Minimax Statistical Learning and Domain Adaptation with Wasserstein Distances

Jaeho Lee† Maxim Raginsky‡

Abstract

As opposed to standard empirical risk minimization (ERM), distributionally robust optimization aims to minimize the worst-case risk over a larger ambiguity set containing the original empirical distribution of the training data. In this work, we describe a minimax framework for statistical learning with ambiguity sets given by balls in Wasserstein space. In particular, we prove a generalization bound that involves the covering number properties of the original ERM problem. As an illustrative example, we provide generalization guarantees for domain adaptation problems where the Wasserstein distance between the source and target domain distributions can be reliably estimated from unlabeled samples.

1 Introduction and problem set-up

In the traditional paradigm of statistical learning [1], we have a class $\mathcal{P}$ of probability measures on a measurable instance space $\mathcal{Z}$ and a class $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}_+$. Each $f \in \mathcal{F}$ quantifies the loss of some decision rule or a hypothesis applied to instances $z \in \mathcal{Z}$, so, with a slight abuse of terminology, we will refer to $\mathcal{F}$ as the hypothesis space. The (expected) risk of a hypothesis $f$ on instances generated according to $P$ is given by

$$R(P, f) := \mathbb{E}_P[f(Z)] = \int_{\mathcal{Z}} f(z) P(dz).$$

Given an $n$-tuple $Z_1, \ldots, Z_n$ of i.i.d. training examples drawn from an unknown $P \in \mathcal{P}$, the objective is to find a hypothesis $\hat{f} \in \mathcal{F}$ whose risk $R(P, \hat{f})$ is close to the minimum risk

$$R^*(P, \mathcal{F}) := \inf_{f \in \mathcal{F}} R(P, f)$$

with high probability. Under suitable regularity assumptions, this objective can be accomplished via Empirical Risk Minimization (ERM) [1, 2]:

$$R(P_n, f) = \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \rightarrow \min_{f \in \mathcal{F}} R(P, f)$$

This work was supported in part by the NSF grant nos. CIF-1527388 and CIF-1302438, and in part by the NSF CAREER award 1254041.

†jlee620@illinois.edu
‡maxim@illinois.edu
where $P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}$ is the empirical distribution of the training examples.

Recently, however, an alternative viewpoint has emerged, inspired by ideas from robust statistics and robust stochastic optimization. In this \textit{distributionally robust framework}, instead of solving the \textit{ERM problem} (1.2), one aims to solve the minimax problem

$$
\sup_{Q \in \mathcal{A}(P_n)} R(Q, f) \rightarrow \min, \; f \in \mathcal{F}
$$

where $\mathcal{A}(P_n)$ is an \textit{ambiguity set} containing the empirical distribution $P_n$ and, possibly, the unknown probability law $P$ either with high probability or almost surely. The ambiguity sets serve as a mechanism for compensating for the uncertainty about $P$ that inherently arises due to having only a finite sample to work with, and they can be constructed in a variety of ways, e.g., via moment constraints [3], $f$-divergence balls [4], and Wasserstein balls [5, 6, 7]. However, with the exception of the recent work by Farnia and Tse [3], the minimizer of (1.3) is still evaluated under the standard statistical risk minimization paradigm. In this work, we instead propose replacing the statistical risk minimization criterion (1.1) with the \textit{local minimax risk}

$$
\inf_{f \in \mathcal{F}} \sup_{Q \in \mathcal{A}(P)} R(Q, f)
$$

at $P$, where the ambiguity set $\mathcal{A}(P)$ is taken to be a Wasserstein ball centered at $P$. (Recently, this modification was also proposed by Farnia and Tse [3] for ambiguity sets defined by a finite number of moment constraints.) As we will argue below, this change of perspective is natural when there is a possibility of \textit{domain drift}, i.e., when the learned hypothesis is evaluated on a distribution $Q$ which may be different from the distribution $P$ that was used to generate the training data.

## 1.1 Wasserstein ambiguity sets and local minimax risk

We assume that the instance space $\mathcal{Z}$ is a Polish space (i.e., a complete separable metric space) with metric $d_{\mathcal{Z}}$. We denote by $\mathcal{P}(\mathcal{Z})$ the space of all Borel probability measures on $\mathcal{Z}$, and by $\mathcal{P}_p(\mathcal{Z})$ with $p \geq 1$ the space of all $P \in \mathcal{P}(\mathcal{Z})$ with finite $p$th moments:

$$
\mathcal{P}_p(\mathcal{Z}) := \{ P \in \mathcal{P}(\mathcal{Z}) : E_P[d_{\mathcal{Z}}^p(\mathcal{Z}, z_0)] < \infty \text{ for some } z_0 \in \mathcal{Z} \}.
$$

The metric structure of $\mathcal{Z}$ can be used to define a family of metrics on the spaces $\mathcal{P}_p(\mathcal{Z})$ [8]:

**Definition 1.1.** For $p \geq 1$, the $p$-Wasserstein distance between $P, Q \in \mathcal{P}_p(\mathcal{Z})$ is

$$
W_p(P, Q) := \inf_{M(\times \mathcal{Z}) = P, M(\mathcal{Z} \times \cdot) = Q} \left( \mathbb{E}_M[d_{\mathcal{Z}}^p(Z, Z')] \right)^{1/p},
$$

(1.4)

where the infimum is over all couplings of $P$ and $Q$, i.e., probability measures $M$ on the product space $\mathcal{Z} \times \mathcal{Z}$ with the given marginals $P$ and $Q$.

**Remark 1.1.** Wasserstein metrics arise in the problem of \textit{optimal transport}: for any coupling $M$ of $P$ and $Q$, the conditional distribution $M_{\mathcal{Z}'|\mathcal{Z}}$ can be viewed as a randomized policy for ‘transporting’ a unit quantity of some material from a random location $Z$ to another location $Z'$, while satisfying the marginal constraint $Z' \sim Q$. If the cost of transporting a unit of material from $z \in \mathcal{Z}$ to $z' \in \mathcal{Z}$ is given by $d_{\mathcal{Z}}^p(z, z')$, then $W_p^p(P, Q)$ is the minimum expected transport cost. \qed
We now consider a learning problem \((P, F)\) with \(P = \mathcal{P}_p(Z)\) for some \(p \geq 1\). Following [5, 6, 7], we let the ambiguity set \(\mathcal{A}(P)\) be the \(p\)-Wasserstein ball of radius \(\varepsilon \geq 0\) centered at \(P\):

\[
\mathcal{A}(P) = B^W_{\varepsilon,p}(P) := \{Q \in \mathcal{P}_p(Z) : W_p(P, Q) \leq \varepsilon\},
\]

where the radius \(\varepsilon > 0\) is a tunable parameter. We then define the local worst-case risk of \(f\) at \(P\),

\[
R_{\varepsilon,p}(P, f) := \sup_{Q \in B^W_{\varepsilon,p}(P)} R(Q, f),
\]

and the local minimax risk at \(P\):

\[
R^*_{\varepsilon,p}(P, F) := \inf_{f \in F} R_{\varepsilon,p}(P, f).
\]

Some inequalities relating the local worst-case risk \(R_{\varepsilon,p}(Q, f)\) and the statistical risk \(R(Q, f)\) are presented in Section 2.

The basic problem of interest can now be stated as follows: given an \(n\)-tuple \(Z_1, \ldots, Z_n\) of i.i.d. training examples drawn from an unknown \(P \in \mathcal{P}_p(Z)\), find a hypothesis \(\hat{f} \in F\), such that

\[
R_{\varepsilon,p}(P, \hat{f}) \approx R^*_{\varepsilon,p}(P, F) \quad \text{with high probability.}
\]

In Section 3, we show that, under some regularity assumptions, this goal can be achieved via a natural Empirical Risk Minimization (ERM) procedure. In particular, Theorem 3.1 provides a high-probability bound on the excess risk \(R_{\varepsilon,p}(P, \hat{f}) - R^*_{\varepsilon,p}(P, F)\).

1.2 Motivating problem: domain adaptation

One of the attractive features of ambiguity sets based on Wasserstein distances is that, because of their intimate connection to the metric geometry of the instance space, they provide a natural mechanism for handling uncertainty due to (possibly randomized) transformations acting on the problem instances. To illustrate this point, we briefly discuss a motivating example of domain adaptation.

In contrast to the standard statistical learning framework, where the risk of the learned hypothesis is evaluated on the same unknown distribution that was used for generating the training examples, the problem of domain adaptation [9] arises when the training data are generated according to an unknown distribution \(P\), but the learned hypothesis is evaluated on another unknown distribution \(Q\). However, it is assumed that these distributions (commonly referred to as problem domains) are somehow related, and some partial information about \(Q\) is also available at training time. In the context of supervised learning, the instances are random couples \(Z = (X, Y)\) consisting of features \(X\) and labels \(Y\), and the training data consist of \(n\) labeled examples \(Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)\) drawn from the source domain \(P\) and \(m\) unlabeled features \(X'_1, \ldots, X'_m\) drawn from the target domain \(Q\). The goal is to learn a hypothesis that would perform well on the target domain \(Q\).

In a recent paper, Courty et al. [10] have introduced an algorithmic framework for domain adaptation based on optimal transport. Their approach revolves around a particular generative model for the drift between the source and the target domains. Let us disintegrate the source domain distribution as \(P = \mu \otimes P_{Y|X}\), where \(\mu \in \mathcal{P}(X)\) is the marginal distribution of the features and \(P_{Y|X}\) is the conditional distribution of the labels given the features. Then there exists an unknown deterministic transformation \(T : \mathcal{X} \rightarrow \mathcal{X}\) of the feature space, such that a sample \((X', Y')\)
from the target domain distribution \( Q = \nu \otimes Q_{Y|X} \) can be generated using the following two-step procedure:

\[
X \xrightarrow{T} X' \xrightarrow{P_{Y|X}} Y',
\]

where the input \( X \) is drawn from \( \mu \). In other words, the domain drift is due solely to an unknown deterministic transformation of the features. If we further assume that \( T \) is the optimal transport map from \( \mu \) to \( \nu \), i.e., that \( W_p^\mu(\mu, \nu) = \mathbb{E}_\mu[d_\mu^p(X, T(X))] \) under some metric \( d_\mu \) on \( X \), and if \( W_p(P, Q) = W_p(\mu, \nu) \), then it is natural to cast the problem of domain adaptation as that of learning a hypothesis \( f \in \mathcal{F} \) that would approximately achieve the local minimax risk \( R_{\varepsilon,p}(P, \mathcal{F}) \), where \( \varepsilon \) is some estimate of \( W_p(\mu, \nu) \) from source and target training data. In Section 4, we present an algorithm based on this idea and provide a quantitative analysis of its performance. In particular, Theorem 4.1 gives a high-probability bound on the excess risk of the learned classifier with respect to the target domain distribution \( Q \). We note that, in contrast to the original methodology of Courty et al. [10], our approach completely bypasses the problem of estimating the transport map \( T \).

2 Local worst-case risk vs. statistical risk

In some situations (see, e.g., Section 4), it is of interest to convert back and forth between local worst-case (or local minimax) risks and the usual statistical risks. In this section, we give a couple of inequalities relating these quantities. The first one is a simple consequence of the Kantorovich duality theorem from the theory of optimal transport [8]:

**Proposition 2.1.** Suppose that \( f \) is \( L \)-Lipschitz, i.e., \( |f(z) - f(z')| \leq L d_Z(z, z') \) for all \( z, z' \in \mathcal{Z} \). Then, for any \( Q \in B_{\varepsilon,p}^W(P) \),

\[
R(Q, f) \leq R_{\varepsilon,p}(P, f) \leq R(Q, f) + 2L \varepsilon.
\]

As an example, consider the problem of binary classification with hinge loss: \( \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \), where \( \mathcal{X} \) is an arbitrary feature space, \( \mathcal{Y} = \{-1, +1\} \), and the hypothesis space \( \mathcal{F} \) consists of all functions of the form \( f(z) = f(x, y) = \max\{0, 1 - y f_0(x)\} \), where \( f_0 : \mathcal{X} \to \mathbb{R} \) is a candidate predictor. Then, since the function \( u \mapsto \max\{0, 1 - u\} \) is Lipschitz-continuous with constant 1, we can write

\[
|f(x, y) - f(x', y')| \leq |y f_0(x) - y' f_0(x')| \leq 2\|f_0\|_\mathcal{X} \mathbf{1}\{y \neq y'\} + |f_0(x) - f_0(x')|,
\]

where \( \|f_0\|_\mathcal{X} := \sup_{x \in \mathcal{X}} |f_0(x)| \). If \( \|f_0\|_\mathcal{X} < \infty \) and if \( f_0 \) is \( L_0 \)-Lipschitz with respect to some metric \( d_\mathcal{X} \) on \( \mathcal{X} \), then it follows that \( f \) is Lipschitz with constant \( \max\{2\|f_0\|_\mathcal{X}, L_0\} \) with respect to the product metric

\[
d_Z(z, z') = d_Z((x, y), (x', y')) := d_\mathcal{X}(x, x') + \mathbf{1}\{y \neq y'\}.
\]

Next we consider the case when the function \( f \) is smooth but not Lipschitz-continuous. Since we are working with general metric spaces that may lack an obvious differentiable structure, we need to first introduce some concepts from metric geometry [11]. A metric space \( (\mathcal{Z}, d_\mathcal{Z}) \) is a geodesic space if for every two points \( z, z' \in \mathcal{Z} \) there exists a path \( \gamma : [0, 1] \to \mathcal{Z} \), such that \( \gamma(0) = z, \gamma(1) = z' \), and \( d_\mathcal{Z}(\gamma(s), \gamma(t)) = (t - s) \cdot d_\mathcal{Z}(\gamma(0), \gamma(1)) \) for all \( 0 \leq s \leq t \leq 1 \) (such a path is called a constant-speed
geodesic). A functional $F : Z \to \mathbb{R}$ is **geodesically convex** if for any pair of points $z, z' \in Z$ there is a constant-speed geodesic $\gamma$, so that

$$F(\gamma(t)) \leq (1 - t)F(\gamma(0)) + tF(\gamma(1)) = (1 - t)F(z) + tF(z'), \quad \forall t \in [0, 1].$$

An **upper gradient** of a Borel function $f : Z \to \mathbb{R}$ is a functional $G_f : Z \to \mathbb{R}_+$, such that for any pair of points $z, z' \in Z$ there exists a constant-speed geodesic $\gamma$ obeying

$$|f(z') - f(z)| \leq \int_0^1 G_f(\gamma(t))dt \cdot d_Z(z, z'). \quad (2.1)$$

With these definitions at hand, we have the following:

**Proposition 2.2.** Suppose that $f$ has a geodesically convex upper gradient $G_f$. Then

$$R(Q, f) \leq R_{\varepsilon,p}(P, f) \leq R(Q, f) + 2\varepsilon \sup_{Q \in B_{\varepsilon,p}(P)} \|G_f(Z)\|_{L^q(Q)},$$

where $1/p + 1/q = 1$, and $\|\cdot\|_{L^q(Q)} := (\mathbb{E}_Q|\cdot|^q)^{1/q}$.

As a simple example, consider the setting of regression with quadratic loss: let $\mathcal{X}$ be a convex subset of $\mathbb{R}^d$, let $\mathcal{Y} = [-B, B]$ for some $0 < B < \infty$, and equip $Z = \mathcal{X} \times \mathcal{Y}$ with the Euclidean metric

$$d_Z(z, z') = \sqrt{|x - x'|^2 + |y - y'|^2}, \quad z = (x, y), z' = (x', y').$$

Suppose that the functions $f \in \mathcal{F}$ are of the form $f(z) = f(x, y) = (y - f_0(x))^2$ with $f_0 \in C^1(\mathbb{R}^d, \mathbb{R})$, such that $\|f_0\|_{\mathcal{X}} \leq M < \infty$ and $\|\nabla f_0(x)\|_2 \leq L\|x\|_2$ for some $0 < L < \infty$. Then Proposition 2.2 leads to the following:

**Proposition 2.3.**

$$R(Q, f) \leq R_{\varepsilon,2}(P, f) \leq R(Q, f) + 4\varepsilon(B + M)(1 + L \sup_{Q \in B_{\varepsilon,1}(P)} \sigma_{Q,X}),$$

where $\sigma_{Q,X} := \mathbb{E}_Q\|X\|_2$ for $Z = (X, Y) \sim Q$.

### 3 Guarantees for empirical risk minimization

Let $Z_1, \ldots, Z_n$ be an $n$-tuple of i.i.d. training examples drawn from $P$. In this section, we analyze the performance of the ERM procedure

$$\hat{f} := \arg\min_{f \in \mathcal{F}} R_{\varepsilon,p}^*(P_n, f). \quad (3.1)$$

The following strong duality result due to Gao and Kleywegt [7] will be instrumental:

**Proposition 3.1.** For any upper semicontinuous function $f : Z \to \mathbb{R}$ and for any $Q \in \mathcal{P}_p(Z)$,

$$R_{\varepsilon,p}(Q, f) = \min_{\lambda \geq 0} \left\{ \lambda \varepsilon^p + \mathbb{E}_Q[\varphi_{\lambda,f}(Z)] \right\},$$

where $\varphi_{\lambda,f}(z) := \sup_{z' \in Z} \left\{ f(z') - \lambda \cdot d_Z^p(z, z') \right\}$. 

5
We begin by imposing some regularity assumptions:

**Assumption 3.1.** The instance space $\mathcal{Z}$ is bounded: $\text{diam}(\mathcal{Z}) := \sup_{z, z' \in \mathcal{Z}} d_Z(z, z') < \infty$.

**Assumption 3.2.** The functions in $\mathcal{F}$ are upper semicontinuous and uniformly bounded: $0 \leq f(z) \leq M < \infty$ for all $f \in \mathcal{F}$ and $z \in \mathcal{Z}$.

**Assumption 3.3.** There exists a hypothesis $f_0 \in \mathcal{F}$, such that, for all $z \in \mathcal{Z}$, $f_0(z) \leq C_0 d_Z^p(z, z_0)$ for some $C_0 > 0$ and $z_0 \in \mathcal{Z}$.

We can now give a performance guarantee for the ERM procedure (3.1):

**Theorem 3.1.** If Assumptions 3.1–3.3 are satisfied, then, with probability at least $1 - \delta$,

$$
R_{\epsilon, p}(P, \hat{f}) - R_{\epsilon, p}^*(P, \mathcal{F}) \leq \frac{24}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, u/2)} du
$$

$$
+ \frac{24C_0(2 \text{diam}(\mathcal{Z}))^p}{\sqrt{n}} \left(1 + \left(\frac{\text{diam}(\mathcal{Z})}{\epsilon}\right)^p\right) + 3M \sqrt{\frac{\log(2/\delta)}{2n}}, \quad (3.2)
$$

where $\mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, \cdot)$ are the covering numbers of $\mathcal{F}$ with respect to the sup norm.

**Remark 3.1.** The first term on the right-hand side of (3.1) is the Dudley entropy integral [12]. We conjecture that this term can be replaced by the expected Rademacher average of the hypothesis class $\mathcal{F}$, but have been unable to prove it.

**Remark 3.2.** The second term on the right-hand side of (3.1) increases as $\epsilon \to 0$. The excess risk bound of Farnia and Tse [3] has the same behavior, where in that case $\epsilon$ is the slack in the moment constraints defining the ambiguity set. This $(1/\epsilon)^p$ scaling can be eliminated if more refined bounds on the optimum dual $\lambda$ are available.

**Proof.** Let $f^* \in \mathcal{F}$ be any achiever of the local minimax risk $R_{\epsilon, p}^*(P, \mathcal{F})$. We start by decomposing the excess risk:

$$
R_{\epsilon, p}(P, \hat{f}) - R_{\epsilon, p}^*(P, \mathcal{F}) = R_{\epsilon, p}(P, \hat{f}) - R_{\epsilon, p}(P, f^*)
$$

$$
\leq R_{\epsilon, p}(P, \hat{f}) - R_{\epsilon, p}(P_n, \hat{f}) + R_{\epsilon, p}(P_n, f^*) - R_{\epsilon, p}(P, f^*),
$$

where the last step follows from the definition of $\hat{f}$. Define

$$
\hat{\lambda} := \arg\min_{\lambda \geq 0} \left\{ \lambda \epsilon^p + \mathbf{E}_{P_n}[\varphi_{\lambda, \hat{f}}(Z)] \right\}, \quad \lambda^* := \arg\min_{\lambda \geq 0} \left\{ \lambda \epsilon^p + \mathbf{E}_P[\varphi_{\lambda, f^*}(Z)] \right\}.
$$

Then, using Proposition 3.1, we can write

$$
R_{\epsilon, p}(P, \hat{f}) - R_{\epsilon, p}(P_n, \hat{f}) = \min_{\lambda \geq 0} \left\{ \lambda \epsilon^p + \int_Z \varphi_{\lambda, \hat{f}}(z) P(dz) \right\} - \left(\hat{\lambda} \epsilon^p + \int_Z \varphi_{\hat{\lambda}, \hat{f}}(z) P_n(dz) \right)
$$

$$
\leq \int_Z \varphi_{\hat{\lambda}, \hat{f}}(z) (P - P_n)(dz)
$$

and, following similar logic,

$$
R_{\epsilon, p}(P_n, f^*) - R_{\epsilon, p}(P, f^*) \leq \int_Z \varphi_{\lambda^*, f^*}(z) (P_n - P)(dz).
$$

(3.3)
By Lemma 3.1 given in Section 3.2, $\hat{\lambda} \in \Lambda := [0, C_0 2^{p-1}(1 + (\text{diam}(\mathcal{Z})/\varepsilon)p)]$. Hence, defining the function class $\Phi := \{\varphi_{\lambda,f} : \lambda \in \Lambda, f \in \mathcal{F}\}$, we have

$$R_{\varepsilon,p}(P, \hat{f}) - R_{\varepsilon,p}(P_n, \hat{f}) \leq \sup_{\varphi \in \Phi} \left[ \int_{\mathcal{Z}} \varphi d(P - P_n) \right].$$

(3.4)

Since all $f \in \mathcal{F}$ take values in $[0, M]$, the same holds for all $\varphi \in \Phi$. Therefore, by a standard symmetrization argument,

$$R_{\varepsilon,p}(P, \hat{f}) - R_{\varepsilon,p}(P_n, \hat{f}) \leq 2 \mathfrak{R}_n(\Phi) + M \sqrt{\frac{2 \log(2/\delta)}{n}}$$

with probability at least $1 - \delta/2$, where

$$\mathfrak{R}_n(\Phi) := \mathbb{E} \left[ \sup_{\varphi \in \Phi} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \varphi(Z_i) \right]$$

is the expected Rademacher average of $\Phi$, with i.i.d. Rademacher random variables $\sigma_1, \ldots, \sigma_n$ independent of $Z_1, \ldots, Z_n$. Moreover, from (3.3) and from Hoeffding’s inequality it follows that

$$R_{\varepsilon,p}(P_n, f^*) - R_{\varepsilon,p}(P, f^*) \leq M \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability at least $1 - \delta/2$. Consequently,

$$R_{\varepsilon,p}(P, \hat{f}) - R_{\varepsilon,p}(P, \mathcal{F}) \leq 2 \mathfrak{R}_n(\Phi) + 3M \sqrt{\frac{\log(2/\delta)}{2n}}$$

(3.5)

with probability at least $1 - \delta$. Using the bound of Lemma 3.2 from Section 3.2 in (3.5), we obtain the statement of the theorem.

3.1 Example bounds

In this subsection, we illustrate the use of Theorem 3.1 when (upper bounds on) the covering numbers for the hypothesis class $\mathcal{F}$ are available. Throughout this section, we let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where the feature space $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq r_0\}$ is a ball of radius $r_0$ in $\mathbb{R}^d$ with center at the origin, the label space $\mathcal{Y} \subseteq [-B, +B]$ for some $0 < B < \infty$. We equip $\mathcal{Z}$ with the Euclidean metric

$$d_{\mathcal{Z}}(z, z') = d_{\mathcal{Z}}((x, y), (x', y')) := \sqrt{\|x - x'\|_2^2 + |y - y'|^2}$$

and take $p = 2$.

We first consider a simple neural network class $\mathcal{F}$ consisting of functions of the form $f(z) = f(x, y) = (y - s(f_0^T x))^2$, where $s : \mathbb{R} \to \mathbb{R}$ is a bounded smooth nonlinearity with $s(0) = 0$ and with bounded first derivative, and where $f_0$ takes values in the unit ball in $\mathbb{R}^d$.

**Corollary 3.1.** For any $P \in \mathcal{P}(\mathcal{Z})$, with probability at least $1 - \delta$,

$$R_{\varepsilon,2}(P, \hat{f}) - R_{\varepsilon,2}^*(P, \mathcal{F}) \leq \frac{C_1}{\sqrt{n}} + \frac{384(r_0 + B)^2}{\sqrt{n}} \left( 1 + \frac{4(r_0^2 + B^2)}{\varepsilon^2} \right) + \frac{6\|s\|_\infty^2 + B^2}{\sqrt{2n}} \sqrt{\log(2/\delta)}$$
where $C_1$ is a constant dependent only on $d, r_0, s$:

$$C_1 = 288\sqrt{dr_0}(B + \|s\|_\infty)\|s'\|_\infty.$$ 

We also consider the case of a massive nonparametric class. Let $(\mathcal{H}_K, \| \cdot \|_K)$ be the Gaussian reproducing kernel Hilbert space (RKHS) with the kernel $K(x_1, x_2) = \exp \{-\|x_1 - x_2\|_2^2/\sigma^2\}$ for some $\sigma > 0$, and let $B_r := \{ h \in \mathcal{H}_K : \|h\|_K \leq r \}$ be the radius-$r$ ball in $\mathcal{H}_K$. Let $\mathcal{F}$ be the class of all functions of the form $f(z) = f(x, y) = (y - f_0(x))^2$, where the predictors $f_0 : \mathcal{X} \to \mathbb{R}$ belong to $I_K(B_r)$, an embedding of $B_r$ into the space $C(\mathcal{X})$ of continuous real-valued functions on $\mathcal{X}$ equipped with the sup norm $\|f\|_{\mathcal{X}} := \sup_{x \in \mathcal{X}} |f(x)|$.

In order to apply Theorem 3.1, we need to control the covering numbers $\mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, \cdot)$. To that end, we need the following estimate due to Cucker and Zhou [13, Thm 5.1] (which was later shown by Kühn [14] to be asymptotically exact up to the double logarithmic factor):

**Proposition 3.2.** For compact $\mathcal{X} \subseteq \mathbb{R}^d$,

$$\log \mathcal{N}(I_K(B_r), \| \cdot \|_{\mathcal{X}}, u) \leq d \left( 32 + \frac{640d(diam(\mathcal{X}))^2}{\sigma^2} \right)^{d+1} \left( \log \frac{r}{u} \right)^{d+1}$$

holds for all $0 < u \leq r/2$.

Using Proposition 3.2, we can prove the following generalization bound for Gaussian RKHS.

**Corollary 3.2.** With probability at least $1 - \delta$, for any $P \in \mathcal{P}(\mathcal{Z})$,

$$R_{s,2}(P, \tilde{f}) - R_{s,2}^*(P, \mathcal{F}) \leq \frac{C_1}{\sqrt{n}} \left( r^2 + Br \right) + \frac{384(r_0 + B)^2}{\sqrt{n}} \left( 1 + \frac{4(r_0^2 + B^2)}{\varepsilon^2} \right) + \frac{6(r^2 + B^2)\sqrt{\log(2/\delta)}}{\sqrt{2n}}$$

where $C_1$ is a constant dependent only on $d, r_0, \sigma$:

$$C_1 = 48\sqrt{d} \left( 2\Gamma \left( \frac{d + 3}{2}, \log 2 \right) + (\log 2)^{\frac{d+1}{2}} \right) \left( 32 + \frac{2560d^2r_0^2}{\sigma^2} \right)^{\frac{d+1}{2}},$$

and $\Gamma(s, v) := \int_v^\infty u^{s-1} e^{-u} du$ is the incomplete gamma function.

### 3.2 Technical lemmas for the proof of Theorem 3.1

**Lemma 3.1.** Fix some $Q \in \mathcal{P}_p(\mathcal{Z})$. Define $\tilde{f} \in \mathcal{F}$ and $\tilde{\lambda} \geq 0$ via

$$\tilde{f} := \arg\min_{f \in \mathcal{F}} R_{s,p}(Q, f) \quad \text{and} \quad \tilde{\lambda} := \arg\min_{\lambda \geq 0} \left\{ \lambda \varepsilon^p + \mathbb{E}_Q [\varphi_{\lambda, f}(\mathcal{Z})] \right\}.$$

Then

$$\tilde{\lambda} \leq C_0 2^{p-1} \left( 1 + \left( \frac{\text{diam}(\mathcal{Z})}{\varepsilon} \right)^p \right). \quad (3.6)$$

**Lemma 3.2.** The expected Rademacher complexity of the function class $\Phi$ satisfies

$$\mathcal{R}_n(\Phi) \leq \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \| \cdot \|_{\infty}, u/2)} du + \frac{12C_0(2 \text{diam}(\mathcal{Z}))^p}{\sqrt{n}} \left( 1 + \left( \frac{\text{diam}(\mathcal{Z})}{\varepsilon} \right)^p \right).$$
4 Application to domain adaptation

As discussed in Section 1.2, the problem of domain adaptation arises when we want to transfer the data or knowledge from a source domain $P \in \mathcal{P}(Z)$ to a different but related target domain $Q \in \mathcal{P}(Z)$ [9]. Suppose that it is possible to estimate the Wasserstein distance $W_p(P, Q)$ between the two domain distributions. Then, as we show below, we can provide a generalization bound for the target domain by combining estimation guarantees for $W_p(P, Q)$ with risk inequalities of Section 2.

We work in the setting considered by Courty et al. [10]: Let $Z = \mathcal{X} \times \mathcal{Y}$, where $(\mathcal{X}, d_\mathcal{X})$ is the feature space and $(\mathcal{Y}, d_\mathcal{Y})$ is the label space. We endow $Z$ with the $\ell_p$ product metric

$$d_Z(z, z') = d_Z((x, y), (x', y')) := (d_\mathcal{X}(x, x') + d_\mathcal{Y}(y, y'))^{1/p}.$$  

Let $P = \mu \otimes P_{Y|X}$ and $Q = \nu \otimes Q_{Y|X}$ be the source and the target domain distributions, respectively. We assume that domain drift is due to an unknown (possibly nonlinear) transformation $T : \mathcal{X} \to \mathcal{X}$ of the feature space that preserves the conditional distribution of the labels given the features. That is, $\nu = T_\# \mu$, the pushforward of $\mu$ by $T$, and for any $x \in \mathcal{X}$ and any measurable set $B \subseteq \mathcal{Y}$

$$P_{Y|X}(B|x) = Q_{Y|X}(B|T(x)).$$  

(4.1)

This assumption leads to the following lemma, which enables us to estimate $W_p(P, Q)$ only from unlabeled source domain data and unlabeled target domain data:

**Lemma 4.1.** Suppose there exists a deterministic and invertible optimal transport map $T : \mathcal{X} \to \mathcal{X}$ such that $\nu = T_\# \mu$, i.e., $W_p^p(\mu, \nu) = E_{\mu}[d_\mathcal{X}^p(X, T(X))]$. Then

$$W_p(P, Q) = W_p(\mu, \nu).$$  

(4.2)

**Remark 4.1.** If $\mathcal{X}$ is a convex subset of $\mathbb{R}^d$ endowed with the $\ell_p$ metric $d_\mathcal{X}(x, x') = \|x - x'\|_p$ for $p \geq 2$, then, under the assumption that $\mu$ and $\nu$ have positive densities with respect to the Lebesgue measure, the (unique) optimal transport map from $\mu$ to $\nu$ is deterministic and a.e. invertible – in fact, its inverse is equal to the optimal transport map from $\nu$ to $\mu$ [8].  

Now suppose that we have $n$ labeled examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from $P$ and $m$ unlabeled examples $X'_1, \ldots, X'_m$ from $\nu$. Define the empirical distributions

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \quad \nu_m = \frac{1}{m} \sum_{j=1}^{m} \delta_{X'_j}. $$

Notice that, by the triangle inequality, we have

$$W_p(\mu, \nu) \leq W_p(\mu, \mu_n) + W_p(\mu_n, \nu_m) + W_p(\nu, \nu_m).$$  

(4.3)

Here, $W_p(\mu_n, \nu_m)$ can be computed from unlabeled data by solving a finite-dimensional linear program [8], and the following convergence result of Fournier and Guillin [15] implies that, with high probability, both $W_p(\mu, \mu_n)$ and $W_p(\nu, \nu_m)$ rapidly converge to zero as $n, m \to \infty$.
Proposition 4.1. Let $\mu$ be a probability distribution on a bounded set $\mathcal{X} \subset \mathbb{R}^d$, where $d > 2p$. Let $\mu_n$ denote the empirical distribution of $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mu$. Then, for any $r \in (0, \infty)$,

$$
\mathbb{P}(W_p(\mu_n, \mu) \geq r) \leq C_a \exp(-C_b nr^{d/p}) 
$$

(4.4)

where $C_a, C_b$ are constants depending on $p, d, \text{diam}(\mathcal{X})$ only.

Based on these considerations, we propose the following domain adaptation scheme:

1. Compute the $p$-Wasserstein distance $W_p(\mu_n, \nu_m)$ between the empirical distributions of the features in the labeled training set from the source domain $P$ and the unlabeled training set from the target domain $Q$.

2. Set the desired confidence parameter $\delta \in (0, 1)$ and the radius

$$
\hat{\varepsilon}(\delta) := W_p(\mu_n, \nu_m) + \left( \frac{\log(4C_a/\delta)}{C_b n} \right)^{p/d} + \left( \frac{\log(4C_a/\delta)}{C_b m} \right)^{p/d}. 
$$

(4.5)

3. Compute the empirical risk minimizer

$$
\hat{f} = \arg \min_{f \in \mathcal{F}} R_{\hat{\varepsilon}(\delta), p}(P_n, f),
$$

(4.6)

where $P_n$ is the empirical distribution of the $n$ labeled samples from $P$.

We can give the following target domain generalization bound for the hypothesis generated according to (4.6):

Theorem 4.1. Suppose that the feature space $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$ with $d > 2p$, take $d_\mathcal{X}(x, x') = \|x - x'\|_p$, and let $\mathcal{F}$ be a family of hypotheses with Lipschitz constant at most $L$. Then with probability at least $1 - \delta$, the empirical risk minimizer $\hat{f}$ from (4.6) satisfies

$$
R(Q, \hat{f}) - R(Q, \mathcal{F}) \leq 2L\hat{\varepsilon}(\delta) + C_1 \sqrt{n} + C_2 \sqrt{n} \left( 1 + \left( \frac{\text{diam}(\mathcal{Z})}{\hat{\varepsilon}(\delta)} \right)^p \right) + \frac{3M \sqrt{\log(4/\delta)}}{\sqrt{2n}},
$$

where

$$
C_1 = 24 \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}; \| \cdot \|_{\infty}, u/2)} du \quad \text{and} \quad C_2 = 24C_0(2\text{diam}(\mathcal{Z}))^p.
$$

Proof. For simplicity, we assume that there exists a hypothesis $f^* \in \mathcal{F}$ that achieves $R(Q, \mathcal{F})$. Then, for any $\varepsilon > 0$ such that $W_p(P, Q) \leq \varepsilon$, Proposition 2.1 implies that

$$
R(Q, \hat{f}) - R(Q, f^*) \leq R_{\varepsilon, p}(P, \hat{f}) - R_{\varepsilon, p}(P, f^*) + 2L\varepsilon
\leq R_{\varepsilon, p}(P, \hat{f}) - R_{\varepsilon, p}(P, \mathcal{F}) + 2L\varepsilon.
$$

From Theorem 3.1, we know that

$$
R_{\varepsilon, p}(P, \hat{f}) - R_{\varepsilon, p}(P, \mathcal{F}) \leq C_1 \sqrt{n} + C_2 \sqrt{n} \left( 1 + \left( \frac{\text{diam}(\mathcal{Z})}{\varepsilon} \right)^p \right) + \frac{3M \sqrt{\log(4/\delta)}}{\sqrt{2n}}.
$$
holds with probability at least $1 - \delta/2$. Thus, it remains to find the right $\varepsilon$, such that that $W_p(P, Q) \leq \varepsilon$ holds with high probability. From Proposition 4.1, we see that each of the following two statements holds with probability at least $1 - \delta/4$:

$$W_p(\mu_n, \mu) \leq \left( \frac{\log(4C_a/\delta)}{C_bn} \right)^{p/d}, \quad W_p(\nu_m, \nu) \leq \left( \frac{\log(4C_a/\delta)}{C_bm} \right)^{p/d}. $$

Since $W_p(P, Q) = W_p(\mu, \nu)$ by Lemma 4.1, we see that $W_p(P, Q) \leq \tilde{\varepsilon}(\delta)$ with probability at least $1 - \delta/2$, where $\tilde{\varepsilon}(\delta)$ is given by Eq. (4.5). The claim of the theorem follows from the union bound. \qed
A Proofs

A.1 Proofs for Section 2

Proof of Proposition 2.1. For $p = 1$, the result follows immediately from the Kantorovich dual representation of $W_1(\cdot, \cdot)$ [8]:

$$W_1(Q, Q') = \sup \left\{ |E_Q F - E_{Q'} F| : \sup_{z, z' \in Z} \frac{|F(z) - F(z')|}{d_z(z, z')} \leq 1 \right\}$$

and from the fact that, for $Q, Q' \in B^W_{\varepsilon, 1}(P)$, $W_1(Q, Q') \leq 2\varepsilon$ by the triangle inequality. For $p > 1$, the result follows from the fact that $W_1(Q, Q') \leq W_p(Q, Q')$ for all $Q, Q' \in \mathcal{P}_p(Z)$. 

Proof of Proposition 2.2. Fix some $Q, Q' \in B^W_{\varepsilon, p}(P)$ and let $M \in \mathcal{P}(Z \times Z)$ achieve the infimum in (1.4) for $W_p(Q, Q')$. Then for $(Z, Z') \sim M$ we have

$$f(Z') - f(Z) \leq \int_0^1 G_f(\gamma(t))dt \cdot d_Z(Z, Z')$$

where the first inequality is from (2.1) and the second one is by the assumed geodesic convexity of $G_f$. Taking expectations of both sides with respect to $M$ and using Hölder’s inequality, we obtain

$$R(Q', f) - R(Q, f) \leq \frac{1}{2} \left( E_M |G_f(Z) + G_f(Z')|^q \right)^{1/q} (E_M d^p_Z(Z, Z'))^{1/p}$$

$$= \frac{1}{2} \left\| G_f(Z) + G_f(Z') \right\|_{L^q(M)} W_p(Q, Q'),$$

where we have used the $p$-Wasserstein optimality of $M$ for $Q$ and $Q'$. By the triangle inequality, and since $Z \sim Q$ and $Z' \sim Q$,

$$\left\| G_f(Z) + G_f(Z') \right\|_{L^q(M)} \leq \left\| G_f(Z) \right\|_{L^q(Q)} + \left\| G_f(Z) \right\|_{L^q(Q')}$$

$$\leq 2 \sup_{Q \in B^W_{\varepsilon, p}(P)} \|G_f(Z)\|_{L^q(Q)}.$$  

Interchanging the roles of $Q$ and $Q'$ and proceeding with the same argument, we obtain the estimate

$$\sup_{Q, Q' \in B^W_{\varepsilon, p}(P)} \left| R(Q, f) - R(Q', f) \right| \leq 2\varepsilon \sup_{Q \in B^W_{\varepsilon, p}(P)} \|G_f(Z)\|_{L^q(Q)},$$

from which it follows that

$$R(Q, f) \leq R_{\varepsilon, p}(P, f)$$

$$= \sup_{Q' \in B^W_{\varepsilon, p}(P)} \left[ R(Q', f) - R(Q, f) + R(Q, f) \right]$$

$$\leq R(Q, f) + 2\varepsilon \sup_{Q \in B^W_{\varepsilon, p}(P)} \|G_f(Z)\|_{L^q(Q)}.$$  

\[\square\]
Proof of Proposition 2.3. As a subset of \( \mathbb{R}^{d+1} \), \( Z \) is a geodesic space: for any pair \( z, z' \in Z \) there is a unique constant-speed geodesic \( \gamma(t) = (1 - t)z + tz' \). We claim that \( G_f(z) = G_f(x, y) = 2(B + M)(1 + L\|\nabla f_0(x)\|_2) \) is a geodesically convex upper gradient for \( f(z) = f(x, y) = (y - f_0(x))^2 \). In this flat Euclidean setting, geodesic convexity coincides with the usual definition of convexity, and the map \( z \mapsto G_f(z) \) is evidently convex:

\[
G_f((1 - t)z + tz') \leq (1 - t)G_f(z) + tG_f(z').
\]

Next, by the mean-value theorem,

\[
f(z') - f(z) = \int_0^1 (z' - z, \nabla f((1 - t)z + t'z))dt \\
\leq \int_0^1 \|\nabla f((1 - t)z + t'z')\|_2 dt \cdot \|z - z'\|_2 \\
= \int_0^1 \|\nabla f((1 - t)z + t'z')\|_2 dt \cdot d_Z(z, z'),
\]

and a simple calculation shows that

\[
\|\nabla f(z)\|_2^2 = \|\nabla f(x, y)\|_2^2 \\
= 4f(z) \left(1 + \|\nabla f_0(z)\|_2^2\right) \\
\leq 4(B + M)^2(1 + L^2\|x\|_2^2).
\]

Therefore, \( \|\nabla f(z)\|_2 \leq G_f(z) \) for \( z = (x, y) \), as claimed. Thus, by Proposition (2.2),

\[
R(Q, f) \leq R^*_{\varepsilon, 2}(P, f) \\
\leq R(Q, f) + 2 \sup_{Q \in B^W_{\varepsilon, 2}(P)} \|G_f(Z)\|_{L^2(Q)} \varepsilon \\
= R(Q, f) + 4(B + M) \left(1 + L \sup_{Q \in B^W_{\varepsilon, 2}(P)} E_Q\|X\|_2\right) \varepsilon \\
= R(Q, f) + 4(B + M) \left(1 + L \sup_{Q \in B^W_{\varepsilon, 2}(P)} \sigma_{Q, X}\right) \varepsilon.
\]

\[\square\]

A.2 Proofs for Section 3

Proof of Corollary 3.1. We first verify the regularity assumptions. Assumption 3.1 is evidently satisfied since \( \text{diam}(Z)^2 = \text{diam}(\mathcal{X})^2 + \text{diam}(\mathcal{Y})^2 \leq 4r_0^2 + 4B^2 \). Each \( f \in \mathcal{F} \) is continuous, and Assumption 3.2 holds with \( M = 2(||s||^2_{\infty} + B^2) \). To verify Assumption 3.3, take \( f_0 = 0 \) and pick an arbitrary \( x_0 \in \mathcal{X} \). Then

\[
f(x, y) = (y - s(0))^2 = y^2 \leq \|x - x_0\|^2 + y^2 = d_Z^2(z, z_0)
\]

with \( z_0 = (x_0, 0) \). Thus, Assumption 3.3 holds with \( C_0 = 1 \) and \( z_0 = 0 \).
To evaluate the Dudley entropy integral in (3.1), we need to estimate the covering numbers $\mathcal{N}(\mathcal{F}, \| \cdot \|_{\infty}, \cdot)$. First observe that, for any two $f, g \in \mathcal{F}$ corresponding to $f_0, g_0 \in \mathbb{R}^d$, we have

$$\sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} |f(x, y) - g(x, y)| = \sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \left| y - s(f_0^T x) - (y - s(g_0^T x)) \right|^2 \leq 2B \sup_{x \in \mathcal{X}} |s(f_0^T x) - s(g_0^T x)| + \sup_{x \in \mathcal{X}} |s^2(f_0^T x) - s^2(g_0^T x)| \leq 2c(B + \|s\|_{\infty}) \|s^2\|_{\infty} \|f_0 - g_0\|_2.$$ 

Since $f_0, g_0$ belong to the unit ball in $\mathbb{R}^d$,

$$\mathcal{N}(\mathcal{F}, \| \cdot \|_{\infty}, u/2) \leq \left( \frac{6D}{u} \right)^d$$

for $0 < u < 2D$, and $\mathcal{N}(\mathcal{F}, \| \cdot \|_{\infty}, u/2) = 1$ for $u \geq 2D$, which gives

$$\int_0^{\infty} \sqrt{\log \mathcal{N}(\mathcal{F}, \| \cdot \|_{\infty}, u/2)} \, du \leq \int_0^{2D} \sqrt{d \log \left( \frac{6D}{u} \right)} \, du = 6D \sqrt{d} \int_0^{1/3} \sqrt{\log \left( \frac{1}{u} \right)} \, du \leq 6cD \sqrt{d},$$

where $c = \frac{1}{6} \left( 2\sqrt{\log 3} + 3\sqrt{\pi} \text{erfc}(\sqrt{\log 3}) \right) < 1$. Substituting this into the bound (3.1), we get the desired estimate. \[\square\]

**Proof of Corollary 3.2.** We start by presenting the following technical lemma.

**Lemma A.1.** For any $f, g \in \mathcal{F}$ we have:

$$\|f\|_{\infty} \leq 2(r^2 + B^2)$$

$$\|f - g\|_{\infty} \leq 2(r + B)\|f_0 - g_0\|_{\mathcal{X}},$$

where $f_0, g_0 \in I_K(\mathcal{B}_r)$ are the predictors satisfying $f(x, y) = (y - f_0(x))^2$ and $g(x, y) = (y - g_0(x))^2$.

**Proof.** First note that for Gaussian kernels, $\sup_{x \in \mathcal{X}} \sqrt{K(x, x)} \leq 1$ by definition. Then, we have the first claim by

$$|f(z)| = (f_0(x) - y)^2 \leq 2f_0^2(x) + 2y^2 \leq 2r^2 + 2B^2,$$

where the first inequality is due to convexity of square function and the second inequality is due to the reproducing kernel property of $K$ and the Cauchy-Schwarz inequality in $\mathcal{H}_K$. Second claim is established similarly:

$$|f(z) - g(z)| = |(f_0(x) - y)^2 - (g_0(x) - y)^2|$$

$$= |f_0(x) + g_0(x) - 2y|f_0(x) - g_0(x)|$$

$$\leq (2 \sup_{\|h\|_{\mathcal{K}} \leq r} |h(x)| + 2|y|)\|f_0 - g_0\|_{\mathcal{X}},$$

where the last inequality is due to Cauchy-Schwarz inequality again. \[\square\]
We now return to the proof of Corollary 3.2. Assumption 3.1 holds since \( \diam(Z)^2 = \diam(X)^2 + \diam(Y)^2 \leq 4r_0^2 + 4B^2 \). The functions in \( \mathcal{F} \) are continuous, and Assumption 3.2 holds with \( M = 2(r^2 + B^2) \) by virtue of the first estimate of Lemma A.1. To verify Assumption 3.3, take \( f_0 = 0 \) and pick an arbitrary \( x_0 \in X \). Then

\[
 f(x, y) = y^2 \leq \|x - x_0\|^2 + y^2 = d^2_Z(z, z_0)
\]

with \( z_0 = (x_0, 0) \). Thus, Assumption 3.3 holds with \( C_0 = 1 \) and \( z_0 = 0 \). Now we proceed to upper-bound the Dudley entropy integral for \( \mathcal{F} \):

\[
 \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \| \cdot \|\infty, \frac{u}{2})} \, du \\
\leq \int_0^\infty \sqrt{\log \mathcal{N}(I_K(B_r), \| \cdot \|\infty, \frac{u}{4(r + B)})} \, du \\
= \int_0^{4(r + B)r} \sqrt{\log \mathcal{N}(I_K(B_r), \| \cdot \|\infty, \frac{u}{4(r + B)})} \, du \\
\leq \int_0^{2(r^2 + Br)} \sqrt{\log \mathcal{N}(I_K(B_r), \| \cdot \|\infty, \frac{u}{4(r + B)})} \, du \\
+ \int_{2(r^2 + Br)}^{4(r^2 + Br)} \sqrt{\log \mathcal{N}(I_K(B_r), \| \cdot \|\infty, \frac{r}{2})} \, du \\
:= T_1 + T_2
\]

where we used the second claim of Lemma A.1 for the first inequality and the monotonicity of covering numbers for the second inequality. Plugging in the estimate from Proposition 3.2,

\[
 T_1 \leq \sqrt{d} \left( 32 + \frac{2560dr_0^2}{\sigma^2} \right)^{\frac{d+1}{2}} \int_0^{2(r^2 + Br)} \left( \log \frac{4(r^2 + Br)}{u} \right)^{\frac{d+1}{2}} \, du \\
= 4\sqrt{d} \left( 32 + \frac{2560dr_0^2}{\sigma^2} \right)^{\frac{d+1}{2}} (r^2 + Br) \int_{\log 2}^{\infty} e^{-u} u^{\frac{d+1}{2}} \, du \\
= 4\sqrt{d} \left( 32 + \frac{2560dr_0^2}{\sigma^2} \right)^{\frac{d+1}{2}} (r^2 + Br) \Gamma \left( \frac{d + 3}{2}, \log 2 \right),
\]

\[
 T_2 \leq 2(r^2 + Br) \cdot \sqrt{d} \left( 32 + \frac{2560dr_0^2}{\sigma^2} \right)^{\frac{d+1}{2}} (\log 2)^{\frac{d+1}{2}},
\]

and hence \( T_1 + T_2 \leq \frac{C_1}{48} (r^2 + Br) \).

\( \square \)

**Proof of Lemma 3.1.** Since \( \varphi_{\lambda,f} \geq 0 \) for all \( \lambda, f \), we arrive at

\[
 \tilde{\lambda} \leq \frac{R_{\varepsilon,p}(Q, \mathcal{F})}{\varepsilon^p}. \quad (A.1)
\]
We proceed to upper-bound the local minimax risk $R_{\varepsilon,p}(Q, \mathcal{F})$:

$$R_{\varepsilon,p}(Q, \mathcal{F}) = \inf_{f \in \mathcal{F}} \min_{\lambda \geq 0} \left\{ \lambda \varepsilon^p + \int_{\mathcal{Z}} \sup_{z' \in \mathcal{Z}} \left[ f(z') - \lambda d^p_{\mathcal{Z}}(z, z') \right] Q(dz') \right\}$$

$$\leq \min_{\lambda \geq 0} \left\{ \lambda \varepsilon^p + \int_{\mathcal{Z}} \sup_{z' \in \mathcal{Z}} \left[ f_0(z') - \lambda d^p_{\mathcal{Z}}(z, z') \right] Q(dz') \right\}$$

$$\leq \min_{\lambda \geq 0} \left\{ \lambda \varepsilon^p + \int_{\mathcal{Z}} \sup_{z' \in \mathcal{Z}} \left[ C_0 d^p_{\mathcal{Z}}(z, z_0) - \lambda d^p_{\mathcal{Z}}(z, z') \right] Q(dz') \right\}.$$

For $\lambda \geq C_0 2^{p-1}$, the integrand can be upper-bounded as follows:

$$\sup_{z' \in \mathcal{Z}} \left[ C_0 d^p_{\mathcal{Z}}(z', z_0) - \lambda d^p_{\mathcal{Z}}(z, z') \right] \leq \sup_{z' \in \mathcal{Z}} \left[ C_0 2^{p-1} d^p_{\mathcal{Z}}(z, z_0) + (C_0 2^{p-1} - \lambda) d^p_{\mathcal{Z}}(z, z') \right]$$

$$\leq C_0 2^{p-1} d^p_{\mathcal{Z}}(z, z_0).$$

Therefore,

$$R_{\varepsilon,p}(Q, \mathcal{F}) \leq \min_{\lambda \geq C_0 2^{p-1}} \left\{ \lambda \varepsilon^p + C_0 2^{p-1} \int_{\mathcal{Z}} d^p_{\mathcal{Z}}(z, z_0) Q(dz) \right\}$$

$$\leq C_0 2^{p-1} \left( \varepsilon^p + (\text{diam} \left( \mathcal{Z} \right))^p \right).$$

Substituting this estimate into (A.1), we obtain (3.6) \hfill \Box

**Proof of Lemma 3.2.** Define the $\Phi$-indexed process $X = (X_\varphi)_{\varphi \in \Phi}$ via

$$X_\varphi := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i \varphi(Z_i),$$

which is clearly zero-mean: $\mathbb{E}[X_\varphi] = 0$ for all $\varphi \in \Phi$. To upper-bound the Rademacher average $\mathbb{R}_n(\Phi)$, we first show that $X$ is a subgaussian process with respect to a suitable pseudometric. For $\varphi = \varphi_{\lambda,f}$ and $\varphi' = \varphi_{\lambda',f'}$, define

$$d_\Phi(\varphi, \varphi') := \|f - f'\|_\infty + (\text{diam} \left( \mathcal{Z} \right))^p |\lambda - \lambda'|,$$

and it is not hard to show that $\|\varphi - \varphi'\|_\infty \leq d_\Phi(\varphi, \varphi')$. Then, for any $t \in \mathbb{R}$, using Hoeffding’s lemma and the fact that $(\sigma_i, Z_i)$ are i.i.d., we arrive at

$$\mathbb{E} \left[ \exp \left( t(X_\varphi - X_{\varphi'}) \right) \right] = \mathbb{E} \left[ \exp \left( \frac{t}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i (\varphi(Z_i) - \varphi'(Z_i)) \right) \right]$$

$$= \left( \mathbb{E} \left[ \exp \left( \frac{t}{\sqrt{n}} \sigma_1 (\varphi(Z_1) - \varphi'(Z_1)) \right) \right] \right)^n$$

$$\leq \exp \left( \frac{t^2 d^2_\Phi(\varphi, \varphi')}{2} \right).$$
Hence, $X$ is subgaussian with respect to $d_\Phi$, and therefore the Rademacher average $\mathfrak{R}_n(\Phi)$ can be upper-bounded by the Dudley entropy integral [12]:

$$\mathfrak{R}_n(\Phi) \leq \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(\Phi, d_\Phi, u)} du,$$

where $\mathcal{N}(\Phi, d_\Phi, \cdot)$ are the covering numbers of $(\Phi, d_\Phi)$. From the definition of $d_\Phi$, it follows that

$$\mathcal{N}(\Phi, d_\Phi, u) \leq \mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, u/2) \cdot \mathcal{N}(\Lambda, | \cdot |, u/2(\text{diam}(Z))^p),$$

and therefore

$$\mathfrak{R}_n(\Phi) \leq \frac{12}{\sqrt{n}} \left( \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, u/2)} du + \int_0^\infty \sqrt{\log \mathcal{N}(\Lambda, | \cdot |, u/2(\text{diam}(Z))^p)} du \right).$$

Since $\Lambda$ is a compact interval, it is straightforward to upper-bound the second integral:

$$\int_0^\infty \sqrt{\log \mathcal{N}(\Lambda, | \cdot |, u/2(\text{diam}(Z))^p)} du \leq 2|\Lambda|(\text{diam}(Z))^p \int_0^{1/2} \sqrt{\log(1/u)} du = 2c|\Lambda|(\text{diam}(Z))^p,$$

where $|\Lambda| = C_0 2^{p-1} (1 + (\text{diam}(Z)/\varepsilon)^p)$ is the length of $\Lambda$ and $c = \frac{1}{2} \left( \sqrt{\log 2} + \sqrt{\text{erfc}(\sqrt{\log 2})} \right) < 1$. Consequently,

$$\mathfrak{R}_n(\Phi) \leq \frac{12}{\sqrt{n}} \left( \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, u/2)} du + 2|\Lambda|(\text{diam}(Z))^p \right) \leq \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \| \cdot \|_\infty, u/2)} du + \frac{12C_0(2\text{diam}(Z))^p}{\sqrt{n}} \left( 1 + \left( \frac{\text{diam}(Z)}{\varepsilon} \right)^p \right).$$

$\square$

### A.3 Proofs for Section 4

**Proof of Lemma 4.1.** First we prove that $W_p(P, Q) \leq W_p(\mu, \nu)$. Define the mapping $\bar{T} : Z \to Z$ by $\bar{T} := T \otimes \text{id}_Z$, i.e., $\bar{T}(z) = \bar{T}(x, y) = (T(x), y)$, and let $\bar{Q} = \bar{T}_# P$, the pushforward of $P$ by $\bar{T}$. We claim that $\bar{Q} \equiv Q$. Indeed, for any measurable sets $A \subseteq X$ and $B \subseteq Y$,

$$\bar{Q}(A \times B) = \bar{T}_# P(A \times B) = P(T^{-1}(A) \times B) = \int_{T^{-1}(A)} \mu(dx)P_{Y|X}(B|x) = \int_A \mu(dx)P_{Y|X}(B|T(x)) = \int_A \nu(dx)Q_{Y|X}(B|x),$$

where we have used the relation (4.1) and the invertibility of $T$. Thus,

$$W_p(P, Q) \leq E_P[d_\mathcal{Z}(Z, \bar{T}(Z))] = E_P[d_X(X, T(X))] = W_p(\mu, \nu).$$
For the reverse inequality, let $M \in \mathcal{P}(\mathbb{Z} \times \mathbb{Z})$ be the optimal coupling of $P$ and $Q$. Then, for $Z = (X, Y)$ and $Z' = (X', Y')$ with $(Z, Z') \sim M$, the marginal $M_{XX'}$ is evidently a coupling of the marginals $\mu$ and $\nu$, and therefore

$$W_p^p(P, Q) = E_M[d_p^p(Z, Z')]$$
$$= E_M[d_X^p(X, X')] + E_M[d_Y^p(Y, Y')]$$
$$\geq E_M[d_X^p(X, X')]$$
$$\geq W_p^p(\mu, \nu).$$
References

[1] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[2] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, 2011.

[3] F. Farnia and D. Tse. A minimax approach to supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4240–4248. 2016.

[4] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: a generalized empirical likelihood approach. arXiv preprint 1610.03425, 2016.

[5] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. arXiv preprint 1505.05116, 2015.

[6] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.

[7] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. arXiv preprint 1604.02199, 2016.

[8] C. Villani. *Topics in optimal transportation*. American Mathematics Society, Providence, RI, 2003.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[10] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. To appear.

[11] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2008.

[12] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer, 2014.

[13] F. Cucker and D. Zhou. *Learning theory: an approximation theory viewpoint*. Cambridge University Press, Cambridge, MA, 2007.

[14] T. Kühn. Covering numbers of gaussian reproducing kernel hilbert spaces. *Journal of Complexity*, pages 489–499, 2011.

[15] N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, pages 1–32, 2014.