Adversarial Risk via Optimal Transport and Optimal Couplings

Muni Sreenivas Pydi and Varun Jog

***Abstract*— Modern machine learning algorithms perform poorly on adversarially manipulated data. Adversarial risk quantifies the error of classifiers in adversarial settings; adversarial classifiers minimize adversarial risk. In this paper, we analyze adversarial risk and adversarial classifiers from an optimal transport perspective. We show that the optimal adversarial risk for binary classification with 0-1 loss is determined by an optimal transport cost between the probability distributions of the two classes. We develop optimal transport plans (probabilistic couplings) for univariate distributions such as the normal, the uniform, and the triangular distribution. We also derive optimal adversarial classifiers in these settings. Our analysis leads to algorithm-independent fundamental limits on adversarial risk, which we calculate for several real-world datasets. We extend our results to general loss functions under convexity and smoothness assumptions.**

***Index Terms*— Machine learning, statistical learning, robustness, couplings, information theory.**

I. Introduction

DEEP learning has had tremendous success in recent times, producing state-of-the-art results in image classification [1], [2], game playing [3]–[5], speech [6], [7] and natural language processing [8], [9]. However, Szegedy *et al.* [10] discovered that these algorithms are surprisingly vulnerable to minute adversarial perturbations. Many *adversarial attacks* [11]–[13] and defenses [14]–[16] have been proposed since. Often, the defenses are subsequently broken or are computationally intractable in practice.

The reason for existence of adversarial examples in deep learning is unknown, but many explanations have been suggested. One line of work hypothesizes that adversarial examples are inevitable in certain high-dimensional settings [17], [18]. Goodfellow *et al.* [13] propose that the reason for adversarial examples may be the linear nature of deep neural networks. Ilyas *et al.* [19] propose that adversarial examples correspond to non-robust features in the data that are highly predictive, but brittle. Moreover, it was recently proposed that adversarial risk may be fundamentally at odds with standard risk—a claim that finds support both in theory [20] and in practice [21].

In this paper, we deviate from algorithm-dependent investigations of adversarial examples and ask two fundamental questions:

Question 1: How much can the optimal adversarial risk differ from optimal standard risk? In the binary classification setting, we may equivalently ask: How much will the classification error increase due to an adversary—an increase that cannot be mitigated by *any* algorithm?

Question 2: How does the optimal adversarial classifier differ from the standard optimal classifier?

Recent works have addressed Question 1 by deriving upper and lower bounds on the optimal adversarial risk with respect to a fixed set of classifiers, by extending the PAC learning theory to encompass adversaries [22], [23]. A related question asks how much adversarial perturbation is sufficient to make the optimal adversarial risk significantly greater than the optimal standard risk. Relevant works in this direction develop robustness metrics that depend on the classifier [24]–[26]. For Question 2, recall that the optimal classifier without an adversary is simply the Bayes optimal classifier. In adversarial settings, the optimal classifier may differ considerably from the Bayes optimal classifier. A recent line of work shows that optimal adversarial classifier can be calculated using non-parametric methods if data are well-separated [27], [28]. The work of Moosavi-Dezfooli *et al.* [29], Cohen *et al.* [30] and Yang *et al.* [31] suggests that the optimal adversarial classifier has smoother boundaries than the optimal standard classifier. Even so, the question of how the optimal adversarial classifier differs from the standard one remains open.

The closest work to ours is Bhagoji *et al.* [32], which develops algorithm-independent lower bounds for learning in the presence of an adversary. Specifically, [32] contains a similar result to our Theorem 2 which gives the optimal adversarial risk for binary classification with 0-1 loss in terms of an optimal transport cost between the probability distributions of the two classes. We provide a new, simpler proof of this characterization by applying the Kantorovich duality of optimal transport for 0-1 cost functions. We shall discuss the results from Bhagoji *et al.* [32] and compare these with our results at appropriate points in the paper.

Manuscript received December 20, 2019; revised December 22, 2020; accepted July 5, 2021. Date of publication July 26, 2021; date of current version August 25, 2021. The work of Varun Jog was supported in part by NSF under Grant CCF-1841190, Grant CCF-1907786, and Grant CCF-1942134; and in part by the Nvidia GPU Grant Program. This article was presented at the 2020 International Conference on Machine Learning. *(Corresponding author: Muni Sreenivas Pydi.)*

Muni Sreenivas Pydi is with the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: pydi@wisc.edu).

Varun Jog is with the Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge CB3 0WB, U.K. (e-mail: vj270@cam.ac.uk).

Communicated by S. Boucheron, Associate Editor for Statistical Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3100107.

Digital Object Identifier 10.1109/TIT.2021.3100107

Our Contributions:

In this paper, we consider *data perturbing adversaries* that manipulate sampled data points, and *distribution perturbing adversaries* that manipulate the data generating distribution itself. We show that these two adversaries are closely related and establish the precise relationship between them. We primarily focus on the binary classification setting under 0-1 loss function. We answer Question 1 by providing algorithm-independent bounds for adversarial risk that are agnostic to the classifier. We answer Question 2 by deriving the optimal adversarial classifier in some special settings, and by providing bounds on the deviation of the optimal adversarial classifier from the standard optimal classifier in more general settings. Our contributions are listed below.

(1) We resolve Question 1 in the binary classification with 0-1 loss setting by deriving a formula for the optimal adversarial risk in terms of an optimal transport cost between the two data distribution. Our proof is novel and simple and connects adversarial machine learning to well-known results in optimal transport theory.
(2) We construct optimal couplings for the optimal transport cost from (1) when the two data distributions are univariate normal, uniform over intervals, and triangular. We resolve Question 2 in these cases by determining the optimal adversarial classifiers using the optimal couplings. Our results indicate that the decision boundary can be sensitive to the adversary's budget.
(3) We calculate the optimal adversarial risk for the CIFAR10, MNIST, Fashion-MNIST, and SVHN datasets. We perform a similar calculation for data-augmented versions of these datasets. The non-zero values resulting from these calculations highlight the impossibility of being completely accurate—even on the training set—in adversarial settings.
(4) We partially address Question 1 for continuous loss functions by deriving upper and lower bounds on the optimal adversarial risk which depend on convexity and smoothness assumptions of the loss with respect to data. We also partially address Question 2 by upper bounding how much the optimal hypothesis with an adversary can deviate from the optimal hypothesis without an adversary. These bounds are in terms of the curvature of the loss function with respect to the parameters of the hypotheses.

Structure: The rest of the paper is structured as follows: In Section II, we introduce the data-perturbing and distribution-perturbing models of adversaries and show that the data-perturbing adversary may be considered to be a special case of the distribution-perturbing adversary. We also discuss related work, especially, other related notions of adversaries studied in the literature. In Section III, we discuss the optimal adversarial risk for binary classification with 0-1 loss. We settle Question 1 in this setting by introducing the D_ϵ optimal transport cost that completely characterizes the optimal risk. In Section IV we present a coupling strategy that achieves the optimal transport cost in special cases of interest in the univariate case. Using this coupling, we obtain the optimal adversarial classifier, thus settling Question 2 for these special cases. In Section V-A, we discuss the optimal risk for general loss functions and present our bounds on the optimal adversarial risk. In Section V-B, we discuss optimal classifiers for general loss functions and present our deviation bounds on the optimal adversarial classifier. Finally, in Section VI, we present adversarial risk lower bounds for real world datasets and evaluate our bounds for 0-1 loss function.

Notation: The complement of a set A is denoted by A^c. The indicator function that maps all the inputs satisfying condition C to 1 and the rest to 0 is denoted by $\mathbb{1}\{C\}$. The set of probability measures over the measure space $(\mathcal{X}, \sigma(\mathcal{X}))$, where $\mathcal{X}$ is a Polish space and $\sigma(\mathcal{X})$ is the Borel sigma algebra over $\mathcal{X}$, is denoted by $\mathcal{P}(\mathcal{X})$. For any two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, the set of all joint probability measures (or couplings) over $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν is denoted by $\Pi(\mu, \nu)$. The total variation distance and p-Wasserstein distance between μ and ν is denoted by $D_{TV}(\mu, \nu)$ and $W_p(\mu, \nu)$, respectively. A norm and its dual are denoted by $\|\cdot\|$ and $\|\cdot\|_*$, respectively. The cumulative distribution function (cdf) of the standard normal distribution is denoted by Φ and its tail distribution is denoted by $Q(x) := 1 - \Phi(x)$.

II. Models in Adversarial Machine Learning

We describe models of adversaries that are commonly invoked in machine learning and highlight connections between them. We use the following convention: Let $(\mathcal{X}, d)$ denote a separable Hilbert space with metric d for the data points and $\mathcal{Y}$ denote the finite set of discrete labels assigned to the data-points. Let ρ be the data distribution which we express as $\rho_y(y)\rho_{x|y}(x)$, where $\rho_y(y)$ is the marginal probability of label $y \in \mathcal{Y}$ and $\rho_{x|y}(\cdot)$ is the conditional distribution of X given $Y = y$. Let the hypothesis class be $\mathcal{W}$. Let $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{W} \to \mathbb{R}^+$ denote a loss function such that $\ell((\cdot, \cdot), w)$ is ρ-measurable for all $w \in \mathcal{W}$. Let $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ and $z := (x, y) \in \mathcal{Z}$.

A. Types of Adversaries: Informal Description

To quantify the impact of an adversary, several notions of adversarial risk have been proposed in the literature. We highlight two most popular notions: (i) adversary perturbs data points, and (ii) adversary perturbs data distributions.

1) Data Perturbing Adversary: A data-perturbing adversary of budget ϵ can perturb $x \in \mathcal{X}$ to any $x' \in \mathcal{X}$ such that $d(x, x') \leq \epsilon$. The adversary wishes to maximize loss, and so would choose x' accordingly. A natural definition for adversarial loss (or robust loss) incurred by a hypothesis $w \in \mathcal{W}$ for an adversary with budget ϵ is [14], [33]:

$$R_\epsilon(\ell, w) = \mathbb{E}_{(x,y)\sim\rho_y\rho_{x|y}}\left[\sup_{d(x,x')\leq\epsilon} \ell((x', y), w)\right]. \quad (1)$$

2) Distribution Perturbing Adversary: The adversarial loss incurred by a hypothesis $w \in \mathcal{W}$ in the presence of a distribution perturbing adversary with a budget ϵ is defined as follows:

$$\widehat{R}_\epsilon(\ell, w) = \sup_{\rho' \in B_\epsilon(\rho)} \mathbb{E}_{z\sim\rho'}\ell(z, w), \quad (2)$$

where $B_\epsilon(\rho)$ may be thought of as a ball of radius ϵ around ρ, the true data generating distribution. The Wasserstein distance

has been one of the more popular metrics used to define $B_\epsilon(\cdot)$ in the space of distributions [34]–[39].

B. Types of Adversaries: Formal Description

1) Data Perturbing Adversary: Formulation (1) is adequate for most practical purposes, but runs into measure-theoretic difficulties due to the arbitrary choice involved in the adversary's perturbations. Even if the adversarial map $x \to x'$ is ρ-measurable, the function $\sup_{d(x,x')\leq\epsilon} \ell((x',y),w)$ may not be ρ-measurable. Moreover, the adversary may not use a deterministic mapping to perturb the data points, and rather do so randomly. In light of these considerations, we redefine a data perturbing adversary to be a collection of Markov kernels indexed by $y \in \mathcal{Y}$ denoted by $\kappa_y : \mathcal{X} \times \sigma(\mathcal{X}) \to [0,1]$. Equivalently, for each $y \in \mathcal{Y}$ the kernel κ_y satisfies:

1) For all $x \in \mathcal{X}$, the map $A \to \kappa_y(x, A)$ is a probability measure denoted by $\kappa_{y,x} \in \mathcal{P}(\mathcal{X})$,
2) For all $A \in \sigma(\mathcal{X})$, the map $x \to \kappa_{y,x}(A)$ is measurable.

Let the collection of kernels indexed by y be $\kappa := \{\kappa_y \mid y \in \mathcal{Y}\}$. It is useful to think of $\kappa_{y,x}$ as the adversary's perturbation strategy after observing the sample x and its label y – the adversary perturbs x to x' where the latter is the result of passing x through the Markov kernel $\kappa_{y,x}$. The collection of kernels κ completely describes the adversary's strategy. We use $\rho^\kappa(x,y,x')$ to denote the joint distribution of (x,y,x') induced by κ. Let the joint distribution of (x,x') conditioned on y be denoted by $\rho^\kappa_{(x,x')|y} \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$, and the conditional distribution of x' given y be denoted by $\rho^\kappa_{x'|y} \in \mathcal{P}(\mathcal{X})$. We say that the adversary κ has a budget of $\epsilon \geq 0$, denoted by $\kappa \in K_\epsilon$, if the following holds ρ_y-almost surely:

$$\operatorname*{ess\,sup}_{(x,x')\sim\rho^\kappa_{(x,x')|y}} d(x,x') \leq \epsilon. \tag{3}$$

Then, we have the following definition for the adversarial risk incurred by a hypothesis $w \in \mathcal{W}$ in the presence of a data perturbing adversary of budget ϵ:

$$R_\epsilon(\ell, w) = \sup_{\kappa\in K_\epsilon} \mathbb{E}_{(x,y,x')\sim\rho^\kappa(x,y,x')} \left[\ell((x',y),w)\right]. \tag{4}$$

We will use the definition in (4) rather than the one in (1) to denote the adversarial loss unless specified otherwise.

The optimal adversarial loss attainable over the hypotheses $w \in \mathcal{W}$ is defined as the *optimal adversarial risk* or *optimal robust risk*,

$$R^*_\epsilon = \inf_{w\in\mathcal{W}} R_\epsilon(\ell, w). \tag{5}$$

The hypothesis attaining the optimal adversarial risk (if it exists) is called the *optimal adversarial hypothesis* and is denoted by w^*_ϵ. Note that for $\epsilon = 0$, we have $x = x'$ almost surely, and so the adversarial risk equals Bayes risk, $R^*_0 = \inf_w \mathbb{E}_{z\sim\rho}\left[\ell((x,y),w)\right]$.

2) Distribution Perturbing Adversary: Since distribution perturbing adversaries considered in this paper rely on optimal transport distances, we introduce optimal transport briefly. Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and let $c : \mathcal{X}\times\mathcal{X} \to \mathbb{R}^+$ denote the cost $c(x,x')$ of transporting unit mass from $x \in \mathcal{X}$ to $x' \in \mathcal{X}$. The optimal transport cost between μ and ν is given by,

$$\mathcal{T}_c(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi} c(x,x'), \tag{6}$$

where $\Pi(\mu,\nu)$ is the set of all couplings between μ and ν. When $c(x,x') = d(x,x')$, where d is a metric over $\mathcal{X}$, the optimal transport cost is the 1-Wasserstein distance; i.e., $W_1(\mu,\nu) = \mathcal{T}_d(\mu,\nu)$. For $p \geq 1$, the p-Wasserstein distance is given by $W_1(\mu,\nu) = (\mathcal{T}_{d^p}(\mu,\nu))^p$. The ∞-Wasserstein distance is defined to be the limit of the p-Wasserstein distances as follows: $W_\infty(\mu,\nu) = \lim_{p\to\infty} W_p(\mu,\nu)$. Alternatively, the ∞-Wasserstein distance is also characterized as follows [40].

$$W_\infty(\mu,\nu) \tag{7}$$
$$= \inf\{\delta > 0 \ : \ \mu(A) \leq \nu(A^\delta) \ \forall \ A \in \sigma(\mathcal{X})\} \tag{8}$$
$$= \inf_{\pi\in\Pi(\mu,\nu)} \operatorname*{ess\,sup}_{(x,x')\sim\pi} d(x,x'). \tag{9}$$

For $1 \leq p \leq q$, we have $W_p(\mu,\nu) \leq W_q(\mu,\nu) \leq W_\infty(\mu,\nu)$. Hence, the W_∞-metric is stronger than any W_p-metric.

An adversary γ is a collection of distributions over $\mathcal{X}$ indexed by y; i.e., $\gamma = \{\rho^\gamma_{x'|y} \in \mathcal{P}(\mathcal{X}) \mid y \in \mathcal{Y}\}$. An adversary γ is said to have budget ϵ in the p-Wasserstein space if, the following holds ρ_y-almost surely:

$$W_p(\rho^\gamma_{x|y}, \rho_{x|y}) \leq \epsilon.$$

This is denoted by $\gamma \in \Gamma^p_\epsilon$. Note that $\Gamma^q_\epsilon \subseteq \Gamma^p_\epsilon$ for $1 \leq p \leq q \leq \infty$. It is useful to think of $\rho^\gamma_{x'|y}$ as the adversary's strategy after observing the sample (x,y). The adversary perturbs x to x' such that $x' \sim \rho^\gamma_{x'|y}$ and x and x' are conditionally independent given the label y. The collection of distributions $\{\rho^\gamma_{x'|y} \mid y \in \mathcal{Y}\}$ completely describes the adversary's strategy. The data distribution after the adversary's action is $(x',y) \sim \rho_y(y)\rho^\gamma_{x'|y}$. Let $\rho^\gamma(x',y) \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ denote this distribution. For a loss function ℓ and hypothesis w, the adversarial risk is defined as

$$\widehat{R}^p_\epsilon(\ell, w) = \sup_{\gamma\in\Gamma^p_\epsilon} \mathbb{E}_{(x',y)\sim\rho^\gamma(x',y)} \ell((x',y),w). \tag{10}$$

The optimal adversarial loss attainable over the hypotheses $w \in \mathcal{W}$ is defined as the *optimal adversarial risk* or *optimal robust risk*,

$$\widehat{R}^{p,*}_\epsilon = \inf_{w\in\mathcal{W}} \widehat{R}^p_\epsilon(\ell, w).$$

The hypothesis attaining the optimal adversarial risk (if it exists) is called the *optimal adversarial hypothesis* and is denoted by $\widehat{w}^*_\epsilon$. Note as before that for $\epsilon = 0$, adversarial risk equals Bayes risk.

3) Relation Between the Two Types of Adversaries: Our goal is to show that the data perturbing adversary is a special case of the distribution perturbing adversary in the following sense: The adversarial loss incurred under K_ϵ is identical to the adversarial loss incurred under a Γ^∞_ϵ. (Note that the adversaries themselves are different due to the conditional independence of x and x' given y for the distribution perturbing adversary.) This is shown in the following theorem.

Theorem 1: Let $\mathcal{X}$ be a separable Hilbert space, let $\mathcal{Y}$ be a discrete set of labels, let $\mathcal{W}$ be a hypothesis class, and let $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}_+$ be a ρ-measurable loss function. Let

$$R_\epsilon(\ell, w) = \sup_{\kappa \in K_\epsilon} \mathbb{E}_{(x,y,x') \sim \rho^\kappa(x,y,x')} \left[\ell((x', y), w)\right]$$

and

$$\widehat{R}^\infty_\epsilon(\ell, w) = \sup_{\gamma \in \Gamma^\infty_\epsilon} \mathbb{E}_{(x',y) \sim \rho^\gamma(x',y)} \left[\ell((x, y), w)\right].$$

Then $R_\epsilon(\ell, w) = \widehat{R}^\infty_\epsilon(\ell, w)$.

Proof: Let $\kappa \in K_\epsilon$. Observe that the adversarial loss depends only on the marginal distribution of (x', y) in the joint distribution of (x, y, x') induced by κ. Recall that we denote the joint distribution of (x, x') conditioned on y by $\rho^\kappa_{(x,x')|y} \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, and the conditional distribution of x' given y by $\rho^\kappa_{x'|y} \in \mathcal{P}(\mathcal{X})$.

Define a distribution perturbing adversary γ as follows: $\gamma = \{\rho^\gamma_{x'|y} := \rho^\kappa_{x'|y} \mid y \in \mathcal{Y}\}$. The key point to note here is that the marginal distribution $\rho^\gamma(x', y)$ is identical to $\rho^\kappa(x', y)$, and so the adversarial risk for both adversaries is identical. For $y \in \mathcal{Y}$,

$$\begin{aligned} W_\infty(\rho_{x|y}, \rho^\gamma_{x'|y}) &= W_\infty(\rho_{x|y}, \rho^\kappa_{x'|y}) \\ &\leq \operatorname*{ess\,sup}_{(x,x') \sim \rho^\kappa_{(x,x')|y}} d(x, x') \\ &\leq \epsilon. \end{aligned}$$

Here the first inequality follows since W_∞ is the infimum of the essential supremum over all couplings and ρ^κ is just one such coupling, and the final inequality follows since $\kappa \in K_\epsilon$. This shows that for every $\kappa \in K_\epsilon$, there is a $\gamma \in \Gamma^\infty_\epsilon$ that achieves the same adversarial risk, and so we conclude $R_\epsilon(\ell, w) \leq \widehat{R}^\infty_\epsilon(\ell, w)$.

We will now prove the above inequality in the reverse. Let $\gamma = \{\rho^\gamma_{x'|y} \mid y \in \mathcal{Y}\} \in \Gamma^\infty_\epsilon$. Then, $W_\infty(\rho_{x|y}, \rho^\gamma_{x'|y}) \leq \epsilon$ for all $y \in \mathcal{Y}$. Fix a $y \in \mathcal{Y}$. By the definition of W_∞, we have that for any positive integer n, there exists a joint probability measure $\pi_n \in \Pi(\rho_{x|y}, \rho^\gamma_{x'|y})$ such that $\operatorname{ess\,sup}_{(x,x') \sim \pi_n} d(x, x') < \epsilon + 1/n$. We now show that the sequence of measures π_n is tight. Given a $\delta > 0$, let $E \subseteq \mathcal{X}$ be a compact set such that $\min\{\rho_{x|y}(E), \rho^\gamma_{x'|y}(E)\} > 1 - \delta/2$. Then,

$$\pi_n((E \times E)^c) \leq \rho_{x|y}(E^c) + \rho^\gamma_{x'|y}(E^c) < \delta.$$

Hence, by Prokhorov's theorem (for reference, see Theorem 5.1 in [41]), there is a subsequence of (π_n) that converges weakly to a probability measure $\pi^* \in \Pi(\rho_{x|y}, \rho^\gamma_{x'|y})$ that satisfies,

$$\operatorname*{ess\,sup}_{(x,x') \sim \pi^*} d(x, x') \leq \epsilon.$$

Since $\mathcal{X}$ is a complete and separable space, there exists a Markov kernel $\lambda_y : \mathcal{X} \times \sigma(\mathcal{X}) \to [0, 1]$ such that for any $A, B \in \sigma(\mathcal{X})$, we have

$$\pi^*(A \times B) = \int_{x \in A} \lambda_{y,x}(B) d\rho_{x|y}.$$

The existence of such a probability kernel is guaranteed by the product-restricted conditional probability property of Radon spaces (see Theorem 3.1 in [42]). Repeating the above argument, we construct a kernel λ_y for each $y \in \mathcal{Y}$ satisfying $\operatorname{ess\,sup}_{(x,x') \sim \rho_{x|y}\lambda_{y,x}} d(x, x') \leq \epsilon$. Then, $\lambda := \{\lambda_y \mid y \in \mathcal{Y}\} \in K_\epsilon$. Moreover, the joint distribution of (x', y) under $\rho^\gamma_{(x',y)}$ is identical to $\rho^\lambda_{(x',y)}$, and so the corresponding adversarial risks are identical. Hence, for every $\gamma \in \Gamma^\infty_\epsilon$ there exists a $\lambda \in K_\epsilon$ that achieves an identical adversarial risk. This leads to the conclusion $\widehat{R}^\infty_\epsilon(\ell, w) \leq R_\epsilon(\ell, w)$, which completes the proof. □

For any $p \geq 1$, a distribution perturbing adversary of budget ϵ in p-Wasserstein space is more powerful than a data perturbing adversary of the same budget, as shown by the following corollary:

Corollary 1: For $p \geq 1$, consider a W_p-data-perturbing adversary of budget ϵ. The following inequality holds:

$$R_\epsilon(\ell, w) \leq \widehat{R}^p_\epsilon(\ell, w), \tag{11}$$

for all $w \in \mathcal{W}$. Moreover, $R^*_\epsilon \leq \widehat{R}^{p,*}_\epsilon$.

Proof: We have the following sequence of inequalities:

$$\begin{aligned} R_\epsilon(\ell, w) &= \sup_{\gamma \in \Gamma^\infty_\epsilon} \mathbb{E}_{(x',y) \sim \rho^\gamma(x',y)} \left[\ell((x', y), w)\right] && (12) \\ &\leq \sup_{\gamma \in \Gamma^p_\epsilon} \mathbb{E}_{(x',y) \sim \rho^\gamma(x',y)} \left[\ell((x', y), w)\right] && (13) \\ &= \widehat{R}^p_\epsilon(\ell, w), \end{aligned}$$

where the equality in (12) follows from Theorem 1 and the inequality in (13) follows from the fact that $\Gamma^\infty_\epsilon \subseteq \Gamma^p_\epsilon$. Taking infimum over $w \in \mathcal{W}$ on both sides of the inequality, we get $R^*_\epsilon \leq \widehat{R}^{p,*}_\epsilon$. □

To see that the inequality in Corollary 1 can be strict, consider the following example of binary classification with 0-1 loss. Let $P(X|Y = 0)$ (denoted by p_0) be a uniform distribution over $[0, 1]$ and $P(X|Y = 1)$ (denoted by p_1) be a constant distribution at $X = 0$. Let both the classes be equally likely. In this case, $W_p(p_0, p_1) = \left(\int_0^1 x^p dx\right)^{\frac{1}{p}} = 1/(p+1)^{\frac{1}{p}}$. Hence, $W_1(p_0, p_1) = 1/2$, while $W_\infty(p_0, p_1) = 1$. Taking the adversarial budget to be $\epsilon = 3/4$, it is easy to see that $\widehat{R}^{1,*}_\epsilon = 1/2$ because both the distributions are within the perturbation budget of the W_1-distribution perturbing adversary. But $R^*_\epsilon = \epsilon/2 = 3/8 < \widehat{R}^{1,*}_\epsilon$, which is achieved by the optimal classifier which declares label 0 on the set $[\epsilon, 1]$ and 0 otherwise.

4) A Remark on the Risk Bounds for Adversaries: All risk bounds proved in this paper are valid for both adversaries. Since the distribution perturbing adversary is stronger than the data perturbing adversary, any lower bound that holds for the latter holds for the former. Analogously, any upper bound for the distribution perturbing adversary holds for the data perturbing adversary.

C. Related Work

1) Related Notions of Adversarial Risk: In this paper, we assume that the true data distribution $\rho(x, y)$ is expressed as $\rho_y(y)\rho_{x|y}(x)$. This model allows for randomness in the label y for a fixed x. A special case of this model is when the existence of a true labelling function is assumed; i.e., there exists a function $c : \mathcal{X} \to \mathcal{Y}$ such that $c(x)$ is the true label of x for any $x \in \mathcal{X}$. That is, $\rho(x, y) = \mathbb{1}\{y = c(x)\}\rho_x(x)$. Under

this model, Gourdeau *et al.* [43] define *constant-in-the-ball risk* as

$$R(h) = \mathbb{P}_{x\sim\rho_x}[\exists x' : d(x,x') \leq \epsilon,\ h(x') \neq c(x)]. \tag{14}$$

Rewriting in terms of expectation, we get the following.

$$\begin{aligned} R(h) &= \mathbb{E}_{x\sim\rho_x} \mathbb{1}\{\exists x' : d(x,x') \leq \epsilon,\ h(x') \neq c(x)\} \\ &= \mathbb{E}_{x\sim\rho_x}\left[\sup_{d(x,x')\leq\epsilon} \mathbb{1}\{h(x') \neq c(x)\}\right]. \end{aligned}$$

Hence, the *constant-in-the-ball risk* defined above is identical to adversarial risk defined in (1) for 0-1 loss function under hypothesis h. The same notion of risk is also called the *corrupted instance risk* in Diochnos *et al.* [44].

A related notion of adversarial risk is the following:

$$R'(h) = \mathbb{E}_{x\sim\rho_x}\left[\sup_{d(x,x')\leq\epsilon} \mathbb{1}\{h(x') \neq c(x')\}\right]. \tag{15}$$

Here, the loss on the perturbed data point is evaluated with respect to the true label at the perturbed data point; i.e., $c(x')$ rather than the true label of the original data point $c(x)$. This notion of adversarial risk is termed *exact-in-the-ball risk* in Gourdeau *et al.* [43] and *error-region risk* in Diochnos *et al.* [44]. A key difference between $R(h)$ in (14) and $R'(h)$ in (15) is that $R'(h)$ is exactly equal to 0 for $h = c$ for any $\epsilon \geq 0$ whereas $R(h)$ may be strictly positive even for $h = c$. Thus, the definition of $R'(h)$ allows for the existence of an optimal classifier whose adversarial risk is 0, while the optimal classifier that minimizes $R(h)$ may still have a non-zero adversarial risk. As noted in Gourdeau *et al.* [43], $R(h)$ measures the sensitivity of the output label to corruptions in the input, while $R'(h)$ measures how well a hypothesis fits the ground-truth even with corrupted inputs.

Tu *et al.* [45] and Staib and Jegelka [46] use the adversarial risk of (1) to establish a version of Theorem 1. However, they implicitly assume that there exists a measurable function that maps $x \to x'$ such that the supremum in (1) is attained for all $x \in \mathcal{X}$. As explained previously, this is not true in general. Pinot *et al.* [47] define a notion of adversarial risk using a class of adversaries that use measurable maps to perturb data points as

$$R'_\epsilon(\ell, w) = \sup_{f\in\mathcal{F}_\epsilon} \mathbb{E}_{(x,y)\sim\rho}[\ell((f(x),y),w)],$$

where $\mathcal{F}_\epsilon$ is the set of all ρ-measurable functions satisfying the budget constraint $\operatorname{ess\,sup}_{x\sim\chi} d(x, f(x)) \leq \epsilon$. This definition is a special case of the definition in (4), where K_ϵ is restricted to the set of probability kernels with $k_x = \delta_{f(x)}$ (i.e. a probability measure with a single atom at $f(x)$) for $f \in \mathcal{F}_\epsilon$.

2) Surrogates for the Adversarial Risk: Before adversarial deep learning, minimax risk was studied in the context of robust classification with linear classifiers and SVMs [48]–[52]. Here, one proposes surrogate robust loss functions that can be tractably minimized. A similar strategy for minimizing adversarial loss may be found in [22]. For a discussion of surrogate losses, we refer the reader to Bao *et al.* [53].

In practice, the inner maximization term in the adversarial risk is approximated using gradient methods like the Fast Gradient Sign Method (FGSM) [12], [13]. This gives rise to several related notions of risk that can be interpreted as a Taylor approximations for adversarial risk in definition (1). [14], [33].

Surrogate loss functions for ensuring Wasserstein distributional robustness have been proposed in [36], [38], and robustness with respect to other optimal transport-based perturbations is studied in [54]. A key idea in these works is the dual formulation of optimal transport distances. As shown in Theorem 1, adversarial robustness is equivalent to W_∞-distributional robustness. However, the recent work on optimal transport-based robustness cannot be readily extended to the W_∞-case because the W_∞-metric does not admit a transport-cost minimizing formulation (for instance, compare (6) with (9)) and so the classic Kantorovich-Rubinstein duality cannot be applied.

3) Related Notions of Distributionally Robust Risk: The adversarial risk formulation under a distribution perturbing adversary has been widely studied in the distributionally robust optimization (DRO) literature [55], [56], with special focus on Wasserstein DRO [36], [38], [54]. The advantage of using Wasserstein metrics is the ability to measure distances between probability distributions with non-overlapping supports, which is not possible for divergence-based measures.

The distributional uncertainty set is typically centered at the empirical distribution of the data points, unlike definition (2) where it is centered around the true data generating distribution. Bertsimas *et al.* [57] note that when the support of the true distribution is unbounded, the W_∞-uncertainty set around the empirical distribution does not contain the true distribution for any ϵ. Hence, W_∞-distributional robustness is not considered in the distributional robustness setting, except for the works of Tu *et al.* [45] and Staib and Jegelka [46] that make a similar observation as our Theorem 1. Distributionally robust risk has also been studied in a minimax statistical learning framework in [58], [59] for deriving generalization error bounds.

4) Connection to Robust Statistics: Finding optimal classifiers under 0-1 loss is equivalent to hypothesis testing, and there are natural connections of adversarial machine learning to robust hypothesis testing. Classical literature on robust hypothesis studies robust versions of the likelihood ratio test under various (non-adversarial) contamination models such as Huber's ϵ-contamination model, the total variation contamination model, or the Levy-Prokhorov metric contamination model [60], [61]. Contamination models based on f-divergences have also been analyzed for the Kullback-Liebler divergence [62] and the squared Hellinger distance [63], [64].

For general loss functions, finding the parameters $w^* \in \mathcal{W}$ is akin to minimax robust estimation. Classical literature on minimax robust estimation studies problems such as density estimation and regression under a parametrized uncertainty set of probability measures [65]. When the uncertainty sets are constructed with the Hellinger distance, methods are known for obtaining nearly optimal estimators [66]–[68].

5) Connection to Concentration of Measure: The concentration of measure phenomenon in high dimensional settings causes the measure of ϵ-expansion of sets like $\mu(A^\epsilon)$ to blow up even for small $\epsilon > 0$ [69]. Several authors suggest that adversarial examples are inevitable by appealing to concentration of measure phenomena in high dimensional [18], [70], [71]. Specifically, it has been shown that any classifier that works on data sampled from concentrated metric probability spaces is susceptible to a high adversarial risk. For instance, when the input distribution is uniform over a high dimensional sphere [70] or Boolean hypercube [44], or when the latent space of the data is a high dimensional Gaussian [72], the adversarial risk can be significantly higher than standard risk, even for small ϵ.

III. Optimal Adversarial Risk via Optimal Transport

In this section, we present our results on adversarial risk under 0-1 loss in the binary classification setting. We first define the optimal transport cost $D_\epsilon(\mu,\nu)$ between two probability measures μ and ν over a metric space $(\mathcal{X},d)$, as follows.

Definition 1 (D_ϵ Transport Cost): For $\epsilon \geq 0$, define the cost function $c_\epsilon : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ as $c_\epsilon(x,y) = \mathbb{1}\{d(x,y) > 2\epsilon\}$. The optimal transport cost D_ϵ is defined as

$$D_\epsilon(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi} c_\epsilon(x,x'). \tag{16}$$

Remark 1: For $\epsilon = 0$, the optimal cost is equivalent to the total variation distance, i.e., $D_0(\mu,\nu) = D_{TV}(\mu,\nu)$. For $\epsilon > 0$, this cost does not define a metric over the space of distributions. This is because $D_\epsilon(\mu,\nu) = 0$ does not imply μ and ν are identical. Moreover, it also does not define a pseudometric since the triangle inequality is not satisfied. To see this, observe that if μ_1, μ_2, and μ_3 are unit point masses at 0, 2ϵ, and 4ϵ, then $D_\epsilon(\mu_1,\mu_3) = 1 > 0 = D_\epsilon(\mu_1,\mu_2) + D_\epsilon(\mu_2,\mu_3)$.

Next, we present the main theorem of this section that gives the optimal risk under the binary classification setup for a data perturbing adversary.

Theorem 2: Consider the binary classification setup with $\mathcal{Y} = \{0,1\}$, where the input $x \in \mathcal{X}$ is drawn with equal probability from two distributions p_0 (for label 0) and p_1 (for label 0). We consider a set of binary classifiers of the form $\mathbb{1}\{x \in A\}$, where $A \subseteq \mathcal{X}$ is a topologically closed set. That is, the classifier corresponding to A assigns the label 1 for all $x \in A$ and the label 0 for all $x \notin A$. Consider the 0-1 loss function $\ell((x,y),A) = \mathbb{1}\{x \in A, y = 0\} + \mathbb{1}\{x \notin A, y = 1\}$. The adversarial risk with the data perturbing adversary is given by

$$R^*_\epsilon = \frac{1}{2}\left[1 - D_\epsilon(p_0,p_1)\right]. \tag{17}$$

Instantiating Theorem 2 for $\epsilon = 0$, we get $R^*_0 = \frac{1}{2}\left[1 - D_0(p_0,p_1)\right] = \frac{1}{2}\left[1 - D_{TV}(p_0,p_1)\right]$, which is the Bayes risk. It is also possible to derive weaker bounds in terms of the p-Wasserstein distance between the distributions of the two data classes, as shown in the following corollary:

Corollary 2: Under the setup considered in Theorem 2, we have the following bound for $p \geq 1$:

$$R^*_\epsilon \geq \frac{1}{2}\left[1 - \left(\frac{W_p(p_0,p_1)}{2\epsilon}\right)^p\right]. \tag{18}$$

Our next result identifies a necessary and sufficient condition for $D_\epsilon(\mu,\nu) = 0$ for probability measures on a bounded support. When this holds, adversarial risk is $1/2$; i.e., no classifier can do better than random choice.

Theorem 3: Let $\mu,\nu \in \mathcal{P}(\mathcal{X})$. Then $D_\epsilon(\mu,\nu) = 0$ if and only if $W_\infty(\mu,\nu) \leq 2\epsilon$.

A. Proofs of Theorems 2 and 3

Proof of Theorem 2: Let $A \subseteq \mathcal{X}$ be a closed set such that the classifier declares 1 on A and 0 on A^c. Suppose the true hypothesis is 0 and the observed sample is x. If the set $B(x,\epsilon) := \{y \in \mathcal{X} \mid d(x,y) \leq \epsilon\}$ has a non-empty intersection with A, then the adversary can push x to $x' \in B(x,\epsilon)\cap A$ such that the $\ell((x',0),A) = 1$. The set of all such x's is given by:

$$A^{\oplus\epsilon} := \cup_{x\in A} B(x,r). \tag{19}$$

An equivalent way to express this is using Minkowski sums:

$$A^{\oplus\epsilon} = \{a + b \mid a \in A, b \in B(0,\epsilon)\}.$$

It is easy to see that $(A^{\oplus\epsilon_1})^{\oplus\epsilon_2} = A^{\oplus(\epsilon_1+\epsilon_2)}$ for $\epsilon_1,\epsilon_2 \geq 0$.

Similarly, if the true hypothesis is 1, then the adversary can ensure a loss of 1 as long as $B(x,\epsilon)\cap A^c \neq \phi$. The set of all such x's is given by

$$(A^c)^{\oplus\epsilon} := \cup_{x\in A^c} B(x,r). \tag{20}$$

We define $A^{\ominus\epsilon}$ as

$$A^{\ominus\epsilon} := ((A^c)^{\oplus\epsilon})^c. \tag{21}$$

The robust risk over the hypothesis class of closed sets is given by

$$\begin{aligned} R^*_\epsilon &= \inf_{A \text{ closed}} \frac{1}{2}\left(p_0(A^{\oplus\epsilon}) + p_1\left((A^c)^{\oplus\epsilon}\right)\right) \\ &= \frac{1}{2}\left(1 - \sup_{A \text{ closed}} \left\{p_1\left(A^{\ominus\epsilon}\right) - p_0(A^{\oplus\epsilon})\right\}\right). \end{aligned}$$

The following Lemma provides basic topological properties of the sets $A^{\oplus\epsilon}$ and $A^{\ominus\epsilon}$.

Lemma 1 (Proof in Appendix A-B): Let $\epsilon > 0$. If A is a closed set, then $A^{\oplus\epsilon}$ and $A^{\ominus\epsilon}$ are also closed sets.

We now consider a slightly different notion of ϵ-expansions of sets similar to our definition of $A^{\oplus\epsilon}$. For $\epsilon > 0$, define

$$A^\epsilon = \{x \in \mathcal{X} \mid d(x,A) \leq \epsilon\}, \tag{22}$$

where

$$d(x,A) = \inf_{a\in A} d(x,a). \tag{23}$$

Our next Lemma shows the equivalence of $A^{\oplus\epsilon}$ and A^ϵ for closed sets A.

Lemma 2 (Proof in Appendix A-C): For a closed set A, we have $A^\epsilon = A^{\oplus\epsilon}$.

Note that $A^{\oplus\epsilon}$ and A^{ϵ} need not be equivalent when A is not closed. For example consider $(\mathcal{X}, d) = (\mathbb{R}, |\cdot|)$ with $A = (0, 1)$. Then, $A^{\oplus\epsilon} = (-\epsilon, 1+\epsilon)$ whereas $A^{\epsilon} = [-\epsilon, 1+\epsilon]$.

The main idea of our proof is to leverage Strassen's theorem (Appendix A-A), which states that

$$D_\epsilon(p_0, p_1) = \sup_{A \text{ closed}} \left\{p_1(A) - p_0(A^{2\epsilon})\right\}.$$

To prove the equality $R^*_\epsilon = \frac{1}{2}[1 - D_\epsilon(p_0, p_1)]$, notice that it is enough to prove that

$$\sup_{A \text{ closed}} p_1(A^{\ominus\epsilon}) - p_0(A^{\oplus\epsilon}) = \sup_{A \text{ closed}} p_1(A) - p_0(A^{2\epsilon}). \tag{24}$$

We need the following lemma:

Lemma 3 (Proof in Appendix A-D): Let A be a closed set. Then $(A^{\ominus\epsilon})^{\oplus\epsilon} \subseteq A$ and $A \subseteq (A^{\oplus\epsilon})^{\ominus\epsilon}$.

Figure 1 illustrates the above lemma when A is a square in $\mathbb{R}^2$ with the Euclidean distance metric.

We have the sequence of inequalities

$$\begin{aligned}
&\sup_{A \text{ closed}} p_1(A) - p_0(A^{2\epsilon}) \\
&\overset{(a)}{\geq} \sup_{A \text{ closed}} p_1(A^{\ominus\epsilon}) - p_0((A^{\ominus\epsilon})^{2\epsilon}) \\
&\overset{(b)}{\geq} \sup_{A \text{ closed}} p_1(A^{\ominus\epsilon}) - p_0(A^{\oplus\epsilon}).
\end{aligned}$$

Here, (a) follows because $A^{\ominus\epsilon}$ is contained in the set of all closed sets by Lemma 1. Inequality (b) follows by the equivalence $(A^{\ominus\epsilon})^{2\epsilon} = (A^{\ominus\epsilon})^{\oplus 2\epsilon}$ from Lemma 2, and Lemma 3 since

$$(A^{\ominus\epsilon})^{2\epsilon} = [(A^{\ominus\epsilon})^{\oplus\epsilon}]^{\oplus\epsilon} \subseteq A^{\oplus\epsilon},$$

and so $p_0((A^{\ominus\epsilon})^{2\epsilon}) \leq p_0(A^{\oplus\epsilon})$.

For the other direction, notice that

$$\begin{aligned}
&\sup_{A \text{ closed}} p_1(A^{\ominus\epsilon}) - p_0(A^{\oplus\epsilon}) \\
&\overset{(a)}{\geq} \sup_{A \text{ closed}} p_1((A^{\oplus\epsilon})^{\ominus\epsilon}) - p_0((A^{\oplus\epsilon})^{\epsilon}) \\
&\overset{(b)}{\geq} \sup_{A \text{ closed}} p_1(A) - p_0(A^{2\epsilon}).
\end{aligned}$$

Here, (a) follows because $A^{\oplus\epsilon}$ is a closed set according to Lemma 1. To see (b), first note that using Lemma 2,

$$(A^{\oplus\epsilon})^{\epsilon} = (A^{\oplus\epsilon})^{\oplus\epsilon} = A^{\oplus 2\epsilon} = A^{2\epsilon}.$$

Thus, $p_0((A^{\oplus\epsilon})^{\epsilon}) = p_0(A^{2\epsilon})$. Moreover, Lemma 3 states that $A \subseteq (A^{\epsilon})^{\ominus\epsilon}$, and so $p_1(A) \leq p_1((A^{\oplus\epsilon})^{\ominus\epsilon})$. This completes the proof. □

Comparison With Bhagoji et al. [32]: We point out that a similar result was obtained recently in [32]. A key difference is that the proof in [32] was established for a larger hypothesis class of measurable sets A; i.e., the following equality was established:

$$\sup_{A \in \sigma(\mathcal{X})} \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}) = \sup_{A \in \sigma(\mathcal{X})} \mu(A) - \nu(A^{2\epsilon}).$$

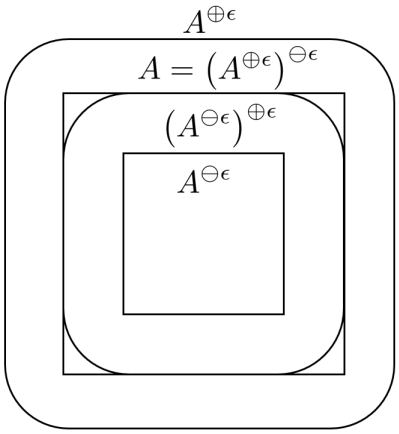


Fig. 1. Illustration of $A, A^{\oplus\epsilon}, A^{\ominus\epsilon}, (A^{\oplus}\epsilon)^{\ominus\epsilon}$, and $(A^{\ominus\epsilon})^{\oplus\epsilon}$ for a closed square in $(\mathbb{R}^2, \|\cdot\|_2)$. Observe that $(A^{\ominus\epsilon})^{\oplus\epsilon} \subseteq A$ and $A \subseteq (A^{\oplus\epsilon})^{\ominus\epsilon}$.

It is not hard to check that A^{ϵ} is closed for any measurable set A, and so

$$\sup_{A \in \sigma(\mathcal{X})} \mu(A) - \nu(A^{2\epsilon}) = \sup_{A \text{ closed}} \mu(A) - \nu(A^{2\epsilon})$$

We may restrict to the smaller hypothesis class of closed sets A and use the result in [32] to obtain an inequality

$$\sup_{A \text{ closed}} \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}) \leq \sup_{A \text{ closed}} \mu(A) - \nu(A^{2\epsilon}).$$

Our result shows that this is, in fact, an equality.

Proof of Corollary 2: From Theorem 2, we have

$$R^*_\epsilon = \frac{1}{2}\left[1 - \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi}[\mathbb{1}\{d(x, x') > 2\epsilon\}]\right].$$

For $p \geq 1$ and any $\pi \in \Pi(\mu, \nu)$, we have the following:

$$\begin{aligned}
&\mathbb{E}_{(x,x')\sim\pi}[\mathbb{1}\{d(x, x') > 2\epsilon\}] \\
&= \mathbb{E}_{(x,x')\sim\pi}[\mathbb{1}\{d(x, x')^p > (2\epsilon)^p\}] \\
&\leq \mathbb{E}_{(x,x')\sim\pi}\left[\left(\frac{d(x, x')}{2\epsilon}\right)^p\right],
\end{aligned}$$

where the last inequality follows from Markov's inequality. Therefore,

$$\begin{aligned}
R^*_\epsilon &= \frac{1}{2}\left[1 - \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi}\left[\left(\frac{d(x, x')}{2\epsilon}\right)^p\right]\right] \\
&\geq \frac{1}{2}\left[1 - \left(\frac{W_p(p_0, p_1)}{2\epsilon}\right)^p\right].
\end{aligned}$$

□

Proof of Theorem 3: Since $W_\infty(\mu, \nu) = \inf\{\delta > 0 \mid \mu(A) \leq \nu(A^\delta)$ for all measurable $A\}$, if $W_\infty(\mu, \nu) \leq 2\epsilon$, then $\mu(A) \leq \nu(A^{2\epsilon})$ for all closed sets A. Hence,

$$D_\epsilon(\mu, \nu) = \sup_{A \text{ closed}} \mu(A) - \nu(A^{2\epsilon}) \leq 0.$$

Since $D_\epsilon(\mu, \nu) \geq 0$, we conclude that $D_\epsilon(\mu, \nu) = 0$.

For the reverse direction, suppose that $D_\epsilon(\mu, \nu) = 0$. This means there exists a sequence of couplings $\{\pi\}_{i\geq 1}$ such that $\mathbb{E}_{\pi_i} c_\epsilon(x, x') \to 0$ where $\pi_i \in \Pi(\mu, \nu)$. Using a strategy as in Theorem 1, we conclude that the sequence $\{\pi_i\}$ is tight,

and thus there exists a subsequence that converges weakly to $\pi^* \in \Pi(\mu,\nu)$. Since c is a lower semicontinuous cost function, the coupling π^* satisfies $\mathbb{E}_{\pi^*} c_\epsilon(x,x') = 0$, or equivalently, $\operatorname{ess\,sup}_{(x,x')\sim\pi^*} d(x,x') \leq 2\epsilon$. Using the definition of W_∞ from (9), we conclude $W_\infty(\mu,\nu) \leq 2\epsilon$. □

IV. Optimal Adversarial Classifiers via Optimal Couplings

In this section, we explicitly compute the optimal risk and optimal classifier for a data perturbing adversary in some special cases. Instead of using D_ϵ, we have shown in Corollary 2 that the optimal adversarial risk can be lower-bounded using other well-understood metrics such as the W_p distances. However, these bounds are often too loose to use in practice, and this motivates us to study the optimal cost D_ϵ directly. In this section, we show that in certain special cases, the optimal coupling corresponding to calculating D_ϵ may be explicitly evaluated. Furthermore, in these cases, we can exactly characterize the optimal classifier and the optimal risk in the presence of an adversary. Given measures μ and ν corresponding to the two (equally likely) data classes, the general strategy we employ consists of the following steps:

(1) Propose a coupling π between μ and ν.
(2) Using this coupling, obtain the upper bound
$$D_\epsilon(\mu,\nu) \leq \mathbb{E}_{(x,x')\sim\pi} c_\epsilon(x,x').$$
(3) Identify a closed set A and compute a lower bound using
$$D_\epsilon(\mu,\nu) \geq \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}).$$
(4) Show that the lower and upper bounds match. This shows that the proposed coupling is optimal, and the sets A and A^c define the two regions of the optimal robust classifier.

In the examples we consider, guessing the set A corresponding to the optimal robust classifier is easy. The challenging part is proposing a coupling and establishing its optimality. Although we shall focus on real-valued random variables, some of our results also naturally extend to higher dimensional distributions.

In the following subsection, we review some results pertaining to optimal transport on the real line. We then present results that help in evaluating D_ϵ cost for real-valued random variables. In the subsequent subsections, we use these results to propose optimal couplings for several univariate distributions.

A. Optimal Transport on the Real Line

For a probability measure μ on $\mathbb{R}$, the cumulative distribution function (cdf) of μ is defined as $F(x) = \mu((-\infty, x])$, and for $t \in [0,1]$, the inverse cdf (or quantile function) is defined as $F^{-1}(t) = \inf\{x \in \mathbb{R} \ : \ F(x) \geq t\}$.

Lemma 4 (Theorem 2.5 in [73]): Let μ and ν be probability measures on the real line, where μ is absolutely continuous with respect to the Lebesgue measure. Then there exists a unique non-decreasing function $T : \mathbb{R} \to \mathbb{R}$ such that $\mu(T^{-1}(A)) = \nu(A)$ for any measurable set $A \subseteq \mathbb{R}$. Moreover, if F and G denote the cumulative distribution functions of μ and ν respectively, then T is given by $T(x) = G^{-1}(F(x))$.

The function T in Lemma 4 that transforms (or "pushes forward") the measure μ into ν is called a *monotone transport map*. Given a monotone transport map, we can define a coupling induced by the monotone map as follows. $(X, X') \sim \Pi(\mu,\nu)$ where $X \sim \mu$ and $X' = T(X) \sim \nu$. This coupling is also known by the name *quantile coupling*.

The following lemma shows that the coupling induced by the monotone transport map is optimal for certain cases of the cost function.

Lemma 5 (Theorem 2.9 in [73]): Let $h : \mathbb{R} \to \mathbb{R}^+$ be a strictly convex function. Let μ and ν be probability measures on the real line, where μ has a density. Consider the cost function $c(x,x') = h(x'-x)$. Suppose $\mathcal{T}_c(\mu,\nu)$ is finite. Then, $\mathcal{T}_c(\mu,\nu) = \mathbb{E}_{x\sim\mu}[c(x,T(x))]$, where T is the monotone transport map from μ to ν.

For the case of $h(x) = |x|^p$ where $p \geq 1$, Lemma IV-A shows that the optimal coupling for p-Wasserstein distance is induced by the optimal transport map. However this may not be the case for ∞-Wasserstein distance. In the following theorem, we use the monotone map from Lemma 4 to present a more concrete condition than Theorem 3 for checking when $D_\epsilon(\mu,\nu) = 0$ for measures over $\mathbb{R}$.

Theorem 4: Let μ and ν be probability measures on $\mathbb{R}$ that are absolutely continuous with respect to the Lebesgue measure with Radon-Nikodyn derivatives $f(\cdot)$ and $g(\cdot)$, respectively. Let F and G denote the cumulative distribution functions of μ and ν respectively. Then $D_\epsilon(\mu,\nu) = 0$ if and only if $\|F^{-1} - G^{-1}\|_\infty \leq 2\epsilon$.

Proof: Consider the monotone transport map from μ to ν given by $T(x) = G^{-1}(F(x))$ as in Lemma 4. We shall show that this map satisfies $|T(x) - x| \leq 2\epsilon$ for all $x \in \mathbb{R}$, and so the optimal transport cost D_ϵ must be 0. To see this, note that

$$\begin{aligned} T(x) - x &= G^{-1}(F(x)) - x \\ &\leq F^{-1}(F(x)) + 2\epsilon - x \\ &= 2\epsilon, \end{aligned}$$

where the last equality is in the μ-almost sure sense. A similar argument shows $x - T(x) \leq 2\epsilon$, and thus $|T(x) - x| \leq 2\epsilon$.

For the converse, suppose that there exists a $t_0 \in (0,1)$ such that $G^{-1}(t_0) - F^{-1}(t_0) > 2\epsilon$. Equivalently, $G^{-1}(t_0) > F^{-1}(t_0) + 2\epsilon$. Applying the G function on both sides,

$$t_0 > G(F^{-1}(t_0) + 2\epsilon).$$

Consider the set $\tilde{A} = (-\infty, F^{-1}(t_0)]$. For this set, notice that

$$\nu(\tilde{A}^{2\epsilon}) = \nu((-\infty, F^{-1}(t_0) + 2\epsilon]) = G(F^{-1}(t_0) + 2\epsilon).$$

Thus, we have

$$\begin{aligned} D_\epsilon(\mu,\nu) &= \sup_A \mu(A) - \nu(A^{2\epsilon}) \\ &\geq \mu(\tilde{A}) - \nu(\tilde{A}^{2\epsilon}) \\ &= t_0 - G(F^{-1}(t_0) + 2\epsilon) \\ &> 0. \end{aligned}$$

A similar argument may also be made for the case when $F^{-1}(t_0) - G^{-1}(t_0) > 2\epsilon$. □

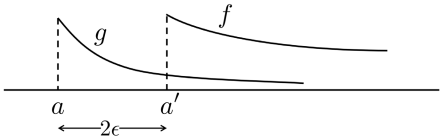


Fig. 2. Figure illustrating the conditions in Lemma 6.

The above argument shows that monotone transport maps are optimal when $D_\epsilon = 0$. But monotone maps are not always optimal for the cost function $c_\epsilon(\cdot,\cdot)$. Consider for example the two measures $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$, and $\epsilon = 0.1$. The monotone map in this case is $T(x) = x + 1$, which gives unit cost of transportation. However, Theorem 5 shows that the optimal transport cost in this example is strictly smaller than 1.

Checking the condition $\|F^{-1} - G^{-1}\| \leq 2\epsilon$ is not always easy. We identify a simple but useful characterization in the following corollary:

Corollary 3: Let μ and ν be as in Theorem 4. Suppose that for every $x \in \mathbb{R}$, we have $F(x) \geq G(x)$ and $F(x) \leq G(x + 2\epsilon)$. Then $D_\epsilon(\mu,\nu) = 0$.

Proof: Applying the G^{-1} function to both sides of both inequalities, we arrive at

$$T(x) \geq x, \quad \text{and} \quad T(x) \leq x + 2\epsilon.$$

This gives $|T(x) - x| \leq 2\epsilon$ for all x, which concludes the proof. □

Theorem 4 may also be applied to finite positive measures μ, ν with $\mu(\mathbb{R}) = \nu(\mathbb{R}) = U < \infty$ with simple scaling. In what follows, we define a notion of optimal transport for finite positive measures that may have unequal masses.

Definition 2: [Optimal Transport Cost for General Measures] Let μ and ν be as in Theorem 4. Suppose that $\mu(\mathbb{R}) = U$ and $\nu(\mathbb{R}) = V$ and $U \leq V$. Let ν' be a measure on $\mathbb{R}$ with Radon-Nikodyn derivative g' such that $\nu'(\mathbb{R}) = U$. We say $\nu' \subseteq \nu$, or ν' is contained in ν, if $g(x) \geq g'(x)$ ν-almost surely. Then the optimal transport cost $D_\epsilon(\mu,\nu)$ is defined as

$$D_\epsilon(\mu,\nu) = \inf_{\nu' \subseteq \nu} D_\epsilon(\mu,\nu').$$

Note that the amount of mass being moved is $\min(U,V) = U$.

In the following two lemmas, we identify conditions under which $D_\epsilon(\mu,\nu) = 0$ for finite positive measures with unequal mass.

Lemma 6: Let μ and ν be as in Theorem 4. Assume that $\mu(\mathbb{R}) = U$ and $\nu(\mathbb{R}) = V$. Suppose the following conditions hold:

1) The support of g is a subset of $[a, +\infty)$ and the support of f is a subset of $[a + 2\epsilon, +\infty) =: [a', +\infty)$.
2) For all $x \in \mathbb{R}$, we have $g(x) \leq f(x + 2\epsilon)$.

Then $D_\epsilon(\mu,\nu) = 0$. A similar result holds if the supports of g and f are subsets of $(-\infty, -a]$ and $(-\infty, -a - 2\epsilon]$ respectively, and $f(-x - 2\epsilon) \geq g(-x)$.

Proof: Consider the transport map $T(x) = x + 2\epsilon$ applied to ν. This map has the effect of "translating" the measure ν by 2ϵ to the right. Call this translated measure η. Since $f(x) \geq g(x - 2\epsilon)$, it is immediate that $\eta \subseteq \mu$. Moreover, the transport cost is $D_\epsilon(\nu,\eta) = 0$. This shows that $D_\epsilon(\mu,\nu) = 0$. □

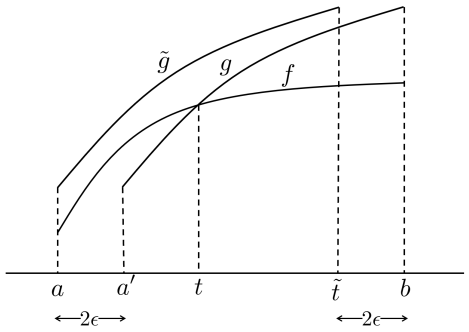


Fig. 3. Figure illustrating the conditions in Lemma 7. Note that in general $\tilde{t}$ need not be equal to $b - 2\epsilon$ as shown in the figure.

Lemma 7: Let μ and ν be as in Theorem 4. Assume that $\mu(\mathbb{R}) = \nu(\mathbb{R}) = U$. Suppose the following conditions hold (see Figure 3 for an illustration):

1) Let $a, b \in \mathbb{R}$ be such that the support of f is a subset of $[a,b]$ and the support of g is a subset of $[a', b] := [a + 2\epsilon, b]$.
2) There exists $t \in [a,b]$ such that $f(x) \geq g(x)$ for $x \in [a,t)$, and $f(x) \leq g(x)$ for $x \in (t,b]$.
3) Let $\tilde{g}(x) = g(x + 2\epsilon)$. Note that the support $\tilde{g}$ is within $[a, b - 2\epsilon]$. There exists $\tilde{t} \in [a, b - 2\epsilon]$ such that $f(x) \leq \tilde{g}(x)$ for $x \in [a,\tilde{t})$, and $f(x) \geq \tilde{g}(x)$ for $x \in (\tilde{t}, b - 2\epsilon]$.

Then $D_\epsilon(\mu,\nu) = 0$. A mirror image of this result also holds: $D_\epsilon(\mu,\nu) = 0$ when the support of f is a subset of $[b, c + 2\epsilon]$, that of g is a subset of $[b,c]$, and $f(x) \leq g(x)$ for $x \in [b,t)$ and $f(x) \geq g(x)$ for $x \in [t, c + 2\epsilon]$; and for $\tilde{g}(x) = g(x + 2\epsilon)$ we have $f(x) \geq \tilde{g}(x)$ for $x \in [b + 2\epsilon, \tilde{t})$ and $f(x) \leq g(x)$ for $x \in [\tilde{t}, c + 2\epsilon]$.

Proof: We first prove $F(x) \geq G(x)$. To see this, consider $H(x) = F(x) - G(x)$. Since the derivative of H is $f - g$, it must be that H is increasing from $[a,t)$ and decreasing from $[t,b]$. Also, we have $H(a) = H(b) = 0$, and so the function H must be non-negative in $[a,b]$. Equivalently, we must have $F(x) \geq G(x)$ for $x \in \mathbb{R}$. We now prove $F(x) \leq G(x + 2\epsilon)$. Consider $\tilde{H}(x) = F(x) - \tilde{G}(x)$. By condition (3), the derivative of this function is negative from $[a,\tilde{t}]$ and positive from $[\tilde{t}, b]$. Thus, the function $\tilde{H}$ decreases on the interval $[a,\tilde{t})$ and increases on the interval $[\tilde{t}, b]$. Note that since $\tilde{H}(a) = \tilde{H}(b) = 0$, the function $\tilde{H}$ must be non-positive in the interval $[a,b]$. Thus, we have $F(x) \leq G(x + 2\epsilon)$. Applying Corollary 3 concludes the proof. □

B. Gaussian Distributions With Identical Variances

Theorem 5: Let $p_0 = \mathcal{N}(\mu_0, \sigma^2)$ and $p_1 = \mathcal{N}(\mu_1, \sigma^2)$ in the metric space $(\mathbb{R}, |\cdot|)$. Assume $\mu_0 < \mu_1$ without loss of generality. Then the following hold:

1) If $\epsilon \geq \frac{|\mu_0 - \mu_1|}{2}$, the optimal robust risk is $1/2$. A constant classifier achieves this risk.
2) If $\epsilon < \frac{|\mu_0 - \mu_1|}{2}$, the optimal classifier satisfies $A = \left[\frac{\mu_1 + \mu_0}{2}, +\infty\right]$, where A is the region where the classifier

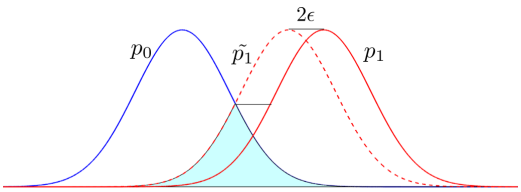


Fig. 4. Optimal coupling for two Gaussians with identical variances. The shaded region within p_0 is translated by 2ϵ to p_1, whereas the remaining is mass in p_0 is moved at a cost of 1 per unit mass.

declares label 1. The optimal risk in this case is $\int_{\frac{\mu_1+\mu_0}{2}-\epsilon}^{\infty} p_0(x)dx = Q\left(\frac{\frac{\mu_1-\mu_0}{2}-\epsilon}{\sigma}\right)$.

The lower bound of $1/2$ on the adversarial risk is trivially achieved by the constant classifier. Part (1) of the theorem states that for large enough ϵ, this is the best one can do. For smaller values of ϵ, the above theorem shows that the most robust classifier is the same as the MLE classifier. For larger values of ϵ, the MLE classifier has a risk *larger than* $1/2$; i.e., it is worse than the constant classifier.

Proof: We shall prove (1) first. Note that if $\epsilon \geq \frac{\mu_1-\mu_0}{2}$, the transport map T defined by $T(x) = x+(\mu_1-\mu_0)$ transports p_0 to p_1. Moreover, this coupling satisfies $|T(x) - x| = \mu_1 - \mu_0 \leq 2\epsilon$. Thus, the optimal transport cost for this coupling is 0, and therefore so is $D_\epsilon(p_0, p_1)$. This gives the lower bound

$$R^*_\epsilon \geq \frac{1}{2}.$$

However, since the constant classifier achieves the lower bound, we conclude $R^*_\epsilon = 1/2$.

For part (2), we consider the following strategy for transporting the mass from p_0 to p_1. As shown in Figure 4, consider the distribution $\tilde{p}_1$ obtained by shifting p_1 to the left by 2ϵ. That is, $\tilde{p}_1(x) = p_1(x + 2\epsilon)$. Define $q : \mathbb{R} \to \mathbb{R}$ as $q(x) = \min(p_0(x), \tilde{p}_1(x))$. It is evident that the overlapping area between $\tilde{p}_1$ and p_0 (i.e., the area under the curve $q(x)$) maybe be translated by 2ϵ to the right so that it lies entirely under the curve $p_1(x)$. More precisely, $q(x - 2\epsilon) \leq p_1(x)$ for all $x \in \mathbb{R}$. Hence, the area under $q(x)$ may be transported to $p_1(x)$ at 0 cost by using the transport map $T(x) = x+2\epsilon$. It is easily verified that the area under $q(x)$ equals $2Q\left(\frac{\frac{\mu_1-\mu_0}{2}-\epsilon}{\sigma}\right)$, and so the total cost of transporting p_0 to p_1 is at most $1-2Q\left(\frac{\frac{\mu_1-\mu_0}{2}-\epsilon}{\sigma}\right)$. Plugging this into the lower bound, we see that

$$R^*_\epsilon \geq Q\left(\frac{\frac{\mu_1-\mu_0}{2}-\epsilon}{\sigma}\right).$$

Since this risk is achieved by the MLE classifier, we conclude that this is the optimal robust risk and the MLE classifier is the optimal robust classifier. □

Theorem 5 can be easily extended to d-dimensional Gaussians with the same identity covariances. Our results may be summarized in the following theorem:

Theorem 6: Let $p_0 = \mathcal{N}(\mu_0, \sigma^2\ I_d)$ and $p_1 = \mathcal{N}(\mu_1, \sigma^2\ I_d)$ in the metric space $(\mathbb{R}, ||\cdot||_2)$. Then the following hold:

1) If $\epsilon \geq \frac{||\mu_0-\mu_1||_2}{2}$, the optimal robust risk is $1/2$. A constant classifier achieves this risk.
2) If $\epsilon < \frac{||\mu_0-\mu_1||_2}{2}$, the optimal classifier is given by the following halfspace:

$$A = \left\{x\ :\ (\mu_1 - \mu_0)\left(x - \frac{\mu_0+\mu_1}{2}\right) \geq 0\right\}. \quad (25)$$

Comparison to Bhagoji et al. [32]: Bhagoji *et al.* also explore optimal classifiers for multivariate normal distributions. In fact, they show a more general version of our Theorems 5 and 6 by considering data distributions $\mathcal{N}(\mu_0, \Sigma)$ and $\mathcal{N}(\mu_1, \Sigma)$, and an adversary that perturbs within ℓ_p-balls.

In the following subsections, we shall generalize Theorem 5 in a different way by considering various interesting examples of univariate distributions and identifying optimal couplings for these.

C. Gaussians With Arbitrary Means and Variances

We shall introduce a general coupling strategy and apply it to the special case of Gaussian random variables. Given two probability measures μ and ν on $\mathbb{R}$, our strategy consists of the following steps:

(1) Partition the real line into $K \geq 1$ intervals $S_i, 1 \leq i \leq K$, and let the restriction of μ to S_i be μ_i.
(2) Partition the real line into $K \geq 1$ intervals $T_i, 1 \leq i \leq K$, and let the restriction of ν to T_i be ν_i.
(3) Transport mass from μ_i to ν_i such that $D_\epsilon(\mu_i, \nu_i) = 0$. (We shall use the definition of mass transport between measures with unequal masses from definition 2.) The transport maps used in these K problems may be arbitrary; however, we shall often use versions of the monotone optimal transport map [74].

Our next lemma is specific to Gaussian pdfs:

Lemma 8: Let f and g be Gaussian pdfs corresponding to $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, respectively. Assume $\sigma_1^2 > \sigma_2^2$. Then the equation $f(x) - g(x) = 0$ has exactly two solutions $s_1 < \mu_2 < s_2$.

Proof: By scaling and translating, we may set $\mu_2 = 0$ and $\sigma_2^2 = 1$. Solving $f(x) - g(x) = 0$ is equivalent to solving the quadratic equation

$$\frac{x^2}{2} - \frac{(x-\mu_1)^2}{2\sigma_1^2} = \log \sigma_1.$$

Simplifying, we wish to solve

$$x^2(\sigma_1^2 - 1) + 2\mu_1\ x - (\mu_1^2 + 2\sigma_1^2 \log \sigma_1) = 0.$$

Since $\sigma_1 > 1$, the above quadratic has two distinct roots: one negative and one positive. This proves the claim. □

We shall call the two points where f and g intersect as the left and right intersection points.

Theorem 7: Let μ and ν be the Gaussian measures $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$, respectively. Assume $\sigma_1^2 > \sigma_2^2$ without loss of generality. Let $m > 0$ be such that $f(m+\epsilon) = g(m-\epsilon)$. Let $A = (-\infty, -m] \cup [m, +\infty)$. Then the optimal transport cost between μ and ν is given by

$$\begin{aligned} D_\epsilon(\mu, \nu) &= \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}) \\ &= 2Q\left(\frac{m+\epsilon}{\sigma_1}\right) - 2Q\left(\frac{m-\epsilon}{\sigma_2}\right). \end{aligned}$$

TABLE I
THE REAL LINE IS PARTITIONED INTO FIVE REGIONS FOR μ AND ν, AS SHOWN IN THE TABLE

μ_{--}	$(-\infty, -m-\epsilon]$
μ_{-}	$(-m-\epsilon, -r]$
μ_{0}	$(-r, +r)$
μ_{+}	$[r, m+\epsilon)$
μ_{++}	$[m+\epsilon, \infty)$

ν_{--}	$(-\infty, -m+\epsilon]$
ν_{-}	$(-m+\epsilon, -r]$
ν_{0}	$(-r, +r)$
ν_{+}	$[r, m-\epsilon)$
ν_{++}	$[m-\epsilon, \infty)$

The corresponding robust risk is

$$R_\epsilon^* = \frac{1 - \mu(A^{\ominus\epsilon}) + \nu(A^{\oplus\epsilon})}{2}.$$

Moreover, if μ corresponds to hypothesis 1, the optimal robust classifier declares label 1 on the set A.

Proof: We shall propose a map that transports μ to ν. (See Figure 5 for an illustration.) The existence of a $m > 0$ such that $f(m+\epsilon) = g(m-\epsilon)$ is guaranteed by Lemma 8. Consider $r \in (0, m-\epsilon)$ whose value will be provided later. First, we partition $\mathbb{R}$ into the five regions for μ and ν, as shown in Table I. For μ, these partitions are $(-\infty, -m-\epsilon]$, $(-m-\epsilon, -r]$, $(-r, +r)$, $[r, m+\epsilon)$, and $[m+\epsilon, \infty)$. Let μ restricted to these intervals be μ_{--}, μ_{-}, μ_0, μ_{+}, and μ_{++}, respectively. The measure ν is also partitioned five ways, but the intervals used in this case are slightly modified to be $(-\infty, -m+\epsilon]$, $(-m+\epsilon, -r]$, $(-r, r)$, $[r, m-\epsilon)$, and $[m-\epsilon, +\infty)$. Call ν restricted to these intervals ν_{--}, ν_{-}, ν_0, ν_{+}, and ν_{++}, respectively.

The transport plan from μ to ν will consist of five maps transporting $\mu_{--} \to \nu_{--}$, $\mu_{-} \to \nu_{-}$, $\mu_0 \to \nu_0$, $\mu_{+} \to \nu_{+}$, and $\mu_{++} \to \nu_{++}$. In each case, we plan to show that $D_\epsilon(\mu_*, \nu_*) = 0$, where $*$ ranges over all possible subscripts in $\{--, -, 0, +, ++\}$. Note that these measures do not necessarily have identical masses, and thus by Definition 2, we are transporting a quantity of mass equal to the minimum mass among the two measures. For this reason, even though the transport cost is $D_\epsilon(\mu_*, \nu_*) = 0$, it does not mean $D_\epsilon(\mu, \nu) = 0$.

Consider μ_{++} and ν_{++}. We have $f(m+\epsilon) = g(m-\epsilon)$ by the choice of m. We argue that for any $t \geq 0$, we must have $f(m+\epsilon+t) \geq g(m-\epsilon+t)$. This is because any two Gaussian pdfs can intersect in at most two points. By Lemma 8, the ϵ-shifted Gaussian pdfs $f(x+\epsilon)$ and $g(x-\epsilon)$ have m as their right intersection point, and there are no additional points of intersection to the right of m. Since the tail of f is heavier, it means that $f(m+\epsilon+t) \geq g(m-\epsilon+t)$ for all $t \geq 0$. By Lemma 6, we can now conclude $D_\epsilon(\mu_{++}, \nu_{++}) = 0$. A similar argument also shows $D_\epsilon(\mu_{--}, \nu_{--}) = 0$.

Before we consider μ_{-} and ν_{-}, we first define r as follows: Pick $r > 0$ such that $\mu([-m-\epsilon, -r)) = \nu([-m+\epsilon, -r))$. To see that such an r must exist, consider the functions $a(t) := \mu([-m-\epsilon, t))$ and $b(t) := \nu([-m+\epsilon, t))$ as t ranges over $(-m+\epsilon, 0)$. When $t = -m+\epsilon$, we have $a(t) > b(t) = 0$. When $t = 0$, we have $a(t) = 1/2 - \mu_{--}(\mathbb{R}) < b(t) = 1/2 - \nu_{--}(\mathbb{R})$. Thus, there must exist a $t_0 \in (-m+\epsilon, 0)$ such that $a(t_0) = b(t_0)$. Pick the smallest (i.e., the leftmost) such t_0, and set $-r = t_0$. Call $f(\cdot)$ restricted to $[-m-\epsilon, -r)$ and $g(\cdot)$ restricted to $[-m+\epsilon, -r)$ as f_- and g_-, respectively, and their corresponding cdfs F_- and G_-, respectively. We claim that μ_- and ν_- satisfy all three conditions from Lemma 7. Since the supports of f_- and g_- are $[-m-\epsilon, -r)$ and $[-m+\epsilon, -r)$, condition (1) is immediately verified. To check condition (2), we break up the interval $[-m-\epsilon, -r)$ into two parts: $[-m-\epsilon, -s)$ and $[-s, -r)$, where s is such that $f(-s) = g(-s)$. Observe that $f_- \geq g_-$ on $[-m-\epsilon, -s)$, whereas $f_- \leq g_-$ on $[-s, -r)$. This shows that condition (2) is satisfied. We have $g_-(-m+\epsilon) = f_-(-m-\epsilon)$. Again, using Lemma 8 the 2ϵ-shifted Gaussian pdf $f(x-2\epsilon)$ and $g(x)$ have $-m+\epsilon$ as their left intersection point, and the right intersection point is to the right of 0. Thus, we have $f(x-2\epsilon) \leq g(x)$ for all $x \in [-m+\epsilon, 0] \supseteq [-m+\epsilon, r)$. Using this domination, we conclude that $f_- \leq \tilde{g}_-$ in the interval $[-m-\epsilon, -r-2\epsilon)$ and $f_- \geq g_- = 0$ in the interval $(-r-2\epsilon, -r]$, and so condition (3) is satisfied. Applying Lemma 7, we conclude $D_\epsilon(\mu_-, \nu_-) = 0$. An essentially identical argument may be used to show $D_\epsilon(\mu_+, \nu_+) = 0$. The minor difference being that r is chosen to satisfy $\mu([r, m+\epsilon)) = \nu([r, m-\epsilon))$, and the mirror image of Lemma 7 is applied.

Finally, consider the interval $(-r, +r)$. In this interval, $f(x) \leq g(x)$ for every point. Hence, a transport map from μ_0 to ν_0 is obtained by simply considering the identity function. Any remaining mass in μ is moved to ν arbitrarily, incurring a cost of at most 1 per unit mass. The total cost of transport is then upper-bounded by

$$\begin{aligned}
&D_\epsilon(\mu, \nu) \\
&\leq 1 - [\min(\mu_{--}, \nu_{--}) + \min(\mu_{-}, \nu_{-}) + \min(\mu_0, \nu_0) \\
&\qquad + \min(\mu_{+}, \nu_{+}) + \min(\mu_{++}, \nu_{++})] \\
&= 1 - [\nu_{--} + \mu_{-} + \mu_0 + \mu_{+} + \nu_{++}] \\
&= 1 - \mu([-m-\epsilon, m+\epsilon]) - 2\nu([m-\epsilon, \infty)) \\
&= \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}) \\
&= 2Q\left(\frac{m+\epsilon}{\sigma_1}\right) - 2Q\left(\frac{m-\epsilon}{\sigma_2}\right).
\end{aligned}$$

where for brevity we have denoted $\mu_*(\mathbb{R})$ as μ_*. However, we also have

$$D_\epsilon(\mu, \nu) \geq \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}).$$

The lower and upper bounds match and this concludes the proof. The robust risk R_ϵ^* is given by Theorem 2. The robust risk of the classifier that declares label 1 on the set A is easily seen to be R_ϵ^*. □

We now extend the above proof strategy to demonstrate the optimal coupling for Gaussians with arbitrary means and arbitrary variances. Our main result is the following:

Theorem 8: Let μ and ν be Gaussian measures $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ respectively. Assume $\sigma_1^2 > \sigma_2^2$ without loss of generality. Let $m_1, m_2 > 0$ be such that $f(-m_1 - \epsilon) = g(-m_1 + \epsilon)$ and $f(m_2 + \epsilon) = g(m_2 - \epsilon)$. Let $A = (-\infty, -m_1] \cup [m_2, \infty)$. Then the optimal transport cost between μ and ν is given by

$$D_\epsilon(\mu, \nu) = \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}).$$

Consequently, the robust risk is given by

$$R_\epsilon^* = \frac{1}{2}(1 - \mu(A^{\ominus\epsilon}) + \nu(A^{\oplus\epsilon})).$$

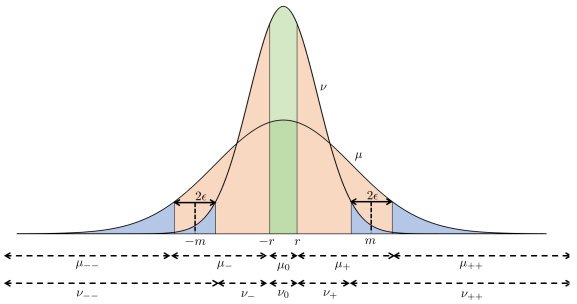


Fig. 5. Optimal transport coupling for centered Gaussian distributions μ and ν. As in the proof of Theorem 7, we divide the real line into five regions. The transport plan from μ to ν consists of five maps transporting $\mu_{--} \to \nu_{--}$ (blue regions to the left), $\mu_- \to \nu_-$ (orange regions to the left), $\mu_0 \to \nu_0$ (green regions in the middle), $\mu_+ \to \nu_+$ (orange regions to the right), and $\mu_{++} \to \nu_{++}$ (blue regions to the right).

If μ corresponds to hypothesis 1, the optimal robust classifier declares label 1 on the set A.

Proof: We first note that the existence of m_1 and m_2 in the theorem statement is guaranteed by Lemma 8. As in the proof of Theorem 7, we shall divide the real line into five regions as shown in Table II where we define r_1 and r_2 shortly. Using an identical strategy as in Theorem 7, we conclude $D_\epsilon(\mu_{--}, \nu_{--}) = D_\epsilon(\mu_{++}, \nu_{++}) = 0$. Define r_1 as the leftmost point where $\mu([-m_1 - \epsilon, r_1)) = \nu([-m_1 + \epsilon, r_1))$. Similarly, define r_2 to be the rightmost point such that $\mu([r_2, m_2 + \epsilon)) = \nu([r_2, m_2 - \epsilon))$. We shall now prove $D_\epsilon(\mu_-, \nu_-) = 0$ by using Lemma 7. Verifying conditions (1) and (2) is exactly as in that of Theorem 7. The novel component of this proof is verifying condition (3), since the domination used in the proof of Theorem 7 does not work in this case due to the asymmetry. Consider the pdfs $f_-(x)$ and $g_-(x + 2\epsilon)$. These two pdfs, being restrictions of Gaussian pdfs to suitable intervals, may only intersect in at most two points. One of these points of intersection is $-m_1 - \epsilon$ by the choice of m_1, so there can be at most one other point of intersection in the interval $[-m_1 - \epsilon, -r_1 - 2\epsilon]$. Note that there may be no point of intersection in this interval. However, the key observation is that in both cases, condition (3) continues to be satisfied. To see this, suppose that there is a point of interaction $\tilde{t}$. In this case, $f_- \le \tilde{g}_-$ in $[-m_1 - \epsilon, \tilde{t})$, and $f_- \ge g_-$ in $(\tilde{t}, -r_1]$. If there is no point of intersection, then $f_- \le \tilde{g}_-$ in $[-m_1 - \epsilon, -r_1 - 2\epsilon)$, and $f_- \ge g_- = 0$ in $(-r_1 - 2\epsilon, -r_1]$. This verifies condition (3). Using Lemma 7, we conclude $D_\epsilon(\mu_-, \nu_-) = 0$. An identical approach gives $D_\epsilon(\mu_+, \nu_+) = 0$. Since $f(x) \le g(x)$ for all points in the interval $(-r_1, r_2)$, the identity map may be used to conclude $D_\epsilon(\mu_0, \nu_0) = 0$.

Any remaining mass in μ is moved to ν arbitrarily, incurring a cost of at most 1 per unit mass. The total cost of transport is then upper-bounded by

$$\begin{aligned} &D_\epsilon(\mu, \nu) \\ &\le 1 - [\min(\mu_{--}, \nu_{--}) + \min(\mu_-, \nu_-) + \min(\mu_0, \nu_0) \\ &\qquad + \min(\mu_+, \nu_+) + \min(\mu_{++}, \nu_{++})] \\ &= 1 - [\nu_{--} + \mu_- + \mu_0 + \mu_+ + \nu_{++}] \end{aligned}$$

TABLE II
THE REAL LINE IS PARTITIONED INTO FIVE REGIONS FOR μ AND ν AS SHOWN IN THE TABLE

μ_{--}	$(-\infty, -m_1 - \epsilon]$	ν_{--}	$(-\infty, -m_1 + \epsilon]$
μ_-	$(-m_1 - \epsilon, -r_1]$	ν_-	$(-m_1 + \epsilon, -r_1]$
μ_0	$(-r_1, +r_2)$	ν_0	$(-r_1, +r_2)$
μ_+	$[r_2, m_2 + \epsilon)$	ν_+	$[r_2, m_2 - \epsilon)$
μ_{++}	$[m_2 + \epsilon, \infty)$	ν_{++}	$[m_2 - \epsilon, \infty)$

$$\begin{aligned} &= 1 - \mu([-m_1 - \epsilon, m_2 + \epsilon]) - \nu((-\infty, -m_1 + \epsilon)) \\ &\qquad - \nu([m_2 - \epsilon, \infty)) \\ &= \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}), \end{aligned}$$

where for brevity we have denoted $\mu_*(\mathbb{R})$ as μ_*, where $*$ ranges over all possible subscripts in $\{--, -, 0, +, ++\}$. The rest of the proof is identical to that of Theorem 7. □

D. Beyond Gaussian Examples

The coupling strategy for Gaussian random variables can also be applied to other univariate examples that share some similarities with the Gaussian case. To illustrate, we describe the optimal classifier and optimal coupling for uniform distributions and triangular distributions.

Theorem 9 (Uniform Distributions): Let μ and ν be uniform measures on closed intervals I and J respectively. Without loss of generality, we assume $|I| \le |J|$. Then the optimal robust risk is $\nu(I^{2\epsilon})$ and the optimal classifier is given by $A = I^\epsilon$.

Proof: See Appendix B-A □

In the following, we present the optimal adversarial risk and optimal classifier for symmetric triangular distributions. For $\delta > 0$, we use $\Delta(m, \delta)$ to denote a triangular distribution with support $[m-\delta, m+\delta]$ and mode at m. The pdf of such a distribution is given by the function $f(x) = \frac{1}{\delta} \max\left\{1 - \frac{|x-m|}{\delta}, 0\right\}$.

The next lemma is similar to Lemma 8, but is specific to symmetric triangular distributions.

Lemma 9: Let μ and ν correspond to the triangular distributions $\Delta(m_1, \delta_1)$ and $\Delta(m_2, \delta_2)$ with pdfs f and g respectively. Assume $\delta_1 < \delta_2$. Then,

1) If $|m_1 - m_2| > \delta_2 + \delta_1$, then the equation $f(x) - g(x) = 0$ has no solutions on the supports of μ or ν.
2) If $\delta_2 - \delta_1 < |m_1 - m_2| \le \delta_2 + \delta_1$, then the equation $f(x) - g(x) = 0$ has exactly one solution u on the support of μ. Further, $u \ge m_1$ if and only if $m_1 \le m_2$.
3) If $|m_1 - m_2| \le \delta_2 - \delta_1$, then the equation $f(x) - g(x) = 0$ has exactly two solutions $l \in [m_1 - \delta_1, m_1]$ and $r \in [m_1, m_1 + \delta_1]$ on the support of μ.

Proof: We may assume that $m_1 \le m_2$, as case of $m_1 \ge m_2$ follows by symmetry.

Suppose $|m_1 - m_2| > \delta_2 + \delta_1$. Then $m_1 + \delta_1 < m_2 - \delta_2$. Hence, the supports of μ and ν are disjoint and the result follows trivially.

Suppose $\delta_2 - \delta_1 < |m_1 - m_2| \le \delta_2 + \delta_1$. Then, $m_1 + \delta_1 \in [m_2 - \delta_2, m_2 + \delta_2]$ and $m_1 - \delta_1 \notin [m_2 - \delta_2, m_2 + \delta_2]$. Hence, the only solution u to $f(x) - g(x) = 0$ occurs at the intersection of the graph of $g(x)$ with the line segment joining the points $(m_1, 1/\delta_1)$ and $(m_1 + \delta_1, 0)$. Clearly, $u \ge m_1$.

Suppose $|m_1 - m_2| \leq \delta_2 - \delta_1$. Then, $m_1 - \delta_1 \geq m_2 - \delta_2$ and $m_1 + \delta_1 \leq m_2 - \delta_2$. Hence, $[m_1 - \delta_1, m_1 + \delta_1] \subset [m_2 - \delta_2, m_2 + \delta_2]$. It follows that $f(m_1 - \delta_1) - g(m_1 - \delta_1) < 0$, $f(m_1) - g(m_1) > 0$ and $f(m_1 + \delta_1) - g(m_1 + \delta_1) < 0$. Since $f(x) - g(x)$ is a continuous function, there must be $l \in [m_1 - \delta_1, m_1]$ and $r \in [m_1, m_1 + \delta_1]$ such that $f(l) - g(l) = 0$ and $f(r) - g(r) = 0$. Moreover, $f(x) > g(x)$ for $x \in (l, r)$ and $f(x) < g(x)$ for $x \in (m_2 - \delta_2, l) \cup (r, m_2 + \delta_2)$. Hence, l and r are the only solutions to $f(x) - g(x) = 0$ on the support of μ. □

Theorem 10 (Triangular Distributions): Let μ and ν correspond to the triangular distributions $\Delta(m_1, \delta_1)$ and $\Delta(m_2, \delta_2)$ with pdfs f and g respectively. Without loss of generality, assume $\delta_1 < \delta_2$ and $m_1 < m_2$ (the case of $m_1 > m_2$ follows from symmetry). Let $2\epsilon \in (0, \min(2\delta_1, \delta_2 - \delta_1))$. Let $l = \sup\{x \leq m_1 : f(x + \epsilon) = g(x - \epsilon)\}$ and $r = \inf\{x \geq m_1 : f(x - \epsilon) = g(x + \epsilon)\}$. Let A be the set defined as follows.

1) If $m_2 - m_1 \geq \delta_2 + \delta_1 + 2\epsilon$, then $A = (-\infty, m_1 + \delta_1 + \epsilon]$.
2) If $\delta_2 - \delta_1 - 2\epsilon \leq m_2 - m_1 < \delta_2 + \delta_1 + 2\epsilon$, then $A = (-\infty, r]$.
3) If $m_2 - m_1 < \delta_2 - \delta_1 - 2\epsilon$, then $A = [l, r]$.

Then $D_\epsilon(\mu, \nu) = \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon})$, and the robust risk is

$$R^*_\epsilon = \frac{1 - \mu(A^{\ominus\epsilon}) + \nu(A^{\oplus\epsilon})}{2},$$

and if μ corresponds to hypothesis 1, then the optimal robust classifier declares label 1 on A.

Proof: See Appendix B-B □

V. Adversarial Risk for Continuous Loss Functions

It is natural to ask if the results for 0-1 loss may be extended to continuous losses. In this section, we present adversarial risk bounds in regression-like settings with continuous losses and investigate Questions 1 and 2 in light of these bounds.

A. Optimal Adversarial Risk

In this section, we prove lower and upper bounds on the optimal adversarial risk for distribution perturbing adversaries with budget $\epsilon \geq 0$. To prove lower bounds, we consider the W_∞-distribution perturbing adversary with budget ϵ, since this bound is valid for all W_p-distribution perturbing adversaries. Similarly, we prove upper bounds for the W_1-distribution perturbing adversary with budget ϵ.

A Trivial Lower Bound: We start by presenting a trivial lower bound on the optimal adversarial risk. We shall assume that for all $\epsilon \geq 0$, the optimal hypothesis w^*_ϵ exists to simplify presentation. The proofs can be easily modified by considering sequences of hypothesis such that $\liminf_i R^*_\epsilon(w_i) = R^*_\epsilon$ in case w^*_ϵ does not exist.

Theorem 11: The optimal adversarial risk is at least as large as the optimal standard risk, that is, $R^*_\epsilon \geq R^*_0$.

Proof: We have the sequence of inequalities:

$$R^*_\epsilon = R_\epsilon(\ell, w^*_\epsilon) \geq R_0(\ell, w^*_\epsilon) \geq R_0(\ell, w^*_0) = R^*_0.$$

The first inequality holds because $R_\epsilon(\ell, w)$ is a non-decreasing function of ϵ for any fixed ℓ and w. The second inequality follows from the fact that the adversarially optimal classifier w^*_ϵ is sub-optimal for minimizing the standard risk. □

Note that the bound in Theorem 11 does not depend on the strength of the adversary ϵ, and hence it may not be very tight for large ϵ. In what follows, we show tighter lower bounds for R^*_ϵ that depend on ϵ.

For the lower bound, we consider loss functions that are convex with respect to the input x, as defined below.

Definition 3 (Convex Loss Function): We say that the loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \to \mathbb{R}^+$ is convex with respect to the input if it satisfies the following condition.

$$\ell((x', y), w) - \ell((x, y), w) \geq \langle \nabla_x \ell((x, y), w), x' - x \rangle. \quad (26)$$

Theorem 12: The adversarial risk for a loss function satisfying (26) is bounded as follows.

$$R^*_\epsilon \geq R^*_0 + \inf_{w \in \mathcal{W}} \mathbb{E}_z \left[\sup_{d(x,x') \leq \epsilon} \langle \nabla_x \ell((x, y), w), x' - x \rangle \right]. \quad (27)$$

Remark 2: The lower bound holds for any p-Wasserstein distribution perturbing adversary with budget ϵ.

Note that adversary's metric $d(\cdot, \cdot)$ may not be the same as the norm on the Hilbert space $\mathcal{X}$. In the special case d corresponds to the norm $\|\cdot\|_{\text{adv}}$, we can tighten the result of Theorem 12 as follows.

Corollary 4: In the setting of Theorem 12, if $d(x, x') = \|x - x'\|_{\text{adv}}$ for $x, x' \in \mathcal{X}$, then the following bound holds:

$$R^*_\epsilon \geq R^*_0 + \epsilon \inf_{w \in \mathcal{W}} \mathbb{E}_z[\|\nabla_x \ell((x, y), w)\|_{\text{adv}^*}], \quad (28)$$

where $\|\cdot\|_{\text{adv}^*}$ is the dual norm of $\|\cdot\|_{\text{adv}}$.

Proof of Theorem 12: Recall the notation $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $z = (x, y)$. Since w^*_ϵ is sub-optimal for minimizing standard risk, we have

$$\mathbb{E}_z[\ell((x, y), w^*_\epsilon)] \geq \mathbb{E}_z[\ell((x, y), w^*_0)].$$

Hence,

$$\begin{aligned}
&R^*_\epsilon - R^*_0 \\
&= \mathbb{E}_z \left[\sup_{d(x,x') \leq \epsilon} \ell((x', y), w^*_\epsilon) \right] - \mathbb{E}_z[\ell((x, y), w^*_0)] \\
&\geq \mathbb{E}_z \left[\sup_{d(x,x') \leq \epsilon} \ell((x', y), w^*_\epsilon) \right] - \mathbb{E}_z[\ell((x, y), w^*_\epsilon)] \\
&= \mathbb{E}_z \left[\sup_{d(x,x') \leq \epsilon} \ell((x', y), w^*_\epsilon) - \ell((x, y), w^*_\epsilon) \right] \\
&\geq \mathbb{E}_z \left[\sup_{d(x,x') \leq \epsilon} \langle \nabla_x \ell((x, y), w^*_\epsilon), x' - x \rangle \right], \\
&\geq \inf_{w \in \mathcal{W}} \mathbb{E}_z \left[\sup_{d(x,x') \leq \epsilon} \langle \nabla_x \ell((x, y), w^*_\epsilon), x' - x \rangle \right].
\end{aligned}$$

□

Proof of Corollary 4: From the proof of Theorem 12, we have

$$R^*_\epsilon - R^*_0 \geq \mathbb{E}_z \left[\sup_{d(x,x')\leq\epsilon} \langle \nabla_x \ell((x',y),w^*_\epsilon), x'-x\rangle \right].$$

Under the condition that $d(x,x') = \|x-x'\|_{\text{adv}}$,

$$\begin{aligned} &\sup_{d(x,x')\leq\epsilon} \langle \nabla_x \ell((x',y),w^*_\epsilon), x'-x\rangle \\ &= \sup_{\|\delta\|_{\text{adv}}\leq\epsilon} \langle \nabla_x \ell((x',y),w^*_\epsilon), \delta\rangle \\ &= \epsilon \|\nabla_x \ell((x',y),w^*_\epsilon)\|_{\text{adv}^*}. \end{aligned}$$

□

Next, we prove an upper bound for the adversarial risk for a W_1-distribution perturbing adversary. As noted earlier, this upper bound also holds for a W_p-distribution perturbing adversary of the same budget, where $1 \leq p \leq \infty$. We make the following assumption on the loss function:

Definition 4 (L_w-Lipschitz Loss Function): We say that the loss function $\ell : \mathcal{Z}\times\mathcal{W} \to \mathbb{R}^+$ is L_w-Lipschitz with respect to the input if it satisfies the following condition.

$$|\ell((x',y),w) - \ell((x,y),w)| \leq L_w \|x'-x\|. \tag{29}$$

Theorem 13: The adversarial risk for a W_1-distribution perturbing adversary with budget ϵ satisfies $\widehat{R}^{1,*}_\epsilon \leq R^*_0 + \epsilon L_{w^*_0}$.

The proof of this result uses an optimal transport idea from [75].

Proof of Theorem 13: Suppose that the infimum for $\widehat{R}^{1,*}_\epsilon$ in equation (5) is attained at $\widehat{w}^*_\epsilon$ and the supremum for $\widehat{R}^1_\epsilon(\ell,\widehat{w}^*_\epsilon)$ in equation (10) is attained for $\gamma^* \in \Gamma^1_\epsilon$. For $y\in\mathcal{Y}$, recall that $\rho^{\gamma^*}_{x'|y} \in \mathcal{P}(\mathcal{X})$ denotes the distribution of the perturbed data point $x'\in\mathcal{X}$. Let $\pi_y \in \Pi(\rho_{x|y}, \rho^{\gamma^*}_{x'|y})$ be such that $W_1(\rho_{x|y},\rho^{\gamma^*}_{x'|y}) = \mathbb{E}_{(x,x')\sim\pi_y} d(x,x')$. Then

$$\begin{aligned} &\widehat{R}^{1,*}_\epsilon - R^*_0 \\ &= \mathbb{E}_{(x',y)\sim\rho_y\rho^{\gamma^*}_{x'|y}} \ell((x',y),\widehat{w}^*_\epsilon) \\ &\qquad - \mathbb{E}_{(x,y)\sim\rho_y\rho_{x|y}} \ell((x,y),w^*_0) \\ &\overset{(a)}{\leq} \mathbb{E}_{(x',y)\sim\rho_y\rho^{\gamma^*}_{x'|y}} \ell((x',y),\widehat{w}^*_0) \\ &\qquad - \mathbb{E}_{(x,y)\sim\rho_y\rho_{x|y}} \ell((x,y),w^*_0) \\ &\overset{(b)}{=} \mathbb{E}_y \mathbb{E}_{(x,x')\sim\pi_y} [\ell((x',y),\widehat{w}^*_0) - \ell((x,y),\widehat{w}^*_0)] \\ &\overset{(c)}{\leq} \mathbb{E}_y \mathbb{E}_{(x,x')\sim\pi_y} d(x,x') \cdot L_{w^*_0} \\ &\overset{(d)}{\leq} \epsilon L_{w^*_0}. \end{aligned}$$

Here, (a) follows from the definition of $\widehat{w}^*_\epsilon$, (b) follows from linearity of expectation since π_y is a coupling of (x,x') that preserves the marginals, (c) follows from the Lipschitz assumption and (d) follows from the fact that $\gamma^* \in \Gamma^1_\epsilon$. □

B. Optimal Adversarial Classifier

In Sections III and V-A, we looked at Question 1 and showed that the adversarial risk can be strictly lower-bounded as a function of adversarial budget ϵ. In this section, we tackle Question 2 and analyze how w^*_ϵ or $\widehat{w}^*_\epsilon$ may deviate from w^*_0. For the case of 0-1 loss, the optimal classifier can change drastically even with small change in the adversarial budget ϵ. For instance, consider the setting of Theorem 5. When ϵ changes from being less than $\frac{|\mu_0-\mu_1|}{2}$ to greater than $\frac{|\mu_0-\mu_1|}{2}$, the optimal classifier changes from a halfspace to a constant classifier. Studying the 0-1 loss is hard because closed sets are not parametrized easily. Hence we focus on the case of convex loss functions–where convexity is with respect to w—to derive bounds in this section. Deriving bounds without strong convexity assumptions appears challenging. To see this, observe that there may be multiple global optima w^*_0 when $\epsilon = 0$. The optimal hypothesis can jump from one global optimal to a different one—possibly far away—even without any adversary.

Since our proof technique uses the upper and lower bounds for adversarial losses obtained in Section V-A, the bounds for deviation of w^*_ϵ and $\widehat{w}^*_\epsilon$ are identical. Now, we prove a theorem on how much the optimal classifier can change in the presence of an adversary.

Theorem 14: For a loss function ℓ that satisfies (29), and is λ-strongly convex with respect to w, the following result holds:

$$\|w^*_\epsilon - w^*_0\| \leq \sqrt{\frac{2\epsilon L_{w^*_0}}{\lambda}}. \tag{30}$$

Proof of Theorem 14: We have the following series of inequalities.

$$\begin{aligned} \epsilon L_{w^*_0} &\overset{(a)}{\geq} R^*_\epsilon - R^*_0 \\ &\overset{(b)}{\geq} R_0(\ell,w^*_\epsilon) - R_0(\ell,w^*_0) \\ &\overset{(c)}{\geq} \frac{\lambda}{2}(\nabla^2_w R_0(\ell,w^*_0))\|w^*_\epsilon - w^*_0\|^2. \end{aligned}$$

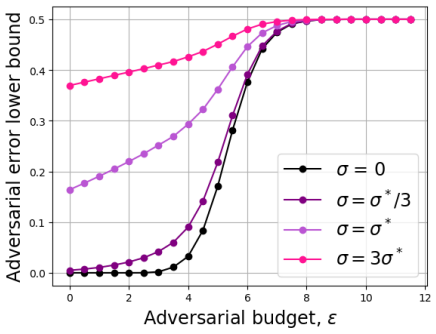
Here, (a) follows from Theorem 13, (b) follows from the fact that w^*_ϵ is sub-optimal for minimizing $R_0(\ell,w)$, and (c) follows from the λ-strong convexity of ℓ with respect to w. □

The above theorem shows that larger values of λ prevent the adversary from changing the hypothesis drastically. If the loss function is merely convex but not strongly convex, adding a quadratic penalty $\frac{\lambda}{2}\|w\|^2$ to the loss function will ensure strong convexity.

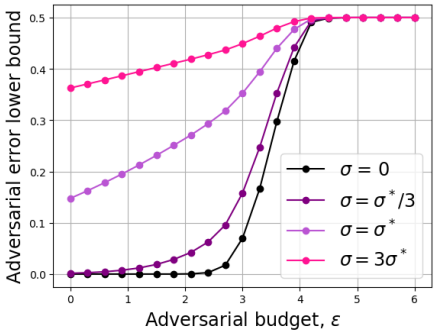
VI. Experiments

In this section, we present lower bounds on the optimal adversarial risk for empirical distributions derived from several real world datasets.

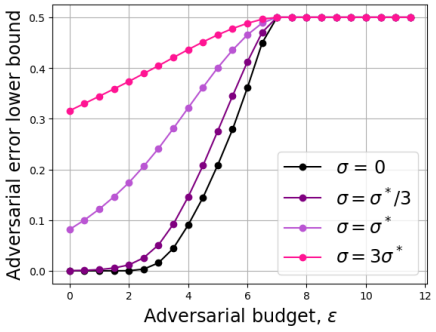
For the case of empirical distributions, the computation of the optimal transport cost in (16) can be formulated as a linear program and solved efficiently. Moreover, when the number of data points in the two empirical distributions is the same, the problem of finding the optimal coupling between the two distributions is reduced to an assignment problem (see Proposition 2.11 in [76]), wherein the task is to optimally match each data point from the first distribution to a distinct data point from the second distribution. Using



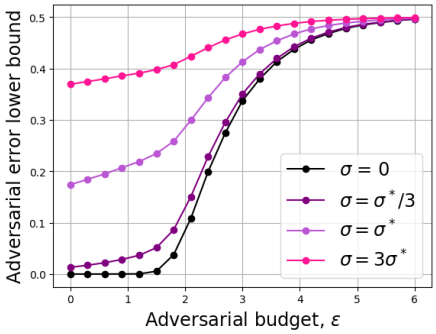
(a) CIFAR10



(b) MNIST

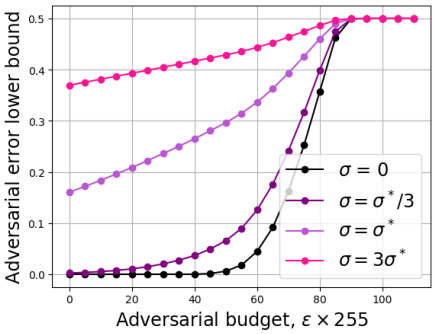


(c) Fashion-MNIST

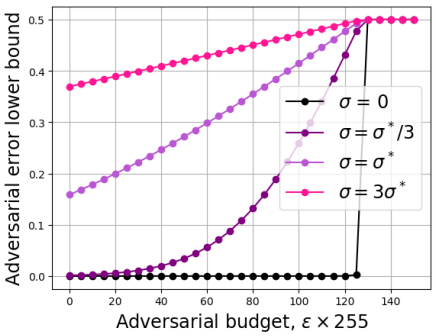


(d) SVHN

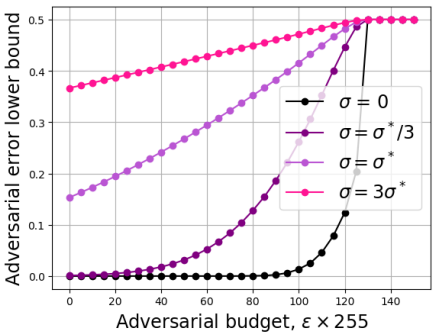
Fig. 6. Lower bounds on adversarial risk computed using Theorem 2 for ℓ_2 adversary. The curves with $\sigma = 0$ give the exact optimal risk for empirical distributions, while the other cuvers give lower bounds on the optimal risk for Gaussian mixtures based on the empirical distributions using the coupling in Theorem 6.



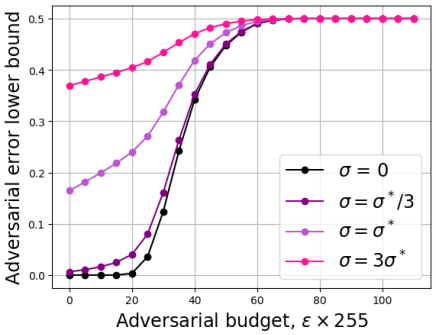
(a) CIFAR10



(b) MNIST



(c) Fashion-MNIST



(d) SVHN

Fig. 7. Lower bounds on adversarial risk computed using Theorem 2 for ℓ_∞ adversary. The curves with $\sigma = 0$ give the exact optimal risk for empirical distributions, while the other cuvers give lower bounds on the optimal risk for Gaussian mixtures based on the empirical distributions using the coupling in Theorem 6.

this methodology, we evaluate the optimal risk for ℓ_2 and ℓ_∞ adversaries for classes 3 and 5 in CIFAR10, MNIST, Fashion-MNIST and SVHN datasets. The results for other pairs of classes are very similar, and are therefore omitted for brevity. For MNIST, Fashion-MNIST and SVHN datasets, we evaluate the optimal adversarial risk given in Theorem 2 by randomly sampling 5000 data points from each class. The results are showing in Figures 6 and 7 with the legend $\sigma = 0$, for ℓ_2 and ℓ_∞ adversaries respectively.

Since a major fraction of the data points in the empirical distributions are well-separated in ℓ_2 and ℓ_∞ metrics, the optimal risk bound remains 0 even for high ϵ. For instance, for CIFAR10 dataset, the optimal risk remains 0 for ϵ as high as $40/255$ for ℓ_∞. Similar results were also obtained in Bhagoji *et al.* [32]. However, the optimal risk bounds for the true distributions may not be 0 for high ϵ, as it is unreasonable to expect a perfectly robust optimal classifier under very strong adversarial perturbations. In addition, a common technique while training for a classifier is to augment the dataset with Gaussian perturbed samples for robustness and generalization [77], [78]. Motivated by this, we also compute optimal risk lower bounds on Gaussian mixture distribution with the data points as the centers with scaled identity covariances. $\sigma = 0$ corresponds to the empirical distribution of the data points from the two classes. As σ increases, the overlap in the probability mass between the two classes increases. This allows for the cost of optimal coupling that achieves D_ϵ to decrease, thus leading to a higher, possibly non-trivial bound for R^*_ϵ.

To compute the optimal risk lower bound for Gaussian mixture, we use a coupling between the mixture distributions in two steps. In the first step, we solve for the optimal coupling that gives the exact optimal risk for the empirical distributions. This gives a pairwise matching of data points between the two empirical distributions. In the second step, we use the optimal coupling for multidimensional Gaussians from Theorem 6 to transport the mass in the Gaussians within each pair. Overall, this transport map gives an upper bound on the D_ϵ optimal transport cost between the two mixture distributions. Using this, we obtain the lower bounds on adversarial risk shown in Figures 6 and 7.

For ℓ_2 and ℓ_∞ adversaries, Figures 6 and 7 show the lower bounds for various values of the variance σ used for the Gaussian mixture, where σ^* is half of the mean distance between data points from the two distributions. As explained previously, we see in Figures 6 and 7 that the lower bound curves for higher values of σ are above those for lower values. For instance, the optimal risk for CIFAR10 dataset under ℓ_2 perturbation with $\epsilon = 3$ is 0.25 for $\sigma = \sigma^*$. That is, the adversarial error rate for CIFAR10 with $\epsilon = 3$ for any algorithm cannot be less than 0.25 even when trained with Gaussian data augmentation (with $\sigma = \sigma^*$). In comparison, the lower bound obtained in Bhagoji *et al.* [32] (which is equivalent to the case of $\sigma = 0$) is 0 for $\epsilon = 3$. Computation of non-trivial lower bounds for higher values of ϵ on adversarial error rate as in Figures 6 and 7 is made possible by our analysis on the optimal coupling to achieve D_ϵ between multivariate Gaussians in section IV-B.

VII. Discussion

In this paper, we have analyzed two notions of *adversarial risk*: one resulting from a distribution perturbing adversary ($\widehat{R}^*_\epsilon$) and the other from a data perturbing adversary (R^*_ϵ). We have introduced the D_ϵ optimal transport distance between probability distributions. Through an application of duality in the optimal transport cost formulation (via Strassen's theorem), we have shown that D_ϵ completely characterizes the optimal adversarial risk R^*_ϵ for the case of binary classification under 0-1 loss function. For general loss functions, we give lower bounds on R^*_ϵ and upper bounds on $\widehat{R}^*_\epsilon$ in terms of the Lipschitz and strong convexity parameters of the loss function.

Our analysis raises several interesting questions: How big is the gap between $\widehat{R}^*_\epsilon$ and R^*_ϵ for different kinds of loss functions? Is it possible to directly lower bound $\widehat{R}^*_\epsilon$ without appealing to its dependence on R^*_ϵ? Does there exist an optimal transport distance akin to D_ϵ that characterizes $\widehat{R}^*_\epsilon$? As evidenced by experiments, our bounds for general loss functions are not particularly tight. Furthermore, we need fairly strong assumptions such as convexity and Lipschitz property for the loss function to state these bounds. It would be interesting to study if these conditions may be relaxed and if tighter bounds could be obtained.

In analysing the adversarial risk for 0-1 loss functions, we give a novel coupling strategy based on monotone mappings that solves the D_ϵ optimal transport problem for symmetric unimodal distributions like Gaussian, triangular, and uniform distributions. Employing the duality in the optimal transport, we also obtain the adversarially optimal classifier under these settings. Our coupling analysis calls for an interesting open question: Is there a general coupling strategy, akin to the maximal coupling strategy to achieve the total variation transport cost, that works for a broader class of distributions? If yes, this gives us a handle on analyzing the nature of optimal decision boundaries in the adversarial setting. Optimal transport between measures with unequal mass has received attention in recent work [79]. We plan to investigate if the version of transport from Definition 2 is useful in other contexts, and whether computational methods as in [76] may be used to compute it in practice.

Our analysis for 0-1 loss reveals how the optimal risk smoothly changes from Bayes risk as the data perturbing budget ϵ is increased. Somewhat more surprisingly, our analysis shows that in some cases, the optimal classifier can change abruptly in the presence of an adversary even for small changes in ϵ. It remains to be seen if these observations on optimal risk and optimal classifier also hold for the distribution perturbing adversary.

Using our characterization of R^*_ϵ in terms of D_ϵ, we obtain the optimal risk attainable for classification of real-world datasets like CIFAR10, MNIST, Fashion-MNIST and SVHN. Moreover, levaraging our optimal coupling strategy for Gaussian distributions, we also obtain lower bounds on optimal risk for Gaussian mixtures based on these datasets. These lower bounds have implications for the limits of data augmentation strategies using Gaussian perturbations. Our bounds on adversarial risk are classifier agnostic, and only depend on the data disributions. In addition, our bounds are efficiently computable for empirical/mixture distributions via reformulation as a linear program. However, our characterization of R^*_ϵ in terms of D_ϵ is limited to the binary classification setting. It is not clear if a similar characterization is possible for multi-class classification, perhaps using multi-marginal optimal transport theory [80].

Finally, we remark that analzing the D_ϵ optimal transport cost may be interesting in itself. The optimal transport cost $c_\epsilon(x, x') = \mathbb{1}\{d(x, x') > 2\epsilon\}$ is discontinuous and does not satisfy triangle inequality. This makes it hard to analyse D_ϵ using standard techniques in optimal transport literature. For instance, it would be interesting tighten the bounds from [81] concerning rates of converges of D_ϵ between empirical distributions converges to D_ϵ between the true data-generating distributions.

APPENDIX A
PROOFS FOR SECTION III

A. Strassen's Theorem

Strassen's theorem is a special case for the Kantorovich duality in the case of a 0-1 loss. The statement provided below is as in Villani [74, Corollary 1.28]:

Lemma 10: Let the input X be drawn from a Polish space $\mathcal{X}$. Let $\Pi(p_0, p_1)$ be the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals p_0 and p_1. Then for $\epsilon \geq 0$ and $A \subseteq \mathcal{X}$,

$$\inf_{\pi \in \Pi(p_0,p_1)} \pi[d(x, x') > \epsilon] = \sup_{A \ closed} \left\{p_0(A) - p_1(A^{\oplus\epsilon})\right\}.$$

B. Proof of Lemma 1

Let A be a closed set and let B be the closed ball of radius ϵ. Fix $\delta > 0$. Let $\{z_i\}_{i\geq 1}$ be a sequence of points in $A^{\oplus\epsilon}$ converging to a limit z. Assume without loss of generality that $d(z_i, z) < \delta/2$. We shall show that $z \in A^{\oplus\epsilon}$ as well. Note that every z_i admits an expression $z_i = a_i + b_i$, where $a_i \in A$ and $b_i \in B$. Since B is a compact set, there exists a subsequence among the $\{b_i\}$ sequence that converges to $b^* \in B$. Fix a $\delta > 0$ and pick a subsequence $\{\tilde{b}_i\}_{i\geq 1}$ such that $\tilde{b}_i \to b^*$ and $|\tilde{b}_i - b^*| < \delta/2$ for all $i > 0$. Denote the corresponding subsequence of $\{a_i\}$ by $\{\tilde{a}_i\}$ and $\{z_i\}$ by $\{\tilde{z}_i\}$. Observe that

$$z - \tilde{a}_i = (z - \tilde{z}_i) + (b^* - \tilde{b}_i) - b^*,$$

and so by the triangle inequality

$$d(z, \tilde{a}_i) < \delta/2 + \delta/2 + \epsilon = \epsilon + \delta.$$

Thus $\tilde{a}_i \in B(z, \epsilon + \delta) \cap A$, which is a compact set, giving a convergent subsequence within the $\{\tilde{a}_i\}$ sequence. Let that subsequence converge to a^*. We must have $a^* \in A$ and $b^* \in B$ since A and B are closed. This means $z = a^* + b^*$ must lies in $A^{\oplus\epsilon}$, which shows that $A^{\oplus\epsilon}$ is closed.

Recall that $A^{\ominus\epsilon} = ((A^c)^{\oplus\epsilon})^c$. Since A^c is an open set, it is enough to show that $C^{\oplus\epsilon}$ is open if C is open. Let $z \in C^{\oplus\epsilon}$, which means $z = c + b$ for some $c \in C$ and $b \in B$. Consider a small open ball of radius δ around c, called $N_\delta(c)$ that lies entirely in C. This is possible since C is assumed to be open. Now observe that $N_\delta(z) \subseteq C^{\oplus\epsilon}$, since $N_\delta(z) = N_\delta(c) + b$. This shows that every point $z \in C^{\oplus\epsilon}$ admits a small ball around it that is contained in $C^{\oplus\epsilon}$, or equivalently, $C^{\oplus\epsilon}$ is open. This completes the proof.

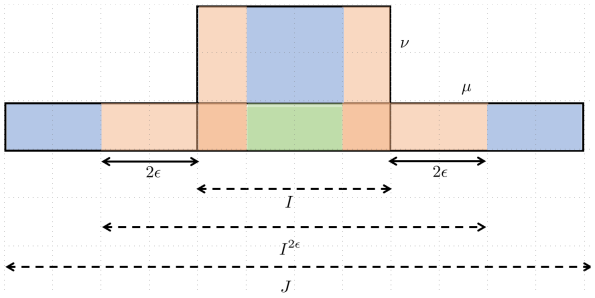


Fig. 8. Optimal coupling for two uniform distributions. The region shaded in green is kept in place (at no cost). The two regions shaded in orange are transported monotonically from either side at a cost not exceeding 2ϵ per unit mass. The remaining region in blue is moved at the cost of 1 per unit mass.

C. Proof of Lemma 2

Let $x \in A^{\oplus\epsilon}$. Then there exists an $a \in A$ such that $d(x, a) \leq \epsilon$, which means $d(x, A) \leq \epsilon$, and so $x \in A^\epsilon$. This shows that $A^{\oplus\epsilon} \subseteq A^\epsilon$.

To prove the reverse direction, suppose $x \in A^\epsilon$. This means we can a sequence of points $\{a_i\}$ such that $a_i \in A$ and $\liminf_i d(x, a_i) \leq \epsilon$. Fix a $\delta > 0$ and assume without loss of generality that $d(x, a_i) \leq \epsilon + \delta$ for all $i > 0$. Then $a_i \in B(x, \epsilon + \delta) \cap A$ for all $i > 0$. As A is closed, the set $B(x_i, \epsilon + \delta) \cap A$ is compact, and there exists a subsequence $\{\tilde{a}_i\}$ that converges to $a^* \in A$. By the triangle inequality, $d(x, a^*) \leq d(x, \tilde{a}_i) + d(\tilde{a}_i, a^*)$. Taking $\liminf_i$ on both sides yields

$$d(x, a^*) \leq \liminf_i d(x, \tilde{a}_i) \leq \epsilon.$$

This implies $x \in A^{\oplus\epsilon}$, and we conclude $A^\epsilon \subseteq A^{\oplus\epsilon}$.

D. Proof of Lemma 3

We claim that a point $x \in A^{\ominus\epsilon}$ if and only if $B(x, \epsilon)$ lies entirely in A. If this were not the case, then we could find a $y \in A^c$ such that $d(x, y) \leq \epsilon$, and so $x \in (A^c)^{\oplus\epsilon}$, which implies $x \notin ((A^c)^{\oplus\epsilon})^c = A^{\ominus\epsilon}$. Conversely, if $B(x, \epsilon) \in A$ then $d(x, y) > \epsilon$ for all $y \in A^c$, and so $x \notin (A^c)^{\oplus\epsilon}$, which means $x \in ((A^c)^{\oplus\epsilon})^c = A^{\ominus\epsilon}$. This observation implies that $(A^{\ominus\epsilon})^{\oplus\epsilon} \subseteq A$.

Using the above logic for $A^{\oplus\epsilon}$, we see that a point $x \in (A^{\oplus\epsilon})^{\ominus\epsilon}$ if and only if $B(x, \epsilon) \subseteq A^{\oplus\epsilon}$. By definition of $A^{\oplus\epsilon}$, every point $x \in A$ satisfies $B(x, \epsilon) \subseteq A^{\oplus\epsilon}$. Thus, if $x \in A$ then $x \in (A^{\oplus\epsilon})^{\ominus\epsilon}$. Equivalently, $A \subseteq (A^{\oplus\epsilon})^{\ominus\epsilon}$.

APPENDIX B
PROOFS FOR SECTION IV

A. Proof of Theorem 9

Like in the proof for Theorem 7, we prove Theorem 9 by partitioning the real line into several regions for μ and ν, and transporting mass between these regions. Figure 8 shows the optimal coupling for the case when $I^{2\epsilon} \subseteq J$.

We first prove a lower bound. Choose the set $A = I$, we have that

$$D_\epsilon(\mu, \nu) \geq \mu(A) - \nu(A^{2\epsilon}) = 1 - \nu(I^{2\epsilon}). \tag{31}$$

To establish the upper bound, we need to find a coupling that transports μ to ν such that the cost of transportation is bounded above by $1-\nu(I^{2\epsilon})$. Without loss of generality, let $I=[-w_1,w_1]$ and $J=[c-w_2,c+w_2]$ for some $c>0$ and $0<w_1\leq w_2$.

Case 1: $2\epsilon < w_2-w_1$. We split the analysis into the following five sub-cases.

Case 1(a): $c\in[w_1+w_2+2\epsilon,\infty)$. In this case, the intervals I and J are separated by at least 2ϵ. Hence, $\nu(I^{2\epsilon})=0$, and therefore, $D_\epsilon(\mu,\nu)\geq 1-\nu(I^{2\epsilon})=1$. Combining this with the fact that $D_\epsilon(\mu,\nu)\leq 1$, we get that $D_\epsilon(\mu,\nu)=1=1-\nu(I^{2\epsilon})$.

Case 1(b): $c\in[-w_1+w_2-2\epsilon, w_1+w_2+2\epsilon)$. In this case, $\nu(I^{2\epsilon})=\nu([c-w_2,w_1+2\epsilon])=(w_1+2\epsilon-c+w_2)/(2w_2)\leq 1$. Since $\mu([-w_1,w_1])=1\geq\nu([c-w_2,w_1+2\epsilon])$, there must exist a $u\in[-w_1,w_1]$ such that $\mu([u,w_1])=\nu([c-w_2,w_1+2\epsilon])$. Solving for u, we get the following.

$$\begin{aligned}\frac{w_1-u}{2w_1}&=\mu([u,w_1])\\&=\nu([c-w_2,w_1+2\epsilon])\\&=\frac{w_1+2\epsilon-c+w_2}{2w_2}\\\implies u&=w_1-\frac{w_1}{w_2}(w_1+2\epsilon-c+w_2).\end{aligned}$$

Since $w_1/w_2<1$, the above equation for u shows that $u>w_1-(w_1+2\epsilon-c+w_2)=c-w_2-2\epsilon$. Hence, $(c-w_2)-u<2\epsilon$.

Let μ_0 be the restriction of μ to $[u,w_1]$ and ν_0 be the restriction of ν to $[c-w_2,w_1+2\epsilon]$. Then, by construction, $\mu_0(\mathbb{R})=\nu_0(\mathbb{R})$. By Lemma 4, we have a monotone transport map $T:[u,w_1]\to[c-w_2,w_1+2\epsilon]$ that transports μ_0 to ν_0 given by $T(x)=\frac{w_1+2\epsilon-c+w_2}{w_1-u}(x-u)+(c-w_2)$. Note that T transports u to $c-w_2$ and w_1 to $w_1+2\epsilon$. Also, $T(x)>x$. Since T has a slope greater than 1, $T(x)-x$ is an increasing function. Moreover, $T(w_1)-w_1=2\epsilon$ and $T(u)-u=(c-w_2)-u<2\epsilon$. Hence, $|T(x)-x|\leq 2\epsilon$ for all $x\in[u,w_1]$. Hence, $D_\epsilon(\mu_0,\nu_0)=0$. Therefore, $D_\epsilon(\mu,\nu)\leq 1-\min(\mu_0,\nu_0)=1-\nu([c,w_1+2\epsilon])=1-\nu(I^{2\epsilon})$. Combining with the lower bound in (31), we conclude that $D_\epsilon(\mu,\nu)=1-\nu(I^{2\epsilon})$.

Case 1(c): $c\in(-w_2+w_1+2\epsilon,-w_1+w_2-2\epsilon)$. In this case, $\nu(I^{2\epsilon})=\nu([-w_1-2\epsilon,w_1+2\epsilon])=(2w_1+4\epsilon)/(2w_2)\leq 1$. Since $\mu([0,w_1])=1/2>\nu(0,w_1+2\epsilon)$, there must exists a $v\in[0,w_1]$ such that $\mu([v,w_1])=\nu([0,w_1+2\epsilon])$. Let μ_+ be the restriction of μ to $[u,w_1]$ and ν_+ be the restriction of ν to $[0,w_1+2\epsilon]$. Then, by construction, $\mu_+(\mathbb{R})=\nu_+(\mathbb{R})$. Similar to the map T in case 1b, there exists a monotone transport map $T_+:[u,w_1]\to[0,w_1+2\epsilon]$ such that $|T_+(x)-x|\leq 2\epsilon$. Hence, $D_\epsilon(\mu_+,\nu_+)=0$. Similarly, let μ_- be the restriction of μ to $[-w_1,-u]$ and ν_+ be the restriction of ν to $[-w_1-2\epsilon,0]$. Then by symmetry, there also exists a monotone transport map $T_-:[-w_1,-u]\to[-w_1-2\epsilon,0]$ such that $|T_-(x)-x|\leq 2\epsilon$. Hence, $D_\epsilon(\mu_-,\nu_-)=0$. Therefore,

$$\begin{aligned}D_\epsilon(\mu,\nu)&\leq 1-[\min(\mu_+,\nu_+)+\min(\mu_-,\nu_-)]\\&=1-[\nu([0,w_1+2\epsilon])+\nu([-w_1-2\epsilon,0])]\\&=1-\nu([-w_1-2\epsilon,w_1+2\epsilon])\\&=1-\nu(I^{2\epsilon}).\end{aligned}$$

Combining with the lower bound in (31), we conclude that $D_\epsilon(\mu,\nu)=1-\nu(I^{2\epsilon})$.

Case 1(d): $c\in(-w_1-w_2-2\epsilon,w_1-w_2+2\epsilon]$. The geometry of this case is a mirror image of that in case 1b. Hence, just as in case 2, we have $D_\epsilon(\mu,\nu)=1-\nu(I^{2\epsilon})$.

Case 1(e): $c\in(-\infty,-w_1-w_2-2\epsilon]$. Like in case 1, the intervals I and J are separated by at least 2ϵ. Hence, similar to case 1, we get that $D_\epsilon(\mu,\nu)=1=1-\nu(I^{2\epsilon})$.

Case 2: $2\epsilon\geq w_2-w_1$. In this case, we have the following sub-cases.

Case 2(a): $c\in[w_1+w_2+2\epsilon,\infty)$. Like in case 1a, the intervals I and J are separated by at least 2ϵ. Hence, $D_\epsilon(\mu,\nu)=1=1-\nu(I^{2\epsilon})$.

Case 2(b): $c\in[w_1-w_2+2\epsilon,w_1+w_2+2\epsilon)$. Since $[w_1-w_2+2\epsilon,w_1+w_2+2\epsilon)\subseteq[-w_1+w_2-2\epsilon,w_1+w_2+2\epsilon)$, the coupling obtained in Case 1b can be directly applied in this case. Hence, we again have $D_\epsilon(\mu,\nu)=1=1-\nu(I^{2\epsilon})$.

Case 2(c): $c\in(-w_1+w_2-2\epsilon,w_1-w_2+2\epsilon)$. In this case, the supports of μ and ν are within 2ϵ of each other. More specifically, $J\subseteq I^{2\epsilon}$. Hence, $\nu(I^{2\epsilon})=1$. Let T denote the monotone transport map from μ and ν as defined in Lemma 4. Then, $T(x)=\frac{w_2}{w_1}(x-w_1)+(c+w_2)$. Note that T maps $[-w_1,w_1]$ to $[c-w_2,c+w_2]$ monotonically. Since the supports of μ and ν are within 2ϵ of each other, we have $|T(x)-x|\leq 2\epsilon$. Hence, $D_\epsilon(\mu,\nu)=0=1-\nu(I^{2\epsilon})$.

Case 2(d): $c\in(-w_1-w_2-2\epsilon,-w_1+w_2-2\epsilon]$. This case is a mirror image of case 2b and hence the result $D_\epsilon(\mu,\nu)=1-\nu(I^{2\epsilon})$ remains the same.

Case 2(e): $c\in(-\infty,-w_1-w_2-2\epsilon]$. Like in case 1a, the intervals I and J are separated by at least 2ϵ. Hence, $D_\epsilon(\mu,\nu)=1=1-\nu(I^{2\epsilon})$.

It is easily checked that the error attained by the proposed classifier also matches the bound, which completes the proof.

B. Proof of Theorem 10

We have the following cases:

Case 1: $m_2-m_1>\delta_1+\delta_2+2\epsilon$.

In this case μ and ν have disjoint supports separated by at least 2ϵ. Moreover, $\mu(A^{\ominus\epsilon})=1$ and $\nu(A^{\oplus\epsilon})=0$. Then,

$$\begin{aligned}D_\epsilon(\mu,\nu)&=\sup_{A\ closed}\mu(A^{\ominus\epsilon})-\nu(A^{\oplus\epsilon})\\&\geq\mu(A^{\ominus\epsilon})-\nu(A^{\oplus\epsilon})\\&=1.\end{aligned}$$

Combining the above inequality with the fact that $D_\epsilon(\mu,\nu)\leq 1$, we get $D_\epsilon(\mu,\nu)=1$.

Case 2: $m_2-m_1<\delta_2-\delta_1-2\epsilon$.

In this case,

$$\begin{aligned}|(m_2+\epsilon)-(m_1-\epsilon)|&=|(m_2-m_1)+2\epsilon|\\&=(m_2-m_1)+2\epsilon<\delta_2-\delta_1,\\|(m_2-\epsilon)-(m_1+\epsilon)|&=|(m_2-m_1)-2\epsilon|\\&\leq|m_2-m_1|+2\epsilon\\&<\delta_2-\delta_1.\end{aligned}$$

Hence, by Lemma 9, the equations $f(x+\epsilon)=g(x-\epsilon)$ and $f(x-\epsilon)=g(x+\epsilon)$ have exactly two solutions each, on the

supports of $\Delta(m_1-\epsilon,\delta_1)$ and $\Delta(m_1+\epsilon,\delta_1)$ respectively. Hence, l must be the minimum of the two solutions to $f(x+\epsilon)=g(x-\epsilon)$ and r must be the maximum of the two solutions to $f(x-\epsilon)=g(x+\epsilon)$. As in the proof of Theorem 8, we divide the real line into five regions as shown in Table III, where l' is the leftmost point such that $\mu([l+\epsilon,l'])=\nu([l-\epsilon,l'])$ and r' is the rightmost point such that $\mu([r',r-\epsilon])=\nu([r',r+\epsilon])$. Observe that by construction, $f(x)\le g(x+2\epsilon)$ for $x\in[r-\epsilon,m_1+\delta_1]$. Hence by Lemma 6, $D_\epsilon(\mu_{++},\nu_{++})=0$. Similarly, we also get $D_\epsilon(\mu_{--},\nu_{--})=0$.

We will now use Lemma 7 to show that $D_\epsilon(\mu_-,\nu_-)=0$. Let $a=l-\epsilon$, $a'=l+\epsilon$, $b=l'$ and $\tilde{t}=l'-2\epsilon$. Let t be the first coordinate of the intersection point of two line segments, one joining $(a,g(a))$ and $(b,g(b))$, and the other joining $(a',f(a'))$ and $(b,f(b))$. The following three conditions are satisfied by μ_- and ν_-. (1) The support of ν_- is $[a,b]$ and the support of μ_- is $[a',b]=[a+2\epsilon,b]$. (2) $g(x)\ge f(x)$ for $x\in[a,t)$ and $f(x)\ge g(x)$ for $x\in(t,b]$. (3) $g(x)\le f(x+2\epsilon)$ for $x\in[a,\tilde{t})$ and the interval $(\tilde{t},b-2\epsilon]$ is empty because $\tilde{t}=b-2\epsilon$. Hence, $D_\epsilon(\mu_-,\nu_-)=0$. Similarly, $D_\epsilon(\mu_+,\nu_+)=0$.

Finally, $D_\epsilon(\mu_0,\nu_0)=0$. This is because $f(x)\ge g(x)$ for $x\in[l',r']$ where $[l',r']$ is the support of both μ_0 and ν_0 and so an identity map $T(x)=x$ may be used to transport all the mass from ν_0 to μ_0 at zero cost.

Like in the proof of Theorem 5, we can upper bound $D_\epsilon(\mu,\nu)$ as follows.

$$\begin{aligned}D_\epsilon(\mu,\nu)&\le 1-[\nu([l-\epsilon,r+\epsilon])+\mu([m_1-\delta_1,l+\epsilon])\\&\qquad+\mu([r-\epsilon,m_1+\delta_1])\\&=\mu(A^{\ominus\epsilon})-\nu(A^{\oplus\epsilon}).\end{aligned}$$

Since $D_\epsilon(\mu,\nu)=\sup_{B\ closed}\mu(B^{\ominus\epsilon})-\nu(B^{\oplus\epsilon})$, the above inequality turns to an equality.

Case 3: $\delta_2-\delta_1-2\epsilon<m_2-m_1<\delta_2+\delta_1+2\epsilon$.

In this case,

$$\begin{aligned}(m_2-\epsilon)-(m_1+\epsilon)&=(m_2-m_1)-2\epsilon<\delta_2+\delta_1,\\(m_1+\epsilon)-(m_2-\epsilon)&=2\epsilon-(m_2-m_1)<\delta_2+\delta_1.\end{aligned}$$

Hence, $|(m_2-\epsilon)-(m_1+\epsilon)|<\delta_2+\delta_1$. By Lemma 9, the equation $f(x-\epsilon)=g(x+\epsilon)$ has either one or two solutions. Therefore, r must be the rightmost solution to $f(x-\epsilon)=g(x+\epsilon)$.

We will split the analysis into three sub-cases.

Case 3(a): $m_2-m_1>\delta_2-\delta_1+2\epsilon$.

We will decompose μ and ν into two mutually singular positive measures each. Let μ_- and μ_+ be the restriction of μ to the intervals $[m_1-\delta_1,r-\epsilon]$ and $[r-\epsilon,m_1+\delta_1]$ respectively. Let ν_- and ν_+ be the restriction of ν to the intervals $[m_2-\delta_2,r+\epsilon]$ and $[r+\epsilon,m_2+\delta_2]$ respectively. The following inequality shows that the support of ν_- is of a smaller length than that of μ_-.

$$\begin{aligned}&[(r+\epsilon)\text{-}(m_2-\delta_2)]-[(r-\epsilon)\text{-}(m_1-\delta_1)]\\&=\delta_2-\delta_1+2\epsilon-(m_2-m_1)\\&<0.\end{aligned}$$

It follows that the support of ν_+ is of a greater length than that of μ_+. By construction, $g(x-2\epsilon)\le f(x)$ for $x\in[m_2-\delta_2,r+\epsilon]$. Hence, by Lemma 6, $D_\epsilon(\mu_-,\nu_-)=0$. A similar analysis shows that $D_\epsilon(\mu_+,\nu_+)=0$. Hence,

$$\begin{aligned}&D_\epsilon(\mu,\nu)\\&\le 1-\min(\mu_-(\mathbb{R}),\nu_-(\mathbb{R}))-\min(\mu_+(\mathbb{R}),\nu_+(\mathbb{R}))\\&=1-\mu([r-\epsilon,\infty))-\nu([r+\epsilon,\infty))\\&=\mu(A^{\ominus\epsilon})-\nu(A^{\oplus\epsilon}).\end{aligned}$$

Since $D_\epsilon(\mu,\nu)=\sup_{B\ closed}\mu(B^{-\epsilon})-\nu(B^{\epsilon})$, the above inequality turns to an equality.

Case 3(b): $\delta_2-\delta_1<m_2-m_1\le\delta_2-\delta_1+2\epsilon$.

Let μ_-,μ_+,ν_- and ν_+ be as defined in case 3(a). The following inequality shows that the support of μ_+ is smaller than that of ν_+.

$$\begin{aligned}&[(m_2+\delta_2)-(r+\epsilon)]-[(m_1+\delta_1)-(r-\epsilon)]\\&=(m_2-m_1)+\delta_2-\delta_1-2\epsilon>0.\end{aligned}$$

Moreover, $f(x)\le g(x+2\epsilon)$ for $x\in[r-\epsilon,m_1+\delta_1]$. Hence by Lemma 6, $D_\epsilon(\mu_+,\nu_+)=0$.

We will now show that $D_\epsilon(\mu_-,\nu_-)=0$ by verifying the conditions of Lemma 7. Since $2\epsilon<2\delta_1$, we have the following.

$$\begin{aligned}\delta_2-\delta_1&<m_2-m_1\\&\le\delta_2-\delta_1+2\epsilon\\&<\delta_2-\delta_1+2\delta_1\\&=\delta_2+\delta_1.\end{aligned}$$

Hence, by Lemma 9, there is exactly one point of intersection of $f(x)$ and $g(x)$ on the support of μ. Let t be the first coordinate of that point. Let $a=m_2-\delta_2-2\epsilon$, $a'=a+2\epsilon$ and $b=r+\epsilon$. Then, (1) the support of μ_- is $[m_1-\delta_1,r-\epsilon]$ which is a subset of $[a,b]$, and the support of ν_- is $[a',b]$. (2) $f(x)\ge g(x)$ for $x\in(a,t]$ and $f(x)\le g(x)$ for $x\in(t,b]$. Hence, the first two conditions of Lemma 7 are verified. To verify, the third condition, we note the following.

$$\begin{aligned}(m_2-2\epsilon)-m_1&=m_2-m_1-2\epsilon<\delta_2-\delta_1,\\m_1-(m_2-2\epsilon)&=m_1-m_2+2\epsilon<2\epsilon<\delta_2-\delta_1.\end{aligned}$$

Hence, by Lemma 9, $f(x)-g(x+2\epsilon)=0$ exactly twice on the support of μ. The greater of the two will be $r-\epsilon$. Let $\tilde{t}$ be the lesser of the two. Then, $\tilde{t}<r-\epsilon=b-2\epsilon$. Further, $f(x)\le g(x+2\epsilon)$ for $x\in[a,\tilde{t})$ and $f(x)\ge g(x+2\epsilon)$ for $x\in(\tilde{t},b-2\epsilon]$. Hence, $D_\epsilon(\mu_-,\nu_-)=0$ by Lemma 7. Therefore, the optimal risk and optimal classifier remain the same as in case 3(a).

Case 3(c): $m_2-m_1\le\delta_2-\delta_1$.

We will partition the real line into four regions as shown in Table IV, where l' is the leftmost point such that $\mu([m_1-\delta_1,l'])=\nu([m_2-\delta_2,l'])$ and r' is as defined in case 2. Since μ_+,ν_+,μ_{++} and ν_{++} are defined in an identical manner to case 2, we get $D_\epsilon(\mu_+,\nu_+)=D_\epsilon(\mu_{++},\nu_{++})=0$.

We will now show $D_\epsilon(\mu_{--},\nu_{--})=0$ using Lemma 7. Let $a=m_1-\delta_1-2\epsilon$, $a'=a+2\epsilon$, $b=l'$ and $\tilde{t}=b-2\epsilon$. Since $m_2-m_1\le\delta_2-\delta_1$, by Lemma 9, $f(x)-g(x)=0$ has exactly two solutions. Let t be the lesser of the two. Then, (1) the support of ν_{--} is $[m_2-\delta_2,b]$ which is a subset of $[a,b]$ and

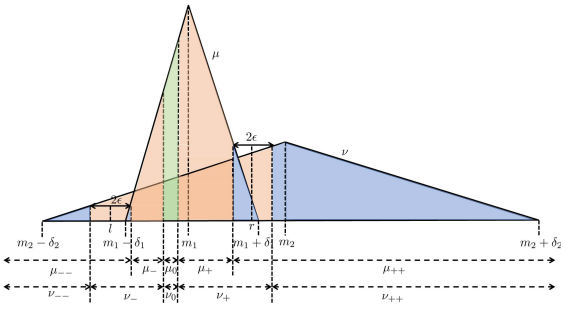


Fig. 9. Optimal transport coupling for triangular distributions μ and ν. As in the proof of Theorem 8, we divide the real line into five regions. The transport plan from μ to ν consists of five maps transporting $\mu_{--} \to \nu_{--}$ (blue regions to the left), $\mu_- \to \nu_-$ (orange regions to the left), $\mu_0 \to \nu_0$ (green regions in the middle), $\mu_+ \to \nu_+$ (orange regions to the right), and $\mu_{++} \to \nu_{++}$ (blue regions to the right).

TABLE III

THE REAL LINE IS PARTITIONED INTO FIVE REGIONS FOR μ AND ν FOR CASE 2

μ_{--}	$(m_1-\delta_1, l+\epsilon]$	ν_{--}	$(m_2-\delta_2, l-\epsilon]$
μ_-	$(l+\epsilon, l']$	ν_-	$(l-\epsilon, l']$
μ_0	(l', r')	ν_0	(l', r')
μ_+	$[r', r-\epsilon)$	ν_+	$[r', r+\epsilon)$
μ_{++}	$[r-\epsilon, m_1+\delta_1)$	ν_{++}	$[r+\epsilon, m_2+\delta_2)$

TABLE IV

THE REAL LINE IS PARTITIONED INTO FOUR REGIONS FOR μ AND ν FOR CASE 3(C)

μ_{--}	$(m_1-\delta_1, l']$	ν_{--}	$(m_2-\delta_2, l']$
μ_-	(l', r')	ν_-	(l', r')
μ_+	$[r', r-\epsilon)$	ν_+	$[r', r+\epsilon)$
μ_{++}	$[r-\epsilon, m_1+\delta_1)$	ν_{++}	$[r+\epsilon, m_2+\delta_2)$

the support of μ_{--} is $[a', b]$. (2) $g(x) \geq f(x)$ for $x \in [a, t)$ and $f(x) \geq g(x)$ for $x \in (t, b]$. (3) $g(x) \leq f(x+2\epsilon)$ for $x \in [a, \tilde{t})$ and the interval $(\tilde{t}, b-2\epsilon]$ is empty because $\tilde{t} = b-2\epsilon$. Hence, $D_\epsilon(\mu_-, \nu_-) = 0$.

Finally, $D_\epsilon(\mu_-, \nu_-) = 0$ because $f(x) \geq g(x)$ for $x \in [l', r']$ and the identity map $T(x) = x$ transports all the mass from ν_- to μ_- at zero cost.

Overall, we have the following inequality.

$$\begin{aligned} &D_\epsilon(\mu, \nu) \\ &\leq 1 - [\nu([m_2-\delta_2, l']) + \nu([l', r']) + \nu([r', r+\epsilon]) \\ &\qquad + \mu([r-\epsilon, m_1+\delta_1])] \\ &= \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon}). \end{aligned}$$

As in Case 2, we conclude that $D_\epsilon(\mu, \nu) = \mu(A^{\ominus\epsilon}) - \nu(A^{\oplus\epsilon})$.

ACKNOWLEDGMENT

The authors would like to thank the AE and anonymous reviewers for their critical feedback that has led to a much improved manuscript.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[4] D. Silver *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, Dec. 2018.

[5] O. Vinyals *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[6] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[7] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[8] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [review article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186.

[10] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10.

[11] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 274–283.

[12] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Nov. 2017, pp. 3–14.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–11.

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–28.

[15] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[16] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 854–863.

[17] A. Shafahi, W. R. Huang, S. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–17.

[18] S. Mahloujifar, D. I. Diochnos, and M. Mahmoody, "The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure," in *Proc. 33rd Conf. Artif. Intell. (AAAI)*, 2019, pp. 4536–4543.

[19] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 125–136.

[20] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–24.

[21] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao, "Is robustness the cost of accuracy?—A comprehensive study on the robustness of 18 deep image classification models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 631–648.

[22] J. Khim and P.-L. Loh, "Adversarial risk bounds via function transformation," 2018, *arXiv:1810.09519*. [Online]. Available: http://arxiv.org/abs/1810.09519

[23] D. Yin, K. Ramchandran, and P. Bartlett, "Rademacher complexity for adversarially robust generalization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 7085–7094.

[24] T.-W. Weng *et al.*, "Evaluating the robustness of neural networks: An extreme value theory approach," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–18.

[25] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, "Efficient neural network robustness certification with general activation functions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 31, 2018, pp. 4939–4948.

[26] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 2266–2276.

[27] Y.-Y. Yang, C. Rashtchian, Y. Wang, and K. Chaudhuri, "Robustness for non-parametric classification: A generic attack and defense," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 941–951.

[28] R. Bhattacharjee and K. Chaudhuri, "When are non-parametric methods robust?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 832–841.

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, "Robustness via curvature regularization, and vice versa," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9078–9086.

[30] J. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1310–1320.

[31] Y.-Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri, "A closer look at accuracy vs. robustness," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 33, 2020, pp. 1–22.

[32] A. N. Bhagoji, D. Cullina, and P. Mittal, "Lower bounds on adversarial robustness from optimal transport," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 7498–7510.

[33] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, Sep. 2018.

[34] D. Wozabal, "Robustifying convex risk measures for linear portfolios: A nonparametric approach," *Oper. Res.*, vol. 62, no. 6, pp. 1302–1315, Dec. 2014.

[35] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *J. Appl. Probab.*, vol. 56, no. 3, pp. 830–857, Sep. 2019.

[36] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," in *Proc. INFORMS Annu. Meeting*, 2016, pp. 1–43.

[37] R. Gao, X. Chen, and A. J. Kleywegt, "Wasserstein distributional robustness and regularization in statistical learning," in *Proc. INFORMS Annu. Meeting*, 2017, pp. 1–27.

[38] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Math. Program.*, vol. 171, nos. 1–2, pp. 115–166, Sep. 2018.

[39] B. Zhu, J. Jiao, and J. Steinhardt, "Generalized resilience and robust statistics," 2019, *arXiv:1909.08755*. [Online]. Available: http://arxiv.org/abs/1909.08755

[40] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," *Michigan Math. J.*, vol. 31, no. 2, pp. 231–240, 1984.

[41] P. Billingsley, *Convergence of Probability Measures*. Hoboken, NJ, USA: Wiley, 2013.

[42] D. Leao, Jr., M. Fragoso, and P. Ruffino, "Regular conditional probability, disintegration of probability and radon spaces," *Proyecciones*, vol. 23, no. 1, pp. 15–29, 2004.

[43] P. Gourdeau, V. Kanade, M. Kwiatkowska, and J. Worrell, "On the hardness of robust classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 7446–7455.

[44] D. I. Diochnos, S. Mahloujifar, and M. Mahmoody, "Adversarial risk and robustness: General definitions and implications for the uniform distribution," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 31, 2018, pp. 10359–10368.

[45] Z. Tu, J. Zhang, and D. Tao, "Theoretical analysis of adversarial learning: A minimax approach," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 12280–12290.

[46] M. Staib and S. Jegelka, "Distributionally robust deep learning as a generalization of adversarial training," in *Proc. NIPS Workshop Mach. Learn. Comput. Secur.*, 2017. [Online]. Available: https://machine-learning-and-security.github.io/papers/mlsec17_paper_30.pdf

[47] R. Pinot, R. Ettedgui, G. Rizk, Y. Chevaleyre, and J. Atif, "Randomization matters. How to defend against strong adversarial attacks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 7717–7727.

[48] G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *J. Mach. Learn. Res.*, vol. 3, pp. 555–582, Dec. 2002.

[49] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *J. Mach. Learn. Res.*, vol. 7, pp. 1283–1314, Jul. 2006.

[50] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *J. Mach. Learn. Res.*, vol. 10, no. 7, pp. 1–26, 2009.

[51] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn, "Distributionally robust logistic regression," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 28, 2015, pp. 1576–1584.

[52] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 517–564, 2018.

[53] H. Bao, C. Scott, and M. Sugiyama, "Calibrated surrogate losses for adversarially robust classification," in *Proc. 33rd Conf. Learn. Theory*, vol. 125, 2020, pp. 408–451.

[54] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Math. Oper. Res.*, vol. 44, no. 2, pp. 565–600, May 2019.

[55] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," *Oper. Res.*, vol. 58, no. 4, pp. 902–917, Aug. 2010.

[56] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Oper. Res.*, vol. 62, no. 6, pp. 1358–1376, 2014.

[57] D. Bertsimas, S. Shtern, and B. Sturt, "A data-driven approach to multi-stage stochastic linear optimization," Tech. Rep., 2020. [Online]. Available: http://www.optimization-online.org/DB_FILE/2018/11/6907.pdf

[58] J. Lee and M. Raginsky, "Minimax statistical learning with Wasserstein distances," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 31, 2018, pp. 2687–2696.

[59] S. Mazuelas, A. Zanoni, and A. Pérez, "Minimax classification with 0–1 loss and performance guarantees," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 33, 2020, pp. 302–312.

[60] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, pp. 1753–1758, Dec. 1965.

[61] P. J. Huber and V. Strassen, "Minimax tests and the Neyman–Pearson lemma for capacities," *Ann. Statist.*, vol. 1, no. 2, pp. 251–263, Mar. 1973.

[62] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 413–421, Jan. 2009.

[63] G. Gul and A. M. Zoubir, "Robust hypothesis testing for modeling errors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 5514–5518.

[64] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5572–5587, Sep. 2017.

[65] P. J. Huber, *Robust Statistics*, vol. 523. Hoboken, NJ, USA: Wiley, 2004.

[66] L. LeCam, "Convergence of estimates under dimensionality restrictions," *Ann. Statist.*, vol. 1, no. 1, pp. 38–53, Jan. 1973.

[67] Y. Baraud, L. Birgé, and M. Sart, "A new method for estimation and model selection: ρ-estimation," *Inventiones Mathematicae*, vol. 207, no. 2, pp. 425–517, 2017.

[68] Y. Baraud and L. Birgé, "Rho-estimators revisited: General theory and applications," *Ann. Statist.*, vol. 46, no. 6B, pp. 3767–3804, Dec. 2018.

[69] M. Ledoux, *The Concentration of Measure Phenomenon*, no. 89. Providence, RI, USA: American Mathematical Society, 2001.

[70] J. Gilmer *et al.*, "Adversarial spheres," in *Proc. Int. Conf. Learn. Represent. (ICLR)-Workshop Track*, 2018, pp. 1–15.

[71] E. Dohmatob, "Limitations of adversarial robustness: Strong no free lunch theorem," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1646–1654.

[72] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 31, 2018, pp. 1178–1187.

[73] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, vol. 55, nos. 58–63. Basel, Switzerland: Birkäuser, 2015, p. 94.

[74] C. Villani, *Topics in Optimal Transportation*. Providence, RI, USA: American Mathematical Society, 2003.

[75] A. Tovar-Lopez and V. Jog, "Generalization error bounds using Wasserstein distances," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[76] G. Peyré and M. Cuturi, "Computational optimal transport," *Found. Trends Mach. Learn.*, vol. 11, nos. 5–6, pp. 355–607, 2019.
[77] L. Holmstrom and P. Koistinen, "Using additive noise in back-propagation training," *IEEE Trans. Neural Netw.*, vol. 3, no. 1, pp. 24–38, Jan. 1992.
[78] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[79] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, "Scaling algorithms for unbalanced optimal transport problems," *Math. Comput.*, vol. 87, no. 314, pp. 2563–2609, Feb. 2018.
[80] B. Pass, "Multi-marginal optimal transport: Theory and applications," *ESAIM, Math. Model. Numer. Anal.*, vol. 49, no. 6, pp. 1771–1790, 2015.
[81] V. Jog, "Reverse Euclidean and Gaussian isoperimetric inequalities for parallel sets with applications," 2020, *arXiv:2006.09568*. [Online]. Available: http://arxiv.org/abs/2006.09568

Muni Sreenivas Pydi received the B.Tech. degree in electrical engineering from IIT Madras in 2014 and the M.S. degree in electrical engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, under the supervision of Prof. Varun Jog.

Varun Jog received the B.Tech. degree in electrical engineering from IIT Bombay in 2010 and the Ph.D. degree in electrical engineering and computer sciences (EECS) from UC Berkeley in 2015.

He was an Assistant Professor with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, from 2016 to 2020. Since January 2021, he has been a Lecturer with the Department of Pure Mathematics and Mathematical Statistics (DPMMS), University of Cambridge. His research interests include information theory, machine learning, and statistics. He was a recipient of the NSF-CAREER Award in 2020, the Eli Jury Award from the Department of Electrical Engineering and Computer Sciences, UC Berkeley, in 2015, and the Jack Keil Wolf Student Paper Award from ISIT in 2015.