Adversarial Risk via Optimal Transport and Optimal Couplings

Muni Sreenivas Pydi [1] **Varun Jog** [1]

Abstract

The accuracy of modern machine learning algorithms deteriorates severely on adversarially manipulated test data. *Optimal adversarial risk* quantifies the best error rate of any classifier in the presence of adversaries, and *optimal adversarial classifiers* are sought that minimize adversarial risk. In this paper, we investigate the optimal adversarial risk and optimal adversarial classifiers from an optimal transport perspective. We present a new and simple approach to show that the optimal adversarial risk for binary classification with $0-1$ loss function is completely characterized by an optimal transport cost between the probability distributions of the two classes. We propose a novel coupling strategy that achieves the optimal transport cost for several univariate distributions like Gaussian, uniform, and triangular. Using the optimal couplings, we obtain the optimal adversarial classifiers in these settings and show how they differ from optimal classifiers in the absence of adversaries. Based on our analysis, we evaluate algorithm-independent fundamental limits on adversarial risk for CIFAR-10, MNIST, Fashion-MNIST and SVHN datasets, and Gaussian mixtures based on them.

1. Introduction

Modern machine learning algorithms based on deep learning have had tremendous success in recent times, producing state-of-the-art results in many domains such as image classification, game playing, speech and natural language processing. Along with the success, it was also discovered that these algorithms exhibit surprising vulnerability to adversarial perturbations that are imperceptible to humans. Since the discovery in (Szegedy et al., 2013), there has been a slew of *adversarial attacks* on neural network based classifiers (Athalye et al., 2018; Carlini & Wagner, 2017; Goodfellow et al., 2014) and defense methods against such attacks (Madry et al., 2018; Papernot et al., 2016; Cisse et al., 2017). Often, the defense methods either fall short of new attacks or are computationally intractable for large neural networks. Recent work has focused on certifiable defenses that are provably robust against a pre-specified class of adversaries (Cohen et al., 2019; Sinha et al., 2017; Raghunathan et al., 2018; Diochnos et al., 2018). Existence of adversarial examples has been attributed to various causes: concentration phenomena in high-dimensional spaces (A. et al., 2019; Mahloujifar et al., 2019; Gilmer et al., 2018; Fawzi et al., 2018); linearity of decision boundaries (Goodfellow et al., 2014); and reliance on non-robust features (Tsipras et al., 2019; Ilyas et al., 2019).

In this paper, we take a step back from deep learning-focused adversarial machine learning. We consider the simplest setting: binary classification with the $0-1$ loss and known data distributions. Our goal is to investigate the notions of adversarial risk and robustness in this setting. In particular, we are interested in analyzing algorithm-independent (information-theoretic) limits on optimal adversarial risk. The first question we investigate is the following:

Question 1. How much can the optimal adversarial risk differ from optimal standard risk?

It is easy to see that the optimal adversarial risk is at least as large as optimal standard risk (see Section 2). Is it possible to derive a tighter lower bound for the optimal adversarial risk? Recent works derive 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. For instance, (Khim & Loh, 2018) and (Yin et al., 2019) develop risk bounds based on a notion of adversarial Rademacher complexity, that is a function of both the data generating distribution and class of classifiers under consideration. In a similar vein, several works (Attias et al., 2018; Cullina et al., 2018) derive sample-complexity bounds for robust learning. More recent works (Diochnos et al., 2019; Gourdeau et al., 2019) specifically focus on lower bounds for sample-complexity, in order to characterize the hardness of robust learning. However, deriving lower bounds on the optimal adversarial risk that are classifier agnostic has not received much attention. Another related question is the fol-

[1]Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI, USA. Correspondence to: Muni Sreenivas Pydi <pydi@wisc.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

lowing: *How much adversarial perturbation is sufficient to make the optimal adversarial risk significantly greater than the optimal standard risk?* Relevant works in this direction have again focused on developing robustness metrics that are specific to the classifier (Weng et al., 2018; Zhang et al., 2018; Hein & Andriushchenko, 2017).

In addition to Question 1, one might also consider the nature of the optimal classifier under the standard and adversarial settings. This motivates the following question:

Question 2. Does the optimal classifier in the adversarial setting differ from that in the standard setting?

(Moosavi-Dezfooli et al., 2019) empirically observed that adversarial training significantly reduces the curvature of the loss function with respect to the input. Another line of work attempts to construct a provably robust classifier from a baseline classifier using randomized smoothing (Cohen et al., 2019). A more recent line of work pursues data-preprocessing strategies that make a classifier more robust (Yang et al., 2020; Bhattacharjee & Chaudhuri, 2020). These works suggest ways in which the optimal adversarial classifier differs from the optimal standard classifier. Even so, many other interesting questions remain. For instance, is the optimal classifier without an adversary approximately the same as the optimal classifier with a small adversary (i.e., small ϵ)? If decision boundaries change, do they change smoothly with increasing strength of an adversary, or do they change drastically?

The closest work to ours is Bhagoji et al. (2019), who develop the first classifier-agnostic lower bounds for learning in the presence of an adversary. Specifically, they present a similar result to our Theorem 2 which gives the optimal adversarial risk 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.

Our Contributions

In this paper, we consider two types of adversaries: (i) data perturbing and (ii) distribution perturbing. We focus on the binary classification setting under $0-1$ loss function. We answer Question 1 by providing *universal* bounds for adversarial risk under the two notions of adversaries that are agnostic to the classifier. We answer Question 2 by deriving the optimal adversarial classifier in some special settings. Our contributions are listed below.

1. We provide a new and simple proof for the characterization of optimal adversarial risk for $0-1$ loss functions in terms of an optimal transport cost between the two data generating distributions, where the transport cost is given by $c_\epsilon(x, x') = \mathbb{1}\{d(x, x') > 2\epsilon\}$ and ϵ is the perturbation budget of the adversary. This completely answers Question 1 for this setting. Our proof establishes connections between adversarial machine learning and well-known results in the theory of optimal transport.

2. We propose a novel coupling strategy that achieves the proposed optimal transport cost between the two class-conditional densities for several univariate distributions like Gaussian, triangular and uniform. Using the analysis of optimal couplings, we obtain the optimal adversarial classifiers for these settings. This answers Question 2 in these settings, and shows how the decision boundary of the optimal classifier changes in the presence of adversaries. In certain cases, we show that the decision boundary can change arbitrarily, even for small changes in the adversary budget ϵ.

3. Using our analysis for $0-1$ loss, we obtain the exact optimal risk attainable for a range of adversarial budgets under ℓ_2-norm and ℓ_∞-norm perturbation of data, for several real-world datasets, namely CIFAR10, MNIST, Fashion-MNIST, and SVHN. In addition, we analyze Gaussian mixtures based on these datasets and compute lower bounds on the optimal adversarial risk for them. These bounds indicate the optimal adversarial error achievable with data augmentation via Gaussian perturbations.

Structure: The rest of the paper is structured as follows: In Section 2, we introduce the two types of adversaries: (1) data perturbing and (2) distribution perturbing, and show that the optimal risk in the data perturbing case is lower. We also present a simple example to show that the optimal risk and optimal classifier can deviate significantly in the presence of adversaries. In Section 3, we deal with Question 1 by introducing the D_ϵ distance that completely characterizes the optimal adversarial risk. In Section 4, we deal with Question 2 by presenting the coupling strategy used to obtain optimal adversarial classifiers in special cases. Finally, in Section 5, we present adversarial risk lower bounds for real world datasets and evaluate our bounds for $0-1$ loss.

Notation: The complement of a set A is denoted by A^c. Define $\mathbb{1}\{C\}$ to be the indicator function that maps all the inputs satisfying condition C to 1 and the rest to 0. For a set A in a Polish space $(\mathcal{X}, d)$, the set A^ϵ denotes the ϵ-expansion of A in $\mathcal{X}$. That is, $A^\epsilon = \{x \in \mathcal{X} : d(x, x') \leq \epsilon \text{ for some } x' \in A\}$. We define $A^{-\epsilon} := ((A^c)^\epsilon)^c$. For any two probability measures μ and ν defined over the Polish space $(\mathcal{X}, d)$, we use $\Pi(\mu, \nu)$ to denote the set of all joint probability measures over $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν, respectively. We use $D_{TV}(\mu, \nu)$ and $W_p(\mu, \nu)$ to denote the total variation distance and p-Wasserstein distance between μ and ν. We use $\|\cdot\|$ to denote a norm and $\|\cdot\|_*$ to denote its

dual norm. The cumulative distribution function (cdf) of the standard normal distribution is denoted by Φ. We use the shorthand w.l.o.g. for 'without loss of generality'.

2. Preliminaries

2.1. Types of Adversaries

Consider a binary classification setting with the $0-1$ loss function. Consider a metric space $(\mathcal{X}, d)$. Let $B_\epsilon(x)$ denote the open ball of radius ϵ around x. Let the output label, $\mathcal{Y} = \{0, 1\}$, 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). Consider a set of binary classifiers of the form $\mathbb{1}\{x \in A\}$, where $A \subseteq \mathcal{X}$. That is, the classifier corresponding to the set A assigns the label 1 for $x \in A$ and the label 0 for $x \notin A$. The $0-1$ loss is given by $\ell((x, y), A) = \mathbb{1}\{x \in A, y = 0\} + \mathbb{1}\{x \notin A, y = 1\}$. For technical reasons that will be made clear in Section 3, we assume that the sets A that define the classifiers are closed.

We shall consider two such notions of adversarial risk that have appeared in the literature: (i) adversary perturbs data points, and (ii) adversary perturbs data distributions.

Data perturbing adversary: Define two functions $T_0, T_1 : \mathcal{X} \to \mathcal{X}$. Given $(x, y) \in \mathcal{X} \times \{0, 1\}$ and a classifier corresponding to set A, the data perturbing adversary of budget ϵ transports x to $T_y(x) \in \operatorname{argmax}_{x' \in B_\epsilon(x)} \ell((x', y), A)$. Hence, $T_0(x) \in \operatorname{argmax}_{x' \in B_\epsilon(x)} \mathbb{1}\{x' \in A\}$ and $T_1(x) \in \operatorname{argmax}_{x' \in B_\epsilon(x)} \mathbb{1}\{x' \notin A\}$. Then the adversarial risk corresponding to A is given by, $R_\epsilon(A) = \frac{1}{2}\left[p_0(A^\epsilon) + p_1\left((A^c)^\epsilon\right)\right]$. The optimal adversarial risk (or the optimal robust risk) is given by,

$$R^*_\epsilon = \inf_{A \subseteq \mathcal{X}} R_\epsilon(A) = \frac{1}{2}\left[1 - \sup_A \left\{p_1\left(A^{-\epsilon}\right) - p_0(A^\epsilon)\right\}\right].$$

Note that for $\epsilon = 0$, $R^*_0 = \frac{1}{2}[1 - \sup_A(p_1(A) - p_0(A))] = \frac{1}{2}[1 - D_{TV}(p_0, p_1)]$, which is the Bayes risk. Moreover, $p_1(A) - p_0(A) \geq p_1(A^{-\epsilon}) - p_0(A^\epsilon)$. Hence, we get the following trivial lower bound on adversarial risk: $R^*_\epsilon \geq R^*_0$.

Distribution perturbing adversary: Given a classifier corresponding to A, the distribution perturbing adversary perturbs the distributions p_0, p_1 to $\tilde{p}_0 \in B_\epsilon(p_0), \tilde{p}_1 \in B_\epsilon(p_1)$, where $B_\epsilon(.)$ denotes 1-Wasserstein ball around a distribution. Then the adversarial risk is given by $\widehat{R}_\epsilon(A) = \sup_{\substack{\tilde{p}_0 \in B_\epsilon(p_0) \\ \tilde{p}_1 \in B_\epsilon(p_1)}} \frac{1}{2}[\tilde{p}_0(A) + \tilde{p}_1(A^c)]$. The optimal robust risk is given by $\inf_{A \subseteq \mathcal{X}} \widehat{R}_\epsilon(A)$ as follows.

$$\widehat{R}^*_\epsilon = \frac{1}{2}\left[1 - \sup_{A \subseteq \mathcal{X}} \inf_{\substack{\tilde{p}_0 \in B_\epsilon(p_0) \\ \tilde{p}_1 \in B_\epsilon(p_1)}} \{\tilde{p}_1(A) - \tilde{p}_0(A)\}\right]. \quad (1)$$

In the following theorem, we will show that a distribution perturbing adversary is stronger than a data perturbing adversary on the same budget.

Theorem 1. *$R_\epsilon(A) \leq \widehat{R}_\epsilon(A)$ for all $A \subseteq \mathcal{X}$. Moreover, $R^*_\epsilon \leq \widehat{R}^*_\epsilon$.*

Proof. Consider a classifier corresponding to the set A. We observe that the transport map T_0 used by the data perturbing adversary satisfies $d(x, T(x)) < \epsilon$ for all $x \in \mathcal{X}$. Hence the push-forward measure $T_{0\sharp}p_0 \in B_\epsilon(p_0)$. Similarly, $T_{1\sharp}p_1 \in B_\epsilon(p_1)$. Hence,

$$\begin{aligned} R_\epsilon(A) &= \frac{1}{2}[T_{0\sharp}p_0(A^\epsilon) + T_{1\sharp}p_1((A^c)^\epsilon)\}] \\ &\leq \sup_{\substack{\tilde{p}_0 \in B_\epsilon(p_0) \\ \tilde{p}_1 \in B_\epsilon(p_1)}} \frac{1}{2}[\tilde{p}_0(A) + \tilde{p}_1(A^c)] = \widehat{R}_\epsilon(A). \end{aligned}$$

Taking infimum over A, we get $R^*_\epsilon \leq \widehat{R}^*_\epsilon$. □

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.

2.2. A Motivating Example

Here, we present a simple binary classification problem with Gaussian class conditional densities which shows that the adversarially optimal classifier indeed differs from the Bayes optimal classifier. We explicitly compute the optimal adversarial risk and the optimal adversarial classifier for a data perturbing adversary as a function of ϵ. The detailed proof of this fact is found in Theorem 7.

Let $x|(y = i) \sim \mathcal{N}(0, \sigma_i^2)$ for $i = 0, 1$ ($\sigma_1 < \sigma_0$). Let the class labels 0 and 1 be equally likely. Consider classifiers parametrized by $\mathcal{W} = \{w \in \mathbb{R} : w > 0\}$ as follows: $f(x) = \mathbb{1}\{x \in [-w, w]\}$. Let $k = \left(\frac{\sigma_0^2 + \sigma_1^2}{\sigma_0^2 - \sigma_1^2}\right)$. Then the optimal adversarial risk and classifier are given by

$$w^*_\epsilon = w^*_0\sqrt{1 + \frac{\epsilon^2(k^2 - 1)}{(w^*_0)^2}} + \epsilon k, \quad (2)$$

$$R^*_\epsilon = \frac{1}{2}\left[1 - 2\left(\Phi\left(\frac{w^*_\epsilon - \epsilon}{\sigma_1}\right) - \Phi\left(\frac{w^*_\epsilon + \epsilon}{\sigma_0}\right)\right)\right]. \quad (3)$$

Equation (3) shows that the optimal adversarial risk increases with increasing power of the adversary (i.e., increasing ϵ). Moreover, we see from (2) the optimal adversarial classifier can differ significantly from the Bayes optimal

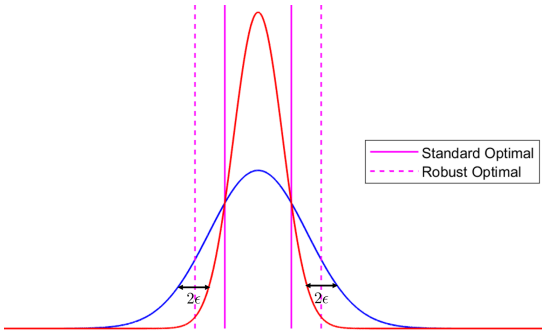


Figure 1. Optimal classifier in the standard setting and (data perturbing) adversarial setting for two centered Gaussian distributions. Here, $\sigma_0 = 1$, $\sigma_1 = 0.5$, and $\epsilon = 0.3$. The optimal adversarial boundary bisects the line segment of length 2ϵ that matches ϕ_0 and ϕ_1.

classifier when σ_0 is close to σ_1 (i.e., when k is large). Figure 1 shows a specific instance of this example.

Evaluating $\widehat{R}^*_\epsilon$ in this case is more challenging, as it involves optimizing over distributions in Wasserstein balls. Suppose we take $B_\epsilon(\rho)$ to be the 2-Wasserstein ball around the distribution ρ. We can exploit the fact that Gaussian measures on $\mathbb{R}$ form a Riemannian manifold with 2-Wasserstein distance as the Riemannian metric (Takatsu, 2011). Moreover, $\{\mathcal{N}(0, \sigma_t^2)\}_{t\in[0,1]}$ is a geodesic from $\mathcal{N}(0, \sigma_0^2)$ to $\mathcal{N}(0, \sigma_1^2)$, where $\sigma_t = (1-t)\sigma_0 + t\sigma_1$. Hence, a suitable choice for the adversarially perturbed distributions $\tilde{p}_0$ and $\tilde{p}_1$ in (1) is to pick $\mathcal{N}(0, (\sigma_0 - \epsilon)^2)$ and $\mathcal{N}(0, (\sigma_1 - \epsilon)^2)$ as the perturbed versions of $\mathcal{N}(0, \sigma_0^2)$ and $\mathcal{N}(0, \sigma_1^2)$ respectively. Fixing this choice of $\tilde{p}_0$ and $\tilde{p}_1$ and maximizing over A in (1), we get the following lower bound on $\widehat{R}^*_\epsilon$.

$$\begin{aligned}\widehat{R}^*_\epsilon &\geq \frac{1}{2}\left[1 - D_{TV}(\mathcal{N}(0, (\sigma_0 - \epsilon)^2), \mathcal{N}(0, (\sigma_1 - \epsilon)^2))\right] \\ &= \frac{1}{2}\left[1 - 2\left(\Phi\left(\frac{\widehat{w}_\epsilon}{\sigma_1 + \epsilon}\right) - \Phi\left(\frac{\widehat{w}_\epsilon}{\sigma_0 - \epsilon}\right)\right)\right], \quad (4)\end{aligned}$$

where $\widehat{w}_\epsilon$ is the positive real number for which the probability densities of $\mathcal{N}(0, \sigma_0^2)$ and $\mathcal{N}(0, \sigma_1^2)$ coincide.

3. Adversarial Risk via Optimal Transport

In this section, we present our results on adversarial risk under $0-1$ loss in the binary classification setting.

Definition 1 (Optimal transport cost)**.** Consider two probability measures μ and ν over a metric space $(\mathcal{X}, d)$. 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

$$D_\epsilon(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,x')\sim\pi} c_\epsilon(x, x'). \quad (5)$$

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. *The adversarial risk with the data perturbing adversary of budget ϵ for the binary classification setting presented in Section 2 is given by*

$$R^*_\epsilon = \frac{1}{2}\left[1 - D_\epsilon(p_0, p_1)\right]. \quad (6)$$

Instantiating Theorem 2 for $\epsilon = 0$, we get $R^*_0 = \frac{1}{2}\left[1 - D_{TV}(p_0, p_1)\right]$, 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 3.1. *Under the setup considered in Theorem 2, we have the following bound for $p \geq 1$:*

$$R^*_\epsilon \geq \frac{1}{2}\left[1 - \left(\frac{W_p(p_0, p_1)}{2\epsilon}\right)^p\right]. \quad (7)$$

Proof sketch of Theorem 2. A key ingredient of our proof is the Strassen's theorem [Corollary 1.28 in (Villani, 2003)], which states that

$$D_\epsilon(p_0, p_1) = \sup_A \left\{p_0(A) - p_1(A^{2\epsilon}) \quad .\right.$$

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

$$\sup_A \mu(A^{-\epsilon}) - \nu(A^\epsilon) = \sup_A \mu(A) - \nu(A^{2\epsilon}). \quad (8)$$

To do so, we make use of the properties of closed sets and their ϵ-expansions. The full proof is included in the supplementary material. □

Comparison with Bhagoji et al.(2019): We note that a similar result was obtained recently in Bhagoji et al.(2019). While the duality in their proof was established for a larger hypothesis class of measurable sets A, our proof relies on Strassen's duality theorem and properties of closed sets. Using closed sets and directly using Strassen's theorem allows us to considerably simplify the technical details as compared with Bhagoji et al.(2019).

4. Adversarial Classifiers via Couplings

Instead of using D_ϵ, we have shown in Corollary 3.1 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. 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 $D_\epsilon(\mu,\nu) \geq \mu(A^{-\epsilon}) - \nu(A^\epsilon)$. (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 univariate random variables, some of our results will also naturally extend to higher dimensional distributions.

4.1. Gaussian Distributions with Identical Variances

Theorem 3. *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 $\mu_0 < \mu_1$ w.l.o.g. Let $A \subseteq \mathbb{R}$ denote the set over which the optimal robust classifier assigns a label 1. If $\epsilon \geq |\mu_0-\mu_1|/2$, then $A = \mathbb{R}$ and $R^*_\epsilon = 1/2$. If $\epsilon < |\mu_0-\mu_1|/2$, then $A = [(\mu_0+\mu_1)/2, \infty)$ and $R^*_\epsilon = p_0(A)$.*

Proof. 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 by moving the mass at each x by 2ϵ. Thus, the optimal transport cost for this coupling is 0, and therefore so is $D_\epsilon(p_0,p_1)$. Hence, $R^*_\epsilon = 1/2$ and the constant classifier achieves it.

For $\epsilon < \frac{\mu_1-\mu_0}{2}$, we consider the coupling shown in Figure 2. Let $\tilde{p}_1$ be the distribution obtained by shifting p_1 to the left by 2ϵ. It is evident that the overlapping area between $\tilde{p}_1$ and p_0 maybe be translated by 2ϵ so that it lies entirely with p_1. This means that the overlapping area may be transported at 0 cost. It is easily verified that the overlapping area contains $2p_0(A)$ mass where $A = [(\mu_0+\mu_1)/2, \infty)$ corresponds to the MLE classifier. So, the total cost of transportation is at most $1 - 2p_0(A)$. Plugging this into Theorem 2, we get $R^*_\epsilon = p_0(A)$, which is the risk achieved by the classifier corresponding to A i.e. the MLE classifier. □

Theorem 3 can be easily extended to d-dimensional Gaussians with the same identity covariances. Our results may be summarized in the following theorem:

Theorem 4. *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 $(\mathbb{R}, ||\cdot||_2)$. If $\epsilon \geq ||\mu_0-\mu_1||_2/2$, then $A = \mathbb{R}$ and $R^*_\epsilon = 1/2$. If $\epsilon < ||\mu_0-\mu_1||_2/2$, then $A = \{x : \|x-\mu_0\| \geq \|x-\mu_1\|\}$ and $R^*_\epsilon = p_0(A)$.*

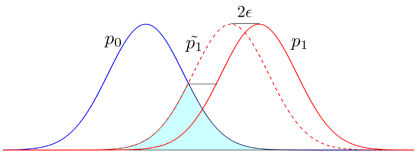


Figure 2. Optimal coupling for two Gaussians with identical variances. The shaded region within p_0 is translated by 2ϵ to p_1, whereas the remaining is mass in p_0 is moved at a cost of 1 per unit mass.

Comparison to Bhagoji et al. (2019): Bhagoji et al. also explore optimal classifiers for multivariate normal distributions. In fact, they show a more general version of our Theorems 3 and 4 by considering data distributions $\mathcal{N}(\mu_0,\Sigma)$ and $\mathcal{N}(\mu_2,\Sigma)$, and an adversary that perturbs within L_p balls. In the following subsections, we shall generalize Theorem 3 in a novel way by considering various interesting examples of univariate distributions and identifying optimal couplings for these.

4.2. 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 partitioning $\mathbb{R}$ into disjoint intervals and defining transport maps between the measures restricted to the intervals. Our first result identifies a necessary and sufficient condition for $D_\epsilon(\mu,\nu) = 0$ for arbitrary measures on $\mathbb{R}^d$.

Theorem 5. *Let μ and ν be finite positive measures on $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure and have bounded supports. Then $D_\epsilon(\mu,\nu) = 0$ if and only if $W_\infty(\mu,\nu) \leq 2\epsilon$. Here, $W_\infty(\mu,\nu) = \lim_{p\to\infty} W_p(\mu,\nu)$.*

Calculating W_∞ is non-trivial in general, but in the univariate case it may be calculated using the monotone transport map, which is known to be optimal.

Theorem 6. *Let μ and ν be finite positive measures on $\mathbb{R}$ that are absolutely continuous with respect to the Lebesgue measure with Radon-Nikodyn derivatives $f(\cdot)$ and $g(\cdot)$, respectively. 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 $F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}$. The cdf $G(\cdot)$ and inverse cdf $G^{-1}(\cdot)$ are defined analogously. Suppose that $\mu(\mathbb{R}) = \nu(\mathbb{R}) = U$. Then $D_\epsilon(\mu,\nu) = 0$ if and only if $\|F^{-1} - G^{-1}\|_\infty \leq 2\epsilon$.*

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_{++}	$[m-\epsilon, \infty)$

Table 1. The real line is partitioned into five regions for μ and ν, as shown in the table.

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 4.1. *Let μ and ν be as in Theorem 6. Suppose that for every $x \in \mathbb{R}$, we have $F(x) \geq G(x)$ and $F(x) \leq G(x + 2\epsilon)$. Then $D_\epsilon(\mu, \nu) = 0$.*

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 $\sigma_1^2 > \sigma_2^2$ w.l.o.g. Let $m > 0$ be such that $f(m+\epsilon) = g(m-\epsilon)$. Then the optimal adversarial classifier decides label 1 on the set $A = (-\infty, -m] \cup [m, +\infty)$. The corresponding optimal adversarial risk is $R_\epsilon^* = (1 - \mu(A^{-\epsilon}) + \nu(A^\epsilon))/2$.*

Proof of Theorem 7. First, we partition $\mathbb{R}$ into the five regions for μ and ν, as shown in Table 1, where $r > 0$ is such that such that $\mu([-m-\epsilon, -r)) = \nu([-m+\epsilon, -r))$. 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 use Corollary 4.1 to show that $D_\epsilon(\mu_*, \nu_*) = 0$, where $*$ ranges over all possible subscripts. Note that these measures do not necessarily have identical masses, and 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 $D_\epsilon(\mu, \nu) = 0$. We upper bound $D_\epsilon(\mu, \nu)$ using this transport plan as follows.

$$\begin{aligned} D_\epsilon(\mu, \nu) &\leq 1 - \sum_{*\in\{--,-,0,+,++\}} \min(\mu_*, \nu_*) \\ &= 1 - \mu([-m-\epsilon, m+\epsilon]) - 2\nu([m-\epsilon, \infty)) \\ &= \mu(A^{-\epsilon}) - \nu(A^\epsilon). \end{aligned}$$

However, we also have $D_\epsilon(\mu, \nu) \geq \mu(A^{-\epsilon}) - \nu(A^\epsilon)$. The lower and upper bounds match and this concludes the proof. R_ϵ^* is given by Theorem 2. The robust risk of the classifier that decides 1 on the set A is easily seen to be R_ϵ^*. □

We extend the above proof strategy to demonstrate the optimal coupling for Gaussians with arbitrary means and arbitrary variances in the following theorem.

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 $\sigma_1^2 > \sigma_2^2$ w.l.o.g. Let $m_1, m_2 > 0$ be such that $f(-m_1 - \epsilon) = g(-m_1 + \epsilon)$ and $f(m_2 + \epsilon) = g(m_2 - \epsilon)$. Then the optimal adversarial classifier decides label 1 on the set $A = (-\infty, -m_1] \cup [m_2, \infty)$. The corresponding optimal adversarial risk is $R_\epsilon^* = (1 - \mu(A^{-\epsilon}) + \nu(A^\epsilon))/2$.*

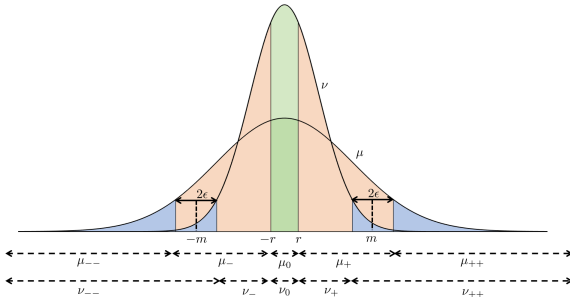


Figure 3. Optimal transport coupling for centered Gaussian distributions μ and ν. The transport plan from μ to ν consists of five maps transporting $\mu_{--} \to \nu_{--}$ (blue regions to the left), $\mu_{-} \to \nu_{-}$ (orange regions to the left), $\mu_0 \to \nu_0$ (green regions in the middle), $\mu_{+} \to \nu_{+}$ (orange regions to the right), and $\mu_{++} \to \nu_{++}$ (blue regions to the right).

4.3. 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. The precise proof details in these cases may be reconstructed from the proofs of Theorems 7 and 8.

Theorem 9 (Uniform distributions)**.** *Let μ and ν be uniform measures on closed intervals I and J respectively. w.l.o.g., we assume $|I| \leq |J|$. Then the optimal robust risk is $\nu(I^{2\epsilon})$ and the optimal classifier is given by $A = I^\epsilon$.*

Proof sketch 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 4 shows the optimal coupling for the case when $I^{2\epsilon} \subseteq J$. □

Theorem 10 (Triangular distributions)**.** *Denote a triangular distribution with support $[m - \delta, \mu + \delta]$ as $\Delta(m, \delta)$. Let μ and ν correspond to the triangular distributions $\Delta(m_1, \delta_1)$ and $\Delta(\mu_2, \delta_2)$ with pdfs f and g respectively. w.l.o.g., assume $\delta_1 < \delta_2$. Let $l < m_1 < r$ be such that $f(l+\epsilon) = g(l-\epsilon)$ and $f(r-\epsilon) = g(r+\epsilon)$. (In case of multiple such points, pick l to be the largest among all such points, and r to smallest.) Then the optimal adversarial classifier decides label 1 on the set $A = [l, r]$. The corresponding optimal adversarial risk is $R_\epsilon^* = (1 - \mu(A^{-\epsilon}) + \nu(A^\epsilon))/2$.*

Proof sketch of Theorem 10. We omit all proof details and

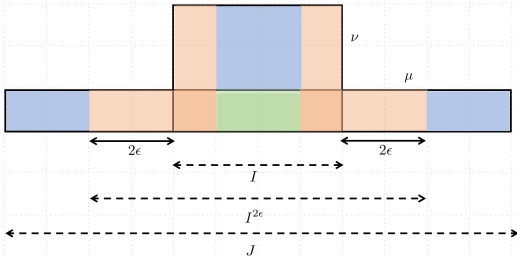


Figure 4. Optimal coupling for two uniform distributions. The region shaded in green is kept in place (at no cost). The two regions shaded in orange are transported monotonically from either side at a cost not exceeding 2ϵ per unit mass. The remaining region in blue is moved at the cost of 1 per unit mass.

point to Figure 5 which shows the coupling in a special case. □

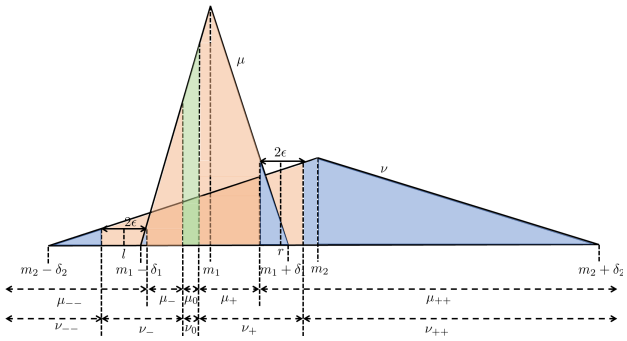
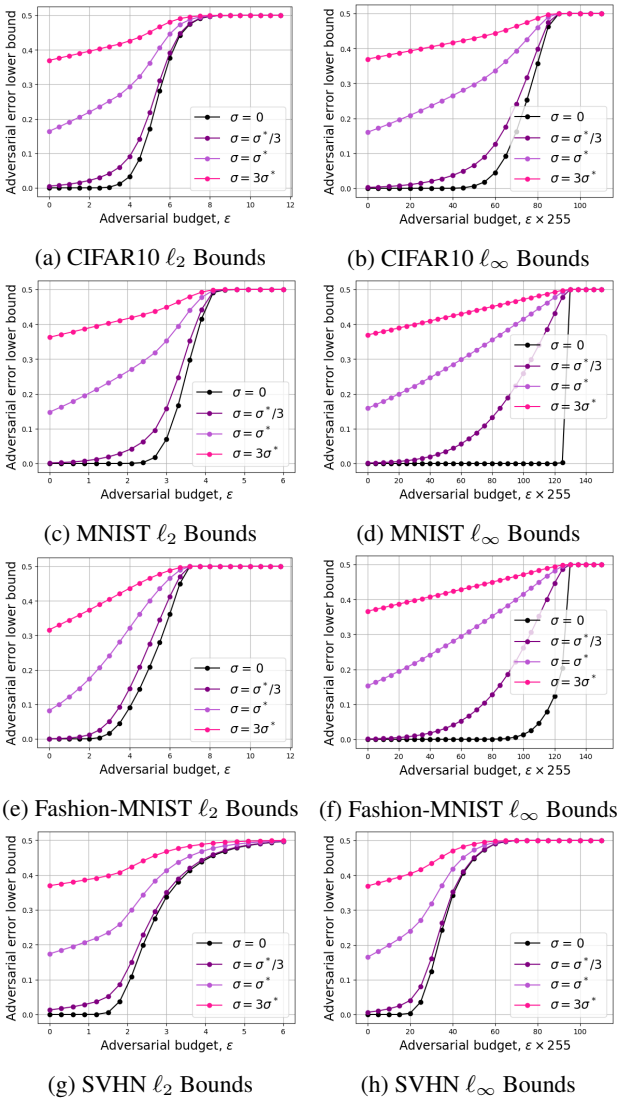


Figure 5. Optimal transport coupling for triangular distributions μ and ν. The transport plan from μ to ν consists of five maps transporting $\mu_{--} \to \nu_{--}$ (blue regions to the left), $\mu_{-} \to \nu_{-}$ (orange regions to the left), $\mu_0 \to \nu_0$ (green regions in the middle), $\mu_{+} \to \nu_{+}$ (orange regions to the right), and $\mu_{++} \to \nu_{++}$ (blue regions to the right).

5. 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 (5) can be formulated as a linear program and solved efficiently. When the number of data points in the two empirical distributions is the same, the problem reduces to an optimal matching problem between the two datasets (see Proposition 2.11 in (Peyré & Cuturi, 2019)). Using 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 Figure 6 with the legend $\sigma = 0$.



(a) CIFAR10 ℓ_2 Bounds (b) CIFAR10 ℓ_∞ Bounds

(c) MNIST ℓ_2 Bounds (d) MNIST ℓ_∞ Bounds

(e) Fashion-MNIST ℓ_2 Bounds (f) Fashion-MNIST ℓ_∞ Bounds

(g) SVHN ℓ_2 Bounds (h) SVHN ℓ_∞ Bounds

Figure 6. Lower bounds on adversarial risk computed using Theorem 2. The curves with $\sigma = 0$ gives the optimal risk for empirical distributions, while the other curves give lower bounds on optimal risk for Gaussian mixtures based on the empirical distributions using the coupling in Theorem 4.

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. (Bhagoji et al., 2019). 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 (Holmstrom & Koistinen, 1992; Goodfellow et al., 2016). 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. The variance $\sigma^2 = 0$ 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. We emphasize that our bounds for Gaussian mixture models on datasets are computed using the exact mixture distribution, rather than drawing samples and evaluating our bounds on the empirical distribution. This exact computation is done using our theoretical results on multi-dimensional Gaussians in Section 4.

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 4 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 Figure 6.

Figure 6 shows 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 Figure 6 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. (Bhagoji et al., 2019) (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 Figure 6 is made possible by our analysis on the optimal coupling to achieve D_ϵ between multivariate Gaussians in Section 4.1.

The bounds in Figure 6 indicate the best error rates achievable on the *training datasets* for CIFAR10, MNIST, Fashion-MNIST and SVHN. The curves for $\sigma > 0$ show the limits on error rate even when trained with Gaussian data augmentation. Moreover, the bounds hold irrespective of the classification algorithm. The code accompanying Figure 6 is made available at `http://github.com/munisreenivas/adv-risk-optimal-transport`.

6. Discussion and Open Problems

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. 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 $\widehat{R}^*_\epsilon$?

In analysing the adversarial risk for $0-1$ loss, 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 optimal transport, we also obtain the optimal robust 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.

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. We note that 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.

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 to see how fast D_ϵ between empirical distributions converges to D_ϵ between the true data-generating distributions. This may be used to obtain finite-sample lower bounds for adversarial error. Recent work (Jog, 2020) in this line of research derives sample complexity bounds for estimating D_ϵ from empirical distributions using a reverse Gaussian isoperimetric inequality for sets of the form A^ϵ. Another recent work (Yu, 2019) implies a sharp threshold for the asymptotics of D_ϵ between product distributions in terms of the 1-Wasserstein metric.

References

A., S., Huang, W. R., Studer, S., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? *International Conference on Learning Representations*, 2019.

Athalye, A., Carlini, N., and A., W. D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. *Algorithmic Learning Theory*, 2018.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. *Conference on Neural Information Processing Systems*, 2019.

Bhattacharjee, R. and Chaudhuri, K. When are non-parametric methods robust? In *International Conference on Machine Learning*, 2020. To appear.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14. ACM, 2017.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, 2017.

Cohen, J., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

Cullina, D., Bhagoji, A. N., and Mittal, P. PAC-learning in the presence of adversaries. In *Conference on Neural Information Processing Systems*, pp. 230–241, 2018.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Conference on Neural Information Processing Systems*, 2018.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. Lower bounds for adversarially robust PAC learning. *arXiv preprint arXiv:1906.05815*, 2019.

Fawzi, A., Fawzi, H., and Fawzi, O. Adversarial vulnerability for any classifier. In *Conference on Neural Information Processing Systems*, 2018.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT Press, 2016.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gourdeau, P., Kanade, V., Kwiatkowska, M., and Worrell, J. On the hardness of robust classification. *Conference on Neural Information Processing Systems*, 2019.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pp. 2266–2276, 2017.

Holmstrom, L. and Koistinen, P. Using additive noise in back-propagation training. *IEEE Transactions on Neural Networks*, 3(1):24–38, 1992.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Conference on Neural Information Processing Systems*, 2019.

Jog, V. Reverse Lebesgue and Gaussian isoperimetric inequalities for parallel sets with applications. *arXiv preprint arXiv:2006.09568*, 2020.

Khim, J. and Loh, P.-L. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *Thirty-Third Conference on Artificial Intelligence (AAAI)*, 2019.

Moosavi-Dezfooli, S.-M., Fawzi, A., Uesato, J., and Frossard, P. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy*, pp. 582–597. IEEE, 2016.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends R in Machine Learning*, 11(5-6): 355–607, 2019.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *International Conference on Learning Representations*, 2018.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Takatsu, A. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 2011.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *International Conference on Learning Representations*, 2019.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. *International Conference on Learning Representations*, 2018.

Yang, Y.-Y., Rashtchian, C., Wang, Y., and Chaudhuri, K. Robustness for non-parametric classification: A generic attack and defense. In *International Conference on Artificial Intelligence and Statistics*, pp. 941–951, 2020.

Yin, D., Ramchandran, K., and Bartlett, P. Rademacher complexity for adversarially robust generalization. *International Conference on Machine Learning*, 2019.

Yu, L. Asymptotics of Strassen's optimal transport problem. *arXiv preprint arXiv:1912.02051*, 2019.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. In *Conference on Neural Information Processing Systems*, 2018.