Geometry and Stability of Supervised Learning Problems

Facundo Mémoli   FACUNDO.MEMOLI@GMAIL.COM
Department of Mathematics
Rutgers University, Piscataway, NJ 08854

Brantley Vose   VOSE.5@OSU.EDU
Department of Mathematics
The Ohio State University, Columbus, OH 43210

Robert C. Williamson   BOB.WILLIAMSON@UNI-TUEBINGEN.DE
University of Tübingen and Tübingen AI Center, 72076 Tübingen, Germany

Editor: Pierre Alquier

Abstract

We introduce a notion of distance between supervised learning problems, which we call the *Risk distance*. This distance, inspired by optimal transport, facilitates stability results; one can quantify how seriously issues like sampling bias, noise, limited data, and approximations might change a given problem by bounding how much these modifications can move the problem under the Risk distance. With the distance established, we explore the geometry of the resulting space of supervised learning problems, providing explicit geodesics and proving that the set of classification problems is dense in a larger class of problems. We also provide two variants of the Risk distance: one that incorporates specified weights on a problem's predictors, and one that is more sensitive to the contours of a problem's risk landscape.

Keywords: supervised learning, stability, metric geometry, optimal transport, risk landscape

1. Introduction

In machine learning, even before beginning work on a problem, we are often forced to accept discrepancies between the problem we want to solve and the problem we actually get to work with. We may put up with noise or bias in the data collection process which distorts our view of the true distribution of observations. We may replace a loss function with a surrogate loss whose computational cost or optimization properties are preferable. We may have access to only a small sample of observations. Such compromises are a necessary reality.

This brings us to our primary motivating questions.

- **Question 1**: How much can such a compromise change our problem and its descriptive features?

- **Question 2**: How much effect can multiple compromises have in conjunction? Can we guarantee that a sequence of small changes will not have drastic effects on the problem to be solved?

1.1 Overview of our approach.

In this paper, we provide a comprehensive framework with which to answer such questions. While many frameworks exist to answer Question 1, these methods are concerned with quantifying changes to one or two aspects of a problem at a time, limiting their ability to answer Question 2. In real world problems, multiple simultaneous corruptions and substitutions are unavoidable. Our framework is broad enough to handle simultaneous changes to many aspects of a problem. To begin, we give a precise definition of "supervised learning problem" and define a notion of distance, dubbed the *Risk distance* and denoted d_R, by which to compare pairs of problems. This gives rise to the *(pseudo)metric space of supervised learning problems*. The Risk distance lets us make geometric sense of Questions 1 and 2; we can measure how much a compromise affects a problem by seeing how far the problem moves under the Risk distance.

To actually define the Risk distance d_R, we draw on the wisdom of metric geometry. In 1975, Edwards constructed a metric on the set of isomorphism classes of compact metric spaces which came to be known as the Gromov-Hausdorff distance (Edwards, 1975; Gromov, 1981). Similarly, the Gromov-Wasserstein distance, introduced by Mémoli (2007, 2011), provides an optimal-transport-based metric on the collection of isomorphism classes of metric spaces equipped with probability measures. The Gromov-Hausdorff and Gromov-Wasserstein distances have become integral to the theory of metric geometry by facilitating a geometric understanding of spaces in which the points themselves are spaces.

Inspired by this tradition, we craft the Risk distance in the image of the Gromov-Wasserstein distance. This sets up the following analogy between supervised learning and metric geometry.

Supervised Learning		Metric Geometry
Problem	$\longleftrightarrow$	Metric measure space
Loss function	$\longleftrightarrow$	Metric
Risk distance	$\longleftrightarrow$	Gromov-Wasserstein distance

1.2 Previous approaches.

Question 1 has previously been explored for various specific kinds of changes to a problem. In our work, a *supervised learning problem* (or just *problem*, for short), is modeled as a 5-tuple $P = (X, Y, \eta, \ell, H)$ where X and Y are the *input space* and *response space* respectively, η is *the joint law*: a probability measure on $X \times Y$, ℓ is the *loss function* and H is the *predictor set*: a collection of functions $h : X \to Y$.

- **Joint law η.** A common assumption in machine learning is that the data is drawn according to an unknown underlying probability measure. Concepts such as noise and bias in data collection or data shift phenomena such as covariate shift, label shift, or concept drift (Huyen, 2022, Ch 8), can be described as changes to this underlying measure. The effects of various kinds of noise (Zhu and Wu, 2004; Natarajan et al., 2013; Menon et al., 2015, 2018; Rooyen and Williamson, 2018; Iacovissi et al., 2023) and data shifts (Shimodaira, 2000) on supervised learning problems is a longstanding area of research. A geometric framework for understanding changes in probability

measures exists as well; information geometry seeks to understand spaces of probability measures from the viewpoint of Riemannian geometry (Amari, 2016; Ay et al., 2017).

- **Loss function ℓ.** A theoretically attractive loss may have poor optimization properties, prompting one to replace it with a so-called surrogate loss. Alternatively, an attractive loss could be expensive to exactly compute, making it a candidate for approximation, like replacing a Wasserstein-based loss with a Sinkhorn-based loss in probability estimation (Cuturi, 2013). Replacing the loss represents a tradeoff between theoretical properties and computational efficiency, and the quantitative details of this tradeoff have been explored in many contexts (Lin, 2004; Zhang, 2004; Bartlett et al., 2006; Steinwart, 2007; Awasthi et al., 2022; Mao et al., 2023).

- **Predictor set H.** Universal approximation theorems, popular in deep learning, establish that certain classes of models can approximate large classes of predictors arbitrarily well (Pinkus, 1999). These can be seen as theorems which compare large, intractable predictor sets to those that can be produced by a given model, aiming to show that there is effectively no difference between selecting a predictor from either set. Approximation theory more broadly is similarly concerned with the approximation power of function classes or transforms (Trefethen, 2019).

- **Input and response spaces X and Y.** Modifications of the input and response spaces are implicit in many of the modifications listed above. Additionally, the process of feature engineering (Huyen, 2022, Ch 5), for which there is little established theory, can be seen as a modification of the input space. Examples of output space modification include the common relaxation of classification to class probability estimation or, less common, the discretization of the response space into a finite set of labels (Langford and Zadrozny, 2005) which served as a motivation for the exact results by Reid and Williamson (2011).

- **Combinations.** Our notion of a learning problem comprises five separate components. There are some existing results that explore how changes to some components affect the others. Some of this work goes under the name of "Machine Learning Reductions" (Langford, 2009). There are results on how changes to the distribution of examples η (due to "noise") has the effect of changing the loss function ℓ (with label noise) (Rooyen and Williamson, 2018) or the model class H (with attribute noise) (Bishop, 1995; Dhifallah and Lu, 2021; Rothfuss et al., 2019; Smilkov et al., 2017; Williamson and Cranko, 2024).

There is precedent for geometric frameworks regarding statistical learning. We have mentioned the rise of information geometry as a way to understand statistical manifolds. Also, much as we aim to do with supervised learning problems, Le Cam developed a notion of distance (Le Cam, 1964; Mariucci, 2016) between the "statistical experiments" of Blackwell (Blackwell, 1951). A modified discrepancy incorporating a loss function was introduced by Torgersen (1991, Complement 37). We will discuss the relationship between supervised learning problems and statistical experiments in Section 3.2.

While much research has sought to quantify and control the effects of various specific corruptions on supervised learning problems, no framework so far has been comprehensive
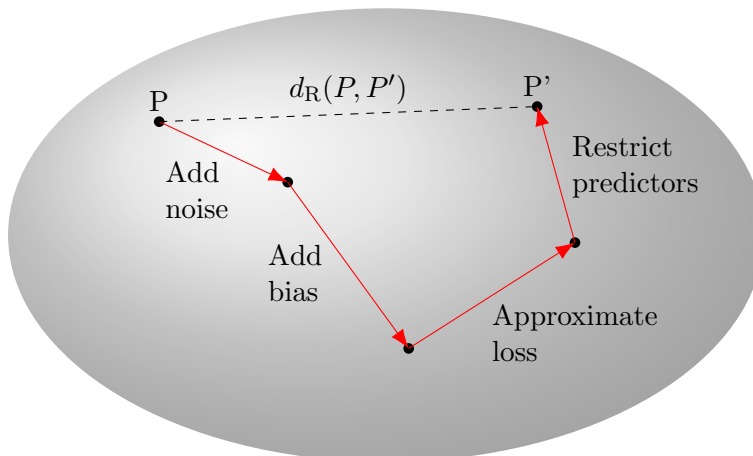
Figure 1: A simple depiction of how a sequence of changes can push a particular problem through the space of problems. An idealized problem P is modified by a sequence of four changes, each change moving the problem some distance in the space of problems. Once the sequence of corruptions has been applied, we call the resulting problem P'. The original and corrupted problems are connected by a dashed line representing the risk distance between them.

enough to consider changes to all aspects of a problem. That is, while many have given answers to Question 1, to the best of our knowledge no attempt has been made at a unified answer to Question 2.

1.3 Our contributions.

We provide a geometric framework with which to answer both Questions 1 and 2 in the form of the Risk distance.

Our primary tool for answering Question 1 with the Risk distance is the central topic of Section 4: stability results. A stability result is an upper bound on how much a certain change to a problem can move that problem under the Risk distance. That is, if we apply a certain change to a problem P to obtain a new problem P', such as a change to the loss function, the data collection process, or the set of potential predictors, a stability result is an upper bound on $d_{\mathrm{R}}(P, P')$. Such a result acts as a quantitative guarantee that a small change to some aspect of the problem will not drastically change the problem as a whole.

Comparing problems via a metric also presents advantages over more general kinds of dissimilarity scores: the triangle inequality lets one string together upper bounds on distances. To wit, let x_0 be a point in a metric space (X, d_X). Suppose we create a sequence of points $x_0, x_1, \ldots, x_k$ by pushing each point x_{i-1} a distance of at most ε to get the next point x_i in the sequence. Then the triangle inequality assures us that

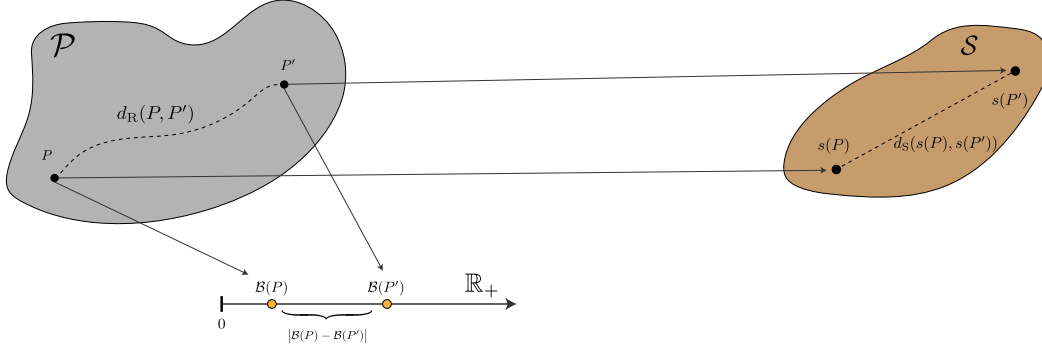$$d_X(x_0, x_k) \leqslant \sum_{i=1}^{k} d_X(x_{i-1}, x_i) \leqslant k\varepsilon.$$

Figure 2: The space $\mathcal{P}$ of all supervised learning problems and a descriptor. A descriptor is modeled as a map $s : \mathcal{P} \to \mathcal{S}$ whose codomain is some metric space $(\mathcal{S}, d_{\mathcal{S}})$. A descriptor s is regarded as stable if there is some constant $C > 0$ such that $C\, d_{\mathrm{R}}(P, P) \geqslant d_{\mathcal{S}}(s(P), s(P'))$ for all $P, P' \in \mathcal{P}$. The figure also depicts the case when $\mathcal{S}$ is the set $\mathbb{R}_+$ of non-negative reals and $s(P)$ is the constrained Bayes risk $\mathcal{B}(P)$ of P. In that case, one of our stability results (Theorem 37) establishes that $d_{\mathrm{R}}(P, P') \geqslant \left| \mathcal{B}(P) - \mathcal{B}(P') \right|$ for all $P, P' \in \mathcal{P}$.

In other words, in a metric space, a short sequence of small pushes cannot add up to a drastic change. In the context of the Risk distance, this feature lets us handle multiple simultaneous changes to a problem by applying the triangle inequality and stringing together stability results. This gives an answer to Question 2. See Figure 1 for an illustration. By design, the Risk distance will be a pseudometric (that is, a metric for which distinct points can be distance zero apart), but the above virtues of metric spaces apply to pseudometric spaces as well.

Along with stability of the Risk distance under various problem modifications, Section 4 will also demonstrate stability of certain *descriptors of learning problems* with respect to the Risk distance. In particular, we prove in Section 4.1 that a descriptor of a problem called its *constrained Bayes risk* is stable under the Risk distance. In other words, if the Risk distance between two problems P and P' is less than ε, then the optimal expected loss across all predictors for P is within ε of that of P'; see Figure 2 for an illustration. We obtain this stability result by first proving stability under the Risk distance of a stronger descriptor, called the *loss profile set* of a problem, which describes the possible distributions of losses incurred by a random observation. We also demonstrate stability under the Risk distance of a descriptor from statistical learning theory known as the *Rademacher complexity* for arbitrarily large sample sizes.

In addition to facilitating stability results, the Risk distance gives us a notion of convergence for supervised learning problems. Just as it is fruitful to discuss the convergence of sequences of functions, random variables, or probability distributions, so too can it be useful to discuss the convergence of a sequence of problems. In Section 5.4, we will demonstrate an analog of the Glivenko–Cantelli theorem by proving that the "empirical problem," a problem constructed using a finite random sample of observations instead of the true distribution

from which it is sampled, converges to the "true problem" under the Risk distance with probability 1.

Lastly, the Risk distance lets us consider the geometry of the space of problems and consequences thereof. In Sections 5.1 and 5.2, we connect the geometric concept of a *geodesic* between two problems with the probabilistic notion of couplings of distributions associated with those problems. This connection lets us write down some explicit geodesics in the space of problems. In Section 5.3, we show that classification problems, meaning those problems with a finite output space, are dense in a larger space of problems under the Risk distance. This density result proves useful when establishing the convergence of empirical problems.

In the final two sections, we also study two variants of the Risk distance. In Section 6, we define a parameterized family of distances, each called the L^p-*Risk distance* for $p \in [1, \infty]$. The L^p-Risk distance incorporates weights in the form of a specified probability measure on the set of possible predictors. The use of this additional data results in improved stability when compared to the Risk distance. In Section 7, we also define a stronger version of the Risk distance, called the *Connected Risk distance*, which is designed to be more sensitive to changes in the risk landscape of a problem. In particular, we show that a topological summary of the problem called its *Reeb graph* is stable under the Connected Risk distance.

1.4 Outlook

We regard the Risk distance as a tool to produce theoretical guarantees. To prove that problems are stable under a certain kind of change with respect to the Risk distance, or that certain descriptors are stable under the Risk distance, we need only bound the Risk distance above or below. The utility of the Risk distance lies not in explicit algorithms and associated computations, but in the theoretical landscape that it facilitates. The current landscape of machine learning is rapidly expanding, particularly in the space of problem formulations. Given the staggering growth of machine learning research (e.g., 27,000 submissions to NeurIPS 2025 (NeurIPS, 2025)), there is clear value in providing a conceptual framework to help organize and interpret this proliferation of methods. This paper takes an initial step toward an organizing 'map' of this space—namely, we endow the class of supervised learning problems with a metric that induces a geometric organization, so existing and emerging methods can be situated and compared within a common landscape. While this paper focuses on supervised learning, many generative models, such as GANs, are implicitly tied to supervised learning problems via divergence measures—e.g., f-divergences—which are in one-to-one correspondence with certain supervised objectives (Nowozin et al., 2016; Williamson and Cranko, 2024).

1.5 Other previous work.

Kleinberg's seminal paper (Kleinberg, 2002) initiated a thread of research into the formal study of clustering problems (Carlsson and Mémoli, 2010; Carlsson and Mémoli, 2013; Ackerman, 2012; Cohen-Addad et al., 2018). In a spirit similar to this thread, the present paper aims to formally examine supervised learning problems.

The Risk distance introduced in this paper participates in a growing thread of research into optimal-transport-based distances, a thread which begins with the metrics discussed in Section 2.3 and runs through more recent variants including optimal transport distances

for graph-based data (Chowdhury and Mémoli, 2019; Chen et al., 2022), matrices (Peyré et al., 2016), partitioned data sets (Chowdhury et al., 2021), labeled data sets (Alvarez-Melis and Fusi, 2020), optimal transport of both data and features jointly (Vayer et al., 2020b), or of data with quite general structures (Vayer et al., 2020a; Patterson, 2020), as well as approximations (Scetbon et al., 2022; Sejourne et al., 2021; Vincent-Cuaz et al., 2021) and regularizations (Peyré et al., 2016) thereof.

1.6 Statement of contribution

The genesis of this project dates back to 2010-2011 when F.M. visited R.W. at NICTA (Canberra, Australia). The formulation of the results, finding the proofs, and the vast majority of the writing of the paper was done by F.M. and B.V. R.W. contributed the initial formulation of the question to be addressed, provided some context from the ML literature, and reviewed and slightly revised the final document.

Contents

2. Background

We begin with some necessary concepts and notation from measure theory, metric geometry, and optimal transport. We refer the reader to the standard reference of Villani (2003) for more on measure-theoretic and optimal transport concepts, and that of Burago, Burago, and Ivanov (2001) for metric geometry concepts.

2.1 Preliminary Notation

We use π_i to represent a projection map $X_1 \times X_2 \times \cdots \times X_n \to X_i$ onto the ith component of a product. More generally, $\pi_{i_1,\ldots,i_k} X_1 \times \cdots \times X_n \to X_{i_1} \times \cdots \times X_{i_k}$ represents the projection onto coordinates $i_1, \ldots, i_k$.

When integrating a measurable function f against a measure μ, we will favor the notation

$$\int_{X \times Y} f(x, y) \, \mu(dx \times dy)$$

if f is a function of two variables, and similarly for more variables. We often suppress the domain of integration under the integral symbol since it is implicit in the measure μ. Rarely, when many variables are necessary and horizontal space is at a premium, we may instead use the notation

$$\int f(x, y) \, d\mu(x, y).$$

Given a measurable space (Z, Σ_Z), we let $\mathrm{Prob}(Z)$ denote the space of probability measures on (Z, Σ_Z). For any $z \in Z$, we let $\delta_z \in \mathrm{Prob}(Z)$ denote the *Dirac probability measure at z* defined, for $A \in \Sigma_Z$, as $\delta_z(A) = 1$ whenever $z \in A$ and as 0 otherwise. Given a probability measure $\alpha \in \mathrm{Prob}(\mathbb{R})$, we let $\mathrm{mean}(\alpha) := \int_{\mathbb{R}} t \, \alpha(dt)$ denote its mean whenever it exists.

2.2 Pushforward Measures, Correspondences, and Couplings

We recall some basic ideas from measure theory and establish notation. Given a measure space (X, Σ_X, μ), a measurable space (Y, Σ_Y), and a measurable map $f : X \to Y$, the *pushforward of μ through f* is the measure on (Y, Σ_Y) defined by

$$f_\sharp \mu(A) := \mu(f^{-1}(A))$$

for all $A \in \Sigma_Y$. If μ is a probability measure, $f_\sharp \mu$ will be a probability measure as well. Hence $f_\sharp$ is a function $\mathrm{Prob}(X) \to \mathrm{Prob}(Y)$. A pushforward has a probabilistic interpretation: if X is a random variable in a measurable space X with law μ, and $f : X \to Y$ is measurable, then the random variable $f(\mathsf{X})$ has law $f_\sharp \mu$.

Given two sets X and Y, a subset $R \subseteq X \times Y$ is called a *correspondence* if

- for all $x \in X$ there is some $y \in Y$ with $(x, y) \in R$, and

- for all $y \in Y$ there is some $x \in X$ with $(x, y) \in R$.

We use $\mathcal{C}(X, Y)$ to denote the set of all correspondences between X and Y.

A *coupling* between two probability measure spaces (X, Σ_X, μ_X) and (Y, Σ_Y, μ_Y) is a probability measure on $X \times Y$ with marginals μ_X and μ_Y. That is, a probability measure γ on $X \times Y$ is a coupling if the pushforward measures satisfy $(\pi_1)_\sharp \gamma = \mu_X$, $(\pi_2)_\sharp \gamma = \mu_Y$. We use $\Pi(\mu_X, \mu_Y)$ to denote the set of all such measures. Couplings can be seen as a probabilistic counterpart to correspondences.

The following are standard methods for constructing correspondences between two sets X and Y.

- The entire product $X \times Y$ is always a correspondence. In particular, this shows that $\mathcal{C}(X, Y)$ cannot be empty.

- If $X = Y$, one can construct the *diagonal correspondence* $\{(x, x) | x \in X\}$.

- If $Y = \{\bullet\}$ is the one point set, then there is a unique correspondence $X \times \{\bullet\}$.

- The graph of any surjection $f : X \to Y$ is a correspondence between X and Y.

Analogous constructions exist for couplings as well. Let (X, μ_X) and (Y, μ_Y) be probability spaces.

- The product measure $\mu_X \otimes \mu_Y$ is always a coupling, sometimes called the *independent coupling*. This shows that $\Pi(\mu_X, \mu_Y)$ cannot be empty.

- Suppose $(X, \mu_X) = (Y, \mu_Y)$. If $\Delta_X : X \to X \times X$ denotes the diagonal map $\Delta_X(x) = (x, x)$, then the *diagonal coupling* is the pushforward measure $(\Delta_X)_\sharp \mu_X$.

- If $Y = \{\bullet\}$ is the one point set, then forcibly $\mu_Y = \delta_\bullet$ and there is a unique coupling $\mu_X \otimes \delta_\bullet$ in this case.

- Suppose $f : X \to Y$ is a measure-preserving map, meaning $f_\sharp \mu_X = \mu_Y$. Let $i_f : X \to X \times Y$ be the map sending X to the graph of f, meaning $i_f(x) := (x, f(x))$. Then the pushforward $(i_f)_\sharp \mu_X$ is a coupling between μ_X and μ_Y.

When we discuss Markov kernels in Section 2.5, we will mention one more method for generating couplings by combining a probability measure $\alpha \in \mathrm{Prob}(X)$ and a Markov Kernel $\beta : X \to \mathrm{Prob}(Y)$ using reverse disintegration.

We will make use of the following standard coupling result, which can be found in any standard optimal transport text, such as that of Villani (2003, Lemma 7.6). Note that a topological space is called *Polish* if it is separable and metrizable by a complete metric. Most familiar spaces are Polish, including $\mathbb{R}^n$ and open subsets thereof, manifolds with or without boundary, and finite discrete spaces. A Borel measurable space arising from a Polish topological space is called Polish as well.

Proposition 1 (The Gluing Lemma) *Let μ_1, μ_2 and μ_3 be probability measures on Polish spaces X_1, X_2, and X_3 respectively. Let $\gamma_{1,2} \in \Pi(\mu_1, \mu_2)$ and $\gamma_{2,3} \in \Pi(\mu_2, \mu_3)$. Then there exists a probability measure $\gamma_{1,2,3}$ on $X_1 \times X_2 \times X_3$ with marginals $\gamma_{1,2}$ on $X_1 \times X_2$ and $\gamma_{2,3}$ on $X_2 \times X_3$.*

We call such a $\gamma_{1,2,3}$ a *gluing* of $\gamma_{1,2}$ and $\gamma_{2,3}$. An analog of the gluing lemma for correspondences (see Burago et al., 2001, Exercise 7.3.26) instead of couplings is included below in order to highlight the parallels between couplings and correspondences.

Proposition 2 (The Gluing Lemma for Correspondences) *Let A_1, A_2 and A_3 be sets, and let $R_{1,2} \in \mathcal{C}(A_1, A_2)$ and $R_{2,3} \in \mathcal{C}(A_2, A_3)$. Then there exists a subset $R_{1,2,3} \subseteq A_1 \times A_2 \times A_3$ which projects to $R_{1,2}$ on $A_1 \times A_2$ and $R_{2,3}$ on $A_2 \times A_3$.*

We call such an $R_{1,2,3}$ a *gluing* of $R_{1,2}$ and $R_{2,3}$.

Proof Selecting $R_{1,2,3} = \{(x, y, z) \in A_1 \times A_2 \times A_3 | (x, y) \in R_{1,2}, (y, z) \in R_{2,3}\}$ provides one such subset. ∎

While some measurable spaces can be quite poorly behaved (for instance, those that fail to exhibit a gluing lemma), the measurable spaces in this paper will all arise from relatively well-behaved topological spaces.

Assumption 3 *We assume that any measurable space that appears will arise from a Polish topological space by equipping it with the Borel σ-algebra. Therefore, when convenient, we will suppress the σ-algebra in our notation for measurable spaces. That is, instead of writing a measurable space as a pair (X, Σ_X) with X a Polish space and Σ_X its Borel σ-algebra, we will simply write X with the understanding that the Borel σ-algebra is being used. We will similarly write measure spaces (X, Σ_X, μ) as pairs (X, μ).*

Bearing Assumption 3 in mind, we define the *support* of the measure μ as the closed set

$$\mathrm{supp}[\mu] := \{x \in X \mid \mu(U) > 0 \,\forall\, U \text{ open s.t. } x \in U\}.$$

2.3 Metric Geometry and Optimal Transport

Here we present four metrics from the field of metric geometry. The following diagram (inspired by Mémoli, 2011, Fig. 2) illustrates the relationships between the four metrics.
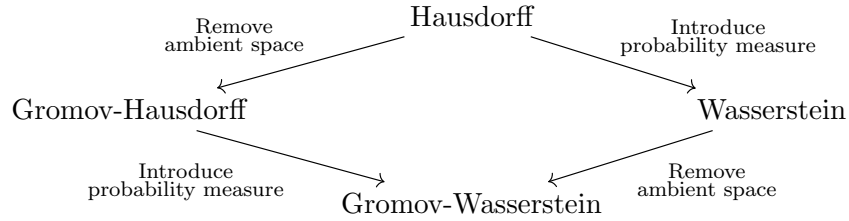
We do not explicitly use the Gromov-Hausdorff or Gromov-Wasserstein distances in this paper. However, these distances provide context by illustrating common threads that run through all four metrics and is shared by the Risk distance. In all four of the metrics, objects are compared by optimally "aligning" the two objects, whether via correspondences or couplings, and numerically scoring the distortion incurred by an optimal alignment. The Risk distance (Definition 26) follows this same template. To emphasize the continuation of this thread, we will take a moment to compare the Risk distance to the Gromov-Wasserstein distance in Section 3.5.

Note that we will often suppress the underlying metric in our notation, writing a metric space (X, d_X) as simply X.

2.3.1 Hausdorff Distance

The first metric is the Hausdorff distance. Introduced by Hausdorff (1914), it provides a notion of distance between compact subsets of a fixed metric space. Given two compact

subsets X, Y of a metric space (Z, d_Z), the Hausdorff distance between them is typically defined to be

$$d_{\mathrm{H}}^Z(X, Y) := \max \left\{ \sup_{x \in X} \inf_{y \in Y} d_Z(x, y), \sup_{y \in Y} \inf_{x \in X} d_Z(x, y) \right\}.$$

To emphasize the commonalities among the four metrics of this section, we will use an alternative equivalent definition provided by Mémoli (2011, Prop. 2.1) in terms of correspondences.

Proposition 4 *Given two compact subsets X and Y of a metric space (Z, d_Z), the Hausdorff distance between X and Y is equal to*

$$d_{\mathrm{H}}^Z(X, Y) = \inf_{R \in \mathcal{C}(X, Y)} \sup_{(x, y) \in R} d_Z(x, y).$$

2.3.2 GROMOV-HAUSDORFF DISTANCE

The second metric of this section is the Gromov-Hausdorff distance, an extension of the Hausdorff distance that removes the need for a common ambient metric space. While the Hausdorff distance compares compact subsets of a common metric space Z, the Gromov-Hausdorff distance compares pairs of compact metric spaces.

Definition 5 *Given two compact metric spaces (X, d_X) and (Y, d_Y), the* Gromov-Hausdorff *distance between them is given by*

$$d_{\mathrm{GH}}(X, Y) := \inf_{R \in \mathcal{C}(X, Y)} \sup_{(x, y), (x', y') \in R} \frac{1}{2} \big| d_X(x, x') - d_Y(y, y') \big|.$$

The function d_{GH} was first discovered by Edwards (1975, Def. III.3) and later rediscovered by its namesake Gromov (1981), both with alternative (but equivalent) definitions different from the one given in Definition 5. The formulation above in terms of correspondences is due to Kalton and Ostrovskii (1997). This formula does indeed define a metric on the set of compact metric spaces up to isometry, meaning that $d_{\mathrm{GH}}(X, Y) = 0$ if and only if X and Y are isometric (Burago et al., 2001, Thm. 7.3.30).

2.3.3 WASSERSTEIN DISTANCE

The third metric of this section is the Wasserstein distance.

Definition 6 (Kantorovich, 1942) *Let $p \in [1, \infty)$. Given two probability measures μ, ν on a metric space (Z, d_Z), the p-Wasserstein distance between μ and ν is*

$$d_{\mathrm{W}, p}^Z(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \left(\int_{X \times X} d_Z^p(x, y) \, \gamma(dx \times dy) \right)^{1/p}.$$

Notice that to construct the Wasserstein distance, we can begin with the Hausdorff metric and replace the correspondences with their probabilistic counterparts: couplings. We then replace the supremum with an L^p-style integral, and our job is done. Moving forward,

since we primarily make use of the 1-Wasserstein distance, we will refer to $d_{W,1}$ simply as the *Wasserstein distance*.

The Wasserstein distance is also called the *earth mover's distance* due to the following interpretation. Suppose (X, d_X) represents the ground of a garden equipped with the 2-d Euclidean metric, and μ describes the distribution of a pile of soil on X. Suppose that ν is a desired distribution of soil. Then a coupling $\gamma \in \Pi(\mu, \nu)$ represents a plan for moving the soil from distribution μ to distribution ν, and $d_{W,1}^X(\mu, \nu)$ is the amount of work that an optimal plan will require. Indeed, the Wasserstein metric seems to have first appeared in the work of Soviet mathematician and economist L. V. Kantorovich, who was studying optimal transportation of resources. For an English translation of Kantorovich's 1942 paper, see Kantorovich (2006), and for a historical survey of this often-rediscovered metric, see Vershik (2006).

2.3.4 GROMOV-WASSERSTEIN DISTANCE

The fourth metric of this section is the Gromov-Wasserstein distance. Just as the Gromov-Hausdorff metric is an ambient-space-free variant of the Hausdorff distance, the Gromov-Wasserstein distance is an ambient-space-free variant of the Wasserstein distance. It can be used to compare probability measures regardless of the underlying metric space.

Definition 7 *A* metric measure space *is a triple (X, d_X, μ_X) such that (X, d_X) is a metric space, and μ_X is a Borel probability measure on X with full support.*

For more on metric measure spaces, we direct the reader to Mikhael Gromov's foundational book (Gromov, 2007, Chapter $3\frac{1}{2}$).

Definition 8 (Mémoli (2007, 2011)) *Let $p \in [1, \infty)$. The p-Gromov-Wasserstein distance between two metric measure spaces (X, d_X, μ_X) and (Y, d_Y, μ_Y) is given by*

$$d_{GW,p}(X, Y) := \inf_{\gamma \in \Pi(\mu_X, \mu_Y)} \frac{1}{2} \left(\int_{X \times Y \times X \times Y} \left| d_X(x, x') - d_Y(y, y') \right|^p \gamma(dx \times dy) \gamma(dx' \times dy') \right)^{1/p}.$$

As with the Wasserstein distance, we will typically take $p = 1$ and refer to $d_{GW,1}$ simply as the *Gromov-Wasserstein distance*.

While preceded by Gromov's box distance and observable distance (Gromov, 1981, Ch. $3\frac{1}{2}$), the first to formulate a "Wasserstein-type" adaptation of the Gromov-Hausdorff distance to the setting of metric measure spaces was Sturm (2006). The definition above was introduced and developed by Mémoli (2007, 2011) in order to apply methods from optimal transport to the computational problem of data comparison and matching. Sturm (2023) extensively explored the geometry of the resulting "space of spaces". The Gromov-Wasserstein distance has spawned variants in many domains (some of which are referenced in Section 1) and applications to various problems in Machine Learning and Data Science such as unsupervised natural language translation (Alvarez-Melis and Jaakkola, 2018), imitation learning (Fickinger et al., 2022), generative modeling (Bunne et al., 2019), and to the study of Graph Neural Networks in the form of the Weisfeiler-Lehman distance and a corresponding version of the Gromov-Wasserstein distance for comparing labelled Markov processes (Chen et al., 2022). See also Peyré et al. (2019); Vayer et al. (2020a, 2019); Peyré et al. (2016);

Scetbon et al. (2022); Chowdhury and Needham (2020); Sejourne et al. (2021); Delon et al. (2022); Beier et al. (2022); Le et al. (2022); Dumont et al. (2024); Arya et al. (2024) for other work related to the Gromov-Wasserstein distance.

2.4 Total Variation Distance

There is one additional metric that we will use. The simplest method of comparing two probability measures on a common measurable space is the *total variation distance*. See Levin et al. (2017, Sec 4.1) for an introduction.

Definition 9 *Let μ, ν be probability measures on the measurable space $(\Omega, \mathcal{S})$. The* total variation distance *between μ and ν is given by*

$$d_{\mathrm{TV}}(\mu, \nu) := \sup_{B \in \mathcal{S}} \left| \mu(B) - \nu(B) \right|.$$

It is not actually necessary to take an absolute value in the definition of d_{TV}, since the value $|\mu(B) - \nu(B)|$ is achieved by either $\mu(B) - \nu(B)$ or $\mu(\Omega \backslash B) - \nu(\Omega \backslash B)$. Hence,

$$d_{\mathrm{TV}}(\mu, \nu) = \sup_{B \in \mathcal{S}} \left(\mu(B) - \nu(B) \right).$$

Additionally, the total variation distance can be seen as the Wasserstein distance when Ω is equipped with the discrete metric. In other words,

$$d_{\mathrm{TV}}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int 1_{x \neq y} \, \gamma(dx \times dy) = \inf_{\gamma \in \Pi(\mu, \nu)} \gamma \left\{ (x, y) | x \neq y \right\}.$$

Note that the above equalities only make sense if the diagonal $\Delta := \{(x, x) | x \in \Omega\}$ is measurable. This is true under Assumption 3; if Ω is a Polish space, then Δ is closed in $\Omega \times \Omega$.

2.5 Markov Kernels

Here we establish definitions and notation surrounding Markov kernels. See Klenke (2014, Ch 14.2) for a comprehensive introduction.

Let (X, Σ_X) and (Y, Σ_Y) be measurable spaces. A map $M : X \times \Sigma_Y \to \mathbb{R}$ is called a *Markov kernel if $M(x, \cdot) \in \mathrm{Prob}(Y)$ for all $x \in X$, and $M(\cdot, A)$ is measurable for all $A \in \Sigma_Y$. One often writes $M(x)(A)$ instead of $M(x, A)$ and thinks of M as a map $X \to \mathrm{Prob}(Y)$.* One can use Markov kernels to represent functions where the output is random; given an input x, the random output is described by the probability measure $M(x)$. Markov kernels also go by the name "communication channel" in information theory since they can be used to model a channel in which random noise affects the output. We will use Markov kernels to model noise in Sections 4.4 and 6.5.

There are several standard methods for constructing Markov kernels. A probability measure can be seen as a simple Markov kernel. Given $\mu \in \mathrm{Prob}(X)$, we can construct a Markov kernel from a singleton $\{\bullet\} \to \mathrm{Prob}(X)$ mapping $\bullet \mapsto \mu$. Measurable functions can be seen as Markov kernels as well. Given a measurable $f : X \to Y$, we can construct the *deterministic Markov kernel induced by f*, denoted δ_f, by setting $\delta_f(x)$ to be the point mass $\delta_{f(x)}$.

We can construct more interesting Markov kernels by the process of *disintegration*. Let $\mu \in \text{Prob}(X \times Y)$ with X-marginal α. Given that X and Y are Polish spaces, there exists a Markov kernel $\beta : X \to \text{Prob}(Y)$ satisfying

$$\mu(A \times B) = \int_A \beta(x)(B) \, \alpha(dx). \tag{1}$$

This characterizes β up to an α-null set. We call such a β the *disintegration of μ along X*. The existence of such a β is shown in, for instance, (Bogachev, 2007, Theorem 10.4.8).

Disintegration can be seen as a measure-theoretic formulation of conditional probability. Indeed, if X and Y are random variables in X and Y respectively with joint law μ, then the disintegration β of μ along X gives the conditional law of Y with respect to X. That is, $\beta(x)(B) = \mathbb{P}(\mathsf{Y} \in B | \mathsf{X} = x)$ for all $x \in X$ and measurable subsets $B \subseteq Y$.

One can also reverse the disintegration process. Just as a joint law of (X, Y) can be recovered from the law of X and the conditional law of Y on X, so too can we recover a coupling in a similar manner. Indeed, Equation (1) shows that the coupling μ can also be recovered from β and α. This is another common method for producing couplings; one selects a probability measure $\alpha \in \text{Prob}(X)$ and a Markov kernel $\beta : X \to \text{Prob}(Y)$ and defines μ as in Equation (1).

Given two Markov kernels $M : X \to \text{Prob}(Y)$ and $N : Y \to \text{Prob}(Z)$, their *Markov kernel composition $M \cdot N : X \to \text{Prob}(Z)$* is given by

$$M \cdot N(x)(C) := \int_Y N(y)(C) \, M(x)(dy)$$

for all measurable $C \subseteq Z$. Given Markov kernels $M : X \to \text{Prob}(Y)$ and $N : W \to \text{Prob}(Z)$, the *independent product of Markov kernels $M \otimes N : X \times W \to \text{Prob}(Y \times Z)$* is given by

$$M \otimes N(x, w) := M(x) \otimes N(w)$$

where the $\otimes$ on the right hand side denotes the product of measures.

2.6 Empirical Measures and Glivenko–Cantelli Classes

If a data set is sampled from an underlying probability measure μ, the sample itself can be modeled as a random measure called an *empirical measure*.

Definition 10 *Given a probability space (Z, μ), the nth empirical measure of μ is the random measure given by*

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}$$

where $Z_1, \ldots, Z_n$ is an i.i.d. sample with law μ.

A practitioner, having access only to a random sample and not the true underlying law μ from which it is drawn, can nevertheless approximate μ with μ_n. One hopes that for large n, μ_n is similar enough to μ for practical purposes. For instance, one would hope that integrating against μ and μ_n yield similar results when n is large. One formalization of this hope taken from statistical learning theory is the notion of a *Glivenko-Cantelli class*.

Definition 11 *Let (X, μ) be a probability space, and let $\mathcal{F}$ be a collection of integrable functions $X \to \mathbb{R}$. Let μ_n denote the empirical measure of μ. We say that $\mathcal{F}$ is a Glivenko-Cantelli class with respect to μ if*

$$\sup_{f \in \mathcal{F}} \left| \int f(x)\mu(dx) - \int f(x)\mu_n(dx) \right| \xrightarrow{a.s.} 0.$$

We say that $\mathcal{F}$ is a Universal Glivenko-Cantelli class if it is a Glivenko-Cantelli class with respect to every $\mu \in \mathrm{Prob}(X)$.

See Van Der Vaart and Wellner (1996) for an overview of Glivenko-Cantelli classes and their relationships to other conditions in the theory of empirical processes.

We will employ Glivenko-Cantelli classes in our analysis of the convergence of empirical problems (Section 5.4), and universal Glivenko-Cantelli classes in our discussion of Rademacher complexity (Section 4.2).

The Glivenko-Cantelli condition may be illuminated by the following sufficient condition taken from the theory of empirical processes involving the popular Vapnik-Chervonenkis (VC) dimension. Given a family of subsets $\mathcal{A}$ of some set Z, the *VC dimension of $\mathcal{A}$* is a number which measures the "flexibility" of the family $\mathcal{A}$. Briefly, the VC dimension of $\mathcal{A}$ is the cardinality of the largest finite set $B \subseteq Z$ such that $\{A \cap B | A \in \mathcal{A}\} = \mathrm{Pow}(B)$, where Pow denotes the power set. If there is no largest such n, we say that $\mathcal{A}$ has *infinite VC dimension*. For an introduction to VC dimension and its place in the theory of statistical learning, see Devroye et al. (1996, Ch 12).

Definition 12 *Let $f : X \to \mathbb{R}$ be any function. The subgraph of f is given by*

$$\mathrm{sg}(f) := \{(x, t) | t < f(x)\} \subseteq X \times \mathbb{R}.$$

Proposition 13 *Let $\mathcal{F}$ be a collection of measurable functions $X \to \mathbb{R}$ that is uniformly bounded below and above. Suppose $\mathcal{F}$ is endowed with a Polish topology, and subsequently the Borel σ-algebra, such that the map $(f, x) \mapsto f(x)$ is jointly measurable in f and x. If the collection of subgraphs $\{\mathrm{sg}(f) | f \in \mathcal{F}\}$ has finite VC dimension, then $\mathcal{F}$ is a universal Glivenko-Cantelli class.*

Proposition 13 follows from standard results in the theory of empirical processes which can be found in, for instance, the reference by Van Der Vaart and Wellner (1996). Specifically, we use a combination of Theorems 2.6.7 and 2.4.3 along with Example 2.3.5 from that source.

3. The Risk Distance

In this section, we introduce the points (problems) and distance function (the Risk distance) which will together define our pseudometric space of supervised learning problems.

3.1 Problems

Before we proceed, we must solidify the informal notion of "supervised learning problem" into a formal definition. Supervised learning problems come in many forms, including

regression, classification, and ranking problems. We would like a definition that is general enough to encompass all of these examples. In all of these problem types, the goal is to examine the relationship between a random variable in an input space X and another in an response space Y. One then aims to select an appropriate function $h : X \to Y$ that predicts the output value given an input. The effectiveness of h is measured using a chosen loss function. (As an aside, even "unsupervised" problems, such as generating samples from a distribution, can have a supervised problem "under-the-hood"; for example, f-GANS rely upon judging suitability of a distribution by use of the f-divergence (Nock et al., 2017), which is intimately related to a Bayes risk (Reid and Williamson, 2011)).

Definition 14 *A* supervised learning problem, *or simply* problem, *is a 5-tuple*

$$P = (X, Y, \eta, \ell, H)$$

where

- *X and Y are Polish topological spaces called the* input space *and* response space *respectively,*[1]

- *η is a Borel probability measure on $X \times Y$ called the* joint law,

- *ℓ is a measurable function $Y \times Y \to \mathbb{R}_{\geq 0}$ called the* loss function,

- *and H is a family of measurable functions $X \to Y$, called the* predictors.[2]

  We require each $h \in H$ to exhibit finite expected loss, meaning

  $$\int_{X \times Y} \ell(h(x), y) \, \eta(dx \times dy) < \infty.$$

Notice that our definition does not explicitly mention the random variables in the input and output spaces. Only the most crucial information about the variables, the joint law η that couples them, is recorded. This strengthens the analogy with metric geometry; a supervised learning problem is analogous to a metric measure space, where the metric corresponds to the loss function and the measure corresponds to the joint law.

The "expected loss" mentioned in Definition 14 is more commonly known as the *risk* of a predictor.

---

1. The Polish condition is put in place to avoid pathologies. In all of the examples we have in mind, X and Y are either open, closed, or convex subsets of $\mathbb{R}^n$, or else discrete sets. Such spaces are Polish. This should also help assuage the reader with set-theoretic concerns. We will speak freely of the "collection of all problems". While this collection is technically too large to be a set, restricting ourselves to the above examples resolves this issue.
2. For greater generality, one could alternatively define a supervised learning problem such that the predictors $h \in H$ are functions into some decision space D which is not necessarily the same as Y. That is, one could define a problem to be a 6-tuple (X, Y, η, ℓ, H, D), where H is a collection of functions $h : X \to D$ and ℓ is a function $\ell : D \times Y \to \mathbb{R}_{\geq 0}$. Definition 14 would then cover the special case $D = Y$. The more general definition would, for instance, encompass class probability estimation problems by taking $D = \text{Prob}(Y)$. Many of the results in this paper would generalize to this more general setting. We leave this investigation to future work.

Definition 15 *Given a problem $P = (X, Y, \eta, \ell, H)$, the* risk *of a predictor $h \in H$ is defined to be*

$$\mathcal{R}_P(h) := \int \ell(h(x), y) \, \eta(dx \times dy).$$

The constrained Bayes risk *of P is defined to be*

$$\mathcal{B}(P) := \inf_{h \in H} \mathcal{R}_P(h) = \inf_{h \in H} \int \ell(h(x), y) \, \eta(dx \times dy).$$

The risk of a predictor is its expected loss on a randomly selected observation. If we measure the performance of a predictor by its risk, then the constrained Bayes risk of a problem represents the optimal possible performance when selecting a predictor from H. The constrained Bayes risk is so called by Williamson and Cranko (2024) since it is a modification of the classical notion of the Bayes risk of a predictor. Note that if the predictor set for a problem P is the set $H_{X,Y}$ of all measurable functions $f : X \to Y$, then $\mathcal{B}(P)$ agrees with the classical Bayes risk, which we will henceforth denote as

$$\mathcal{B}^*(P) := \inf_{h \in H_{X,Y}} \mathcal{R}_P(h).$$

We emphasize that a problem does not include a method for *choosing* a predictor. We think of a problem as something that a practitioner must *solve* by selecting, through some algorithmic procedure, a function from the predictor set that performs well according to some useful functional built on the loss function, such as the risk $\mathcal{R}_P : H \to \mathbb{R}_{\geqslant 0}$. In this regard, in Section 7 we consider a certain compressed representation of $\mathcal{R}_P : H \to \mathbb{R}_{\geqslant 0}$ which permits gaining insight into the complexity associated to this selection task.

Notation 16 *To avoid tediously re-writing these 5-tuples, we establish some persistent notation. We assume, unless otherwise stated, that a problem with the name P has components (X, Y, η, ℓ, H), a problem named P' has components $(X', Y', \eta', \ell', H')$, and a problem named P'' has components $(X'', Y'', \eta'', \ell'', H'')$. Similarly, for any integer $i \geqslant 0$, the problem P_i has components $(X_i, Y_i, \eta_i, \ell_i, H_i)$.*

Notation 17 *Given any problem $P = (X, Y, \eta, \ell, H)$, we use the shorthand ℓ_h for the function $\ell_h : X \times Y \to \mathbb{R}_{\geqslant 0}$ given by $\ell_h(x, y) := \ell(h(x), y)$.*

Example 18 *Consider the problem of predicting whether a college student will successfully graduate given their high school GPA and ACT scores. We can model this problem as follows. We represent each student with a point in $\mathbb{R}^2 \times \{0, 1\}$, where the coordinates in $\mathbb{R}^2$ are determined by the student's GPA and ACT score, and their graduation status is represented with a 0 for failure and 1 for success. We think of students as being sampled from an unknown probability measure η on $\mathbb{R}^2 \times \{0, 1\}$. Our aim is then to select a predictor $\mathbb{R}^2 \to \{0, 1\}$. We might restrict ourselves to predictors with a linear decision boundary:*

$$h_{a,b,c}(x_1, x_2) := \begin{cases} 1 & ax_1 + bx_2 \geqslant c \\ 0 & ax_1 + bx_2 < c \end{cases}.$$

We may also decide that all incorrect predictions carry the same penalty. Any incorrect guess incurs a loss of 1, while a correct guess has loss 0.

We have now constructed a problem

$$P := (\mathbb{R}^2, \{0, 1\}, \eta, \ell, H),$$

where $\ell(y, y') = 1_{y \neq y'}$, and $H = \{h_{a,b,c} | a, b, c \in \mathbb{R}\}$.

The risk of an arbitrary predictor $h_{a,b,c}$ is then

$$\mathcal{R}_P(h_{a,b,c}) = \int \ell(h_{a,b,c}(x), y) \, \eta(dx \times dy)$$

$$= \int_{\mathbb{R}^2 \times \{0\}} h_{a,b,c}(x) \, \eta(dx \times dy) + \int_{\mathbb{R}^2 \times \{1\}} (1 - h_{a,b,c}(x)) \, \eta(dx \times dy)$$

$$= \eta\big(\{x \in \mathbb{R}^2 | h_{a,b,c}(x) = 1\} \times \{0\}\big) + \eta\big(\{x \in \mathbb{R}^2 | h_{a,b,c}(x) = 0\} \times \{1\}\big).$$

That is, the risk of $h_{a,b,c}$ is its misclassification probability. The constrained Bayes risk $\mathcal{B}(P)$ is then the infimal misclassification probability across all predictors with linear decision boundaries. Under additional normality conditions, the predictor achieving the risk $\mathcal{B}(P)$ can be found with the Fisher linear discriminant (Fisher, 1936).

Example 19 *Suppose we are faced with a linear regression problem. That is, we have a random vector X in $\mathbb{R}^n$ and a random number Y in $\mathbb{R}$ and are tasked with finding an affine linear functional $\mathbb{R}^n \to \mathbb{R}$ that most closely resembles their joint law. We can formalize this problem as follows. We let $\eta \in \mathrm{Prob}(\mathbb{R}^n \times \mathbb{R})$ be the unknown joint law. We may select the squared difference $\ell(y, y') := (y - y')^2$ as our loss function. The resulting problem is then*

$$P := (\mathbb{R}^n, \mathbb{R}, \eta, \ell, H)$$

where H is the collection of affine linear functions $\mathbb{R}^n \to \mathbb{R}$.

The risk of an arbitrary predictor $h(x) = a^T x + b$ is the mean squared error of h:

$$\mathcal{R}_P(h) = \int \ell(h(x), y) \, \eta(dx \times dy) = \int (a^T x + b - y)^2 \, \eta(dx \times dy).$$

The constrained Bayes risk $\mathcal{B}(P)$ is then the infimal mean squared error across all linear functionals $h \in H$.

Example 20 *By making trivial choices for each component, one can define the* one-point problems *to be*

$$P_\bullet(c) := (\{\bullet\}, \{\bullet\}, \delta_{(\bullet, \bullet)}, c, \{\mathrm{Id}_\bullet\})$$

for any $c \geq 0$. We will most often set $c = 0$ and use the notation $P_\bullet := P_\bullet(0)$. The constrained Bayes risk $\mathcal{B}(P_\bullet(c))$ is easily seen to be c, since the loss function of $P_\bullet(c)$ is identically c.

3.2 Comparison with Blackwell's Statistical Experiments

Our definition of a supervised learning problem may seem superficially similar to the established notion of a statistical experiment which is prevalent in mathematical statistics. Since there already exist established quantitative methods for comparing statistical experiments such as Le Cam's deficiency distance (Le Cam, 1964; Mariucci, 2016), let us take a moment to contrast statistical experiments and supervised learning problems. Originally defined by Blackwell (Blackwell, 1951), a statistical experiment consists of

- a measurable space X,

- a set of probability measures $\{P_\theta, \theta \in \Theta\}$ indexed by some set Θ, and

- a distinguished $\theta \in \Theta$, thought to be the "true" parameter, or the parameter "chosen by nature".

A practitioner is then tasked with estimating the true value of θ and hence the underlying probability measure.

Supervised learning can be placed into Blackwell's framework (c.f. Williamson and Cranko, 2024) by taking X to be the input space of the problem, taking Θ to be the response space Y, and letting the map $y \mapsto P_y$ be the disintegration of the joint law $\eta \in \mathrm{Prob}(X \times Y)$ along Y.[3] When predicting the label for a particular observation x, the "true" value y_0 represents the unknown true response, and we assume the observed input x was drawn from the probability measure P_{y_0}. One's goal, of course, is to then estimate y_0 using an observed $x \in X$. A process for doing so defines a predictor $X \to Y$. Hence the superivised learning problem is encoded as a statistical experiment $E := (X, \{P_y | y \in Y\}, y_0)$.

In contrast, our supervised learning problems have three additional pieces of information:

(1) the Y-marginal of η,

(2) a loss function ℓ on the response space Y, and

(3) an explicit family H of functions $X \to Y$.

Indeed, using the Y-marginal of η and the disintegration $y \mapsto P_y$, we can recover the joint probability measure $\eta \in \mathrm{Prob}(X \times Y)$ via reverse-disintegration (see Section 2.5 for details). Hence from the statistical experiment and the listed additional information, one can recover all five components of the supervised learning problem (X, Y, η, ℓ, H).

3.3 Motivating Properties

Our goal is to quantitatively compare problems by developing a notion of distance on the collection of all problems. This notion will take the form of a pseudometric d_R which we will call the *Risk distance*. To design such a distance, we might begin by deciding on some desirable properties. When we introduce d_R, we will see that it is characterized as the largest pseudometric satisfying these conditions.

The first property that one needs to decide when designing a pseudometric is what its zero set should be. In other words one needs to postulate a notion of *isomorphism* that

---

3. In the context of classification, the measure P_y is called the *class-conditional distribution*.

we expect the notion of distance to be compatible with. We start by introducing one such notion.

Definition 21 *Two problems P and P' are said to be* strongly isomorphic *if there exist measurable bijections*

$$f_1 : X' \to X \qquad f_2 : Y' \to Y$$

and a bijection between predictor sets

$$\varphi : H' \to H$$

such that the following conditions are satisfied:

1. *$(f_1 \times f_2)_\sharp \eta' = \eta$,*

2. *For every $h' \in H'$, the diagram*

commutes. That is, $\ell'_{h'} = \ell_{\varphi(h')} \circ (f_1 \times f_2)$.

Here the notation $f_1 \times f_2$ denotes the product map: $(f_1 \times f_2)(x', y') := (f_1(x'), f_2(y'))$. (We choose the notation $f_1 \times f_2$ over the alternative (f_1, f_2) because the former emphasizes that each component of the output depends only on the corresponding component of the input.) In other words, a strong isomorhism consists of bijections between the components of a problem that preserve the structure of the problem, and in particular preserve the performance of every predictor; if $f_1(x') = x$, $f_2(y') = y$ and $\phi(h') = h$, then $\ell'_{h'}(x', y') = \ell_h(x, y)$. Despite being natural, the notion of strong isomorphism is rather rigid. For instance, Condition 2 requires that an equality involving the predictors and loss hold for every point in $X' \times Y'$, even those outside of the support of η'. Such points could reasonably be considered irrelevant to the supervised learning problem, since an observation will appear there with probability 0. In order to motivate a more flexible notion of isomorphism which characterizes the pairs of problems that should be distance zero apart, we collect potential modifications to problems that one might reasonably consider immaterial.

Example 22 *Consider the linear regression problem $P = (\mathbb{R}^n, \mathbb{R}, \eta, \ell, H)$ from Example 19. We chose H to be the set of all affine linear functionals $\mathbb{R}^n \to \mathbb{R}$. As is often done, we could alternatively formulate the linear regression problem as follows. We place our data in $\mathbb{R}^{n+1}$ by setting the first coordinate of every observation to a constant 1. Instead of an affine linear map, our task is now to select a (non-affine) linear map $\mathbb{R}^{n+1} \to \mathbb{R}$. Formally, we construct a problem*

$$P' := (\mathbb{R}^{n+1}, \mathbb{R}, i_\sharp \eta, \ell, H')$$

where

21

- $i : \mathbb{R}^n \to \mathbb{R}^{n+1}$ is given by $i(x_1, \ldots, x_n) := (1, x_1, \ldots, x_n)$, and

- H' is the set of all (non-affine) linear functionals $\mathbb{R}^{n+1} \to \mathbb{R}$.

The problems P and P' represent two common formulations of linear regression, the latter stated in terms of linear functionals, the former in terms of affine linear functionals. While P and P' are not strongly isomorphic, the formulations they represent are considered equivalent. Whatever our notion of distance between problems, it is desirable that P and P' be distance zero apart in order to reflect this equivalence.

Example 23 *Consider the problem*

$$P := (\mathbb{R}, \{a, b\}, \eta, \ell, H),$$

where

- η is any probability measure on $\mathbb{R} \times \{a, b\}$.

- ℓ is the 0-1 loss (also called the discrete metric) on $\{a, b\}$. Namely, $\ell(a, a) = \ell(b, b) = 0$ and $\ell(a, b) = \ell(b, a) = 1$.

- H consists of functions $\mathbb{R} \to \{a, b\}$ with finitely many discontinuities.

This P represents a classification problem; a predictor $h \in H$ assigns to each input either the label a or the label b. Now construct a second problem P' by adding to P a third label, say, c. We design the remaining components of P' so as to make the labels b and c functionally identical. More precisely, we define

$$P' := (\mathbb{R}, \{a, b, c\}, \eta', \ell', H')$$

where

- η' is any measure such that η is the pushforward of η' under the map $\{a, b, c\} \to \{a, b\}$ sending $a \mapsto a$ and $b, c \mapsto b$.

- ℓ' extends ℓ by setting
$$\ell'(b, c) = \ell'(c, b) = \ell'(c, c) = 0$$
$$\ell'(a, c) = \ell'(c, a) = 1.$$
  That is, ℓ' considers b and c to be indistinguishable.

- H' consists of all functions $\mathbb{R} \to \{a, b, c\}$ with finitely many discontinuities.

The problems P and P' are both classification problems. By design, P' is P with an extraneous label. If we solve P by choosing a function $h \in H$, the same h will provide a solution to P' with identical performance. Conversely, if we solve P' by selecting some $h' \in H'$, we can produce an identically-performing solution for P by composing h' with the map sending $a \mapsto a$ and $b, c \mapsto b$. While P and P' are not strongly isomorphic, they differ only superficially, and a meaningful distance should set P and P' to be distance zero apart.

Examples 22 and 23 demonstrate that any of the five components of a problem can change without modifying the problem's essential nature. This motivates the following definition which arises as a relaxation of Definition 21.

Definition 24 *Given two problems P and P', we say that P' is a* simulation *of P (or that P' simulates P), and denote it by $P' \rightsquigarrow P$, if there exist measurable functions*

$$f_1 : X' \to X \qquad f_2 : Y' \to Y$$

such that the following are satisfied.

1. $(f_1 \times f_2)_\sharp \eta' = \eta$,

2. *For every $h \in H$, there exists an $h' \in H'$ such that the diagram*

commutes (that is, $\ell'_{h'} = \ell_h \circ (f_1 \times f_2)$) η'-almost everywhere. Similarly, for every $h' \in H'$, there exists an $h \in H$ such that the above diagram commutes η'-almost everywhere.

Note that if P and P' are strongly isomorphic then P and P' are simulations of each other. It is, however, clear that the notion of simulation is weaker than strong isomorphism. Indeed, the former does not require the maps f_1 and f_2 to be bijections, nor does it require Condition 2 to be satisfied through a bijection between H and H'. Furthermore, Condition 2 encodes an additional, more subtle, relaxation of the second condition of strong isomorphism: whereas the latter requires the diagram to commute everywhere on $X' \times Y'$, the former only requires this to take place η'-almost everywhere within $X' \times Y'$. In Example 22, the problem P' is a simulation of P via the projection map $\mathbb{R}^{n+1} \to \mathbb{R}^n$ that drops the first coordinate, and the identity map on the output space $\mathbb{R}$. Likewise, in Example 23, P' is a simulation of P via the identity map on the input space $\mathbb{R}$ and the map on the output space sending $a \mapsto a$ and $b, c \mapsto b$.

The term "simulation" is meant to suggest that when a problem P' is a simulation of P, then P' is, in a sense, richer than P. The terminology is inspired by similar concepts appearing in the study of automata and Markov chains (Larsen and Skou, 1989; Milner, 1980, 1989).

Motivated by Examples 22 and 23, we posit that if P' is a simulation of P, then it is reasonable to consider P and P' to be functionally identical. As these examples demonstrate, if P' simulates P, then P' can be solved "in terms of P" using the maps f_1 and f_2 which witness the simulation. That is, one could take an i.i.d. sample $\{(x'_i, y'_i)\}_{i=1}^n$ drawn from η' and apply the maps f_1 and f_2 to the x_i's and y_i's respectively to get a sample $\{(x_i, y_i)\}_{i=1}^n$ in $X \times Y$. The pushforward condition in the definition of a simulation guarantees that

23

the transformed sample will be an i.i.d. sample drawn from η. One could then solve P by using this data to select an appropriate predictor $h \in H$, then solve P' with a corresponding $h' \in H'$. Point 2 in the simulation definition implies that $\ell'(h'(x_i'), y_i') = \ell(h(x_i), y_i)$. In other words, we are guaranteed that h and h' have identical performance on corresponding observations. For instance, both problems have the same constrained Bayes risk. Indeed, if P' is a simulation of P via the maps $f_1 : X \to X'$ and $f_2 : Y \to Y'$, then

$$
\begin{aligned}
\mathcal{B}(P') &= \inf_{h' \in H'} \int \ell'_{h'}(x', y') \, \eta'(dx' \times dy') \\
&= \inf_{h \in H} \int \ell_h(f_1(x'), f_2(y')) \, \eta'(dx' \times dy') \\
&= \inf_{h \in H} \int \ell_h(x, y) \, ((f_1 \times f_2)_\sharp \eta')(dx \times dy) \\
&= \inf_{h \in H} \int \ell_h(x, y) \, \eta(dx \times dy) = \mathcal{B}(P).
\end{aligned}
$$

This sort of relationship between problems that allows one to solve one problem in terms of another is reminiscent of John Langford's "Machine Learning Reductions" research program (Langford, 2009).

If P' is a simulation of P, we would like our distance d_R to consider them identical.

> **Condition 1:** If P' is a simulation of P, then $d_\mathrm{R}(P, P') = 0$.

Symmetrizing the simulation relationship produces a more general relationship which we call *weak isomorphism*, modeled after *weak isomorphism of networks*, defined by Chowdhury and Mémoli (2019); Chowdhury and Mémoli (2023).

Definition 25 *Define two problems P and P' to be* weakly isomorphic *if there exists a third problem $\widetilde{P}$ that is a simulation of both P and P'.*

In other words, weak isomorphism between problems P and P' arises from the existence of a simultaneous simulation of both problems, as illustrated via the diagram[4]

$$
\begin{array}{ccc}
& \widetilde{P} & \\
& \swarrow \quad \searrow & \\
P & & P'.
\end{array}
$$
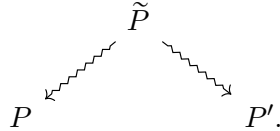
It is clear that strong isomorphism implies weak isomorphism. As we will see below (Theorem 27 and Proposition 30), under fairly general conditions, the Risk distance $d_\mathrm{R}(P, P')$ will vanish if and only if P and P' are weakly isomorphic.

---

4. Simulations, weak isomorphisms and their connections to the Risk distance suggest the existence of a "category of problems" in which simulations are morphisms, weak isomorphisms are spans, and from which the Risk distance can be derived. We leave the exploration of such a categorical perspective to future work.

The second desirable condition for d_R is simpler to state. Let P be a problem, and suppose that P' is a problem with all the same components, except possibly for the loss function ℓ'. If there exists $\alpha \geqslant 0$ such that, for all $h \in H$, we have

$$\int_{X \times Y} \left| \ell(h(x), y) - \ell'(h(x), y) \right| \eta(dx \times dy) \leqslant \alpha,$$

then we demand $d_\mathrm{R}(P, P') \leqslant \alpha$.

Condition 2: If
$$P = (X, Y, \eta, \ell, H)$$
$$P' = (X, Y, \eta, \ell', H),$$

then
$$d_\mathrm{R}(P, P') \leqslant \sup_{h \in H} \int_{X \times Y} \left| \ell(h(x), y) - \ell'(h(x), y) \right| \eta(dx \times dy).$$

3.4 The Risk Distance

The following is a pseudometric that satisfies Conditions 1 and 2.

Definition 26 *Let P and P' be problems. For any coupling $\gamma \in \Pi(\eta, \eta')$ and correspondence $R \in \mathcal{C}(H, H')$, define the* risk distortion *of γ and R to be*

$$\mathrm{dis}_{P, P'}(R, \gamma) := \sup_{(h, h') \in R} \int \left| \ell(h(x), y) - \ell'(h'(x'), y') \right| \gamma(dx \times dy \times dx' \times dy').$$

The Risk distance *between P and P' is then given by*

$$d_\mathrm{R}(P, P') := \inf_{\substack{\gamma \in \Pi(\eta, \eta') \\ R \in \mathcal{C}(H, H')}} \mathrm{dis}_{P, P'}(R, \gamma).$$

The risk distortion (and subsequently the Risk distance) is so named because of its similarity to the risk of a predictor. If one believes that the most salient numerical descriptor of a predictor is the loss that it incurs, then it is natural to describe a predictor by its loss on an average observation. It is similarly natural to compare predictors by comparing their respective losses for an average observation.

The idea behind the Risk distance is that we should consider two problems P and P' to be similar if we can find a correspondence between the predictor sets and a coupling between the underlying spaces such that pairs of corresponding predictors *incur similar losses* on corresponding observations.

Notably, the Risk distortion and Risk distance are not invariant to scaling and shifting of the loss functions. While we often think of loss substitutions such as $\ell(y, y') \mapsto a\ell(y, y') + b$ for $a, b \in \mathbb{R}$, $a > 0$ as immaterial modifications in supervised learning, such changes do, in some sense, affect the problem. For instance, such a substitution will change the Bayes risk of the problem, so any notion of distance which controls the Bayes Risk (see Theorem 37 below) cannot be invariant to such modifications. In order to use the Risk distance on specific

problems, one must select loss functions which are meaningful in their specific context. For instance, one may elect to "normalize" their loss functions by subtracting the Bayes risk $\mathcal{B}(P)$ or the minimal possible loss $\min_{y,y' \in Y} \ell(y, y')$, or by scaling the loss to have meaningful units such as bits or nats. Our formalism supports such decisions.

By design, the Risk distance is not a proper metric, since Condition 1 implies that distinct problems could be distance zero apart. This is desirable behavior. Reformulating a problem in an equivalent way as in Example 22 or adding extraneous labels as in Example 23 should not move a problem at all under the Risk distance since such modifications do not meaningfully change the problem.

Theorem 27 *The function d_{R} is a pseudometric on the collection $\mathcal{P}$ of all problems. Furthermore, $d_{\mathrm{R}}(P, P') = 0$ whenever P and P' are weakly isomorphic.*

That d_{R} is pseudometric entails that d_{R} vanishes on the diagonal, that is $d_{\mathrm{R}}(P, P) = 0$ for all problems P, satisfies symmetry, that is $d_{\mathrm{R}}(P, P') = d_{\mathrm{R}}(P', P)$ for all problems P and P', and satisfies the triangle inequality: $d_{\mathrm{R}}(P, P'') \leqslant d_{\mathrm{R}}(P, P') + d_{\mathrm{R}}(P', P'')$ for all problems P, P' and P''.

That d_{R} satisfies the triangle inequality is nontrivial to prove, and the proof is relegated to Appendix A. Not only does d_{R} satisfy Conditions 1 and 2, it is characterized by them in the sense that it is *uniquely determined by these conditions.*

Theorem 28 *The Risk distance satisfies Conditions 1 and 2. That is,*

1. *If P' is a simulation of P, then $d_{\mathrm{R}}(P, P') = 0$.*

2. *For any problems $P = (X, Y, \eta, \ell, H)$ and $P' = (X, Y, \eta, \ell', H)$ which differ only in their loss functions, we have*

$$d_{\mathrm{R}}(P, P') \leqslant \sup_{h \in H} \int_{X \times Y} \left| \ell(h(x), y) - \ell'(h(x), y) \right| \eta(dx \times dy).$$

Furthermore, the Risk distance is the largest pseudometric simultaneously satisfying Condition 1 and Condition 2.

Note that the claim from Theorem 27 that $d_{\mathrm{R}}(P, P') = 0$ whenever P and P' are weakly isomorphic can now be obtained from Theorem 28 as follows. By Definition 25, there exists a problem P'' which is is joint simulation of both P and P'. The claim can now be obtained through the facts that d_{R} satisfies both the triangle inequality and Condition 1 so that

$$d_{\mathrm{R}}(P, P') \leqslant d_{\mathrm{R}}(P, P'') + d_{\mathrm{R}}(P', P'') = 0.$$

The statement of Theorem 28 implicitly asserts uniqueness of the largest such pseudometric. Indeed, if d_1 and d_2 were two different pseudometrics satisfying Conditions 1 and 2, their pointwise maximum $d' := \max(d_1, d_2)$ would also satisfy Conditions 1 and 2, would be a pseudometric, but would be larger than d_1 and d_2.[5] The proof of this theorem can also be found in Appendix A.

While Theorem 27 states that weakly isomorphic problems are distance zero apart, we can say more; weak isomorphism is equivalent to the existence of a coupling and correspondence between the problems with zero risk distortion.

---

5. This would follow from the assumption that there exist $P, P' \in \mathcal{P}$ such that $d_1(P, P') \neq d_2(P, P')$.

Proposition 29 *Let P and P' be problems. Then $P \cong P'$ if and only if there exist a correspondence $R \in \mathcal{C}(H, H')$ and a coupling $\gamma \in \Pi(\eta, \eta')$ such that*

$$\text{dis}_{P,P'}(R, \gamma) = 0.$$

The proof is in Appendix A. Note that if $\text{dis}_{P,P'}(R, \gamma) = 0$ then $d_{\text{R}}(P, P') = 0$ by definition of the Risk distance, but the converse does not necessarily hold; it is possible that the infimal risk distortion is not achieved by any particular coupling and correspondence. Hence weak isomorphism between problems P and P' is generally stronger than saying $d_{\text{R}}(P, P') = 0$. We will discuss the existence of such "optimal" couplings and correspondences in Section 5.2, where we will prove that, under certain conditions, weak isomorphism is equivalent to identification under the Risk distance. While the statement (Corollary 61) is in terms of definitions we have not yet established, it is equivalent to the following:

Proposition 30 *Let $P = (X, Y, \eta, \ell, H)$ and $P' = (X', Y', \eta', \ell', H')$ be problems. Suppose that for all $h \in H$, $h' \in H'$, the functions ℓ_h and ℓ'_h are continuous almost everywhere (with respect to η and η' respectively), and the sets of functions*

$$\{\ell_h | h \in H\} \subseteq L^1(\eta) \quad and \quad \{\ell'_{h'} | h' \in H'\} \subseteq L^1(\eta')$$

are compact under their respective L^1 norms. Then $d_{\text{R}}(P, P') = 0$ if and only if $P \cong P'$.

The continuity condition is satisfied by a typical regression problem, such as Example 19, since the predictors and loss for such a problem tend to be continuous. A classification problem can satisfy the continuity condition as well under some assumptions on η. In particular, we need only assume that a random observation falls on a decision boundary with probability zero. We will discuss the continuity condition in more detail in Section 5.2.

The compactness condition is implied if one assumes that a problem's predictor set H is parameterized by some compact parameter space $\Theta \to H$ and that the composition $\Theta \to H \to L^1(\eta)$ is continuous. Consider, for example, the problem $P = (\mathbb{R}, \{0, 1\}, \eta, \ell, H)$ where ℓ is the 0-1 loss and H is the set of indicators $1_{(a,\infty)}$ for a in the compact interval $[-\infty, \infty]$. That is, P is a binary classification problem for which the predictors are given by hard cutoff values. Under modest assumptions on η (namely that its first marginal is absolutely continuous with respect to the Lebesgue measure), the composition $[-\infty, \infty] \to H \to L^1(\eta)$ is continuous, and hence P satisfies the compactness condition. Under the same assumptions, we also have that each ℓ_h is η-almost everywhere continuous. Hence P satisfies both conditions of Proposition 30.

While classification problems are more likely to satisfy the compactness condition of Proposition 30, regression problems can be made to do so as well. Consider the linear regression problem in Example 19. The predictor set for that problem is parameterized by the non-compact space $\mathbb{R}^{n+1}$. However, it can be made compact with regularization. Indeed, regularized regression methods like ridge regression and lasso can be seen as restrictions of the parameter space to a compact region $A \subset \mathbb{R}^{n+1}$ (James et al., 2021, Sec 6.2). Linear regression problems with such restrictions will satisfy both the continuity and compactness conditions.

Adding a constant $\alpha > 0$ to the loss function of a problem moves it a distance of α under the Risk distance. That is, if P is a problem with loss function ℓ, and P' is the same

problem but with loss function $\ell + \alpha$, then $d_\mathrm{R}(P, P') = \alpha$. This may appear to be undesirable behavior, since the losses ℓ and $\ell + \alpha$ are practically identical from an optimization point of view. However, adding the constant α to the loss function also increases the constrained Bayes risk of the problem by α. From this perspective, it is reasonable to expect the distance between the two problems to be α.

Example 31 *Let P be an arbitrary problem and $P_\bullet = P_\bullet(0)$ be the one-point problem of Example 20. There is only one correspondence in $\mathcal{C}(H, \{\mathrm{Id}_\bullet\})$ and one coupling in $\Pi(\eta, \delta_{(\bullet, \bullet)})$, so one can compute*

$$d_\mathrm{R}(P, P_\bullet) = \mathrm{dis}_{P, P_\bullet}\left(H \times \{\mathrm{Id}_{\{\bullet\}}\}, \eta \otimes \delta_{(\bullet, \bullet)}\right)$$
$$= \sup_{h \in H} \int \ell(h(x), y)\, \eta(dx \times dy).$$

Note the similarity with the constrained Bayes risk $\mathcal{B}(P)$. While $\mathcal{B}(P)$ is equal to the risk of an optimal *predictor, the distance $d_\mathrm{R}(P, P_\bullet)$ is equal to the risk of a* pessimal *predictor.*

If we additionally assume that the loss ℓ is bounded above by a constant ℓ_max, one can similarly compute the distance from P to the one-point problem $P_\bullet(\ell_\mathrm{max})$ to be

$$d_\mathrm{R}(P, P_\bullet(\ell_\mathrm{max})) = \sup_{h \in H} \int \left(\ell_\mathrm{max} - \ell(h(x), y)\right) \eta(dx \times dy)$$
$$= \ell_\mathrm{max} - \inf_{h \in H} \int \ell(h(x), y)\, \eta(dx \times dy)$$
$$= \ell_\mathrm{max} - \mathcal{B}(P).$$

Hence $\mathcal{B}(P) = \ell_\mathrm{max} - d_\mathrm{R}(P, P(\ell_\mathrm{max}))$, showing that the constrained Bayes risk can be written in terms of the Risk distance if the loss is bounded.

Remark 32 *The Risk distance is not designed to be explicitly computed for arbitrary pairs of problems. Much like the Gromov-Hausdorff distance, its usefulness lies in the stability results which it facilitates. However, for completeness, we briefly remark on computational considerations. Given that the computation of the structurally similar but seemingly simpler Gromov-Hausdorff distance is known to be NP-hard (Agarwal et al., 2015; Schmiedl, 2017; Mémoli et al., 2023), we expect that estimation of the exact Risk distance for problems with finite input and output spaces is also NP-hard in general. In this direction, we will soon prove, as Corollary 34, that deciding whether $d_\mathrm{R}(P, P') = 0$ for arbitrary problems P and P' with finite input and output spaces is at least as hard as graph isomorphism.*

In later sections, we will also explore two variants of the Risk distance that are constructed by tweaking the treatment of the correspondences $\mathcal{C}(H, H')$ in Definition 26. First, by endowing the predictor sets H and H' with probability measures λ and λ', one can replace $\mathcal{C}(H, H')$ with $\Pi(\lambda, \lambda')$ and the supremum over R with an L^p-style integral over a coupling. This gives rise to the L^p-*Risk distance* introduced in Section 6. Alternatively, by restricting the set of possible correspondences $\mathcal{C}(H, H')$ to the subset $\mathcal{C}_\mathrm{con}(H, H')$ consisting only of those satisfying a certain topological condition, we arrive at the *Connected Risk distance* of Section 7.

3.5 Connections to the Gromov-Wasserstein Distance

The Hausdorff, Gromov-Hausdorff, Wasserstein, and Gromov-Wasserstein distances all follow a common paradigm. One first considers the space of all "alignments" between the objects in question, whether "alignment" means a correspondence between sets or a coupling between measures. One then settles on a way to numerically score the distortion of a correspondence or coupling, whether through a supremum or integral. The distance between the objects is then the infimal possible score. The Risk distance follows the same paradigm. An "alignment" of two problems consists of a correspondence between their predictor sets and a coupling between their joint laws. We use the loss function to numerically score the coupling and correspondence, then infimize over all choices.

Indeed, if we think of a loss function as a notion of distance, then the definition of d_{R} most closely resembles that of the Gromov-Wasserstein distance. The only extra component is a supremum over the predictor correspondence. To make the comparison more explicit, suppose (X, d_X, μ_X) and $(X', d_{X'}, \mu_{X'})$ are metric measure spaces. We can construct the following problems:

$$P := (X, X, \mu_X \otimes \mu_X, d_X, \{\mathrm{Id}_X\})$$
$$P' := (X', X', \mu_{X'} \otimes \mu_{X'}, d_{X'}, \{\mathrm{Id}_{X'}\}).$$

We then write

$$d_{\mathrm{R}}(P, P') = \inf_{\gamma \in \Pi(\mu_X \otimes \mu_X, \mu_{X'} \otimes \mu_{X'})} \int \left| d_X(x, y) - d_{X'}(x', y') \right| \gamma(dx \times dy \times dx' \times dy').$$

Here $\otimes$ represents the product of measures. This is exactly twice the Gromov-Wasserstein distance (Definition 8) except for one difference: the measure γ. The infimum runs over all of $\Pi(\mu_X \otimes \mu_X, \mu_{X'} \otimes \mu_{X'})$, while in the true Gromov-Wasserstein distance, γ must take the specific form $\gamma = \sigma \otimes \sigma$ for some $\sigma \in \Pi(\mu_X, \mu_{X'})$. Indeed, the expression above (divided by 2) is identified as the "second lower bound" of the Gromov-Wasserstein distance by Mémoli (2007).

Another connection between the Gromov-Wasserstein and Risk distances can be drawn as well. Specifically, one can encode compact metric measure spaces X as supervised learning problems $i(X)$ in a natural way which preserves the identifications induced by the two pseudometrics.

Proposition 33 *Given a compact metric measure space (X, d_X, μ_X), define a problem*

$$i(X) := (X, X, (\Delta_X)_\sharp \mu_X, d_X, H_X),$$

where

- *$\Delta_X : X \to X \times X$ is the diagonal map, and*

- *H_X is the set of all constant functions $X \to X$.*

If X and X' are metric measure spaces, then

$$d_{\mathrm{GW},1}(X, X') = 0 \iff d_{\mathrm{R}}(i(X), i(X')) = 0.$$

The proof can be found in Appendix A.1.1 where we also discuss a connection with a variant of the Gromov-Wasserstein distance by Hang et al. (2019, Section 4.1) that is recovered as $d_{\mathrm{R}}(i(X), i(Y))$. It is a folklore result in optimal transport that deciding whether $d_{\mathrm{GW},1}(X, X') = 0$ for finite metric measure spaces X and X' is at least as hard as the graph isomorphism problem. This is not difficult to prove; let $G = (V, E)$ and $G' = (V', E')$ be finite simple graphs. We can encode G as a metric measure space $V = (V, d_V, \mu_V)$, where d_V is the shortest path metric and μ_V is the uniform probability measure. Similarly construct $V' = (V', d_{V'}, \mu_{V'})$. Then $d_{\mathrm{GW},1}(V, V') = 0$ if and only if G and G' are isomorphic. Indeed, if G and G' are isomorphic then the isomorphism induces the requisite coupling between μ_V and $\mu_{V'}$. Conversely, if $d_{\mathrm{GW},1}(V, V') = 0$, then V and V' are isomorphic as metric measure spaces, and in particular are isometric as metric spaces. An isometry between V and V' that preserves the shortest path metric is in fact a graph isomorphism.

Hence Proposition 33 yields the following corollary.

Corollary 34 *The problem of deciding whether $d_{\mathrm{R}}(P, P') = 0$ for two problems with finite input and response spaces is at least as hard as graph isomorphism.*

It is desirable to draw a stronger connection between the Risk distance and Gromov-Wasserstein distance by finding a function i such that $d_{\mathrm{GW},1}(X, X') = d_{\mathrm{R}}(i(X), i(X'))$. We will nearly accomplish this goal in Section 6.2 by linking the L^p-Risk distance with a slightly modified Gromov-Wasserstein distance which is known in the literature.

4. Stability

One major use of the Risk distance is to facilitate stability results. By "stability result," we mean a result of one of the two following forms. Suppose we modify a problem P via some process to produce another problem $f(P)$. We say that d_{R} *is stable under f* if every problem P satisfies an inequality of the form

$$d_{\mathrm{R}}(P, f(P)) \leqslant J(f)$$

for some fixed functional J which measures how dissimilar f is from the identity function $\mathrm{id} : \mathcal{P} \to \mathcal{P}$ and satisfies $J(\mathrm{id}) = 0$. We would say that "the Risk distance is stable under f". Stability results of this form assure us that small modifications to a problem do not move that problem far under d_{R}. For instance, Condition 2 from Section 3.3 is a stability result. In particular, Condition 2 implies that d_{R} is stable under changes to the loss function.

Alternatively, a "stability result" could mean the following. Suppose that, from any problem P, we can compute a descriptor $g(P)$ which lies in a metric space (Z, d_Z). We say that g is *stable under d_{R}* if there is some non-negative constant C satisfying

$$d_Z(g(P), g(P')) \leqslant C d_{\mathrm{R}}(P, P')$$

for all problems P and P'. Stability results of this form act as quantitative guarantees that problems "close" under d_{R} have similar descriptors.

4.1 Stability of the Constrained Bayes Risk and Loss Profile

From a machine learning perspective, the most important feature of a predictor is the loss that it incurs. This information can be captured by a predictor's *loss profile*, a probability measure

on the real line describing the loss that a predictor incurs on a random observation. If P is a problem and $h \in H$ is a predictor, the *loss profile of h* is the pushforward $(\ell_h)_\sharp \eta \in \text{Prob}(\mathbb{R})$.

Note that the loss profile of h is the distribution of the random variable $\ell(h(\mathsf{X}), \mathsf{Y})$, where X and Y are distributed according to η. This is closely related to the random variable $\ell(h(\mathsf{X}), \mathsf{Y}) - \mathcal{B}(P)$, called the *excess loss random variable* in the literature, differing only by an additive constant. The distribution of the excess loss is related to convergence rates of empirical risk minimization (Van Erven et al., 2015; Grünwald and Mehta, 2020).

Letting h range over all of H, we get an informative description of a problem.

Definition 35 *Given a problem P, the* loss profile set of P *is the set*

$$L(P) := \{(\ell_h)_\sharp \eta | h \in H\}.$$

Note that the loss profile $L(P)$ of P contains enough information to recover the constrained Bayes risk. This follows immediately from the definition (we omit the proof).

Proposition 36 *Given a problem P, one has*

$$\mathcal{B}(P) = \inf_{\alpha \in L(P)} \text{mean}(\alpha).$$

Since probability measures on the real line, such as loss profiles, can be compared using the Wasserstein metric $d_{\mathrm{W},1}$, *collections* of probability measures on $\mathbb{R}$ form subsets of the metric space $(\text{Prob}(\mathbb{R}), d_{\mathrm{W},1})$, and therefore can be compared using the Hausdorff distance d_{H} with underlying metric $d_{\mathrm{W},1}$. That is, if $A, B \subseteq \text{Prob}(\mathbb{R})$, the Hausdorff distance between A and B is given by

$$d_{\mathrm{H}}^{d_{\mathrm{W},1}}(A, B) = \inf_{R \in \mathcal{C}(A,B)} \sup_{(\mu,\nu) \in R} d_{\mathrm{W},1}(\mu, \nu).$$

Now that we have a way to compare loss profile sets, we can prove their stability under d_{R}. The theorem below establishes this stability property, and in addition, also proves that the constrained Bayes risk is stable in the sense of the Risk distance.

Theorem 37 (Stability of the Constrained Bayes Risk and Loss Profile) *If P and P' are problems, then*

$$|\mathcal{B}(P) - \mathcal{B}(P')| \leqslant d_{\mathrm{H}}^{d_{\mathrm{W},1}}(L(P), L(P')) \leqslant d_{\mathrm{R}}(P, P')$$

The proof, which is relegated to Appendix A, proceeds in two steps. We establish the first inequality using the fomulation of the constrained Bayes risk presented in Proposition 36. To prove the second inequality, we notice that the quantity in the middle defines a pseudometric on the collection of all problems, then apply the maximality claim in Theorem 28 to show that the Risk distance is larger. By ignoring the middle quantity in Theorem 37, we obtain stability the constrained Bayes risk.

Example 38 (Constrained Bayes Risk and Bounded Loss) *Here we remark that if the problems P and P' have bounded loss functions, then the stability of the constrained Bayes risk, namely that if P and P' are problems, then*

$$|\mathcal{B}(P) - \mathcal{B}(P')| \leqslant d_{\mathrm{R}}(P, P'),$$

can be easily and directly obtained through arguments involving the metric properties of the Risk distance. Indeed, if $c > \max(\sup \ell, \sup \ell')$, then, by Example 31 and the triangle inequality for the Risk distance, we have that

$$c - \mathcal{B}(P) = d_{\mathrm{R}}(P, P_{\bullet}(c)) \leqslant d_{\mathrm{R}}(P, P') + d_{\mathrm{R}}(P', P_{\bullet}(c)) = d_{\mathrm{R}}(P, P') + c - \mathcal{B}(P')$$

so that $\mathcal{B}(P') - \mathcal{B}(P) \leqslant d_{\mathrm{R}}(P, P')$. By swapping the roles of P and P' we obtain the inequality $\mathcal{B}(P) - \mathcal{B}(P') \leqslant d_{\mathrm{R}}(P', P)$ which then, together with the fact that $d_{\mathrm{R}}(P', P) = d_{\mathrm{R}}(P, P')$, implies the stability claim. Note that Theorem 37 implies the same claim without assuming the boundedness condition on the loss functions.

4.2 Stability of Rademacher Complexity

In statistical learning theory, the *Rademacher complexity* quantifies the flexibility of a predictor set and controls generalization error (Mohri et al., 2018). We establish the standard definition of Rademacher complexity of a function class, along with a related version stated in the language of supervised learning problems. In what follows, if μ is a measure on a measurable space X, let $\mu^{\otimes m}$ represent the m-fold product measure $\mu \otimes \cdots \otimes \mu \in \mathrm{Prob}(X^m)$.

Definition 39 *Let $m \geqslant 1$. Given a probability space (X, μ) and a family $\mathcal{F}$ of measurable real-valued functions on X, the classical mth Rademacher complexity of $\mathcal{F}$ is given by*

$$R_m(\mathcal{F}) := \int \int \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \, \mathrm{Rad}^{\otimes m}(d\sigma) \, \mu^{\otimes m}(d\overline{x}),$$

where Rad is the Rademacher distribution (that is, the uniform distribution on $\{-1, 1\}$), $\sigma = (\sigma_1, \ldots, \sigma_m)$, $\overline{x} = (x_1, \ldots, x_m)$. Similarly, the mth Rademacher complexity of a problem P is given by

$$R_m(P) := \int \int \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell_h(x_i, y_i) \, \mathrm{Rad}^{\otimes m}(d\sigma) \, \eta^{\otimes m}(d\overline{x} \times d\overline{y}).$$

That is, $R_m(P)$ is the mth Rademacher complexity of the function class $\{\ell_h | h \in H\}$.

The Rademacher complexity of a problem P is high if, given a random sample $((x_i, y_i))_{i=1}^{m}$ of m observations and a vector σ of random ± 1 noise, we can typically find some predictor h whose vector of losses $(\ell_h(x_i, y_i))_{i=1}^{m}$ is highly correlated with the noise σ. This would suggest that H is quite flexible, and possibly that empirical risk minimization over H runs the risk of overfitting on a random sample of size m.

For a fixed m, the Rademacher complexity R_m need not be stable under the Risk distance.

Example 40 *For $n \geqslant 2$, define the problem*

$$P_n = ([n], \{0, 1\}, \mathrm{unif}([n]) \otimes \delta_0, \ell, H_n)$$

where $[n] := \{1, \ldots, n\}$, ℓ is the 0-1 loss, and H_n is the set of indicator functions of singletons $1_k : [n] \to \{0, 1\}$ for all $k \in [n]$. Recall also the one-point problem $P_{\bullet} = P_{\bullet}(0)$ from Example 20.

We claim that $R_1(P_\bullet) = 0$, $R_1(P_n) = 1/2$ for all n, but $d_R(P_n, P_\bullet) \to 0$ as $n \to \infty$. This shows that the Rademacher complexity R_1 is not continuous with respect to the Risk distance. One can easily see that $R_1(P_\bullet) = 0$, since the loss function of $P_\bullet$ is identically 0. We can also easily compute

$$R_1(P_n) = \int \left(\frac{1}{2} \sup_{k \in [n]} \ell(1_k(x), y) - \frac{1}{2} \inf_{k \in [n]} \ell(1_k(x), y) \right) (\mathrm{unif}([n]) \otimes \delta_0)(dx \times dy)$$

$$= \int \left(\frac{1}{2}(1) - \frac{1}{2}(0) \right) (\mathrm{unif}([n]) \otimes \delta_0)(dx \times dy) = \frac{1}{2}.$$

Meanwhile, since there is only one coupling with the point mass $\delta_{(\bullet, \bullet)}$ and only one correspondence with the singleton $\{\mathrm{Id}_{\{\bullet\}}\}$, we can compute

$$d_R(P_n, P_\bullet) = \mathrm{dis}_{P_n, P_\bullet}(H_n \times \{\mathrm{Id}_{\{\bullet\}}\}, \mathrm{unif}([n]) \otimes \delta_0 \otimes \delta_{(\bullet, \bullet)})$$

$$= \sup_{k \in [n]} \int |\ell(1_k(x), 0) - 0| \, \mathrm{unif}([n])(dx)$$

$$= \sup_{k \in [n]} \int 1_k(x) \, \mathrm{unif}([n])(dx) = \sup_{k \in [n]} \frac{1}{n} = \frac{1}{n}.$$

Hence, even though $P_n \to P$ in the Risk distance, the Rademacher complexities $R_1(P_n)$ do not converge to $R_1(P_\bullet)$.

Example 40 shows that R_1 is not even continuous with respect to the Risk distance. This suggests more generally that we should not hope for straightforward stability of R_m under the Risk distance when m is small. If we instead take m to be arbitrarily large, we can still obtain a kind of stability.

Theorem 41 (Stability of Rademacher Complexity) *Let P and P' be problems and consider the family*

$$\mathcal{F} := \left\{ |\ell_h(-, -) - \ell'_{h'}(-, -)| \, \big| \, h \in H, h' \in H' \right\}$$

of functions $X \times Y \times X' \times Y' \to \mathbb{R}$. Then

$$\left| R_m(P) - R_m(P') \right| \leqslant d_R(P, P') + 2R_m(\mathcal{F}).$$

If P and P' have bounded loss functions and $\mathcal{F}$ forms a universal Glivenko-Cantelli class, then

$$\limsup_m \left| R_m(P) - R_m(P') \right| \leqslant d_R(P, P').$$

The proof can be found in Appendix A. The conditions under which $R_m(\mathcal{F}) \to 0$ as $m \to \infty$, and the rate of said convergence, are classical subjects of research in statistical learning theory. Results in this direction typically provide convergence rates of order $O(\sqrt{1/m})$ (Bartlett and Mendelson, 2002) or $O(\sqrt{\ln(m)/m})$ (Mohri et al., 2018, Chapter 3) under specific assumptions.

One can apply Proposition 13 to produce at a weaker form of Theorem 41 which does not reference the Glivenko-Cantelli property. Specifically, we can replace the universal Glivenko-Cantelli condition Theorem 41 with the following pair of sufficient conditions:

1. The sets H and H' can be endowed with Polish topologies, and subsequently Borel σ-algebras, such that the maps $(h, x, y) \mapsto \ell_h(x, y)$ and $(h', x', y') \mapsto \ell'_{h'}(x', y')$ are measurable.

2. The family of subgraphs $\{\mathrm{sg}(g_{h,h'}) \big| h \in H, h' \in H'\}$ (see Definition 12) has finite VC dimension, where

$$g_{h,h'}(x, y, x', y') := \big|\ell_h(x, y) - \ell'_{h'}(x', y')\big|.$$

In fact, we can state an even simpler sufficient condition involving the two problems P and P' separately.

Corollary 42 *Let P and P' be problems with bounded loss functions. Assume that H and H' are equipped with Polish topologies such that the maps $(h, x, y) \mapsto \ell_h(x, y)$ and $(h', x', y') \mapsto \ell'_{h'}(x', y')$ are measurable.*

If both $\{\ell_h | h \in H\}$ and $\{\ell'_{h'} | h' \in H'\}$ lie in finite-dimensional subspaces of $L^1(X \times Y)$ and $L^1(X' \times Y')$ respectively, then

$$\limsup_m \big|R_m(P) - R_m(P')\big| \leqslant d_{\mathrm{R}}(P, P').$$

The subspace condition in Corollary 42, while simple, may seem strange. We claim the condition is not as odd as it may at first appear.

Example 43 *Consider the generic linear regression problem P from Example 19. For that problem, the functions ℓ_h are of the form*

$$\ell_h(x, y) = (a^T x + b - y)^2$$

for $a \in \mathbb{R}^n$, $b \in \mathbb{R}$. These functions are polynomials in x and y of degree at most 2. Hence the collection $\{\ell_h | h \in H\} \subseteq L^1(\mathbb{R}^n \times \mathbb{R})$ lies within the finite-dimensional subspace of polynomials of degree at most 2. Hence P satisfies the subspace condition of Corollary 42.

More generally, any problem

$$P = (\mathbb{R}^n, \mathbb{R}^m, \eta, \ell, H)$$

will satisfy the subspace condition if ℓ is a polynomial loss function and H is of the form

$$H = \{a_1 f_1(x) + \cdots + a_k f_k(x) | a_1, \ldots, a_k \in \mathbb{R}\}$$

for some fixed $f_1, \ldots f_k : \mathbb{R}^n \to \mathbb{R}^m$. This is because the set $\{\ell_h | h \in H\}$ will fall within a finite-dimensional space of polynomials in the variables $f_1, \ldots f_k$ and y.

Proof [Corollary 42] Let

$$\mathcal{S} := \mathrm{span}(\{\ell_h(-, -) - \ell'_{h'}(-, -) \big| h \in H, h' \in H'\}),$$

where the span is taken in the space of all measurable functions $X \times Y \times X' \times Y' \to \mathbb{R}$. Consider

$$\max(\mathcal{S}, \mathcal{S}) := \{\max(f, g) | f, g \in \mathcal{S}\},$$

where the maximum $\max(f, g)$ is taken pointwise. Since $\mathcal{S}$ is finite-dimensional, it follows (Van Der Vaart and Wellner, 1996, Lemma 2.6.15 and Lemma 2.6.18(ii)) that the collection of subgraphs of functions in $\max(\mathcal{S}, \mathcal{S})$ has finite VC dimension. The set $\{g_{h,h'} \big| h \in H, h' \in H'\}$ lies in $\max(\mathcal{S}, \mathcal{S})$, so Proposition 13 finishes the proof. ■

4.3 Stability under Sampling Bias

A fundamental source of error in learning problems of any type is bias in the collection of training data. Suppose η is the true joint law that we are trying to approximate with a predictor. While we would like to sample from η directly, an imperfect collection process may over-represent certain regions of the data space. One would hope that minor bias would only have a minor effect on the problem. More formally, we might ask if the Risk distance is stable under small modifications to a problem's joint law. This section provides conditions for such stability results.

The Risk distance is stable under changes to the joint law with respect to the Wasserstein distance with a certain underlying metric. Under certain conditions, it is also stable with respect to the total variation distance (Definition 9). Define the pseudometric $s_{\ell,H}$ on $X \times Y$ by

$$s_{\ell,H}((x,y),(x',y')) := \sup_{h \in H} |\ell(h(x),y) - \ell(h(x'),y')|.$$

Proposition 44 *Let P be a problem, and let P' be a problem that is identical to P except possibly for its joint law η'. Then*

$$d_{\mathrm{R}}(P,P') \leqslant d_{\mathrm{W},1}^{s_{\ell,H}}(\eta,\eta').$$

Suppose furthermore that the loss function ℓ is bounded above by some constant ℓ_{max}. Then

$$d_{\mathrm{R}}(P,P') \leqslant \ell_{\mathsf{max}} d_{\mathrm{TV}}(\eta,\eta').$$

Proof One can bound $d_{\mathrm{R}}(P,P')$ by choosing the diagonal correspondence between H and itself to get

$$d_{\mathrm{R}}(P,P') \leqslant \inf_{\gamma \in \Pi(\eta,\eta')} \sup_{h \in H} \int |\ell(h(x),y) - \ell(h(x'),y')|\, \gamma(dx \times dy \times dx' \times dy')$$

$$\leqslant \inf_{\gamma \in \Pi(\eta,\eta')} \int \sup_{h \in H} |\ell(h(x),y) - \ell(h(x'),y')|\, \gamma(dx \times dy \times dx' \times dy') = d_{\mathrm{W},1}^{s_{\ell,H}}(\eta,\eta'),$$

proving the first part of the proposition. Now, if ℓ is bounded by ℓ_{max} then $s_{\ell,H}$ is bounded by $\ell_{\mathsf{max}} d$, where d is the discrete metric. Hence

$$d_{\mathrm{W},1}^{s_{\ell,H}}(\eta,\eta') \leqslant d_{\mathrm{W},1}^{\ell_{\mathsf{max}} d}(\eta,\eta') = \ell_{\mathsf{max}} d_{\mathrm{TV}}(\eta,\eta')$$

as desired. ∎

In the presence of sampling bias, the data collection process gives inordinate weight to some regions of the data space $X \times Y$ at the cost of other regions. We can think of this biased data collection as sampling not from the true probability measure η, but instead from a scaled probability measure given by

$$f\eta(A) := \int_A f(x,y)\, \eta(dx \times dy)$$

for all measurable A, where $f : X \times Y \to \mathbb{R}_{\geqslant 0}$ is a density function with respect to η, meaning f is a measurable function with $\int f(x,y)\,\eta(dx \times dy) = 1$. Here f is larger than 1 in regions of $X \times Y$ that are overrepresented by the data collection process, and less than 1 in regions underrepresented.

In extreme cases of sampling bias, some regions of the data space $X \times Y$ may be completely missing from the training data. In this case, instead of being sampled from the true joint law η, the data is sampled from some subset $A \subseteq X \times Y$ with $\eta(A) > 0$ according to the probability measure

$$\eta|_A(B) := \frac{1}{\eta(A)} \eta(A \cap B).$$

Using Proposition 44, we can show that d_R is stable under the above formulations of sampling bias.

Corollary 45 (Stability Under Sampling Bias) *Let P be a problem.*

(1) Let $f : X \times Y \to \mathbb{R}_{\geqslant 0}$ be a density function with respect to η. Define

$$P_f := (X, Y, f\eta, \ell, H).$$

Then

$$d_R(P, P_f) \leqslant \frac{1}{2} \int \big|1 - f(x,y)\big|\, \eta(dx \times dy).$$

(2) Let $A \subseteq X \times Y$ be a measurable subset with $\eta(A) > 0$. Define

$$P|_A := (X, Y, \eta|_A, \ell, H).$$

Then

$$d_R(P, P|_A) \leqslant 1 - \eta(A).$$

Proof First recall that

$$d_{\mathrm{TV}}(\eta, \eta') = \sup_{B \subseteq X \times Y} \eta(B) - f\eta(B) = \sup_{B \subseteq X \times Y} f\eta(B) - \eta(B),$$

where the infima range over all measurable subsets of $X \times Y$. Then

$$d_{\mathrm{TV}}(\eta, \eta') = \frac{1}{2} \left(\sup_{B \subseteq X \times Y} \eta(B) - f\eta(B) + \sup_{B \subseteq X \times Y} f\eta(B) - \eta(B) \right)$$

$$= \frac{1}{2} \left(\sup_{B \subseteq X \times Y} \int_B 1 - f(x,y)\, \eta(dx \times dy) + \sup_{B \subseteq X \times Y} \int_B f(x,y) - 1\, \eta(dx \times dy) \right)$$

$$= \frac{1}{2} \left(\int_{\{f<1\}} 1 - f(x,y)\, \eta(dx \times dy) + \int_{\{f>1\}} f(x,y) - 1\, \eta(dx \times dy) \right)$$

$$= \frac{1}{2} \int \big|1 - f(x,y)\big|\, \eta(dx \times dy).$$

To finish the proof, apply part (1) when $f = \frac{1}{\eta(A)} 1_A$. ∎

36

4.4 Stability under Sampling Noise

Another source of error in sampling is the presence of noise in the process of collecting data. While we would like to sample from η directly, our data is rarely collected with perfect fidelity. One would hope that a small amount of noise would not change the problem much. In this section, we model noise with a *noise kernel* and show that we can often guarantee that the Risk distance is stable under the application of such noise.

First we consider the case where the loss function ℓ is bounded by a metric on Y. Our bounds are in terms of a generalization of the Wasserstein metric for Markov kernels.

Definition 46 *Let X be a measure space and let $M : X \to \mathrm{Prob}(Y)$, $N : X \to \mathrm{Prob}(Z)$ be Markov kernels. A* coupling *between M and N is a Markov kernel $\tau : X \to \mathrm{Prob}(Y \times Z)$ such that for a.e. $x \in X$,*

$$(\pi_1)_\sharp(\tau(x)) = M(x),$$
$$(\pi_2)_\sharp(\tau(x)) = N(x).$$

Just as we use $\Pi(\mu, \nu)$ to denote the set of all couplings of probability measures μ and ν, we use $\Pi(M, N)$ to denote the set of all couplings of Markov kernels M and N.

Note that if X is a single point, we recover the classical definition of a coupling of probability measures.

Definition 47 (Wasserstein Distance for Markov Kernels) *Let (X, μ) be a measure space and let (Y, d_Y) be a metric space. Let $M, N : X \to \mathrm{Prob}(Y)$ be a pair of Markov kernels. For $1 \leqslant p < \infty$, the p-Wasserstein distance between M and N is given by*

$$d_{\mathrm{W},p}(M, N) := \left(\inf_{\tau \in \Pi(M,N)} \int_X \int_{Y \times Y} d_Y^p(y, y') \tau(x)(dy \times dy') \mu(dx) \right)^{1/p}.$$

Definitions 46 and 47 are due to Patterson (2020, Defs 5.1 and 5.2). Patterson notes that terminology conflicts between "products" and "couplings" of Markov kernels are present in the literature. As with the Wasserstein distance, we will have little use for $d_{\mathrm{W},p}$ for $p \neq 1$, and so establish the convention that the *Wasserstein distance for Markov kernels* refers to $d_{\mathrm{W},1}$. An alternative, equivalent definition of the metric $d_{\mathrm{W},1}$ may be more intuitive:

$$d_{\mathrm{W},1}(M, N) = \int_X d_{\mathrm{W},1}^{d_Y}(M(x), N(x)) \, \mu(dx).$$

Equivalence of these definitions was noted by Patterson (2020), and follows from results in the theory of optimal transport (Villani, 2008, Corollary 5.22). A distance based on a generalization of the above alternative definition was independently discovered and studied by Kitagawa and Takatsu (2023) under the name "disintegrated Monge-Kantorovich distance".

We are abusing notation by using the same symbol for the Wasserstein distances between measures and between Markov kernels. We justify this by noting that the former is a special case of the latter. Indeed, recall that if $\mu, \nu \in \mathrm{Prob}(X)$ are probability measures on a metric space X, we can think of μ and ν as Markov kernels from the one-point measure space $(\{\bullet\}, \delta_\bullet)$. The Wasserstein distance between these Markov kernels then agrees with the Wasserstein distance between the measures μ and ν.

The Wasserstein distance for Markov kernels also generalizes the L^1 distance on measurable functions. Indeed, if we have measurable functions $f, g : X \to Y$, then the distance $d_{\mathrm{W},1}(\delta_f, \delta_g)$ between the induced deterministic Markov kernels is just the L^1 distance between f and g.

A simple kind of noise can be modeled with a Markov kernel $M : Y \to \mathrm{Prob}(Y)$. We interpret this noise as follows. When we make an observation that would have label $y \in Y$, we instead observe a random corrupted label with law $M(y)$. More specifically, let P be a problem whose joint law has X-marginal α and X-disintegration $\beta : X \to \mathrm{Prob}(Y)$. Without the presence of noise, an observation has a random input component x chosen with law α, and a random label chosen with law $\beta(x)$. The presence of noise represented by M creates a new problem where the labels are instead chosen according to the law $(\beta \cdot M)(x)$. In other words, we get a problem with the same X-marginal α, but with disintegration $\beta' : X \to \mathrm{Prob}(Y)$ given by

$$\beta'(x) := (\beta \cdot M)(x).$$

We might call this type of noise "X-independent Y-noise," (otherwise known as "simple label noise") since the noise M is applied in the Y direction, and the same noise kernel M is applied for all $x \in X$.

A more general kind of noise would allow for dependence on X. Instead of a Markov kernel $M : Y \to \mathrm{Prob}(Y)$, we model our noise as a Markov kernel $N : X \times Y \to \mathrm{Prob}(Y)$. We write N_x to mean the Markov kernel $N(x, -) : Y \to \mathrm{Prob}(Y)$. In the presence of this kind of noise, the problem P becomes a new problem P' with the same X-marginal α, but X-disintegration β' now given by

$$\beta'(x) := (\beta \cdot N_x)(x).$$

We might call this kind of noise "X-dependent Y-noise".

Consider a noise kernel $N = \delta_{\pi_2}$, which is the Markov kernel sending (x, y) to the point mass δ_y. Then for all $x \in X$, we have $N_x(y) = \delta_y$ and hence $\beta \cdot N_x = \beta$. In other words, N represents no noise at all, and applying N to a problem would leave the problem unchanged. Our next stability result states that, if any kernel N is close to this "no noise" case under $d_{\mathrm{W},1}$, then applying N will not move the problem much in the Risk distance, provided that ℓ is well-controlled by some metric on Y.

Lemma 48 *Suppose that P and P' are problems given by*

$$P = (X, Y, \eta, \ell, H)$$

$$P' = (X, Y, \eta', \ell, H)$$

where η and η' have the same X-marginal $\alpha \in \mathrm{Prob}(X)$, but different disintegrations along X, called β and β' respectively. Suppose also that Y is equipped with a metric d_Y with respect to which ℓ is Lipschitz in the second argument, meaning there is a constant $C > 0$ such that

$$\left|\ell(z, y) - \ell(z, y')\right| \leqslant C d_Y(y, y')$$

for all $y, y', z \in Y$. Then

$$d_{\mathrm{R}}(P, P') \leqslant C d_{\mathrm{W},1}^{d_Y}(\beta, \beta').$$

Theorem 49 (Stability under X-dependent Y-noise) *Let P be a problem whose output space Y is equipped with a metric d_Y, with respect to which the loss function ℓ is C-Lipschitz in the second argument. Let P' be the problem obtained by applying the noise kernel $N : X \times Y \to \mathrm{Prob}(Y)$ to P. Then*

$$d_{\mathrm{R}}(P, P') \leqslant C d_{\mathrm{W},1}^{d_Y}(N, \delta_{\pi_2}).$$

The proofs are relegated to Appendix A. We illustrate Theorem 49 with an example regarding label noise in binary classification.

Example 50 *Let $P = (X, \{0, 1\}, \eta, \ell, H)$ be a binary classification problem where ℓ is the 0-1 loss $\ell(y, y') = 1_{y \neq y'}$. Suppose we apply noise to P by taking each observation (x, y) and doing the following:*

- *With probability ε, re-assign its label y uniformly at random.*

- *With probability $(1 - \varepsilon)$, leave the label y as it is.*

Through this process we are effectively applying the noise kernel

$$N_\varepsilon := \varepsilon u + (1 - \varepsilon)\delta_{\pi_2},$$

where u is the uniform measure on $\{0, 1\}$. Let P'_ε be the problem P with the noise kernel N_ε applied. Since the 0-1 loss is a metric, we can apply Theorem 49 with $d_Y = \ell$ and $C = 1$ to get

$$\begin{aligned}
d_{\mathrm{R}}(P, P'_\varepsilon) &\leqslant d_{\mathrm{W},1}(\varepsilon u + (1 - \varepsilon)\delta_{\pi_2}, \delta_{\pi_2}) \\
&= \int d_{\mathrm{W},1}^\ell(\varepsilon u + (1 - \varepsilon)\delta_y, \delta_y) \, \eta(dx \times dy).
\end{aligned}$$

Recall (Section 2.4) that when the underlying metric is the discrete metric, the Wasserstein distance is the same as the total variation distance. Furthermore, one can show that $d_{\mathrm{TV}}(\varepsilon u + (1 - \varepsilon)\delta_y, \delta_y) = \varepsilon d_{\mathrm{TV}}(u, \delta_y) = \varepsilon/2$. Hence we get

$$d_{\mathrm{R}}(P, P'_\varepsilon) \leqslant \varepsilon/2.$$

One could generalize further; suppose that the probability ε of reassignment depends on x. That is, instead of selecting a constant ε, we select a measurable function $\varepsilon : X \to [0, 1]$. Then, for an observation (x, y), with probability $\varepsilon(x)$ we re-assign y uniformly at random. Then the same argument above instead gives us the bound

$$d_{\mathrm{R}}(P, P'_\varepsilon) \leqslant \frac{1}{2} \int \varepsilon(x)\alpha(dx),$$

where α is the X-marginal of η. That is, the risk distance between P and the noised problem P'_ε is bounded by half the average *value of $\varepsilon(x)$.*

In Section 6.5, we will consider an even more general noise model, allowing for noise in the X and Y directions simultaneously. We will show that a modification of the Risk distance, called the L^p-Risk distance, is stable under such noise in Theorem 91, which is analogous to Theorem 49.
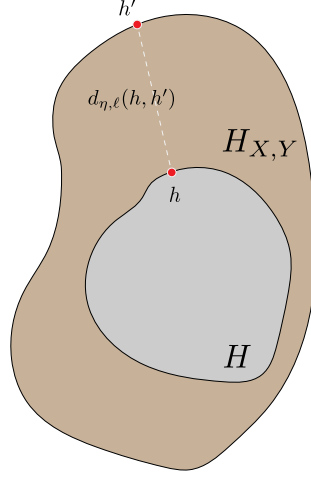
Figure 3: Distance between H and $H_{X,Y}$; see Example 52.

4.5 Stability under Modification of Predictor Set

We have shown that d_R is stable under changes to the loss function and certain changes to η. We round out this section by demonstrating stability of d_R under changes to the predictor set of a problem. To state such a stability result, we first need a way to quantify changes to the predictor set. The most salient aspect of a predictor h is its performance with respect to the chosen loss function. In other words, more crucial than the predictor h itself is the function $\ell_h(x, y) := \ell(h(x), y)$. With this perspective in mind, we can map H into $L^1(\eta)$ via the map $h \mapsto \ell_h$, and compare predictors using the L^1 norm there. That is, we compare two predictors h and h' via the quantity

$$d_{\ell,\eta}(h, h') := \|\ell_h - \ell_{h'}\|_{L^1(\eta)} = \int \big|\ell_h(x, y) - \ell_{h'}(x, y)\big|\, \eta(dx \times dy).$$

This defines a pseudometric on the set of all measurable functions $h : X \to Y$ for which $\|\ell_h\|_{L^1(\eta)} < \infty$, which any predictor satisfies since the definition of a problem mandates that predictors have finite risk. The map $h \mapsto \ell_h$ identifies two predictors h and h' if they η-almost always incur the same loss on a random observation in $X \times Y$.

We can now state a stability theorem in terms of the Hausdorff distance on predictor sets with underlying metric $d_{\ell,H}$, i.e. the metric

$$d_{\mathrm{H}}^{d_{\ell,\eta}}(H, H') = \inf_{R \in \mathcal{C}(H,H')} \sup_{(h,h') \in R} \|\ell_h - \ell_{h'}\|_{L^1(\eta)}.$$

Theorem 51 *Let P be a problem, and let P' be a problem that is identical to P except possibly for the predictor set H'. Then*

$$d_\mathrm{R}(P, P') \leqslant d_{\mathrm{H}}^{d_{\ell,\eta}}\big(H, H'\big).$$

Example 52 *Note that if* $P = (X, Y, \eta, \ell, H)$ *and* $P' = (X, Y, \eta, \ell, H_{X,Y})$, *where* $H_{X,Y}$ *stands for the set of all measurable functions from* X *to* Y, *Theorem 37 implies that*

$$\mathcal{B}^*(P) \leqslant \mathcal{B}(P) \leqslant \mathcal{B}^*(P) + \sup_{h' \in H_{X,Y}} \inf_{h \in H} d_{\ell,\eta}(h, h').$$

The right-hand side contains the worst case $d_{\ell,\eta}$ *distance between an arbitrary measurable function* h' *and the (constrained) predictor set* H. *In other words, this quantity expresses how far* H *is from the superclass* $H_{X,Y}$ *and therefore controls the* approximation error*: the difference between the constrained and unconstrained Bayes risks; see Figure 3.*

Proof [Theorem 51] We bound $d_{\mathrm{R}}(P, P')$ by choosing the diagonal coupling between η and itself. That is,

$$d_{\mathrm{R}}(P, P') \leqslant \inf_{R \in \mathcal{C}(H, H')} \sup_{(h, h') \in R} \int \left| \ell(h(x), y) - \ell(h'(x), y) \right| \eta(dx \times dy)$$

$$= \inf_{R \in \mathcal{C}(H, H')} \sup_{(h, h') \in R} \| \ell_h - \ell_{h'} \|_{L^1(\eta)} = d_{\mathrm{H}}^{d_{\ell,\eta}}(H, H'),$$

which is the desired bound. ∎

We illustrate the metric $d_{\mathrm{H}}^{d_{\ell,\eta}}$ and Theorem 51 with a toy example.

Example 53 *Let* $P = (\mathbb{R}^2, \{0, 1\}, \eta, \ell, H)$ *be the classification problem from Example 18. Recall that the predictors in* H *were the functions of the form*

$$h_{a,b,c}(x_1, x_2) := \begin{cases} 1 & ax_1 + bx_2 \geqslant c \\ 0 & ax_1 + bx_2 < c \end{cases}.$$

We modify H *by adding to each predictor a small area on which they always predict class 1. That is, we define a new predictor set* $H'_t := \left\{ h'_{a,b,c,t} \middle| a, b, c \in \mathbb{R} \right\}$ *where* $h'_{a,b,c,t} : \mathbb{R}^2 \to \{0, 1\}$ *is given by*

$$h'_{a,b,c,t}(x_1, x_2) := \begin{cases} 1 & x \in B_t(0) \\ h_{a,b,c}(x) & otherwise \end{cases}.$$

Here $B_t(0)$ *represents the ball of radius* t *centered at 0. We can construct a correspondence between* H *and* H'_t *by coupling each* $h_{a,b,c}$ *with* $h'_{a,b,c,t}$. *That is, we define*

$$R := \left\{ (h_{a,b,c}, h'_{a,b,c,t}) \middle| a, b, c \in \mathbb{R} \right\}.$$

Then $R \in \mathcal{C}(H, H'_t)$, *so we can write the bound*

$$d_{\mathrm{H}}^{d_{\ell,\eta}}(H, H'_t) \leqslant \sup_{(h_{a,b,c}, h'_{a,b,c,t}) \in R} \| \ell_{h_{a,b,c}} - \ell_{h_{a,b,c,t}} \|_{L^1(\eta)}$$

$$= \sup_{a,b,c \in \mathbb{R}} \int \left| \ell(h_{a,b,c}(x), y) - \ell(h_{a,b,c,t}(x), y) \right| \eta(dx \times dy)$$

$$= \sup_{a,b,c \in \mathbb{R}} \int_{B_t(0) \times \{0,1\}} \left| \ell(h_{a,b,c}(x), y) - \ell(h_{a,b,c,t}(x), y) \right| \eta(dx \times dy)$$

$$\leqslant \eta(B_t(0) \times \{0, 1\}).$$

41

Furthermore, assuming $\eta(0 \times \{0, 1\}) = 0$, we can take the limit as $t \to 0$ to get $\lim_{t \to 0} d_{\mathrm{H}}^{d_{\ell, \eta}}(H, H_t') = 0$. If we define $P_t := (\mathbb{R}^2, \{0, 1\}, \eta, \ell, H_t')$ for all $t > 0$, By Theorem 51, P_t converges to P under the Risk distance as $t \to 0$.

We now provide a more serious application of Theorem 51. Different classes of predictors are often compared in the context of *universal approximation theorems*, which generally state that a certain kind of function can be approximated arbitrarily well using a certain class of model. Using Theorem 51, one can translate universal approximation theorems into statements about the risk distance. While there are many universal approximation theorems (Pinkus, 1999), we will demonstrate with a classical example (see Leshno et al., 1993).

Proposition 54 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous non-polynomial function. Let $P = (X, \mathbb{R}, \eta, \ell, H)$ be a problem where*

- *X is a compact subset of $\mathbb{R}^m$,*

- *η is any probability distribution on X,*

- *ℓ is a uniformly continuous loss function, and*

- *H is the set of all functions $X \to \mathbb{R}$ which can come from a feedforward network with a single hidden layer of arbitrary width and with activation function ϕ. That is, H is the set of all functions of the form*

$$(x_1, \ldots, x_m) \mapsto \sum_{i=1}^{M} a_i \phi \left(\sum_{j=1}^{m} w_{ij} x_j + b_i \right)$$

   where the parameters a_i, w_{ij}, and b_j range over $\mathbb{R}$, and M ranges over the positive integers.

Let P' be the same problem, but with predictor set H' the set of all continuous functions $X \to \mathbb{R}$. Then $d_{\mathrm{R}}(P, P') = 0$.

Proof Let $h' : X \to \mathbb{R}$ be continuous and let $\varepsilon > 0$ be arbitrary. By uniform continuity of ℓ, there exists a $\delta > 0$ such that, if $|y - y'| < \delta$, then $|\ell(y, y'') - \ell(y', y'')| < \varepsilon$ for any $y'' \in \mathbb{R}$. By (Leshno et al., 1993, Thm 1), there exists an $h \in H$ with $|h(x) - h'(x)| < \delta$ for all $x \in X$. Then

$$d_{\ell, \eta}(h, h') = \int \left| \ell(h(x), y) - \ell(h'(x), y) \right| \eta(dx \times dy) \leqslant \int \varepsilon \, \eta(dx \times dy) = \varepsilon.$$

Since $H \subseteq H'$, this implies the Hausdorff distance $d_H^{d_{\ell, \eta}}(H, H')$ is less than ε. By Theorem 51, $d_{\mathrm{R}}(P, P') < \varepsilon$ as well. ∎

In other words, under certain conditions, the risk distance does not distinguish between a model and the class of functions in which it is dense.
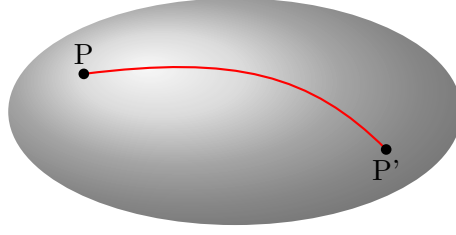
Figure 4: A simple depiction of two problems P and P' in the space of problems, joined by a geodesic.

5. Geometry of the Space of Problems

We now shift our attention from stability properties of d_{R} to its geometric properties. We will explore geodesics and their connection to optimal couplings and correspondences. We also show that classification problems (that is, those with finite output spaces) are dense in a larger class of problems. We will use this density to study the effect of limited sampling by studying the convergence of "empirical problems," which are problems whose joint laws are given by empirical measures gleaned from finite random samples.

5.1 Geodesics

We begin our exploration of the geometry of the space of problems by searching for geodesics. A *geodesic* in a metric space (X, d_X) is a continuous function from the unit interval $f : [0, 1] \to X$ such that, for all $s, t \in [0, 1]$,

$$d_X(f(t), f(s)) = d_X(f(0), f(1))|s - t|.$$

Some would call such an f a "shortest path" (Burago et al., 2001, Chapter 2.5) and reserve the term "geodesic" for a path that *locally* satisfies the above property. However, the above definition is standard in the field of optimal transport (Ambrosio et al., 2021, Section 9.2, Eq. 9.4).

While valuable in illuminating the geometry of a space, geodesics are not solely of geometric interest. A geodesic between two points gives us a way to interpolate or average between them. If a family $\{P_t\}_{t \in [0,1]}$ of problems defines a geodesic in the space of problems, then $P_{1/2}$ could reasonably be considered an average of P_0 and P_1. If we would rather take a weighted average, we could choose smaller or larger values of t to give more weight to P_0 or to P_1 respectively.

We will show that geodesics can be supplied by certain couplings and correspondences.

Definition 55 *Let P and P' be problems. If there exists a coupling $\gamma \in \Pi(\eta, \eta')$ and a correspondence $R \in \mathcal{C}(H, H')$ such that*

$$d_{\mathrm{R}}(P, P') = \mathrm{dis}_{P,P'}(R, \gamma),$$

then we call γ an optimal coupling *and R an* optimal correspondence *for P and P'. That is, γ and R are called optimal if they achieve the infima in the definition of $d_{\mathrm{R}}(P, P')$.*

In Section 5.2, we will explore sufficient conditions for the existence of optimal couplings and correspondences. Presently, we motivate our interest in optimal couplings and correspondences by showing that they are closely related to the geometry of the space of problems; optimal couplings and correspondences provide us with explicit geodesics in the space of problems.

Theorem 56 *Suppose P_0 and P_1 admit an optimal coupling γ and an optimal correspondence R. Then the following family of problems is a geodesic between P_0 and P_1:*

$$P_t := (X_0 \times X_1, Y_0 \times Y_1, \gamma, \ell_t, R_\times)$$

for all $t \in [0, 1]$, where

- $\ell_t((y_0, y_1), (y_0', y_1')) := (1 - t)\ell_0(y_0, y_0') + t\ell_1(y_1, y_1')$.

- $R_\times := \{h_0 \times h_1 | h_0 \in H_0, h_1 \in H_1\}$.

It may at first appear that we have given P_0 two conflicting definitions in the above statement, once as the given

$$P_0 := (X_0, Y_0, \eta_0, \ell_0, H_0)$$

and again by setting $t = 0$ in our parameterized family, yielding

$$P_0 := (X_0 \times X_1, Y_0 \times Y_1, \gamma, \ell_0, R_\times).$$

However, the latter problem is a simulation of the former via the projection maps $X_0 \times X_1 \to X_0$ and $Y_0 \times Y_1 \to Y_0$. Hence the two problems are distance zero apart under the Risk distance and can be safely identified. A similar line of thought applies to P_1.

The proof that this family is a geodesic is very similar to the proofs of analogous results for the Gromov-Hausdorff and Gromov-Wasserstein distances (Chowdhury and Mémoli, 2018; Sturm, 2023).

Proof Define $C := d_\mathrm{R}(P_0, P_1)$. Let $0 \leqslant s \leqslant t \leqslant 1$. If we can show

$$d_\mathrm{R}(P_s, P_t) \leqslant C(t - s),$$

then equality will follow. Indeed, if the inequality were strict for some $0 \leqslant s_0 \leqslant t_0 \leqslant 1$, then we could use the triangle inequality to write

$$C = d_\mathrm{R}(P_0, P_1) \leqslant d_\mathrm{R}(P_0, P_{s_0}) + d_\mathrm{R}(P_{s_0}, P_{t_0}) + d_\mathrm{R}(P_{t_0}, P_1)$$
$$< s_0 C + (t_0 - s_0)C + (1 - s_0)C = C,$$

which is a contradiction.

Now, we can bound $d_{\mathrm{R}}(P_s, P_t)$ by selecting the diagonal coupling from $\Pi(\gamma, \gamma)$ and diagonal correspondence from $\mathcal{C}(R_\times, R_\times)$ to get

$$d_{\mathrm{R}}(P_s, P_t)$$

$$\leqslant \sup_{h_0 \times h_1 \in R_\times} \int \left| \ell_s((h_0(x_0), h_1(x_1)), (y_0, y_1)) - \ell_t((h_0(x_0), h_1(x_1)), (y_0, y_1)) \right| \gamma(dx_0 \times dy_0 \times dx_1 \times dy_1)$$

$$= \sup_{h_0 \times h_1 \in R_\times} \int \left| (t-s)\big(\ell_0(h_0(x_0), y_0) + \ell_1(h_1(x_1), y_1)\big) \right| \gamma(dx_0 \times dy_0 \times dx_1 \times dy_1)$$

$$= (t-s) \sup_{(h_0, h_1) \in R} \int \left| \ell_0(h_0(x_0), y_0) + \ell_1(h_1(x_1), y_1) \right| \gamma(dx_0 \times dy_0 \times dx_1 \times dy_1)$$

$$= (t-s)C.$$

The last equality is because γ and R are an optimal coupling and correspondence of P_0 and P_1. ∎

5.2 Existence of Optimal Couplings and Correspondences

Motivated by Theorem 56, we search for conditions under which optimal couplings and correspondences exist. We will achieve this by imposing some continuity and compactness conditions on our problems.

Definition 57 *A problem P is called* continuous *if, for all $h \in H$, ℓ_h is continuous η-almost everywhere.*

Note that the above condition applies to many relevant problems. A typical linear regression problem will be continuous. For instance, the linear regression problem in Example 19 is continuous, since the loss is continuous, as is every predictor. Under reasonable assumptions, a classification problem can be continuous as well; one only needs the possible decision boundaries to each have measure zero in the X-marginal. Take the classification problem P in Example 18. In that problem, ℓ_h is discontinuous for every predictor h. However, if we assume that any line in $\mathbb{R}^2$ has measure zero with respect to the X-marginal of η (as would be the case if the X-marginal of η is given by a continuous probability distribution), then any linear decision boundary has measure zero. Consequently, ℓ_h will be continuous (indeed, constant) η-almost everywhere.

Let P, P' be continuous problems. Recall that the risk distortion is the function

$$\mathrm{dis}_{P,P'} : \mathcal{C}(H, H') \times \Pi(\eta, \eta') \to \mathbb{R}$$

$$\mathrm{dis}_{P,P'}(R, \gamma) := \sup_{(h, h') \in R} \int \left| \ell_h(x, y) - \ell_{h'}(x', y') \right| \gamma(dx \times dy \times dx' \times dy').$$

We would like to discuss the potential continuity of this function, and hence require a suitable topology on its domain. We endow $\Pi(\eta, \eta')$ with the weak topology. We give the set $\{\ell_h | h \in H\}$ the $L^1(\eta)$ topology, and endow H with the initial topology induced by the map $h \mapsto \ell_h$. Equivalently, we endow H with the pseudometric $d_{\ell, \eta}$ defined in

Section 4.5 and similarly equip H' with $d_{\ell',\eta'}$. We then endow $H \times H'$ with the product metric $\delta((g,g'),(h,h')) := \max\{d_{\ell,\eta}(g,h), d_{\ell',\eta'}(g',h')\}$, and place the resulting Hausdorff distance d_{H}^{δ} on $\mathcal{C}(H,H')$.

Define

$$\widetilde{\mathrm{dis}}_{P,P'} : H \times H' \times \Pi(\eta,\eta') \to \mathbb{R}$$

$$\widetilde{\mathrm{dis}}_{P,P'}(h,h',\gamma) := \int \big| \ell_h(x,y) - \ell_{h'}(x',y') \big| \, \gamma(dx \times dy \times dx' \times dy').$$

Hence

$$\mathrm{dis}_{P,P'}(R,\gamma) = \sup_{(h,h') \in R} \widetilde{\mathrm{dis}}_{P,P'}(h,h',\gamma).$$

We proceed with a sequence of lemmas. We begin by showing that, under certain conditions, $\widetilde{\mathrm{dis}}_{P,P'}$ is lower semi-continuous, then leverage this result to obtain lower semi-continuity of $\mathrm{dis}_{P,P'}$. Recall that a function $f : S \to \mathbb{R}$ on a topological space S is *lower semi-continuous* if, for all $x_0 \in S$, we have $\liminf_{x \to x_0} f(x) \geqslant f(x_0)$. Lower semi-continuity is a useful notion because such functions satisfy a limited version of the extreme value theorem. Specifically, a lower semi-continuous function on a compact space achieves its minimum.

Lemma 58 *Let P and P' be continuous problems. Then $\widetilde{\mathrm{dis}}_{P,P'}$ is lower semi-continuous.*

Lemma 59 *Let P and P' be continuous problems. Then $\mathrm{dis}_{P,P'}$ is lower semi-continuous.*

Using the fact that a lower semi-continuous function on a compact set achieves its minimum, we can prove the following Theorem.

Theorem 60 *If P and P' are continuous problems with compact predictor sets, then there exists an optimal coupling and correspondence for P and P'. That is, there exists $R \in \mathcal{C}(H,H')$ and $\gamma \in \Pi(\eta,\eta')$ such that*

$$d_{\mathrm{R}}(P,P') = \mathrm{dis}_{P,P'}(R,\gamma).$$

The proofs can be found in Appendix A. In turn, Theorem 60 and the characterization of weak isomorphism provided by Theorem 29 give us conditions under which weak isomorphism is equivalent to being Risk distance zero apart.

Corollary 61 *If P and P' are continuous problems with compact predictor sets, then P and P' are weakly isomorphic if and only if $d_{\mathrm{R}}(P,P') = 0$.*

5.3 Density of Classification Problems

One benefit of equipping a set with a pseudometric is that, in return, we receive a notion of convergence. One can ask whether a sequence of points converges, how quickly it does so, and what the limit might be. Density results pave the way for limiting arguments. It is often possible, by a limiting argument, to transfer knowledge about a "simple" dense subspace to the space as a whole.

In this section, we show that classification problems (that is, problems whose output space is a finite collection of labels) are dense in a broader class of functions under d_{R}. We will leverage this information in Section 5.4 to prove the main result of the section: the convergence of empirical problems (Theorem 69).
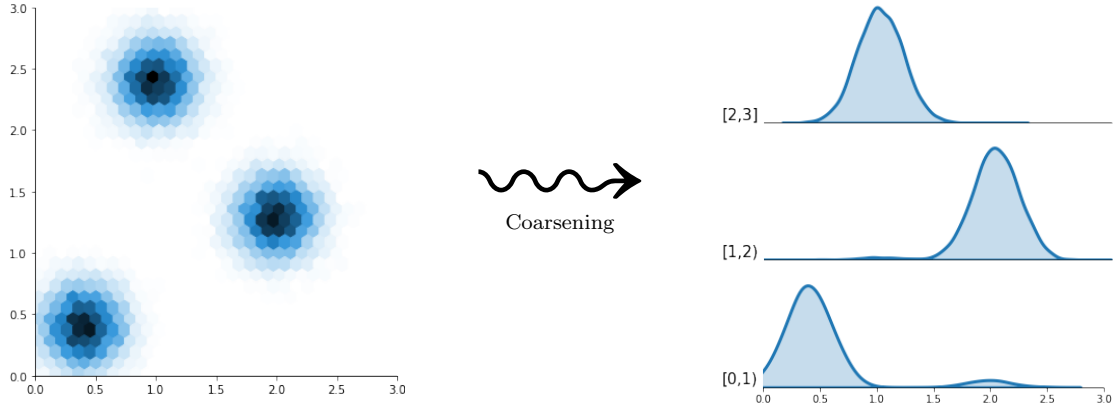
Figure 5: A depiction of the coarsening process in Example 64. On the left is a heat map depicting the joint law η of a problem P whose input and output spaces are both the interval $X = Y = [0, 3]$. We consider the partition $Q = \{[0, 1), [1, 2), [2, 3]\}$ of $Y = [0, 3]$. The coarsened problem $P_Q := (X, Q, \eta_Q, \ell_Q, H_Q)$ has three points in the output space Q, represented by the vertical axis. The joint law η_Q is then a probability measure on three copies of $[0, 3]$. We depict η_Q on the right with density plots of the three class-conditional probability measures.

Definition 62 (Classification Problem) *A problem is called a* classification problem *if its response space is finite. A problem is called* compact *if its response space is compact and its loss function is continuous.*

A classification problem is so called because we can think of its response space as a finite collection of labels. In Appendix B, we characterize the problems that are weakly isomorphic to classification problems.

Given a compact problem P, we can produce a related classification problem P_Q by clustering the response space into a finite partition Q and applying a procedure we call *coarsening*. The definition of a coarsening makes use of a certain map associated to a partition. Given a set A and a partition Q of A, consider the map $\pi_Q : A \to Q$ which sends each $a \in A$ to the unique partition block $q \in Q$ which contains a. That is, π_Q is defined such that $a \in \pi_Q(a)$ for all $a \in A$. This map π_Q is often called the *quotient map* associated to Q.

Definition 63 (Coarsening) *Let P be a problem, let Q be a finite partition of Y into measurable sets, and let $\pi_Q : Y \to Q$ be the quotient map associated to Q.*

The coarsening of P with respect to Q *is the* classification *problem*

$$P_Q = (X, Q, \eta_Q, \ell_Q, H_Q)$$

where

- $\eta_Q := (\mathrm{Id}_X \times \pi_Q)_\sharp \eta$

- $\ell_Q(q, q') := \sup_{y \in q, y' \in q'} \ell(y, y')$,

- $H_Q := \{\pi_Q \circ h \,|\, h \in H\}$.

Note that if Q is a finite partition of Y, then Q is itself a finite set of partition blocks. The response space Q of the coarsened problem P_Q is a finite set with one element for each block of the partition Q, and hence P_Q is indeed a classification problem.

Example 64 *An example of the coarsening procedure is depicted in Figure 5. We imagine a problem*

$$P = ([0,3], [0,3], \eta, \ell, H)$$

where η is as pictured, $\ell(y_1, y_2) := |y_1 - y_2|$ is the Euclidean distance, and H is the set of increasing continuous bijections $[0,3] \to [0,3]$. We choose the partition $Q = \{[0,1), [1,2), [2,3]\}$ of P and consider the coarsened problem

$$P_Q = ([0,3], Q, \eta_Q, \ell_Q, H_Q).$$

Then P_Q is a classification problem with three points in its output space $Q = \{q_1, q_2, q_3\}$, where $q_1 := [0,1)$, $q_2 := [1,2)$, and $q_3 := [2,3]$. Then η_Q is a probability measure on $[0,3] \times Q$, a space which consists of three line segments: $[0,3] \times \{q_1\}$, $[0,3] \times \{q_2\}$, and $[0,3] \times \{q_3\}$.

The coarsened loss

$$\ell_Q(q_i, q_j) := \sup_{y \in q_i, y' \in q_j} \ell(y, y') \text{ for } i, j = 1, 2, 3,$$

can be written in the form of a table:

ℓ_Q	q_1	q_2	q_3
q_1	1	2	3
q_2	2	1	2
q_3	3	2	1

The predictor set H_Q is $\{\pi_Q \circ h \,|\, h \in H\}$, where $\pi_Q : Y \to Q$ is the quotient map. Then H_Q can also be written as the set of all functions $g : [0,3] \to Q$ of the form

$$g(t) = \begin{cases} q_1 & t < a \\ q_2 & a \leqslant t < b \\ q_3 & b \leqslant t \end{cases}$$

for all $0 < a \leqslant b < 3$.

Coarsening is reminiscent of the "probing reduction" (Langford and Zadrozny, 2005), which takes a probability estimation problem and solves it with an ensemble of classifier learners, hence "reducing" the problem of probability estimation to the problem of binary classification. Indeed, one can view the probing reduction as a practical method of approximating a problem whose output space is $Y = [0,1]$ by iteratively constructing finer and finer partitions of Y and solving the resulting coarsened problems with binary classifiers.

Theorem 65 *Let P be a compact problem. For any $\varepsilon > 0$, there is a partition Q of the response space such that*

$$d_{\mathrm{R}}(P, P_Q) < \varepsilon.$$

The proof is relegated to Appendix A.

Remark 66 *One notable feature of the proof of Theorem 65 is that the partition Q provided by the theorem does not depend on the joint law η of P. More precisely, suppose P is a problem. For any probability measure $\eta' \in \mathrm{Prob}(X \times Y)$, we can create a new problem*

$$P_{\eta'} := (X, Y, \eta', \ell, H).$$

The proof of Theorem 65 shows that, not only can we find a partition Q of Y such that $d_{\mathrm{R}}(P, P_Q) < \varepsilon$, but we can choose Q such that for all $\eta' \in \mathrm{Prob}(X \times Y)$, we have $d_{\mathrm{R}}((P_{\eta'})_Q, P_{\eta'}) < \varepsilon$. We will use this stronger fact to prove Theorem 69.

Theorem 65 immediately gives us a density result.

Corollary 67 (Classification Density Theorem) *Under d_{R}, the set of classification problems is dense in the space of compact problems.*

5.4 Convergence of Empirical Problems

With the density of classification problems in hand, we are ready to examine the effect that limited sampling can have on a problem. One often does not have direct knowledge of the underlying joint law from which data is sampled; one only has access to a finite data set. In terms of our supervised learning problems, instead of accessing the joint law η of a problem P directly, we can only access the empirical measure η_n of η. Namely, given an i.i.d. set $\{(x_i, y_i)\}_{i=1}^n$ sampled from the probability measure η, we set

$$\eta_n := \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}.$$

The resulting problem with joint law η_n is called the *empirical problem*.

Definition 68 *If P is a problem, the nth empirical problem induced by P is the random problem defined by*
$$P_n := (X, Y, \eta_n, \ell, H).$$

Being dependent on a random sample, η_n is a random measure. Consequently, P_n is a random problem. One would hope that as the size of the sample grows, the empirical problem P_n converges to the true problem P with probability 1. Under certain conditions, this is indeed true.

Theorem 69 (Convergence of the Empirical Problem) *Let P be a problem satisfying the following:*

1. *H is compact with respect to $d_{\ell,\eta}$.*

2. *Y is compact.*

3. *ℓ is continuous.*

4. The collection $\{|\ell_h - \ell_{h'}|\}_{h,h' \in H}$ forms a Glivenko-Cantelli class with respect to η.

Let P_n be the nth empirical problem induced by P. Then $d_{\mathrm{R}}(P, P_n) \to 0$ almost surely.

The proof, found in Appendix A, uses Theorem 65 to reduce to the case where P is a classification problem. We briefly illuminate the Glivenko-Cantelli condition in Theorem 69 by providing simpler sufficient conditions.

Proposition 70 *Define $g_{h,h'}(x,y) := |\ell_h(x,y) - \ell_{h'}(x,y)|$. Theorem 69 still holds if Condition 4 is replaced with either of the following conditions.*

4a. The collection of subgraphs $\{\mathrm{sg}(g_{h,h'})|h, h' \in H\}$ (see Definition 12) has finite VC dimension.

4b. For any $h_0, h_0' \in H$ and $\varepsilon > 0$, there exists an open neighborhood $(h_0, h_0') \in U \subseteq H \times H$ such that

$$\int \sup_{(h,h') \in U} g_{h,h'}(x,y)\eta(dx \times dy) - \varepsilon < \int g_{h_0,h_0'}(x,y)\eta(dx \times dy)$$
$$< \int \inf_{(h,h') \in U} g_{h,h'}(x,y)\eta(dx \times dy) + \varepsilon.$$

Conditions 2 and 4a together imply condition 4, since the compactness of Y implies that the $g_{h,h'}$ are uniformly bounded, and hence that Proposition 13 applies. Note that Condition 4a does not depend on the joint law η. Conditions 1 and 4b together imply condition 4 as well by a theorem of Gaenssler (1975).

Theorem 69 is a convergence result. A quantitative result about the *rate* of convergence would be more useful. Certainly the convergence rate will depend on properties of P and P'; the proof of Theorem 69 suggests that the proximity of P and P' to less complex problems is crucial.

Question 71 *How does the rate of the convergence given by Theorem 69 depend on the properties of P and P'? Can one bound the convergence rate for specific classes of problems?*

6. The L^p-Risk Distance: Weighting Problems

In this section we introduce a variant of the Risk distance called the *L^p-Risk distance* which incorporates probability measures on the predictor sets. This additional information lets us prioritize certain predictors over others by weighting them more heavily, subsequently softening the Risk distance and improving its stability. The weighting can be interpreted as a Bayesian prior, so the perspective adopted in this section is broadly aligned with the principles of Bayesian statistics. While approaches such as Bayesian decision theory (Berger, 1985) incorporate elements like loss functions, prior work does not operate on a space of problems, nor does it attempt to endow such a space with a metric structure as we do in this paper.

6.1 Weighting Predictor Sets and the L^p-Risk Distance

Many of our results thus far have required assumptions of compact predictor sets or bounded loss functions. Without these assumptions, the Risk distance tends to be poorly-behaved. For instance, problems such as the linear regression in Example 19 tend to be infinitely far away from problems of bounded loss under the Risk distance. One reason for this phenomenon is the supremum in the definition of the Risk distance. The supremum treats unreasonable predictors, such as those whose graphs are far from the support of the joint law, with the same consideration as more reasonable options. One might reasonably complain about this design decision.

One solution to this problem is to *weight* the predictors, de-prioritizing extreme options by giving them less weight. These weights could come from prior knowledge or belief about which predictors might perform well or from a measure of complexity of each predictor (so that more complex predictors receive lower weights). Indeed, considering these weights as part of the specification of a problem can be seen as a proxy for adding a regularization term to the constrained Bayes risk.

More formally, we provide weights on the predictor set in the form of a probability measure. This lets us replace the supremum in the definition of d_{R} with an L^p-type integral, effectively softening the Risk distance.

Definition 72 *A* weighted problem *is a pair* (P, λ) *where*

- *P is a problem whose predictor set H comes equipped with a Polish topology, and*

- *λ is a probability measure on H.*

When there is no chance of confusion, we may suppress the measure λ and denote the weighted problem by P.

Assumption 73 *We make the following measurability assumption. Given a weighted problem (P, λ), consider the map*

$$H \times X \times Y \to \mathbb{R}_{\geqslant 0}$$

given by

$$(h, x, y) \mapsto \ell_h(x, y).$$

Since H, X, and Y are topological spaces, we can equip each space with the corresponding Borel σ-algebra. We can then equip the product $H \times X \times Y$ with the product σ-algebra. For any weighted problem (P, λ), we henceforth assume that the above map $(h, x, y) \mapsto \ell_h(x, y)$ is measurable with respect to this product σ-algebra.

Definition 74 *Let (P, λ) and (P', λ') be weighted problems. For $p \in [1, \infty]$ and for any couplings $\gamma \in \Pi(\eta, \eta')$ and $\rho \in \Pi(\lambda, \lambda')$, define the L^p-risk distortion of γ and ρ to be*

$$\mathrm{dis}_{P,P',p}(\rho, \gamma) := \left(\int \left(\int \left| \ell(h(x), y) - \ell'(h'(x'), y') \right| d\gamma(x, y, x', y') \right)^p d\rho(h, h') \right)^{1/p} \quad p \in [1, \infty),$$

$$\mathrm{dis}_{P,P',\infty}(\rho, \gamma) := \sup_{(h,h') \in \mathrm{supp}[\rho]} \int \left| \ell(h(x), y) - \ell'(h'(x'), y') \right| d\gamma(x, y, x', y') \qquad p = \infty.$$

The L^p-Risk distance *between weighted problems* (P, λ) *and* (P', λ') *is then defined to be*

$$d_{\mathrm{R},p}((P, \lambda), (P', \lambda')) := \inf_{\rho, \gamma} \mathrm{dis}_{P, P', p}(\rho, \gamma)$$

where γ *ranges over all couplings* $\Pi(\eta, \eta')$ *and* ρ *ranges over all couplings* $\Pi(\lambda, \lambda')$.

Theorem 75 *The L^p-Risk distance is a pseudometric on the collection of all weighted problems. Furthermore, we have*

(i) $d_{\mathrm{R},p} \leqslant d_{\mathrm{R},q} \leqslant d_{\mathrm{R},\infty}$ *for all* $1 \leqslant p \leqslant q \leqslant \infty$;

(ii) $d_{\mathrm{R}} \leqslant d_{\mathrm{R},\infty}$.

Remark 76 *Property (i) is analogous to a similar property enjoyed by the Gromov-Wasserstein distance (Mémoli, 2011, Theorem 5.1 (h)) and it follows from (Mémoli, 2011, Lemma 2.2). Property (ii) states that the L^∞-Risk distance is never smaller than the standard Risk distance from Definition 26. This is analogous to the fact that the Gromov-Hausdorff distance is bounded above by the L^∞-Gromov-Wasserstein distance (Mémoli, 2011, Theorem 5.1 (b)). The proof of the triangle inequality for the L^p-Risk distance is identical to the proof of the same for the Risk distance, save for an application of Minkowski's inequality for the outer integral. As indicated above, the proof of the other claimed properties follows similar steps as those for the Gromov-Wasserstein distance and therefore we omit them.*

Example 77 *Just as we defined the one-point problems*

$$P_\bullet(c) := (\{\bullet\}, \{\bullet\}, \delta_{(\bullet, \bullet)}, c, \{\mathrm{Id}_{\{\bullet\}}\})$$

for each $c \geqslant 0$ in Example 20, we can define a corresponding one-point weighted problem *by endowing $\{\mathrm{Id}_{\{\bullet\}}\}$ with the point-mass probability measure. We will still use $P_\bullet$ to denote $P_\bullet(0)$. Then, in a manner similar to Example 31, if (P, λ) is a weighted problem, one can show*

$$d_{\mathrm{R},p}(P, P_\bullet) = \left(\int_H \left(\int_{X \times Y} \ell(h(x), y) \, \eta(dx \times dy) \right)^p \lambda(dh) \right)^{1/p} = \left(\int_H \mathcal{R}_P^p(h) \, \lambda(dh) \right)^{1/p}.$$

This can be seen as a kind of weighted constrained Bayes risk of (P, λ). While the constrained Bayes risk of an unweighted problem P is the infimal possible risk, the above expression $d_{\mathrm{R},p}(P, P_\bullet)$ represents the pth moment of possible risks when a predictor is chosen at random according to λ. In particular, $d_{\mathrm{R},1}(P, P_\bullet)$ is simply the average risk of a predictor chosen randomly according to λ.

One can establish a connection with the usual constrained Bayes risk as follows. Assume that the problem P is such that its loss function ℓ is bounded above by a constant $\ell_{\max} > 0$. Then,

$$d_{\mathrm{R},p}(P, P_\bullet(\ell_{\max})) = \left(\int_H \left(\ell_{\max} - \mathcal{R}_P(h) \right)^p \lambda(dh) \right)^{1/p}.$$

If the support of λ is the whole H, we obtain, as $p \to \infty$, that $d_{\mathrm{R},p}(P, P_\bullet(\ell_{\max})) \to \ell_{\max} - \mathcal{B}(P)$.

The distance $d_{\mathrm{R},p}(P, P_\bullet)$ from Example 77 will appear again in Theorem 82, so we establish a definition.

Definition 78 *Given a weighted problem $P = (P, \lambda)$ and $1 \leqslant p < \infty$, the p-diameter of P is given by*

$$d_{\mathrm{R},p}(P, P_\bullet) = \left(\int_H \left(\int_{X \times Y} \ell(h(x), y) \, \eta(dx \times dy) \right)^p \lambda(dh) \right)^{1/p} = \left(\int_H \mathcal{R}_P^p(h) \, \lambda(dh) \right)^{1/p}.$$

The word "diameter" is chosen in analogy with metric geometry, where the diameter of a metric space can be found from its Gromov-Hausdorff distance from the one-point metric space.

6.2 Connection to the Gromov-Wasserstein Distance

In Section 3.5, we saw two ways to turn metric measure spaces into problems, each of which relates the Gromov-Wasserstein distance to the Risk distance. In this section, we will draw a similar connection between the Gromov-Wasserstein distance and the L^p-Risk distance. Specifically, we will provide a way of turning metric measure spaces into weighted problems such that the L^1-Risk distance between the resulting problems aligns with a well-known modification of the Gromov-Wasserstein distance between the original spaces.

Given any metric measure space (X, d_X, μ_X), define a weighted problem (P_X, μ_X),

$$P_X := (X, X, (\Delta_X)_\sharp \mu_X, d_X, H_X),$$

where

- $\Delta_X : X \to X \times X$ is the diagonal map $x \mapsto (x, x)$, and

- H_X is the set of all constant functions $X \to X$.

Here we think of μ_X as a measure on both X and H_X since the two are in natural bijection.

Proposition 79 *Let (X, d_X, μ_X) and (Y, d_Y, μ_Y) be two metric measure spaces. Then,*

$$d_{\mathrm{R},1}(P_X, P_Y) = \inf_{\substack{\gamma \in \Pi(\mu_X, \mu_Y) \\ \rho \in \Pi(\mu_X, \mu_Y)}} \int \int \left| d_X(x_2, x_1) - d_Y(y_2, y_1) \right| \gamma(dx_1 \times dy_1) \, \rho(dx_2 \times dy_2). \quad (\star)$$

The proof of this proposition can be found in Appendix A.

Use $\mathcal{F}(\gamma, \rho)$ to denote the double integral in the expression $(\star)$. If we add the additional constraint that $\gamma = \rho$, then the expression $(\star)$ becomes $2d_{\mathrm{GW},1}(X, Y)$. As written, $(\star)$ is a lower bound on $2d_{\mathrm{GW},1}(X, Y)$, which could be arrived at via a "bilinear relaxation" of the optimization problem posed by $d_{\mathrm{GW},1}$. That is, one could begin with the objective function for the Gromov-Wasserstein optimization problem

$$\mathcal{F}'(\gamma) := \int \int \left| d_X(x_2, x_1) - d_Y(y_2, y_1) \right| \gamma(dx_1 \times dy_1) \, \gamma(dx_2 \times dy_2),$$

53

which is quadratic in the transport plan γ, and decouple the two instances of γ to get the objective $\mathcal{F}(\gamma, \rho)$, which is bilinear in two transport plans. This "relaxation" was considered by (Mémoli, 2011, Section 7) for estimating the Gromov-Wasserstein distance via an alternate minimization procedure. Similar optimal transport problems, which simultaneously optimize two transport plans in a Gromov-Wasserstein-like objective function, appear in the literature. Indeed, when X and Y are finite, the expression $(\star)$ is a special case of CO-Optimal transport (Vayer et al., 2020b). Sejourne et al. (2021) apply this relaxation technique, which they call the "biconvex relaxation," to an unbalanced variant of the Gromov-Wasserstein distance. This inspired Beier et al. (2023) to explore a similar relaxation for a "multi-marginal" variant of the same problem. A theoretical study of the metric properties of the right-hand side of $(\star)$ appears in (Chen et al., 2022, Appendix A.4).

6.3 Stability of Loss Profile Distribution

Being softer than the Risk distance, the loss profile is less stable under the L^p-Risk distance. To obtain a stability result, one cannot compare the loss profile sets directly, but instead compare certain probability distributions on the loss profile sets.

Definition 80 *Let (P, λ) be a weighted problem. Define $F_P : H \to \mathrm{Prob}(\mathbb{R})$ by*

$$F_P(h) := (\ell_h)_\sharp \eta.$$

The loss profile distribution *of P is then defined to be $\mathcal{L}(P) := (F_P)_\sharp \lambda$.*

Note that $\mathcal{L}(P) \in \mathrm{Prob}(\mathrm{Prob}(\mathbb{R}))$ is a probability measure on the space of real probability measures. More specifically, $F_P(h)$ is the loss profile of h, so $\mathcal{L}(P)$ describes what the loss profile of a random predictor chosen according to λ is likely to be.

While Theorem 37 shows that $L(P)$ is stable under the Risk distance with respect to the Hausdorff distance, the following analogous theorem shows that the measure $\mathcal{L}(P)$ is stable under the L^p-Risk distance with respect to the Wasserstein distance.

Theorem 81 *Let P and P' be weighted problems. Then*

$$d_{\mathrm{W},p}^{d_{\mathrm{W},1}}(\mathcal{L}(P), \mathcal{L}(P')) \leqslant d_{\mathrm{R},p}(P, P')$$

where $d_{\mathrm{W},p}^{d_{\mathrm{W},1}}$ is the p-Wasserstein metric on $\mathrm{Prob}(\mathrm{Prob}(\mathbb{R}))$ with underlying metric $d_{\mathrm{W},1}$ on $\mathrm{Prob}(\mathbb{R})$.

The proof is in Appendix A.

6.4 Improved Density

A useful theorem in measure theory states that any probability measure μ on a Polish space X is *tight*, meaning that for any $\varepsilon > 0$, there is a compact subset $K \subseteq X$ with $\mu(K) > 1 - \varepsilon$. In a slogan, "Polish probability spaces are almost compact." Applying this principle to our setting suggests that any weighted problem should be close to one with a compact predictor set. This can be made rigorous with a density theorem.

Theorem 82 *Let $p \geqslant 1$ and let (P, λ) be a weighted problem of finite p-diameter. Then there exists a compact weighted problem (P', λ') with $d_{\mathrm{R},p}(P, P') < \varepsilon$.*

Additionally, a result analogous to Theorem 65 still holds for the L^p-Risk distance.

Proposition 83 *Let $p \geqslant 1$ and let (P, λ) be a compact weighted problem. Then there exists a weighted classification problem (P', λ') with $d_{\mathrm{R},p}(P, P') < \varepsilon$.*

The proofs of both Theorem 82 and Proposition 83 can be found in Appendix A. Chaining these two density results together proves the following corollary.

Corollary 84 *Under $d_{\mathrm{R},p}$ for $p \geqslant 1$, weighted classification problems are dense in the space of weighted problems of finite p-diameter.*

Example 85 *Consider the linear regression problem P from Example 19. If we endow the predictor set of P with a joint law satisfying some finite moment assumptions, the resulting weighted problem has finite p-diameter. Corollary 84 then tells us that P can be approximated arbitrarily well by weighted classification problems under $d_{\mathrm{R},p}$.*

More precisely, let P be a linear regression problem, and suppose the predictor set

$$H = \left\{ x \mapsto a^T x + b \,\middle|\, (a, b) \in \mathbb{R}^n \times \mathbb{R} \right\}$$

is equipped with a probability measure λ. We can think of both λ and η as probability measures on $\mathbb{R}^n \times \mathbb{R}$. If

1. *all $n + 1$ marginals of η have finite second moment, and*

2. *all $n + 1$ marginals of λ have finite $2p$-th moment,*

then $\operatorname{diam}_p(P) < \infty$.

To prove this, we can write

$$\left(\operatorname{diam}_p(P) \right)^{1/2} = \left(\int \left(\int \ell(a^T x + b, y) \, \eta(dx \times dy) \right)^p \lambda(da \times db) \right)^{1/2p}$$

$$= \left(\int \left(\left(\int (a^T x + b - y)^2 \, \eta(dx \times dy) \right)^{1/2} \right)^{2p} \lambda(da \times db) \right)^{1/2p}.$$

By liberally applying Minkowski's inequality and simplifying, we can bound this expression by the sum

$$\sum_{i=1}^n \left(\int a_i^{2p} \, \lambda(da \times db) \right)^{1/2p} \left(\int x_i^2 \, \eta(dx \times dy) \right)^{1/2}$$

$$+ \left(\int (b)^{2p} \, \lambda(da \times db) \right)^{1/2p} + \left(\int y^2 \, \eta(dx \times dy) \right)^{1/2}.$$

These integrals are the moments of η and λ that we have assumed to be finite.

55

6.5 Improved Stability

The L^p-Risk distance enjoys many of the same stability and density results as the Risk distance. Indeed, we will soon see that many of these results can be *strengthened* for the L^p-Risk distance. That is, the L^p-Risk distance enjoys *stronger* stability and density properties than the Risk distance.

First we will establish that the L^p-Risk distance inherits some stability results from the Risk distance, saving the improvements for later. Many of our bounds on the Risk distance are proven using the bound

$$d_{\mathrm{R}}(P, P') \leqslant \inf_{\gamma \in \Pi(\eta, \eta')} \mathrm{dis}_{P,P'}(R_{\mathrm{diag}}, \gamma) \tag{2}$$

where R_{diag} is the diagonal correspondence, which is valid whenever P and P' share a predictor set. Hence we can extend the results of those sections using the following lemma.

Lemma 86 *Let (P, λ) and (P', λ) be weighted problems that share a predictor set H and predictor weights λ. Then*

$$d_{\mathrm{R},p}((P, \lambda), (P', \lambda)) \leqslant \inf_{\gamma \in \Pi(\eta, \eta')} \mathrm{dis}_{P,P'}(R_{\mathrm{diag}}, \gamma)$$

where R_{diag} is the diagonal correspondence of H with itself.

Proof First write

$$d_{\mathrm{R},p}((P, \lambda), (P', \lambda)) \leqslant \mathrm{dis}_{P,P',p}(\rho_{\mathrm{diag}}, \gamma)$$

where ρ_{diag} is the diagonal coupling of λ with itself.

Next we bound the L^p-risk distortion by the L^∞-risk distortion as follows:

$$\begin{aligned}
\mathrm{dis}_{P,P',p}(\rho_{\mathrm{diag}}, \gamma) &\leqslant \mathrm{dis}_{P,P',\infty}(\rho_{\mathrm{diag}}, \gamma) \\
&= \sup_{(h,h')\in\mathrm{supp}[\rho_{\mathrm{diag}}]} \int \left| \ell(h(x), y) - \ell'(h'(x'), y') \right| d\gamma(x, y, x', y') \\
&\leqslant \sup_{(h,h')\in R_{\mathrm{diag}}} \int \left| \ell(h(x), y) - \ell'(h'(x'), y') \right| d\gamma(x, y, x', y') \\
&= \mathrm{dis}_{P,P'}(R_{\mathrm{diag}}, \gamma),
\end{aligned}$$

where the second inequality follows from the fact that the support of ρ_{diag} lies within the diagonal $R_{\mathrm{diag}} \subseteq H \times H$. ∎

Condition 2 from Theorem 28, the Risk distance bounds of Sections 4.3 (Stability under bias), 4.4 (Stability under noise), and 5.4 (Convergence of empirical problems) are all proven via the diagonal bound (2).[6] By Lemma 86, these results still hold if we replace the Risk distance with the L^p-Risk distance and replace both problems with weighted problems that share a common weight measure λ. That is, in the above results, we can replace each instance of $d_{\mathrm{R}}(P, P')$ with $d_{\mathrm{R},p}((P, \lambda), (P', \lambda))$ and they will still hold true.

---

6. In addition to the diagonal bound, the proof of Theorem 69 also uses the density of classification problems (Theorem 67). However, the proof of the analogous result for weighted problems is still valid if we replace this step with an application of the density of weighted classification problems (Theorem 83).

Moreover, some of our results for the Risk distance can be strengthened for the L^p-Risk distance. Indeed, we have already seen that the density results of Section 5.3 can be strengthened into Theorem 82. We can strengthen some stability results in the same manner. For instance, the L^p-Risk distance satisfies the following version of Condition 2 in which the supremum is replaced by an integral.

Proposition 87 *Let (P, λ) and (P', λ) be weighted problems with*

$$P = (X, Y, \eta, \ell, H)$$
$$P' = (X, Y, \eta, \ell', H)$$

Then

$$d_{\mathrm{R},p}(P, P') \leqslant \left(\int \left(\int \left| \ell(h(x), y) - \ell'(h(x), y) \right| \eta(dx \times dy) \right)^p \lambda(dh) \right)^{1/p}.$$

Proof Bound the L^p-Risk distance by selecting the diagonal couplings. ∎

We can also modify Theorem 51 by replacing the Hausdorff distance with a Wasserstein distance.

Proposition 88 *Let (P, λ) and (P, λ') be weighted problems with*

$$P := (X, Y, \eta, \ell, H)$$
$$P' := (X, Y, \eta, \ell, H').$$

Let $i : H \to H \cup H'$ and $j : H' \to H \cup H'$ be the inclusion maps, and endow $H \cup H'$ with the pseudometric $d_{\ell,\eta}(h_1, h_2) := \|\ell_{h_1} - \ell_{h_2}\|_{L^1(\eta)}$ introduced in Section 5.2.

Then

$$d_{\mathrm{R},p}(P, P') \leqslant d_{\mathrm{W},p}^{d_{\ell,\eta}}(i_\sharp \lambda, j_\sharp \lambda').$$

Proof Selecting the diagonal coupling between η and itself gives us the bound

$$d_{\mathrm{R},p}(P, P') \leqslant \inf_{\rho \in \Pi(\lambda, \lambda')} \left(\int \left(\int \left| \ell_h(x, y) - \ell_{h'}(x, y) \right| \eta(dx \times dy) \right)^p \rho(dh \times dh') \right)^{1/p}$$

$$= \inf_{\rho \in \Pi(i_\sharp \lambda, j_\sharp \lambda')} \left(\int \left(\int \left| \ell_h(x, y) - \ell_{h'}(x, y) \right| \eta(dx \times dy) \right)^p \rho(dh \times dh') \right)^{1/p}$$

$$= \inf_{\rho \in \Pi(i_\sharp \lambda, j_\sharp \lambda')} \left(\int \left(d_{\ell,\eta}(h, h') \right)^p \rho(dh \times dh') \right)^{1/p} = d_{\mathrm{W},p}^{d_{\ell,\eta}}(i_\sharp \lambda, j_\sharp \lambda')$$

as desired. ∎

Proposition 44 can be modified for weighted problems as well. We first construct an analog of the metric $s_{\ell,H}$ from Section 4.3. Replacing the supremum over H in the definition of $s_{\ell,H}$ with an L^p-style integral against λ, we arrive at the definition

$$s_{\ell,\lambda,p}((x, y), (x', y')) := \left(\int_H \left| \ell(h(x), y) - \ell(h(x'), y') \right|^p \lambda(dh) \right)^{1/p}.$$

57

In other words, to determine the distance between (x, y) and (x', y'), we ask how different the loss on the observation (x, y) will be from the loss on (x', y') for an average predictor. We can now state our stability result.

Proposition 89 *Let (P, λ) be a weighted problem, and let (P', λ) be a weighted problem that is identical to P except possibly for its joint law η'. Then*

$$d_{\mathrm{R},p}(P, P') \leqslant d_{\mathrm{W},1}^{s_{\ell,\lambda,p}}(\eta, \eta').$$

Proof By selecting the diagonal coupling between λ and itself and applying Minkowski's inequality for integrals, we can bound $d_{\mathrm{R},p}$ by

$$d_{\mathrm{R},p}(P, P') \leqslant \inf_{\gamma \in \Pi(\eta, \eta')} \left(\int \left(\int |\ell(h(x), y) - \ell(h(x'), y')| \, \gamma(dx \times dy \times dx' \times dy') \right)^p \lambda(dh) \right)^{1/p}$$

$$\leqslant \inf_{\gamma \in \Pi(\eta, \eta')} \int \left(\int |\ell(h(x), y) - \ell(h(x'), y')|^p \, \lambda(dh) \right)^{1/p} \gamma(dx \times dy \times dx' \times dy')$$

$$= d_{\mathrm{W},1}^{s_{\ell,\lambda,p}}(\eta, \eta'),$$

giving us the desired bound. ∎

Lastly, we turn our attention to stability under noise. We have shown in Section 4.4 that the Risk distance is stable with respect to noise in the joint law when ℓ is a metric. By contrast, for a general ℓ, the Risk distance d_{R} can be highly sensitive to changes in the joint law. The following example demonstrates this sensitivity.

Example 90 *Define a problem*

$$P = (\{\bullet\}, \mathbb{R}, \delta_0, \ell, H)$$

where $\ell(y, y') = (y - y')^2$, and H is the collection of all functions $\{\bullet\} \to \mathbb{R}$, making H itself homeomorphic to $\mathbb{R}$.[7] Now, for any $\varepsilon > 0$, define

$$P_\varepsilon = \left(\{\bullet\}, \mathbb{R}, \frac{1}{2}(\delta_0 + \delta_\varepsilon), \ell, H \right).$$

We can think of P_ε as P with some noise; when making an observation, there is a 50% of observing the correct y-value of 0, and a 50% chance of making a small error and reading a y-value of ε. We would hope that P_ε would converge to P under d_{R} as $\varepsilon \to 0$, but this is not true. There is only one coupling between δ_0 and $1/2(\delta_0 + \delta_\varepsilon)$, so the Risk distance is

$$d_{\mathrm{R}}(P, P_\varepsilon) = \inf_{R \in \mathcal{C}(H, H)} \sup_{(a,b) \in R} \int \int |(a - y)^2 - (b - y')^2| \, \delta_0(dy) \left(\frac{1}{2}\delta_0 + \frac{1}{2}\delta_\varepsilon \right)(dy')$$

$$= \inf_{R \in \mathcal{C}(H, H)} \sup_{(a,b) \in R} \left[\frac{1}{2}|a^2 - b^2| + \frac{1}{2}|a^2 - (b - \varepsilon)^2| \right]$$

---

7. In other words, we can identify H with $\mathbb{R}$.

$$\geqslant \frac{1}{2} \inf_{R \in \mathcal{C}(H,H)} \sup_{(a,b) \in R} \left| b^2 - (b-\varepsilon)^2 \right|$$

$$= \frac{1}{2} \sup_{b \in \mathbb{R}} \left| 2b\varepsilon - \varepsilon^2 \right| = \infty.$$

The metric $d_{\mathrm{R},p}$ enjoys a stronger stability than d_{R} with respect to noise in a problem's joint law. We use a more general noise model than the one outlined in Section 4.4, allowing for noise in the X direction as well as the Y direction. More specifically, we let our noise be given by any Markov kernel $N : X \times Y \to \mathrm{Prob}(X \times Y)$. That is, given a problem P, define P_N to be the same problem but with joint law on $X \times Y$ given by

$$\eta \cdot N(A) := \int_{X \times Y} N(x,y)(A) \, \eta(dx \times dy)$$

for all measurable $A \subseteq X \times Y$. We justify the use of Markov kernel composition notation by thinking of the measure η as a Markov kernel $\eta : \{\bullet\} \to \mathrm{Prob}(X \times Y)$.

We can now state our improved result about stability under noise.

Theorem 91 *Let (P, λ) be a weighted problem, and $N : X \times Y \to \mathrm{Prob}(X \times Y)$ a Markov kernel. Then*

$$d_{\mathrm{R},p}(P, P_N) \leqslant d_{\mathrm{W},1}^{s_{\ell,\lambda,p}}(\delta_{\mathrm{Id}_{X \times Y}}, N).$$

Here $\delta_{\mathrm{Id}_{X \times Y}}$ is the Markov kernel sending (x,y) to $\delta_{(x,y)}$, and $d_{\mathrm{W},1}^{s_{\ell,\lambda,p}}$ is the Wasserstein metric on Markov kernels $X \times Y \to \mathrm{Prob}(X \times Y)$ when $X \times Y$ is equipped with the joint law η and the metric $s_{\ell,\lambda,p}$.

The proof can be found in Appendix A.

Example 92 *Let (P, λ) be a weighted problem with $P = (\mathbb{R}, \mathbb{R}, \eta, \ell, H)$, where*

- *$\ell(y, y') := (y - y')^2$,*

- *$H := \{f_a | a \in \mathbb{R}\}$, where $f_a(x) := ax$,*

- *λ is the standard normal distribution.*

Let $N : \mathbb{R} \times \mathbb{R} \to \mathrm{Prob}(\mathbb{R} \times \mathbb{R})$ be a noise kernel that applies x-dependent noise in the vertical direction. That is, if $i_x^y : \mathbb{R} \to \mathbb{R}^2$ is the function $i_x^y(y') = (x, y + y')$, set

$$N(x, y) = (i_x^y)_\sharp M(x)$$

for some Markov kernel $M : \mathbb{R} \to \mathbb{R}$. Assume that for all x, $M(x)$ is a symmetric measure with finite 2nd moment $v(x)$ and 4th moment $k(x)$. Theorem 91 gives us $d_{R,2}(P, P_N) \leqslant d_{\mathrm{W},1}^{s_{\ell,\lambda,2}}(\delta_{\mathrm{Id}_{\mathbb{R}^2}}, N)$. The right hand side is difficult to exactly compute, so we will compute the larger quantity $d_{\mathrm{W},2}^{s_{\ell,\lambda,2}}(\delta_{\mathrm{Id}_{\mathbb{R}^2}}, N)$.

We first produce an explicit formula for $s_{\ell,\lambda,2}((x, y), (x', y'))$.

$$s_{\ell,\lambda,2}^2((x, y), (x', y')) = \int \big(\ell_h(x, y) - \ell_h(x', y')\big)^2 \, \lambda(dh)$$

$$= \int_{-\infty}^{\infty} \big((ax - y)^2 - (ax' - y')^2\big)^2 \frac{e^{-a^2/2}}{\sqrt{2\pi}} da.$$

After expanding the polynomial term, the integral can be split into five terms that represent various moments of the standard normal distribution. We arrive at the polynomial

$$s_{\ell,\lambda,2}^2((x,y),(x',y')) = 3(x^2 - x'^2)^2 + 4(xy - x'y')^2 + (y^2 - y'^2)^2 + 2(x^2 - x'^2)(y^2 - y'^2).$$

Next we can compute $d_{W,2}^{s_{\ell,\lambda,2}}(\delta_{(x,y)}, N(x,y))$ for any fixed $(x,y) \in \mathbb{R}^2$. Indeed, there is only one coupling between $\delta_{(x,y)}$ and $N(x,y)$, so we can write

$$\left(d_{W,2}^{s_{\ell,\lambda,2}}\right)^2 (\delta_{(x,y)}, N(x,y)) = \int s_{\ell,\lambda,2}^2((x,y),(x',y'))\, N(x,y)(dx' \times dy')$$

$$= \int s_{\ell,\lambda,2}^2((x,y),(x',y'))\, (i_x^y)_\sharp M(x)(dx' \times dy')$$

$$= \int s_{\ell,\lambda,2}^2((x,y),(x,y'+y))\, M(x)(dy')$$

$$= \int 4x^2 y'^2 + y'^2(y'+2y)^2\, M(x)(dy').$$

Again, expanding the polynomial terms and splitting the integral reduces the problem to moment computations. Recalling that $M(x)$ is symmetric and hence has vanishing odd moments, we arrive at the expression

$$\left(d_{W,2}^{s_{\ell,\lambda,2}}\right)^2 (\delta_{(x,y)}, N(x,y)) = 4v(x)x^2 + 4v(x)y^2 + k(x).$$

Integrating against η gives us an expression for $\left(d_{W,2}^{s_{\ell,\lambda,2}}\right)^2 (\delta_{\mathrm{Id}_{\mathbb{R}^2}}, N)$, and hence an upper bound for $d_{R,p}^2(P, P_N)$.

$$d_{R,p}^2(P, P_N) \leqslant \int \left[4v(x)\left(\int y^2\, \beta(x)(dy)\right) + 4x^2 v(x) + k(x) \right] \alpha(dx)$$

where α is the first marginal of η, and β is the disintegration. We take a moment to interpret the integrand of the outer integral. The integral $\int y^2 \beta(x)(dy)$ is the second moment of $\beta(x)$. Hence the term $v(x)\int y^2 \beta(x)(dy)$ measures the correlation between the spread of $\beta(x)$ and the noise $\sigma^2(x)$ applied at x. The term $x^2\sigma^2(x)$ measures the application of noise far from the origin. Hence this bound suggests that noise has an effect on the squared Risk distance proportional to its 4th moment, with an additional effect proportional to its 2nd moment if applied at x-values far from zero, or where $\beta(x)$ is already spread out.

7. The Connected Risk Distance: Risk Landscapes

Having discussed the weighted variant of the Risk distance, we now introduce and study one more variant, called the *Connected Risk distance*. This variant is motivated by the observation that the Risk distance is insensitive to the contours of a problem's risk landscape. We will show that by strengthening the Risk distance into the Connected Risk distance, a topological descriptor of the risk landscape called the *Reeb graph* can be made to exhibit stability under the Connected Risk distance without breaking many of the stability results that the Risk distance enjoys. The Reeb graph of the risk landscape of a problem P encodes information about the inherent *complexity* associated to solving P.

7.1 Insensitivities of the Risk distance

We saw in section 4.1 that the constrained Bayes risk of a problem is stable under the Risk distance. The constrained Bayes risk $\mathcal{B}(P)$ of a problem P is a sufficiently descriptive invariant if we plan to solve P by exhaustively searching the predictor set H for a predictor with minimal risk. If we instead wish to use a more practical search method like gradient descent, then $\mathcal{B}(P)$ is a woefully incomplete descriptor. The specific structure of the risk function $\mathcal{R}_P : H \to \mathbb{R}_{\geqslant 0}$ becomes crucially important, rather than just its infimum. Is $\mathcal{R}_P$ rife with local minima? What are the risks of the local minima? What are their basins like? The Risk distance is not sensitive to these kinds of concerns.

Example 93 *Define a problem*

$$P := ([0,1], \{0,1\}, \eta, \ell, H)$$

and a family of problems

$$P_t := ([0,1], \{0,1\}, \eta, \ell, H_t), t \in (0,1],$$

where

1. *ℓ is the 0-1 loss $\ell(y, y') := 1_{y \neq y'}$.*

2. *η is the uniform distribution on $[0,1] \times \{0\}$.*

3. *$H := \left\{ 1_{[0,a)} \middle| a \in [0,1] \right\} \cup \left\{ 1_{(a,1]} \middle| a \in [0,1] \right\}$.*

4. *$H_t := \left\{ 1_{[0,a)} \middle| a \in [t,1] \right\} \cup \left\{ 1_{(a,1]} \middle| a \in [0,1] \right\}$.*

Here we take the convention that $[0,0) = (1,1] = \varnothing$. We claim that $d_{\mathrm{R}}(P, P_t) \leqslant t$. Indeed, by Theorem 51,

$$
\begin{aligned}
d_{\mathrm{R}}(P, P_t) &\leqslant d_{\mathrm{H}}^{L^1(\eta)}\left(\{\ell_h | h \in H\}, \{\ell_{h'} | h' \in H_t\} \right) \\
&= \sup_{h \in H} \inf_{h' \in H_t} \|\ell_h - \ell_{h'}\|_{L^1(\eta)} \\
&= \sup_{h \in H \setminus H_t} \inf_{h' \in H_t} \int_0^1 \left| \ell(h(x), 0) - \ell(h'(x), 0) \right| dx \\
&= \sup_{a \in [0,t)} \inf_{h' \in H_t} \int_0^1 \left| 1_{1_{[0,a)}(x) \neq 0} - 1_{h'(x) \neq 0} \right| dx \\
&= \sup_{a \in [0,t)} \inf_{h' \in H_t} \int_0^1 \left| 1_{[0,a)}(x) - h'(x) \right| dx.
\end{aligned}
$$

Choosing $h' = 0$ gives us the upper bound bound

$$\leqslant \sup_{a \in [0,t)} \int_0^1 1_{[0,a)}(x) \, dx = t.$$

Hence $d_{\mathrm{R}}(P_t, P) \to 0$ as $t \to 0$. At the same time, the risk landscapes of P and P_t look very different for any $t > 0$. These risk landscapes are depicted in Figure 6. Under the metric
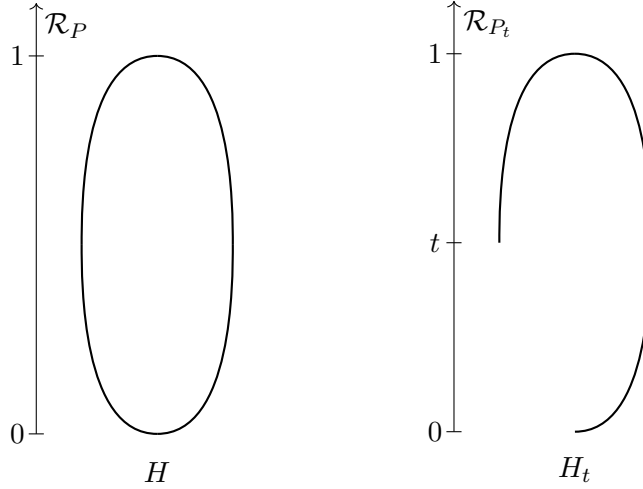
Figure 6: The risk landscapes for P and P_t in Example 93.

$d_{\ell,\eta}$, H is topologically a circle. The risk function $\mathcal{R}_P : H \to \mathbb{R}_{\geqslant 0}$ has one local minimum, which is achieved at the function $h(x) \equiv 0$. For $t > 0$, H_t is formed from H by removing a small segment of the circle, making H_t topologically a closed interval under $d_{\ell,\eta}$. Then $\mathcal{R}_{P_t}$ has a local minimum at each end of the endpoints of the interval, one at $h(x) = 0$ and the other at $h(x) = 1_{[0,t)}$.

In this section, we strengthen the Risk distance to exert more control over the risk landscape. Inspired by Bauer et al. (2021), we restrict the correspondences in the definition of d_{R} to only those satisfying a certain connectivity property.

7.2 The Connected Risk distance

If H, H' are topological spaces, define $\mathcal{C}_{\mathrm{con}}(H, H')$ to be the set of correspondences $R \subseteq H \times H'$ such that the projections $R \to H$ and $R \to H'$ are *inverse connected*, meaning the preimage of any connected set is connected.

Definition 94 (Connected Risk distance)

$$d_{\mathrm{RC}}(P, P') := \inf_{\substack{\gamma \in \Pi(\eta, \eta') \\ R \in \mathcal{C}_{\mathrm{con}}(H, H')}} \mathrm{dis}_{P,P'}(R, \gamma)$$

$$= \inf_{\substack{\gamma \in \Pi(\eta, \eta') \\ R \in \mathcal{C}_{\mathrm{con}}(H, H')}} \sup_{(h, h') \in R} \int \left| \ell_h(x, y) - \ell'_{h'}(x', y') \right| \gamma(dx \times dy \times dx' \times dy').$$

Note that the definition of d_{RC} is the same as that of d_{R} (Definition 26), except that the correspondence R is chosen from $\mathcal{C}_{\mathrm{con}}(H, H')$. We claim d_{RC} is a pseudometric. The only nontrivial property is the triangle inequality, which requires a modified version of the gluing lemma.

Lemma 95 *Let H_1, H_2, H_3 be topological spaces, and let $R_{1,2} \in \mathcal{C}_{\mathrm{con}}(H_1, H_2)$, $R_{2,3} \in \mathcal{C}_{\mathrm{con}}(H_2, H_3)$. Then there exists an $R \subseteq H_1 \times H_2 \times H_3$ such that the projection maps*

onto each of the three coordinates are all inverse connected, and such that $\mathrm{proj}_{1,2}(R) = R_{1,2}$ *and* $\mathrm{proj}_{2,3}(R) = R_{2,3}$.

With this lemma in hand, we can prove the triangle inequality in exactly the same way that we did for d_{R}. The proof of Lemma 95 can be found in Appendix A.

Remark 96 *Note that d_{RC} is a larger metric than d_{R}, since R is infimized over $\mathcal{C}_{\mathrm{con}}(H_1, H_2) \subseteq \mathcal{C}(H_1, H_2)$. Therefore we would expect d_{RC} to be less stable than d_{R}. Some of the stability results of d_{R} were proven by selecting R to be the diagonal coupling between a predictor set H and itself. Since $\mathcal{C}_{\mathrm{con}}(H, H)$ still contains the diagonal coupling, these proofs apply just as well to d_{RC}, just as we saw with the L^p-Risk distance in Section 6.5. Specifically, Condition 2 from Section 3.3, as well as the results of Sections 4.3 (stability under bias) and 4.4 (stability under noise) still apply to the Connected Risk distance.*

Additionally, since d_{RC} is larger than d_{R}, the problem descriptors stable under the former are also stable under the latter. In particular, the results of Section 4.1 and Section 4.2 still hold for d_{RC}; still stable under the Connected Risk distance are the loss profile set and consequently the Bayes risk, as well as the Rademacher complexity for an arbitrarily large sample size.

7.3 Risk Landscapes and Reeb Graphs

We will show that a certain topological summary of the risk landscape of a problem, known as the *Reeb graph*, is stable under the Connected Risk distance.

Definition 97 (Induced Reeb Graph) *Let X be a topological space and $f : X \to \mathbb{R}$ a continuous function. Define an equivalence relation on X by declaring that $x \sim y$ if and only if x and y belong to the same connected component of some level set of f. Define $F := X/\!\sim$. Let $q : F \to \mathbb{R}$ be the unique continuous function that factors through f. The pair (F, q) is called the* Reeb graph *induced by f.*

Typically, we picture a Reeb graph (F, q) by thinking of the function q as a "height" function; we imagine that F is situated in space so that $q(x)$ is the altitude of the point x. See Figure 7. In the computer science literature, many authors define the Reeb graph by declaring that $x \sim y$ if and only if x and y lie in the same *path-connected* component instead of the same *connected* component. While these definitions are not equivalent in general, they are equivalent under reasonable topological assumptions, such as when the level sets of f are locally path-connected.

Reeb graphs are used as summaries of landscapes induced by real functions. While originally designed by Georges Reeb to study functions on smooth manifolds (Reeb, 1946) and independently rediscovered by Kronrod (1950), Reeb graphs were first popularized as a computational tool by Shinagawa et al. (1991). Since then, Reeb graphs have grown in popularity, finding applications in the comparison (Hilaga et al., 2001; Escolano et al., 2013), parameterization (Patanè et al., 2004; Zhang et al., 2005), and denoising (Wood et al., 2004) of shapes, in symmetry detection (Thomas and Natarajan, 2011), and in sketching of static (Chazal and Sun, 2013; Ge et al., 2011) and time-varying (Edelsbrunner et al., 2004) data, among many others. See Biasotti et al. (2008); Yan et al. (2021) for surveys.
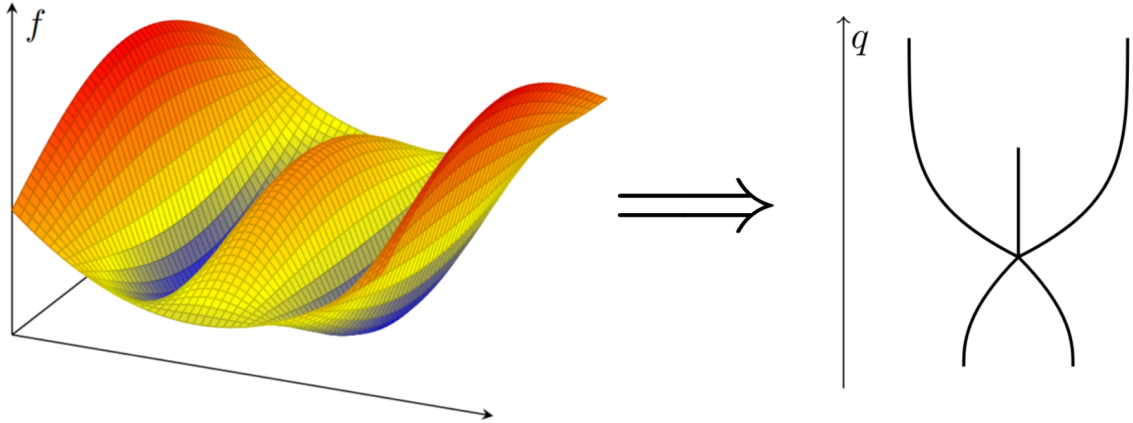
Figure 7: An illustration of a function f on the unit square and its induced Reeb graph. The two local minima of the function produce downward-pointing leaves of equal height on the Reeb graph. Similarly, the three local maxima produce upward-pointing leaves, with the lower maximum producing a lower leaf.

While it is easiest to imagine "Reeb graphs" as *topological graphs*[8] with attached height functions, one should be wary that, in general, the Reeb graph of a function need not admit the structure of a topological graph. Thankfully, many classes of spaces and functions are known to induce Reeb graphs which are also topological graphs. One list of sufficient conditions is provided by de Silva et al. (2016, Example 2.2).

Definition 98 *Let P be a problem. The* Reeb graph of P, *denoted* $\mathrm{Reeb}(P)$, *is the Reeb graph induced by the risk function* $\mathcal{R}_P : H \to \mathbb{R}_{\geqslant 0}$.

The Reeb graph of a problem P is a summary of the risk landscape. (See Figure 8.) For instance, the height of the lowest point of $\mathrm{Reeb}(P)$ is the constrained Bayes risk. If $\mathcal{R}_P$ has a unique minimizer, then $\mathrm{Reeb}(P)$ will have a unique lowest point. Conversely, if $\mathrm{Reeb}(P)$ has a unique lowest point, then the minimizers of $\mathcal{R}_P$ form a connected set. Local minima appear as downward-pointing leaves, with the height of the edges corresponding to the depth of the basins around those minima. Similarly, local maxima are represented by upward-pointing leaves. The Reeb graph of P therefore captures information that can be useful for quantifying the complexity of the problem P.

Reeb graphs can be compared via any of several related metrics. The largest such metric in popular use is called the *universal distance* (Bauer et al., 2021).

Definition 99 (Universal Distance on Reeb graphs) *The* universal distance *between two Reeb graphs (F, f) and (G, g) is given by*

$$d_{\mathrm{U}}((F, f), (G, g)) := \inf_{Z, \widehat{f}, \widehat{g} : Z \to \mathbb{R}} \|\widehat{f} - \widehat{g}\|_\infty$$

---

8. The following definition will suffice: A *topological graph* is a topological space produced from a graph by realizing the vertices as isolated points and the edges as arcs connecting them; see (Hatcher, 2005, Chapter 0) for details.
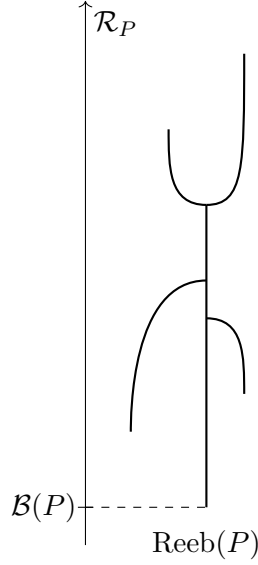
Figure 8: The Reeb graph of some problem P. Based on the downward-pointing leaves of the Reeb graph, this problem's risk landscape seems to have three basins. The height of the lowest point in the Reeb graph corresponds to the constrained Bayes risk $\mathcal{B}(P)$.

where the infimum ranges over all spaces Z and functions $\widehat{f}, \widehat{g} : Z \to \mathbb{R}$ such that $\widehat{f}$ induces the Reeb graph (F, f) and $\widehat{g}$ induces the Reeb graph (G, g).

Even with respect to the universal distance, the Reeb graph of a problem is stable under the Connected Risk distance.

Theorem 100 *Let P and P' be problems. Then*

$$|\mathcal{B}(P) - \mathcal{B}(P')| \leqslant d_{\mathrm{U}}(\mathrm{Reeb}(P), \mathrm{Reeb}(P')) \leqslant d_{\mathrm{RC}}(P, P').$$

Hence the Reeb graph of a problem is a stronger descriptor than the constrained Bayes risk while remaining stable under the Connected Risk distance. The proof can be found in Appendix A.

Example 101 *Let P be a problem whose loss ℓ is a metric, such as a 0-1 loss in a classification problem, mean absolute error in a regression problem, or a Wasserstein distance in a probability estimation problem. Suppose that we expect noise in the labels of our data such that the average distance under ℓ between the true label and the observed label is ε_1. Write η' for the noisy distribution from which the observed data is drawn. Suppose also that the loss function ℓ is difficult to exactly compute or has poor optimization properties, but there is an approximation ℓ' that does not share the same issues and is ε_2-close to ℓ in the supremum norm. Replacing η with the noisy η' and ℓ with the approximation ℓ' could introduce new local minima to the risk function. We would like to understand how bad these*

local minima can be. Let $P' = (X, Y, \eta', \ell, H)$ and $P'' = (X, Y, \eta', \ell', H)$. We aim to bound the universal distance between the Reeb graphs of the idealized problem P and the corrupted problem P''. Since Theorem 49 applies to the Connected Risk distance (Remark 96), we have

$$d_{\mathrm{RC}}(P, P') \leqslant \varepsilon_1.$$

Since d_{RC} also satisfies Condition 2 from Section 3.3, we also have

$$d_{\mathrm{RC}}(P', P'') \leqslant \max_{h \in H} \int \left| \ell(h(x), y) - \ell'(h(x), y) \right| \eta(dx \times dy) \leqslant \varepsilon_2.$$

Then, using the triangle inequality for d_{RC} and Theorem 100, we can write the bound

$$d_{\mathrm{U}}(\mathrm{Reeb}(P), \mathrm{Reeb}(P')) \leqslant d_{\mathrm{RC}}(P, P'') \leqslant d_{\mathrm{RC}}(P, P') + d_{\mathrm{RC}}(P', P'') \leqslant \varepsilon_1 + \varepsilon_2.$$

In other words, if we replace the idealized problem P with the corrupted problem P'', the Reeb graph can change by at most $\varepsilon_1 + \varepsilon_2$ in the universal distance. Hence any downward-pointing leaf of $\mathrm{Reeb}(P)$ of depth more than $2(\varepsilon_1 + \varepsilon_2)$ from top to bottom has a corresponding downward-pointing leaf in $\mathrm{Reeb}(P'')$, and the difference in the heights of the tips of two corresponding leaves is at most $\varepsilon_1 + \varepsilon_2$. We conclude that a basin of depth greater than $2(\varepsilon_1 + \varepsilon_2)$ in the risk landscape of the original problem corresponds to some basin in the corrupted problem, and the risk at the bottom of two corresponding basins differ by at most $\varepsilon_1 + \varepsilon_2$. Similarly, any new downward-pointing leaves in the Reeb graph created by corrupting P to get P'' can be of length at most $2(\varepsilon_1 + \varepsilon_2)$ from top to bottom, so any new basins introduced by the corruption can be of depth at most $2(\varepsilon_1 + \varepsilon_2)$ from the deepest to highest point of the basin.

8. Discussion

Machine learning is currently engaged in a rapid and wide-ranging exploration of its problem space. This paper proposes a guiding and organizing principle, focusing on supervised learning. As the field grows explosively, such structure is increasingly needed to make sense of emerging approaches. In this work, we argue that, given a notion of distance between supervised learning problems, various corruptions and modifications of problems can be seen from a geometric perspective. We propose the Risk distance and two variants thereof as reasonable choices for this distance notion, bolstering our claims with a garniture of stability results. The development in this paper can be seen as a direct descendent of earlier work in the machine learning community on reductions between learning problems, surrogate losses and different forms of regularization.

The present work has opened the way for further understanding of machine learning analogous to several productive, related research programs. The idea of the *function space* lies at the center of the abstract theory of functional analysis (Birkhoff and Kreyszig, 1984, page 259), the development of which was highly productive for 20th century mathematics (Dieudonné, 1981). Similarly productive for metric geometry was the work of Gromov (1981), which created the metric space of all metric spaces. Other similar examples include category theory, which is concerned with putting relational structures on collection of objects (Mac Lane, 2013), and the theories of optimal transport (Villani, 2008) and information geometry (Amari, 2016), which both propose geometric spaces of probability measures.

Common among all of these examples is the viewpoint that *to understand an object, it is productive to place it in a larger space of similar objects.* Following this viewpoint, the present work opens the door to similar techniques by establishing candidates for *spaces of supervised learning problems.* In particular, the definition of the Risk distance directly echoes the motifs of an established lineage of distances from metric geometry and optimal transport (Section 2.3). We have shown that the geometry is meaningful; the geometric notion of a geodesic is related to optimal couplings and correspondences (Section 5.1), and that the topological density of classification problems (Section 5.3) is useful in proving the convergence of empirical problems (Section 5.4). Throughout Section 6, the weighting λ placed on the set of predictors can be interpreted as a Bayesian prior—an approach aligned with the principles of Bayesian statistics. However, classical Bayesian frameworks do not consider a space of problems or attempt to endow it with a metric structure, as we do here. While Bayesian decision theory (Berger, 1985) incorporates elements such as loss functions, it does not address the geometry of the problem space. It is conceivable that the present work may offer new perspectives in that direction.

More generally, we are optimistic that further geometric, topological, and even categorical studies of the space of problems can shed light on questions of interest in supervised learning. For example, the following concrete research directions appear to be natural candidates for exploration through the lens of the ideas developed in this paper:

- *Convergence Rates for the Empirical Problem.* We posed Question 71 aimed at making Theorem 69, on the convergence of the empirical problem, more explicit and quantitative.

- *Leverage the framework to design broader, practical reductions between learning problems*, allowing more components of ML tasks to vary. As shown in Section 5.3, classification problems are dense—in the sense of the Risk distance—within the space of all compact problems. This suggests a structural connection between classification and regression problems, despite their apparent differences. Conversely, it is natural to investigate reductions in the opposite direction: for example, reducing a classification problem with severe label noise to a regression problem with biased sampling, where both problems are close in Risk distance. Such a reduction would enable robust regression techniques to be repurposed for noisy classification tasks, facilitating the transfer of insights across distinct problem classes.

- *Analyze the robustness of other theoretical results—e.g., fast rates of convergence—through the lens of the Risk distance (or suitable variants thereof).* In particular, is "fast convergence" a stable property under appropriate structural assumptions on the problems?

- *Assess the empirical value of the optimal couplings and correspondences produced by our theory.* An optimal coupling and correspondence (as studied in Section 5.2) between two problems provide a kind of dictionary relating the two problems. What new insights can be drawn from such a dictionary?

- *Investigate concrete models of* sampling bias *within our framework*, such as distribution shift (Quiñonero-Candela et al., 2009) and selection bias (Meng, 2018), which arise in many practical learning scenarios.

Acknowledgments

Appendix A. Relegated Proofs

A.1 Proofs and Details from Section 3

Proof [Theorem 27] The only nontrivial pseudometric axiom to prove is the triangle inequality. Furthermore, the claim that d_R vanishes on weakly isomorphic problems will follow from the triangle inequality and the fact that d_R satisfies Condition 1 (Theorem 28). That is, if P'' is common simulation of P and P', then $d_\mathrm{R}(P, P') \leqslant d_\mathrm{R}(P, P'') + d_\mathrm{R}(P'', P') = 0 + 0$. Hence we need only prove that d_R satisfies the triangle inequality.

Let $P_i = (X_i, Y_i, \eta_i, \ell_i, H_i)$ for $i = 1, 2, 3$. Let $\gamma_{1,2} \in \Pi(\eta_1, \eta_2)$, $\gamma_{2,3} \in \Pi(\eta_2, \eta_3)$, $R_{1,2} \in \mathcal{C}(H_1, H_2)$, and $R_{2,3} \in \mathcal{C}(H_2, H_3)$ be arbitrary. We apply the the gluing lemma (Lemma 1) and its counterpart for correspondences (Lemma 2). Let γ be a gluing of $\gamma_{1,2}$ and $\gamma_{1,3}$, and let $\gamma_{1,3}$ be its marginal on the first and third coordinates. Similarly, let R be a gluing of $R_{1,2}$ and $R_{2,3}$, and let $R_{1,3}$ be its image on its first and third components. We can then write

$$
\begin{aligned}
\mathrm{dis}_{P,P'}(R_{1,3}, \gamma_{1,3}) &= \sup_{(h_1,h_3)\in R_{1,3}} \int \left|\ell_1(h_1(x_1), y_1) - \ell_3(h_3(x_3), y_3)\right| \gamma_{1,3}(dx_1{\times}dy_1{\times}dx_3{\times}dy_3) \\
&= \sup_{(h_1,h_2,h_3)\in R} \int \left|\ell_1(h_1(x_1), y_1) - \ell_3(h_3(x_3), y_3)\right| \gamma(dx_1{\times}dy_1{\times}dx_2{\times}dy_2{\times}dx_3{\times}dy_3) \\
&\leqslant \sup_{(h_1,h_2,h_3)\in R} \int \left|\ell_1(h_1(x_1), y_1) - \ell_2(h_2(x_2), y_2)\right| \gamma(dx_1{\times}dy_1{\times}dx_2{\times}dy_2{\times}dx_3{\times}dy_3) \\
&\quad + \sup_{(h_1,h_2,h_3)\in R} \int \left|\ell_2(h_2(x_2), y_2) - \ell_3(h_3(x_3), y_3)\right| \gamma(dx_1{\times}dy_1{\times}dx_2{\times}dy_2{\times}dx_3{\times}dy_3) \\
&= \sup_{(h_1,h_2)\in R_{1,2}} \int \left|\ell_1(h_1(x_1), y_1) - \ell_2(h_2(x_2), y_2)\right| \gamma_{1,2}(dx_1{\times}dy_1{\times}dx_2{\times}dy_2) \\
&\quad + \sup_{(h_2,h_3)\in R_{2,3}} \int \left|\ell_2(h_2(x_2), y_2) - \ell_3(h_3(x_3), y_3)\right| \gamma_{2,3}(dx_2{\times}dy_2{\times}dx_3{\times}dy_3) \\
&= \mathrm{dis}_{P,P'}(R_{1,2}, \gamma_{1,2}) + \mathrm{dis}_{P,P'}(R_{2,3}, \gamma_{2,3}).
\end{aligned}
$$

Since we can construct such a $\gamma_{1,3}$ and $R_{1,3}$ from any given $\gamma_{1,2}, \gamma_{2,3}, R_{1,2}, R_{2,3}$, we can take the infimum of each side to get the desired inequality. ∎

Proof [Theorem 28] First we demonstrate that the Risk distance does indeed satisfy Conditions 1 and 2. Let P' be a simulation of P via the maps $f_1 : X' \to X$ and $f_2 : Y' \to Y$.

68

Define $f := f_1 \times f_2$. Since $f_\sharp \eta' = \eta$, f induces a coupling $\gamma_f \in \Pi(\eta', \eta)$. Define $R \subseteq H' \times H$ to be all the pairs (h', h) such that $h \circ f_1 = f_2 \circ h'$. Since f induces a simulation, R is a correspondence. By definition of the Risk distance, we can now write

$$d_{\mathrm{R}}(P, P') \leqslant \sup_{(h', h) \in R} \int \left| \ell'(h'(x'), y') - \ell(h(x), y) \right| \gamma_f(dx' \times dy' \times dx \times dy)$$

$$= \sup_{(h', h) \in R} \int \left| \ell'(h'(x'), y') - \ell(h \circ f_1(x'), f_2(y')) \right| \eta'(dx' \times dy').$$

We claim the integrand is identically zero. Indeed,

$$\ell(h \circ f_1(x'), f_2(y')) = \ell(f_2 \circ h'(x'), f_2(y')) = \ell'(h'(x'), y').$$

Hence $d_{\mathrm{R}}(P, P') = 0$.

Let $P = (X, Y, \eta, \ell, H)$ and $P' = (X, Y, \eta, \ell', H)$ be any pair of problems that differ only in their loss functions. We can bound $d_{\mathrm{R}}(P, P')$ by selecting the diagonal coupling between η and itself, and the diagonal correspondence between H and itself. This directly yields Condition 2.

Now suppose d is a pseudometric satisfying Conditions 1 and 2, and let P_1 and P_2 be problems. Let $\gamma \in \Pi(\eta_1, \eta_2)$ and $R \in \mathcal{C}(H_1, H_2)$ be arbitrary. Define two new problems

$$P_1' := (X_1 \times X_2, Y_1 \times Y_2, \gamma, \ell_1', R_\times)$$
$$P_2' := (X_1 \times X_2, Y_1 \times Y_2, \gamma, \ell_2', R_\times),$$

where

- $\ell_1'((y_1, y_2), (y_1', y_2')) := \ell_1(y_1, y_1')$,

- $\ell_2'((y_1, y_2), (y_1', y_2')) := \ell_2(y_2, y_2')$,

- $R_\times := \{h_1 \times h_2 | (h_1, h_2) \in R\}$.

Then P_1' is a simulation of P_1 via the projection maps $X_1 \times X_2 \to X_1$ and $Y_1 \times Y_2 \to Y_1$, so by Condition 1, $d(P_1, P_1') = 0$. Similarly, $d(P_2, P_2') = 0$. Combining this with Condition 2, we get

$$d(P_1, P_2) = d(P_1', P_2')$$

$$\leqslant \max_{h_1 \times h_2 \in R_\times} \int \left| \ell_1'((h_1 \times h_2)(x_1, x_2), (y_1, y_2)) - \ell_2'((h_1 \times h_2)(x_1, x_2), (y_1, y_2)) \right| \gamma(dx_1 \times dx_2 \times dy_1 \times dy_2)$$

$$= \max_{(h_1, h_2) \in R} \int \left| \ell_1(h_1(x_1), y_1) - \ell_2(h_2(x_2), y_2) \right| \gamma(dx_1 \times dx_2 \times dy_1 \times dy_2)$$

$$= \mathrm{dis}_{P_1, P_2}(R, \gamma).$$

Taking an infimum over all choices of γ and R finishes the proof. ∎

Proof [Proposition 29] Let P and P' be weakly isomorphic, and let P'' be a common simulation of both P and P' via the maps

Define the map

$$m : X'' \times Y'' \to X \times Y \times X' \times Y'$$
$$(x'', y'') \mapsto (f_1(x''), f_2(y''), g_1(x''), g_2(y'')),$$

which, we note, makes the following diagram commute:

Define $\gamma := m_\sharp \eta''$. Note that, since the above diagram commutes, γ is a coupling between η and η'. Also, let $\phi : H'' \to H$ be such that $\ell''_{h''} = \ell_{\phi(h'')} \circ (f_1 \times f_2)$ holds η''-almost everywhere. Define $\psi : H'' \to H'$ similarly. Then

$$R := \big\{ (\phi(h''), \psi(h'')) \big| h'' \in H'' \big\}$$

is a correspondence between H and H'. Finally, compute

$$\text{dis}_{P,P'}(R, \gamma) = \sup_{(h,h') \in R} \int \big| \ell_h(x, y) - \ell'_{h'}(x', y') \big| \, \gamma(dx \times dy \times dx' \times dy')$$

$$= \sup_{(h,h') \in R} \int \big| \ell_h \circ \pi_{12}(x, y, x', y') - \ell'_{h'} \circ \pi_{34}(x, y, x', y') \big| \, \gamma(dx \times dy \times dx' \times dy')$$

$$= \sup_{(h,h') \in R} \int \big| \ell_h \circ \pi_{12} \circ m(x'', y'') - \ell'_{h'} \circ \pi_{34} \circ m(x'', y'') \big| \, \eta''(dx'' \times dy'')$$

$$= \sup_{(h,h') \in R} \int \big| \ell_h(f_1(x''), f_2(y'')) - \ell'_{h'}(g_1(x''), g_2(y'')) \big| \, \eta''(dx'' \times dy'')$$

$$= \sup_{h'' \in H''} \int \big| \ell_{\phi(h'')}(f_1(x''), f_2(y'')) - \ell'_{\psi(h'')}(g_1(x''), g_2(y'')) \big| \, \eta''(dx'' \times dy'')$$

$$= \sup_{h'' \in H''} \int \big| \ell''_{h''}(x'', y'') - \ell''_{h''}(x'', y'') \big| \, \eta''(dx'' \times dy'') = 0.$$

Now suppose that there exist $R \in \mathcal{C}(H, H')$ and $\gamma \in \Pi(\eta, \eta')$ such that $\text{dis}_{P,P'}(R, \gamma) = 0$. Construct a new problem

$$P'' = (X \times X', Y \times Y', \tau_\sharp \gamma, \ell'', R_\times)$$

where

70

- $\tau : X \times Y \times X' \times Y' \to X \times X' \times Y \times Y'$ swaps the second and third components,

- $\ell''((y_1, y_1'), (y_2, y_2')) := \ell(y_1, y_2)$

- $R_\times := \{h \times h' | (h, h') \in R\}$.

Then the projection maps $f_1 : X \times X' \to X$ and $f_2 : Y \times Y' \to Y$ show that P'' is a simulation of P. Indeed, since γ is a coupling of η and η', we have

$$(f_1 \times f_2)_\sharp \tau_\sharp \gamma = ((f_1 \times f_2) \circ \tau)_\sharp \gamma = (\pi_{12})_\sharp \gamma = \eta.$$

Furthermore, for any $h \in H$, we can consider some function $(h \times h') \in R_\times$ and write

$$\ell''_{(h \times h')}((x, x'), (y, y')) = \ell''((h(x), h'(x')), (y, y')) = \ell(h(x), y) = \ell_h(f_1(x, x'), f_2(y, y')).$$

Similarly, P'' is a simulation of P' via the projection maps $g_1 : X \times X' \to X'$ and $g_2 : Y \times Y' \to Y'$. Indeed, an almost identical calculation to the one above shows that $(g_1 \times g_2)_\sharp \tau_\sharp \gamma = \eta'$. Furthermore, if $h' \in H'$ and $(h \times h') \in R_\times$, we have already calculated that $\ell''_{(h \times h')}((x, x'), (y, y')) = \ell(h(x), y)$. The fact that R and γ have Risk distortion zero impliex that the latter is equal $\tau_\sharp \gamma$-almost everywhere to $\ell'(h'(x'), y') = \ell'_{h'}(g_1(x, x'), g_2(y, y'))$, completing the proof. ∎

A.1.1 CONNECTING THE RISK DISTANCE WITH THE GROMOV-WASSERSTEIN DISTANCE

Here we discuss a connection with a variant of the Gromov-Wasserstein distance considered by Hang et al. (2019). Given two metric measure spaces X, X', a correspondence $R \in \mathcal{C}(X, X')$ and a coupling $\mu \in \Pi(\mu_X, \mu_{X'})$, one considers the following notion of distortion:

$$\overline{\mathrm{dis}}(R, \mu) := \sup_{(x_2, x_2') \in R} \int_{X \times X'} |d_X(x_1, x_2) - d_{X'}(x_1', x_2')| \mu(dx_1 \times dx_1'),$$

which leads to the definition

$$\overline{d_{\mathrm{GW}}}(X, X') := \inf_{R, \mu} \overline{\mathrm{dis}}(X, X').$$

This definition, which blends an L^∞ matching (expressed through the correspondence R) with an L^1 one (expressed via the coupling μ) was considered by Hang et al. (2019) due its beneficial interaction with stability questions surrounding Fréchet functions in the persistent homology setting.[9] With the notation and definitions from Section 3.5, we have the following statement.

Proposition 102 *For all compact metric measure spaces X and X': $d_{\mathrm{R}}(i(X), i(X')) = \overline{d_{\mathrm{GW}}}(X, X')$.*

---

9. The precise variant considered in (Hang et al., 2019) uses "continuous" correspondences. The version we consider here is structurally identical but slightly softer due to our use of arbitrary correspondences.

Proof First note that any correspondence R between H_X and $H_{X'}$ can be regarded as a correspondence between X and X'. This is so because all functions in H_X (resp. in $H_{X'}$) are constant.

We then have the following calculation:

$$
\begin{aligned}
\operatorname{dis}_{i(X),i(X')}(R,\gamma) &= \sup_{(h,h')\in R} \int \left|d_X(h(x_0),x_1) - d_{X'}(h'(x_0'),x_1')\right| \gamma(dx_0 \times dx_1 \times dx_0' \times dx_1') \\
&= \sup_{(x_2,x_2')\in R} \int \left|d_X(x_2,x_1) - d_{X'}(x_2',x_1')\right| \gamma(dx_0 \times dx_1 \times dx_0' \times dx_1') \\
&= \sup_{(x_2,x_2')\in R} \int \left|d_X(x_2,x_1) - d_{X'}(x_2',x_1')\right| (\pi_{2,4})_\sharp\gamma(dx_0 \times dx_1 \times dx_0' \times dx_1') \\
&= \overline{\operatorname{dis}}(R,(\pi_{2,4})_\sharp\gamma).
\end{aligned}
$$

The above equality immediately implies that $d_R(i(X),i(X')) \geqslant \overline{d_{\mathrm{GW}}}(X,X')$. The reverse inequality can be obtained from the following fact. Let $\tau : X \times X' \to X \times X \times X' \times X'$ be the map $(x,x') \mapsto (x,x,x',x')$. Then, if $\mu \in \Pi(\mu_X,\mu_{X'})$, $\gamma_\mu := \tau_\sharp\mu$ is an element of $\Pi((\Delta_X)_\sharp\mu_X,(\Delta_{X'})_\sharp\mu_{X'})$. Since $\pi_{2,4} \circ \tau = \mathrm{id} : X \times X' \to X \times X'$, the identity map, $(\pi_{2,4})_\sharp\gamma_\mu = \mu$ and the calculation above gives $\operatorname{dis}_{i(X),i(X')}(R,\gamma_\mu) = \overline{\operatorname{dis}}(\mu)$. From this we obtain that $d_R(i(X),i(X')) \leqslant \overline{d_{\mathrm{GW}}}(X,X')$. ∎

Proof [Proposition 33] If $d_{\mathrm{GW},p}(X,X') = 0$, then X and X' are isomorphic as metric measure spaces. Using the isomorphism $X \to X'$, it is not hard to construct a coupling and correspondence between $i(X)$ and $i(X')$ of risk distortion zero. Conversely, let $d_R(i(X),i(X')) = 0$. It is not a priori obvious that there exists a correspondence $R \in \mathcal{C}(H_X,H_{X'})$ and a coupling $\gamma \in \Pi((\Delta_X)_\sharp\mu_X,(\Delta_{X'})_\sharp\mu_{X'})$ that together achieve the infimal risk distortion of zero. In Section 5.2, we will explore the existence of such couplings and correspondences. In this case, since d_X and $d_{X'}$ are continuous functions and X and X' are assumed to be compact, Theorem 60 does guarantee that there exists an R and a γ with $\operatorname{dis}_{i(X),i(X')}(R,\gamma) = 0$. Furthermore, recall that a correspondence R between H_X and $H_{X'}$ is equivalent to a correspondence between X and X'. Hence, as in the proof of Proposition 102, we have $0 = \operatorname{dis}_{i(X),i(X')}(R,\gamma) = \overline{\operatorname{dis}}(R,(\pi_{2,4})_\sharp\gamma)$. Set $R' := \operatorname{supp}((\pi_{2,4})_\sharp\gamma)$. Since the integrand defining $\overline{\operatorname{dis}}(R,(\pi_{2,4})_\sharp\gamma)$ is continuous and non-negative, we can conclude that

$$d_X(x_2,x_1) = d_{X'}(x_2',x_1') \tag{3}$$

for all $(x_2,x_2') \in R$ and $(x_1,x_1') \in R'$. Furthermore, since $(\pi_{2,4})_\sharp\gamma$ is a coupling between the fully supported measures μ_X and $\mu_{X'}$, its support R' is a correspondence between X and X' (see Mémoli, 2011, Lemma 2.2). We claim that $R = R'$. Indeed, if $(x_2,x_2') \in R$, selecting any pair $(x_2,x_1') \in R'$ yields $d_X(x_2,x_2) = d_{X'}(x_2',x_1')$ and hence $x_2' = x_1'$. This shows $R \subseteq R'$, and a similar argument shows the opposite inclusion. We furthermore claim that R is the graph of a bijection. Let $(x,x'),(x,x'') \in R$. Then $d_X(x,x) = d_{X'}(x',x'')$, showing $x' = x''$. Again, a similar argument holds when reversing the roles of X and X'. Let $f : X \to X'$ be the bijection whose graph is R. Then equation (3) becomes $d_X(x_2,x_1) = d_{X'}(f(x_2),f(x_1))$, showing that f is an isometry. Additionally, since $(\pi_{2,4})_\sharp\gamma$ is a coupling of μ_X and $\mu_{X'}$ which

is supported on R, the projections $R \to X$ and $R \to X'$ are measure-preserving bijections. It follows that the composition of the natural maps $X \to R \to X'$, which agrees with f, is a measure-preserving map. Hence f is an isomorphism of metric measure spaces, and $d_{\mathrm{GW},1}(X, X') = 0$. ■

A.2 Proofs from Section 4

Proof [Theorem 37] Let $\mathrm{Prob}_1(\mathbb{R}_{\geqslant 0}) \subseteq \mathrm{Prob}(\mathbb{R}_{\geqslant 0})$ be the subset of probability measures with finite mean. The mean function

$$\mathrm{mean} : \mathrm{Prob}_1(\mathbb{R}_{\geqslant 0}) \to \mathbb{R}_{\geqslant 0}$$

is 1-Lipschitz when the domain is endowed with the 1-Wasserstein distance $d_{\mathrm{W},1}$ and the codomain is given the Euclidean distance. We will now "lift" the mean function to a function on power sets by applying it elementwise. That is, we define

$$\mathrm{Pow}(\mathrm{mean}) : \mathrm{Pow}(\mathrm{Prob}_1(\mathbb{R}_{\geqslant 0})) \to \mathrm{Pow}(\mathbb{R}_{\geqslant 0})$$
$$\{\alpha_i | i \in I\} \mapsto \{\mathrm{mean}(\alpha_i) | i \in I\}.$$

We also endow the domain and codomain with the Hausdorff distance, where the underlying metrics are $d_{\mathrm{W},1}$ and the Euclidean distance respectively. Since the mean function is 1-Lipschitz, so is $\mathrm{Pow}(\mathrm{mean})$. The infimum function $\inf : \mathrm{Pow}(\mathbb{R}_{\geqslant 0}) \to \mathbb{R}_{\geqslant 0}$ 1-Lipschitz as well, so the composition $\inf \circ \mathrm{Pow}(\mathrm{mean})$ is 1-Lipschitz. In other words, for any $M, N \subseteq \mathrm{Prob}_1(\mathbb{R}_{\geqslant 0})$, we have

$$\left| \inf \{\mathrm{mean}(\mu) | \mu \in M\} - \inf \{\mathrm{mean}(\nu) | \nu \in N\} \right| \leqslant d_{\mathrm{H}}^{d_{\mathrm{W},1}}(M, N).$$

By Proposition 36, applying this inequality when M and N are loss profile sets gives the first desired inequality.

To prove the second inequality, observe that $d(P, P') := d_{\mathrm{H}}^{d_{\mathrm{W},1}}(L(P), L(P'))$ is a pseudo-metric on the collection of all problems. By Theorem 28, we need only show that d satisfies Conditions 1 and 2. Beginning with Condition 1, let P' be a simulation of P via the maps $f_1 : X' \to X$ and $f_2 : Y' \to Y$. Let $f := f_1 \times f_2$. If $h \in H$ and $h' \in H'$ are such that $f_2 \circ h = h' \circ f_2$, then

$$\ell'_{h'} \circ f(x, y) = \ell'(h' \circ f_1(x), f_2(y)) = \ell'(f_2 \circ h(x), f_2(y)) = \ell(h(x), y) = \ell_h(x, y).$$

Hence the loss profiles of h and h' are equal:

$$(\ell'_{h'})_\sharp \eta' = (\ell'_{h'} \circ f)_\sharp \eta = (\ell_h)_\sharp \eta.$$

Such an $h' \in H'$ exists for every $h \in H$ and vice versa by definition of a simulation, so we conclude that $L(P) = L(P')$ and $d(P, P') = 0$.

For Condition 2, let $h \in H$. We can bound the Wasserstein distance between the loss profiles $(\ell_h)_\sharp \eta$ and $(\ell'_h)_\sharp \eta$ by selecting the coupling $(\ell_h, \ell'_h)_\sharp \eta$, which yields

$$d_{\mathrm{W},1}((\ell_h)_\sharp \eta, (\ell'_h)_\sharp \eta) \leqslant \int |r - s| (\ell_h, \ell'_h)_\sharp \eta(dr \times ds)$$

$$= \int |\ell_h(x, y) - \ell'_h(x, y)| \, \eta(dx \times dy).$$

73

Since this holds for all h, Condition 2 is satisfied as well. ∎

Proof [Theorem 41] Let $R \in \mathcal{C}(H, H')$ and $\gamma \in \Pi(\eta, \eta')$ be arbitrary. Then

$$
\begin{aligned}
&\left| R_m(P) - R_m(P') \right| \\
&= \left| \iint \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell_h(x_i, y_i) \operatorname{Rad}^{\otimes m}(d\sigma) \, \eta^{\otimes m}(d\overline{x} \times d\overline{y}) \right. \\
&\quad \left. - \iint \sup_{h' \in H'} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell'_{h'}(x'_i, y'_i) \operatorname{Rad}^{\otimes m}(d\sigma) \, \eta^{\otimes m}(d\overline{x}' \times d\overline{y}') \right| \\
&\leqslant \iint \sup_{(h,h') \in R} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell_h(x_i, y_i) - \frac{1}{m} \sum_{i=1}^{m} \sigma_i \ell'_{h'}(x'_i, y'_i) \right| \operatorname{Rad}^{\otimes m}(d\sigma) \, \gamma^{\otimes m}(d\overline{x} \times d\overline{y} \times d\overline{x}' \times d\overline{y}') \\
&\leqslant \int \sup_{(h,h') \in R} \frac{1}{m} \sum_{i=1}^{m} \left| \ell_h(x_i, y_i) - \ell'_{h'}(x'_i, y'_i) \right| \gamma^{\otimes m}(d\overline{x} \times d\overline{y} \times d\overline{x}' \times d\overline{y}').
\end{aligned}
$$

Let $\widehat{\gamma}$ be the unique probability measure on the countable product $(X \times Y)^{\mathbb{N}}$ whose marginal on the first m components is $\gamma^{\otimes m}$. Then the above integral can be written

$$
\int \sup_{(h,h') \in R} \frac{1}{m} \sum_{i=1}^{m} \left| \ell_h(x_i, y_i) - \ell'_{h'}(x'_i, y'_i) \right| \widehat{\gamma}(d\overline{x} \times d\overline{y} \times d\overline{x}' \times d\overline{y}'). \tag{4}
$$

We now aim to establish the bound

$$
\begin{aligned}
&\int \sup_{(h,h') \in R} \frac{1}{m} \sum_{i=1}^{m} \left| \ell_h(x_i, y_i) - \ell'_{h'}(x'_i, y'_i) \right| \widehat{\gamma}(d\overline{x} \times d\overline{y} \times d\overline{x}' \times d\overline{y}') \\
&\quad - \sup_{(h,h') \in R} \int \left| \ell_h(x_i, y_i) - \ell'_{h'}(x'_i, y'_i) \right| \gamma(d\overline{x} \times d\overline{y} \times d\overline{x}' \times d\overline{y}') \\
&\leqslant 2\operatorname{Rad}_m(\mathcal{F})
\end{aligned}
$$

since infimizing this inequality over γ and R will give us the result. It will suffice to establish

$$
\begin{aligned}
&\int \sup_{(h,h') \in R} \left[\frac{1}{m} \sum_{i=1}^{m} \left| \ell_h(x_i, y_i) - \ell'_{h'}(x'_i, y'_i) \right| - \int \left| \ell_h(x_i, y_i) - \ell'_{h'}(x'_i, y'_i) \right| \right] \widehat{\gamma}(d\overline{x} \times d\overline{y} \times d\overline{x}' \times d\overline{y}') \\
&\leqslant 2\operatorname{Rad}_m(\mathcal{F}),
\end{aligned}
$$

which, in probabilistic notation, is written

$$
\mathbb{E} \sup_{(h,h') \in R} \left[\frac{1}{m} \sum_{i=1}^{m} \left| \ell_h(X_i, Y_i) - \ell'_{h'}(X'_i, Y'_i) \right| - \mathbb{E} \left| \ell_h(X, Y) - \ell'_{h'}(X', Y') \right| \right] \leqslant 2\operatorname{Rad}_m(\mathcal{F}) \tag{5}
$$

where the first expectation is taken over the tuples (X_i, Y_i, X'_i, Y'_i) each sampled i.i.d. from γ, and the second expectation is over (X, Y, X', Y') also sampled from γ independently of the other tuples. The expression on the left hand side of Equation 5 is well-studied in the

statistical learning theory literature, and the bound in equation 5 is easily obtained by a symmetrization argument like the one laid out in (Liao, 2020, Theorem 4.1).

Now assume $\mathcal{F}$ is a universal Glivenko-Cantelli class. Return to the expression (4) and let $m \to \infty$. By the Glivenko-Cantelli assumption, the integrand of (4) converges almost everywhere to

$$\sup_{(h,h') \in R} \int \left| \ell_h(x,y) - \ell'_{h'}(x',y') \right| \gamma(dx \times dy \times dx' \times dy') = \mathrm{dis}_{P,P'}(R,\gamma).$$

Furthermore, since ℓ and ℓ' are bounded, by the dominated convergence theorem, the limit of (4) itself is $\mathrm{dis}_{P,P'}(R,\gamma)$ as well. Infimizing over the choices of R and γ completes the proof. $\blacksquare$

Proof [Lemma 48] We aim to bound $d_{\mathrm{R}}(P,P')$. To this end, define

$$\Delta : X \times Y \times Y \to X \times Y \times X \times Y$$
$$(x,y,y') \mapsto (x,y,x,y').$$

Let Π_g be the collection of all gluings of η and η' along X. That is,

$$\Pi_g := \left\{ \lambda \in \mathrm{Prob}(X \times Y \times Y) \big| (\pi_{1,2})_\sharp \lambda = \eta, (\pi_{1,3})_\sharp \lambda = \eta' \right\}.$$

By the gluing lemma (Lemma 1), Π_g is nonempty. Note that if $\lambda \in \Pi_g$, then $\Delta_\sharp \lambda \in \Pi(\eta,\eta')$.

We can then write

$$d_{\mathrm{R}}(P,P') \leqslant \inf_{\gamma \in \Pi(\eta,\eta')} \int \left| \ell(h(x),y) - \ell(h(x'),y') \right| \gamma(dx \times dy \times dx' \times dy')$$

$$\leqslant \inf_{\lambda \in \Pi_g} \int \left| \ell(h(x),y) - \ell(h(x'),y') \right| \Delta_\sharp \lambda(dx \times dy \times dx' \times dy')$$

$$= \inf_{\lambda \in \Pi_g} \int \left| \ell(h(x),y) - \ell(h(x),y') \right| \lambda(dx \times dy \times dy')$$

$$\leqslant C \inf_{\lambda \in \Pi_g} \int d_Y(y,y') \, \lambda(dx \times dy \times dy').$$

Now suppose $\tau \in \Pi(\beta,\beta')$. Then one can show that the induced coupling $\lambda \in \mathrm{Prob}(X \times Y \times Y)$, which has X-marginal α and X-disintegration $\tau : X \to \mathrm{Prob}(Y \times Y')$, lies in Π_g. Hence we can continue our string of inequalities with

$$\leqslant C \inf_{\tau \in \Pi(\beta,\beta')} \int \int d_Y(y,y') \, \tau(x)(dy \times dy') \alpha(dx) = C d_{\mathrm{W},1}^{d_Y}(\beta,\beta'),$$

the desired bound. $\blacksquare$

Proof [Theorem 49] By Lemma 48, we obtain

$$d_{\mathrm{R}}(P,P') \leqslant C d_{\mathrm{W},1}^{d_Y}(\beta,\beta').$$

We can write β and β' as

$$\beta = \Delta_X \cdot (\delta_{\mathrm{Id}_X} \otimes \beta) \cdot \delta_{\pi_2}$$
$$\beta' = \Delta_X \cdot (\delta_{\mathrm{Id}_X} \otimes \beta) \cdot N.$$

Here Δ_X is the diagonal kernel $\Delta_X : X \to X \times X$ given by $\Delta_X(x) = \delta_{(x,x)}$, and $\otimes$ represents the product of Markov kernels. Note that

$$\alpha \cdot \Delta_X \cdot (\delta_{\mathrm{Id}_X} \otimes \beta) = \eta.$$

Hence $\Delta_X \cdot (\delta_{\mathrm{Id}_X} \otimes \beta)$ is an example of a measure non-increasing Markov kernel (Patterson, 2020, Proposition 6.1), so we can write

$$d_{\mathrm{W},1}^{d_Y}(\Delta_X \cdot (\delta_{\mathrm{Id}_X} \otimes \beta) \cdot \delta_{\pi_2}, \Delta_X \cdot (\delta_{\mathrm{Id}_X} \otimes \beta) \cdot N) \leqslant d_{\mathrm{W},1}^{d_Y}(\delta_{\pi_2}, N),$$

completing the proof. ∎

A.3 Proofs from Section 5

Proof [Lemma 58] The following proof uses ideas similar to those in the proof of Theorem 4.1 of Villani (2008), with the additional wrinkle of an integrand that depends on n and an additional trick of "hot-swapping" an integrand with a lower semi-continuous version.

Since the topology on the domain is induced by a pseudometric, it is enough to consider an arbitrary convergent sequence $(h_n, h_n', \gamma_n) \to (h, h', \gamma)$ in $H \times H' \times \Pi(\eta, \eta')$ and check that,

$$\widetilde{\mathrm{dis}}_{P,P'}(h, h', \gamma) \leqslant \liminf_n \widetilde{\mathrm{dis}}_{P,P'}(h_n, h_n', \gamma_n).$$

To that end, let $\varepsilon > 0$ be arbitrary. We define a function

$$f : X \times Y \times X' \times Y' \to \mathbb{R}_{\geqslant 0}$$
$$(x_0, y_0, x_0', y_0') \mapsto \liminf_{(x,y,x',y') \to (x_0,y_0,x_0',y_0')} |\ell_h(x,y) - \ell_{h'}'(x',y')|.$$

That is, f is a lower semi-continuous version of the function $|\ell_h - \ell_{h'}'|$, constructed by modifying the latter on its points of discontinuity. Indeed, f and $|\ell_h - \ell_{h'}'|$ agree wherever the latter is continuous, which is γ-almost everywhere on $X \times Y \times X' \times Y'$. To wit, if $A \subseteq X \times Y$ and $A' \subseteq X' \times Y'$ are the points at which ℓ_h and $\ell_{h'}'$ are discontinuous respectively, then the points of discontinuity for $|\ell_h - \ell_{h'}'|$ lie within $A \times X' \times Y' \cup X \times Y \times A'$, and one can check

$$\gamma(A \times X' \times Y' \cup X \times Y \times A') \leqslant \gamma(A \times X' \times Y') + \gamma(X \times Y \times A') = \eta(A) + \eta'(A') = 0.$$

Hence f is a lower semi-continuous function which agrees γ-almost everywhere with $|\ell_h - \ell_{h'}'|$.

We can now proceed with our convergence argument; let $\varepsilon > 0$ and pick N large enough that, for all $n \geqslant N$,

- $d_{\ell,\eta}(h_n, h) < \varepsilon/3$,

- $d_{\ell',\eta'}(h_n', h') < \varepsilon/3$,

- and[10]

$$\int f(x,y,x',y')\,d\gamma < \int f(x,y,x',y')\,d\gamma_n + \varepsilon/3.$$

The latter is possible since f is a lower semi-continuous, non-negative function (Villani, 2008, Lemma 4.3). Then for large n we have

$$\widetilde{\mathrm{dis}}_{P,P'}(h,h',\gamma) = \int \left| \ell_h(x,y) - \ell'_{h'}(x',y') \right| d\gamma$$

$$= \int f(x,y,x',y')\,d\gamma$$

$$\leqslant \int f(x,y,x',y')\,d\gamma_n + \varepsilon/3$$

$$= \int \left| \ell_h(x,y) - \ell'_{h'}(x',y') \right| d\gamma_n + \varepsilon/3$$

$$\leqslant \int \left| \ell_h(x,y) - \ell_{h_n}(x,y) \right| d\gamma_n$$

$$+ \int \left| \ell_{h_n}(x,y) - \ell'_{h'_n}(x',y') \right| d\gamma_n$$

$$+ \int \left| \ell_{h'_n}(x',y') - \ell'_{h'}(x',y') \right| d\gamma_n + \varepsilon/3$$

$$= d_{\ell,\eta}(h,h_n) + \widetilde{\mathrm{dis}}_{P,P'}(h_n,h'_n,\gamma_n) + d_{\ell',\eta'}(h',h'_n) + \varepsilon/3 < \widetilde{\mathrm{dis}}_{P,P'}(h_n,h'_n,\gamma_n) + \varepsilon$$

as desired. In the last equality, we replace γ_n with η and η' in the first and last integrals respectively since γ_n is a coupling between η and η'. ∎

In order to prove Lemma 59 we will require the following technical lemma.

Lemma 103 *Let X be a pseudometric space and $f : X \to \mathbb{R}$ a lower semi-continuous function. Then the map*

$$F : \mathrm{Pow}(X) \to \mathbb{R}$$
$$A \mapsto \sup_{a \in A} f(a)$$

is lower semi-continuous as well when the power set $\mathrm{Pow}(X)$ is endowed with the Hausdorff metric.

Proof Let the sequence of subsets $A_n \subseteq X$ converge to $A \subseteq X$ in the Hausdorff distance. Select $a' \in A$ such that

$$\sup_{a \in A} f(a) < f(a') + \varepsilon/2.$$

Select also a sequence $a'_n \in X$ such that $a'_n \in A_n$ for all n, and a'_n converges to a'. Then for sufficiently large n,

$$f(a') < f(a'_n) + \varepsilon/2$$

---

10. In this proof, for conciseness, we will write $\int f(x,y,x',y')d\gamma$ instead of $\int f(x,y,x',y')\,\gamma(dx \times dy \times dx' \times dy')$, etc.

by lower semi-continuity of f. Hence for large n,

$$F(A) = \sup_{a \in A} f(a) < f(a') + \varepsilon/2 < f(a'_n) + \varepsilon \leqslant \sup_{a \in A_n} f(a) + \varepsilon = F(A_n) + \varepsilon.$$

Hence F is lower semi-continuous. ∎

Proof [Lemma 59] Write

$$\mathrm{dis}_{P,P'}(R, \gamma) = \sup_{((h,\nu),(h',\nu')) \in R \times \{\gamma\}} \widetilde{\mathrm{dis}}_{P,P'}(h, h', \nu).$$

Using the above equation, we can think of $\mathrm{dis}_{P,P'}$ as a composition of two maps:

$$\mathcal{C}(H, H') \times \Pi(\eta, \eta') \to \mathrm{Pow}(H \times H' \times \Pi(\eta, \eta'))$$
$$(R, \gamma) \mapsto R \times \{\gamma\},$$

and

$$\mathrm{Pow}(H \times H' \times \Pi(\eta, \eta')) \to \mathbb{R}$$
$$A \mapsto \sup_{(h,h',\nu) \in A} \widetilde{\mathrm{dis}}_{P,P'}(h, h', \nu).$$

The second map is lower semi-continuous by Lemma 103. The first map is an isometric embedding. Indeed, let d_w be a metric on $\Pi(\eta, \eta')$ that induces the weak topology. The topology on the domain $\mathcal{C}(H, H') \times \Pi(\eta, \eta')$ is induced by the metric

$$d_1((R_1, \gamma_1), (R_2, \gamma_2)) := \max\{d_{\mathrm{H}}(R_1, R_2), d_w(\gamma_1, \gamma_2)\}.$$

Meanwhile, unwrapping the metric on the codomain $\mathrm{Pow}(H \times H' \times \Pi(\eta, \eta'))$ gives us

$$d_2(R_1 \times \{\gamma_1\}, R_2 \times \{\gamma_2\})$$

$$:= \max \left\{ \sup_{(h_1,h'_1,\nu_1) \in R_1 \times \{\gamma_1\}} \inf_{(h_2,h'_2,\nu_2) \in R_2 \times \{\gamma_2\}} \max\{d_{H \times H'}((h_1, h'_1)(h_2, h'_2)), d_w(\nu_1, \nu_2)\}, \right.$$

$$\left. \sup_{(h_2,h'_2,\nu_2) \in R_2 \times \{\gamma_2\}} \inf_{(h_1,h'_1,\nu_1) \in R_1 \times \{\gamma_1\}} \max\{d_{H \times H'}((h_1, h'_1)(h_2, h'_2)), d_w(\nu_1, \nu_2)\} \right\}$$

$$= \max \left\{ \sup_{(h_1,h'_1) \in R_1} \inf_{(h_2,h'_2) \in R_2} d_{H \times H'}((h_1, h'_1)(h_2, h'_2)), \right.$$

$$\left. \sup_{(h_2,h'_2) \in R_2} \inf_{(h_1,h'_1) \in R_1} d_{H \times H'}((h_1, h'_1)(h_2, h'_2)), d_w(\gamma_1, \gamma_2) \right\}$$

$$= \max\{d_{\mathrm{H}}(R_1, R_2), d_w(\gamma_1, \gamma_2)\},$$

which is the definition of $d_1((R_1, \gamma_1), (R_2, \gamma_2))$. Hence the composition is lower semi-continuous. ∎

Proof [Theorem 60] Throughout this proof, let $\mathcal{C}_c(H, H') \subseteq \mathcal{C}(H, H')$ be the collection of topologically closed correspondences between H and H'.

Note that $\Pi(\eta, \eta')$ is compact under the weak topology, since η and η' are probability measures on Polish spaces. Compactness of such coupling spaces is proved, for instance, by Sturm (2023, Lemma 1.2), Chowdhury and Mémoli (2019, Lemma 2.2) and in a proof by Villani (2003, Prop 2.1). Similarly, $\mathcal{C}_c(H, H')$ is compact under the Hausdorff distance when H and H' are compact. (Here we are equipping H and H' with the pseudometrics $d_{\ell,\eta}$ and $d_{\ell',\eta'}$ respectively, introduced in Section 4.5.) As noted by Chowdhury and Mémoli (2018, Section 2), compactness of such correspondence spaces follows from a theorem of Blaschke (Burago et al., 2001, Theorem 7.3.8). Indeed, Blaschke's theorem implies that the space of all closed subsets of $H \times H'$ is compact. Since $\mathcal{C}_c(H, H')$ is closed in the set of all closed subsets of $H \times H'$, being the intersection of the preimages of H and H' under the projection maps onto the first and second coordinates respectively, we get that $\mathcal{C}_c(H, H')$ is compact as well.

The lower semi-continuity of $\mathrm{dis}_{P,P'}$ (Lemma 59) in particular implies that for any correspondence R and coupling γ, $\mathrm{dis}(R, \gamma) = \mathrm{dis}(\mathrm{cl}(R), \gamma)$, where cl represents topological closure. Indeed, since R and its closure are distance zero apart in the Hausdorff distance, the constant sequence R converges to $\mathrm{cl}(R)$ and vice versa. Hence, when computing the Risk distance, we can restrict our attention to closed correspondences:

$$d_{\mathrm{R}}(P, P') = \inf_{\substack{R \in \mathcal{C}(H,H') \\ \gamma \in \Pi(\eta,\eta')}} \mathrm{dis}_{P,P'}(R, \gamma) = \inf_{\substack{R \in \mathcal{C}_c(H,H') \\ \gamma \in \Pi(\eta,\eta')}} \mathrm{dis}_{P,P'}(R, \gamma).$$

Since $\mathrm{dis}_{P,P'}$ is lower semi-continuous, and the latter infimum is over a pair of compact sets, the infimum must be achieved. $\blacksquare$

Proof [Theorem 65] First consider an arbitrary finite partition Q of Y. Let $\pi : Y \to Q$ be the quotient map. Since $(\mathrm{Id}_X \times \pi)_\sharp \eta = \eta_Q$, the map $\mathrm{Id}_X \times \pi$ induces a coupling $\gamma_{(\mathrm{Id}_X \times \pi)}$ between η and η_Q. We also define a correspondence R_π between H and H_Q by pairing each $h \in H$ with $\pi \circ h \in H_Q$. Selecting this coupling and correspondence gives us a bound on the Risk distance:

$$
\begin{aligned}
d_{\mathrm{R}}(P, P_Q) &\leqslant \mathrm{dis}_{P,P_Q}(R_\pi, \gamma_{\mathrm{Id}_X \times \pi}) \\
&\leqslant \sup_{(h,\pi \circ h) \in R_\pi} \int \left| \ell(h(x), y) - \ell'(\pi \circ h(x'), y') \right| \gamma_{(\mathrm{Id}_X \times \pi)}(dx \times dy \times dx' \times dy') \\
&= \sup_{h \in H} \int \left| \ell(h(x), y) - \ell'(\pi \circ h(x), \pi(y)) \right| \eta(dx \times dy) \\
&= \sup_{h \in H} \int \sup_{\substack{y' \in \pi \circ h(x) \\ y'' \in \pi(y)}} \ell(y', y'') - \ell(h(x), y) \, \eta(dx \times dy) \\
&\leqslant \sup_{h \in H} \int \sup_{\substack{y' \in \pi \circ h(x) \\ y'' \in \pi(y)}} \ell(y', y'') - \inf_{\substack{y' \in \pi \circ h(x) \\ y'' \in \pi(y)}} \ell(y', y'') \, \eta(dx \times dy)
\end{aligned}
$$

79

$$\leqslant \sup_{h \in H} \sup_{(x,y) \in X \times Y} \left[\sup_{\substack{y' \in \pi \circ h(x) \\ y'' \in \pi(y)}} \ell(y', y'') - \inf_{\substack{y' \in \pi \circ h(x) \\ y'' \in \pi(y)}} \ell(y', y'') \right]$$

$$\leqslant \max_{q,q' \in Q} \left[\sup_{\substack{y' \in q' \\ y'' \in q}} \ell(y', y'') - \inf_{\substack{y' \in q' \\ y'' \in q}} \ell(y', y'') \right]. \tag{6}$$

The bound (6) holds for an arbitrary finite partition Q of Y.

Since P is a compact problem, we can pick a metric d_Y on Y with respect to which Y is compact. (Recall that we required our response spaces to be Polish, which makes Y metrizable.) In particular, $Y \times Y$ is a compact metric space under the product metric, given by

$$d_{Y \times Y}((y_1, y_1'), (y_2, y_2')) := \max\{d_Y(y_1, y_2), d_Y(y_1', y_2')\}.$$

By compactness, ℓ is uniformly continuous with respect to $d_{Y \times Y}$. Hence we can pick $\delta > 0$ such that, if $y_1, y_1', y_2, y_2' \in Y$ and

$$d_{Y \times Y}((y_1, y_1'), (y_2, y_2')) < \delta,$$

then

$$\left| \ell(y_1, y_1') - \ell(y_2, y_2') \right| < \varepsilon.$$

Again using compactness, we choose a partition Q of Y such that $\mathrm{diam}_{d_Y}(q) < \delta$ for all $q \in Q$. Then for any $q, q' \in Q$, $y_1, y_2 \in q$, $y_1', y_2' \in q'$, we have

$$d_{Y \times Y}((y_1, y_1'), (y_2, y_2')) = \max\{d_Y(y_1, y_2), d_Y(y_1', y_2')\} < \delta.$$

Hence

$$\left| \ell(y_1, y_1') - \ell(y_2, y_2') \right| < \varepsilon.$$

Taking a supremum over all choices of $q, q', y_1, y_2, y_1', y_2'$, combined with inequality (6), gives us the result. ∎

Proof [Theorem 69] We will proceed by approximating P and P_n by problems with finite input and output spaces. The process of approximation will proceed in steps depicted as follows:

$$P \xrightarrow{\text{Restrict to finite } H} P' \xrightarrow{\text{Replace } Y \text{ with partition } Q_Y} P'_{Q_Y} \xrightarrow{\text{Replace } X \text{ with partition } Q_X} P'_{Q_Y, Q_X}$$

A similar sequence of approximations will be constructed for the empirical problem P_n to produce $P_{Q_Y, Q_X, n}$. We then show that $P_{Q_Y, Q_X, n}$ converges a.s. to P_{Q_Y, Q_X}.

To wit, let $\varepsilon > 0$ and construct new problems as follows. Since H is compact, we can find a finite ε-net $H' \subseteq H$. Define

$$P' := (X, Y, \eta, \ell, H') \qquad\qquad P_n' := (X, Y, \eta_n, \ell, H').$$

By Theorem 65, we can select a partition Q_Y of Y such that the resulting coarsened problem P'_{Q_Y} is ε-close to P'. Furthermore, by Remark 66, we can choose Q_Y in a way that does not depend on the joint law η. We can coarsen P'_n with respect to the same partition Q_Y, giving us two new problems:

$$P'_{Q_Y} := (X, Q_Y, (\mathrm{Id}_X \times \pi_{Q_Y})_\sharp \eta, \ell_{Q_Y}, H'_{Q_Y}) \quad P'_{Q_Y,n} := (X, Q_Y, (\mathrm{Id}_X \times \pi_{Q_Y})_\sharp \eta_n, \ell_{Q_Y}, H'_{Q_Y}).$$

Recall that $\pi_{Q_Y} : Y \to Q_Y$ is the canonical quotient map, $\ell_{Q_Y}(q, q') := \sup_{y \in q, z \in q'} \ell(q, q')$, and $H'_{Q_Y} := \{\pi_{Q_Y} \circ h | h \in H'\}$.

For the third and final round of approximation, notice that $A := \{h^{-1}(q) | h \in H'_{Q_Y}, q \in Q_Y\}$ is a finite collection of subsets of X. Let Q_X be the partition of X generated by A. That is, Q_X is the partition of X induced by the equivalence relation $\sim$ where $x \sim x'$ means that $h(x) = h(x')$ for all $h \in H'_{Q_Y}$. Then define

$$P'_{Q_Y,Q_X} := (Q_X, Q_Y, (\pi_{Q_X} \times \pi_{Q_Y})_\sharp \eta, \ell_{Q_Y}, \widetilde{H'_{Q_Y}})$$
$$P'_{Q_Y,Q_X,n} := (Q_X, Q_Y, (\pi_{Q_X} \times \pi_{Q_Y})_\sharp \eta_n, \ell_{Q_Y}, \widetilde{H'_{Q_Y}}),$$

where $\pi_{Q_X} : X \to Q_X$ is the canonical quotient map, and $\widetilde{H'_{Q_Y}}$ is the collection of functions sending $\pi_{Q_X}(x) \mapsto h(x)$ as h ranges over H'_{Q_Y}.

We apply the triangle inequality to write

$$d_{\mathrm{R}}(P, P_n) \leqslant d_{\mathrm{R}}(P, P') \tag{1}$$
$$+ d_{\mathrm{R}}(P', P'_{Q_Y}) \tag{2}$$
$$+ d_{\mathrm{R}}(P'_{Q_Y}, P'_{Q_Y,Q_X}) \tag{3}$$
$$+ d_{\mathrm{R}}(P'_{Q_Y,Q_X}, P'_{Q_Y,Q_X,n}) \tag{4}$$
$$+ d_{\mathrm{R}}(P'_{Q_Y,Q_X,n}, P'_{Q_Y,n}) \tag{5}$$
$$+ d_{\mathrm{R}}(P'_{Q_Y,n}, P'_n) \tag{6}$$
$$+ d_{\mathrm{R}}(P'_n, P_n). \tag{7}$$

We consider each of these terms individually.

(1) Since P and P' have the same joint law, we can bound $d_{\mathrm{R}}(P, P')$ above by selecting the diagonal coupling. Additionally, since H' is an ε-net for H, we can select a correspondence $R \in \mathcal{C}(H, H')$ such that $d_{\ell,\eta}(h, h') < \varepsilon$ for all $(h, h') \in R$. This produces the bound

$$d_{\mathrm{R}}(P, P') \leqslant \sup_{(h,h') \in R} \int |\ell(h(x), y) - \ell(h'(x), y)| \, \eta(dx \times dy) = \sup_{(h,h') \in R} d_{\ell,\eta}(h, h') \leqslant \varepsilon.$$

(2) We chose Q_Y so that P' and P'_{Q_Y} would be ε-close under d_{R}, so this term is at most ε.

(3) We bound $d_{\mathrm{R}}(P'_{Q_Y}, P'_{Q_Y,Q_X})$ by selecting the coupling induced by $\pi_{Q_X} \times \mathrm{Id}_{Q_Y} : X \times Y \to Q_X \times Y$. We also select the correspondence R between H'_{Q_Y} and $\widetilde{H'_{Q_Y}}$ that pairs each

$h \in H'_{Q_Y}$ with its counterpart $h' \in \widetilde{H'_{Q_Y}}$ which sends $\pi_{Q_X}(x) \mapsto h(x)$. Then we obtain the bound

$$d_{\mathrm{R}}(P'_{Q_Y}, P'_{Q_Y,Q_X}) \leqslant \sup_{(h,h') \in R} \int \left| \ell(h(x), q) - \ell(h'(\pi_{Q_X}(x)), q) \right| (\pi_{Q_X} \times \mathrm{Id}_{Q_Y})_\sharp \eta(dx \times dq)$$

$$= \sup_{(h,h') \in R} \int \left| \ell(h(x), q) - \ell(h(x)), q) \right| (\pi_{Q_X} \times \mathrm{Id}_{Q_Y})_\sharp \eta(dx \times dq) = 0.$$

Alternatively, one could note that P'_{Q_Y} is a simulation of P'_{Q_Y,Q_X} via the maps π_{Q_X} and Id_{Q_Y} and apply Theorem 28.

(4) For brevity, define

$$\nu := (\pi_{Q_X} \times \pi_{Q_Y})_\sharp \eta$$
$$\nu_n := (\pi_{Q_X} \times \pi_{Q_Y})_\sharp \eta_n$$

We apply the total variation bound from Proposition 44 to get

$$d_{\mathrm{R}}(P'_{Q_Y,Q_X}, P'_{Q_Y,Q_X,n}) \leqslant \ell_{\mathsf{max}} d_{\mathrm{TV}}(\nu, \nu_n).$$

Note that ν_n is the empirical measure for ν, which is a measure supported on a finite set. Hence by the law of large numbers, the total variation distance between ν and ν_n converges to 0 a.s. as $n \to \infty$. Hence, with probability 1, for sufficiently large n, we have $d_{\mathrm{R}}(P'_{Q_Y,Q_X}, P'_{Q_Y,Q_X,n}) < \varepsilon$.

(5) The same calculation we applied to term (3) shows that $d_{\mathrm{R}}(P'_{Q_Y,Q_X,n}, P'_{Q_Y,n}) = 0$.

(6) Since we chose Q_Y in a way that did not depend on the joint law η, and since P' and P'_n only differ in their joint laws, we get $d_{\mathrm{R}}(P'_n, P'_{Q_Y,n}) < \varepsilon$.

(7) Consider the (random) function on $H \times H$ given by

$$d_{\ell,\eta_n}(h, h') := \int \left| \ell_h(x, y) - \ell_{h'}(x, y) \right| \eta_n(dx \times dy).$$

Let R be the same as in our calculations for term (1). Then we can write

$$d_{\mathrm{R}}(P'_n, P_n) \leqslant \sup_{(h,h') \in R} \int \left| \ell_h(x, y) - \ell_{h'}(x, y) \right| \eta_n(dx \times dy) = \sup_{(h,h') \in R} d_{\ell,\eta_n}(h, h').$$

Furthermore, since we assumed the integrands $|\ell_h - \ell_{h'}|$ are a Glivenko-Cantelli class with respect to η, we have that for sufficiently large n,

$$\sup_{(h,h') \in R} d_{\ell,\eta_n}(h, h') - \varepsilon \leqslant \sup_{(h,h') \in R} d_{\ell,\eta_n}(h, h') - d_{\ell,\eta}(h, h')$$
$$\leqslant \sup_{(h,h') \in R} \left| d_{\ell,\eta_n}(h, h') - d_{\ell,\eta}(h, h') \right| < \varepsilon.$$

Here the Glivenko-Cantelli assumption is used in the last inequality. Hence

$$d_{\mathrm{R}}(P'_n, P_n) \leqslant \sup_{(h,h') \in R} d_{\ell,\eta_n}(h, h') < 2\varepsilon.$$

Combining our bounds for all seven terms gives us that, with probability 1, for sufficiently large n we have

$$d_{\mathrm{R}}(P, P_n) < \varepsilon + \varepsilon + 0 + \varepsilon + 0 + \varepsilon + 2\varepsilon = 6\varepsilon.$$

Since ε was arbitrary, we conclude that $d_{\mathrm{R}}(P, P_n) \to 0$ with probability 1. ∎

A.4 Proofs from Section 6

Proof [Proposition 79] Let (X, d_X, μ_X) and (Y, d_Y, μ_Y) be two metric measure spaces. The map

$$\Pi((\Delta_X)_\sharp \mu_X, (\Delta_Y)_\sharp \mu_Y) \to \Pi(\mu_X, \mu_Y)$$
$$\gamma \mapsto (\pi_{2,4})_\sharp \gamma$$

is a bijection. Here, $\pi_{2,4} : X \times X \times Y \times Y \to X \times Y$ is the map $(x, x', y, y') \mapsto (x', y')$. Using this fact, we can write

$$d_{\mathrm{R},1}(P_X, P_Y)$$
$$= \inf_{\substack{\gamma \in \Pi((\Delta_X)_\sharp \mu_X, (\Delta_Y)_\sharp \mu_Y) \\ \rho \in \Pi(\mu_X, \mu_Y)}} \int \int \left| d_X(h(x_0), x_1) - d_Y(h'(y_0), y_1) \right| \gamma(dx_0 \times dx_1 \times dy_0 \times dy_1) \rho(dh \times dh')$$
$$= \inf_{\substack{\gamma \in \Pi((\Delta_X)_\sharp \mu_X, (\Delta_Y)_\sharp \mu_Y) \\ \rho \in \Pi(\mu_X, \mu_Y)}} \int \int \left| d_X(x_2, x_1) - d_Y(y_2, y_1) \right| \gamma(dx_0 \times dx_1 \times dy_0 \times dy_1) \rho(dx_2 \times dy_2)$$
$$= \inf_{\substack{\gamma \in \Pi((\Delta_X)_\sharp \mu_X, (\Delta_Y)_\sharp \mu_Y) \\ \rho \in \Pi(\mu_X, \mu_Y)}} \int \int \left| d_X(x_2, x_1) - d_Y(y_2, y_1) \right| (\pi_{2,4})_\sharp \gamma(dx_1 \times dy_1) \rho(dx_2 \times dy_2)$$
$$= \inf_{\substack{\gamma \in \Pi(\mu_X, \mu_Y) \\ \rho \in \Pi(\mu_X, \mu_Y)}} \int \int \left| d_X(x_2, x_1) - d_Y(y_2, y_1) \right| \gamma(dx_1 \times dy_1) \rho(dx_2 \times dy_2).$$

as desired. ∎

Proof [Theorem 81] First, for any $h \in H$, $h' \in H'$, we can bound the Wasserstein distance between their loss profiles by

$$d_{\mathrm{W},1}((\ell_h)_\sharp \eta, (\ell'_{h'})_\sharp \eta') = \inf_{\tau \in \Pi((\ell_h)_\sharp \eta, (\ell'_{h'})_\sharp \eta')} \int |a - b| \, \tau(da \times db)$$
$$\leqslant \inf_{\gamma \in \Pi(\eta, \eta')} \int |a - b| \, (\ell_h \times \ell'_{h'})_\sharp (da \times db)$$
$$= \inf_{\gamma \in \Pi(\eta, \eta')} \int |\ell_h(x, y) - \ell'_{h'}(x', y')| \, \gamma(dx \times dy \times dx' \times dy').$$

83

Similarly, we can bound

$$d_{W,p}^{d_{W,1}}(\mathcal{L}(P), \mathcal{L}(P')) = \inf_{\tau \in \Pi((F_P)\sharp\lambda, (F_P)\sharp\lambda')} \left(\int d_{W,1}^p(\mu, \nu) \, \tau(d\mu \times d\nu) \right)^{1/p}$$

$$\leqslant \inf_{\rho \in \Pi(\lambda, \lambda')} \left(\int d_{W,1}^p(\mu, \nu) \, (F_P \times F_{P'})\sharp\rho(d\mu \times d\nu) \right)^{1/p}$$

$$= \inf_{\rho \in \Pi(\lambda, \lambda')} \left(\int d_{W,1}^p((\ell_h)\sharp\eta, (\ell'_{h'})\sharp\eta') \, \rho(dh \times dh') \right)^{1/p}.$$

Combining these bounds gives

$$d_{W,p}^{d_{W,1}}(\mathcal{L}(P), \mathcal{L}(P'))$$

$$\leqslant \inf_{\rho \in \Pi(\lambda, \lambda')} \left(\int \left(\inf_{\gamma \in \Pi(\eta, \eta')} \int |\ell_h(x, y) - \ell'_{h'}(x', y')| \, \gamma(dx \times dy \times dx' \times dy') \right)^p \rho(dh \times dh') \right)^{1/p}$$

$$\leqslant \inf_{\rho, \gamma} \left(\int \left(\int |\ell_h(x, y) - \ell'_{h'}(x', y')| \, \gamma(dx \times dy \times dx' \times dy') \right)^p \rho(dh \times dh') \right)^{1/p}$$

$$= d_{R,p}(P, P'),$$

as desired. ■

Proof [Theorem 82] The support of a probability measure on a Polish space is σ-compact, so without loss of generality, we can assume that Y is σ-compact by replacing Y with the support of the Y-marginal of η. Write Y as an increasing sequence of compact subsets $(Y_n)_{n=1}^\infty$. For each n, define $\pi_n : Y \to Y_n \cup \{\bullet\}$ by

$$\pi_n(y) := \begin{cases} y & y \in Y_n \\ \bullet & \text{else} \end{cases}.$$

We then define a sequence of weighted problems (P_n, λ_n). Define

$$P_n := (X, Y_n \cup \{\bullet\}, \eta_n, \ell_n, H_n)$$

where

- $\eta_n := (\mathrm{Id}_X \times \pi_n)_\sharp \eta$

- $\ell_n(y, y') := 1_{Y_n \times Y_n}(y, y')\ell(y, y')$

- $H_n := \{\pi_n \circ h | h \in H\}$.

Furthermore, if we let $\pi_{n*} : H \to H_n$ be given by $\pi_{n*}(h) := \pi_n \circ h$, we can define $\lambda_n := (\pi_{n*})_\sharp \lambda$.

We claim that $d_{R,p}(P, P_n)$ converges to 0. Since $\eta_n = (\mathrm{Id}_X \times \pi_n)_\sharp \eta$, the function $\mathrm{Id}_X \times \pi_n$ induces a coupling between η and η_n. Similarly, since $\lambda_n = (\pi_{n*})_\sharp \lambda$, the function π_{n*} induces

a coupling between λ and λ_n. By definition of $d_{\mathrm{R},p}$, choosing these couplings produces a bound on $d_{\mathrm{R},p}(P, P_n)$.

$$d_{\mathrm{R},p}(P, P_n) \leq \left(\int \left(\int \left| \ell(h(x), y) - \ell'(\pi \circ h(x), \pi(y)) \right| \eta(dx \times dy) \right)^p \lambda(dh) \right)^{1/p}$$

$$= \left(\int \left(\int 1_{(Y_n \times Y_n)^c}(h(x), y) \, \ell(h(x), y) \, \eta(dx \times dy) \right)^p \lambda(dh) \right)^{1/p}.$$

Here the superscript c represents set complement. Notice that the indicator $1_{(Y_n \times Y_n)^c}(h(x), y)$ converges pointwise to 0 as $n \to \infty$. We apply the dominated convergence theorem twice to finish the proof. Indeed, by assumption,

$$\mathrm{diam}_p(P) = \left(\int \left(\int \ell(h(x), y) \, \eta(dx \times dy) \right)^p \lambda(dh) \right)^{1/p} < \infty. \tag{7}$$

In particular, (7) tells us that $\ell(h(x), y)$ is integrable as a function of (x, y) for λ-a.e. $h \in H$. Since $\ell(h(x), y)$ dominates $1_{(Y_n \times Y_n)^c}(h(x), y)\ell(h(x), y)$, the dominated convergence theorem tells us that

$$\int 1_{(Y_n \times Y_n)^c}(h(x), y)\ell(h(x), y) \, \eta(dx \times dy) \xrightarrow{n \to \infty} 0$$

for λ-a.e. $h \in H$. Additionally, (7) tells us that

$$\left(\int \ell(h(x), y) \, \eta(dx \times dy) \right)^p$$

is an integrable function of h. Using the dominated convergence theorem again,

$$\left(\int \left(\int 1_{(Y_n \times Y_n)^c}(h(x), y)\ell(h(x), y) \, \eta(dx \times dy) \right)^p \lambda(dh) \right)^{1/p} \xrightarrow{n \to \infty} 0$$

as desired. ∎

Proof [Proposition 83] Let $\varepsilon > 0$. Let Q be the partition of Y provided by Theorem 65, and let

$$P_Q = (X, Q, \eta_Q, \ell_Q, H_Q)$$

be the coarsened problem. Let $R_\pi \in \mathcal{C}(H, H_Q)$ be the correspondence induced by the surjection $\pi_*(h) = \pi \circ h$, and let $\gamma_{(\mathrm{Id}_X \times \pi)} \in \Pi(\eta, \eta_Q)$ be the coupling induced by the map $\mathrm{Id}_X \times \pi$. Recall the proof of Theorem 65, in which we demonstrated that $d_{\mathrm{R}}(P, P_Q) < \varepsilon$ by showing

$$\mathrm{dis}_{P, P_Q}(R_\pi, \gamma_{(\mathrm{Id}_X \times \pi)}) < \varepsilon.$$

Define a measure $\lambda_Q = (\pi_*)_\sharp \lambda$ on H_Q and let $\rho_\pi \in \Pi(\lambda, \lambda_Q)$ be the coupling induced by π_*. Then, since R_π is the support of ρ_π,

$$d_{\mathrm{R},p}((P, \lambda), (P_Q, \lambda_Q)) \leqslant \mathrm{dis}_{P,P_Q,p}(\rho_\pi, \gamma_{\mathrm{Id}_X \times \pi})$$

$$= \left(\int \left(\int \left| \ell(h(x), y) - \ell_Q(h'(x'), y') \right| \gamma_{\mathrm{Id}_X \times \pi}(dx \times dy \times dx' \times dy') \right)^p \rho_\pi(dh \times dh') \right)^{1/p}$$

$$\leqslant \sup_{(h,h') \in R_\pi} \int \left| \ell(h(x), y) - \ell_Q(h'(x'), y') \right| \gamma_{\mathrm{Id}_X \times \pi}(dx \times dy \times dx' \times dy')$$

$$= \mathrm{dis}_{P,P_Q}(R_\pi, \gamma_{\mathrm{Id}_X \times \pi}) < \varepsilon,$$

as desired. ∎

Proof [Theorem 91] Proposition 89 provides the bound $d_{\mathrm{R}}(P, P_N) \leqslant d_{\mathrm{W},1}^{s_{\ell,\lambda,p}}(\eta, \eta \cdot N)$. One can check that if τ is a coupling of the Markov kernels $\delta_{\mathrm{Id}_{X \times Y}}$ and N, then $\eta \cdot \tau \in \Pi(\eta, \eta \cdot N)$. Therefore we can write

$$d_{\mathrm{W},1}^{s_{\ell,\lambda,p}}(\eta, \eta \cdot N)$$

$$= \inf_{\gamma \in \Pi(\eta, \eta \cdot N)} \int_{(X \times Y)^2} s_{\ell,\lambda,p}((x, y), (x', y')) \, \gamma(dx \times dy \times dx' \times dy')$$

$$\leqslant \inf_{\tau \in \Pi(\delta_{\mathrm{Id}_{X \times Y}}, N)} \int_{(X \times Y)^2} s_{\ell,\lambda,p}((x, y), (x', y')) \, (\eta \cdot \tau)(dx \times dy \times dx' \times dy')$$

$$= \inf_{\tau \in \Pi(\delta_{\mathrm{Id}_{X \times Y}}, N)} \int_{X \times Y} \int_{(X \times Y)^2} s_{\ell,\lambda,p}((x_1, y_1), (x_2, y_2)) \, \tau(x, y)(dx_1 \times dy_1 \times dx_2 \times dy_2) \, \eta(dx \times dy)$$

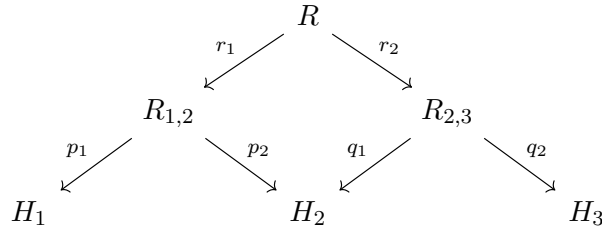$$= d_{\mathrm{W},1}^{s_{\ell,\lambda,p}}(\delta_{\mathrm{Id}_{X \times Y}}, N),$$

which is the desired bound. ∎

A.5 Proofs from Section 7

Proof [Lemma 95] Let

$$R := \{(h_1, h_2, h_3) \in H_1 \times H_2 \times H_3 | (h_1, h_2) \in R_{1,2}, (h_2, h_3) \in R_{2,3}\}.$$

We have the following diagram.

Every pictured map is a projection map. We need only check that they are inverse connected, meaning that the preimage of any connected set is connected. Let $A \subseteq H_1$ be a connected

set. Then $p_1^{-1}(A)$ is connected, and so is $B := p_2(p_1^{-1}(A))$. Similarly, $C := q_2(q_1^{-1}(B))$ is connected. Noting finally that the preimage of A in R is

$$r_1^{-1}(p_1^{-1})(A) = A \times B \times C,$$

we see that it is a product of connected sets and hence connected. The proof for preimages of connected sets in H_2 and H_3 is similar. ∎

Proof [Theorem 100] For the first inequality, suppose $\widehat{f}, \widehat{g} : Z \to \mathbb{R}$ are continuous functions inducing the Reeb graphs $\mathrm{Reeb}(P)$ and $\mathrm{Reeb}(P')$ respectively. Then

$$\|\widehat{f} - \widehat{g}\|_\infty \geqslant \left| \inf_{z \in Z} \widehat{f}(z) - \inf_{z \in Z} \widehat{g}(z) \right| = \left| \inf_{h \in H} \mathcal{R}_P(h) - \inf_{h' \in H'} \mathcal{R}_{P'}(h') \right| = |\mathcal{B}(P) - \mathcal{B}(P')|$$

where the equality comes from the fact that the infimum of a function can be determined from the Reeb graph that it induces.

Now suppose $d_{\mathrm{RC}}(P, P') < a$. Then we can pick a correspondence $R \in \mathcal{C}_{\mathrm{con}}(H, H')$ that witnesses this fact, giving us the following diagram.

$$
\begin{array}{ccc}
& R & \\
{\scriptstyle p_1} \swarrow & & \searrow {\scriptstyle p_2} \\
H & & H' \\
\downarrow {\scriptstyle R_P} & & \downarrow {\scriptstyle R_{P'}} \\
\mathbb{R} & & \mathbb{R}
\end{array}
$$

Recall $\mathcal{R}_P$ is the risk function associated with problem P. We can then induce two Reeb graphs from functions on R: one induced by $\mathcal{R}_P \circ p_1$ and the other induced by $\mathcal{R}_{P'} \circ p_2$. The former Reeb graph is isomorphic to $\mathrm{Reeb}(P)$ since p_1 is a continuous surjection with connected fibers, and similarly the latter is isomorphic to $\mathrm{Reeb}(P')$. Hence

$$
\begin{aligned}
d_{\mathrm{U}}(\mathrm{Reeb}(P), \mathrm{Reeb}(P')) &\leqslant \sup_{(h,h') \in R} |\mathcal{R}_P \circ p_1(h, h') - \mathcal{R}_P \circ p_1(h, h')| \\
&= \sup_{(h,h') \in R} |\mathcal{R}_P(h) - \mathcal{R}_P(h')| \\
&\leqslant \inf_{\gamma \in \Pi(\eta, \eta')} \sup_{(h,h') \in R} \int \left| \ell_h(x, y) - \ell_{h'}(x', y') \right| \gamma(dx \times dy \times dx' \times dy') < a.
\end{aligned}
$$

Hence $d_{\mathrm{U}}(\mathrm{Reeb}(P), \mathrm{Reeb}(P')) < a$ as well. ∎

Appendix B. Weak Isomorphism with Classification Problems

While the infimum in the definition of the Risk distance makes it difficult to characterize when two problems are distance zero apart in general, the closely related notion of weak isomorphism is simpler to deal with. For instance, we can characterize which problems are weakly isomorphic to classification problems.

Definition 104 *Call a problem P rectangular if there exists a finite partition $Y_1, \ldots, Y_m$ of the response space Y and, for any $h \in H$, a finite partition $X_1, \ldots, X_n$ of the input space X such that*

$$\ell_h(x, y) = \sum_{ij} a_{ij} 1_{X_i^h}(x) 1_{Y_j}(y) \quad \eta\text{-}a.e..$$

A problem P is rectangular, for instance, if the loss is a sum of indicators on rectangles. That is, if

$$\ell(y_1, y_2) = \sum_{ij} a_{ij} 1_{Y_i}(y_1) 1_{Y_j}(y_2)$$

for some partition $Y_1, \ldots, Y_n \subseteq Y$, then P is rectangular.

Theorem 105 *A problem is weakly isomorphic to a classification problem if and only if it is rectangular.*

The proof makes use of the notion of a *generated partition*. Let S be a set and let $T_1, \ldots, T_n$ be a collection of subsets of S. Then the *partition of S generated by $T_1, \ldots, T_n$* is the partition of S induced by the equivalence relation $\sim$ where $s \sim s'$ means that $s \in T_i \iff s' \in T_i$ for all $1 \leqslant i \leqslant n$. Note that since there are finitely many T_i, the generated partition has finitely many blocks as well.

Proof Suppose P is rectangular and let $Q := \{Y_1, \ldots, Y_m\}$, $X_1, \ldots, X_n$, and $(a_{ij})_{ij}$ be as in the definition of a rectangular problem, and let $\pi_Q : Y \to Q$ be the quotient map. If $m \neq n$, pad the partitions appropriately with copies of the empty set and pad the matrix $(a_{ij})_{ij}$ with zeros to make $m = n$. Define a classification problem

$$P' := (X, Q, (\mathrm{Id}_X \times \pi_Q)_\sharp \eta, \ell', H')$$

where

- $\ell'(Y_i, Y_j) := a_{ij}$

- $H' = \{\phi(h) | h \in H\}$, where $\phi(h) : X \to Q$ sends x to Y_i if $x \in X_i^h$.

We claim P' is a simulation of P via the maps Id_X and π_Q. Indeed, for any $h \in H$, we can calculate that $\ell'_{\phi(h)}(x, \pi_Q(y)) = a_{ij}$ if $x \in X_i^h$ and $y \in Y_j$. Hence $\ell'_{\phi(h)}(x, \pi_Q(y)) = \ell_h(x, y)$ for any $h \in H$, showing that P' is a simulation of P. In particular, the two are weakly isomorphic.

For the right hand implication, we proceed in two parts. First suppose that P is a simulation of a classification problem P', whose response space is $Y' = \{y'_1, \ldots, y'_m\}$. Let the simulation be via the maps $f_1 : X \to X'$ and $f_2 : Y \to Y'$. We will show that P is rectangular. By the definition of simulation, for any $h \in H$, we can select an $h' \in H'$ such that $\ell_h(x, y) = \ell'_{h'}(f_1(x), f_2(y))$ η-almost everywhere. Since the function ℓ' is on a finite set $Y' \times Y'$, we can write it with a finite sum of indicators to get that, η-almost everywhere,

$$\ell_h(x, y) = \ell'(h' \circ f_1(x), f_2(y)) = \sum_{ij} \ell'(y'_i, y'_j) 1_{Y_i}(h' \circ f_1(x)) 1_{Y_j}(f_2(y))$$

$$= \sum_{ij} \ell'(y'_i, y'_j) 1_{h' \circ f_1^{-1}(Y_i)}(x) 1_{f_2^{-1}(Y_j)}(y).$$

Hence P is rectangular.

Next we prove that if P is rectangular and simulates P', then P' must be rectangular as well. Let the simulation again be via maps $f_1 : X \to X'$ and $f_2 : Y \to Y'$, and let $\{X_i\}_{i=1}^m$, $\{Y_i\}_{i=1}^m$ and $(a_{ij})_{ij}$ be as in the definition of a rectangular problem. Consider the family

$$\mathcal{Y} := \{f_2[Y_i] | 1 \leqslant i \leqslant m\}$$

of subsets of Y'. The sets in $\mathcal{Y}$ do not necessarily form a partition of Y', so consider the partition $Q_{Y'}$ generated by $\mathcal{Y}$. Now let $h' \in H'$ be arbitrary and, using the definition of a simulation, pick $h \in H$ such that the diagram

commutes. Consider the family

$$\mathcal{X}^{h'} := \left\{ f_1[X_i^h] \Big| 1 \leqslant i \leqslant m \right\}.$$

Again, these sets do not necessarily partition X', so let $Q_{X'}^{h'}$ be the partition of X' generated by $\mathcal{X}^{h'}$.

We claim that, if $A \in Q_{X'}^{h'}$ and $B \in Q_{Y'}$, then $\ell_{h'}'$ is constant η'-almost everywhere on $A \times B$. Indeed, by construction of $Q_{X'}^{h'}$ and $Q_{Y'}$, we have that $f_1[X_i^h] \supseteq A$ and $f_2[Y_j] \supseteq B$ for some i and j. Since ℓ_h is constant on $X_i^h \times Y_j$ η-almost everywhere, by commutativity of the above diagram we also have that $\ell_{h'}'$ is constant η'-almost everywhere on $(f_1 \times f_2)[X_i^h \times Y_j] \supseteq A \times B$. Therefore, $\ell_{h'}'$ can be written as a linear combination of indicators on such rectangles:

$$\ell_{h'}'(x', y') = \sum_{ij} a_{ij} 1_{A_i}(x) 1_{B_j}(y)$$

for some $(a_{ij})_{ij}$. Here we have numbered the blocks in the partitions $Q_{X'}^{h'} = \{A_1, \ldots, A_m\}$ and $Q_{Y'} = \{B_1, \ldots, B_n\}$. Hence P' is rectangular.

Now, if P is weakly isomorphic to a classification problem P' via a common simulation P'', we have shown that P'', as a simulation of a classification problem P', must be rectangular. Furthermore, since P'' is rectangular and simulates P, we have shown that P must be rectangular as well. ∎

References

Margareta Ackerman. *Towards Theoretical Foundations of Clustering*. PhD thesis, University of Waterloo, 2012.

Pankaj K. Agarwal, Kyle Fox, Abhinandan Nath, Anastasios Sidiropoulos, and Yusu Wang. Computing the Gromov-Hausdorff distance for metric trees. In Khaled Elbassioni and Kazuhisa Makino, editors, *Algorithms and Computation*, Lecture Notes in Computer Science, pages 529–540, Berlin, Heidelberg, 2015. Springer. ISBN 978-3-662-48971-0. doi: 10.1007/978-3-662-48971-0_45.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21428–21439, 2020.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214.

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, Tokyo, 2016. ISBN 978-4-431-55977-1 978-4-431-55978-8. doi: 10.1007/978-4-431-55978-8.

Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*, volume 130 of *UNITEXT*. Springer International Publishing, Cham, 2021. ISBN 978-3-030-72161-9 978-3-030-72162-6. doi: 10.1007/978-3-030-72162-6.

Shreya Arya, Arnab Auddy, Ranthony A Clark, Sunhyuk Lim, Facundo Mémoli, and Daniel Packer. The Gromov–Wasserstein distance between spheres. *Foundations of Computational Mathematics*, pages 1–56, 2024.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1117–1174. PMLR, June 2022. ISSN: 2640-3498.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete 34*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-56477-7 978-3-319-56478-4. doi: 10.1007/978-3-319-56478-4.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. ISSN 1533-7928.

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, March 2006. ISSN 0162-1459. doi: 10.1198/016214505000000907.

Ulrich Bauer, Claudia Landi, and Facundo Mémoli. The Reeb graph edit distance is universal. *Foundations of Computational Mathematics*, 21(5):1441–1464, October 2021. ISSN 1615-3383. doi: 10.1007/s10208-020-09488-3.

Florian Beier, Robert Beinert, and Gabriele Steidl. On a linear Gromov–Wasserstein distance. *IEEE Transactions on Image Processing*, 31:7292–7305, 2022.

Florian Beier, Robert Beinert, and Gabriele Steidl. Multi-marginal Gromov–Wasserstein transport and barycentres. *Information and Inference: A Journal of the IMA*, 12(4): 2753–2781, 2023.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985.

Silvia Biasotti, Daniela Giorgi, Michela Spagnuolo, and Bianca Falcidieno. Reeb graphs for shape analysis and applications. *Theoretical Computer Science*, 392(1):5–22, February 2008. ISSN 0304-3975. doi: 10.1016/j.tcs.2007.10.018.

Garrett Birkhoff and Erwin Kreyszig. The establishment of functional analysis. *Historia Mathematica*, 11(3):258–321, 1984.

Chris M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 93–102. University of California Press, Berkeley-Los Angeles, Calif., 1951.

Vladimir I. Bogachev. *Measure Theory*. Springer, 2007. ISBN 978-3-540-34513-8 978-3-540-34514-5. doi: 10.1007/978-3-540-34514-5.

Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 851–861. PMLR, 2019.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

Gunnar Carlsson and Facundo Mémoli. Classifying clustering schemes. *Foundations of Computational Mathematics*, 2013. doi: 10.1007/s10208-012-9141-9.

Gunnar E Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11(Apr): 1425–1470, 2010.

Frédéric Chazal and Jian Sun. Gromov-Hausdorff approximation of metric spaces with linear structure, May 2013. arXiv:1305.1172.

Samantha Chen, Sunhyuk Lim, Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Weisfeiler-Lehman meets Gromov-Wasserstein. In *International Conference on Machine Learning*, pages 3371–3416. PMLR, 2022.

Samir Chowdhury and Facundo Mémoli. Distances and isomorphism between networks: stability and convergence of network invariants. *Journal of Applied and Computational Topology*, 7(2):243–361, Jun 2023. ISSN 2367-1734. doi: 10.1007/s41468-022-00105-6.

Samir Chowdhury and Facundo Mémoli. Explicit geodesics in Gromov-Hausdorff space. *Electronic Research Announcements*, 25(0):48–59, June 2018. ISSN 1935-9179. doi: 10.3934/era.2018.25.006. Publisher: Electronic Research Announcements.

Samir Chowdhury and Facundo Mémoli. The Gromov-Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4): 757–787, December 2019. ISSN 2049-8772. doi: 10.1093/imaiai/iaz026.

Samir Chowdhury and Tom Needham. Gromov-Wasserstein averaging in a Riemannian framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 842–843, 2020.

Samir Chowdhury, David Miller, and Tom Needham. Quantized Gromov-Wasserstein. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, Lecture Notes in Computer Science, pages 811–827, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86523-8. doi: 10.1007/978-3-030-86523-8_49.

Vincent Cohen-Addad, Varun Kanade, and Frederik Mallmann-Trenn. Clustering redemption–beyond the impossibility of Kleinberg's axioms. *Advances in Neural Information Processing Systems*, 31, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Vin de Silva, Elizabeth Munch, and Amit Patel. Categorified Reeb graphs. *Discrete & Computational Geometry*, 55(4):854–906, Jun 2016. ISSN 1432-0444. doi: 10.1007/s00454-016-9763-9.

Julie Delon, Agnes Desolneux, and Antoine Salmona. Gromov–Wasserstein distances between gaussian distributions. *Journal of Applied Probability*, 59(4):1178–1198, 2022.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, New York, NY, 1996. ISBN 978-1-4612-6877-2 978-1-4612-0711-5. doi: 10.1007/978-1-4612-0711-5.

Oussama Dhifallah and Yue Lu. On the inherent regularization effects of noise injection during training. In *International Conference on Machine Learning*, pages 2665–2675. PMLR, 2021.

Jean Dieudonné. *History of Functional Analysis*. North-Holland, 1981.

Théo Dumont, Théo Lacombe, and François-Xavier Vialard. On the existence of Monge maps for the Gromov–Wasserstein problem. *Foundations of Computational Mathematics*, pages 1–48, 2024.

Herbert Edelsbrunner, John Harer, Ajith Mascarenhas, and Valerio Pascucci. Time-varying reeb graphs for continuous space-time data. In *Proceedings of the twentieth annual symposium on Computational geometry*, SCG '04, pages 366–372, New York, NY, USA,

June 2004. Association for Computing Machinery. ISBN 978-1-58113-885-6. doi: 10.1145/997817.997872.

David A. Edwards. The structure of superspace. In Nick M. Stavrakas and Keith R. Allen, editors, *Studies in Topology*, pages 121–133. Academic Press, January 1975. ISBN 978-0-12-663450-1. doi: 10.1016/B978-0-12-663450-1.50017-7.

Francisco Escolano, Edwin R. Hancock, and Silvia Biasotti. Complexity fusion for indexing Reeb digraphs. In Richard Wilson, Edwin Hancock, Adrian Bors, and William Smith, editors, *Computer Analysis of Images and Patterns*, Lecture Notes in Computer Science, pages 120–127, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-40261-6. doi: 10.1007/978-3-642-40261-6_14.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport, April 2022. arXiv:2110.03684.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Peter Gaenssler. Around the Glivenko–Cantelli theorem. In Révész Pál, editor, *Limit theorems of probability theory*, pages 93 – 103. North-Holland, 1975. ISBN 0720428343.

Xiaoyin Ge, Issam Safa, Mikhail Belkin, and Yusu Wang. Data skeletonization via Reeb graphs. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Mikhael Gromov. *Structures Métriques pour les Variétés Riemanniennes*. Birkhäuser, 1981.

Mikhael Gromov. Convergence and concentration of metrics and measures. In *Metric Structures for Riemannian and Non-Riemannian Spaces*, Modern Birkhäuser Classics, pages 113–237. Birkhäuser, Boston, MA, 2007. ISBN 978-0-8176-4583-0. doi: 10.1007/978-0-8176-4583-0_4.

Peter D. Grünwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: From erm to generalized Bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. ISSN 1533-7928.

Haibin Hang, Facundo Mémoli, and Washington Mio. A topological study of functional data and Fréchet functions of metric measure spaces. *Journal of Applied and Computational Topology*, 3(4):359–380, 2019.

Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2005.

Felix Hausdorff. *Grundzüge der Mengenlehre*. Von Veit, 1914. ISBN 978-3-11-098985-4.

Masaki Hilaga, Yoshihisa Shinagawa, Taku Kohmura, and Tosiyasu L. Kunii. Topology matching for fully automatic similarity estimation of 3D shapes. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '01, pages 203–212, New York, NY, USA, August 2001. Association for Computing Machinery. ISBN 978-1-58113-374-5. doi: 10.1145/383259.383282.

Chip Huyen. *Designing Machine Learning Systems*. O'Reilly Media, Inc., May 2022. ISBN 9781098107963.

Laura Iacovissi, Nan Lu, and Robert C. Williamson. A general framework for learning under corruption: Label noise, attribute noise, and beyond, July 2023. arXiv:2307.08643.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. doi: 10.1007/978-1-0716-1418-1.

Nigel Kalton and Mikhail Ostrovskii. Distances between Banach spaces. *Forum Mathematicum*, 11, 09 1997. doi: 10.1515/form.11.1.17.

Leonid Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133 (4):1381–1382, March 2006. ISSN 1573-8795. doi: 10.1007/s10958-006-0049-2.

Jun Kitagawa and Asuka Takatsu. Two new families of metrics via optimal transport and barycenter problems, December 2023. arXiv:2311.15874.

Jon Kleinberg. An impossibility theorem for clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

Achim Klenke. Probability measures on product spaces. In Achim Klenke, editor, *Probability Theory: A Comprehensive Course*, Universitext, pages 273–293. Springer, London, 2014. ISBN 978-1-4471-5361-0. doi: 10.1007/978-1-4471-5361-0_14.

Aleksandr Kronrod. On functions of two variables. *Uspekhi Mat. Nauk*, 5:24–134, 1950.

John Langford. Machine learning reductions. 2009. URL https://hunch.net/~jl/projects/reductions/reductions.html.

John Langford and Bianca Zadrozny. Estimating class membership probabilities using classifier learners. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 198–205. PMLR, 06–08 Jan 2005.

Kim G Larsen and Arne Skou. Bisimulation through probabilistic testing (preliminary report). In *Proceedings of the 16th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 344–352, 1989.

Khang Le, Dung Q Le, Huy Nguyen, Dat Do, Tung Pham, and Nhat Ho. Entropic Gromov–Wasserstein between Gaussian distributions. In *International Conference on Machine Learning*, pages 12164–12203. PMLR, 2022.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964. ISSN 00034851.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, January 1993. ISSN 0893-6080. doi: 10.1016/S0893-6080(05) 80131-5.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times: Second Edition.* American Mathematical Society, 2017.

Renjie Liao. Notes on Rademacher complexity, May 2020.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73–82, June 2004. ISSN 0167-7152. doi: 10.1016/j.spl.2004.03.002.

Saunders Mac Lane. *Categories for the Working Mathematician*, volume 5. Springer Science & Business Media, 2013.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23803–23828. PMLR, July 2023. ISSN: 2640-3498.

Ester Mariucci. Le Cam theory on the comparison of statistical models, 2016. arXiv:1605.03301.

Facundo Mémoli. On the use of Gromov-Hausdorff Distances for Shape Comparison. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007. ISBN 978-3-905673-51-7. doi: 10.2312/SPBG/SPBG07/081-090.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, August 2011. ISSN 1615-3375, 1615-3383. doi: 10.1007/s10208-011-9093-5.

Facundo Mémoli, Zane Smith, and Zhengchao Wan. The Gromov-Hausdorff distance between ultrametric spaces: its structure and computation. *Journal of Computational Geometry*, 14(1):78–143, Sep. 2023. doi: 10.20382/jocg.v14i1a4.

Xiao-Li Meng. Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 125–134. PMLR, June 2015. ISSN: 1938-7228.

Aditya Krishna Menon, Brendan van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, September 2018. ISSN 1573-0565. doi: 10.1007/s10994-018-5715-3.

Robin Milner. *A Calculus of Communicating Systems.* Springer, 1980.

Robin Milner. *Communication and Concurrency*, volume 84. Prentice hall Englewood Cliffs, 1989.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning, second edition*. Adaptive Computation and Machine Learning series. MIT Press, 2018. ISBN 978-0-262-35136-2.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

NeurIPS. AI research summit: NeurIPS 2025 receives record-breaking 27,000 paper submissions, 2025. URL `https://www.ctol.digital/news/ai-research-summit-neurips-2025-receives-record-breaking-27000-paper-submissions/`. Accessed 9 June 2025.

Richard Nock, Zac Cranko, Aditya K. Menon, Lizhen Qu, and Robert C. Williamson. f-GANs in an information geometric nutshell. *Advances in Neural Information Processing Systems*, 30, 2017.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 271–279, 2016.

Giuseppe Patanè, Michela Spagnuolo, and Bianca Falcidieno. Para-Graph: Graph-based parameterization of triangle meshes with arbitrary genus. *Computer Graphics Forum*, 23 (4):783–797, 2004. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2004.00808.x.

Evan Patterson. Hausdorff and Wasserstein metrics on graphs and other structured data. *Information and Inference: A Journal of the IMA*, iaaa025, 2020. doi: 10.1093/imaiai/iaaa025.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2664–2672. PMLR, June 2016. ISSN: 1938-7228.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, January 1999. ISSN 1474-0508, 0962-4929. doi: 10.1017/S0962492900002919. Publisher: Cambridge University Press.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. Neural Information Processing. The MIT Press, Cambridge, MA, 2009. ISBN 978-0262170055.

Georges Reeb. Sur les points singuliers d'une forme de Pfaff completement integrable ou d'une fonction numerique [on the singular points of a completely integrable Pfaff form or of a numerical function]. *Comptes Rendus Acad. Sciences Paris*, 222:847–849, 1946.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011.

Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018. ISSN 1533-7928.

Jonas Rothfuss, Fabio Ferreira, Simon Boehm, Simon Walther, Maxim Ulrich, Tamim Asfour, and Andreas Krause. Noise regularization for conditional density estimation. *arXiv preprint arXiv:1907.08982*, 2019.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov-Wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

Felix Schmiedl. Computational aspects of the Gromov–Hausdorff distance and its application in non-rigid shape matching. *Discrete & Computational Geometry*, 57(4):854–880, June 2017. ISSN 1432-0444. doi: 10.1007/s00454-017-9889-4.

Thibault Sejourne, Francois-Xavier Vialard, and Gabriel Peyré. The unbalanced Gromov-Wasserstein distance: Conic formulation and relaxation. In *Advances in Neural Information Processing Systems*, volume 34, pages 8766–8779, 2021.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000. ISSN 0378-3758. doi: 10.1016/S0378-3758(00)00115-4.

Y. Shinagawa, T.L. Kunii, and Y.L. Kergosien. Surface coding based on Morse theory. *IEEE Computer Graphics and Applications*, 11(5):66–78, September 1991. ISSN 1558-1756. doi: 10.1109/38.90568. Conference Name: IEEE Computer Graphics and Applications.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, August 2007. ISSN 1432-0940. doi: 10.1007/s00365-006-0662-3.

Karl-Theodor Sturm. On the geometry of metric measure spaces. *Acta Mathematica*, 196 (1):65–131, July 2006. ISSN 1871-2509. doi: 10.1007/s11511-006-0002-8.

Karl-Theodor Sturm. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*, volume 290. American Mathematical Society, 2023.

Dilip Mathew Thomas and Vijay Natarajan. Symmetry in scalar field topology. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2035–2044, December 2011. ISSN 1941-0506. doi: 10.1109/TVCG.2011.236. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

Erik N. Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991. ISBN 978-1-107-08722-4.

Lloyd N. Trefethen. *Approximation Theory and Approximation Practice, Extended Edition.* Society for Industrial and Applied Mathematics, 2019. ISBN 978-1-61197-593-2.

Aad W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes.* Springer Series in Statistics. Springer, New York, NY, 1996. ISBN 978-1-4757-2547-6 978-1-4757-2545-2. doi: 10.1007/978-1-4757-2545-2.

Tim Van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *The Journal of Machine Learning Research*, 16(1):1793–1861, January 2015. ISSN 1532-4435.

titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced gromov-wasserstein. *Advances in Neural Information Processing Systems*, 32, 2019.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13:212, 2020a.

Titouan Vayer, Ievgen Redko, Rémi Flamary, and Nicolas Courty. CO-optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pages 17559–17570, 2020b.

Anatoly Vershik. Kantorovich metric: Initial history and little-known applications. *Journal of Mathematical Sciences*, 133(4):1410–1417, March 2006. ISSN 1573-8795.

Cédric Villani. *Topics in Optimal Transportation Theory.* American Mathematical Society, 01 2003. doi: 10.1090/gsm/058.

Cédric Villani. *Optimal Transport – Old and New.* Springer-Verlag, 01 2008. doi: 10.1007/978-3-540-71050-9.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed Gromov Wasserstein divergence with applications on graphs. *ArXiv*, abs/2110.02753, 2021.

Robert C. Williamson and Zac Cranko. Information processing equalities and the information–risk bridge. *Journal of Machine Learning Research*, 25(103):1–53, 2024. ISSN 1533-7928.

Zoë Wood, Hugues Hoppe, Mathieu Desbrun, and Peter Schröder. Removing excess topology from isosurfaces. *ACM Transactions on Graphics*, 23(2):190–208, April 2004. ISSN 0730-0301. doi: 10.1145/990002.990007.

Lin Yan, Talha Bin Masood, Raghavendra Sridharamurthy, Farhan Rasheed, Vijay Natarajan, Ingrid Hotz, and Bei Wang. Scalar field comparison with topological descriptors: Properties and applications for scientific visualization. *Computer Graphics Forum*, 40(3):599–633, 2021. ISSN 1467-8659. doi: 10.1111/cgf.14331.

Eugene Zhang, Konstantin Mischaikow, and Greg Turk. Feature-based surface parameterization and texture mapping. *ACM Transactions on Graphics*, 24(1):1–27, January 2005. ISSN 0730-0301. doi: 10.1145/1037957.1037958.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004. ISSN 0090-5364. Publisher: Institute of Mathematical Statistics.

Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210, November 2004. ISSN 1573-7462. doi: 10.1007/s10462-004-0751-8.