Journal of Machine Learning Research 27 (2026) 1-73 Submitted 5/24; Revised 9/25; Published 5/26

Corruptions of Supervised Learning Problems: Typology and Mitigations

Laura Iacovissi LAURA.IACOVISSI@UNI-TUEBINGEN.DE
Tübingen AI Center, University of Tübingen

Nan Lu NAN.LU@UNI-TUEBINGEN.DE
Tübingen AI Center, University of Tübingen

Robert C. Williamson BOB.WILLIAMSON@UNI-TUEBINGEN.DE
Tübingen AI Center, University of Tübingen

Editor: Ambuj Tewari

Abstract

Corruption is notoriously widespread in data collection. Despite extensive research, the existing literature predominantly focuses on specific settings and learning scenarios, lacking a unified view of corruption modelization and mitigation. In this work, we develop a general theory of corruption, which incorporates all modifications to a supervised learning problem, including changes in model class and loss. Focusing on changes to the underlying probability distributions via Markov kernels, our approach leads to three novel opportunities. First, it enables the construction of a novel, provably *exhaustive* corruption framework, distinguishing among different corruption types. This serves to unify existing models and establish a consistent nomenclature. Second, it facilitates a systematic analysis of corruption consequences on learning tasks, by considering Bayes risks in the clean and corrupted scenarios. Notably, while label corruptions affect only the loss function, attribute corruptions additionally influence the hypothesis class. Third, building upon these results, we investigate mitigations for various corruption types. We expand existing loss-correction methods for label corruption to handle dependent corruption types. Our findings highlight the necessity to generalize this classical corruption-corrected learning framework to a new paradigm with weaker requirements to encompass more corruption types. We provide such a paradigm as well as loss correction formulas in the attribute and joint corruption cases.

Keywords: learning theory, Markov kernels, Markovian corruption, noisy data, loss correction

1. Introduction

Machine learning starts with data. The most widespread conception of data defines it 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 understood as identifying and synthesizing patterns embedded in these 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 practice,

corruption regularly occurs in data collection, creating a mismatch between training and test distributions and forcing models to learn from imperfect facts.

We should thus doubt the view of data as static facts, and consider them as a dynamic element of a learning task (Williamson, 2020). In addition to the traditional emphasis on prediction models and loss functions in machine learning, one may focus on the data dynamic itself, so as to understand how different processes may have led us to the observation of certain data, and furthermore, how they 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 the field of machine learning, changes in such dynamic process are often referred to as *distribution shift* or *noisy data*. Here, we adopt a more inclusive term *corruption*, drawing from the computer science literature. Our conceptualization of corruption goes beyond traditional notions: it encompasses all modifications to a learning problem, including changes to the loss function, hypothesis class, or probability distribution from which data are drawn. We interpret corruption not as inherently pejorative, but as a general *modification process*. Whether the corruption is positive, negative, or neutral, depends on the specific context in which it is applied.

A similar stance has been taken in the recent work from Mémoli et al. (2025): they define corruption in an analogous general fashion, while additionally allowing for changes in the attribute and label sets. Their goal is to quantify the distance between learning problems, which they achieve by defining a suitable pseudo-metric. On the other hand, in the present paper we specifically focus on modifications of the probability distribution, aiming to address the lack of understanding of *data as a process* and provide a complete description of this class of corruptions.

Recognizing corruption as a dynamic element of learning has led to considerable research into specific data corruption models (Angluin and Laird, 1988; Zhang et al., 2013; Natarajan et al., 2013; Patrini et al., 2017; Shimodaira, 2000; Quiñonero-Candela et al., 2008; Zhang et al., 2020b). However, these approaches cannot, even in principle, answer questions regarding the comparison of different types of corruption. Inconsistent naming of the same corruption model across different works further slows down progress, and highlights the need for a comprehensive framework.

Whilst there have been attempts to build such a framework, certain limitations persist in terms of homogeneity and exhaustiveness. A famous early endeavor is Quiñonero-Candela et al. (2008), which groups together works about the multi-faceted topic of dataset shift, yet not in a unifying or comprehensive manner. Later on, several studies sought to offer a more homogeneous view of corruption (Moreno-Torres et al., 2012; Kull and Flach, 2014; Sáez, 2022; Subbaswamy et al., 2022); nonetheless, these frameworks typically rely on the assumption of a corruption-invariant marginal or conditional probability. The extent of their exhaustiveness in representing all potential corruption models within their framework is merely conjectured, or left unexplored.

Therefore, the primary objective of this work is to improve the existing understanding of corruption by introducing a novel perspective, while making use of the classical probabilistic approaches. Probability distributions are the only representation of data that will

be used in this work, so the terms "data" and "data distribution" are used interchangeably. This approach allows us to systematically study and compare the possible types of corruption in supervised learning problems, and provides a general framework for analyzing their mitigation.

1.1 Motivations, Approach, and Contributions

We observed in the previous section a continued surge of research papers dedicated to specific models of corruption. In Tab. 1, we provide a non-comprehensive list of such models that can be conceptualized as corruption in our sense. Rather than adding to this already diverse landscape, we propose a taxonomy to systematically organize them. This taxonomy serves as a comprehensive map of probabilistic corruption models, currently absent in the field. Our approach distinguishes itself from the majority of papers in this line, which often propose new corruption models and tailored mitigation algorithms. We, on the other hand, explore their mitigation without the aim of introducing a new algorithm, but for gaining a deeper understanding of theory behind the existing ones. Details about our contributions are summarized in the following.

C1 Understanding Corruptions and Their Types. A common definition of corruption found in the literature is the one of distributional shift. We shape our notion of Markovian corruption inspired by such a concept and making use of Markov kernels. However, attributing failures in learning solely to changes in probability distributions is restrictive. For this reason we broaden the concept of corruption to a general one that includes changes in model class and loss function (Definition 9). Focusing on Markovian corruption (Definition 10), we establish a taxonomy grounded in its dependence on the input and output spaces (Figures 1 and 2). This allows us to uncover commonalities among different models of corruptions (refer to Tab. 2 for the correspondence of Tab. 1 in our taxonomy), thus transcending the diverse terminologies used by different authors. Our resulting framework is proven to be exhaustive for all possible one-step probabilistic corruptions. More generally, we prove that every change in probability distribution can be represented via a one-step Markovian corruption or a non-factorized one (Proposition 17). The statement has interesting consequences in terms of how we think of alternative corruption models, i.e., not within the taxonomy. For instance, arguments may be made for non-probabilistic corruptions that change the probability associated with events in a manner not adhering to probability principles (e.g., Boyd et al. (2023)). In this context, § 3.2 analyses two popular corruption models—selection bias and mutually contaminated distributions—and demonstrates that they are not one-step Markovian in their original definition; however, they have a one-step Markovian representation. For both of them we gain new insights by relating them to our framework.

C2 Consequences of Corruption on Learning Problems. Recently, Williamson and Cranko (2024) deepened the understanding of the relationship between information and Bayes risk of a statistical decision problem, and yielded Information Processing *Equalities* for a certain class of simple corruptions. We connect to this work and relate the Bayes risk of clean and corrupted supervised learning problems through equality results for all corruptions in our taxonomy. Such equalities, illustrated in § 4 (Theorems 22 to 27), effectively prove the equivalence between two learning problems: the former corrupted in a *Markovian fashion*,

Table 1: Examples of models proposed in the literature that capture data corruptions with probabilistic descriptions. Here, X represents the attribute, and Y represents the label. Details and references are included in § C.

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

the latter via a *general corruption* changing only model class and loss function through a Markov kernel. A characteristic of this analysis is its neat avoidance of dependence on specific algorithms, which provides an agnostic means of comparison for corruption types. Such comparison is, in our results, only qualitative, and lays the foundation for future quantitative studies. One of our main findings is the understanding that label corruptions only affects the loss function, while the model class remains untouched by the corruption

kernel; however, for more intricate cases also involving attribute corruptions, both the loss function and the model class are modified.

C3 A Systematic Analysis of Kernel-Based Mitigations. Applying the Bayes risk equalities, we derive in Section § 5 corruption-corrected loss functions for each of the different corruption instances within our framework. We first identify the need of generalizing the concept of classical corruption-corrected learning since it becomes outmoded when considering forms of corruption beyond label corruption. Within the proposed generalized corruption-corrected learning framework, we find a hierarchy-induced set of results on how the optimization problem changes under various corruptions, and how to abstractly compute their loss corrections in Theorems 34 and 36. We conclude that more complex corruptions are more detrimental, and require more sophisticated designs than mitigation via classical loss correction.

2. Technical Background

2.1 Markov Kernels

We now introduce the mathematical machinery used for modeling corruption in learning problems; that is, Markov kernels and some of their relevant properties. The material reported here is drawn from (Klenke, 2007; Çinlar, 2011; van Rooyen and Williamson, 2018; Kallenberg, 2017; Johnston, 2023); the reader can refer to them for a comprehensive understanding of kernels in probability and learning theory.

Definition 1 (Klenke (2007)) *Let $(X_1, \mathcal{X}_1)$ and $(X_2, \mathcal{X}_2)$ be Polish spaces equipped with Borel σ-algebras, i.e., standard Borel measurable spaces. Let κ be a mapping from $X_1 \times \mathcal{X}_2$ into $[0, +\infty]$. Then, κ is called a* ***transition kernel*** *from $(X_1, \mathcal{X}_1)$ to $(X_2, \mathcal{X}_2)$ if*

1. *the mapping $x_1 \in X_1 \to \kappa(x_1, B)$ is $\mathcal{X}_1$-measurable for every set $B \in \mathcal{X}_2$, and*
2. *the mapping $B \in \mathcal{X}_2 \to \kappa(x_1, B)$ is a measure on $(X_2, \mathcal{X}_2)$ for every $x_1 \in X_1$.*

A kernel is said to be a ***Markov kernel*** *if $\kappa(x_1, X_2) = 1 \ \forall x_1 \in X_1$, i.e., it maps to a probability measure; this is denoted by the compact notation $\kappa \colon X_1 \rightsquigarrow X_2$.*[1] *The set X_1 is said to be the domain of κ, and X_2 its image, i.e.,*

$$D(\kappa) = X_1 \,, \quad I(\kappa) = X_2 \,.$$

We refer to the set of kernels as $\mathcal{T}(X_1, X_2)$ and its subset of Markov kernels as $\mathcal{M}(X_1, X_2)$.

To better grasp the concept of Markov kernel, we can think of it as a parameterized family $\kappa(x_1, \cdot)\,, x_1 \in X_1$ of probability measures on the space $(X_2, \mathcal{X}_2)$. It can be interpreted as an *observation channel*, a concept rooted in information theory and properly formalized in (Csiszár, 1972). In this context, a Markov kernel serves as a detailed probabilistic description of the generative process leading from a "hidden value" X_1 to *observed* distribution on X_2.[2] As such, for finite spaces they can be represented as stochastic matrices.

1. This notation is borrowed from category theory, see (Parzygnat, 2020) for a primer.
2. In fact, under our assumptions they are the *regular conditional probabilities* associated to the coupling of the spaces $(X_1, \mathcal{X}_1)$ and $(X_2, \mathcal{X}_2)$, see Çinlar (2011).

Example 1 *Consider a set $Z = \{0, 1\}$, a kernel $\kappa \in \mathcal{M}(Z, Z)$, and random variables $\mathsf{Z}, \tilde{\mathsf{Z}}$ on $(Z, \mathcal{Z})$. We can conveniently write the kernel as the matrix of the conditional probabilities:*

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

Special types of kernels, which will be extensively used in our analysis, are:

- A **trivial** Markov kernel, defined on the trivial domain space $\{*\}$ and taking values in the set $\{\nu\}$, ν being a probability distribution. This kernel will therefore be equivalent to the probability distribution itself. We can formally write

$$\kappa_\nu \colon \{*\} \rightsquigarrow X_1\,, \quad \kappa_\nu \equiv \nu\,, \quad \nu \in \mathcal{P}(X_1)\,, \tag{1}$$

  with $(\{*\}, \{\{*\}, \emptyset\})$ a measurable space with only one element.[3] In the text, we will simplify the notation by directly using ν instead of κ_ν;

- A Dirac delta kernel, i.e., an **identity** kernel, defined as $\delta_{X_1} \colon X_1 \rightsquigarrow X_1$, such that for all $A \in \mathcal{X}_1$, we have $\delta_{X_1}(x, A) = 1$ if $x \in X_1$, $\delta_{X_1}(x, A) = 0$ otherwise.

2.1.1 Kernel Actions

A Markov kernel naturally induces two useful functionals, one on distributions and one on functions. They are defined as:

$$\begin{aligned} \cdot\,\kappa \colon \mathcal{P}(X_1) \to \mathcal{P}(X_2) \qquad & \mu\kappa(B) \coloneqq \int_{X_1} \mu(dx_1)\, \kappa(x_1, B) \qquad \forall B \in \mathcal{X}_2\,, \\ \kappa\,\cdot \colon L^0(X_2, \mathbb{R}) \to L^0(X_1, \mathbb{R}) \qquad & \kappa f(x_1) \coloneqq \int_{X_2} \kappa(x_1, dx_2)\, f(x_2) \qquad \forall x_1 \in X_1\,, \end{aligned}$$

provided the integral exists and assuming that $\mathcal{P}(X)$ refers to the set of probabilities on a set X. We refer to these operators as the *actions* of kernels on distributions and functions, respectively.

Equipped with their action, Markov kernels can now be seen as a point-wise probabilistic description of the distortion process applied to a probability distribution μ on X_1, transforming it into another *observed* distribution on X_2; equivalently, we can make a similar comment for functions f of X_2. Again, Markov kernels are nothing else than *observation channels* (Csiszár, 1972).

2.1.2 Kernel Operations

Kernels can be combined through different operations. We introduce them here briefly, mainly inspired by (Kallenberg, 2017; Johnston, 2023), covering all the necessary properties for this work. We will use henceforth the notation $(X_i, \mathcal{X}_i)$ for a standard Borel

3. This set should be regarded as a placeholder. The value of $*$ does not influence the output of the kernel, which will in any case be ν.

measurable space. We remark that specifying the kernel action operator κf for all measurable f effectively defines a kernel as $\kappa(x_1, B) \coloneqq \kappa\chi_B(x_1)$ (Çinlar, 2011, Remark 6.4), where $\chi_B(x_2)$ is the indicator function for $x_2 \in X_2$, $B \in \mathcal{X}_2$.

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.*

P 1 *Given $\kappa\colon X_1 \rightsquigarrow X_2$ and $\lambda\colon X_2 \rightsquigarrow X_3$, their **chain composition** is a kernel $\kappa\circ\lambda\colon X_1 \rightsquigarrow X_3$ uniquely determined by the following kernel action:*

$$(\kappa \circ \lambda)f(x_1) \coloneqq \int_{X_2} \kappa(x_1, dx_2) \int_{X_3} \lambda(x_2, dx_3) f(x_3) \ ,$$

where $f\colon X_3 \to \mathbb{R}$ is a positive $\mathcal{X}_3$-measurable function.

P 2 *Given $\kappa\colon X_1 \rightsquigarrow X_2$ and $\lambda\colon X_1 \times X_2 \rightsquigarrow X_3$, their **product composition** is a kernel $\kappa \times \lambda\colon X_1 \rightsquigarrow X_2 \times X_3$ uniquely determined by the following kernel action:*

$$(\kappa \times \lambda)f(x_1) \coloneqq \int_{X_2} \kappa(x_1, dx_2) \int_{X_3} \lambda((x_1, x_2), dx_3)\, f(x_2, x_3) \ ,$$

for every f positive $\mathcal{X}_2 \times \mathcal{X}_3$-measurable.

The operations defined above can naturally understood using well-known probability theory results. Consider the trivial Markov kernel

$$\kappa_\nu\colon \{*\} \rightsquigarrow X_1, \ \ \nu \in \mathcal{P}(X_1).$$

In this setting, the operations P1 and P2 apply to κ_ν when composed with a kernel $\lambda_1\colon X_1 \rightsquigarrow X_2$ (for the chain composition) and $\lambda_2\colon \{*\} \times X_1 \rightsquigarrow X_2$ (for the product composition).

This allows us to write distribution-kernel combinations using the same notation as kernel-kernel ones, i.e., $\kappa_\nu \circ \lambda_1$ and $\kappa_\nu \times \lambda_2$.[4] Both of these constructions result in *new probability measures*:

- The composition
$$\nu \circ \lambda_1 \coloneqq \kappa_\nu \circ \lambda_1 \ \in \mathcal{P}(X_2)$$
is equivalent to the kernel action on probabilities $\lambda_1\nu$, and corresponds to the Law of Total Probability. We adopt the $\circ$ notation for it from now on.

- The product
$$\nu \times \lambda_2 \coloneqq \kappa_\nu \times \lambda_2 \ \in \mathcal{P}(X_1 \times X_2)$$
corresponds to the Bayesian decomposition of a joint probability into a marginal ν and a conditional probability λ_2.

4. Strictly speaking, $\kappa_\nu \circ \lambda_1 \in \mathcal{M}(\{*\}, X_2)$, while $\lambda_1\nu \in \mathcal{P}(X_2)$. So this identification holds up to a suitable "projection". However, we avoid this level of technicality to keep the presentation clear and simple.

Notice that λ_2 is essentially of the same type as λ_1, apart from a dummy variable over the singleton set $\{*\}$. In what follows, we will overload the notation and also write $\nu \times \lambda_1 \in \mathcal{P}(X_1 \times X_2)$.

The second set of operations defined here can be referred to as *parallel operations*. Compared to in-series operations as in P1 and P2, it allows for more flexible combinations of kernels.

P 3 *Given $\kappa\colon X_1 \rightsquigarrow X_2$ and $\lambda\colon X_3 \rightsquigarrow X_4$, their **superposition** is a kernel $\kappa \otimes \lambda\colon X_1 \times X_3 \rightsquigarrow X_2 \times X_4$ uniquely determined by the following kernel action:*

$$(\kappa \otimes \lambda) f(x_1, x_3) \coloneqq \int_{X_2} \kappa(x_1, dx_2) \int_{X_4} \lambda(x_3, dx_4)\, f(x_2, x_4) \ ,$$

where $f\colon X_2 \times X_4 \to \mathbb{R}$ is positive $\mathcal{X}_2 \times \mathcal{X}_4$-measurable.

Remark 2 *Observe that no restriction is imposed on the parameter spaces to be equal, e.g., $X_1 = X_3$, or Cartesian products with some space in common, e.g., $X_1 = Y_1 \times Y_2, X_3 = Y_1 \times Y_3$. When this happens, the actions of the two kernels "superpose" on the same space. It is possible for the superposition operation to produce a kernel*

$$(\kappa \otimes \lambda) f(x_1) \coloneqq \int_{X_2} \kappa(x_1, dx_2) \int_{X_3} \lambda(x_1, dx_3)\, f(x_2, x_3) \ ,$$

with $\kappa\colon X_1 \rightsquigarrow X_2$ and $\lambda\colon X_1 \rightsquigarrow X_3$, so that $\kappa \otimes \lambda\colon X_1 \rightsquigarrow X_2 \times X_3$. Another possible case is $\kappa'\colon X_1 \rightsquigarrow X_2$ and $\lambda'\colon X_1 \times X_2 \rightsquigarrow X_3$, 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 P 3 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 P1 and the superposition P3.

P 4 *Given $\kappa\colon X_1 \times X_2 \rightsquigarrow X_3$ and $\lambda\colon X_1 \times X_3 \rightsquigarrow X_4$, their **partial chain composition** is a kernel $\kappa \circ_{X_3} \lambda\colon X_1 \times X_2 \rightsquigarrow X_4$ uniquely determined by the following kernel action:*

$$(\kappa \circ_{X_3} \lambda) f(x_1, x_2) \coloneqq \int_{X_3} \kappa((x_1, x_2), dx_3) \int_{X_4} \lambda((x_1, x_3), dx_4)\, f(x_4) \ ,$$

where $f\colon X_4 \to \mathbb{R}$ is a positive $\mathcal{X}_4$-measurable function.

Essentially, this operation only chains the kernels on the specified space, here X_3, while superposing them on the common parameter, X_1.

2.2 Statistical Experiments and Supervised Learning

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). We summarize the key concepts crucial to our analysis and direct readers to relevant books for a more comprehensive perspective.

2.2.1 The General Learning Problem

In statistical decision theory, a general learning 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) act choosing an *observation* $o \in O$ given some hidden *state* $\theta \in \Theta$. The stochastic process generating o given θ is referred to as the *experiment* E.

Definition 3 *An* ***experiment*** $E \colon \Theta \rightsquigarrow O$ *is a Markov kernel from the hidden state space to the observation space.*

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

Definition 4 *Consider the product space* $(\Theta \times O \times A,\ \Omega \times \mathcal{O} \times \mathcal{A})$*, where* (Θ, Ω)*,* $(O, \mathcal{O})$*, and* $(A, \mathcal{A})$ *are standard Borel measurable spaces for parameters, observations and decisions. We refer to the space* $(\Theta \times O,\ \Theta \times \mathcal{O})$ *as the* ***data space*** *on which we aim to learn. A* ***general learning problem*** *on such a product space is a pair* $(\mathcal{L}, \mathcal{C})$*, where* $\mathcal{L}$ *denotes the learning context and* $\mathcal{C}$ *specifies the learning criterion. Specifically, the* ***learning context*** *is defined as* $\mathcal{L} = (\ell, \mathcal{H}, P)$*, where:*

- $\ell \colon \mathcal{P}(A) \times \Theta \to \mathbb{R}$ *is a* ***loss function*** *in* $L^0(\mathcal{P}(A) \times \Theta, \mathbb{R})$*,*
- $\mathcal{H} \subseteq \mathcal{M}(O, A)$ *is a decision class, or* ***model class****,*
- *and* $P \coloneqq \pi_\theta \times E$ *is the* ***data-generating probability distribution****, with associated experiment* $E \in \mathcal{M}(\Theta, O)$ *modeling the stochastic observations, and prior distribution* $\pi_\theta \in \mathcal{P}(\Theta)$ *modeling the likelihood of the parameters to manifest.*

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

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

2.2.2 Supervised Learning through Risk Minimization

In the specific setup of *supervised learning*, the observation space O is the attribute space $X \subset \mathbb{R}^d, d \geq 1$, while both states Θ and actions A correspond to the label space Y. Then, the experiment $E\colon Y \rightsquigarrow X$ identifies a probability associated with the attribute X, given the state Y. Here we focus on the classification task that assumes the label space to be *finite*, while no constraint is imposed on X apart from being a compact subset of $\mathbb{R}^d$. 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 5 *A **posterior kernel** for a data space $(X \times Y, \mathcal{X} \times \mathcal{Y})$ is a Markov kernel $F \in \mathcal{M}(X, Y)$ from the label space $(Y, \mathcal{Y})$ to the attribute space $(X, \mathcal{X})$.*

Notice that, as we already noticed when introducing Markov kernels, the product operation can be used to write the Bayes decomposition theorem for a join probability $P \in \mathcal{P}(X \times Y)$. Hence, given $F \in \mathcal{M}(X, Y)$ and $E \in \mathcal{M}(Y, X)$, we can write

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

for some suitable marginal probabilities $\pi_X \in \mathcal{P}(X)$ and $\pi_Y \in \mathcal{P}(Y)$.[5]

Definition 6 *Let $(X \times Y, \mathcal{X} \times \mathcal{Y})$ be a data space. Given a loss $\ell \in L^0(\mathcal{P}(Y) \times Y, \mathbb{R}_{\geq 0})$, a model class $\mathcal{H} \subseteq \mathcal{M}(X, Y)$ and a joint probability distribution $P \coloneqq \pi_Y \times E \in \mathcal{P}(X \times Y)$, with $\pi_Y \in \mathcal{P}(Y)$ and $E \in \mathcal{M}(Y, X)$. Then, the **Bayes risk** (BR) is defined as*

$$\begin{aligned} \mathrm{BR}_{\ell,\mathcal{H}}(\pi_Y \times E) &\coloneqq \inf_{h \in \mathcal{H}} \mathrm{R}_{\pi_Y \times E}(\ell \circ h)\,, \\ \mathrm{R}_{\pi_Y \times E}(\ell \circ h) &\coloneqq \mathbb{E}_{\mathsf{Y} \sim \pi_Y} \mathbb{E}_{\mathsf{X} \sim E_{\mathsf{Y}}} \ell(h_{\mathsf{X}}, \mathsf{Y})\,. \end{aligned}$$

*Here, R is known as the **risk**; the notation h_{X} and E_{Y} denotes evaluation of a kernel (such as h or E) at a random variable (such as X or Y) and will be used consistently throughout.*

Definition 7 *Let $(X \times Y, \mathcal{X} \times \mathcal{Y})$ be a data space, consisting of a finite set of labels Y and a set of continuous attributes $X \subset \mathbb{R}^d$. A **supervised learning problem** on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ is a general learning problem, where:*

- *the loss is $\ell \in L^0(\mathcal{P}(Y) \times Y, \mathbb{R}_{\geq 0})$,*
- *the model class is $\mathcal{H} \subseteq \mathcal{M}(X, Y)$,*
- *and the data-generating distribution is $P \coloneqq \pi_Y \times E \in \mathcal{P}(X \times Y)$ with $\pi_Y \in \mathcal{P}(Y)$ and $E \in \mathcal{M}(Y, X)$.*

5. To be precise, $\pi_X \times F \in \mathcal{P}(X \times Y)$ and $\pi_Y \times E \in \mathcal{P}(Y \times X)$ by P2. 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.

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 simply using $\mathcal{L} = (\ell, \mathcal{H}, P)$, with $\mathcal{C}$ implicitly given by risk minimization.*

The definition above fits in the general learning problem framework by considering the specific diagram $Y \overset{E}{\rightsquigarrow} X \overset{h}{\rightsquigarrow} Y$, where h is a decision rule chosen in $\mathcal{H}$, therefore choosing a probability on Y associated to a point in X.[6]

For some cases, the formulation of learning problem in terms of the experiment, or loss and model class, can be restrictive; for this reason, we introduce some alternative ways of writing $\mathcal{L}$. This can be naturally justified by considering the following simple proposition.

Proposition 8 *A supervised learning problem $\mathcal{L} = (\ell, \mathcal{H}, P = \pi_Y \times E)$ on the data space $(X \times Y, \mathcal{X} \times \mathcal{Y})$ can be equivalently expressed*

1. *using the minimization set $\ell \circ \mathcal{H} \coloneqq \{(x, y) \mapsto \ell(h(x), y) \mid h \in \mathcal{H}\}$, i.e., as a couple $\mathcal{L} = (\mathcal{F} \coloneqq \ell \circ \mathcal{H}, P)$.*
2. *using the posterior kernel, i.e., as $P = \pi_X \times F$ for some prior $\pi_X \in \mathcal{P}(X)$ on the attribute space. We will then refer to it as $\mathcal{L} = (\ell, \mathcal{H}, P = \pi_X \times F)$ on $(X \times Y, \mathcal{X} \times \mathcal{Y})$;*
3. *using the joint distribution P, **agnostic** regarding its factorization. We will then refer to it as $\mathcal{L} = (\ell, \mathcal{H}, P)$ on the general data space space $(Z, \mathcal{Z})$.*

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

We remark that, in the literature, when $\mathcal{H} = \mathcal{M}(X, Y)$, we talk about unconstrained learning problem and unconstrained Bayes risk. Since our focus in the following will exclusively be on constrained supervised learning problems, we will simply term them *learning problems* and never make use of the other cases.

3. A Taxonomy of Corruptions in Supervised Learning

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 corruptions can take, 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, through a careful examination, we analyze the relationships between our taxonomy of corruption and existing corruption models, elucidating novel insights generated by our framework.

3.1 Corruption Definition and Types

We have defined in the previous section that a learning problem comprises three key components: the loss function ℓ, the model class $\mathcal{H}$, and the probability distribution P from which we draw the data. We now turn our attention to how such a problem may be corrupted.

6. This is, considering the hypothesis, or decision, as stochastic. Several techniques exist to obtain a deterministic labels from a stochastic decision rule, with different consequences (Cotter et al., 2019).

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 $(\ell, \mathcal{H}, P)$.

Definition 9 *Let $(Z, \mathcal{Z}), (Z', \mathcal{Z}')$ be data spaces. A **general corruption** is a mapping sending a learning problem $\mathcal{L} = (\ell, \mathcal{H}, P)$ into another learning problem $\tilde{\mathcal{L}} = (\tilde{\ell}, \tilde{\mathcal{H}}, \tilde{P})$, where $\mathcal{L}, \tilde{\mathcal{L}}$ are defined on $(Z, \mathcal{Z}), (Z', \mathcal{Z}')$ respectively.*

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.[7] This definition subsumes a significant portion of existing literature, including classical works on distribution shift and noisy data. As such, our attention now turns 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. This, in turn, suggests that future work must extend beyond this subcase, as we have identified certain examples that are not covered by this definition (see § 3.2).

Definition 10 *A **Markovian corruption** is a general corruption that maps a learning problem $\mathcal{L} = (\ell, \mathcal{H}, P)$ defined on $(Z, \mathcal{Z})$ into another learning problem $\tilde{\mathcal{L}}$ on $(Z', \mathcal{Z}')$ through the action of a Markov kernel $\kappa \colon Z \rightsquigarrow Z'$, i.e., such that $\tilde{\mathcal{L}} = (\ell, \mathcal{H}, \tilde{P} = P \circ \kappa)$. Two important subcases are:*

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

We will use the term "corruption" throughout the remainder of the paper as shorthand for "Markovian corruption".

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 P, i.e., $\tilde{P} = P \circ \kappa$. The kernel is therefore not uniquely identified by the corruption definition, since multiple Markov kernels can generate $\tilde{P}$ from P. However, for the analysis carried on in the rest of the paper, 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), as a point-wise 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/Nature), consequently giving rise to corruptions (distorted data distribution).

7. While Markov kernels have been utilized in formalizing corruption (van Rooyen and Williamson, 2018; Williamson and Cranko, 2024), their primary foci were solely on label corruption, attribute corruption, or simple joint corruption.

For formal statements, we abuse the kernel notation and refer to the corruption induced by a kernel as the kernel itself, i.e., $\kappa\colon Z \rightsquigarrow Z'$, or equivalently $\kappa \in \mathcal{M}(Z, Z')$ for some suitable sets Z, Z'.

3.1.1 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 $X \times Y$ induced by $\kappa\colon X \times Y \rightsquigarrow X \times Y$, when one space (either X or Y space) is absent in the image or domain of the corruption kernel, we obtain a partial corruption. In Fig. 1, we present all possible types of partial corruption, with the exception of those that are identical (Dirac delta kernels) or trivial ($D(\kappa) = \{*\}$) kernels, as they can be seen as obvious subcases of other partial corruptions. By construction, we can state:

Proposition 11 *There are no partial Markovian corruptions outside of those listed in Fig. 1.*

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

3.1.2 Constructing Joint Corruption as a Combination of Partial Ones

We now enumerate all possible ways of constructing joint corruptions, i.e., of the type $\kappa\colon X \times Y \rightsquigarrow X \times Y$, by combining the nodes in Fig. 1 through the superposition operation P3. To this end, we introduce an additional condition to be imposed on the combinations of partial corruptions.

Definition 12 *A Markovian corruption, induced by a kernel $\kappa \in \mathcal{M}(X \times Y, X \times Y)$, 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 P1–4.*

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.

Remark 13 *It is also possible for a corruption kernel $\kappa \in \mathcal{M}(X \times Y, X \times Y)$ to not respect the one-step condition in Definition 12. 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 P1–4 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*

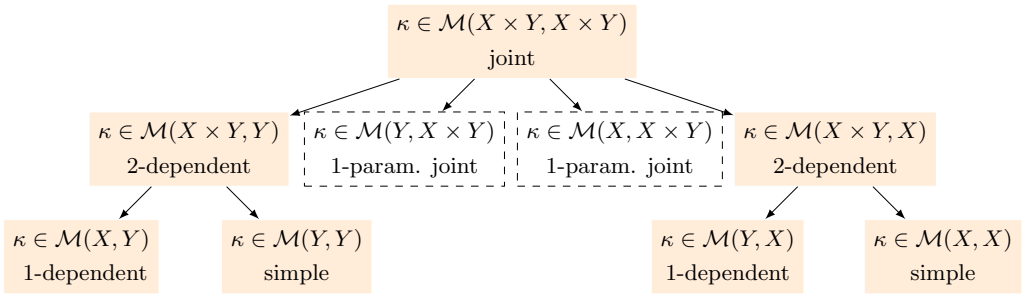


Figure 1: *Hierarchy of partial corruption types.* The partial corruption types are hierarchically organized based on their dependence on the instance X and label Y space, as depicted through a tree structure. At the root of the tree lies the most general form of corruption, where the domain and image spaces are both the joint one, i.e., $D(\kappa) = I(\kappa) = X \times Y$. 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.

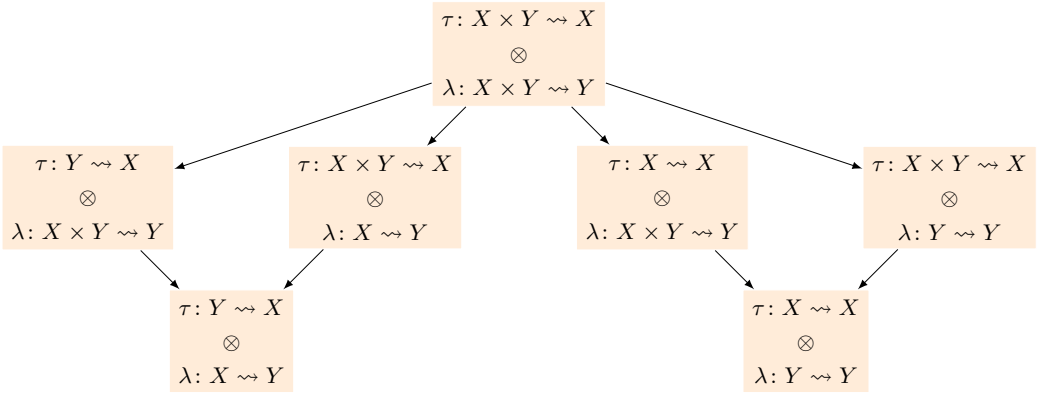


Figure 2: *Feasible combinations of partial corruptions.* Joint corruptions, i.e., of type $\kappa \colon X \times Y \rightsquigarrow X \times Y$, are obtained by combining two compatible partial corruptions in Fig. 1. The graph structure is induced by that of the partial corruption types. Notice that we can only combine a partial corruption with $I(\tau) = X$ with another such that $I(\lambda) = Y$, following Proposition 14. 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.

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

Proof According to Definition 10, we must map a joint probability distribution on $X \times Y$ 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 $(\tau \otimes \lambda)(x, d\tilde{x}, d\tilde{y}) = \tau(x, d\tilde{x}, d\tilde{y}) \otimes \lambda(x, d\tilde{y})$ or of more than 2 kernels are not allowed because, according to P3, the measure on the corrupted labels (or in general, space) would be ill-defined. By taking λ and τ from the partial corruption in Fig. 1, enumerating all of their possible combinations, and checking which of them are feasible, we can see that only the ones with $I(\tau) = X$ and $I(\lambda) = Y$ or $I(\tau) = Y$ and $I(\lambda) = X$ 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. 2. Proposition 14 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 label and attributes, such as 1-parameter joint ones, are incorporated in the class of joint non-factorized corruption, i.e., $\mathcal{M}(X \times Y, X \times Y)$.[8] 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 14.

Corollary 15 *There are no one-step Markovian corruptions outside of those listed in Fig. 2, and therefore it constitutes a* **complete** *taxonomy.*

3.1.3 A Practical Example

Here, we present an illustration of a one-step Markovian corruption (in the finite case) within a practical scenario to facilitate for 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–predict 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 modeling the observed outcome

8. Note that a 1-parameter joint corruption can be seen as a subcase of a joint one, as $\kappa(x, y, d\tilde{x}d\tilde{y}) = \lambda(x, d\tilde{x}d\tilde{y})\,\mathbf{1}(y)$, where $\mathbf{1}(y)$ only trivially depends on y since it is the matrix with all entries equal to 1.

"rearrest", denoted as $\tilde{\mathsf{Y}} \in Y \coloneqq \{+1, -1\}$, as a corrupted version of the true outcome "reoffense", denoted as $\mathsf{Y} \in Y \coloneqq \{+1, -1\}$, depending on the attribute $\mathsf{X} \in X = \{b, w\}$. Specifically, the disparity between Y and $\tilde{\mathsf{Y}}$ can be captured by a higher probability of flipping the reoffense label ($\mathsf{Y} = +1$) to the no rearrest label ($\tilde{\mathsf{Y}} = -1$) for white population ($\mathsf{X} = w$) compared to the black population ($\mathsf{X} = b$):

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

where p is a conditional probability density.

Moreover, we assume 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 ($\mathsf{Y} = -1$) to the rearrest label ($\tilde{\mathsf{Y}} = +1$) is zero for both the black and white populations:

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

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

$$\delta_X \coloneqq \begin{bmatrix} p'(\tilde{\mathsf{X}} = b \,|\, \mathsf{X} = b) & p'(\tilde{\mathsf{X}} = b \,|\, \mathsf{X} = w) \\ p'(\tilde{\mathsf{X}} = w \,|\, \mathsf{X} = b) & p'(\tilde{\mathsf{X}} = w \,|\, \mathsf{X} = w) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\lambda_{\mathsf{X}}\big|_{\mathsf{X}=x} \coloneqq \begin{bmatrix} p(\tilde{\mathsf{Y}} = +1 \,|\, \mathsf{Y} = +1, \mathsf{X} = x) & 0 \\ p(\tilde{\mathsf{Y}} = -1 \,|\, \mathsf{Y} = +1, \mathsf{X} = x) & 1 \end{bmatrix} = \begin{bmatrix} 1 - \alpha(x) & 0 \\ \alpha(x) & 1 \end{bmatrix},$$

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

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

From Definition 10 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 additionally consider

$$q = [1/4, 1/4, 1/4, 1/4]^\top \in \mathcal{P}(X \times Y),$$

where the specific order is assumed to be

$$q \coloneqq [\, q(\mathsf{X} = b, \mathsf{Y} = +1),\, q(\mathsf{X} = b, \mathsf{Y} = -1),\, q(\mathsf{X} = w, \mathsf{Y} = +1),\, q(\mathsf{X} = w, \mathsf{Y} = -1) \,]^\top.$$

Note that in finite spaces, the superposition operation P3 reduces to the Kronecker product, hence we write the joint corruption kernel $\delta_X \otimes \lambda \colon X \times Y \rightsquigarrow X \times Y$ as

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

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 $\tilde{\mathsf{X}} = \tilde{x}, \tilde{\mathsf{Y}} = \tilde{y} \,|\, \mathsf{X} = x, \mathsf{Y} = y$. Then, we can obtain the *a corrupted joint probability* $\tilde{q} \in \mathcal{P}(X \times Y)$ in the following manner:

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

In this finite case, the chain composition operation P1 reduces to matrix multiplication. After applying the corruption kernel $\delta_X \otimes \lambda$, the original clean learning problem, characterized by the joint probability q, is transformed into a distinct, corrupted problem governed by $\tilde{q}$.

3.1.4 On the Exhaustiveness of Markovian Corruption

We now know from Corollary 15 that the taxonomy of one-step corruptions is complete. However, we also noticed that one-step corruptions are not the only possible type of corruptions, because kernels can also come in different forms. By leveraging a well-known property of Markov kernels—their bijection with coupling of probability spaces—we define an *exhaustive* taxonomy.

Definition 16 *We say that a taxonomy of Markovian corruptions is* ***exhaustive*** *if, for every fixed pair of distributions $(P, \tilde{P}) \in \mathcal{P}(Z) \times \mathcal{P}(Z)$, there exists a Markovian corruption from $\mathcal{L}$ to $\tilde{\mathcal{L}}$ s.t. $\tilde{P} = \kappa \circ P$ for all loss ℓ and model class $\mathcal{H}$.*

Proposition 17 *The set of feasible one-step Markovian joint corruptions, illustrated in Fig. 2, together with the set of non-factorized ones, constitutes an exhaustive taxonomy of Markovian corruptions.*

Proof A coupling is formally defined for two probability spaces $(Z_1, \mathcal{Z}_1, P_1)$, $(Z_2, \mathcal{Z}_2, P_2)$ as a probability space $(Z_1 \times Z_2, \mathcal{Z}_1 \times \mathcal{Z}_2, P)$, such that the marginal probabilities associated to P w.r.t. Z_i, $i \in \{1, 2\}$, are the respective P_i (Wang, 2012, Definition 1.1). By construction, Markov kernels with fixed input and output probabilities $P, \tilde{P}$ are in bijection with all the possible couplings existent on $Z \times Z$ with two *fixed* probability measures; for us, $P, \tilde{P}$ (see details in § E). This, by definition, proves the exhaustiveness of the taxonomy, as it implies that every corruption that is *not* one-step Markovian, still sends P into $\tilde{P}$, has a non-factorized representation. ■

3.2 Scope and Contributions of Our Taxonomy

As presented, our taxonomy introduces a unified, kernel-based framework for understanding and organizing data corruption models in machine learning. It focuses on one-step Markovian corruptions, formally defined in Definition 12, and is proven to be complete (by itself) and exhaustive (adding non-factorized ones). The scope is broad enough to encompass a

Table 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 § C.

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

large variety of existing models, which we organize hierarchically based on their dependence on the instance and label spaces.

A primary contribution of our framework is the reformulation of existing corruption models from Tab. 1, which are aligned with our taxonomy as depicted in Tab. 2.[9] Prior categorizations, e.g., (Quiñonero-Candela et al., 2008), typically rely on invariance-based definitions (as shown in Tab. 1), that is, specifying which parts of the data distribution remains unchanged. By design, they do not support a hierarchical or compositional interpretation of corruption. In contrast, our taxonomy enables a comparative view of corruption models by analyzing the domain and image of their associated kernels. This allows us to order corruptions by type complexity: a kernel with more dependencies induces a more

9. Details about the relationships with these corruption instances are given in § C, and discussions on the relationships with other data corruption taxonomies are given in § D.

intricate corruption, while simpler corruptions emerge as subcases of more complex ones.[10] Such a view provides a theoretical foundation for the systematic analysis and resolution of corruptions; an instance how that can be developed is proposed in Sections § 4 and § 5.

As a second point, our work reveals that some corruption models in Tab. 1 correspond to multiple corruption types or combinations of partial corruptions in our taxonomy. This is because their definitions in Tab. 1 often consider the corruption of either the probability in the X space or the 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. By observing this, we gain the insight that some corruption models are way more general than others, and can even be regarded as taxonomies of their own. This additionally translates into the impossibility of having a one-on-one correspondence between our taxonomy and existing ones. We elaborate further on this point in § D.

Third, our framework helps identifying corruption types that have been taken for granted to be probabilistic in nature, but, according to our definition of probabilistic corruption within the Markovian framework, are not. These include complex cases that are non-one-step (Definition 12) or even non-Markovian (Definition 10), and therefore require different treatment. We discuss examples of such corruptions below.

3.2.1 Multi-Step Corruption

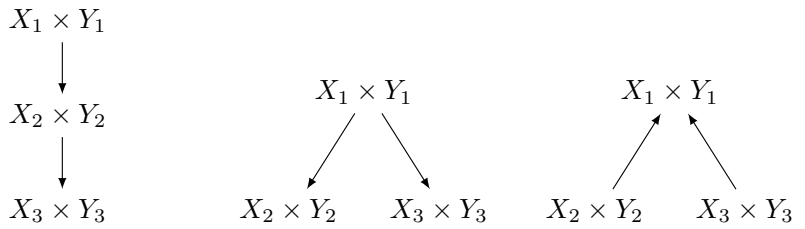
In machine learning research, the corruption process typically involves two environments—training and test time—but other settings are also possible. For example, scenarios involving more than two spaces arise when learning from multiple domains (Ben-David et al., 2010), or in the presence of concept 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" (P1-P2) or "parallel" (P3) 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 $X_i \times Y_i$, $i = 1, 2, 3$. 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 possible presence of a non-trivial Markov kernel between two spaces, the absence of an arrow indicates that the two spaces must be related by a trivial kernel–in other words, they are independent.

We focus on three corruption configurations depicted in Fig. 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. 3(a), where the spaces influence each other sequentially. A concrete example is when $\kappa_1 \colon X \rightsquigarrow W$ models a feature extractor for the attribute space X, and $\kappa_2 \colon W \rightsquigarrow X$ represents a corruption depending

10. However, the concept of "complexity" of a corruption should not be interpreted as anything more than a structural complexity, given by its type. We are not proving here any quantitative complexity result, but organizing corruptions hierarchically for *qualitative* comparison.



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

Figure 3: Possible non-degenerate relations among three probability spaces. Arrows represent a non-trivial Markov kernel $\kappa \colon X_i \times Y_i \rightsquigarrow X_j \times Y_j$.

on the latent features only. Although this structure does not strictly satisfy our one-step corruption definition as per Definition 12, 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 possibility is that the spaces relate according to the *triangular* structures shown in Fig. 3(b) and (c). In particular, case (b) 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 other two, i.e., $\kappa := \kappa_1 \otimes \kappa_2$, with $\kappa_1 \in \mathcal{M}(X_1 \times Y_1, X_2 \times Y_2)$ and $\kappa_2 \in \mathcal{M}(X_1 \times Y_1, X_3 \times Y_3)$. As for case (c), it can be used to model scenarios with merged (noisy or clean) 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 via $\kappa \in \mathcal{M}(X_2 \times Y_2 \times X_3 \times Y_3, X_1 \times Y_1)$, with a decoupled joint probability as input, e.g., $P = P_2 \times \kappa_{P_3}$ where κ_{P_3} is a trivial kernel and P_2, P_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 $\lambda \in \mathcal{M}(X \times Y \times X, Y)$ and $\tau \in \mathcal{M}(X, X)$. Let A be an element of the Borel sigma algebra on $X \times Y$.

We can obtain a corrupted measure $\tilde{P}$ as

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

where P is the joint clean probability and π the associated clean 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 $(X \times Y \times X)$, where the latter X is meant to be equipped with the *corrupted* marginal probability $\pi \circ \tau$, and $X \times Y$ has marginal P. This modeling choice amounts to a corruption that does not act in a "parallel" fashion on X and 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 X.

Notice that, in general, we cannot write the combined action of λ and τ as a unique kernel κ acting on P; the available operations P1, P2 and P4 do not permit this. Consequently, it cannot be expressed as a single graphical model, although each individual step can. However, we can write $\tilde{P}(A) = [\tilde{F} \times \tilde{\pi}](A)$, with $\tilde{\pi} \coloneqq \pi \circ \tau$ and $\tilde{F} \coloneqq P \circ_{X\times Y} \lambda$. This shows that the resulting corruption is **non-Markovian** in how it acts on the clean probability P, *but has (almost) Markovian components.*[11]

3.2.2 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. 1. Fix a measurable instance space X, and denote by P a distribution over $X \times [K]$ for $[K] \coloneqq \{1, 2, \cdots, K\}$ with random variables $(\mathsf{X}, \mathsf{Y}) \sim P$.

Definition 18 (Katz-Samuels et al. (2019)) *Let $P(\mathsf{X} = x \,|\, \mathsf{Y} = k)$ be class-conditional distribution and $\pi_k \coloneqq P(\mathsf{Y} = k)$ be the base rate, both for $k \in [K]$. 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 $(\ell, \mathcal{H}, P)$ to $(\ell, \mathcal{H}, \tilde{P})$,*

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

such that the corrupted class-conditional distributions are of the form

$$\tilde{P}(\mathsf{X} = x \,|\, \mathsf{Y} = m) \coloneqq \sum_{k=1}^{K} \pi_{m,k} P(\mathsf{X} = x \,|\, \mathsf{Y} = k) \quad \forall x \in X \,,$$

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

This definition has some clear differences with our definition of corruption. First, Definition 18 uses the class conditional probabilities $P(\mathsf{X} = x \,|\, \mathsf{Y} = k)$ 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:

$$\tilde{P}(A) = \sum_{Y} \int_A \int_X \delta_X(x, d\tilde{x})\, \kappa_M(\tilde{y}, dy)\, E(y, dx)\, \tilde{\pi}(d\tilde{y}) \tag{2}$$

$$= \int_A \left[\big(\kappa_M \circ (E \circ \delta_X)\big) \times \tilde{\pi} \right] (d\tilde{x}, d\tilde{y}) \neq \int_A \left[\big(\delta_X \otimes \kappa_M\big) \circ (\pi \times E) \right] (d\tilde{x}, d\tilde{y}) \,, \tag{3}$$

where now $\tilde{P}$ and P are joint probabilities, $\kappa_M(\tilde{y}, dy) = \mathbf{\Pi}_{\tilde{y},y} \in \mathcal{M}(Y, Y)$, and $Y \coloneqq [\max(K, M)]$ to get a square matrix with added entries filled with zeros. In particular, we remark 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 as the one provided as input, i.e., as clean distribution. Generally, we can always write the following relationship:

$$\tilde{\pi}(d\tilde{y}) = \int_Y \lambda_M(\hat{y}, d\tilde{y})\, \pi(d\hat{y}) \,, \tag{4}$$

where $\lambda_M \colon Y \rightsquigarrow Y$, so the variable $\hat{y}$ is defined on the clean probability space $(Y, \mathcal{Y}, \pi)$, $\mathcal{Y}$ being a suitable σ-algebra.

Mutually Contaminated Distributions Model is Non-Markovian. The formula derived above classifies MCD as multi-step, because plugging Eq. (4) in Eq. (2) violates Definition 12. This gets even clearer when looking at Eq. (3): the right-hand side would imply that there exists a single kernel in $(\delta_X \otimes \kappa_M)(x, \tilde{y}, d\tilde{x}, dy) \in \mathcal{M}(X \times Y, X \times Y)$ 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 remark that the existence of $(\delta_X \otimes \kappa_M)(x, \tilde{y}, d\tilde{x}, dy)$ would still *not make a viable Markovian corruption* in the sense of Definition 10 because of the variables not being compatible with the probability P for generating $\tilde{P}(d\tilde{x}, d\tilde{y}) = \big(P \circ (\delta_X \otimes \kappa_M)\big)(d\tilde{x}, d\tilde{y}) = \int_{X\times Y} P(dx, dy) \big(\delta_X \otimes \kappa_M\big)(x, \tilde{y}, d\tilde{x}, dy)$, the latter being an ill-posed integral since we have two measures on 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., $\lambda_M(\hat{y}, d\tilde{y}) = \tilde{\pi}(d\tilde{y})$. 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. That because, having $\lambda_M(\hat{y}, d\tilde{y}) = \delta_Y(\hat{y}, d\tilde{y})$ we write the marginal

$$\pi(d\tilde{y}) = \int_Y \lambda_M(\hat{y}, d\tilde{y})\, \pi(d\hat{y}) = \int_Y \delta_Y(\hat{y}, d\tilde{y})\, \pi(d\hat{y})\,.$$

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 in the above. However, Menon et al. (2015) prove in their Section 2.3 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 multi-class setting by taking

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

where λ_C is the Markov kernel associated to CCN, and $\delta_X \otimes \lambda_C$ 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., $\lambda_M = \delta_Y$. 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.

3.2.3 Selection Bias

Another corruption model that has been widely studied in the literature is selection bias. Over the years, multiple definitions have been proposed, as we briefly discuss in § C 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 19 (Chapter 3.2, Quiñonero-Candela et al. (2008)) *Let $Z \subseteq \mathbb{R}^d$ and the Borel σ-algebra $\mathcal{Z}$ on Z form a measurable space. Consider a clean probability space $(Z, \mathcal{Z}, P)$ and a corrupted one $(Z, \mathcal{Z}, \tilde{P})$, from which we aim to learn. We define **selection bias** as a general corruption such that $\mathcal{L} = (\ell, \mathcal{H}, P)$ and $\tilde{\mathcal{L}} = (\ell, \mathcal{H}, \tilde{P})$, and that fulfills the following conditions:*

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

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

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 $\kappa \in \mathcal{M}(Z, Z)$ on the input probability P, we get the condition

$$\int_A \int_Z \kappa(z, d\tilde{z}) P(dz) = \int_A \alpha(z) P(dz) \quad \forall A \in \mathcal{Z} \, .$$

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, i.e., $\kappa(z, Z) = 1$ for all $z \in Z$, unless $\alpha(z) = 1 \; \forall z \in Z$. This kernel is defined such that P is corrupted into $\tilde{P}$, 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 $\kappa \in \mathcal{M}(Z, Z)$. It can be rewritten through its density w.r.t. a suitable measure, i.e.,

$$\tilde{P}(A) = \int_A \int_Z \kappa(z, d\tilde{z}) \, P(dz) = \int_A \int_Z k(z, \tilde{z}) \, \nu(d\tilde{z}) \, P(dz) \quad \forall A \in \mathcal{Z} \, ,$$

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

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

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 P and $\tilde{P}$, since they are probability measures and our exhaustiveness argument holds. Therefore, we can represent selection bias via a 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 $P_{tr}(\mathsf{X} = x, \mathsf{Y} = y \,|\, \mathsf{S} = 1)$ as the one generating the training samples, where S is some Bernoulli selection variable that determines whether a data point is excluded or included (i.e., corrupted). The test data are instead drawn from the uncorrupted distribution $P \in \mathcal{P}(X \times Y)$. In other words, they assume two joint probability spaces on $(X \times Y \times \{0, 1\})$: one clean, where S and the data are independent, and the marginal data distribution is P; one with dependence, and the data distribution is some $\tilde{P}$. It follows that these objects exist:

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

2. $\kappa_\pi \colon X \times Y \rightsquigarrow \{0,1\}$, a trivial kernel that assumes the value π regardless of the point in $X \times Y$ it is evaluated on.[12] This is the kernel description of the random sampling case, where S is independent of (X, Y);

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

$$\int_{A \subseteq X \times Y} \kappa_{tr}(s = 1, dx, dy) = P_{tr}\Big((\mathsf{X}, \mathsf{Y}) \in A \mid \mathsf{S} = 1\Big)$$

Thus, we can write

$$\tilde{P}(A) \coloneqq (P \circ \kappa_\pi \circ \kappa_{tr})(A) =: (P \circ \kappa)(A),$$

for $A \subseteq X \times Y$. $\tilde{P}$ is the training distribution and P 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 Definition 19. Having proved that Definition 19 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) $\tilde{P}$ does not at all depend on (the support of) P. The corruption process is only dependent on S and it 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.

4. Consequences of Corruption: Data Processing Equalities

Having identified all the types of one-step Markovian corruptions, a natural subsequent question is how to systematically compare their effects. 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 $P \in \mathcal{P}(X \times Y)$. More formally, their equalities are of the form

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

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 $\mathcal{M}(Y, X)$, which is computed either as $E \circ \tau \,, \tau \in \mathcal{M}(X, X)$ or as $\lambda \circ E \,, \lambda \in \mathcal{M}(Y, Y)$. We recall that the minimization set is defined as $\ell \circ \mathcal{H} \coloneqq \{\, (x, y) \mapsto \ell(h_x, y) \,|\, h \in \mathcal{H} \subseteq \mathcal{M}(X, Y) \,\}$ (as introduced in Proposition 8),

12. This is equivalent to the definition of trivial kernel we gave in Eq. (1).

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 22.

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 here as simple X and 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 explicit formulas describing how the corruptions act on prior and posterior distributions, as well as on the minimization set $\ell \circ \mathcal{H}$. Accordingly, the main goal of this section is to present qualitative results in terms of conserved "entropy"[13] 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, as they enable a precise differentiation of the loss correction techniques based on the corruption type.

4.1 Preliminary Properties

When introducing Markov kernels in Definition 1, 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 as modification of the same set of objects while rearranging the probability measure defined on it. Hence, a corruption from $X \times Y$ to Y has to be considered as a *parameterized version* of the corruption on Y, where the parameter is x. For this reason, we also introduced the operation P4, which allows chaining while keeping a specified free parameter. A degenerate sub-case takes place when we deal with a kernel from X to Y. We will make use of the notation κ_y, κ_x to express the parameterization, which is a shortcut for $\kappa_{\mathsf{Y}=y} = \kappa_y$, $\kappa_{\mathsf{X}=x} = \kappa_x$ respectively.

Markov kernels prove themselves as 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., the one described in Section 19.2 of Aliprantis and Border (2006). Here we choose a less

13. 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).

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 § 2.2; positivity and boundedness together ensure the Fubini-Tonelli's theorem to be applied safely in the following proofs. We can now give a formal statement of the aforementioned property of kernels.

Lemma 20 (Data Processing Equality in Terms of Risk) *Consider a bounded loss ℓ, a model $h \in \mathcal{M}(X, Y)$, and a probability distribution P on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ standard Borel. Let κ be a joint corruption kernel in $\mathcal{M}(X \times Y, X \times Y)$. Then,*

$$\mathrm{R}_P[\kappa(\ell \circ h)] = \mathrm{R}_{P \circ \kappa}[\ell \circ h].$$

Proof We know that $\mathrm{R}_P[f] \coloneqq \int f dP$ and the kernel actions definition from § 2.1.1. Being the loss and kernel both bounded and positive, Fubini-Tonelli's theorem holds, and

$$\begin{aligned}
\int (\ell \circ h) dP &= \int_{(x,y)} \int_{(\tilde{x},\tilde{y})} \kappa(x, y, d\tilde{x}, d\tilde{y})\, \ell(h(\tilde{x}), \tilde{y})\, P(dx, dy) \\
&= \int_{(\tilde{x},\tilde{y})} \ell(h(\tilde{x}), \tilde{y}) \int_{(x,y)} \kappa(x, y, d\tilde{x}, d\tilde{y})\, P(dx, dy) \\
&= \int_{(\tilde{x},\tilde{y})} \ell(h(\tilde{x}), \tilde{y})(P \circ \kappa)(d\tilde{x}, d\tilde{y}).
\end{aligned}$$

■

This result is central to establishing the Data Processing Equality for constrained Bayes risk in § 4.3. What remains unclear is how the *form* of equality changes, specifically in terms of the clean distribution P and the minimization set of $\ell \circ \mathcal{H}$, 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 $I(\tau) = X$ and $I(\lambda) = Y$. Their full signature will be provided in each theorem. Also, we use the notation $\kappa\mathcal{F} \coloneqq \{\kappa f,\ \forall\, f \in \mathcal{F}\}$ for the action of a kernel on a compatible set of functions.

Remark 21 *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. However, Lemma 20 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).*

4.2 Existing Result: Data Processing Equalities for Simple Corruption

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 § F. In our taxonomy, their "combined noise" corresponds to our "combined simple corruption".

Proposition 22 (Combined simple noise, Williamson and Cranko (2024)) *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, P)$, $E\colon Y \rightsquigarrow X$ its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y, and $F\colon X \rightsquigarrow Y$ its associated posterior such that $P = \pi_X \times F$ for a suitable π_X. Let $(\tau\colon X \rightsquigarrow X) \otimes (\lambda\colon Y \rightsquigarrow Y)$ be a corruption acting on this problem. Then, the kernel action on P can be rewritten as*

$$P \circ (\tau \otimes \lambda) = (\pi_Y \times E) \circ (\tau \otimes \lambda) = \pi_Y \circ \big[(E \circ \tau) \otimes \lambda\big] \tag{5}$$

or, equivalently,

$$P \circ (\tau \otimes \lambda) = (\pi_X \times F) \circ (\tau \otimes \lambda) = \pi_X \circ \big[\tau \otimes (F \circ \lambda)\big]. \tag{6}$$

Then, the BR *Data Processing Equality,*

$$\Big(\ell \circ \mathcal{H}, P \circ (\tau \otimes \lambda)\Big) \equiv_{\text{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), P\Big),$$

holds such that the functions contained in the new minimization set are defined as

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

A Fundamental Difference Between Label and Attribute Corruption. Starting from the simpler result in Proposition 22, we can easily observe some properties of corruption that are conserved also in the next cases. Consider its hypotheses, i.e., $\tau \otimes \lambda = \delta_X \otimes \lambda$ being a corruption acting only on labels. The formula in Eq. (6) therefore tells:

$$\text{BR}_{\ell \circ \mathcal{H}}\big[\pi_X \circ \big(\delta_X \otimes (F \circ \lambda)\big)\big] = \text{BR}_{(\lambda\ell \circ \mathcal{H})}\big[\pi_X \times F\big].$$

When looking at the right-hand side, we see that the λ component only modifies the loss function, and leaves the model class untouched. On the other hand, when considering $\tau \otimes \lambda = \tau \otimes \delta_Y$, we obtain

$$\text{BR}_{\ell \circ \mathcal{H}}\big[\pi_Y \circ \big((E \circ \tau) \otimes \delta_Y\big)\big] = \text{BR}_{\tau(\ell \circ \mathcal{H})}\big[\pi_Y \times E\big]\ ,$$

and in this case notice that the action of $\kappa = \tau \otimes \lambda$ affects the whole minimization set when considering the Bayes risk on the clean distribution. These simple label (Markovian) corruptions are *equivalent in Bayes risk to loss corruptions*, which is non-Markovian in the sense of Definition 10, although induced by a Markov kernel.

4.3 Novel Data Processing Equalities for Other Corruptions

We now present the results for each of the remaining corruption combinations in Fig. 2. From now on, we will refer to the following formula as BR **Data Processing Equality**:

$$\Big(\ell \circ \mathcal{H}, P \circ (\tau \otimes \lambda)\Big) \equiv_{\text{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), P\Big). \tag{7}$$

The following two results are a strict generalization of Proposition 22.

Theorem 23 (2-dependent τ, simple λ) *Let ℓ be a bounded loss function. Consider the learning problem $(\ell, \mathcal{H}, P)$ and suppose $E\colon Y \rightsquigarrow X$ is its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y. Let $(\tau\colon X \times Y \rightsquigarrow X) \otimes (\lambda\colon Y \rightsquigarrow Y)$ be a corruption acting on this problem. Then,*

$$P \circ (\tau \otimes \lambda) = (\pi_Y \times E) \circ (\tau \otimes \lambda) = \pi_Y \circ \big[(E \circ_X \tau) \otimes \lambda\big]$$

and the BR *Data Processing Equality in Eq.* (7) *holds such that the functions contained in the new minimization set are defined as*

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

Here in Theorem 23 we have shown the BR equality for the experiment E. However, for some corruptions the equalities cannot be stated with E and the generative formulation of a learning problem, unless ignoring the joint corruption factorization formula. We hence use the posterior kernel F, i.e., the discriminative formulation of a learning problem, and gain more insights about the minimization set.

Theorem 24 (simple τ, 2-dependent λ) *Let ℓ be a bounded loss function. Consider the learning problem $(\ell, \mathcal{H}, P)$ and suppose $F\colon X \rightsquigarrow Y$ is its associated posterior such that $P = \pi_X \times F$ for a suitable π_X. Let $(\tau\colon X \rightsquigarrow X) \otimes (\lambda\colon X \times Y \rightsquigarrow Y)$ be a corruption acting on this problem. Then, the kernel action on P can be written as*

$$P \circ (\tau \otimes \lambda) = (\pi_X \times F) \circ (\tau \otimes \lambda) = \pi_X \circ \big[\tau \otimes (F \circ_Y \lambda)\big],$$

and the BR *Data Processing Equality in Eq.* (7) *holds such that the functions contained in the new minimization set are defined as*

$$\tau(\lambda\ell \circ \mathcal{H}) \coloneqq \{(x, y) \mapsto [\tau(\lambda\ell_{(x,y)} \circ h)](x), h \in \mathcal{H}\}.$$

We can notice, thanks to Theorems 23 and 24, that when corruption involves dependent structures in the factorization, the loss function or the whole minimization set are modified in a parameterized, *dependent* way. Consider, for instance, the action of $\lambda\colon X \times Y \rightsquigarrow Y$ on the minimization set, when $\tau = \delta_X$. By definition, it generates the measurable functions

$$\lambda\ell \circ \mathcal{H} = \{(x, y) \mapsto (\lambda\ell)(h_x, x, y) \,|\, h \in \mathcal{H}\} = \{(x, y) \mapsto (\lambda\ell_{(x,y)} \circ h)(x)\},$$

which is a strong change in the definition of the induced loss function $\tilde{\ell}\colon \mathcal{P}(Y) \times X \times Y \to \mathbb{R}_{\geq 0}$ considered, although a still valid choice; for example, it has been employed in Steinwart and Christmann (2008). We additionally underline here that *corruptions on 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 25 (1-dependent, 2-dependent) *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, P)$, suppose $E\colon Y \rightsquigarrow X$ is its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y, and $F\colon X \rightsquigarrow Y$ its associated posterior such that $P = \pi_X \times F$ for a suitable π_X.*

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

$$P \circ (\tau \otimes \lambda) = (\pi_Y \times E) \circ (\tau \otimes \lambda) = \pi_Y \circ \big[\tau \otimes (E \circ_X \lambda)\big] \tag{8}$$

The BR *Data Processing Equality in Eq.* (7) *holds such that the functions contained in the new minimization set are defined as*

$$\tau(\lambda\ell \circ \mathcal{H}) \coloneqq \{(x,y) \mapsto [\tau(\lambda\ell_{(x,y)} \circ h)](y), h \in \mathcal{H}\}.$$

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

$$P \circ (\tau \otimes \lambda) = (\pi_X \times F) \circ (\tau \otimes \lambda) = \pi_X \circ \big[(F \circ_Y \tau) \otimes \lambda\big]$$

The BR *Data Processing Equality in Eq.* (7) *holds such that the functions contained in the new minimization set are defined as*

$$\tau(\lambda\ell \circ \mathcal{H}) \coloneqq \{(x,y) \mapsto [\tau(\lambda\ell_x \circ h)](x,y), h \in \mathcal{H}\}.$$

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 to hold.

Corollary 26 (1-dependent τ, general λ) *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, P)$, $E\colon Y \rightsquigarrow X$ its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y, and $F\colon X \rightsquigarrow Y$ its associated posterior such that $P = \pi_X \times F$ for a suitable π_X. Let $(\tau\colon Y \rightsquigarrow X) \otimes (\lambda\colon X \rightsquigarrow Y)$ be a corruption acting on the problem. Then,*

$$P \circ (\tau \otimes \lambda) = (\pi_Y \times E) \circ (\tau \otimes \lambda) = \pi_Y \circ \big[\tau \otimes (E \circ \lambda)\big].$$

or, equivalently,

$$P \circ (\tau \otimes \lambda) = (\pi_X \times F) \circ (\tau \otimes \lambda) = \pi_X \circ \big[(F \circ \tau) \otimes \lambda\big]. \tag{9}$$

The BR *Data Processing Equality in Eq.* (7) *holds such that the functions contained in the new minimization set are defined as*

$$\tau(\lambda\ell \circ \mathcal{H}) \coloneqq \{(x,y) \mapsto [\tau(\lambda\ell_x \circ h)](y), h \in \mathcal{H}\}.$$

In all the Theorems involving a 1-dependent corruption, the minimization set is heavily modified. To better understand how, we take a closer look at the functions contained in the clean and corrupted minimization sets. First, we slightly rework the notation for the minimization set by considering the loss function $\ell(\cdot, y)$ as a parameterized one, i.e., $\ell_y\colon \mathcal{P}(Y) \to \mathbb{R}_{\geq 0}$; then, the set $\ell \circ \mathcal{H} \coloneqq \{\, (x,y) \mapsto \ell(h_x, y) \,|\, h \in \mathcal{H} \subseteq \mathcal{M}(X,Y) \,\}$ can be equivalently rewritten as $\{\, (x,y) \mapsto (\ell_y \circ h)(x) \,|\, h \in \mathcal{H} \subseteq \mathcal{M}(X,Y) \,\}$.

In Eq. (8), we have again the kernel $\lambda \in \mathcal{M}(X \times Y, Y)$ acting on the loss; hence, we obtain $\tilde{\ell}_{(x,y)} = \tilde{\ell}(\cdot, x, y) \coloneqq (\lambda\ell)(\cdot, x, y)$. Therefore, we are again inducing a more general notion of loss, namely $\tilde{\ell}\colon \mathcal{P}(Y) \times X \times Y \to \mathbb{R}_{\geq 0}$. 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 $\tau \in \mathcal{M}(Y, X)$, which "swaps" the input

$\tilde{x} \in X$ with $y \in Y$ in addition to modifying the function itself. Combining them together, we get the new minimization set containing functions of the form $f(x,y) = \big[\tau\big(\tilde{\ell}_{(x,y)} \circ h\big)\big](y)$, which is not anymore comparable with the initial form $\ell_{\tilde{y}} \circ h(\tilde{x})$, nor interpretable as a performance evaluation for the model h.

A similar strong modification is observed for the minimization set in Eq. (9), which contains functions of the form $f(x,y) = \big[\tau\big(\tilde{\ell}_x \circ h\big)\big](y) \coloneqq \big[\tau\big((\lambda\ell)_x \circ h\big)\big](y)$. That is caused both by the action of $\lambda \in \mathcal{M}(X,Y)$ on $\ell(\cdot,y)$, which results in a new loss function $(\lambda\ell)_x(\cdot) \coloneqq \lambda\ell(\cdot,x)$, as well as the action of τ on $\ell \circ h$.

The final result of the factorization, involving $\tau\colon X \times Y \rightsquigarrow X$ and $\lambda\colon X \times Y \rightsquigarrow Y$, yields a negative implication as detailed in the following.

Theorem 27 (2-dependent κ, general λ) *Let ℓ be a bounded loss function. Consider the clean learning problem $(\ell, \mathcal{H}, P)$, and let $(\tau\colon X \times Y \rightsquigarrow X) \otimes (\lambda\colon X \times Y \rightsquigarrow Y)$ be a corruption acting on the problem. Then:*

1. *the action of such corruption on the joint probability P is equivalent to the one of the non-factorized joint corruption;*
2. *the* BR *Data Processing Equality in Eq.* (7) *holds;*
3. *the functions contained in the new minimization set are defined as*

$$\tau(\lambda\ell \circ \mathcal{H}) \coloneqq \big\{(x,y) \mapsto [\tau(\lambda\ell_{(x,y)} \circ h)](x,y), h \in \mathcal{H}\big\}.$$

This result is due to the full dependence on the joint space $X \times Y$ for both τ and λ, making it impossible in general to derive a meaningful decomposition of the action on P via P1, P2 and P4. However, we can still distinguish the effect of λ on the loss, as achieved in all previous cases, and of τ on the full minimization set. For a detailed analysis and proof, see § F.

5. Loss-correction Approaches to Corruption Mitigation

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

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

In our chosen formalization, the training data are drawn from a corrupted distribution $\tilde{P} := P \circ \kappa$, while evaluation is performed on data drawn from the original clean distribution P. We define **accurate learning** as the property of a learning problem for which the optimal model, obtained minimizing risk over a constrained model class using corrupted training data, *coincides* with the model obtained training on clean data with the same model class.

While prior work has already theoretically explored corruption mitigation, often by constructing corrected loss functions tailored to specific types of corruption, such methods are typically limited to a specific corruption 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.

5.1 Existing Work on Corruption-corrected Learning for Label Corruption

A vast amount of theoretical research on corruption-corrected learning has been carried in the fields of *learning with noisy labels* and *learning under distribution shift*. Their goal has been to achieve *unbiased learning* from biased data,[14] where "biased data" means corrupted data in our context, while "unbiased" refers to the use of a corrected loss $\tilde{\ell}$ that yields

$$\mathrm{R}_{\tilde{P}}[\tilde{\ell} \circ h] = \mathrm{R}_P[\ell \circ h], \tag{10}$$

with $\mathrm{R}_P[f] \coloneqq \int f dP$. This equality can be obtained as a direct consequence of Lemma 20 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

$$h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathrm{R}_{\tilde{P}}[\tilde{\ell} \circ h] \quad \text{and} \quad h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathrm{R}_P[\ell \circ h], \tag{11}$$

and therefore constitutes a stronger requirement. We refer to the family of approaches for achieving the above goals collectively as **corruption-corrected learning** (CL).

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

5.1.1 Reconstruction-based Method

The first method achieves 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 $\lambda \colon Y \rightsquigarrow Y$ 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

$$\tilde{\ell}(h_x, \tilde{y}) \coloneqq \lambda^* \ell(h_x, \tilde{y}) \coloneqq \sum_y \lambda^*_{\tilde{y}y} \ell(h_x, y), \quad \lambda \colon Y \rightsquigarrow Y, \tag{12}$$

where λ^* is the reconstruction matrix and $\lambda^* \ell$ is the matrix acting on the loss vector, indexed by the label. This yields unbiased learning as per Eq. (10), since for all x,

$$\begin{aligned} \mathbb{E}_{\tilde{\mathsf{Y}} \sim F \circ \lambda}[\tilde{\ell}(h_x, \tilde{\mathsf{Y}})] &\overset{L20}{=} \mathbb{E}_{\mathsf{Y} \sim F} \lambda(\tilde{\ell} \circ h_x)(\mathsf{Y}) \\ &\overset{(12)}{=} \mathbb{E}_{\mathsf{Y} \sim F} \lambda^* \lambda(\ell \circ h_x)(\mathsf{Y}) \\ &= \mathbb{E}_{\mathsf{Y} \sim F}[\ell(h_x, \mathsf{Y})]. \end{aligned}$$

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

We note that 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., $\delta_X \otimes \lambda$, $\lambda \colon Y \rightsquigarrow Y$, and do not generalize to more complex joint corruption cases.

14. Existing literature defines unbiased learning via unbiasedness of the empirical risk estimator. This notion is trivially implied by our Eq. (10).

5.1.2 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 $I(\tau) = X$ while the conditional distribution F remains unchanged. Under model misspecification (White, 1981),[15] IW provides a principled correction by requiring the clean data distribution to be absolutely continuous w.r.t. the corrupted one, i.e., $P \ll \tilde{P}$. The corrected loss takes the form

$$\tilde{\ell}(h(\boldsymbol{x}), \tilde{y}) \coloneqq w(\boldsymbol{x})\, \ell(h(\boldsymbol{x}), \tilde{y}),\ w(\boldsymbol{x}) \coloneqq \frac{dP}{d\tilde{P}}(\boldsymbol{x}), \tag{13}$$

where $w(\boldsymbol{x})$ is called the importance weight, typically expressed via densities w.r.t. the Lebesgue measure. This guarantees the unbiased learning goal in Eq. (10).

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 $w(x, y) \coloneqq \frac{dP}{d\tilde{P}}(x, y)$ 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 P and $\tilde{P}$), which is essential for the importance weights to be well-defined.

At first glance, the IW correction formula Eq. (13) 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 $\frac{dP}{d\tilde{P}}$, 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 § 3.2.3.

5.2 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 assumptions, 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.

5.2.1 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.

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

Definition 28 *Let $P \in \mathcal{P}(Z_1)$ be a probability distribution, and $\kappa \in \mathcal{M}(Z_1, Z_2)$ be a Markov kernel with the property $P \circ \kappa = \tilde{P}$ for some $\tilde{P} \in \mathcal{P}(Z_2)$. The* Bayesian inverse *of κ is defined as the Markov kernel $\kappa^\dagger \in \mathcal{M}(Z_2, Z_1)$ such that it induces together with $\tilde{P}$ the same coupling induced by κ and P, i.e.,*

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

By taking A or B equal to the Z_2 or Z_1, we respectively get the property of $\kappa^\dagger$ and κ being a "weak" inverse of each other.

Proposition 29 *Let $\kappa \colon Z_1 \rightsquigarrow Z_2$ be a Markov kernel with the property $\tilde{P} = P \circ \kappa$ for $P \in \mathcal{P}(Z_1), \tilde{P} \in \mathcal{P}(Z_2)$, and $\kappa^\dagger$ its Bayesian inverse. Then, it reverses the action on the fixed input and output probabilities, i.e.,*

$$P(A) = (\tilde{P} \circ \kappa^\dagger)(A), \ \forall\, A \in \mathcal{Z}_2.$$

Remark 30 *Recalling how we have defined the posterior kernel $F \in \mathcal{M}(X, Y)$ in Theorem 5, i.e., given $\pi_X \in \mathcal{P}(X)$ and $\pi_Y \in \mathcal{P}(Y)$,*

$$\pi_X \times F = \pi_Y \times E = P \in \mathcal{P}(X \times Y), \tag{14}$$

we notice that $F = E^\dagger$.

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 Z_1 and 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 § E). This is a weaker existence condition w.r.t. the one given for the reconstruction matrix.[16]

Remark 31 *In the discrete case, the Bayesian inverse always exists and is defined by Bayes rule. Furthermore, it is uniquely defined $\tilde{P}$-a.s. This easy to see by unfolding Definition 28 into:*

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

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

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

This formula ensures the uniqueness of $\kappa^\dagger$ within the support of $\tilde{P}$, as all components are unique. However, outside the support when $\tilde{P}$ is zero, the uniqueness may not hold, requiring a non-fixed value for $z_2 \in Z_2$ where $\tilde{P}(z_2) = 0$.

16. 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 P; having a different P induces a different Bayesian inverse. This point becomes clearer when considering the discrete case, as explained in Remark 31.

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

Proposition 32 (Inversed Data Processing Equality in Terms of Risk) *Consider a learning problem $(\ell, \mathcal{H}, P)$ on $(X \times Y, \mathcal{X} \times \mathcal{Y})$, a corruption $\kappa \in \mathcal{M}(X \times Y, X \times Y)$, and a function $f \in \ell \circ \mathcal{M}(X, Y)$. Let ℓ be a bounded loss function. Then,*

$$\mathrm{R}_P[f(\mathsf{Z})] = \mathrm{R}_{P \circ \kappa}[\kappa^\dagger f(\tilde{\mathsf{Z}})], \tag{15}$$

and

$$\mathrm{BR}_{\ell \circ \mathcal{H}}\big(P\big) = \mathrm{BR}_{\kappa^\dagger(\ell \circ \mathcal{H})}\big(P \circ \kappa\big), \tag{16}$$

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

Proof The first claim follows by applying Lemma 20 for $P \circ \kappa \circ \kappa^\dagger = P$, where the equality holds because of Proposition 29. The second is proved by taking the infimum over $\ell \circ \mathcal{H}$ of both sides of Eq. (15). ■

Remark 33 *Notice that Proposition 32 above does not imply $\ell \circ h^* = \kappa^\dagger(\ell \circ h^*)$, but only that their risks coincide, i.e., $\mathbb{E}_P(\ell \circ h^*) = \mathbb{E}_{P \circ \kappa}[\kappa^\dagger(\ell \circ h^*)]$. Under certain conditions on the learning problem, there exists some $h^*, \tilde{h}^* \in \mathcal{H}$ such that $\ell \circ h^* = \kappa^\dagger(\ell \circ \tilde{h}^*)$, but it is not necessarily the case that $\tilde{h}^* = h^*$.*

In the following we will make use of these facts assuming $Z_1 = Z_2 = X \times Y =: Z$ to match the setting of our taxonomy of corruption.

5.2.2 Label Corruption Correction with Bayesian Inverse

Similarly to the reconstruction-matrix and importance-weighting approaches, we enforce the condition in Eq. (10) 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 $\lambda \colon Y \rightsquigarrow Y$. Consider its associated cleaning kernel sending $\tilde{\mathcal{L}} = (\tilde{\ell}, \mathcal{H}, \tilde{P})$ to the clean one $\mathcal{L} = (\ell, \mathcal{H}, P)$, i.e., the Bayesian inverse $\lambda^\dagger \colon Y \rightsquigarrow Y$. By considering a corruption $\delta_X \otimes \lambda$ and its inverse $\delta_X \otimes \lambda^\dagger$, Proposition 32 allows us to write

$$\mathrm{R}_{\tilde{P}}(\lambda^\dagger \ell \circ h) = \mathrm{R}_P(\ell \circ h), \quad \text{where } \tilde{P} := P \circ (\delta_X \otimes \lambda), \quad \forall\, h \in \mathcal{H}\,.$$

This directly yields the loss correction formula:

$$\tilde{\ell}(h_x, \tilde{y}) \coloneqq \lambda^\dagger \ell(h_x, \tilde{y}), \quad \lambda^\dagger \colon Y \rightsquigarrow Y. \tag{17}$$

The corrected loss satisfies the unbiased learning criterion in Eq. (10), and hence also achieves the accurate learning goal in Eq. (11).

Notably, in this simple label corruption case, our kernel-based correction in Eq. (17) resembles the reconstruction-matrix correction in Eq. (12). However, the two differ fundamentally, as illustrated below.

Example 2 *Following van Rooyen and Williamson (2018, Sec. 4.1.2), consider symmetric label corruption in binary classification setting where the clean distribution is given by* $P = (p_1, p_2)^\top$. *Suppose* $\sigma \in (0, 0.5)$*; let the corruption kernel be*

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

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 $\lambda \colon Y \rightsquigarrow Y$ as discussed above, the Bayesian inverse naturally takes the form $\lambda^\dagger : Y \rightsquigarrow Y$. 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 only assume that $\kappa^\dagger$ is **one-step Markovian** and therefore admits the form $\tau \otimes \lambda$.

We now extend this paradigm to the $\lambda \colon X \times Y \rightsquigarrow Y$ case, as from Theorem 24 we see that also in the dependent case, label corruption only affects the loss function.

Theorem 34 *Let* $(\ell, \mathcal{H}, P)$ *be a clean learning problem with* ℓ *being a bounded loss function, and* $\kappa = \delta_X \otimes \lambda$ *a one-step Markovian corruption on it. Let* $\kappa^\dagger$ *be the cleaning kernel inverting* κ*, such that* $(\kappa^\dagger(\ell \circ \mathcal{H}), P \circ \kappa)$ *is its associated corrected problem. When* $\lambda \in \mathcal{M}(Y, Y)$*, we have*

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

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

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

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

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

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

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

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

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

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

where $(h(\tilde{x}), \tilde{x}, \tilde{y}) \in \mathcal{P}(Y) \times X \times Y$. 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 X do not change the hypothesis function. Lastly, by definition of Markov transition, we have that when $\lambda \in \mathcal{M}(Y, Y)$

$$\tilde{\ell}(h(\tilde{x}), \tilde{y}) \coloneqq \sum_{y \in Y} \int_{x \in X} \ell(h(x), y)\, \delta(\tilde{x}, dx)\, \lambda(\tilde{y}, dy) = (\lambda\ell)(h(\tilde{x}), \tilde{y}), \tag{18}$$

which respects the classical definition of loss function, as $(h(\tilde{x}), \tilde{y}) \in \mathcal{P}(Y) \times Y$. This proves the thesis, as by construction $\tilde{\ell}(h(\tilde{x}), \tilde{x}, \tilde{y}) \in \kappa^\dagger(\ell \circ \mathcal{H})$. Proposition 32 implies that $\kappa^\dagger(\ell \circ \mathcal{H})$ makes Eq. (10) hold. ■

For the dependent label corruption 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).

5.2.3 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 34. This is because, when applying the risk conservation formula in Proposition 32, the model class itself is also affected by the corruption kernel. Formally, assume $\kappa^\dagger = \tau \otimes \lambda$ with $I(\tau) = X$ and $I(\lambda) = Y$. Then, Eq. (10) becomes

$$\mathrm{R}_P(\ell \circ h) = \mathrm{R}_{P \circ \kappa}\Big[\kappa^\dagger(\ell \circ h)\Big] = \mathrm{R}_{P \circ \kappa}\Big[(\tau \otimes \lambda)(\ell \circ h)\Big], \quad \forall h \in \mathcal{M}(X, Y), \tag{19}$$

which does not necessarily imply Eq. (11). In Eq. (19), the corruption effect on loss and model class is in general *indistinguishable*; we are not able to rewrite the set $(\tau \otimes \lambda)(\ell \circ \mathcal{H})$ 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 weakened 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

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathrm{R}_{\tilde{P}}[\tilde{\ell} \circ h] \quad \text{and} \quad h^* \in \arg\min_{h \in \mathcal{H}} \mathrm{R}_P[\ell \circ h] \tag{20}$$

with the additional *factorization requirement* of

$$\tilde{\ell}(h, x, y) = \kappa^\dagger(\ell \circ h)(x, y), \quad \forall h \in \mathcal{M}(X, Y), x \in X, y \in Y, \tag{21}$$

where $\tilde{\ell}$ is a *generalized loss* $\tilde{\ell}\colon \mathcal{H} \times X \times Y \to \mathbb{R}_{\geq 0}$.[17]

A key assumption made in the above paradigm is the existence of a minimizer for the problems $(\ell, \mathcal{H}, P)$ and $(\tilde{\ell}, \mathcal{H}, P)$; 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 for the implying well-posedness of GCL.

A 1 *Given the clean learning problem $(\ell, \mathcal{H}, P)$, there exists a minimizer $h^* \in \mathcal{H}$ that attains the Bayes risk value associated to it, i.e.,* $\mathrm{BR}_{\ell \circ \mathcal{H}}(P) = \mathrm{R}_P[\ell \circ h^*]$.

Lemma 35 *Let $(\ell, \mathcal{H}, P)$ be a clean learning problem respecting A 1, and $\kappa = \tau \otimes \lambda$ a Markovian corruption kernel in $\mathcal{M}(X \times Y, X \times Y)$ on it. Denote the risk minimizer for the clean problem as $h^* \in \mathcal{H}$. Then, if Eq.* (21) *is fulfilled by some $\tilde{\ell}$, h^* is also the minimizer of the* GCL*-corrected problem $(\tilde{\ell}, \mathcal{H}, P \circ \kappa)$, as per Eq.* (20)*.*

Proof Consider the problem $(\kappa^\dagger(\ell \circ \mathcal{H}), P \circ \kappa)$ and assume that there is some (possibly generalized) loss function $\tilde{\ell}$ such that Eq. (21). By Proposition 32 and Eq. (21),

$$\mathrm{R}_P(\ell \circ h) \overset{P32}{=} \mathrm{R}_{P \circ \kappa}\Big[(\tau \otimes \lambda)(\ell \circ h)\Big] \overset{(21)}{=} \mathrm{R}_{P \circ \kappa}(\tilde{\ell} \circ h), \tag{22}$$

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

Hence, finding loss correction formulas respecting Eq. (21) and A1 allows us to use the GCL correction paradigm.

We now state and prove the correction results for all the corruption case lying within the GCL paradigm. Recall that the notation $f\#\mu$ stands for the push-forward probability measure of the distribution μ through the function f, defined as $(f\#\mu)(A) \coloneqq \mu(f^{-1}(A))$ for a suitable set A. In the following we will use such notation for kernels. By definition of kernel, τ s.t. $I(\tau) = X$ is a measure when fixing $\tilde{x} \in X$, and considering the kernel $h \in \mathcal{M}(X, Y)$ as its associated function $h\colon X \to \mathcal{P}(Y)$ yields a measurable function; hence, we can write

$$(h\#\tau)(\tilde{x})(A) \coloneqq \tau(\tilde{x}, h^{-1}(A)),\ A \subset \mathcal{P}(Y),$$

that is a family of distributions defined on a set of probability measures on Y, evaluated on 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 $\mathcal{P}(Y)$, given $\tilde{x}$.

Theorem 36 *Let $(\ell, \mathcal{H}, P)$ be a clean learning problem with ℓ being a bounded loss function and such that A 1 holds. Let $\kappa^\dagger = \tau \otimes \lambda \in \mathcal{M}(X \times Y, X \times Y)$ be the one-step Markovian cleaning kernel inverting κ, such that $(\kappa^\dagger(\ell \circ \mathcal{H}), P \circ \kappa)$ is its associated corrected problem. Then, we can find a generalized loss $\tilde{\ell}\colon \mathcal{H} \times X \times Y \to \mathbb{R}_{\geq 0}$ respecting the* GCL *paradigm. In particular:*

17. 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.

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

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

   When both corruptions are simple, the $\lambda\ell$ formula remains unchanged. When λ is 2-dependent, it induces $\lambda\ell\,(\mathsf{u}, \tilde{x}, \tilde{y})$. Lastly, we get $(h\#\tau)(\tilde{x})$ to be replaced by $(h\#\tau)(\tilde{x}, \tilde{y})$ when τ is 2-dependent.

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

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

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

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

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

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

Proof Let us consider a general function in the set $\kappa^\dagger(\ell \circ \mathcal{H})$,

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

By Theorem 35, we are ensured that finding a loss correction formula $\tilde{\ell}$ that respects Eq. (21) is enough for ensuring that such loss respects the GCL paradigm. Hence, we consider here the function in the sets $\kappa^\dagger(\ell \circ \mathcal{H})$ and define an ad-hoc $\tilde{\ell}$ depending on how $\kappa^\dagger$ affects $\ell \circ \mathcal{H}$.

We have already seen in Theorem 34 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 the $\kappa^\dagger$ from such that $\kappa^\dagger = \tau \otimes \lambda$ with $\tau \in \mathcal{M}(X, X)$ and $\lambda \in \mathcal{M}(X \times Y, Y)$. We can define the loss correction $\tilde{\ell}$ as

$$\begin{aligned}
\tilde{\ell}(h, \tilde{x}, \tilde{y}) &\coloneqq \sum_{y\in Y} \int_{x\in X} \ell(h(x), y)\,\tau(\tilde{x}, dx)\,\lambda(\tilde{x}, \tilde{y}, dy) \\
&= \sum_{y\in Y} \int_{x\in h(X)} \ell(u, y)\,\tau(\tilde{x}, h^{-1}(du))\,\lambda(\tilde{x}, \tilde{y}, dy) \\
&= \int_{x\in h(X)} (\lambda\ell)_{\tilde{x}}(u, \tilde{y})\,\tau(\tilde{x}, h^{-1}(du)), \qquad (23)
\end{aligned}$$

where $u = u(dy) \in \mathcal{P}(Y)$ and $h(X) \coloneqq \{h(x) \,|\, x \in X\}$. The following equality holds:

$$\mathbb{E}_{\mathsf{u}\sim\tau(\tilde{x},h^{-1}(\cdot))}[\mathsf{u}] = \int_{x\in h^*(X)} u\,\tau(\tilde{x}, h^{-1}(du)) = (h\#\tau)(\tilde{x}) \in \mathcal{P}(Y),$$

by definition of $\mathcal{H}$ as a subset of $\mathcal{M}(X,Y)$ and using the definition of $h\#\tau$. We remark that $\tau(\tilde{x},(h)^{-1}(du))$ is then a probability in $\mathcal{P}(\mathcal{P}(Y))$. Hence we can rewrite Eq. (23) as

$$\tilde{\ell}(h,\tilde{x},\tilde{y}) \coloneqq \int_{u\in\mathcal{P}(Y)} (\lambda\ell)_{\tilde{x}}(u,\tilde{y})\,\tau(\tilde{x},h^{-1}(du)) = \mathbb{E}_{\mathsf{u}\sim(h\#\tau)(\tilde{x})}[(\lambda\ell)(\mathsf{u},\tilde{x},\tilde{y})].$$

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 from $\tilde{y}$. Therefore we obtain $\tilde{\ell}(h,\tilde{x},\tilde{y}) = \mathbb{E}_{\tau(\tilde{x},\tilde{y},h^{-1}(\cdot))}[(\lambda\ell)(\mathsf{u},\tilde{x},\tilde{y})]$, where only the simple label corruption can be considered, given the missing result for the BR equality in the $D(\tau) = D(\lambda) = X \times Y$ 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 25 and 27, and obtain the thesis by fulfilling the GCL requirement in Eq. (21). ■

5.2.4 Discussion of the GCL Paradigm and Its Correction Formulas

Theorem 36 provides a constructive proof for the existence of GCL corrected losses, i.e., that respect both Eq. (20) and Eq. (21), 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. (20) is well-posed. Within our setup, we could fulfill the assumption by requiring the loss ℓ not only to be positive, measurable and bounded but also *continuous* on $\mathcal{P}(Y)$ in the Euclidean topology, and $\mathcal{H}$ to be *compact*. This would imply continuity of the functional $h \in \mathcal{H} \mapsto \mathrm{R}_P[\ell(h(x),y)]$, and therefore the existence of a minimizer on the compact set $\mathcal{H}$.

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. (21). 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 3 *Consider a simple neural network with one hidden layer of two ReLU neurons and a softmax output,*

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

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

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

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 $X \subset \mathbb{R}^2$ for which the ReLU activation makes the model $h(x)$ 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*

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

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

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

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

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

Imposing $\tilde{\ell}(h(\boldsymbol{x}), y) = \ell(h(f(\boldsymbol{x})), y)$ as per the CL *definition (i.e., stretching Eq.* (17) *to the attribute corruption case), with $\tilde{\ell}\colon \mathcal{P}(Y) \times Y \to \mathbb{R}_{\geq 0}$, implies that*

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

for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{S}_h$ and for all $y \in Y$. This yields a contradiction. Indeed, since $h(x)$ is constant on $\mathcal{S}_h$, any standard corrected loss of the form $\tilde{\ell}(h(x), y)$ 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 $h(f(\boldsymbol{x}_1)) \neq h(f(\boldsymbol{x}_2))$, which implies

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

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

The factorization requirement in Eq. (21) is the key property that defines GCL, as it implies

$$\mathrm{R}_{P \circ \kappa}(\tilde{\ell} \circ h^*) = \mathrm{R}_{P \circ \kappa}\left[\kappa^\dagger(\ell \circ h^*) \right],$$

and by Proposition 32,

$$\mathrm{R}_{P \circ \kappa}(\tilde{\ell} \circ h^*) = \mathrm{R}_P(\ell \circ h^*).$$

Since the factorization holds for every $h \in \mathcal{H}$, it follows that h^* is also a minimizer for the factorized corrected problem $(\tilde{\ell}, \mathcal{H}, P \circ \kappa)$. Investigating which conditions on the learning problem ensure the factorization property is beyond the scope of this analysis. The conclusion to be taken from our loss correction formulas is not that they provide a new

practical tool, but rather that they demonstrate how, even under optimistic assumptions, loss correction techniques become far more complex and involved when dealing with attribute corruption kernels.

Indeed, all the novel correction formulas found in this chapter are more complex than the ones defined in previous works (van Rooyen and Williamson, 2018; Patrini et al., 2017), which only consider a label corruption scenario similar to our CL for simple label corruption. Our version, Theorem 34, further extends the setting by including the dependent label corruption. The second set of results, included in Theorem 36, 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.

6. Conclusions

We proposed a comprehensive and unified framework for general corruption, extending its definition also to model class and loss function changes. We did so by using Markov kernels, and systematically studying corruption in three key aspects: typology, consequence, and correction. The choice of working with Markov kernels enables the use of information-theoretical tools, and provides an alternative interpretation of corruption as an *observation channel* through which we get to see our data distribution. This mathematical modelization allows one to consider data as a dynamic element of a learning problem, as opposed to the view of data as static facts and true representations of reality.

We established a new taxonomy for Markovian corruption of learning problems, yielding qualitative comparisons between corruption types in terms of their hierarchy. To gain a deeper understanding of corruption, we analyzed their consequences by proving Data Processing Equalities for Bayes risk. Given different possible factorizations of a corruption of the joint space, the learning problem is affected in different ways. Furthermore, we applied the equalities for obtaining loss correction formulas. Such an application is rather conventional, and usually leads to a proposed mitigation for the specific model considered. This work does not propose any mitigation algorithm, but analyzes the fundamental difference between label and attribute corruption. The Data Processing Equality results together with the analysis carried out in Section § 5 lead us to the following conclusions:

- Label and attribute corruption differ in how they change the learning problem. The former does not influence the model class; the latter changes model class and loss function in a way that generally cannot be disentangled.
- Classical corruption-corrected learning (CL) is not an adequate paradigm to study general corruption. For cases involving non-identical attribute corruption, we introduce a more general framework named generalized corruption-corrected learning (GCL).
- Loss correction formulas for attribute corruptions involve the notion of generalized loss and an expectation over the set of all $h\#\tau$ predictions. This suggests a *negative* result, as standard loss correction techniques may not guarantee accurate learning when dealing with general attribute corruption.

6.1 Limitations and Future Work

We considered data as probability distributions, implicitly assuming that each dataset has an associated probabilistic generative process. For many applications of machine learning, such an assumption is not warranted. Corruption is being induced by a Markov kernel, under the strong assumption of having full access to their actions. We note that in some cases Markov kernels can be estimated from corrupted data (Liu and Tao, 2015; Scott, 2015), but this question is in general still open and needs further investigation. The consequences of corruption are analyzed through Bayes risk without accounting for sampling or imperfect optimization. Bridging the gap between the distributional-level and the sample-level results would be the next step for this study, which requires tailored ad-hoc analyses. Other directions for making this framework more practically usable include developing quantitative methods to compare corruption severity and investigating the effects of optimization algorithms on the analysis.

From a more theoretical point of view, future work includes investigating the non-Markovian and multi-step classes of corruptions. As we pointed out in § 3, model misspecification lies within the general corruption class, and might be studied alone or as an additional corruption "chained" to a Markovian one. Similarly, changes in loss function can be analyzed further. Additionally, the topic of non-probabilistic corruption (Meng, 2022; Boyd et al., 2023), only superficially touched in the present work, needs deeper analysis. It is unclear whether the current theoretical tools, deployed when dealing with distributional changes, are enough for characterizing and potentially mitigating their consequences on learning problems.

Acknowledgments

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645, as well as by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Laura Iacovissi. Additional thanks to Armando J. Cabrera Pacheco, Nicolò Zottino and Jack Brady for helpful discussions, as well as Christian Fröhlich, Wenkai Xu and one anonymous reviewer for giving thorough and valuable feedback on earlier versions.

Appendix A. Glossary

$(Z, \mathcal{Z})$	Measurable space
$(X, \mathcal{X})$	Attribute space
$(Y, \mathcal{Y})$	Label space
μ and ν	Positive measures
P	Probability measure
$\mathcal{P}(Z)$	Set of probability measures on Z with Borel σ-algebra
$L^0(Z, \mathbb{R})$	Set of measurable functions w.r.t. $\mathcal{Z}$ (specified in text, usually Borel), from Z to the reals
Z	Random variable
$\kappa \colon Z \rightsquigarrow W$	Markov kernel from $(Z, \mathcal{Z})$ to $(W, \mathcal{W})$
$\mathcal{M}(Z, W)$	Set of Markov kernels from $(Z, \mathcal{Z})$ to $(W, \mathcal{W})$
$D(\kappa)$	Domain of a Markov kernel
$I(\kappa)$	Image of a Markov kernel
τ s.t. $I(\tau) = X$	Attribute corruption kernel
λ s.t. $I(\lambda) = Y$	Label corruption kernel
$\kappa_\nu \equiv \nu, \nu \in \mathcal{P}(Z)$	Degenerate kernel, constantly equal to a probability distribution
$\delta_Z \colon Z \rightsquigarrow Z$	Dirac delta kernel from $(Z, \mathcal{Z})$ to $(Z, \mathcal{Z})$
κ_z and κ_{Z}	Markov kernel evaluated on a point z (resp. random variable Z)
κf	Kernel action on functions
$\mu\kappa$ or $\kappa \circ \mu$	Kernel action on probabilities

$\kappa^\dagger \colon W \rightsquigarrow Z$	Bayesian Inversion of Markov kernel κ
$\#$	Push forward measure
$\circ$	Kernel chain composition
$\times$	Kernel product composition
$\otimes$	Kernel superposition
$\circ_Z$	Kernel partial chain composition w.r.t. Z
$E \colon Y \rightsquigarrow X$	Experiment
$F \colon X \rightsquigarrow Y$	Posterior kernel
π	Prior probability measure
$\ell \colon \mathcal{P}(Y) \times Y \to \mathbb{R}_{\geq 0}$	Loss function
$\ell \colon \mathcal{H} \times X \times Y \to \mathbb{R}_{\geq 0}$	Generalized loss function
$\mathcal{H} \subseteq \mathcal{M}(X, Y)$	Model class
$(\ell, \mathcal{H}, P)$	Learning context; mostly used as learning problem assuming criterion to be risk minimization
$\tilde{z}\,, \tilde{Z}\,,$ and $\tilde{P}$	Corrupted object, set, and probability
BR	Bayes risk
$\equiv_{\text{BR}}$	Equivalence relation of the set of learning problems w.r.t. Bayes risk

Table 4: Corruption types, Bayes risk equalities, and loss correction formulas. Each corruption kernel κ is factorized as τ (acting on inputs X) and λ (acting on labels Y), modifying both the data distribution P and the minimization set $\ell \circ \mathcal{H}$. The corresponding Data Processing Equality for Bayes risk is $\Big(\ell \circ \mathcal{H}, P \circ (\tau \otimes \lambda)\Big) \equiv_{\mathrm{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), P\Big)$ (§ 4), and the resulting loss correction formulas (§ 5) are summarized. All operators are expressed using the kernel operations defined in § 2.1.2: P1 chain composition ($\circ$), P2 product composition ($\times$), P3 superposition ($\otimes$), and P4 partial chain composition ($\circ_{(.)}$).

Corruption type $\kappa \coloneqq \tau \otimes \lambda$	Corruptions action $P \to \tilde{P} \coloneqq P \circ (\tau \otimes \lambda)$	Corruption action $\tau(\lambda\ell \circ \mathcal{H})$ on the minimization set $\ell \circ \mathcal{H} \coloneqq \{(x,y) \mapsto \ell(h_x, y) \,|\, h \in \mathcal{H}\}$	Loss correction formula $\tilde{\ell}(h, \tilde{x}, \tilde{y}),\ \forall (\tilde{x}, \tilde{y}) \in X \times Y$
$(\tau\colon X \rightsquigarrow X) \otimes$ $(\lambda\colon Y \rightsquigarrow Y)$	$\tilde{P} = \pi_Y \circ \big[(E \circ \tau) \otimes \lambda\big]$ $\tilde{P} = \pi_X \circ \big[\tau \otimes (F \circ \lambda)\big]$	$\{(x,y) \mapsto [\tau(\lambda\ell_y \circ h)](x), h \in \mathcal{H}\}$	$\mathbb{E}_{\mathsf{u} \sim (h \# \tau)(\tilde{x})}[\lambda\ell\,(\mathsf{u}, \tilde{y})]$
$(\tau\colon X \times Y \rightsquigarrow X) \otimes$ $(\lambda\colon Y \rightsquigarrow Y)$	$\pi_Y \circ \big[(E \circ_X \tau) \otimes \lambda\big]$	$\{(x,y) \mapsto [\tau(\lambda\ell_y \circ h)](x,y), h \in \mathcal{H}\}$	$\mathbb{E}_{\mathsf{u} \sim (h \# \tau)(\tilde{x}, \tilde{y})}[\lambda\ell\,(\mathsf{u}, \tilde{y})]$
$(\tau\colon X \rightsquigarrow X) \otimes$ $(\lambda\colon X \times Y \rightsquigarrow Y)$	$\pi_X \circ \big[\tau \otimes (F \circ_Y \lambda)\big]$	$\{(x,y) \mapsto [\tau(\lambda\ell_{(x,y)} \circ h)](x), h \in \mathcal{H}\}$	$\mathbb{E}_{\mathsf{u} \sim (h \# \tau)(\tilde{x})}[\lambda\ell\,(\mathsf{u}, \tilde{x}, \tilde{y})]$
$(\tau\colon Y \rightsquigarrow X) \otimes$ $(\lambda\colon X \times Y \rightsquigarrow Y)$	$\pi_Y \circ \big[\tau \otimes (E \circ_X \lambda)\big]$	$\{(x,y) \mapsto [\tau(\lambda\ell_{(x,y)} \circ h)](y), h \in \mathcal{H}\}$	$\mathbb{E}_{\mathsf{u} \sim (h \# \tau)(\tilde{y})}[\lambda\ell\,(\mathsf{u}, \tilde{x}, \tilde{y})]$
$(\tau\colon X \times Y \rightsquigarrow X) \otimes$ $(\lambda\colon X \rightsquigarrow Y)$	$\pi_X \circ \big[(F \circ_Y \tau) \otimes \lambda\big]$	$\{(x,y) \mapsto [\tau(\lambda\ell_x \circ h)](x,y), h \in \mathcal{H}\}$	$\mathbb{E}_{\mathsf{u} \sim (h \# \tau)(\tilde{x}, \tilde{y})}[\lambda\ell\,(\mathsf{u}, \tilde{x})]$
$(\tau\colon Y \rightsquigarrow X) \otimes$ $(\lambda\colon X \rightsquigarrow Y)$	$\pi_Y \circ \big[\tau \otimes (E \circ \lambda)\big]$ $\pi_X \circ \big[(F \circ \tau) \otimes \lambda\big]$	$\{(x,y) \mapsto [\tau(\lambda\ell_x \circ h)](y), h \in \mathcal{H}\}$	$\mathbb{E}_{\mathsf{u} \sim (h \# \tau)(\tilde{y})}[\lambda\ell\,(\mathsf{u}, \tilde{x})]$
$(\tau\colon X \times Y \rightsquigarrow X) \otimes$ $(\lambda\colon X \times Y \rightsquigarrow Y)$	$P \circ (\tau \otimes \lambda)$	$\{(x,y) \mapsto [\tau(\lambda\ell_{(x,y)} \circ h)](x,y), h \in \mathcal{H}\}$	$\mathbb{E}_{\mathsf{u} \sim (h \# \tau)(\tilde{x}, \tilde{y})}[\lambda\ell\,(\mathsf{u}, \tilde{x}, \tilde{y})]$

Appendix C. Related Existing Paradigms

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 Markov corruptions, thereby unveiling their relationships within the corruption hierarchy presented in Fig. 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. 2, while additional technical intricacies regarding correspondences and relationships are elucidated in subsequent sections.

C.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 X or the label space Y. We discuss in the following various examples of simple corruptions, i.e., in the sets $\mathcal{M}(X, X)$ and $\mathcal{M}(Y, Y)$, as defined in Fig. 1.

C.1.1 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. $(\mathsf{X}+\mathsf{N}, \mathsf{Y})$ with some independent noise random variable $\mathsf{N} \perp\!\!\!\perp \mathsf{X}$; in the second case, one's observation of X contains missing values.

Let $\mathbf{X} = (x_{ij})_{1\leq i\leq n, 1\leq j\leq d}$ be the complete input matrix, with $|X| = n$, and $\mathbf{M} = (m_{ij})_{1\leq i\leq n, 1\leq j\leq d}$ 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):[18]

- Missing completely at random (MCAR): the cause of missingness is entirely random, i.e., $p(\mathsf{M} \,|\, \mathsf{X}) = p(\mathsf{M})$ does not depend on X_o or X_m. This corresponds to having a trivial Markov kernel acting on the clean distribution, $\tau \colon \{*\} \rightsquigarrow X \equiv \mu \in \mathcal{P}(X)$.

18. Assume the rows x_i, m_i are assigned a joint distribution. and X and M are treated as random variables.

- Missing not at random (MNAR): the cause of missingness depends on both observed variables and missing variables, i.e., $p(\mathsf{M} \mid \mathsf{X}) = p(\mathsf{M} \mid \mathsf{X}_o, \mathsf{X}_m)$. This case corresponds to our non-trivial $\tau \colon X \rightsquigarrow X$.

- Missing at random (MAR): the cause of missingness depends on observed variables but not on missing variables, i.e., $p(\mathsf{M} \mid \mathsf{X}) = p(\mathsf{M} \mid \mathsf{X}_o)$. This case is a sub-case of the non-trivial $\tau \colon X \rightsquigarrow X$, which is not directly specifiable by our taxonomy because of the different premises it is built on.

C.1.2 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, $\mathcal{M}(Y, Y)$, that can be formulated as a corrupted posterior. For classification tasks, Y is assumed to be a finite space. Therefore the corruption $\lambda \colon Y \rightsquigarrow Y$ can be represented by a column-stochastic matrix $\mathbf{T} = (\rho_{ij})_{1 \leq i \leq |Y|, 1 \leq j \leq |Y|}$ which specifies the probability of the clean label $\mathsf{Y} = j$ being flipped to the corrupted label $\tilde{\mathsf{Y}} = i$, i.e., $\forall i, j,\ \rho_{ij} = p(\tilde{\mathsf{Y}} = i \mid \mathsf{Y} = j)$. The corrupted joint distribution can be rewritten as $\tilde{P} = \sum_Y p(\tilde{\mathsf{Y}} \mid \mathsf{Y})\, p(\mathsf{Y} \mid \mathsf{X})\, p(\mathsf{X})$. 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 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 $\lambda \colon Y \rightsquigarrow Y$ remains constant w.r.t. its parameters. Recently, some 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.

C.2 Dependent Corruptions

Although simple corruptions have been well studied and understood, more complexities arise in dependent cases, yet they receive relatively less attention and understanding. We discuss in the following examples of the dependent corruptions in the sets $\mathcal{M}(Y, X)$, $\mathcal{M}(X, Y)$, $\mathcal{M}(X \times Y, X)$ and $\mathcal{M}(X \times Y, Y)$, as defined in Fig. 1a.

C.2.1 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 $\tilde{\mathsf{X}}$ respectively, style transfer serves as an illustrative example of $\tau \colon Y \rightsquigarrow X$

"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 also 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.

C.2.2 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 models' 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 $\tau \in \mathcal{M}(X \times Y, X)$ corruption that can be formulated as a corrupted experiment.

C.2.3 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 $\lambda \in \mathcal{M}(X, Y)$ corruption that can be formulated as a corrupted experiment.

C.2.4 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 $\lambda \in \mathcal{M}(X \times Y, Y)$ corruption that can be formulated as a corrupted posterior. Compared to the matrix representation $\mathbf{T}$ of the ccn corruption $\kappa_{Y\tilde{Y}}$, the iln corruption $\kappa_{XY\tilde{Y}}$ 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 $\mathsf{X} = x$ with the clean label $\mathsf{Y} = j$ being flipped to the corrupted label $\tilde{\mathsf{Y}} = i$, i.e., $\forall i, j,\ \rho_{ij}(x) = p(\tilde{\mathsf{Y}} = i \,|\, \mathsf{Y} = j, \mathsf{X} = x)$. 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).

C.3 Combined Corruptions

Given the simple and dependent corruptions, we can combine them to generate 2-parameter joint corruptions, i.e., $\mathcal{M}(X \times Y, X \times Y)$. Below, we discuss some examples of combined noise models illustrated in Fig. 1b.

C.3.1 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 $\tau\colon X \rightsquigarrow X \otimes \lambda\colon Y \rightsquigarrow Y$ corruption that can be formulated as a corrupted experiment.

C.3.2 Target Shift

In the literature, target shift, also known as prior probability shift, refers to the situation where the prior probability $p(\mathsf{Y})$ is changed while the conditional distribution $p(\mathsf{X} \mid \mathsf{Y})$ 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 $p(\mathsf{Y})$ may cause changes in both $p(\mathsf{X})$ and $p(\mathsf{Y} \mid \mathsf{X})$ 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 $\tau\colon Y \rightsquigarrow X \otimes \lambda\colon X \times Y \rightsquigarrow Y$ 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 $\tilde{P} = (\pi_X \times F) \circ (\tau \otimes \lambda)$.

C.3.3 Covariate Shift

In the literature, covariate shift refers to the situation where the marginal distribution $p(\mathsf{X})$ is changed while the class-posterior probability $p(\mathsf{Y} \mid \mathsf{X})$ 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 $p(\mathsf{X})$ may lead to changes in $p(\mathsf{Y})$ and $p(\mathsf{X} \mid \mathsf{Y})$ due to the Bayes rule. Covariate shift is a subcase of $\tau\colon X \times Y \rightsquigarrow X \otimes \lambda\colon X \rightsquigarrow Y$ 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 Definition 19, 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 ($\mathsf{S} = 1$) or excluded from it ($\mathsf{S} = 0$). The corrupted distribution by selection bias can be expressed as $\tilde{P}(\mathsf{X}, \mathsf{Y}) = P(\mathsf{X}, \mathsf{Y} \,|\, \mathsf{S} = 1)$. By assuming the missing at random structure, where S is independent of Y given X: $P(\mathsf{S} \,|\, \mathsf{X}, \mathsf{Y}) = P(\mathsf{S} \,|\, \mathsf{X})$, we recover covariate shift where $P(\mathsf{X} \,|\, \mathsf{S} = 1) \neq P(\mathsf{X})$ and $P(\mathsf{Y} \,|\, \mathsf{X}, \mathsf{S}) = P(\mathsf{Y} \,|\, \mathsf{X})$.

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 Definition 19, requiring for selection bias the support condition $\tilde{P} \ll P$.

More generally, 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 $p(\mathsf{X})$ and the invariance of the class-posterior probability $p(\mathsf{Y} \,|\, \mathsf{X})$. As a result, there exist covariate shift scenarios that cannot be attributed to selection bias when $\tilde{P} \ll P$ is not the case.

C.3.4 Generalized Target Shift

In the literature, generalized target shift refers to the situation where the prior probability $p(\mathsf{Y})$ and the conditional distribution $p(\mathsf{X} \,|\, \mathsf{Y})$ 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 $\tau \colon X \times Y \rightsquigarrow X \otimes \lambda \colon X \times Y \rightsquigarrow Y$ 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 $\tau \colon X \rightsquigarrow X$ corruption, as it would violate the requirement of corrupting the label distribution.

C.3.5 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 $p_{t_0}(\mathsf{X}, \mathsf{Y}) \neq p_{t_1}(\mathsf{X}, \mathsf{Y})$ (Widmer and Kubat, 1996; Gama et al., 2014; Lu et al., 2019). In our words, $p_{t_1}(\mathsf{X}, \mathsf{Y})$ can be seen as a corrupted version of $p_{t_0}(\mathsf{X}, \mathsf{Y})$. Given its generality, this case can be associated with every corruption in our framework; therefore, the most general correspondence is the $\tau \colon X \times Y \rightsquigarrow X \otimes \lambda \colon X \times Y \rightsquigarrow Y$ 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, $p(\mathsf{Y} \,|\, \mathsf{X})$ changes over time, and such changes can occur with or without changes on $p(\mathsf{X})$, often referred to as concept shift; in our framework, this is a subcase of $\tau \colon X \times Y \rightsquigarrow X \otimes \lambda \colon X \times Y \rightsquigarrow Y$ corruption. More details in Tab. 5.

- Sampling shift (Tsymbal, 2004; Widmer and Kubat, 1993; Salganicoff, 1997): here, $p(\mathsf{X})$ changes over time while $p(\mathsf{Y} \mid \mathsf{X})$ remains invariant, also known as virtual drift; in our framework, this is a subcase of $\tau\colon X \times Y \rightsquigarrow X \otimes \lambda\colon X \rightsquigarrow Y$ corruption. More details are provided in Tab. 5.

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 $p(\mathsf{Y} \mid \mathsf{X})$ changing while $p(\mathsf{X})$ remains invariant or $p(\mathsf{X} \mid \mathsf{Y})$ changing while $p(\mathsf{Y})$ remains invariant. Similar to instance- and label-dependent noise and covariate shift, they are examples of $\lambda\colon X \times Y \rightsquigarrow Y$ corruption and $\tau\colon X \times Y \rightsquigarrow X \otimes \lambda\colon Y \rightsquigarrow Y$ corruption that involve more corrupted spaces at different time points.

Appendix D. 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 they are too far from our point of view and objective.

D.1 Traditional and Causal Taxonomies

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. 5. 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.

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)

Table 5: Traditional taxonomies resume.

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[19] when $\mathsf{Y} \to \mathsf{X}$		subcase of $\kappa\colon X \rightsquigarrow X$
at least one among $\{\pi_Y, F, E\}$, according to compatibility	π_X	concept shift[20] when $\mathsf{X} \to \mathsf{Y}$		subcase of $\lambda\colon Y \rightsquigarrow Y$
at least π_Y, causing π_X or F to change	E	prior probability shift[21] when $\mathsf{Y} \to \mathsf{X}$		at least a $\lambda\colon Y \rightsquigarrow Y$ subcase, at most $\kappa\colon X \times Y \rightsquigarrow X$ $\otimes$ $\lambda\colon X \times Y \rightsquigarrow Y$
at least π_X, causing π_Y or E to change	F	covariate shift[22] when $\mathsf{X} \to \mathsf{Y}$		at least a $\kappa\colon X \rightsquigarrow X$ subcase, at most $\kappa\colon X \times Y \rightsquigarrow X$ $\otimes$ $\lambda\colon X \times Y \rightsquigarrow Y$

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 *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. 5. In the case described by Definition 10 we restrict it to take values in $\{0, 1\}$, the clean and corrupted environments. This representation is again missing some 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. 1. An additional limitation of the causal approach lies in the causal assumption itself; we are forced, in this setting, to only consider one conditional probability between E and F to be a valid representation of the generative process, while in our framework we are not inherently forced to make this choice. We although can still make use of causal information in case it is available, as well as have more favorable causal relationship between X and Y depending

on which corruption type we want to analyze. This is apparent in the Data Processing Equalities we prove.

In both the described classes of taxonomies, it not 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 corruption model can facilitate their chaining. Nevertheless, feasibility rules are rather complex and unclear to understand, given the overlapping nature of the corruptions and identifiability problems for causal representations (Pearl, 2009).

Appendix E. 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 interested readers to refer to (Mac Lane, 2013).

E.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 evaluation function $ev_B : \mathcal{P}(X) \to [0,1] = P \mapsto P(B)$ measurable, for B a measurable subset in X;
- the multiplication of the monad, $\mu : \mathcal{P}^2 \Rightarrow \mathcal{P}$, is defined by
$$\mu_X(Q)(B) = \int_{q \in \mathcal{P}(X)} ev_B(q) dQ\,;$$
- the unit of the monad, $\delta : Id \Rightarrow \mathcal{P}$, sends a point $x \in X$ to the Dirac measure at x.

Table 6: Comparison of categorical concepts in Dahlqvist et al. (2016) and probabilistic concepts in this paper.

Categorical	Probabilistic
Kleisli category of Giry monad $\mathtt{G}$, $K\ell$	measurable spaces as objects and Markov kernels as arrows
arrows in category $1 \downarrow \mathcal{K}\ell$	$\mathcal{M}(X, Y)$ where X and Y have marginals p and q, respectively
arrows in category $1 \downarrow F$	subset of the above $\mathcal{M}(X, Y)$ with measure-preserving maps induced by identical kernels δ
Kleisli composition $\circ_{\mathtt{G}}$	chain composition $\circ$ in P1 with transitional kernels
$\alpha_Y^X : \mathrm{Hom}_{\mathtt{Krn}}(X, -) \to \Gamma(X, -)$	product composition $\times$ in P2 with a kernel and a probability

This equips the endofunctor $\mathcal{P} : \mathtt{Mes} \to \mathtt{Mes}$ into a monad, that is, the Giry monad $\mathtt{G} \coloneqq (\mathcal{P}, \mu, \delta)$ on measurable spaces.

The Kleisli category of $\mathtt{G}$, denoted by $\mathcal{K}\ell$, has the same objects as $\mathtt{Mes}$, and the morphism $\kappa \colon X \rightsquigarrow Y$ in $\mathcal{K}\ell$ is a kernel $\kappa \colon X \to \mathcal{P}(Y)$ in $\mathtt{Mes}$. The Kleisli composition of kernel $\kappa \colon X \rightsquigarrow Y$ with $\lambda \colon Y \rightsquigarrow Z$ 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 $\mathcal{P}(\kappa)(\cdot) \coloneqq (\cdot) \circ \kappa^{-1}$, defined on a suitable space of probabilities. Hence, $\circ_{\mathtt{G}}$ is the same as the chain composition we defined in P1.

E.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 P4 in a more formal way.

The category of Markov kernel considered here is the one of ***typed kernel***. Their definition is tied to a fixed probability p on X and a fixed probability q on Y, so that $\kappa \circ_{\mathtt{G}} p = q$, instead of being characterized for every probability on 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 couple on the clean and corrupted distribution. This remark is also crucial for understanding our notion of exhaustiveness in § E.3.

The key object for building the inversion operation is the $\mathtt{Krn}$ category, similar to our notion of space of Markov kernels $\mathcal{M}(X, Y)$, but with an equivalence relation acting on it. We describe its construction in the following steps.

1. Let $F \colon \mathtt{Mes} \to \mathcal{K}\ell$ be the functor embedding $\mathtt{Mes}$ into $\mathcal{K}\ell$ which acts identically on spaces and maps measurable arrows $\kappa \colon X \to Y$ to Kleisli arrows $F(\kappa) = \delta_Y \circ \kappa$. This

means that $F(\kappa)$ only allows one possible jump at each x in X, with δ_Y an identical jump (i.e., a deterministic kernel).

2. It further induces the category $1 \downarrow F$ of probabilities $p : 1 \rightsquigarrow \mathcal{P}(X)$, denoted by (X, p), and morphisms $\kappa\colon (X, p) \rightsquigarrow_\delta (Y, q)$ as degenerate arrows $F(\kappa) : X \rightsquigarrow Y$ s.t. $q = F(\kappa) \circ_{\mathtt{G}} p = \mathtt{P}(\kappa)(p) = p \circ \kappa^{-1}$. In more familiar terms, this is saying that q is the push-forward of p 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 (X, p) to (Y, q) is any Kleisli arrow $\kappa\colon X \rightsquigarrow Y$ s.t. $q = \kappa \circ_{\mathtt{G}} p$, and the arrows are what we denoted as $\mathcal{M}(X, Y)$, where X has marginal probability p and Y has marginal probability q.

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 $\kappa\colon (X, p) \rightsquigarrow (Y, q)$ there are p-negligibly many points jumping to q-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 37 *For all objects $(X, p), (Y, q)$, $R_{(X,p),(Y,q)}$ is the smallest equivalence relation on $Hom_{1\downarrow\mathcal{K}\ell}(X, Y)$ such that*

$$(\kappa, \kappa') \in R_{(X,p),(Y,q)} \quad \Leftrightarrow \quad \kappa = \kappa' \ \ p\text{-a.s.}$$

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 38 *The category* $\mathtt{Krn}$ *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 $\mathrm{Hom}_{\mathtt{Krn}}((X, p), (Y, q))$ and $\mathrm{Hom}_{\mathtt{Krn}}((Y, q), (X, p))$. There are two mappings between the $\mathtt{Krn}$ category and the space of couplings associated to $(X, p), (Y, q)$. The first is equivalent to the product composition we defined in P2 applied to a kernel (i.e., conditional probability) and a probability, and is formally written as

$$\alpha_Y^X : \mathrm{Hom}_{\mathtt{Krn}}((X, p), (Y, q)) \to \Gamma((X, p), (Y, q)) \ \text{ s.t. } \ \alpha_Y^X(\kappa)(B_X \times B_Y) \coloneqq \int_{x \in B_X} \kappa(x)(B_Y) dp\,,$$

with $\Gamma((X, p), (Y, q)) \subset \mathcal{P}(X \times Y)$ the typed couplings associated to the marginals $(X, p), (Y, q)$. 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.,

$$\begin{aligned} &D_Y^X : \Gamma((X, p), (Y, q)) \to \mathrm{Hom}_{\mathtt{Krn}}((X, p), (Y, q)) \\ &\text{s.t.} \quad D_Y^X(\gamma) \coloneqq \mathtt{P}(\pi_Y) \circ \pi_X^\dagger, \ \gamma \in \Gamma((X, p), (Y, q)), \\ &\text{and} \quad \gamma(B_X \times B_Y) \coloneqq \int_{x \in B_X} D_Y^X(\gamma)(x)(B_Y) dp\,, \end{aligned}$$

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 39 *The Bayesian inverse of a typed kernel κ from (X,p) to (Y,q), is defined as*

$$(\cdot)^\dagger : \kappa \mapsto \kappa^\dagger \coloneqq \left(D_X^Y \circ \mathtt{P}(\pi_Y \times \pi_X) \circ \alpha_Y^X\right)(\kappa)\,,$$

with $\mathtt{P}(\pi_Y \times \pi_X) : \Gamma((X,p),(Y,q)) \to \Gamma((Y,q),(X,p))$ being the permutation map.

As the Bayesian inverse has been defined as a bijection between $\mathrm{Hom}_{\mathtt{Krn}}((X,p),(Y,q))$ and $\mathrm{Hom}_{\mathtt{Krn}}((Y,q),(X,p))$, it is always guaranteed to exist in this setting.

Proposition 40 (Bayesian Inversion Theorem) *The Bayesian inverse of a typed kernel κ from (X,p) to (Y,q) exists and is equivalently one of the following objects:*

1. *$\kappa^\dagger : (Y,q) \to (X,p) \in \mathtt{Krn}$ when κ is seen as element of $\mathtt{Krn}$, such that $(\kappa^\dagger \circ_{\mathtt{G}} \kappa) \circ_{\mathtt{G}} q = \delta_Y \circ_{\mathtt{G}} q$ and $(\kappa \circ \kappa^\dagger) \circ_{\mathtt{G}} p = \delta_X \circ_{\mathtt{G}} p$;*
2. *$\kappa^\dagger : Y \rightsquigarrow X \in \mathcal{M}(Y,X)$ when κ is seen as element of $\mathcal{M}(X,Y)$, such that $(\kappa^\dagger \circ_{\mathtt{G}} \kappa) \circ_{\mathtt{G}} q \equiv_R \delta_Y \circ_{\mathtt{G}} q$ and $(\kappa \circ_{\mathtt{G}} \kappa^\dagger) \circ_{\mathtt{G}} p \equiv_R \delta_X \circ_{\mathtt{G}} p$.*

Here, $\delta_{(\cdot)}$ indicates the identical 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 Definition 37. ■

Remark 41 *We can understand the Bayesian inverse of a corruption kernel $\kappa \in \mathcal{M}(Z,Z)$ from $(Z,\mathcal{Z},P)$ to $(Z,\mathcal{Z},\tilde{P})$ that distorts $\tilde{P}(A) = \int_A \int_Z \kappa(z,d\tilde{z})\,P(dz)\ \ \forall A \in \mathcal{Z}$ as a Markov kernel $\kappa^\dagger \in \mathcal{M}(Z,Z)$ satisfying*

$$\int_B \kappa(z,A)\,P(dz) = \int_A \kappa^\dagger(\tilde{z},B)\,\tilde{P}(d\tilde{z}) \quad \forall A \in \mathcal{Z},\, B \in \mathcal{Z}.$$

This formulation extends the discrete Bayes' rule $P(\tilde{z}\,|\,z)P(z) = P(z\,|\,\tilde{z})P(\tilde{z})\,\forall z,\tilde{z} \in Z$. Hence, in the discrete case, the Bayesian inverse always exists and can be expressed as

$$\kappa^\dagger(z\,|\,\tilde{z}) \coloneqq \frac{P(z)\kappa(\tilde{z}\,|\,z)}{\tilde{P}(\tilde{z})} \text{ for } z,\tilde{z} \in Z \text{ with } \tilde{P}(\tilde{z}) \neq 0.$$

This formula ensures the uniqueness of $\kappa^\dagger$ within the support of $\tilde{P}$, as all components are unique. However, outside the support when $\tilde{P}$ is zero, the uniqueness may not hold, requiring a non-fixed value for $\tilde{z} \in Z$ where $\tilde{P}(\tilde{z}) = 0$.

In the continuous case, the Bayesian inverse may not exist. To ensure $\kappa^\dagger \in \mathcal{M}(Z,Z)$ is well-defined, it must satisfy the conditions of being a Markov kernel, as defined in Definition 1, where the mapping $\tilde{z} \to \kappa^\dagger(\tilde{z},B)$ is $\mathcal{Z}$-measurable for every set $B \in \mathcal{Z}$, and the mapping $B \to \kappa^\dagger(\tilde{z},B)$ is a probability measure on $(Z,\mathcal{Z})$ for every $\tilde{z} \in Z$, for the standard Borel space $(Z,\mathcal{Z})$. *Under this condition, the Bayesian inverse always exists, and it is uniquely defined within the support of $\tilde{P}$, where uniqueness is represented by an equivalence class of kernels that are $\tilde{P}$-a.s. equal.*

E.3 Exhaustiveness of Markovian Corruption

As we noticed in § C, 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 α_Y^X and D_Y^X for typed kernels, where one is the inverse of the other by construction (Appendix E.2). They are the operations representing the bijection between the space of Markov kernels typed for p, q and the space of couplings with marginals p, q. 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 F. Proofs for Data Processing Equality Results

Recall that π_Y is a prior distribution on Y, and the notation κ_{X} stands for the kernel κ evaluated on the parameter X, e.g., $E_{\mathsf{Y}}, F_{\mathsf{X}}$, and $\kappa_x \coloneqq \kappa_{\mathsf{X}=x}$. The kernel δ_Z denotes a kernel induced by the Dirac delta measure from $(Z, \mathcal{Z})$ to $(Z, \mathcal{Z})$.

In the proofs we will use a continuous notation for measures on Y, for the sake of simplicity and homogeneity. However, notice that all the λ kernels are actually (parameterized) stochastic matrices $\Lambda = [\Lambda_{\tilde{y}y}]$, where $\Lambda_{\tilde{y}y} = p(\tilde{\mathsf{Y}} = \tilde{y} \,|\, \mathsf{Y} = y)$ for simple corruptions and $\Lambda_{\tilde{y}y}(x) = p(\tilde{\mathsf{Y}} = \tilde{y} \,|\, \mathsf{Y} = y, \mathsf{X} = x)$ for dependent corruptions. Note that both y and $\tilde{y}$ range in Y, and thus they are squared matrices. In Theorem 24 and Lemma 44, the kernel λ acting on the function $\ell \circ \mathcal{H}$ is actually the transpose of the stochastic matrix Λ:

$$\sum_{\tilde{y} \in Y} \lambda_x(y, d\tilde{y})\, \ell(h(x), \tilde{y}) = \sum_{\tilde{y} \in Y} \Lambda_{\tilde{y}y}^{\top}(x)\, \ell_{\tilde{y}}(h(x)) = \widetilde{(\ell_y \circ h)}(x)\,.$$

Below, Theorems 23 and 24 are proved based on the Lemmas concerning BR changes under dependent X and Y corruptions, respectively.

Lemma 42 (X corruption) *Consider the learning problem $(\ell, \mathcal{H}, P)$, with ℓ being a bounded loss, and $E\colon Y \rightsquigarrow X$ its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y. Let $\tau \otimes \delta_Y$ be a corruption acting on this problem, with $\tau \in \mathcal{M}(X \times Y, X)$. Then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\pi_Y \times E) \circ (\tau \otimes \delta_Y)\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big((E \circ_X \tau) \otimes \delta_Y\big)\Big) \equiv_{\mathrm{BR}} \Big(\tau(\ell \circ \mathcal{H}), \pi_Y \times E\Big)\,.$$

Moreover, if $\tau \in \mathcal{M}(X, X)$, we have

$$\Big(\ell \circ \mathcal{H}, (\pi_Y \times E) \circ (\tau \otimes \delta_Y)\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big((E \circ \tau) \otimes \delta_Y\big)\Big) \equiv_{\mathrm{BR}} \Big(\tau(\ell \circ \mathcal{H}), \pi_Y \times E\Big)\,.$$

Proof Let $A \in \mathcal{X} \times \mathcal{Y}$, and π_y be the y-th entry of the π_Y probability vector. By definition of all the objects involved, the action of $\tau \otimes \delta_Y$ on P is

$$\begin{aligned}
\tilde{P}(A) &= \int_{(\tilde{x},\tilde{y})\in A} \int_{(x,y)\in X\times Y} \tau_y(x, d\tilde{x})\, \delta_y(d\tilde{y})\, P(dx, dy) \\
&= \int_{(\tilde{x},\mathtt{y})\in \mathtt{A}} \left[\sum_{y\in Y} \left(\int_{x\in X} \tau_y(x, d\tilde{x})\, E_y(dx) \right) \delta_y(d\tilde{y})\, \pi_y \right] \\
&= \int_{(\tilde{x},\tilde{y})\in A} \left[\sum_{y\in Y} (E \circ_X \tau)_y(d\tilde{x})\, \delta_y(d\tilde{y})\, \pi_y \right] \\
&= \left[\pi_y \circ \left((E \circ_X \tau) \otimes \delta_Y \right) \right](A)\,.
\end{aligned}$$

We can hence rewrite the risk w.r.t. $\tilde{P} \coloneqq \pi_y \circ [(E \circ_X \tau) \otimes \delta_Y]$ as

$$\begin{aligned}
\mathbb{E}_{(\tilde{\mathsf{X}},\tilde{\mathsf{Y}})\sim\tilde{P}} \left[\ell(h_{\tilde{\mathsf{X}}}, \tilde{\mathsf{Y}})) \right] &= \sum_{\substack{y\in Y,\\ \tilde{y}\in Y}} \left[\int_{\tilde{x}\in X} \ell(h_{\tilde{x}}, \tilde{y}) \left(\int_{x\in X} \tau_y(x, d\tilde{x})\, E_y(dx) \right) \right] \delta_y(d\tilde{y})\, \pi_y \\
&= \sum_{y\in Y} \left[\int_{x\in X} \left(\int_{\tilde{x}\in X} (\delta\ell_y \circ h)(\tilde{x})\, \tau_y(x, d\tilde{x}) \right) E_y(dx) \right] \pi_y \\
&= \sum_{y\in Y} \left[\int_{x\in X} [\tau(\delta\ell_y \circ h)](x, y)\, E_y(dx) \right] \pi_y \qquad (24) \\
&= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_Y\times E)} \left[\left(\tau(\delta\ell_{\mathsf{Y}} \circ h) \right)(\mathsf{X}, \mathsf{Y}) \right]\,.
\end{aligned}$$

Let $\tilde{E}_y(d\tilde{x}) \coloneqq (E \circ_X \tau)_y(d\tilde{x})$. We have that the associated BR is

$$\begin{aligned}
\inf_{h\in\mathcal{H}} \mathbb{E}_{(\tilde{\mathsf{X}},\tilde{\mathsf{Y}})\sim\pi_Y\circ[(E\circ_X\tau)\otimes\delta_Y]} \left[\ell(h_{\tilde{\mathsf{X}}}, \tilde{\mathsf{Y}})) \right] &= \inf_{h\in\mathcal{H}} \mathbb{E}_{\tilde{\mathsf{Y}}\sim\pi_Y} \mathbb{E}_{\tilde{\mathsf{X}}\sim\tilde{E}_{\tilde{\mathsf{Y}}}} \left[\ell(h_{\tilde{\mathsf{X}}}, \tilde{\mathsf{Y}})) \right] \\
&= \mathrm{BR}_{\ell\circ\mathcal{H}}[\pi_Y \circ (\tilde{E} \otimes \delta_Y)]\,,
\end{aligned}$$

$$\begin{aligned}
\inf_{h\in\mathcal{H}} \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_Y\times E)} \left[\left(\tau(\delta\ell_{\mathsf{Y}} \circ h) \right)(\mathsf{X}, \mathsf{Y}) \right] &= \inf_{f\in\tau(\ell\circ\mathcal{H})} \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_Y\times E)} \left[f(\mathsf{X}, \mathsf{Y}) \right] \\
&= \mathrm{BR}_{\tau(\ell\circ\mathcal{H})}[\pi_Y \times E]\,, \qquad (25)
\end{aligned}$$

which are equal given the previous computations. We have defined and used in Eq. (25) that $f(x, y) \coloneqq [\tau(\delta\ell_y \circ h)](x, y), h \in \mathcal{H}$. Such functions are the ones populating the minimization set $\tau(\ell \circ \mathcal{H})$, 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. (24) lead to a slightly different model class:

$$\begin{aligned}
\mathrm{BR}_{\ell\circ\mathcal{H}} \left[\pi_Y \circ \left((E \circ \tau) \otimes \delta_Y \right) \right] &= \inf_{h\in\mathcal{H}} \sum_{y\in Y} \left[\int_{x\in X} [\tau(\delta\ell_y \circ h)](x)\, E_y(dx) \right] \pi_y \\
&= \inf_{f\in\tau(\ell\circ\mathcal{H})} \sum_{y\in Y} \left[\int_{x\in X} f(x, y)\, E_y(dx) \right] \pi_y = \mathrm{BR}_{\tau(\ell\circ\mathcal{H})}(\pi_Y \times E)\,.
\end{aligned}$$

■

Theorem 43 (2-dependent τ, simple λ, Theorem 23) *Consider the learning problem $(\ell, \mathcal{H}, P)$, with ℓ being a bounded loss, and $E\colon Y \rightsquigarrow X$ its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y. Let $(\tau\colon X \times Y \rightsquigarrow X) \otimes (\lambda\colon Y \rightsquigarrow Y)$ be a corruption acting on this problem, then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\pi_Y \times E) \circ (\tau \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big((E \circ_X \tau) \otimes \lambda\big)\Big) \equiv_{\mathrm{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), \pi_Y \times E\Big)\,.$$

The functions contained in the new minimization set are defined as

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

Proof With this corruption formulation, we can replicate the proof of Lemma 42 up to Eq. (24) by simply plugging in λ instead of δ_Y. Therefore, we obtain the thesis. ■

We remark that in this case $\tilde{P} \neq \pi_Y \times \tilde{E}$ with $\tilde{E}_y(d\tilde{x}) \coloneqq (E \circ_X \tau)_y(d\tilde{x})$, 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 $\tilde{P} = \pi_y \circ [(E \circ_X \tau) \otimes \lambda]$. We conclude that, in this more general case, it does not make sense to distinguish the effect of corruption on E and π.

Lemma 44 (Y corruption) *Consider the learning problem $(\ell, \mathcal{H}, P)$, with ℓ being a bounded loss, and $F\colon X \rightsquigarrow Y$ its associated posterior such that $P = \pi_X \times F$ for a suitable π_X. Let $\delta_X \otimes \lambda$ be a corruption acting on this problem, with $\lambda \in \mathcal{M}(X \times Y, Y)$. Then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\pi_X \times F) \circ (\delta_X \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_X \circ \big(\delta_X \otimes (F \circ_Y \lambda)\big)\Big) \equiv_{\mathrm{BR}} \Big(\lambda\ell \circ \mathcal{H}, \pi_X \times F\Big)\,.$$

Moreover, if $\lambda \in \mathcal{M}(Y, Y)$, we have

$$\Big(\ell \circ \mathcal{H}, (\pi_X \times F) \circ (\delta_X \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_X \circ \big(\delta_X \otimes (F \circ \lambda)\big)\Big) \equiv_{\mathrm{BR}} \Big(\lambda\ell \circ \mathcal{H}, \pi_X \times F\Big)\,.$$

Proof Let $A \in \mathcal{X} \times \mathcal{Y}$. By definition of all the objects involved, the action of $\tau \otimes \delta_Y$ on P is

$$\begin{aligned}
\tilde{P}(A) &= \int_{(\tilde{x},\tilde{y})\in A} \int_{(x,y)\in X\times Y} \delta_x(d\tilde{x})\, \lambda_x(y, d\tilde{y})\, P(dx, dy) \\
&= \int_{(\tilde{x},\tilde{y})\in A} \left[\int_{x\in X} \left(\sum_{y\in Y} \lambda_x(y, d\tilde{y})\, F_x(dy)\right) \delta_x(d\tilde{x})\, \pi_X(dx)\right] \\
&= \int_{(\tilde{x},\tilde{y})\in A} \left[\int_{x\in X} (F \circ_Y \lambda)_x(d\tilde{y})\, \delta_x(d\tilde{x})\, \pi_X(dx)\right] \\
&= \left[\pi_X \circ \big((F \circ_Y \lambda) \otimes \delta_X\big)\right](A)\,.
\end{aligned}$$

We can hence rewrite the risk w.r.t. $\tilde{P} \coloneqq \pi_X \circ [(F \circ_Y \lambda) \otimes \delta_X]$ as

$$\begin{aligned}
\mathbb{E}_{(\tilde{\mathsf{X}},\tilde{\mathsf{Y}})\sim\tilde{P}}\left[\ell(h_{\tilde{\mathsf{X}}},\tilde{\mathsf{Y}}))\right] &= \int_{\substack{x\in X,\\ \tilde{x}\in X}} \left[\sum_{\tilde{y}\in Y} \ell(h_{\tilde{x}},\tilde{y}) \left(\sum_{y\in Y} \lambda(x,y,d\tilde{y})\, F_x(dy)\right)\right] \delta_x(d\tilde{x})\, \pi_X(dx) \\
&= \int_{\substack{x\in X,\\ \tilde{x}\in X}} \left[\sum_{y\in Y} \left(\sum_{\tilde{y}\in Y} \ell(h_{\tilde{x}},\tilde{y})\, \lambda(x,y,d\tilde{y})\right) F_x(dy)\right] \delta_x(d\tilde{x})\, \pi_X(dx) \\
&= \int_{x\in X} \left[\sum_{y\in Y} \Big(\int_{\tilde{x}\in X} (\lambda\ell)(h_{\tilde{x}},x,y)\delta_x(d\tilde{x})\Big) F_x(dy)\right] \pi_X(dx) \\
&= \int_{x\in X} \left[\sum_{y\in Y} \Big((\lambda\ell)(h_x,x,y)\Big) F_x(dy)\right] \pi_X(dx) \qquad (26) \\
&= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_Y\times E)}\left[(\lambda\ell)(h_{\mathsf{X}},\mathsf{X},\mathsf{Y})\right] \qquad (27) \\
&= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_Y\times E)}\left[(\lambda\ell_{(\mathsf{X},\mathsf{Y})}\circ h)(\mathsf{X})\right] .
\end{aligned}$$

Similarly to the proof provided for Lemma 42, we can switch to BR and obtain

$$\mathrm{BR}_{\ell\circ\mathcal{H}}[\pi_X \circ \big((F \circ_Y \lambda) \otimes \delta_X\big)] = \mathrm{BR}_{\lambda\ell\circ\mathcal{H}}(\pi_X \times F)\,,$$

with functions $\lambda\ell(h_x,x,y) = (\lambda\ell_{(x,y)}\circ h)(x) \in \lambda\ell\circ\mathcal{H}$. If λ is simple, then Eq. (26) leads to a simpler model class:

$$\begin{aligned}
\mathrm{BR}_{\ell\circ\mathcal{H}}[\pi_X \circ \big((F\circ\lambda)\otimes\delta_X\big)] &= \inf_{h\in\mathcal{H}} \int_{x\in X} \left[\sum_{y\in Y} (\lambda\ell)(h_x,y)F_x(dy)\right] \pi_X(dx) \\
&= \inf_{f\in\lambda(\ell\circ\mathcal{H})} \int_{x\in X} \left[\sum_{y\in Y} f(x,y)F_x(dy)\right] \pi_X(dx) \\
&= \mathrm{BR}_{\lambda(\ell\circ\mathcal{H})}(\pi_X\times F)\,.
\end{aligned}$$

■

Theorem 45 (simple τ, 2-dependent λ, Theorem 24) *Consider the learning problem $(\ell,\mathcal{H},P)$, with ℓ being a bounded loss, and $F\colon X \rightsquigarrow Y$ its associated posterior such that $P = \pi_X \times F$ for a suitable π_X. Let $(\tau\colon X \rightsquigarrow X) \otimes (\lambda\colon X\times Y \rightsquigarrow Y)$ be a corruption acting on this problem, then, we obtain*

$$\Big(\ell\circ\mathcal{H}, (\pi_X\times F)\circ(\tau\otimes\lambda)\Big) = \Big(\ell\circ\mathcal{H}, \pi_X\circ\big(\tau\otimes(F\circ_Y\lambda)\big)\Big) \equiv_{\mathrm{BR}} \Big(\tau(\lambda\ell\circ\mathcal{H}), \pi_X\times F\Big)\,.$$

The functions contained in the new minimization set are defined as

$$\tau(\lambda\ell\circ\mathcal{H}) \coloneqq \{(x,y)\mapsto[\tau(\lambda\ell_{(x,y)}\circ h)](x), h\in\mathcal{H}\}\,.$$

Proof With this corruption formulation, we can replicate the proof of Lemma 44 up to Eq. (26) by simply plugging in τ instead of δ_x. Therefore, we obtain the thesis. ■

Theorem 46 (1-dependent and a 2-dependent, Theorem 25) *Consider the clean learning problem $(\ell, \mathcal{H}, P)$, with ℓ being a bounded loss, $E\colon Y \rightsquigarrow X$ its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y, and $F\colon X \rightsquigarrow Y$ its associated posterior such that $P = \pi_X \times F$ for a suitable π_X.*

1. *Let $(\tau\colon Y \rightsquigarrow X) \otimes (\lambda\colon X \times Y \rightsquigarrow Y)$ be a corruption acting on the problem, then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\pi_Y \times E) \circ (\tau \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big(\tau \otimes (E \circ_X \lambda)\big)\Big) \equiv_{\mathrm{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), \pi_Y \times E\Big)\,.$$

   The functions contained in the new minimization set are defined as

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

2. *Let $(\tau\colon X \times Y \rightsquigarrow X) \otimes (\lambda\colon X \rightsquigarrow Y)$ be a corruption acting on the problem, then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\pi_X \times F) \circ (\tau \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_X \circ \big((F \circ_Y \tau) \otimes \lambda\big)\Big) \equiv_{\mathrm{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), \pi_X \times F\Big)\,.$$

   The functions contained in the new minimization set are defined as

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

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 P is

$$\begin{aligned}
\tilde{P}(A) &= \int_{(\tilde{x},\tilde{y}) \in A} \left[\sum_{y \in Y} \left(\int_{x \in X} \lambda_y(x, d\tilde{y})\; E_y(dx)\right) \tau(y, d\tilde{x})\, \pi_y\right] \\
&= \int_{(\tilde{x},\tilde{y}) \in A} \left[\sum_{y \in Y} (E \circ_X \lambda)(y, d\tilde{y})\, \tau(y, d\tilde{x})\, \pi_y\right] \\
&= \int_{(\tilde{x},\tilde{y}) \in A} \left[\sum_{y \in Y} \big(\tau \otimes (E \circ_X \lambda)\big)(y, d\tilde{x}, d\tilde{y})\, \pi_y\right] = \left[\pi_Y \times \big(\tau \otimes (E \circ_X \lambda)\big)\right](A)\,.
\end{aligned}$$

We can then write the associated risk w.r.t. $\tilde{P} \coloneqq \pi_Y \times \big(\tau \otimes (E \circ_X \lambda)\big)$ as

$$\begin{aligned}
\mathbb{E}_{(\tilde{\mathsf{X}},\tilde{\mathsf{Y}})\sim\tilde{P}}[\ell(h_{\tilde{\mathsf{X}}},\tilde{\mathsf{Y}})] &= \sum_{\substack{y\in Y,\\ \tilde{y}\in Y}} \left[\int_{\tilde{x}\in X} \ell(h_{\tilde{x}},\tilde{y}) \left(\int_{x\in X} \lambda(x,y,d\tilde{y})\, E_y(dx)\right)\right] \tau(y,d\tilde{x})\, \pi_y \\
&= \sum_{y\in Y} \left[\int_{x\in X} \left(\int_{\tilde{x}\in X} \left(\sum_{\tilde{y}\in Y} \lambda(x,y,d\tilde{y})\, \ell(h_{\tilde{x}},\tilde{y})\right) \tau(y,d\tilde{x})\right) E_y(dx)\right] \pi_y \\
&= \sum_{y\in Y} \left[\int_{x\in X} \left(\int_{\tilde{x}\in X} (\lambda\ell_{(x,y)} \circ h)(\tilde{x})\, \tau(y,d\tilde{x})\right) E_y(dx)\right] \pi_y \\
&= \sum_{y\in Y} \left[\int_{x\in X} [\tau(\lambda\ell_{(x,y)} \circ h)](y)\, E_y(dx)\right] \pi_y \\
&= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_Y\times E)} \left[\big(\tau(\lambda\ell_{(\mathsf{X},\mathsf{Y})} \circ h)\big)(\mathsf{Y})\right] ,
\end{aligned}$$

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 P as

$$\begin{aligned}
\tilde{P}(A) &= \int_{(\tilde{x},\tilde{y})\in A} \left[\int_{x\in X} \lambda(x,d\tilde{y}) \left(\sum_{y\in Y} \tau(x,y,d\tilde{x})\, F(x,dy)\right) \pi_X(dx)\right] \\
&= \int_{(\tilde{x},\tilde{y})\in A} \left[\int_{x\in X} \lambda(x,d\tilde{y})\, (F \circ_Y \tau)(x,d\tilde{x})\, \pi_X(dx)\right] = \left[\pi_X \times \big(\lambda \otimes (F \circ_Y \tau)\big)\right](A) ,
\end{aligned}$$

and repeat a similar argument but for the F 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 47 (1-dependent τ and λ, Corollary 26) *Consider the clean learning problem $(\ell,\mathcal{H},P)$, with ℓ being a bounded loss, $E\colon Y \rightsquigarrow X$ its associated experiment such that $P = \pi_Y \times E$ for a suitable π_Y, and $F\colon X \rightsquigarrow Y$ its associated posterior such that $P = \pi_X \times F$ for a suitable π_X. Let $(\tau\colon Y \rightsquigarrow X) \otimes (\lambda\colon X \rightsquigarrow Y)$ be a corruption acting on the problem, then, we obtain*

$$\Big(\ell \circ \mathcal{H}, (\pi_Y \times E) \circ (\tau \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_Y \circ \big(\tau \otimes (E \circ \lambda)\big)\Big) \equiv_{\text{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), \pi_Y \times E\Big) .$$

or, equivalently,

$$\Big(\ell \circ \mathcal{H}, (\pi_X \times F) \circ (\tau \otimes \lambda)\Big) = \Big(\ell \circ \mathcal{H}, \pi_X \circ \big((F \circ \tau) \otimes \lambda\big)\Big) \equiv_{\text{BR}} \Big(\tau(\lambda\ell \circ \mathcal{H}), \pi_X \times F\Big) .$$

The functions contained in the new minimization set are defined as

$$\tau(\lambda\ell \circ \mathcal{H}) \coloneqq \{(x,y) \mapsto [\tau(\lambda\ell_x \circ h)](y), h \in \mathcal{H}\} .$$

Proof We can replicate the proof of Theorem 25 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 $f(x,y) \coloneqq \tau[(\lambda\ell)_x \circ h](y)$, i.e., comparing a point x with a kernel on $\mathcal{P}(X)$ parameterized by y. Therefore, we obtain the thesis. ■

Theorem 48 (2-dependent κ and λ, Theorem 27) *Consider the clean learning problem $(\ell, \mathcal{H}, P)$, with ℓ being a bounded loss, and let $(\tau\colon X \times Y \rightsquigarrow X) \otimes (\lambda\colon X \times Y \rightsquigarrow Y)$ be a corruption acting on the problem. Then:*

1. *the action of such corruption on the joint probability P is equivalent to the one of the non-decomposed joint corruption;*
2. *the action on the minimization set $\ell \circ \mathcal{H}$ induces the following* BR*-equivalence*
$$\big(\ell, \mathcal{H}, P \circ (\tau \otimes \lambda)\big) \equiv_{\text{BR}} \big(\tau(\lambda\ell \circ \mathcal{H}), P\big)\,;$$
3. *the functions contained in the new minimization set are defined as*
$$\tau(\lambda\ell \circ \mathcal{H}) \coloneqq \{(x,y) \mapsto [\tau(\lambda\ell_{(x,y)} \circ h)](x,y), h \in \mathcal{H}\}\,.$$

Proof Let $A \in \mathcal{X} \times \mathcal{Y}$. By definition of all the objects involved, the action of $\tau \otimes \lambda$ on P is

$$\begin{aligned}\tilde{P}(A) &= \int_{(\tilde{x},\tilde{y})\in A} \left[\sum_{y\in Y} \left(\int_{x\in X} \tau_y(x, d\tilde{x})\, \lambda_x(y, d\tilde{y})\, E_y(dx)\right) \pi_y\right] \\ &= \int_{(\tilde{x},\tilde{y})\in A} \left[\int_{x\in X} \left(\sum_{y\in Y} \tau_y(x, d\tilde{x})\, \lambda_x(y, d\tilde{y})\, F_x(dy)\right) \pi_X(dx)\right]\,.\end{aligned}$$

In both the formulations above, obtained by factorizing the joint probability P in two different ways, we cannot isolate the action of one between λ and τ on F 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 P1, 2 and 4 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

$$\begin{aligned}\mathbb{E}_{(\tilde{\mathsf{X}},\tilde{\mathsf{Y}})\sim\tilde{P}}[\ell(h_{\tilde{\mathsf{X}}}, \tilde{\mathsf{Y}})] &= \int_{\tilde{x}\in X, \tilde{y}\in Y} \ell(h_{\tilde{x}}, \tilde{y}) \left[\sum_{y\in Y} \left(\int_{x\in X} \tau(x,y,d\tilde{x})\, \lambda(x,y,d\tilde{y})\, E_y(dx)\right) \pi_y\right] \\ &= \sum_{y\in Y} \left[\int_{x\in X} \left(\int_{\tilde{x}\in X} \left(\sum_{\tilde{y}\in Y} \ell(h_{\tilde{x}}, \tilde{y})\, \lambda(x,y,d\tilde{y})\right) \tau(x,y,d\tilde{x})\, E_y(dx)\right)\right] \pi_y \\ &= \sum_{y\in Y} \left[\int_{x\in X} \left(\int_{\tilde{x}\in X} \lambda\ell(h_{\tilde{x}}, x, y)\, \tau(x,y,d\tilde{x})\, E_y(dx)\right)\right] \pi_y \\ &= \sum_{y\in Y} \left[\int_{x\in X} \tau[\lambda\ell_{(x,y)} \circ h](x,y)\, E_y(dx)\right] \pi_y \qquad (28) \\ &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_Y \times E)} \left[\big(\tau(\lambda\ell_{(\mathsf{X},\mathsf{Y})} \circ h)\big)(\mathsf{X},\mathsf{Y})\right]\,.\end{aligned}$$

Following the same reasoning, we can also write

$$\mathbb{E}_{(\tilde{\mathsf{X}},\tilde{\mathsf{Y}})\sim\tilde{P}}[\ell(h_{\tilde{\mathsf{X}}},\tilde{\mathsf{Y}})] = \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim(\pi_X\times F)}\left[\left(\tau(\lambda\ell_{(\mathsf{X},\mathsf{Y})}\circ h)\right)(\mathsf{X},\mathsf{Y})\right] .$$

We prove point 2 and 3 minimizing both the obtained risk equalities w.r.t. $h \in \mathcal{H}$. ■

References

Rocío Alaiz-Rodríguez and Nathalie 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.

Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, 3rd edition, 2006.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2: 343–370, 1988.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79: 151–175, 2010.

David Blackwell. Comparison of experiments. *Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102, 1951.

Gilles Blanchard and Clayton Scott. Decontamination of mutually contaminated models. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1–9. PMLR, 2014.

Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10(2):2780–2824, 2016.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory (COLT)*, pages 92–100, 1998.

Robin J. Boyd, Gary D. Powney, and Oliver L. Pescott. We need to talk about nonprobability samples. *Trends in Ecology & Evolution*, 38(6):521–531, 2023. ISSN 0169-5347.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Online learning of noisy data with kernels. In *Conference on Computational Learning Theory (COLT)*, pages 218–230, 2010.

Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International conference on machine learning (ICML)*, pages 1789–1799. PMLR, 2020.

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7):938–971, 2019.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2010.

Andrew Cotter, Maya Gupta, and Harikrishna Narasimhan. On making stochastic classifiers deterministic. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019.

Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4):191–213, 1972.

Fredrik Dahlqvist, Vincent Danos, Ilias Garnier, and Ohad Kammar. Bayesian inversion by ω-complete cone duality. In *International Conference on Concurrency Theory*, 2016.

Jun Du and Zhihua Cai. Modelling class noise with symmetric and asymmetric distributions. In *AAAI Conference on Artificial Intelligence*, 2015.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems (NeurIPS)*, 2014.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning (ICML)*, pages 1386–1394. PMLR, 2015.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220, 2008.

Arnold M Faden. The existence of regular conditional probabilities: necessary and sufficient conditions. *The Annals of Probability*, 13(1):288–298, 1985.

Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. In *Advances in neural information processing systems (NeurIPS)*, pages 11996–12007, 2020.

Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Generalizing importance weighting to a universal solver for distribution shift problems. In *Advances in neural information processing systems (NeurIPS)*, 2023.

Kilian Fatras, Hiroki Naganuma, and Ioannis Mitliagkas. Optimal Transport meets Noisy Label Robust Loss and MixUp Regularization for Domain Adaptation. In Sarath Chandar, Razvan Pascanu, and Doina Precup, editors, *Proceedings of The 1st Conference on Lifelong Learning Agents*, pages 966–981. PMLR, 2022.

Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. Fairness evaluation in presence of biased noisy labels. In *Artificial intelligence and statistics (AISTATS)*, pages 2325–2336. PMLR, 2020.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys*, 46(4): 1–37, 2014.

Sergio Hernan Garrido Mejia, Elke Kirschbaum, Armin Kekić, and Atalanti Mastakouri. Estimating joint interventional distributions from marginal interventional data. *arXiv preprint arXiv:2409.01794*, 2024.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

Sally A. Goldman and Robert H. Sloan. Can PAC learning algorithms tolerate random attribute noise? *Algorithmica*, 14(1):70–84, 1995.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning (ICML)*, pages 2839–2848. PMLR, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015.

Luigi Gresele, Julius Von Kügelgen, Jonas Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. In *International conference on machine learning (ICML)*. PMLR, 2022.

Eric Grinstein, Ngoc QK Duong, Alexey Ozerov, and Patrick Pérez. Audio style transfer. In *IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 586–590, 2018.

Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, 32(4):1367–1433, 2004.

Haibo He and Edwardo A García. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. *Advances in neural information processing systems (NeurIPS)*, 2017.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *International conference on machine learning (ICML)*, pages 2971–2980. PMLR, 2019.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, page 694, 2016.

David Johnston. *Statistical Causal Modelling and Decision Theory*. PhD thesis, The Australian National University, 2023.

Olav Kallenberg. *Random measures, theory and applications*. Springer, 2017.

Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott. Decontamination of mutual contamination models. *Journal of machine learning research*, 20(41):1–72, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems (NeurIPS)*, 2017.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, 2007.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning (ICML)*, pages 5637–5664. PMLR, 2021.

Meelis Kull and Peter Flach. Patterns of dataset shift. In *First International Workshop on Learning over Multiple Contexts (LMCE) at ECML-PKDD*, 2014.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning (ICML)*, pages 3122–3130. PMLR, 2018.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Jiabin Liu, Bo Wang, Zhiquan Qi, Yingjie Tian, and Yong Shi. Learning from label proportions with generative adversarial networks. *Advances in neural information processing systems (NeurIPS)*, 2019.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12): 2346–2363, 2019.

Saunders Mac Lane. *Categories for the working mathematician*. Springer Science & Business Media, 2013.

Andrey Malinin, Neil Band, Yarin Gal, Mark Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Facundo Mémoli, Brantley Vose, and Robert C Williamson. Geometry and Stability of Supervised Learning Problems. *Journal of Machine Learning Research*, 26:1–99, 2025.

Xiao-Li 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.

Xiao-Li 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.

Aditya Menon, Brendan van Rooyen, Cheng Soon Ong, and Robert C Williamson. Learning from corrupted binary labels via class-probability estimation. In *International conference on machine learning (ICML)*, pages 125–134. PMLR, 2015.

Aditya Krishna Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, 2018.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45 (1):521–530, 2012.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems (NeurIPS)*, 2013.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

Arthur Parzygnat. Kleisli categories and probability - 03 - markov kernels. `https://youtu.be/psUDrasc21o?si=we87QEeKiGOa0_eN`, 2020.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen 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.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, Chichester, UK, 2016.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Mary Poovey. *A history of the modern fact: Problems of knowledge in the sciences of wealth and society*. University of Chicago Press, 1998.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. In *International conference on machine learning (ICML)*, pages 776–783, 2008.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2022.

Negar Rostamzadeh, Ben Hutchinson, Christina Greer, and Vinodkumar Prabhakaran. Thinking beyond distributions in testing machine learned models. In *NeurIPS Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Jonathan Rothwell. How the war on drugs damages black social mobility. *The Brookings Institution*, 30, 2014.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

José A. Sáez. Noise models in classification: Unified nomenclature, extended taxonomy and pragmatic categorization. *Mathematics*, 10(20), 2022.

Marcos Salganicoff. Tolerating concept and sampling shift in lazy learning using prediction error context switching. *Artificial Intelligence Review*, 11:133–155, 1997.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora 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.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics (AISTATS)*, pages 838–846. PMLR, 2015.

Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. *Advances in neural information processing systems (NeurIPS)*, 33:22256–22267, 2020.

George Shackelford and Dennis Volper. Learning k-DNF with noise in the attributes. In *First annual workshop on Computational Learning Theory (COLT)*, pages 97–103, 1988.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Albert N Shiryaev and Vladimir G Spokoiny. *Statistical Experiments And Decision, Asymptotic Theory*. World Scientific, 2000.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

Adarsh Subbaswamy, Bryant Chen, and Suchi Saria. A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, 10(1): 64–89, 2022.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision (ECCV)*, pages 709–726. Springer, 2022.

Yuting Tang, Nan Lu, Tianyi Zhang, and Masashi Sugiyama. Multi-class classification from multiple unlabeled datasets with partial risk regularization. In *Asian Conference on Machine Learning (ACML)*, pages 990–1005. PMLR, 2023.

Erik Torgersen. *Comparison of statistical experiments*. Cambridge University Press, 1991.

Alexey 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.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of machine learning research*, 18(228):1–50, 2018.

Brendan van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems (NeurIPS)*, 28, 2015.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 839–847, 2017.

Peter Vorburger and Abraham Bernstein. Entropy-based concept shift detection. In *Sixth International Conference on Data Mining (ICDM)*, pages 1113–1118. IEEE, 2006.

Feng-Yu Wang. Coupling and applications. In *Stochastic Analysis and Applications to Finance: Essays in Honour of Jia-An Yan*, pages 411–424. World Scientific, 2012.

Qizhou Wang, Bo Han, Tongliang Liu, Gang Niu, Jian Yang, and Chen Gong. Tackling instance-dependent label noise via a universal probabilistic model. In *AAAI Conference on Artificial Intelligence*, pages 10183–10191, 2021.

Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, and John R Leathwick. Presence-only data and the EM algorithm. *Biometrics*, 65(2):554–563, 2009.

Halbert White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981.

Gerhard Widmer and Miroslav Kubat. Effective learning in dynamic environments by explicit context tracking. In *European Conference on Machine Learning (ECML)*, volume 6, pages 227–243, 1993.

Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23:69–101, 1996.

Robert C Williamson. Process and Purpose, Not Thing and Technique: How to Pose Data Science Research Challenges. *Harvard Data Science Review*, 2(3), 2020.

Robert C. Williamson and Zac 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.

Keisuke Yamazaki, Motoaki Kawanabe, Sumio Watanabe, Masashi Sugiyama, and Klaus-Robert Müller. Asymptotic bayesian generalization error when training and test distributions are different. In *International conference on machine learning (ICML)*, pages 1079–1086, 2007.

Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4409–4420, 2021.

Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

Xiyu Yu, Tongliang Liu, Mingming Gong, Kun Zhang, Kayhan Batmanghelich, and Dacheng Tao. Label-noise robust domain adaptation. In *International conference on machine learning (ICML)*, pages 10913–10924. PMLR, 2020.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning (ICML)*, pages 819–827. PMLR, 2013.

Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, Qingsong Liu, and Clark Glymour. Domain adaptation as a problem of inference on graphical models. *Advances in neural information processing systems (NeurIPS)*, pages 4965–4976, 2020a.

Tianyi Zhang, Ikko Yamane, Nan Lu, and Masashi Sugiyama. A one-step approach to covariate shift adaptation. In *Asian Conference on Machine Learning (ACML)*, pages 65–80. PMLR, 2020b.

Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *The Artificial Intelligence Review*, 22(3):177, 2004.