \dot{E}(y)(dx) \right) \pi_y \right] \\
&= \sum_{y \in \mathcal{Y}} \left[\int_{x \in \mathcal{X}} \left(\int_{\tilde{x} \in \mathcal{X}} \left(\sum_{\tilde{y} \in \mathcal{Y}} \ell(\dot{h}(\tilde{x}), \tilde{y})\, \lambda(x, y, d\tilde{y}) \right) \tau(x, y, d\tilde{x})\, \dot{E}(y)(dx) \right) \right] \pi_y \\
&= \sum_{y \in \mathcal{Y}} \left[\int_{x \in \mathcal{X}} \left(\int_{\tilde{x} \in \mathcal{X}} \hat{\lambda}\ell(\dot{h}(\tilde{x}), x, y)\, \tau(x, y, d\tilde{x})\, \dot{E}(y)(dx) \right) \right] \pi_y \\
&= \sum_{y \in \mathcal{Y}} \left[\int_{x \in \mathcal{X}} \hat{\tau}[\hat{\lambda}\ell_{(x,y)} \circ h](x, y)\, \dot{E}(y)(dx) \right] \pi_y \\
&= \mathbb{E}_{(X,Y) \sim (\pi_Y \times E)} \left[\left(\hat{\tau}(\hat{\lambda}\ell_{(X,Y)} \circ h)\right)(X, Y) \right].
\end{aligned}$$

Following the same reasoning, we can also write

$$\mathbb{E}_{(\tilde{X},\tilde{Y})\sim\tilde{\phi}}[\ell(\dot{h}(\tilde{X}),\tilde{Y})] = \mathbb{E}_{(X,Y)\sim(\pi_X\times E^\dagger)}\left[\left(\hat{\tau}(\hat{\lambda}\ell_{(X,Y)}\circ h)\right)(X,Y)\right] .$$

We prove point 2 and 3 minimizing both the obtained risk equalities w.r.t. $h \in \mathcal{H}$. □

Appendix E

Proof for Generalized Corruption-Corrected Learning with Bayesian Inverse

In this section, we give the proofs for the generalized loss correction theorem stated in § 5.2. Recall that we define the push-forward probability distribution of h through τ as

$$(\dot{h}\#\tau)(\tilde{x})(\mathcal{A}) \coloneqq \tau(\tilde{x}, \dot{h}^{-1}(\mathcal{A})),\ \mathcal{A} \subseteq \Delta_{\mathrm{ca}}(\mathcal{Y}).$$

Theorem (Theorem 5.18)**.** *Let $(\ell, \mathcal{H}, \phi)$ be a clean learning problem with $\ell\colon \Delta_{\mathrm{ca}}(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ being a bounded loss function and such that A1 holds. Let $\kappa^\dagger = \tau \otimes \lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$ be the Markovian cleaning kernel reversing κ, such that $(\hat{\kappa}^\dagger(\ell \circ \mathcal{H}), \check{\kappa}\phi)$ is its associated corrected problem. Then, we can find a generalized loss $\tilde{\ell}\colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ respecting the GCL paradigm. In particular:*

1. *When $\kappa^\dagger$ is of the form $(\tau\colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$, or $(\tau\colon \mathcal{X} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$, or $(\tau\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{Y} \rightsquigarrow \mathcal{Y})$, we have*

   $$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{x}, \tilde{y})}[\hat{\lambda}\ell(u, \tilde{x}, \tilde{y})] \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\,.$$

   When λ is simple, it induces $\hat{\lambda}\ell(u, \tilde{x}, \tilde{y}) = \hat{\lambda}\ell(u, \tilde{y})$. Lastly, we get $(\dot{h}\#\tau)(\tilde{x}, \tilde{y}) = (\dot{h}\#\tau)(\tilde{x})$ when τ is simple. All of the above cases define a corrected loss as $\ell\colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.

2. *When $\kappa^\dagger$ is of the form $(\tau\colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \rightsquigarrow \mathcal{Y})$, we have*

   $$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{y})}[\hat{\lambda}\ell(u, \tilde{x})] \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\,,$$

   with $\ell\colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.

3. *When $\kappa^\dagger$ is of the form $(\tau\colon \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$, or $(\tau\colon \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes$*

$(\lambda : \mathcal{X} \rightsquigarrow \mathcal{Y})$, *we respectively have*

$$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{x}, \tilde{y})] \quad \forall\,(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\,,$$
$$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{x}, \tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{x})] \quad \forall\,(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\,,$$

with $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.

4. *When* $\kappa^\dagger$ *is of the form* $(\tau : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}) \otimes (\lambda : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y})$, *we have*

$$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{x}, \tilde{y})}[\hat{\lambda}\ell\,(u, \tilde{x}, \tilde{y})] \quad \forall\,(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\,,$$

with $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.

Proof. Let us consider a general function in the set $\hat{\kappa}^\dagger(\ell \circ \mathcal{H})$,

$$\hat{\kappa}^\dagger(\ell \circ h) \coloneqq \sum_{y \in Y} \int_{x \in X} \ell(\dot{h}(x), y)\, \kappa^\dagger(\tilde{x}, \tilde{y}, dxdy) = \sum_{y \in Y} \int_{x \in X} \ell(\dot{h}(x), y)\, (\tau \otimes \lambda)(\tilde{x}, \tilde{y}, dxdy)\,.$$

By Lemma 5.17, we are ensured that finding a loss correction formula $\tilde{\ell}$ that respects Eq. (5.13) is enough to ensure that such a loss respects the GCL paradigm. Hence, we consider the functions in the set $\hat{\kappa}^\dagger(\ell \circ \mathcal{H})$ and define an ad hoc $\tilde{\ell}$ depending on how $\kappa^\dagger$ affects $\ell \circ \mathcal{H}$.

We have already seen in Theorem 5.16 how loss correction formulas are identified by a λ that is simple, 1- or 2-dependent. For this reason, we compute step by step the simplest case of attribute corruption (a simple τ) combined with a more complex label corruption (2-dependent λ), and understand how attribute corruption influences the loss correction formulas. All remaining cases can be computed with small variations.

Consider $\kappa^\dagger$ such that $\kappa^\dagger = \tau \otimes \lambda$ with $\tau \in \mathcal{M}(\mathcal{X}, \mathcal{X})$ and $\lambda \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}, \mathcal{Y})$. We can define the loss correction $\tilde{\ell}$ as

$$\begin{aligned}
\tilde{\ell}(h, \tilde{x}, \tilde{y}) &\coloneqq \sum_{y \in \mathcal{Y}} \int_{x \in \mathcal{X}} \ell(\dot{h}(x), y)\, \tau(\tilde{x}, dx)\, \lambda(\tilde{x}, \tilde{y}, dy) \\
&= \sum_{y \in \mathcal{Y}} \int_{x \in \dot{h}(\mathcal{X})} \ell(u, y)\, \tau(\tilde{x}, \dot{h}^{-1}(du))\, \lambda(\tilde{x}, \tilde{y}, dy) \\
&= \int_{x \in \dot{h}(\mathcal{X})} (\hat{\lambda}\ell)_{\tilde{x}}(u, \tilde{y})\, \tau(\tilde{x}, \dot{h}^{-1}(du)),
\end{aligned} \tag{E.1}$$

where $u = u(dy) \in \Delta_{\mathrm{ca}}(\mathcal{Y})$ and $\dot{h}(\mathcal{X}) \coloneqq \{h(x) \,|\, x \in \mathcal{X}\}$. The following equality holds:

$$\mathbb{E}_{u \sim \tau(\tilde{x}, \dot{h}^{-1}(\cdot))}[u] = \int_{x \in \dot{h}^*(\mathcal{X})} u\, \tau(\tilde{x}, (\dot{h})^{-1}(du)) = (\dot{h}\#\tau)(\tilde{x}) \in \Delta_{\mathrm{ca}}(\mathcal{Y}),$$

by definition of $\mathcal{H}$ as a subset of $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ and using the definition of $\dot{h}\#\tau$. We remark that

$\tau(\tilde{x}, (\dot{h})^{-1}(du))$ is then a probability in $\Delta_{\text{ca}}(\Delta_{\text{ca}}(\mathcal{Y}))$. Hence we can rewrite Eq. (E.1) as

$$\tilde{\ell}(h, \tilde{x}, \tilde{y}) \coloneqq \int_{u \in \Delta_{\text{ca}}(\mathcal{Y})} (\hat{\lambda}\ell)_{\tilde{x}}(u, \tilde{y})\, \tau(\tilde{x}, (\dot{h})^{-1}(du)) = \mathbb{E}_{u \sim (\dot{h}\#\tau)(\tilde{x})}[(\hat{\lambda}\ell)(u, \tilde{x}, \tilde{y})].$$

This is the same correction formula found in point 1.

As for more dependent attribute corruptions, i.e. $\tau(\tilde{x}, \tilde{y}, dx)$, the action on the hypothesis will be dependent on $\tilde{y}$. Therefore, we obtain $\tilde{\ell}(h, \tilde{x}, \tilde{y}) = \mathbb{E}_{\tau(\tilde{x}, \tilde{y}, \dot{h}^{-1}(\cdot))}[(\hat{\lambda}\ell)(u, \tilde{x}, \tilde{y})]$, where only the simple label noise can be considered, given the missing result for the **BR** equality in the $D(\tau) = D(\lambda) = \mathcal{X} \times \mathcal{Y}$ case. Hence, for the remaining points 2, 3 and 4, we follow the same procedure deployed in the above, using the action formula of dependent corruptions as described in the proof of Theorems 5.6 and 5.8, and obtain the thesis by fulfilling the GCL requirement in Eq. (5.13). □