Adversarial Robustness in Machine Learning: An Optimal Transport Perspective

by
Muni Sreenivas Pydi

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
(Electrical and Computer Engineering)

at the
University of Wisconsin-Madison
2022

Date of Final Oral Exam: August 8, 2022

The dissertation is approved by the following members of the Final Oral Committee:
Varun Jog (Chair), Lecturer, Pure Mathematics & Mathematical Statistics (University of Cambridge)
Robert Nowak, Professor, Electrical & Computer Engineering
Kangwook Lee, Assistant Professor, Electrical & Computer Engineering
Xiaojin (Jerry) Zhu, Professor, Computer Science
Nicolás García Trillos, Assistant Professor, Statistics

Adversarial Robustness in Machine Learning: An Optimal Transport Perspective

Abstract

Deep learning based classification algorithms perform poorly on adversarially perturbed data. Adversarial risk quantifies the performance of a classifier in the presence of an adversary. Numerous definitions of adversarial risk—not all mathematically rigorous and differing subtly in the details—have appeared in the literature. Adversarial attacks are designed to increase the adversarial risk of classifiers, and robust classifiers are sought that can resist such attacks. It was hitherto unknown what the theoretical limits on adversarial risk are, and whether there is an equilibrium in the game between the classifier and the adversary.

In this thesis, we establish a mathematically rigorous foundation for adversarial robustness, derive algorithm-independent bounds on adversarial risk, and provide alternative characterizations based on distributional robustness and game theory. Key to these results are the numerous connections we discover between adversarial robustness and optimal transport theory. We begin by examining various definitions for adversarial risk, and laying down conditions for their measurability and equivalences. In binary classification with 0-1 loss, we show that the optimal adversarial risk is determined by an optimal transport cost between the probability distributions of the two classes. Using the couplings that achieve this cost, we derive the optimal robust classifiers for several univariate distributions. Using our results, we compute lower bounds on adversarial risk for several real-world datasets. We extend our results to general loss functions under convexity and smoothness assumptions.

We close with alternative characterizations for adversarial adversarial robustness that lead to the proof of a pure Nash equilibrium in the two-player game between the adversary and the classifier. We show that adversarial risk is identical to the minimax risk in a robust hypothesis testing problem with Wasserstein uncertainty sets. Moreover, the optimal adversarial risk is the Bayes error between a worst-case pair of distributions belonging to these sets. Our theoretical results lead to several algorithmic insights for practitioners and motivate further study on the intersection of adversarial robustness and optimal transport.

Contents

Listing of figures

To my mom and dad, for all their sacrifices.

Acknowledgments

I MUST begin by thanking my advisor Prof. Varun Jog, whose brilliant technical capability is only outshined by his patience to tolerate me for all these years. I still fondly remember that one afternoon in my third year where I came to him clueless with a theorem of Dudley[25] and said that the proof looks like black magic to me. It used the marriage lemma and some other notation that might as well be from the devil's handbook. Varun took two hours to work through the proof with me. I am still a bit confused about marriages, but less so about the marriage lemma. Varun taught me to not be afraid of mere symbols, to think clearly and write clearly. I will forever cherish that. I must also thank Prof. Po-Ling Loh, who along with Varun advised me throughout my masters program and kick-started my research at UW-Madison. Through her kindness and patience, Po-Ling has been a constant source of support for me at UW-Madison.

I was fortunate to have the pleasure of interacting with the many excellent faculty at UW-Madison. Prof. Nicolas Garcia Trillos's course on optimal transport was a turning point in my PhD. Discussing my research with Nicolas was always a pleasure. His insights on my problems have led to many of the results in this thesis. I want to thank Prof. Robert Nowak, Prof. Dimitris Papailiopoulos and Prof. Jerry Zhu for their excellent courses on machine learning. Having been initially admitted to UW-Madison for a masters program, their teaching and the passion they had for research convinced me that I needed to continue at UW-Madison for PhD. I would also like to thank Prof. Kangwook Lee for agreeing to join my dissertation committee on short notice, and for the pleasant interactions during my time at Madison.

I would like to thank my amazing friends at Madison, who made even the harshest of Madison's winters bearable. My lab-mates Zheng, Xiaomin, Jinnian, Duzhe, Adrian, Zhili, Ashley, Amir and Ankit—thank you for pleasant memories in the Loh-Jog research group. Special thanks to Ankit who hails from Neemuch and whose niceness is nearly too much. To my uncharacteristically long list of room-mates at Madison—Bhargav, Arjun, Ravi, Nitesh, Shashank, Itzel (except Hans, who I hope gets the help he needs)—I am thankful for your company, and in awe of

your ability to endure my presence. I propose a death match between Bhargav and Ravi to decide who amongst you is the most worthy. To the "Settlers of Madison"—Aarati, Shantanu, Kaushik, Neha, Bhumesh, Bhavya, Michael, Vishnu and Pragathi—thanks for the boardgames and Friday nights. I could not have done this without you all.

I would also like to thank "the trippies"—Madki, Ranji and Hemanshi; my "long-distance room-mates"—Srihari, Jagadish, Nunna and Bobba; my "IIT-M buddies"—J (Manoj), JV (Vegnesh), Pavan, Avinash and Vishwa; "the book club"—Ananyaa, Anurag, Satwik, Sharaf and Srivatsan; and my therapist Sarah from Edelweiss. Your conversations and Zoom / WhatsApp calls got me through some difficult times in the pandemic.

I would like to acknowledge some *adversarial* players who contributed to my growth. To one Ms. M***n A******n who hit me with her car on the night of November 1, 2019, and fractured two of my vertebrae—thank you for bringing me closer to my friends and family, and teaching me about the fragility of this gift of life that I am given. I never got to see you, but you occupied an unwanted amount of my mind-space right before my qualifying exam. I want to thank that one VP in Deutsche Bank, Mumbai, who told me I am not suitable for a career in Finance, and one interviewer in Ola Cabs, Bengaluru, who turned me down for a data science position for undisclosed reasons—both of you contributed to my decision of pursuing graduate studies. You remind me that every closed door opens a new one, although in this case the new door led to a grueling five years of struggle in a strange country away from everyone I knew. But hey, I like to believe that it led to something worthwhile. This thesis is a case in point.

Madison, Wisconsin—thank you for being my home for the past five years. My parents were worried sick about sending off their only son to a cold place like you, but you took good care of me. Watching you utterly transform from one season to the next gave me such great joy that I will forgive you for the winters. Thank you for the sublime sunsets, for the fall colors and for the countless walks along your (not one, not two, but three!) lake-shores.

Finally, I want to thank my dad Chandra Sekhar, my mom Radhika, my sweet little sister Aruna, and my grandma, Rajamma. You are my strength. Thanks also to Nagendra uncle and Padmaja aunty for nursing me back to health when I was sick. There is *one* left to thank, but as the scripture reads, "As all water falling from the sky eventually reaches the sea, so do salutations to everything divine ultimately go to the One".

Darkness there was at first, by darkness hidden;
Without distinctive marks, this all was fluid;
That which, becoming, by the void was covered;
That One by force of contemplation, came into being.

Nāsadīya Sūkta (Hymn of Creation), Rig Veda

1 Introduction

1.1 Motivation

Deep learning has had tremendous success in recent times, producing state-of-the-art results in image classification [49,41], game playing [76,77,89], speech [42,35] and natural language processing [95,22]. However, Szegedy et al. [82] discovered that these algorithms are surprisingly vulnerable to minute adversarial perturbations. Many *adversarial attacks* [1,16,33] and defenses [58,65,20] have been proposed since. Often, the defenses are subsequently broken or are computationally intractable in practice. Despite the demonstrated vulmerabilities of deep learning, the adoption of deep learning in safety-critical applications such as autonomous driving [37,64], medical imaging [36,57,55] and law [50,17] is on the rise. Hence, it has become all the more important to understand the limitations of machine learning algorithms in the presence of adversarial entities.

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 [71,59]. Goodfellow et al. [33] propose that the reason for adversarial examples may be the linear nature of deep

neural networks. Ilyas et al.[47] 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[85] and in practice[81].

Complementary to the theoretical investigations on *why* the adversarial examples occur, there have been many works on *how* to mitigate such adversarial attacks with provable guarantees if possible. The starting point for such investigations is to define a notion of *adversarial risk* that quantifies the robustness of an algorithm to adversarial attacks. A classification algorithm with high accuracy (low risk) in the absence of an adversary may have poor accuracy (high risk) when an adversary is present. Thus, a modified notion known as *adversarial risk* is used to quantify the adversarial robustness of algorithms. Algorithms that minimize adversarial risk are deemed adversarially robust.

In this thesis, we deviate from algorithm-dependent investigations on *how* to mitigate adversarial attacks. We also deviate from theoretical investigations on *why* adversarial examples exist. Our focus in this work is on *what* adversarial risk really is, *what* its fundamental limitations are, and *what* other ways there are to characterize adversarial robustness. The motivation for this focus is explained in more detail below.

1.1.1 Rigorous Foundations

There is no universally agreed upon definition of adversarial risk. Even the simplest setting of binary classification in $\mathbb{R}^d$ with an ℓ_2 adversary admits various definitions involving set expansions[23,34], transport maps[67], Markov kernels[68], and couplings[62]. These works broadly interpret adversarial risk as a measure of robustness to small perturbations, but their definitions differ in subtle details such as the class of adversaries and algorithms considered, budget constraints placed on the adversary, assumptions on the loss function, and the geometries of decision boundaries.

The diversity of definitions for adversarial risk makes it challenging to compare approaches. Theoretical results pertaining to one definition of adversarial risk may not apply to another. Moreover, different definitions are motivated from different view points and the insights that led to the development of algorithms that minimize one type of adversarial risk may not extend to other definitions. Hence, a common framework for comparing the various definitions of adversarial risk is useful both for theoreticians and practitioners. This is the motivation for the following question.

Question 1. How are the various formulations of adversarial risk related to each other?

In addition to the problem of numerous definitions, there is the problem of rigor. Not all approaches for defin-

ing adversarial risk are rigorous. For instance, the classes of adversarial strategies and classifier algorithms are often unclear, and issues of measurability are ignored. Although this may be harmless for applied research, it has led to incorrect proofs and insufficient assumptions in some theoretical works. Thus, we seek to answer the following question.

Question 2. What assumptions are needed to make adversarial risk well-defined?

1.1.2 Fundamental Limits

Once a definition for adversarial risk is established, algorithms that minimize it are sought and deemed adversarially robust. Procedures for finding them have been effective in practice [58,87,65], spurring numerous theoretical investigations into adversarial risk and its minimizers. For instance, a popular method for defending against adversarial attacks is adversarial training, wherein a model is trained using the gradient of the loss computed at the worst-case perturbation of each training data point. The goal in adversarial training is to minimize adversarial risk defined as expected worst-case loss incurred by a model in the presence of a data-perturbing adversary. Another approach knows as randomized smoothing aims to make a classifier more robust by "smoothening" its decision boundary through random averaging. This is done by outputting the majority decision of a classifier on a number of randomly sampled data points close to the test data point.

In light of these investigations on minimizing adversarial risk, a quantity of interest is the *optimal adversarial risk*, which is the minimum value for adversarial risk over the space of all possible decision regions of a classifier. Optimal adversarial risk is analogous to the notion of Bayes risk in a standard learning setup. In a binary classification setup with equal priors, it is known that the Bayes risk is determined by the total variation distance between the true probability distributions of the two classes. It was unclear how this result extends to the adversarial setting. In the presence of a data-perturbing adversary, the risk incurred by any classifier is at least as big as its standard risk. But it was unknown how much the optimal adversarial risk differs from the optimal standard risk, i.e. the Bayes risk. This motivates the second major open question addressed by this thesis.

Question 3. How much can the optimal adversarial risk differ from the Bayes risk?

A related question concerns the optimal classifier in the adversarial setting. A recent line of work shows that non-parametric methods that converge to the Bayes classifier asymptotically can yield non-robust classifiers in the presence of an adversary [93,8]. Hence, the Bayes classifier is not necessarily optimal in the presence of an adversary.

The work of Moosavi-Dezfooli et al.[63], Cohen et al.[21] and Yang et al.[94] 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. Hence, we have the following question.

Question 4. How does the optimal adversarial classifier differ from the Bayes classifier?

1.1.3 Alternative Characterizations

Finding optimal classifiers under 0-1 loss is equivalent to hypothesis testing, and there are natural connections of adversarially robust classification to robust hypothesis testing. Classical literature on robust hypothesis testing 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[44,46]. Contamination models based on f-divergences have also been analyzed for the Kullback-Liebler divergence[54] and the squared Hellinger distance[39,40]. These models study robustness to perturbations in data generating distributions rather than the data itself.

Another line of work studies distributionally robust optimization (DRO)[31,91], wherein the parameters of an optimization problem are sampled from an unknown probability distribution inside an uncertainty set. In Wasserstein DRO[28,27,11], the uncertainty sets are constructed using the Wasserstein metrics. 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.

At first glance, distributional robustness may seem distinct from adversarial robustness. Some recent works suggest connections between the two, when the distributional uncertainty is measured in the Wasserstein metrics[78,86,79]. However, it is unclear whether the adversarial robustness model can be reduced to a distributional robustness model. This motivates the following question.

Question 5. What is the precise relationship between adversarial robustness and distributional robustness?

Optimal adversarial risk is most commonly defined as the minimax risk under adversarial contamination[58,73]. Some recent works consider a two-player zero-sum game between the adversary and the classifier where the payoff is defined as the adversarial risk incurred for a particular choice of contamination by the adversary and a particular choice of decision region by the classifier[62,67,13,14]. In this game, the minimax risk corresponds to the best payoff attainable by the classifier when the adversary makes the first move. A natural question to ask is if this risk matches

with the best payoff attainable by the adversary when the classifier makes the first move. An equality between the two payoffs determines the value of the game and proves the existence of a Nash equilibrium. Hence, we would like to answer the following question.

Question 6. Is there a Nash equilibrium in the zero-sum game between the adversary and the classifier?

1.2 Contributions and Thesis Outline

This thesis addresses all the six questions raised in the preceding section. Our contributions are listed below.

- **Rigorous foundations:** We address Questions 1 and 2 in Chapter 2.
  - **Conditions for well-defined adversarial risk:** We examine the definitions of adversarial risk based on set expansions in a binary classification setup with 0-1 loss. For Polish spaces, we observe that adversarial risk is not Borel measurable, and hence, not well-defined when the decision region is an arbitrary Borel set (or, when the loss function is an arbitrary Borel measurable function). We show that the problem can be resolved by considering a Polish space equipped with the universal completion of the Borel σ-algebra and restricting the decision regions to Borel sets (or by restricting the loss function to be upper semi-analytic, which is stronger than Borel measurability and weaker than universal measurability). For the Euclidean space with the Lebesgue σ-algebra, we show that adversarial risk is well-defined for any Lebesgue measurable decision region. Our key lemma (Lemma 2.3.3) shows that the Lebesgue σ-algebra is preferred over the Borel σ-algebra because set expansions are Lebesgue measurable but not necessarily Borel measurable. For general loss functions, we show that the expected supremum formulation of adversarial risk is well-defined for upper semi-analytic loss functions. These results resolve Question 1, and are contained in Section 2.3.
  - **Equivalences between adversarial risk definitions:** We show that the definition of adversarial risk using set expansions is identical to a notion of risk that appears in robust hypothesis testing with ∞-Wasserstein uncertainty sets. We prove this result in Polish spaces using the theory of measurable selections[5,90]. In $\mathbb{R}^d$, we are able to use the powerful theory of Choquet capacities[19] (in particular, Huber and Strassen's 2-alternating capacities[46]) to establish results of a similar nature. In addition, we derive the conditions under which this notion of adversarial risk is equivalent to alternative no-

tions defined using transport maps and Markov kernels. These results address Question 2, and are contained in Section 2.4.

- **Fundamental limits:** We address Questions 3 and 4 in Chapter 3.
  - **Optimal adversarial risk via optimal transport:** We resolve Question 3 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. We show that the formula can be extended to the case of unequal priors by considering an unbalanced optimal transport cost between scaled data distributions. The main tool we use is Theorem 9 in which we generalize a classical result of Strassen on excess-cost optimal transport [80,88] from probability measures to finite measures with possibly unequal mass. These results are contained in Section 3.1.
  - **Optimal adversarial classifiers via optimal couplings:** We construct optimal couplings for the optimal transport cost that determines the optimal adversarial risk when the two data distributions are univariate normal, uniform over intervals, and triangular. We resolve Question 4 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. These results are contained in Section 3.2.
  - **Results for continuous loss functions:** For continuous loss functions, we derive 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 4 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. These results are contained in Section 3.3.
  - **Adversarial risk bounds for real-world datasets:** 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. These results are contained in Section 3.4.

- **Alternative characterizations:** We address Questions 5 and 6 in Chapter 4.

  - **Distributional robustness perspective:** We consider a robust hypothesis testing setup where the true distributions are contaminated in ε-balls measured in the ∞-Wasserstein metric. We prove that there exists a least favorable pair of distributions (LFDs) in the uncertainty sets that determine the optimal error achievable in this setting. For Polish spaces with a *midpoint property*, our proof uses a novel characterization of the D_ε optimal transport cost from Section 3.1 in terms of the shortest total variation distance between W_∞ probability balls. For $\mathbb{R}^d$, our proof borrows from the results of Huber and Strassen [46] on 2-alternating capacities. These results are contained in Section 4.1.

  - **Game theoretic perspective:** We consider the setup of a zero-sum game between the adversary and the classifier. We show that the value of this game (adversarial risk) is equal to the minimum Bayes error between a pair of distributions belonging to the ∞-Wasserstein uncertainty sets centered around true data-generating distributions. We prove the existence of a pure Nash equilibrium in this game for $\mathbb{R}^d$ and for Polish spaces with a *midpoint property*. This extends/strengthens existing results [62,67,13] to non-parametric classifiers. These results are contained in Section 4.2.

[T]he requirement of rigor, which has become proverbial in mathematics, corresponds to a universal philosophical necessity of our understanding; . . . A new problem, especially when it comes from the world of outer experience, is like a young twig, which thrives and bears fruit only when it is grafted carefully and in accordance with strict horticultural rules upon the old stem, the established achievements of our mathematical science.

David Hilbert, Lecture at the ICM - Paris, 1900

2

Rigorous Foundations

2.1 Notation and Preliminaries

2.1.1 Notation

Throughout the thesis, we use $\mathcal{X}$ to denote a Polish space (a complete, separable metric space) with metric d and Borel σ-algebra $\mathcal{B}(\mathcal{X})$. For $x \in \mathcal{X}$ and $A \subseteq \mathcal{X}$, we define the distance of x from A as, $d(x, A) = \inf_{a \in A} d(x, a)$. For $x \in \mathcal{X}$ and $r \geq 0$, let $B_r(x)$ denote the closed ball of radius r centered at x. We use $\mathcal{P}(\mathcal{X})$ and $\mathcal{M}(\mathcal{X})$ to denote the space of probability measures and finite measures defined on the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, respectively. Let $\overline{\mathcal{B}}(\mathcal{X})$ denote the universal completion of $\mathcal{B}(\mathcal{X})$. Let $\overline{\mathcal{P}}(\mathcal{X})$ and $\overline{\mathcal{M}}(\mathcal{X})$ denote the space of probability measures and finite measures defined on the complete measure space $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$. For $\mu, \nu \in \mathcal{M}(\mathcal{X})$, we say ν *dominates* μ if $\mu(A) \leq \nu(A)$ for all $A \in \mathcal{B}(\mathcal{X})$ and write $\mu \preceq \nu$. When $\mathcal{X}$ is $\mathbb{R}^d$, we use $\mathcal{L}(\mathcal{X})$ to denote the Lebesgue σ-algebra and λ to denote the d-dimensional Lebesgue measure. For a positive integer n, we use $[n]$ to denote the finite set $\{1, \ldots, n\}$.

2.1.2 Metric Space Topology

Let $\varepsilon \geq 0$ and $A \in \mathcal{B}(\mathcal{X})$. We introduce the following three notions of expansion of the set A by ε.

Definition 1. (Minkowski set expansion) The *ε-Minkowski expansion* of A is given by $A^{\oplus\varepsilon} := \cup_{a\in A} B_\varepsilon(a)$.

Definition 2. (Closed set expansion) The *ε-closed expansion* of A is defined as $A^{\varepsilon} := \{x \in \mathcal{X} : d(x, A) \leq \varepsilon\}$.

Definition 3. (Open set expansion) The *ε-open expansion* of A is defined as $A^{\varepsilon)} := \{x \in \mathcal{X} : d(x, A) < \varepsilon\}$.

We use the notation $A^{-\varepsilon}$ to denote $((A^c)^{\varepsilon})^c$. Similarly, $A^{\ominus\varepsilon} := ((A^c)^{\oplus\varepsilon})^c$. For example, consider the set $A = (0, 1]$ in the space $(\mathcal{X}, d) = (\mathbb{R}, |\cdot|)$ and $\varepsilon > 0$. Then $A^{\oplus\varepsilon} = (-\varepsilon, 1+\varepsilon]$, $A^{\varepsilon} = [-\varepsilon, 1+\varepsilon]$ and $A^{\varepsilon)} = (-\varepsilon, 1+\varepsilon)$. For any $A \in \mathcal{B}(\mathcal{X})$, A^{ε} is closed and $A^{\varepsilon)}$ is open. Hence, $A^{\varepsilon}, A^{\varepsilon)} \in \mathcal{B}(\mathcal{X})$. Moreover, $A^{\varepsilon)} \subseteq A^{\oplus\varepsilon} \subseteq A^{\varepsilon}$. However, $A^{\oplus\varepsilon}$ may not be in $\mathcal{B}(\mathcal{X})$ (see Lemma 2.3.1). In general, the Minkowski sum of two Borel sets need not be Borel[26], and that of two Lebesgue measurable sets need not be Lebesgue measurable[75].

2.1.3 Optimal Transport

Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. A *coupling* between μ and ν is a joint probability measure $\pi \in \mathcal{P}(\mathcal{X}^2)$ with marginals μ and ν. The set $\Pi(\mu, \nu) \subseteq \mathcal{P}(\mathcal{X}^2)$ denotes the set of all couplings between μ and ν. The *optimal transport cost* between μ and ν under a cost function $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is defined as,

$$\mathcal{T}_c(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x, x')d\pi(x, x'). \tag{2.1}$$

For a positive integer p, the *p-Wasserstein distance* between μ and ν is defined as, $W_p(\mu, \nu) = (\mathcal{T}_{d^p}(\mu, \nu))^{\frac{1}{p}}$. The *$\infty$-Wasserstein metric* is defined as $W_\infty(\mu, \nu) = \lim_{p\to\infty} W_p(\mu, \nu)$. It can also be expressed in the following ways[30].

$$W_\infty(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \operatorname*{ess\,sup}_{(x,x')\sim\pi} d(x, x') = \inf\{\delta > 0 : \mu(A) \leq \nu(A^\delta)\forall A \in \mathcal{B}(\mathcal{X})\}. \tag{2.2}$$

Given a $\mu \in \mathcal{P}(\mathcal{X})$ and a measurable function $f : \mathcal{X} \to \mathcal{X}$, the *push-forward* of μ by f is defined as a probability measure $f_{\sharp\mu} \in \mathcal{P}(\mathcal{X})$ given by $f_{\sharp\mu} = \mu(f^{-1}(A))$ for all $A \in \mathcal{B}(\mathcal{X})$.

2.1.4 Submodular Capacities

We introduce the following definitions from the work of Huber and Strassen[46].

Definition 4 (Capacity)**.** A set function $v : \mathcal{B}(\mathcal{X}) \to [0,1]$ is a *capacity* if it satisfies the following conditions.

1. $v(\varphi) = 0$ and $v(\mathcal{X}) = 1$, where φ denotes the null set.
2. For $A, B \subseteq \mathcal{X}, A \subseteq B \implies v(A) \leq v(B)$.
3. $A_n \uparrow A \implies v(A_n) \uparrow v(A)$.
4. $F_n \downarrow F, F_n$ closed $\implies v(F_n) \downarrow v(F)$.

Definition 5 (2-Alternating Capacity)**.** A capacity v defined on the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is called 2-alternating if it satisfies the following condition for any $A, B \subseteq \mathcal{X}$.

$$v(A \cup B) + v(A \cap B) \leq v(A) + v(B). \tag{2.3}$$

Note that the above 2-alternating condition is equivalent to the submodularity condition for defining a submodular set function. Hence, 2-alternating capacities are also known as submodular capacities.

For any compact set of probability measures $\Xi \subseteq \mathcal{P}(\mathcal{X})$, the upper probability defined as $v(A) = \sup_{\mu \in \Xi} \mu(A)$ is a capacity[46]. The upper probability of ε-neighborhoods of a $\mu \in \mathcal{P}(\mathcal{X})$ defined using either the total variation metric or the Levy-Prokhorov metric can be shown to be a 2-alternating capacity[46].

2.2 A Survey of the Many Faces of Adversarial Risk

In this section, we review several definitions for adversarial risk that are found in the literature. First, we consider a setting of general loss functions, where classifiers are parametrized by parameter w in a hypothesis class $\mathcal{W}$. Next, we consider a binary classification setting with the 0-1 loss function, where non-parametric classifiers of the form $f_A(x) = 1\{x \in A\}$ correspond to decision regions $A \subseteq \mathcal{X}$.

2.2.1 General Loss Setting

Let $\mathcal{X}$ be the feature space, a Polish space equipped with a distance metric d. Let $\mathcal{Y}$ be a finite set of labels. Let ρ be the true data distribution of labeled data points $(x, y) \in \mathcal{X} \times \mathcal{Y}$, which can be expressed as $\rho(x, y) = \rho_y(y)\rho_{x|y}(x)$ where $\rho_y(y)$ is the marginal probability of label $y \in \mathcal{Y}$ and $\rho_{x|y}(x)$ is the conditional probability of $x \in \mathcal{X}$ given

the label y. Let $\mathcal{W}$ denote the hypothesis class. Let $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{W} \to [0, \infty]$ denote a loss function that is measurable with respect to $\mathcal{B}(\mathcal{X})$ for all $w \in \mathcal{W}$.

Consider a data-perturbing adversary of budget $\varepsilon \geq 0$ that perturbs any data point $x \in \mathcal{X}$ to $x' \in \mathcal{X}$ such that $d(x, x') \leq \varepsilon$. The adversarial risk of a classifier $w \in \mathcal{W}$ under a loss function ℓ in the presence of such an adversary is given by,

$$R_{\oplus\varepsilon}(\ell, w) = \mathbb{E}_{(x,y)\sim\rho}\left[\sup_{d(x,x')\leq\varepsilon} \ell((x', y), w)\right]. \tag{2.4}$$

If the loss function $\ell(\cdot, w)$ is upper semi-continuous and bounded above for all $w \in \mathcal{W}$, Meunier et al. ([62]) show that $R_{\oplus\varepsilon}(\ell, w)$ is well-defined. But in general, it may not be so.

One way to resolve measurability issues is to restrict the adversary to use measurable transport maps for data perturbation. Let $F := \{f_y : \mathcal{X} \to \mathcal{X}, f_y \text{ is } \rho_y - \text{measurable} \mid y \in \mathcal{Y}\}$ denote a collection of measurable maps for each label $y \in \mathcal{Y}$. We say that F is of budget ε (denoted by $F \in F_\varepsilon$) if $d(x, f_y(x)) \leq \varepsilon$ with probability 1 for $(x, y) \sim \rho$. Under such an adversary, the adversarial risk may be defined as follows.

$$R_{F_\varepsilon}(\ell, w) = \sup_{F\in F_\varepsilon} \mathbb{E}_{(x,y)\sim\rho}\left[\ell((f_y(x), y), w)\right]. \tag{2.5}$$

The above definition was used for the binary classification setting in Pinot et al.[67] A more general definition for adversarial risk was proposed in Pydi and Jog[68] using Markov kernels. Let κ denote a set of Markov kernels κ_y for $y \in \mathcal{Y}$. Let $\rho^\kappa_{(x,y,x')}$ denote the joint distribution of (x, y, x') induced by κ. We say that the Markov kernel adversary κ has a budget ε (denoted by $\kappa \in K_\varepsilon$) if $d(x, x') \leq \varepsilon, \rho^\kappa_{(x,x')|y}$-a.s. where $\rho^\kappa_{(x,x')|y} \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ denotes the conditional distribution of (x, x') given $y \in \mathcal{Y}$ and x' is the perturbation of the data point x with label y using the Markov kernel $\kappa_y \in \kappa$. Under such a Markov kernel adversary, adversarial risk is defined as the following in Pydi and Jog[68].

$$R_{K_\varepsilon}(\ell, w) = \sup_{\kappa\in K_\varepsilon} \mathbb{E}_{(x,y,x')\sim\rho^\kappa_{(x,y,x')}}\left[\ell((x', y), w)\right]. \tag{2.6}$$

Another way to define adversarial risk is by considering perturbations to the input data distributions rather than individual data points. Optimal transport-based perturbations, in particular the ∞-Wasserstein metric (denoted by W_∞) has been used to define such perturbations ([68,62]). Let an adversary γ be defined as a collection of perturbed

probability distributions for each label i.e., $\gamma := \{\rho^{\gamma}_{x'|y} \in \mathcal{P}(\mathcal{X}) | y \in \mathcal{Y}\}$. We say that the adversary γ has a budget ε (denoted by Γ_ε) if $W_\infty(\rho_{x|y}, \rho^{\gamma}_{x'|y}) \leq \varepsilon$ for all $y \in \mathcal{Y}$. Under such a distribution perturbing adversary, the adversarial risk is defined as,

$$R_{\Gamma_\varepsilon}(\ell, w) = \sup_{\gamma \in \Gamma_\varepsilon} \mathbb{E}_{(x',y)\sim \rho_y \rho^{\gamma}_{x'|y}} \left[\ell((x', y), w)\right]. \tag{2.7}$$

The use of ∞-Wasserstein metric for defining adversarial risk is motivated by the following fact: For $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $W_\infty(\mu, \nu) \leq \varepsilon$ if and only if there exists a coupling (a joint probability distribution) $\pi \in \Pi(\mu, \nu)$ such that $d(x, x') \leq \varepsilon$ with probability 1 for $(x, x') \sim \pi$. That means, all the probability mass under the distribution μ may be transported to ν without transporting any mass by more than ε almost surely.

The following inequality is an immediate consequence of the above definitions of adversarial risk:

$$R_{F_\varepsilon}(\ell, w) \leq R_{K_\varepsilon}(\ell, w) \leq R_{\Gamma_\varepsilon}(\ell, w). \tag{2.8}$$

We shall investigate conditions for equality in the above inequality and relations between the above three formulations of adversarial risk and the classical formulation $R_{\oplus\varepsilon}(\ell, w)$.

2.2.2 Binary Classification with 0-1 Loss Setting

In this subsection, we consider a binary classification setting where the label space $\mathcal{Y} = \{0, 1\}$. Let $p_0, p_1 \in \mathcal{P}(\mathcal{X})$ be the data-generating distributions for labels 0 and 1, respectively. Let the prior probabilities for labels 0 and 1 be in the ratio $T : 1$ where we assume $T \geq 1$ without loss of generality. For any set $A \in \mathcal{B}(\mathcal{X})$, we may consider a classifier $f_A(x) := 1\{x \in A\}$ which labels any point in the set A as 1 and any point in A^c as 0. We say that such a classifier has a decision region A. The error (standard risk) incurred by such a classifier under the 0-1 loss function is, $R_{\oplus 0}(\ell_{0/1}, A) = \frac{T}{T+1} p_0(A) + \frac{1}{T+1} p_1(A^c)$.

An adversary of budget $\varepsilon > 0$ can perturb any $x \in \mathcal{X}$ to $x' \in B_\varepsilon(x)$. It follows that any $x \in A$ can be perturbed to $x' \in \cup_{a \in A} B_\varepsilon(a) = A^{\oplus\varepsilon}$. Hence, adversarial risk could be defined as

$$R_{\oplus\varepsilon}(\ell_{0/1}, A) = \frac{T}{T+1} p_0(A^{\oplus\varepsilon}) + \frac{1}{T+1} p_1((A^c)^{\oplus\varepsilon}). \tag{2.9}$$

The above formulation is a special case of (2.4) for the 0-1 loss function. Indeed, for $x \in \mathcal{X}$ and $y \in \{0, 1\}$,

$\ell_{0/1}((x,y),A) = 1\{x \in A, y = 0\} + 1\{x \in A^c, y = 1\}$. Hence,

$$\begin{aligned} & R_{\oplus\varepsilon}(\ell_{0/1}, A) \\ &= \frac{T}{T+1}\mathbb{E}_{p_0}\left[\sup_{d(x,x')\leq\varepsilon} 1\{x' \in A\}\right] + \frac{1}{T+1}\mathbb{E}_{p_1}\left[\sup_{d(x,x')\leq\varepsilon} 1\{x' \in A^c\}\right] \\ &= \frac{T}{T+1}p_0(A^{\oplus\varepsilon}) + \frac{1}{T+1}p_1((A^c)^{\oplus\varepsilon}). \end{aligned}$$

A problem with the formulation in equation (2.9) is the ambiguity over the measurability of the sets $A^{\oplus\varepsilon}$ and $(A^c)^{\oplus\varepsilon}$. Even when $A \in \mathcal{B}(\mathcal{X})$, it is not guaranteed that $A^{\oplus\varepsilon}, (A^c)^{\oplus\varepsilon} \in \mathcal{B}(\mathcal{X})$ (see Appendix A for an example). Hence, $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is not well-defined for all $A \in \mathcal{B}(\mathcal{X})$. It is shown in Pydi and Jog[68] that $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is well-defined when A is either closed or open. A simple fix to this measurability problem is to use closed set expansion instead of the Minkowski set expansion, as done in Mahloujifar et al.[59] This leads to the following formulation.

$$R_{\varepsilon}(\ell_{0/1}, A) = \frac{T}{T+1}p_0(A^{\varepsilon}) + \frac{1}{T+1}p_1((A^c)^{\varepsilon}). \tag{2.10}$$

The above definition is well-defined for any $A \in \mathcal{B}(\mathcal{X})$ because A^{ε} and $(A^c)^{\varepsilon}$ are both closed and hence, measurable. However, under the above definition, a point $x \in A$ may be perturbed to $x' \in A^{\varepsilon}$ such that $d(x,x') > \varepsilon$. For example, when $A = (0,1)$, we have $A^{\varepsilon} = [-\varepsilon, \varepsilon]$ and an adversary may transport $x = \delta > 0$ to $x' = -\varepsilon$, violating the budget constraint at x. Another problem with this definition is the fact that we do not recover the standard risk by setting $\varepsilon = 0$ in the above definition.

Alternatively, one can also fix the measurability problem by considering open set expansions, as done in Bungent et al.[15], leading to the following formulation.

$$R_{\varepsilon)}(\ell_{0/1}, A) = \frac{T}{T+1}p_0(A^{\varepsilon)}) + \frac{1}{T+1}p_1((A^c)^{\varepsilon)}). \tag{2.11}$$

Like $R_{\varepsilon}(\ell_{0/1}, A)$, $R_{\varepsilon)}(\ell_{0/1}, A)$ is also well-defined for any $A \in \mathcal{B}(\mathcal{X})$. As remarked in Bungent et al.[15], a drawback with this approach is that setting $\varepsilon = 0$ does not recover the standard risk definition.

Remark 1. *The formulations in equations* (2.4), (2.9) *and* (2.10) *can give a strictly positive adversarial risk even for a "perfect" (i.e., Bayes optimal) classifier. This is consistent with the literature on adversarial examples where even a perfect classifier is forced to make errors in the presence of evasion attacks. These formulations of adversarial risk*

correspond to "constant-in-the-ball" risk of Gourdeau et al.[34], and "corrupted-instance" risk in Diochnos et al.[23] and Mahloujifar et al.[59] Here, an adversarial risk of zero is only possible if the supports of p_0 and p_1 are non-overlapping and separated by at least 2ε. This is not the case with other formulations of adversarial risk such as "exact-in-the-ball" risk[34], "prediction-change risk and "error-region" risk[23,59]. We focus on the "corrupted-instance" family of risks in this work.

Another approach to defining adversarial risk is by explicitly defining the class of adversaries of budget ε as measurable transport maps $f : \mathcal{X} \to \mathcal{X}$ that push-forward the true data distribution such that no point is transported by more than a distance of ε; i.e., $d(x, f(x)) \leq \varepsilon$. The transport map-based adversarial risk[67] is formally defined as follows:

$$R_{F_\varepsilon}(\ell_{0/1}, A) = \sup_{\substack{f_0, f_1 : \mathcal{X} \to \mathcal{X} \\ \forall x \in \mathcal{X}, d(x, f_i(x)) \leq \varepsilon}} \frac{T}{T+1} f_{0\sharp p_0}(A) + \frac{1}{T+1} f_{1\sharp p_1}((A^c)). \tag{2.12}$$

It is easy to see that the above definition is a special case of the definition in equation (2.5) for the 0-1 loss function. Yet another definition uses the robust hypothesis testing framework with W_∞ uncertainty sets. In this approach, an adversary perturbs the true distribution p_i to a corrupted distribution p_i' such that $W_\infty(p_i, p_i') \leq \varepsilon$. From (2.2), this is equivalent to the existence of a coupling $\pi \in \Pi(p_i, p_i')$ such that $\operatorname{ess\,sup}_{(x,x') \sim \pi} d(x, x') \leq \varepsilon$. The adversarial risk with such an adversary is given by

$$R_{\Gamma_\varepsilon}(\ell_{0/1}, A) = \sup_{W_\infty(p_1, p_1'), W_\infty(p_0, p_0') \leq \varepsilon} \frac{T}{T+1} p_0'(A) + \frac{1}{T+1} p_1'((A^c)). \tag{2.13}$$

Clearly, $R_{F_\varepsilon}(\ell_{0/1}, A) \leq R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$, but conditions for equality were not studied in prior work. Moreover, their relation to set expansion-based definitions in (2.9) and (2.10) was also unknown.

2.2.3 Other Related Notions of Adversarial Risk

2.2.3.1 Error Region Adversarial Risk

In this thesis, 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,

$\rho(x,y) = 1\{y = c(x)\}\rho_x(x)$. Under this model, Gourdeau et al.[34] define *constant-in-the-ball risk* as

$$R(h) = \mathbb{P}_{x\sim\rho_x}[\exists x' : d(x,x') \leq \varepsilon,\ h(x') \neq c(x)]. \tag{2.14}$$

Rewriting in terms of expectation, we get the following.

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

Hence, the *constant-in-the-ball risk* defined above is identical to adversarial risk defined in (2.4) for 0-1 loss function under hypothesis h. The same notion of risk is also called the *corrupted instance risk* in Diochnos et al.[23].

A related notion of adversarial risk is the following:

$$R'(h) = \mathbb{E}_{x\sim\rho_x}\left[\sup_{d(x,x')\leq\varepsilon} 1\{h(x') \neq c(x')\}\right]. \tag{2.15}$$

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.[34] and *error-region risk* in Diochnos et al.[23]. A key difference between $R(h)$ in (2.14) and $R'(h)$ in (2.15) is that $R'(h)$ is exactly equal to 0 for $h = c$ for any $\varepsilon \geq 0$ 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.[34], $R(h)$ 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.

2.2.3.2 Distributionally Robust Risk

The adversarial risk formulation under a distribution perturbing adversary has been widely studied in the distributionally robust optimization (DRO) literature[31,91], with special focus on Wasserstein DRO[28,27,11]. 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 (2.7) where it is centered around the true data generating distribution. Bertsimas et al.[6] 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.[86] and Staib and Jegelka[79] that make a similar observation as our Theorem 5. Distributionally robust risk has also been studied in a minimax statistical learning framework in [53,61] for deriving generalization error bounds.

In the setting of Section 2.2.1, we may extend the $R_{\Gamma_\varepsilon}(\ell, w)$ definition of adversarial risk to a p-Wasserstein distributionally robust risk. As before, let an adversary γ be defined as a collection of perturbed probability distributions for each label i.e., $\gamma := \{\rho^\gamma_{x'|y} \in \mathcal{P}(\mathcal{X}) | y \in \mathcal{Y}\}$. We say that the adversary γ has a budget ε in p-Wasserstein metric (denoted by $\Gamma^{(p)}_\varepsilon$) if $W_p(\rho_{x|y}, \rho^\gamma_{x'|y}) \leq \varepsilon$ for all $y \in \mathcal{Y}$. Under such a distribution perturbing adversary, the p-Wasserstein distributionally robust risk is defined as,

$$R_{\Gamma^{(p)}_\varepsilon}(\ell, w) = \sup_{\gamma \in \Gamma^{(p)}_\varepsilon} \mathbb{E}_{(x',y)\sim \rho_y \rho^\gamma_{x'|y}} \left[\ell((x', y), w)\right]. \tag{2.16}$$

The $R_{\Gamma_\varepsilon}(\ell, w)$ definition of adversarial risk is therefore a special case of the p-Wasserstein distributionally robust risk, $R_{\Gamma^{(p)}_\varepsilon}(\ell, w)$, with $p = \infty$.

For any $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and integers p, q satisfying $1 \leq p \leq q < \infty$, we have the inequality $W_\infty(\mu, \nu) \leq W_q(\mu, \nu) \leq W_p(\mu, \nu)$. Hence, $\Gamma_\varepsilon \subseteq \Gamma^{(q)}_\varepsilon \subseteq \Gamma^{(p)}_\varepsilon$. This leads to the following inequality.

$$R_{\Gamma_\varepsilon}(\ell, w) \leq R_{\Gamma^{(q)}_\varepsilon}(\ell, w) \leq R_{\Gamma^{(p)}_\varepsilon}(\ell, w). \tag{2.17}$$

2.2.3.3 Surrogate Losses for Adversarial Risk

Before adversarial deep learning, minimax risk was studied in the context of robust classification with linear classifiers and SVMs [51,74,92,72,18]. Here, one proposes surrogate robust loss functions that can be tractably minimized. A similar strategy for minimizing adversarial loss may be found in [48]. For a discussion of surrogate losses, we refer the reader to Bao et al.[2].

In practice, the inner maximization term in the adversarial risk is approximated using gradient methods like the Fast Gradient Sign Method (FGSM) [33,16]. This gives rise to several related notions of risk that can be interpreted as a Taylor approximations for adversarial risk in definition $R_{\oplus\varepsilon}(\ell, w)$. [58,73].

Surrogate loss functions for ensuring Wasserstein distributional robustness have been proposed in [28,27], and ro-

bustness with respect to other optimal transport-based perturbations is studied in [11]. A key idea in these works is the dual formulation of optimal transport distances. As we will see in the next section, 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 (2.1) with (2.2)) and so the classic Kantorovich-Rubinstein duality cannot be applied.

2.2.3.4 Adversaries in 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 [44,46]. Contamination models based on f-divergences have also been analyzed for the Kullback-Liebler divergence [54] and the squared Hellinger distance [39,40].

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 [45]. When the uncertainty sets are constructed with the Hellinger distance, methods are known for obtaining nearly optimal estimators [52,4,3].

2.3 Conditions for Well-Defined Adversarial Risk

In this section, we discuss the conditions under which the definitions for adversarial risk presented in Section 2.2 are well-defined. In Subsection 2.3.1 we present the results for the binary classification setting under 0-1 loss and in Subsection 2.3.2 we discuss the setting of more general loss functions.

2.3.1 Binary Classification with 0-1 Loss Setting

As stated in Section 2.2, $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ may not be well-defined for some decision regions $A \in \mathcal{B}(\mathcal{X})$ because of the non-measurability of the sets $A^{\oplus\varepsilon}$ and $(A^c)^{\oplus\varepsilon}$. Specifically, we have the following lemma.

Lemma 2.3.1. *For any $\varepsilon > 0$, there exists $A \in \mathcal{B}(\mathcal{X})$ such that $A^{\oplus\varepsilon} \notin \mathcal{B}(\mathcal{X})$.*

The proof of Lemma 2.3.1 is in Appendix B.1.1.

In this section, we lay down the conditions under which the ambiguity on the measurability of $A^{\oplus\varepsilon}$ can be resolved. We begin by presenting a lemma that shows that $A^{\oplus\varepsilon}$ is an analytic set (i.e., a continuous image of a Borel set) whenever A is Borel. It is known that analytic sets are universally measurable; i.e., they belong in $\overline{\mathcal{B}}(\mathcal{X})$, the universal completion of the Borel σ-algebra $\mathcal{B}(\mathcal{X})$, and are measurable with respect to any finite measure defined on the complete measure space, $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$.

Lemma 2.3.2. *Let $A \in \mathcal{B}(\mathcal{X})$. Then, $A^{\oplus\varepsilon}$ is an analytic set. Consequently, $A^{\oplus\varepsilon} \in \overline{\mathcal{B}}(\mathcal{X})$.*

The proof of Lemma 2.3.2 is in Appendix B.1.1. By virtue of Lemma 2.3.2, we have the following.

Theorem 1. *Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$. Then for any $A \in \mathcal{B}(\mathcal{X})$, $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is well-defined.*

The proof of Theorem 1 is in Appendix B.1.1. For the special case of $\mathcal{X} = \mathbb{R}^d$, we can further strengthen Theorem 1 to include all Lebesgue measurable sets $\mathcal{L}(\mathcal{X})$ instead of just Borel sets $\mathcal{B}(\mathcal{X})$. For this, we use the concept of porous sets.

Definition 6 (Porous set)**.** A set $E \subseteq \mathcal{X}$ is said to be porous if there exists $\alpha \in (0,1)$ and $r_0 > 0$ such that for every $r \in (0, r_0]$ and every $x \in \mathcal{X}$, there is an $x' \in \mathcal{X}$ such that $B_{\alpha r}(x') \subseteq B_r(x) \backslash E$.

Porous sets are a subclass of nowhere dense sets. Importantly, $\lambda(E) = 0$ for any porous set $E \subseteq \mathbb{R}^d$ [99]. By the following lemma, the set difference between the closed/open set expansions is porous.

Lemma 2.3.3. *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$ and $A \in \mathcal{L}(\mathcal{X})$. Then $E = A^{\varepsilon} \backslash A^{\varepsilon)}$ is porous.*

The proof of Lemma 2.3.3 is in Appendix B.1.1. Lemma 2.3.3 plays a crucial role in proving that $A^{\oplus\varepsilon} \in \mathcal{L}(\mathcal{X})$ whenever $A \in \mathcal{L}(\mathcal{X})$. We recall that $A^{\oplus\varepsilon}$ is the Minkowski sum of A with the closed ε-ball. In general, the Minkowski sum of two Lebesgue measurable sets is not always Lebesgue measurable [75,29]. So the fact that one of them is a closed ball in case of $A^{\oplus\varepsilon}$ is important. In the following theorem, we use Lemma 2.3.3 to prove the measurability of $A^{\oplus\varepsilon}$ and in turn prove that $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is well-defined for any $A \in \mathcal{L}(\mathcal{X})$.

Theorem 2. *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \geq 0$. Then for any $A \in \mathcal{L}(\mathcal{X})$, $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is well-defined.*

Proof. By Lemma 2.3.3 $A^{\varepsilon} \backslash A^{\varepsilon)}$ is porous, and so $\lambda(A^{\varepsilon} \backslash A^{\varepsilon)}) = 0$. Hence, $\lambda(A^{\varepsilon}) = \lambda(A^{\varepsilon)})$. Using the fact that $A^{\varepsilon)} \subseteq A^{\oplus\varepsilon} \subseteq A^{\varepsilon}$, we have $A^{\oplus\varepsilon} \backslash A^{\varepsilon)} \subseteq A^{\varepsilon} \backslash A^{\varepsilon)}$. Hence, $\lambda(A^{\oplus\varepsilon} \backslash A^{\varepsilon)}) = 0$. Therefore, $A^{\oplus\varepsilon} \in \mathcal{L}(\mathcal{X})$ and $\lambda(A^{\oplus\varepsilon}) = \lambda(A^{\varepsilon}) = \lambda(A^{\varepsilon)})$. Since $A^{\oplus\varepsilon}, (A^c)^{\oplus\varepsilon} \in \mathcal{L}(\mathcal{X})$, $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is well-defined. □

2.3.2 General Loss Setting

In the expected-supremum formulation of adversarial risk shown in (2.4), the worst-case loss function given by $\sup_{d(x,x')\leq\varepsilon} \ell((x',y),w)$ may not be measurable even when $\ell((x',y),w)$ is measurable for every $x' \in \mathcal{X}$ because the supremum is taken over an uncountable family of measurable functions. In this subsection, we resolve this ambiguity over the measurability of the worst-case loss function.

A real-valued function $\varphi : \mathcal{X} \to \mathbb{R}$ is called upper semi-analytic if the set $\{x \in \mathcal{X} : \varphi(x) > t\}$ is an analytic set for every $t \in \mathbb{R}$. Since every Borel set is an analytic set, it follows that every Borel measurable function is upper semi-analytic. However, the converse is not true in general. Nevertheless, upper semi-analytic functions are universally measurable owing to the fact that analytic sets are universally measurable. We now present a lemma that shows that the worst-case loss function $\sup_{d(x,x')\leq\varepsilon} \ell((x',y),w)$ is universally measurable if $\ell((\cdot,y),w)$ is upper semi-analytic for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$.

Lemma 2.3.4. *If the loss function $\ell((\cdot,y),w)$ is upper semi-analytic for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, then the worst-case loss function $\sup_{d(x,x')\leq\varepsilon} \ell((x',y),w)$ is also upper semi-analytic and hence universally measurable. Therefore, $R_{\oplus\varepsilon}(\ell,w)$ is well-defined on the measure space $(\mathcal{X},\overline{\mathcal{B}}(\mathcal{X}))$.*

The proof of Lemma 2.3.4 is in Appendix B.1.2. For the special case of $\mathcal{X} = \mathbb{R}^d$, we can further extend the measurability of the worst-case loss function from upper semi-analytic functions to the more general Lebesgue measurable functions, as shown in the following lemma.

Lemma 2.3.5. *Let $(\mathcal{X},d) = (\mathbb{R}^d, \|\cdot\|)$. Then, $R_{\oplus\varepsilon}(\ell,w)$ is well-defined for any loss function $\ell : (\mathcal{X}\times\mathcal{Y})\times\mathcal{W} \to [0,\infty]$ for which $\ell((\cdot,y),w)$ is Lebesgue measurable for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$.*

The proof of Lemma 2.3.5 is in Appendix B.1.2.

Now that we have established the conditions for which $R_{\oplus\varepsilon}$ is well-defined, in the next section, we explore its relation to other notions of adversarial risk.

2.4 Equivalences between Adversarial Risk Definitions

In this section, we show the conditions under which $R_{\oplus\varepsilon}(\ell_{0/1},A)$ is equivalent to other notions of adversarial risk based on transport maps and W_∞ robustness. The equivalences established in this section are summarized in

Table 2.1: Equivalences among adversarial risk formulations for 0-1 loss. $R_{\oplus\varepsilon}(A), R_{\varepsilon}(A), R_{F_\varepsilon}(A)$ and $R_{\Gamma_\varepsilon}(A)$ denote adversarial risk for 0-1 loss function ($\ell_{0/1}$) for a binary classifier with decision region A (i.e. $f_A(x) = 1\{x \in A\}$), defined using Minkowski set expansions, closed set expansions, transport maps and ∞-Wasserstein metric respectively. $\mathcal{B}(\mathcal{X})$ and $\mathcal{L}(\mathcal{X})$ denote the Borel and Lebesgue σ-algebras. $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$ denotes the universal completion of the Borel measure space, $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Equivalences in Adversarial Risk	Conditions
$R_{\oplus\varepsilon}(A) = R_{\Gamma_\varepsilon}(A)$	$\mathbb{R}^d$: $A \in \mathcal{L}(\mathcal{X})$ or $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$: $A \in \mathcal{B}(\mathcal{X})$
$R_{\oplus\varepsilon}(A) = R_{\Gamma_\varepsilon}(A) = R_{F_\varepsilon}(A)$	$(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$: $A \in \mathcal{B}(\mathcal{X})$
$R_{\oplus\varepsilon}(A) = R_{\Gamma_\varepsilon}(A) = R_{F_\varepsilon}(A) = R_{\varepsilon}(A) = R_{\varepsilon)}(A)$	$\mathbb{R}^d$: $A \in \mathcal{L}(\mathcal{X})$ and p_0, p_1 have densities

Table 2.2: Equivalences among adversarial risk formulations for general loss. $R_{\oplus\varepsilon}(w), R_{F_\varepsilon}(w), R_{K_\varepsilon}(w)$ and $R_{\Gamma_\varepsilon}(w)$ denote adversarial risk for a loss function ℓ for a classifier parametrized by $w \in \mathcal{W}$, defined using expected supremum, transport maps, Markov kernels and ∞-Wasserstein metric respectively. $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$ denotes the universal completion of the Borel measure space.

Equivalences in Adversarial Risk	Conditions
$R_{\oplus\varepsilon}(w) = R_{\Gamma_\varepsilon}(w)$	$\mathbb{R}^d$: $\ell((x,y),w)$ Lebesgue measurable in x, or $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$: $\ell((x,y),w)$ upper semi-analytic in x
$R_{\oplus\varepsilon}(w) = R_{\Gamma_\varepsilon}(w) = R_{F_\varepsilon}(w) = R_{K_\varepsilon}(w)$	$(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$: $\ell((x,y),w)$ upper semi-continuous in x

Tables 2.1 and 2.2. In Subsection 2.4.1, we consider general Polish spaces and in Subsection 2.4.2, we consider the Euclidean space.

2.4.1 Results on Polish Spaces via Measurable Selections

We begin by proving the equivalence between the definitions for adversarial risk based on the three notions of set expansions.

Theorem 3. *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ be absolutely continuous with respect to the Lebesgue measure. Let $\varepsilon \geq 0$. Then for any $A \in \mathcal{L}(\mathcal{X})$,*

$$R_{\varepsilon)}(\ell_{0/1}, A) = R_{\oplus\varepsilon}(\ell_{0/1}, A) = R_{\varepsilon}(\ell_{0/1}, A).$$

Proof. By Lemma 2.3.3 $A^\varepsilon \backslash A^{\varepsilon)}$ is porous, and so $\lambda(A^\varepsilon \backslash A^{\varepsilon)}) = 0$. Hence, $\lambda(A^\varepsilon) = \lambda(A^{\oplus\varepsilon})\lambda(A^{\varepsilon)})$. Therefore, the desired result follows from the assumption that p_0 and p_1 are absolutely continuous with respect to the Lebesgue measure. □

We now present a lemma that links the measure of ε-Minkowsi set expansion to the worst case measure over a

W_∞ probability ball of radius ε.

Lemma 2.4.1. *Let $\mu \in \overline{\mathcal{P}}(\mathcal{X})$ and $A \in \mathcal{B}(\mathcal{X})$. Then $\sup_{W_\infty(\mu,\mu')\le\varepsilon} \mu'(A) = \mu(A^{\oplus\varepsilon})$. Moreover, the supremum in the previous equation is achieved by a $\mu^* \in \mathcal{P}(\mathcal{X})$ that is induced from μ via a measurable transport map $\varphi : \mathcal{X} \to \mathcal{X}$ (i.e. $\mu^* = \varphi_\sharp\mu$) satisfying $d(x, \varphi(x)) \le \varepsilon$ for all $x \in \mathcal{X}$.*

The proof of Lemma 2.4.1 is in Appendix B.2.1. A crucial step in the proof of Lemma 2.4.1 is finding a measurable transport map φ such that $\varphi^{-1}(A) = A^{\oplus\varepsilon}$ and $d(x, \varphi(x)) \le \varepsilon$ for all $x \in \mathcal{X}$. In the following theorem, we use Lemma 2.4.1 to establish the equivalence between three different notions of adversarial risk introduced in section 2.2.

Theorem 4. *Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and $A \in \mathcal{B}(\mathcal{X})$. Then $R_{\oplus\varepsilon}(\ell_{0/1}, A) = R_{F_\varepsilon}(\ell_{0/1}, A) = R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$. In addition, the supremum over f_0 and f_1 in $R_{F_\varepsilon}(\ell_{0/1}, A)$ is attained. Similarly, the supremum over p_0' and p_1' in $R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$ is attained.*

Proof. Since $A \in \mathcal{B}(\mathcal{X})$, $A^c \in \mathcal{B}(\mathcal{X})$ and by Lemma 2.3.2, $A^{\oplus\varepsilon}, (A^c)^{\oplus\varepsilon} \in \overline{\mathcal{B}}(\mathcal{X})$. Therefore $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is well-defined. By Lemma 2.4.1, we have

$$\begin{aligned}
R_{\Gamma_\varepsilon}(\ell_{0/1}, A) &= \sup_{\substack{W_\infty(p_0,p_0')\le\varepsilon \\ W_\infty(p_1,p_1')\le\varepsilon}} \frac{T}{T+1} p_0'(A) + \frac{1}{T+1} p_1'((A^c)) \\
&= \frac{T}{T+1} \left(\sup_{W_\infty(p_0,p_0')\le\varepsilon} p_0'(A) \right) + \frac{1}{T+1} \left(\sup_{W_\infty(p_1,p_1')\le\varepsilon} p_1'((A^c)) \right) \\
&= \frac{T}{T+1} p_0(A^{\oplus\varepsilon}) + \frac{1}{T+1} p_1((A^c)^{\oplus\varepsilon}) \\
&= R_{\oplus\varepsilon}(\ell_{0/1}, A).
\end{aligned}$$

By Lemma 2.4.1 again, the supremum over p_0' and p_1' in $R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$ is attained by measures pushed forward from p_0 and p_1 via some measurable maps f_0 and f_1. From this, the remaining assertions of the theorem follow. □

We will now extend the above result to more general loss functions. The following lemma plays a critical role in doing this.

Lemma 2.4.2. *Let $\mu \in \overline{\mathcal{P}}(\mathcal{X})$. Then for any real-valued upper semi-analytic function function $\varphi : \mathcal{X} \to [0, \infty)$,*

$$\sup_{W_\infty(\mu,\mu')\le\varepsilon} \mathbb{E}_{x\sim\mu'}[\varphi(x)] = \mathbb{E}_{x\sim\mu} \left[\sup_{d(x,x')\le\varepsilon} \varphi(x') \right]. \tag{2.18}$$

Moreover, if the function φ is upper semi-continuous, then the supremum on the left hand side in the previous equation is achieved by a $\mu^* \in \overline{\mathcal{P}}(\mathcal{X})$ that is induced from μ via a universally measurable transport map $m : \mathcal{X} \to \mathcal{X}$ (i.e. $\mu^* = m_\sharp \mu$) satisfying $d(x, m(x)) \leq \varepsilon$ for all $x \in \mathcal{X}$.

The proof of Lemma 2.4.2 is in Appendix B.2.1. Using Lemma 2.4.2, we prove the following theorem, which generalizes Theorem 4 to more general loss functions.

Theorem 5. *If the loss function $\ell((\cdot, y), w)$ is upper semi-analytic for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, then $R_{\oplus\varepsilon}(\ell, w) = R_{\Gamma_\varepsilon}(\ell, w)$. If in addition, $\ell((\cdot, y), w)$ is upper semi-continuous for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, then $R_{\oplus\varepsilon}(\ell, w) = R_{F_\varepsilon}(\ell, w) = R_{K_\varepsilon}(\ell, w) = R_{\Gamma_\varepsilon}(\ell, w)$.*

Proof.

$$\begin{aligned} R_{\Gamma_\varepsilon}(\ell, w) &= \sup_{\gamma \in \Gamma_\varepsilon} \mathbb{E}_{(x', y) \sim \rho_y \rho^\gamma_{x'|y}} \left[\ell((x', y), w) \right] \\ &= \mathbb{E}_{(x,y) \sim \rho_y \rho_{x|y}} \left[\sup_{d(x,x') \leq \varepsilon} \ell((x', y), w) \right] \\ &= R_{\oplus\varepsilon}(\ell, w), \end{aligned}$$

where the second inequality follows from Lemma 2.4.2 because of the assumption that $\ell((\cdot, y), w)$ is upper semi-analytic for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$.

With the stronger assumption that $\ell((\cdot, y), w)$ is upper semi-continuous for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, Lemma 2.4.2 shows that for every $y \in \mathcal{Y}$, there exists a universally measurable transport map $m_y : \mathcal{X} \to \mathcal{X}$ satisfying $d(x, m(x)) \leq \varepsilon$ for all $x \in \mathcal{X}$ such that the following holds.

$$\begin{aligned} R_{\Gamma_\varepsilon}(\ell, w) &= \sup_{\gamma \in \Gamma_\varepsilon} \mathbb{E}_{(x', y) \sim \rho_y \rho^\gamma_{x'|y}} \left[\ell((x', y), w) \right] \\ &= \mathbb{E}_{(x,y)) \sim \rho_y \rho_{x|y}} \left[\ell((m_y(x), y), w) \right] \\ &\leq \sup_{F \in F_\varepsilon} \mathbb{E}_{(x,y) \sim \rho} \left[\ell((f_y(x), y), w) \right] \\ &= R_{F_\varepsilon}(\ell, w). \end{aligned}$$

Combining the above inequality with (2.8), we have $R_{\oplus\varepsilon}(\ell, w) = R_{F_\varepsilon}(\ell, w) = R_{K_\varepsilon}(\ell, w) = R_{\Gamma_\varepsilon}(\ell, w)$.

□

2.4.2 Results in $\mathbb{R}^d$ via Submodular Capacities

In this subsection, we establish a connection between adversarial risk and Choquet capacities[19] in $\mathbb{R}^d$. This connection allows us to extend Theorem 4 from Borel sets to the broader class of Lebesgue measurable sets. We will again use this connection for proving minimax theorems and existence of Nash equilibria in Chapter 4. We begin with the following definitions.

Definition 7 (Capacity)**.** A set function $v : \mathcal{B}(\mathcal{X}) \to [0,1]$ is a *capacity* if it satisfies the following conditions: (1) $v(\varnothing) = 0$ and $v(\mathcal{X}) = 1$; (2) For $A, B \in \mathcal{B}(\mathcal{X}), A \subseteq B \implies v(A) \leq v(B)$; (3) $A_n \uparrow A \implies v(A_n) \uparrow v(A)$; and (4) $F_n \downarrow F, F_n$ closed $\implies v(F_n) \downarrow v(F)$.

Definition 8 (2-Alternating Capacity)**.** A capacity v defined on the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is called 2-alternating if $v(A \cup B) + v(A \cap B) \leq v(A) + v(B)$ for all $A, B \in \mathcal{B}(\mathcal{X})$.

For any compact set of probability measures $\Xi \subseteq \mathcal{P}(\mathcal{X})$, the upper probability defined as $v(A) = \sup_{\mu \in \Xi} \mu(A)$ is a capacity[46]. The upper probability of ε-neighborhoods of a $\mu \in \mathcal{P}(\mathcal{X})$ defined using either the total variation metric or the Levy-Prokhorov metric can be shown to be a 2-alternating capacity[46]. The following lemma shows that $A \mapsto \mu(A^{\oplus \varepsilon})$ is a 2-alternating capacity under some conditions.

Lemma 2.4.3. *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $\mu \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \geq 0$. Define a set function v on $\mathcal{X}$ such that for any $A \in \mathcal{L}(\mathcal{X})$, $v(A) := \mu(A^{\oplus \varepsilon})$. Then v is a 2-alternating capacity.*

The proof of Lemma 2.4.3 is included in Appendix B.2.2.

Now we relate the capacity defined in Lemma 2.4.3 to the W_∞ metric. Since the ε-neighborhood of a $\mu \in \mathcal{P}(\mathcal{X})$ in W_∞ metric is a compact set of probability measures[98], the upper probability over this W_∞ ε-ball is a capacity. The following lemma shows that it is a 2-alternating capacity, and identifies it with the capacity defined in Lemma 2.4.3.

Lemma 2.4.4. *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $\mu \in \overline{\mathcal{P}}(\mathcal{X})$. Then for any $A \in \mathcal{L}(\mathcal{X})$, $\sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mu'(A) = \mu(A^{\oplus \varepsilon})$. Moreover, the supremum in the previous equation is attained.*

The proof of Lemma 2.4.4 is included in Appendix B.2.2. Lemma 2.4.4 plays a similar role to Lemma 2.4.1 in proving the following equivalence between adversarial robustness and W_∞ robustness.

Theorem 6. *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \geq 0$. Then for any $A \in \mathcal{L}(\mathcal{X})$, $R_{\oplus\varepsilon}(\ell_{0/1}, A) = R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$, and the supremum over p_0' and p_1' in $R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$ is attained.*

Proof. Observe that

$$\begin{aligned} R_{\Gamma_\varepsilon}(\ell_{0/1}, A) &= \frac{T}{T+1}\left(\sup_{W_\infty(p_0, p_0')\leq\varepsilon} p_0'(A)\right) + \frac{1}{T+1}\left(\sup_{W_\infty(p_1, p_1')\leq\varepsilon} p_1'((A^c))\right) \\ &\overset{(*)}{=} \frac{T}{T+1} p_0(A^{\oplus\varepsilon}) + \frac{1}{T+1} p_1((A^c)^{\oplus\varepsilon}) \\ &= R_{\oplus\varepsilon}(\ell_{0/1}, A), \end{aligned}$$

where $(*)$ follows from Lemma 2.4.4. By Lemma 2.4.4 again, the supremum over p_0' and p_1' in $R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$ is attained. □

Unlike Theorem 4, Theorem 6 does not show the equivalence of $R_{F_\varepsilon}(\ell_{0/1}, A)$ with the other definitions under the relaxed assumption of $A \in \mathcal{L}(\mathcal{X})$. This is because Lemma 2.4.4 does not provide a push-forward map φ such that $\mu^* = \varphi_{\sharp\mu}$ with μ^* attaining the supremum over the W_∞ ball.

Maturity, the way I understand it, is knowing what your limitations are.

Kurt Vonnegut, Cat's Cradle

3 Fundamental Limits

3.1 Binary Classification: Optimal Adversarial Risk via Optimal Transport

In this section, we present one of the main results of this thesis that links the D_ε cost to the optimal adversarial risk in binary classification. We begin this section by introducing the Kantorovich duality for optimal transport in section 3.1.1. In Section 3.1.2, we present the result for the case of equal priors, where the proof relies heavily on Strassen's theorem. In Section 3.1.3, we extend the result to the case of unequal priors using a notion of optimal transport between measures of unequal mass.

3.1.1 Preliminaries on Duality in Optimal Transport

Recall from Chapter 2 that the optimal transport cost between two probability measures μ and ν under a cost function $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is defined as,

$$\mathcal{T}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x, x')d\pi(x, x').$$

The Kantorovich duality theorem[88] states that the above minimization problem is equivalent to the following maximization problem.

$$\mathcal{T}_c(\mu,\nu) = \sup_{\varphi(x)+\psi(y)\leq c(x,y)} \int \varphi d\mu + \int \psi d\nu,$$

where $\varphi, \psi : \mathcal{X} \to \mathbb{R}$ are continuous and bounded functions on $\mathcal{X}$ satisfying the constraint, $\varphi(x) + \psi(y) \leq c(x,y)$ for all $x, y \in \mathcal{X}$.

An important special case of the Kantorovich duality theorem for $\{0,1\}-$valued cost functions is the Strassen's theorem. Before we state the Strassen's theorem, we introduct the following definition for an optimal transport cost involving a $\{0,1\}-$valued cost function.

Definition 9 (D_ε cost)**.** Let $c_\varepsilon : \mathcal{X}^2 \to \{0,1\}$ be such that $c_\varepsilon(x,x') = 1\{(x,x') \in \mathcal{X}\times\mathcal{X} : d(x,x') > 2\varepsilon\}$. Then for $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $\varepsilon \geq 0$, the D_ε transport cost between μ and ν is defined as,

$$D_\varepsilon(\mu,\nu) := \mathcal{T}_{c_\varepsilon}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{P}_{(x,x')\sim\pi}(d(x,x') > 2\varepsilon). \tag{3.1}$$

Remark 1. *For $\varepsilon = 0$, the optimal cost is equivalent to the total variation distance, i.e., $D_0(\mu,\nu) = D_{TV}(\mu,\nu)$. For $\varepsilon > 0$, this cost does not define a metric over the space of distributions. This is because $D_\varepsilon(\mu,\nu) = 0$ 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 $D_\varepsilon(\mu_1,\mu_3) = 1 > 0 = D_\varepsilon(\mu_1,\mu_2) + D_\varepsilon(\mu_2,\mu_3)$.*

Stassen's theorem stated below gives a duality formula for D_ε cost that based on the measures of expansions of closed sets.

Proposition 1. *Let $c_\varepsilon : \mathcal{X}^2 \to \{0,1\}$ be such that $c_\varepsilon(x,x') = 1\{(x,x') \in \mathcal{X}\times\mathcal{X} : d(x,x') > 2\varepsilon\}$. Then for $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $\varepsilon \geq 0$,*

$$D_\varepsilon(\mu,\nu) = \sup_{A\in\mathcal{B}(\mathcal{X})} \mu(A) - \nu(A^{2\varepsilon}). \tag{3.2}$$

3.1.2 The case of Equal Priors: Balanced Transport

The following theorem relates the optimal adversarial risk with D_ε cost for the case of equal priors.

Theorem 7. *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 $p_0 \in \overline{\mathcal{P}}(\mathcal{X})$ (for label 0) and $p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ (for label 0). We consider a set of binary classifiers of the form $1\{x \in A\}$, 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 $\ell((x, y), A) = 1\{x \in A, y = 0\} + 1\{x \notin A, y = 1\}$. Then the optimal adversarial risk with the data perturbing adversary of budget $\varepsilon \geq 0$ is given by*

$$\inf_{\substack{A \in \mathcal{B}(\mathcal{X}) \\ A \text{ closed}}} = \frac{1}{2}\left[1 - D_\varepsilon(p_0, p_1)\right]. \tag{3.3}$$

Instantiating Theorem 7 for $\varepsilon = 0$, 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 3.1.1. *Under the setup considered in Theorem 7, we have the following bound for $p \geq 1$:*

$$R_\varepsilon^* \geq \frac{1}{2}\left[1 - \left(\frac{W_p(p_0, p_1)}{2\varepsilon}\right)^p\right]. \tag{3.4}$$

Our next result identifies a necessary and sufficient condition for $D_\varepsilon(\mu, \nu) = 0$ 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 8. *Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Then $D_\varepsilon(\mu, \nu) = 0$ if and only if $W_\infty(\mu, \nu) \leq 2\varepsilon$.*

We now include the complete proof of Theorem 7 below, as the proof provides intuition on how Strassen's theorem is crucial to proving equality (3.3). Further, our proof is much simpler compared to the proof of a similar statement that appears in the contemporary work of Bhagoji et al.[7]

Proof of Theorem 7. The optimal adversarial risk over the hypothesis class of closed sets is given by

$$\inf_{A \text{ closed}} \frac{1}{2}\left(p_0(A^{\oplus\varepsilon}) + p_1\left((A^c)^{\oplus\varepsilon}\right)\right) = \frac{1}{2}\left(1 - \sup_{A \text{ closed}}\left\{p_1\left(A^{\ominus\varepsilon}\right) - p_0(A^{\oplus\varepsilon})\right\}\right).$$

The main idea of our proof is to leverage Strassen's theorem (Proposition 1), which states that

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

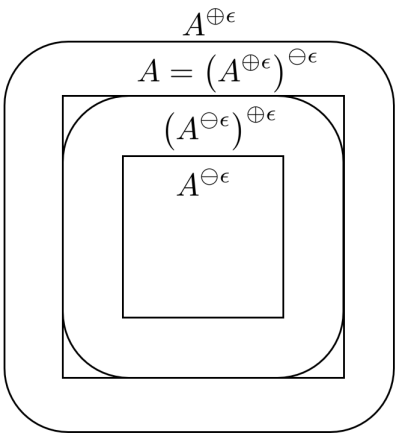


Figure 3.1: Illustration of $A, A^{\oplus\varepsilon}, A^{\ominus\varepsilon}, (A^{\oplus}\varepsilon)^{\ominus\varepsilon}$, and $(A^{\ominus\varepsilon})^{\oplus\varepsilon}$ for a closed square in $(\mathbb{R}^2, \|\cdot\|_2)$. Observe that $(A^{\ominus\varepsilon})^{\oplus\varepsilon} \subseteq A$ and $A \subseteq (A^{\oplus}\varepsilon)^{\ominus\varepsilon}$.

To prove the desired equality, notice that it is enough to prove that

$$\sup_{A \text{ closed}} p_1(A^{\ominus\varepsilon}) - p_0(A^{\oplus\varepsilon}) = \sup_{A \text{ closed}} p_1(A) - p_0(A^{2\varepsilon}). \tag{3.5}$$

We have the sequence of inequalities

$$\sup_{A \text{ closed}} p_1(A) - p_0(A^{2\varepsilon}) \overset{(a)}{\geq} \sup_{A \text{ closed}} p_1(A^{\ominus\varepsilon}) - p_0((A^{\ominus\varepsilon})^{2\varepsilon}) \overset{(b)}{\geq} \sup_{A \text{ closed}} p_1(A^{\ominus\varepsilon}) - p_0(A^{\oplus\varepsilon}).$$

Here, (a) follows because $A^{\ominus\varepsilon}$ is contained in the set of all closed sets by Lemma A.0.4. Inequality (b) follows by the equivalence $(A^{\ominus\varepsilon})^{2\varepsilon} = (A^{\ominus\varepsilon})^{\oplus 2\varepsilon}$ from Lemma A.0.5, and Lemma A.0.6 since $(A^{\ominus\varepsilon})^{2\varepsilon} = [(A^{\ominus\varepsilon})^{\oplus\varepsilon}]^{\oplus\varepsilon} \subseteq A^{\oplus\varepsilon}$, and so $p_0((A^{\ominus\varepsilon})^{2\varepsilon}) \leq p_0(A^{\oplus\varepsilon})$.

For the other direction, notice that

$$\sup_{A \text{ closed}} p_1(A^{\ominus\varepsilon}) - p_0(A^{\oplus\varepsilon}) \overset{(a)}{\geq} \sup_{A \text{ closed}} p_1((A^{\oplus\varepsilon})^{\ominus\varepsilon}) - p_0((A^{\oplus\varepsilon})^{\varepsilon}) \overset{(b)}{\geq} \sup_{A \text{ closed}} p_1(A) - p_0(A^{2\varepsilon}).$$

Here, (a) follows because $A^{\oplus\varepsilon}$ is a closed set according to Lemma A.0.4. To see (b), first note that using Lemma A.0.5, $(A^{\oplus\varepsilon})^{\varepsilon} = (A^{\oplus\varepsilon})^{\oplus\varepsilon} = A^{\oplus 2\varepsilon} = A^{2\varepsilon}$. Thus, $p_0((A^{\oplus\varepsilon})^{\varepsilon}) = p_0(A^{2\varepsilon})$. Moreover, Lemma A.0.6 states that $A \subseteq (A^{\varepsilon})^{\ominus\varepsilon}$, and so $p_1(A) \leq p_1((A^{\oplus\varepsilon})^{\ominus\varepsilon})$. This completes the proof. □

Remark 2. *A similar result to Theorem 7 appeared in the contemporary work of Bhagoji et al.[7] A key difference is that the proof in Bhagoji et al.[7] was established for a larger hypothesis class of measurable sets $A \in \mathcal{B}(\mathcal{X})$; i.e., the*

following equality was established:

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

It is not hard to check that A^{ε} is closed for any $A \in \mathcal{B}(\mathcal{X})$, and so

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

We may restrict to the smaller hypothesis class of closed sets A and use the result in[7] *to obtain an inequality*

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

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

3.1.3 The case of Unequal Priors: Unbalanced Transport

In this subsection, we present a theorem analogous to Theorem 7 for the case of unequal priors in binary classification. We show that the optimal adversarial risk in this case is determined by an *unbalanced* optimal transport cost between two finite measures that are related to the data-generating probability measures. We begin by introducing unbalanced optimal transport.

Definition 10 (Coupling between finite measures)**.** Let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ with $\mu(\mathcal{X}) \leq \nu(\mathcal{X})$. A coupling between μ and ν is a measure $\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{X})$ such that for any measurable set $A \subseteq \mathcal{X}$, $\pi(A \times \mathcal{X}) = \mu(A)$ and $\pi(\mathcal{X} \times A) \leq \nu(A)$. The set $\Pi(\mu, \nu)$ denotes the set of all couplings between μ and ν.

Definition 11 (Optimal transport cost between finite measures)**.** Let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ be such that $\mu(\mathcal{X}) \leq \nu(\mathcal{X})$. Let $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ denote a cost function. Then the optimal transport cost between μ and ν under the cost c is defined as,

$$T_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, x') d\pi(x, x'). \tag{3.6}$$

Recall that for $\mu, \nu \in \mathcal{M}(\mathcal{X})$, we say that ν (i.e. $\mu \preceq \nu$) *dominates* μ if and only if for all measurable sets $A \in \mathcal{B}(\mathcal{X}), \mu(A) \leq \nu(A)$.

In the following theorem, we generalize Proposition 1 to the case when μ, ν are finite measures, in place of probability measures.

Theorem 9 (Generalized Strassen's theorem)**.** *Let $\mu, \nu \in \mathcal{M}(\mathcal{X})$ be such that $0 < M = \mu(\mathcal{X}) \leq \nu(\mathcal{X})$. Let $\varepsilon > 0$. Let $c_\varepsilon : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ be such that $c_\varepsilon(x, x') = 1\{(x, x') \in \mathcal{X} \times \mathcal{X} : d(x, x') > 2\varepsilon\}$. Then,*

$$\sup_{A \in \mathcal{B}(\mathcal{X})} \mu(A) - \nu(A^{2\varepsilon}) = T_{c_\varepsilon}(\mu, \nu) = M \inf_{\substack{\nu' \in \mathcal{P}(\mathcal{X}): \\ \nu' \preceq \nu/M}} D_\varepsilon\left(\mu/M, \nu'\right). \tag{3.7}$$

T_{c_ε} in (3.7) is a generalization of D_ε to finite positive measures. When $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $T_{c_\varepsilon}(\mu, \nu) = D_\varepsilon(\mu, \nu)$ and (3.7) of Theorem 9 reduces to (3.2) of Proposition 1. A central idea in our proof of Theorem 9 is the duality in linear programming. Using strong duality, we first show the result in (3.7) for discrete measures on a finite support. We then apply the discrete result on a sequence of measures supported on a countable dense subset of the Polish space $\mathcal{X}$. Using the tightness of finite measures on $\mathcal{X}$, we construct an optimal coupling that achieves the cost $T_{c_\varepsilon}(\mu, \nu)$ in (3.7). We then show that the constructed coupling satisfies (3.7). This proof strategy is adapted from the works of Dudley [25] and Schay [70].

Like the Strassen's theorem, generalized Strassen's theorem involves closed set expansions. The following lemma allows us to switch to Minkowski set expansions.

Lemma 3.1.1. *Let $\mu, \nu \in \overline{\mathcal{M}}(\mathcal{X})$ and let $\varepsilon \geq 0$. Then,*

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

Moreover, the supremum on the right hand side of the above equality can be replaced by a supremum over closed sets.

The proof of Lemma 3.1.1 is contained in Appendix C. Using Lemma 3.1.1 and the generalized Strassen's theorem, we show the following result on optimal adversarial risk for unequal priors, generalizing Theorem 7.

Theorem 10. *Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \geq 0$. Then,*

$$\inf_{A \in \mathcal{B}(\mathcal{X})} R_{\oplus\varepsilon}(\ell_{0/1}, A) = \frac{1}{T+1}\left[1 - \inf_{q \in \mathcal{P}(\mathcal{X}): q \preceq T p_0} D_\varepsilon(q, p_1)\right]. \tag{3.8}$$

Moreover, the infimum on the left hand side can be replaced by an infimum over closed sets.

Proof.

$$\begin{aligned}
\inf_{A\in\mathcal{B}(\mathcal{X})} R_{\oplus\varepsilon}(\ell_{0/1}, A) &= \inf_{A\in\mathcal{B}(\mathcal{X})} \frac{1}{T+1}\left[Tp_0(A^{\oplus\varepsilon}) + p_1((A^c)^{\oplus\varepsilon})\right] \\
&= \frac{1}{T+1}\left[1 - \sup_{A\in\mathcal{B}(\mathcal{X})} \left(p_1(A^{\ominus\varepsilon}) - Tp_0(A^{\oplus\varepsilon})\right)\right] \\
&\overset{(i)}{=} \frac{1}{T+1}\left[1 - \sup_{A\in\mathcal{B}(\mathcal{X})} \left(p_1(A) - Tp_0(A^{2\varepsilon})\right)\right] \\
&\overset{(ii)}{=} \frac{1}{T+1}\left[1 - \inf_{\substack{q\in\mathcal{P}(\mathcal{X}):\\ q\preceq Tp_0}} D_\varepsilon(q, p_1)\right],
\end{aligned}$$

where (i) follows from Lemma 3.1.1 and (ii) follows from Theorem 9. □

Theorem 10 extends Theorem 7 in two ways: (1) the infimum is taken over all sets for which $R_{\oplus\varepsilon}(\ell_{0/1}, A)$ is well-defined, instead of restricting to closed sets, and (2) the priors on both labels can be unequal. We also note that for $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$, (3.8) holds with the infimum on the left hand side taken over all $A \in \mathcal{L}(\mathcal{X})$.

3.2 Binary Classification: 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 7 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

$$D_\varepsilon(\mu, \nu) \leq \mathbb{E}_{(x,x')\sim\pi} c_\varepsilon(x, x').$$

(3) Identify a closed set A and compute a lower bound using

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

(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.

3.2.1 Preliminaries on 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 3.2.1. *[Theorem 2.5 in [69]] 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 3.2.1 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 3.2.2 (Theorem 2.9 in [69])**.** *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 3.2.1 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 3.2.1 to present a more concrete condition than Theorem 8 for checking when $D_\varepsilon(\mu,\nu) = 0$ for measures over $\mathbb{R}$.

Theorem 11. *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 $D_\varepsilon(\mu,\nu) = 0$ if and only if $\|F^{-1} - G^{-1}\|_\infty \leq 2\varepsilon$.*

Proof. Consider the monotone transport map from μ to ν given by $T(x) = G^{-1}(F(x))$ as in Lemma 3.2.1. We shall show that this map satisfies $|T(x) - x| \leq 2\varepsilon$ for all $x \in \mathbb{R}$, and so the optimal transport cost D_ε must be 0. To see this, note that

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

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

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

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

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

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

Thus, we have

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

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

The above argument shows that monotone transport maps are optimal when $D_\varepsilon = 0$. But monotone maps are not always optimal for the cost function $c_\varepsilon(\cdot,\cdot)$. Consider for example the two measures $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$, and $\varepsilon = 0.1$. The monotone map in this case is $T(x) = x + 1$, which gives unit cost of transportation. However, Theorem 12 shows that the optimal transport cost in this example is strictly smaller than 1.

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

Corollary 3.2.1. *Let μ and ν be as in Theorem 11. Suppose that for every $x \in \mathbb{R}$, we have $F(x) \geq G(x)$ and $F(x) \leq G(x + 2\varepsilon)$. Then $D_\varepsilon(\mu,\nu) = 0$.*

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

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

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

Theorem 11 may also be applied to finite measures μ, ν with $\mu(\mathbb{R}) = \nu(\mathbb{R}) = U < \infty$ with simple scaling.

Let μ and ν be finite measures of unequal mass on the real line such that $0 < M = \mu(\mathbb{R}) \leq \nu(\mathbb{R})$. For the purpose of keeping the notation concise, we will use $D_\varepsilon(\mu,\nu)$ to denote the unbalanced optimal transport cost between μ and ν for the cost function c_ε that appears in the generalized Strassen's theorem (Theorem 9). That is,

$$D_\varepsilon(\mu,\nu) := M \inf_{\substack{\nu' \in \mathcal{P}(\mathcal{X}): \\ \nu' \preceq \nu/M}} D_\varepsilon\left(\mu/M, \nu'\right). \tag{3.9}$$

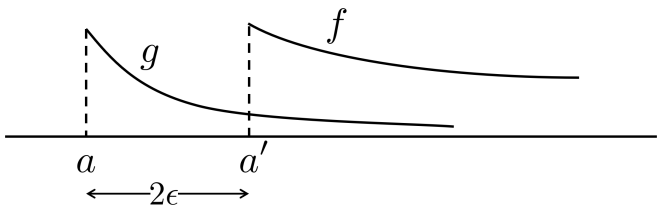


Figure 3.2: Figure illustrating the conditions in Lemma 3.2.3.

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

Lemma 3.2.3. *Let μ and ν be finite 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. 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 $[a+2\varepsilon,+\infty) =: [a',+\infty)$.*
2. *For all $x \in \mathbb{R}$, we have $g(x) \le f(x+2\varepsilon)$.*

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

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

Lemma 3.2.4. *Let μ and ν be as in Lemma 3.2.3. Assume that $\mu(\mathbb{R}) = \nu(\mathbb{R}) = U$. Suppose the following conditions hold (see Figure 3.3 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 $[a',b] := [a+2\varepsilon,b]$.*
2. *There exists $t \in [a,b]$ such that $f(x) \ge g(x)$ for $x \in [a,t)$, and $f(x) \le g(x)$ for $x \in (t,b]$.*
3. *Let $\tilde{g}(x) = g(x+2\varepsilon)$. Note that the support $\tilde{g}$ is within $[a,b-2\varepsilon]$. There exists $\tilde{t} \in [a,b-2\varepsilon]$ such that $f(x) \le \tilde{g}(x)$ for $x \in [a,\tilde{t})$, and $f(x) \ge \tilde{g}(x)$ for $x \in (\tilde{t},b-2\varepsilon]$.*

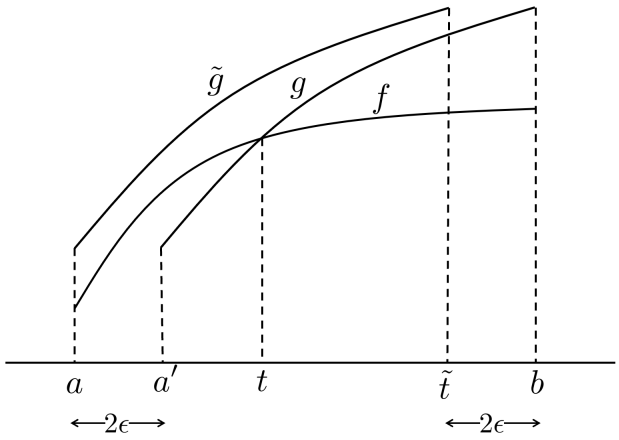


Figure 3.3: Figure illustrating the conditions in Lemma 3.2.4. Note that in general $\tilde{t}$ need not be equal to $b - 2\varepsilon$ as shown in the figure.

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

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 $F(x) \leq G(x + 2\varepsilon)$. 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 $F(x) \leq G(x + 2\varepsilon)$. Applying Corollary 3.2.1 concludes the proof. □

3.2.2 Gaussian distributions with identical variances

Theorem 12. *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 $\varepsilon \geq \frac{|\mu_0 - \mu_1|}{2}$, the optimal robust risk is $1/2$. A constant classifier achieves this risk.*
2. *If $\varepsilon < \frac{|\mu_0 - \mu_1|}{2}$, the optimal classifier satisfies $A = \left[\frac{\mu_1 + \mu_0}{2}, +\infty\right]$, where A is the region where the classifier declares label 1. The optimal risk in this case is $\int_{\frac{\mu_1+\mu_0}{2} - \varepsilon}^{\infty} p_0(x)dx = Q\left(\frac{\frac{\mu_1 - \mu_0}{2} - \varepsilon}{\sigma}\right)$.*

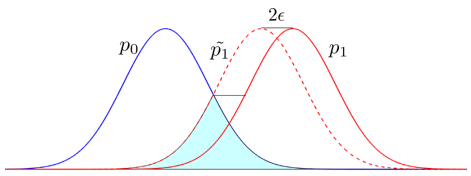


Figure 3.4: Optimal coupling for two Gaussians with identical variances. The shaded region within p_0 is translated by 2ε to p_1, whereas the remaining is mass in p_0 is moved at a cost of 1 per unit mass.

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 optimal adversarially 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 $\varepsilon \geq \frac{\mu_1 - \mu_0}{2}$, the transport map T defined by $T(x) = x + (\mu_1 - \mu_0)$ transports p_0 to p_1. Moreover, this coupling satisfies $|T(x) - x| = \mu_1 - \mu_0 \leq 2\varepsilon$. Thus, the optimal transport cost for this coupling is 0, and therefore so is $D_\varepsilon(p_0, p_1)$. This gives the following lower bound on $R^*_{\oplus\varepsilon} := \inf_{A \in \mathcal{B}(\mathcal{X})} R_\varepsilon(\ell_{0/1}, A)$.

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

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

For part (2), we consider the following strategy for transporting the mass from p_0 to p_1. As shown in Figure 3.4, consider the distribution $\tilde{p}_1$ obtained by shifting p_1 to the left by 2ε. That is, $\tilde{p}_1(x) = p_1(x + 2\varepsilon)$. 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, $q(x - 2\varepsilon) \leq p_1(x)$ 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 $T(x) = x + 2\varepsilon$. It is easily verified that the area under $q(x)$ equals $2Q\left(\frac{\frac{\mu_1 - \mu_0}{2} - \varepsilon}{\sigma}\right)$, and so the total cost of transporting p_0 to p_1 is at most $1 - 2Q\left(\frac{\frac{\mu_1 - \mu_0}{2} - \varepsilon}{\sigma}\right)$. Plugging this into the lower bound, we see that

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

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

Theorem 13. *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 $\varepsilon \geq \frac{||\mu_0 - \mu_1||_2}{2}$, the optimal robust risk is $1/2$. A constant classifier achieves this risk.*
2. *If $\varepsilon < \frac{||\mu_0 - \mu_1||_2}{2}$, the optimal classifier is given by the following halfspace:*

$$A = \left\{ x \,:\, (\mu_1 - \mu_0)\left(x - \frac{\mu_0 + \mu_1}{2}\right) \geq 0 \right\}. \tag{3.10}$$

Remark 3. *Bhagoji et al.[7] also explore optimal classifiers for multivariate normal distributions. In fact, they show a more general version of our Theorems 12 and 13 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 12 in a different way by considering various interesting examples of univariate distributions and identifying optimal couplings for these.

3.2.3 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 $D_\varepsilon(\mu_i, \nu_i) = 0$. (Here, we are using the definition of D_ε for finite measures from (3.9).) The transport maps used in these K problems may be arbitrary; however, we shall often use versions of the monotone optimal transport map[88].

Our next lemma is specific to Gaussian pdfs:

Lemma 3.2.5. *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 14. *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 $f(m+\varepsilon) = g(m-\varepsilon)$. Let $A = (-\infty, -m] \cup [m, +\infty)$. Then the optimal transport cost between μ and ν is given by*

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

The corresponding robust risk is

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

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 3.5 for an illustration.) The existence of a $m > 0$ such that $f(m+\varepsilon) = g(m-\varepsilon)$ is guaranteed by Lemma 3.2.5. Consider $r \in (0, m-\varepsilon)$ whose value will be provided later. First, we partition $\mathbb{R}$ into the five regions for μ and ν, as shown in Table 3.1. For μ, these partitions are $(-\infty, -m-\varepsilon]$, $(-m-\varepsilon, -r]$, $(-r, +r)$, $[r, m+\varepsilon)$, and $[m+\varepsilon, \infty)$. 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 $(-\infty, -m+\varepsilon]$, $(-m+\varepsilon, -r]$, $(-r, r)$, $[r, m-\varepsilon)$, and $[m-\varepsilon, +\infty)$. Call ν restricted to these intervals ν_{--}, ν_-, ν_0, ν_+, and ν_{++}, respectively.

μ_{--}	$(-\infty, -m-\varepsilon]$
μ_{-}	$(-m-\varepsilon, -r]$
μ_0	$(-r, +r)$
μ_{+}	$[r, m+\varepsilon)$
μ_{++}	$[m+\varepsilon, \infty)$

ν_{--}	$(-\infty, -m+\varepsilon]$
ν_{-}	$(-m+\varepsilon, -r]$
ν_0	$(-r, +r)$
ν_{+}	$[r, m-\varepsilon)$
ν_{++}	$[m-\varepsilon, \infty)$

Table 3.1: The real line is partitioned into five regions for μ and ν, as shown in the table.

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 $D_\varepsilon(\mu_*, \nu_*) = 0$, where $*$ ranges over all possible subscripts in $\{--, -, 0, +, ++\}$. Note that these measures do not necessarily have identical masses, and thus 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 $D_\varepsilon(\mu_*, \nu_*) = 0$, it does not mean $D_\varepsilon(\mu, \nu) = 0$.

Consider μ_{++} and ν_{++}. We have $f(m+\varepsilon) = g(m-\varepsilon)$ by the choice of m. We argue that for any $t \geq 0$, we must have $f(m+\varepsilon+t) \geq g(m-\varepsilon+t)$. This is because any two Gaussian pdfs can intersect in at most two points. By Lemma 3.2.5, the ε-shifted Gaussian pdfs $f(x+\varepsilon)$ and $g(x-\varepsilon)$ 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 $f(m+\varepsilon+t) \geq g(m-\varepsilon+t)$ for all $t \geq 0$. By Lemma 3.2.3, we can now conclude $D_\varepsilon(\mu_{++}, \nu_{++}) = 0$. A similar argument also shows $D_\varepsilon(\mu_{--}, \nu_{--}) = 0$.

Before we consider μ_{-} and ν_{-}, we first define r as follows: Pick $r > 0$ such that $\mu([-m-\varepsilon, -r)) = \nu([-m+\varepsilon, -r))$. To see that such an r must exist, consider the functions $a(t) := \mu([-m-\varepsilon, t))$ and $b(t) := \nu([-m+\varepsilon, t))$ as t ranges over $(-m+\varepsilon, 0)$. When $t = -m+\varepsilon$, 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 $t_0 \in (-m+\varepsilon, 0)$ 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 $[-m-\varepsilon, -r)$ and $g(\cdot)$ restricted to $[-m+\varepsilon, -r)$ 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 3.2.4. Since the supports of f_- and g_- are $[-m-\varepsilon, -r)$ and $[-m+\varepsilon, -r)$, condition (1) is immediately verified. To check condition (2), we break up the interval $[-m-\varepsilon, -r)$ into two parts: $[-m-\varepsilon, -s)$ and $[-s, -r)$, where s is such that $f(-s) = g(-s)$. Observe that $f_- \geq g_-$ on $[-m-\varepsilon, -s)$, whereas $f_- \leq g_-$ on $[-s, -r)$. This shows that condition (2) is satisfied. We have $g_-(-m+\varepsilon) = f_-(-m-\varepsilon)$. Again, using Lemma 3.2.5 the 2ε-shifted Gaussian pdf $f(x-2\varepsilon)$ and $g(x)$ have $-m+\varepsilon$ as their left intersection point, and the right intersection point is to the right of 0. Thus, we have $f(x-2\varepsilon) \leq g(x)$ for all

$x \in [-m+\varepsilon, 0] \supseteq [-m+\varepsilon, r)$. Using this domination, we conclude that $f_- \leq \tilde{g}_-$ in the interval $[-m-\varepsilon, -r-2\varepsilon)$ and $f_- \geq g_- = 0$ in the interval $(-r-2\varepsilon, -r]$, and so condition (3) is satisfied. Applying Lemma 3.2.4, we conclude $D_\varepsilon(\mu_-, \nu_-) = 0$. An essentially identical argument may be used to show $D_\varepsilon(\mu_+, \nu_+) = 0$. The minor difference being that r is chosen to satisfy $\mu([r, m+\varepsilon)) = \nu([r, m-\varepsilon))$, and the mirror image of Lemma 3.2.4 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

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

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

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

The lower and upper bounds match and this concludes the proof. The optimal adversarial risk $R^*_{\oplus\varepsilon}$ is given by Theorem 7. The optimal adversarial risk of the classifier that declares label 1 on the set A is easily seen to be $R^*_{\oplus\varepsilon}$. □

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 15. *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 $f(-m_1 - \varepsilon) = g(-m_1 + \varepsilon)$ and $f(m_2 + \varepsilon) = g(m_2 - \varepsilon)$. Let $A = (-\infty, -m_1] \cup [m_2, \infty)$. Then the optimal transport cost between μ and ν is given by*

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

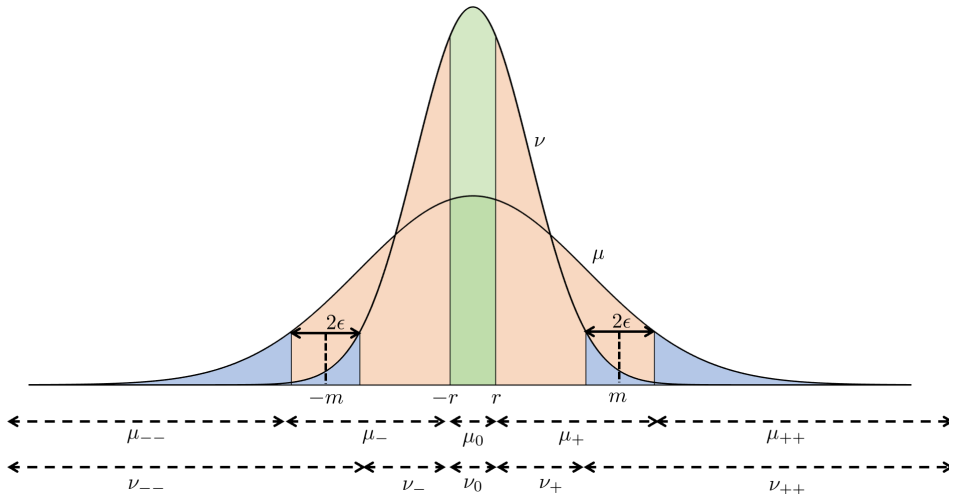


Figure 3.5: Optimal transport coupling for centered Gaussian distributions μ and ν. As in the proof of Theorem 14, we divide the real line into five regions. The transport plan from μ to ν consists of five maps transporting $\mu_{--} \to \nu_{--}$ (blue regions to the left), $\mu_- \to \nu_-$ (orange 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).

Consequently, the optimal adversarial risk is given by

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

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 3.2.5. As in the proof of Theorem 14, we shall divide the real line into five regions as shown in Table 3.2 where we define r_1 and r_2 shortly. Using an identical strategy as in Theorem 14, we conclude $D_\varepsilon(\mu_{--}, \nu_{--}) = D_\varepsilon(\mu_{++}, \nu_{++}) = 0$. Define r_1 as the leftmost point where $\mu([-m_1 - \varepsilon, r_1)) = \nu([-m_1 + \varepsilon, r_1))$. Similarly, define r_2 to be the rightmost point such that $\mu([r_2, m_2 + \varepsilon)) = \nu([r_2, m_2 - \varepsilon))$. We shall now prove $D_\varepsilon(\mu_-, \nu_-) = 0$ by using Lemma 3.2.4. Verifying conditions (1) and (2) is exactly as in that of Theorem 14. The novel component of this proof is verifying condition (3), since the domination used in the proof of Theorem 14 does not work in this case due to the asymmetry. Consider the pdfs $f_-(x)$ and $g_-(x + 2\varepsilon)$. 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 $-m_1 - \varepsilon$ by the choice of m_1, so there can be at most one other point of intersection in the interval $[-m_1 - \varepsilon, -r_1 - 2\varepsilon]$. 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

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

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

Table 3.2: The real line is partitioned into five regions for μ and ν as shown in the table.

case, $f_- \leq \tilde{g}_-$ in $[-m_1 - \varepsilon, \tilde{t})$, and $f_- \geq g_-$ in $(\tilde{t}, -r_1]$. If there is no point of intersection, then $f_- \leq \tilde{g}_-$ in $[-m_1 - \varepsilon, -r_1 - 2\varepsilon)$, and $f_- \geq g_- = 0$ in $(-r_1 - 2\varepsilon, -r_1]$. This verifies condition (3). Using Lemma 3.2.4, we conclude $D_\varepsilon(\mu_-, \nu_-) = 0$. An identical approach gives $D_\varepsilon(\mu_+, \nu_+) = 0$. Since $f(x) \leq g(x)$ for all points in the interval $(-r_1, r_2)$, the identity map may be used to conclude $D_\varepsilon(\mu_0, \nu_0) = 0$.

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

$$\begin{aligned} D_\varepsilon(\mu, \nu) &\leq 1 - \left[\min(\mu_{--}, \nu_{--}) + \min(\mu_-, \nu_-) + \min(\mu_0, \nu_0) + \min(\mu_+, \nu_+) + \min(\mu_{++}, \nu_{++})\right] \\ &= 1 - \left[\nu_{--} + \mu_- + \mu_0 + \mu_+ + \nu_{++}\right] \\ &= 1 - \mu([-m_1 - \varepsilon, m_2 + \varepsilon]) - \nu((-\infty, -m_1 + \varepsilon)) - \nu([m_2 - \varepsilon, \infty)) \\ &= \mu(A^{\ominus\varepsilon}) - \nu(A^{\oplus\varepsilon}), \end{aligned}$$

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 14. □

3.2.4 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 16 (Uniform distributions)**.** *Let μ and ν be uniform measures on closed intervals I and J respectively. Without loss of generality, we assume $|I| \leq |J|$. Then the optimal robust risk is $\nu(I^{2\varepsilon})$ and the optimal classifier is given by $A = I^\varepsilon$.*

The proof of Theorem 16 is in Appendix C.2

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 3.2.5, but is specific to symmetric triangular distributions.

Lemma 3.2.6. *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 $\delta_2 - \delta_1 < |m_1 - m_2| \leq \delta_2 + \delta_1$, then the equation $f(x) - g(x) = 0$ has exactly one solution u on the support of μ. Further, $u \geq m_1$ if and only if $m_1 \leq m_2$.*
3. *If $|m_1 - m_2| \leq \delta_2 - \delta_1$, 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 $m_1 \leq m_2$, as case of $m_1 \geq m_2$ 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 $\delta_2 - \delta_1 < |m_1 - m_2| \leq \delta_2 + \delta_1$. 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, $u \geq m_1$.

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 17 (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 $2\varepsilon \in (0, \min(2\delta_1, \delta_2 - \delta_1))$. Let $l = \sup\{x \leq m_1 : f(x + \varepsilon) = g(x - \varepsilon)\}$ and $r = \inf\{x \geq m_1 : f(x - \varepsilon) = g(x + \varepsilon)\}$. Let A be the set defined as follows.*

1. *If* $m_2 - m_1 \geq \delta_2 + \delta_1 + 2\varepsilon$, *then* $A = (-\infty, m_1 + \delta_1 + \varepsilon]$.

2. *If* $\delta_2 - \delta_1 - 2\varepsilon \leq m_2 - m_1 < \delta_2 + \delta_1 + 2\varepsilon$, *then* $A = (-\infty, r]$.

3. *If* $m_2 - m_1 < \delta_2 - \delta_1 - 2\varepsilon$, *then* $A = [l, r]$.

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

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

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

The proof of Theorem 17 is in Appendix C.2

3.3 Continuous Loss Functions

It is natural to ask if the results for 0-1 loss in the preceding sections may be extended to continuous losses. Unlike for 0-1 loss, the optimal adversarial risk for general loss functions no longer admits a dual formulation based on optimal transport. However, we can still derive upper and lower bounds on the optimal adversarial risk for loss functions that are convex and smooth. In this section, we present such results on adversarial risk bounds in regression-like settings with continuous losses.

For this section, we assume that the feature space $\mathcal{X}$ is a Hilbert space, i.e., $\mathcal{X}$ is equipped with an inner product and a norm induced by the inner product.

3.3.1 Bounds on Optimal Adversarial Risk

Recall from Theorem 5 that for a loss function $\ell((\cdot, y), w)$ that is upper semi-analytic for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, we have $R_{\oplus\varepsilon}(\ell, w) = R_{\Gamma_\varepsilon}(\ell, w)$. From equation (2.17), we have $R_{\Gamma_\varepsilon}(\ell, w) \leq R_{\Gamma_\varepsilon^{(p)}}(\ell, w)$. In this section, we prove lower bounds on $R_{\Gamma_\varepsilon}(\ell, w)$ and upper bounds on $R_{\Gamma_\varepsilon^{(p)}}(\ell, w)$. Overall, this leads to lower and upper bounds on p-Wasserstein distributionally robust risk, which includes the adversarial risk, $R_{\Gamma_\varepsilon}(\ell, w)$ as a special case.

Throughout this section, we assume that the loss function satisfies the assumptions on Theorem 5 so that $R_{\oplus\varepsilon}(\ell, w) = R_{\Gamma_\varepsilon}(\ell, w)$. For a hypothesis class $\mathcal{W}$, we define the optimal adversarial as follows.

$$R^*_{\oplus\varepsilon} := \inf_{w \in \mathcal{W}} R_{\oplus\varepsilon}(\ell, w) = \inf_{w \in \mathcal{W}} R_{\Gamma_\varepsilon}(\ell, w). \tag{3.11}$$

Similarly, we define $R^*_{\Gamma_\varepsilon^{(p)}} := \inf_{w\in\mathcal{W}} R_{\Gamma_\varepsilon^{(p)}}(\ell, w)$. As a result of equation (2.17), we have the following inequality for integers p and q with $1 \le p \le q < \infty$.

$$R^*_{\oplus\varepsilon} \le R^*_{\Gamma_\varepsilon^{(q)}} \le R^*_{\Gamma_\varepsilon^{(p)}}. \tag{3.12}$$

We denote the standard risk (i.e. the risk when $\varepsilon = 0$) by $R_0(\ell, w)$ and define $R^*_0 = \inf_{w\in\mathcal{W}} R_0(\ell, w)$.

A trivial lower bound

We start by presenting a trivial lower bound on the optimal adversarial risk. To simplify presentation, we shall assume that for all $\varepsilon \ge 0$, there exists an optimal hypothesis $w^*_\varepsilon \in \mathcal{W}$ that attains the infimum in $\inf_{w\in\mathcal{W}} R_{\oplus\varepsilon}(\ell, w)$. The proofs can be easily modified by considering sequences of hypothesis such that $\liminf_i R^*_{\oplus\varepsilon}(w_i) = R^*_\varepsilon$ in case w^*_ε does not exist.

Theorem 18. *The optimal adversarial risk is at least as large as the optimal standard risk, that is, $R^*_{\oplus\varepsilon} \ge R^*_0$.*

Proof. We have the sequence of inequalities:

$$R^*_{\oplus\varepsilon} = R_{\oplus\varepsilon}(\ell, w^*_\varepsilon) \ge R_0(\ell, w^*_\varepsilon) \ge R_0(\ell, w^*_0) = R^*_0.$$

The first inequality holds because $R_{\oplus\varepsilon}(\ell, w)$ 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 $R_0(\ell, w)$. □

Note that the bound in Theorem 18 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 $R^*_{\oplus\varepsilon}$ 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 12 (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.

$$\ell((x', y), w) - \ell((x, y), w) \ge \langle \nabla_x \ell((x, y), w), x' - x\rangle. \tag{3.13}$$

Theorem 19. *The adversarial risk for a loss function satisfying* (3.13) *is bounded as follows.*

$$R^*_{\oplus\varepsilon} \geq R^*_0 + \inf_{w\in\mathcal{W}} \mathbb{E}_{(x,y)\sim\rho}\left[\sup_{d(x,x')\leq\varepsilon} \langle \nabla_x \ell((x,y),w), x'-x\rangle\right]. \tag{3.14}$$

Remark 2. *The lower bound holds for any p-Wasserstein distribution perturbing adversary with budget ε because $R^*_{\oplus\varepsilon} \leq R^*_{\Gamma^{(p)}_\varepsilon}$.*

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 19 as follows.

Corollary 3.3.1. *In the setting of Theorem 19, if $d(x,x') = \|x-x'\|_{adv}$ for $x, x' \in \mathcal{X}$, then the following bound holds:*

$$R^*_{\oplus\varepsilon} \geq R^*_0 + \varepsilon \inf_{w\in\mathcal{W}} \mathbb{E}_z[\|\nabla_x \ell((x,y),w)\|_{adv^*}], \tag{3.15}$$

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

Proof of Theorem 19. 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

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

Hence,

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

□

Proof of Corollary 3.3.1. From the proof of Theorem 19, we have

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

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

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

□

Next, we prove an upper bound for the adversarial risk for a W_1-distribution perturbing adversary. From equation 3.12, 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 13 (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.

$$|\ell((x',y), w) - \ell((x,y), w)| \leq L_w \|x' - x\|. \tag{3.16}$$

Theorem 20. *The adversarial risk for a W_1-distribution perturbing adversary with budget ε satisfies, $R^*_{\Gamma^{(p)}_\varepsilon} \leq R^*_0 + \varepsilon L_{w^*_0}$. Naturally, we also have $R^*_{\oplus\varepsilon} \leq R^*_0 + \varepsilon L_{w^*_0}$.*

The proof of this result uses an optimal transport idea from Tovar-Lopez and Jog[83].

Proof of Theorem 20. Recall that,

$$R^*_{\Gamma^{(p)}_\varepsilon} = \inf_{w\in\mathcal{W}} R_{\Gamma^{(p)}_\varepsilon}(\ell, w) = \inf_{w\in\mathcal{W}} \sup_{\gamma\in\Gamma^{(p)}_\varepsilon} \mathbb{E}_{(x',y)\sim\rho_y\rho^\gamma_{x'|y}} \left[\ell((x',y), w) \right].$$

Suppose that the infimum for $R^*_{\Gamma^{(p)}_\varepsilon}$ is attained at $\widehat{w}^*_\varepsilon$ and the supremum for $\widehat{R}^1_\varepsilon(\ell, \widehat{w}^*_\varepsilon)$ is attained for $\gamma^* \in \Gamma^{(1)}_\varepsilon$. 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

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

Here, (a) follows from the definition of $\widehat{w}^*_\varepsilon$, (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 $\gamma^* \in \Gamma^{(1)}_\varepsilon$. □

3.3.2 Bound on the evolution of Optimal Adversarial Classifier

Let w^*_ε and $\widehat{w}^*_\varepsilon$ denote the hypotheses in $\mathcal{W}$ that are optimal for $R^*_{\oplus\varepsilon}$ and $R^*_{\Gamma^{(p)}_\varepsilon}$ respectively. In this section, we analyze how w^*_ε or $\widehat{w}^*_\varepsilon$ 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 12. 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 $\varepsilon = 0$. 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 3.3.1, the bounds for deviation of w^*_ε and $\widehat{w}^*_\varepsilon$ are identical. Now, we prove a theorem on how much the optimal classifier can change in the presence of an adversary.

Theorem 21. *For a loss function ℓ that satisfies* (3.16)*, and is λ-strongly convex with respect to w, the following result*

holds:

$$\|w_\varepsilon^* - w_0^*\| \leq \sqrt{\frac{2\varepsilon L_{w_0^*}}{\lambda}}. \tag{3.17}$$

Proof of Theorem 21. We have the following series of inequalities.

$$\begin{aligned}
\varepsilon L_{w_0^*} &\overset{(a)}{\geq} R^*_{\Gamma_\varepsilon^{(p)}} - R_0^* \\
&\overset{(b)}{\geq} R_0(\ell, w_\varepsilon^*) - R_0(\ell, w_0^*) \\
&\overset{(c)}{\geq} \frac{\lambda}{2}(\nabla_w^2 R_0(\ell, w_0^*))\|w_\varepsilon^* - w_0^*\|^2.
\end{aligned}$$

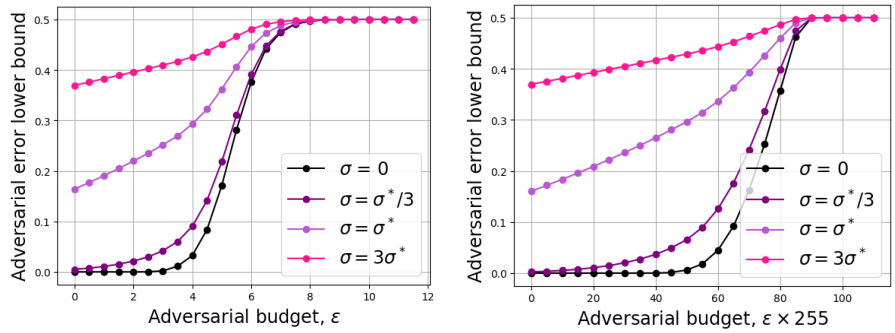
Here, (a) follows from Theorem 20, (b) follows from the fact that w_ε^* is sub-optimal for minimizing $R_0(\ell, w)$, and (c) follows from the λ-strong convexity of ℓ with respect to w. □

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.

3.4 Adversarial Risk Bounds for Real-world Datasets

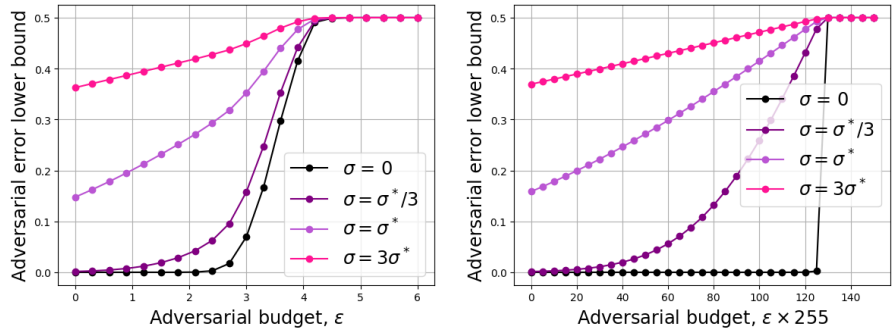
In this section, we present lower bounds on the optimal adversarial risk for empirical distributions derived from several real world datasets.

For the case of empirical distributions, the computation of the optimal transport cost in (3.1) 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 Peyré and Cuturi[66]), wherein the task is to optimally match each data point from the first distribution to a distinct data point from the second distribution. 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 7 by randomly sampling 5000 data points from each class. The results are showing in Figure 3.6 with the legend $\sigma = 0$.



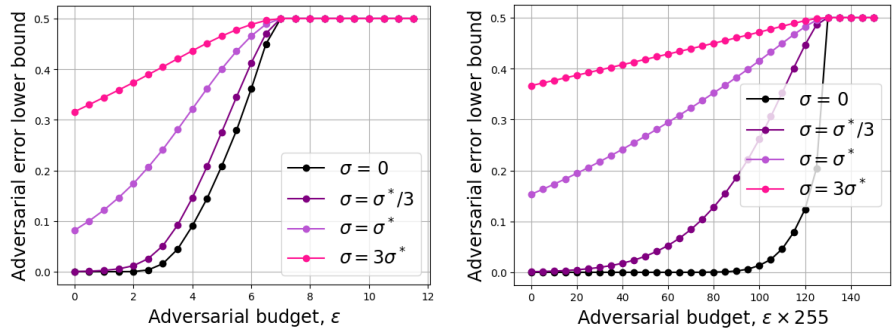
(a) CIFAR10 ℓ_2

(b) CIFAR10 ℓ_∞



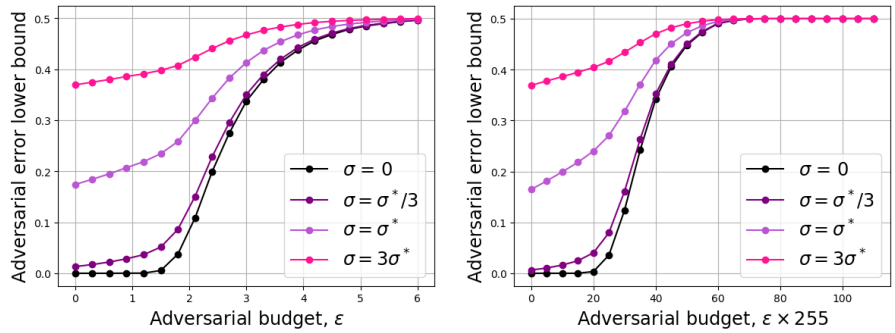
(c) MNIST ℓ_2

(d) MNIST ℓ_∞



(e) Fashion-MNIST ℓ_2

(f) Fashion-MNIST ℓ_∞



(g) SVHN ℓ_2

(h) SVHN ℓ_∞

Figure 3.6: Lower bounds on adversarial risk computed using Theorem 7. The curves with $\sigma = 0$ gives 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 13.

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.[7]. 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[43,32]. 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 $R^*_{\oplus\varepsilon}$.

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 13 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 3.6.

Figure 3.6 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 3.6 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 $\varepsilon = 3$ is 0.25 for $\sigma = \sigma^*$. That is, the adversarial error rate for CIFAR10 with $\varepsilon = 3$ 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.[7] (which is equivalent to the case of $\sigma = 0$) is 0 for $\varepsilon = 3$. Computation of non-trivial lower bounds for higher values of ε on adversarial error rate as in Figure 3.6 is made possible by our analysis on the optimal coupling to achieve D_ε between multivariate Gaussians in section 3.2.2.

We look up at the same stars and see such different things.

George R.R. Martin, A Storm of Swords

4

Alternative Characterizations

4.1 Distributional Robustness Perspective

Recall from Section 2.2 that the $R_{\Gamma_\varepsilon}(\ell_{0/1}, A)$ definition of adversarial risk is motivated from the robust hypothesis testing framework. In this approach, an adversary perturbs the true distribution p_i of class label $i \in \{0,1\}$ to a corrupted distribution p_i' such that $W_\infty(p_i, p_i') \leq \varepsilon$. The adversarial risk with such an adversary is given by

$$R_{\Gamma_\varepsilon}(\ell_{0/1}, A) = \sup_{W_\infty(p_1,p_1'),W_\infty(p_0,p_0')\leq\varepsilon} \frac{T}{T+1}p_0'(A) + \frac{1}{T+1}p_1'((A^c)),$$

and the optimal adversarial risk is given by,

$$R^*_{\Gamma_\varepsilon} := \inf_{A\in\mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_1,p_1'),W_\infty(p_0,p_0')\leq\varepsilon} \frac{T}{T+1}p_0'(A) + \frac{1}{T+1}p_1'((A^c)). \tag{4.1}$$

Consider the case of equal priors i.e., $T = 1$. For a particular choice of the adversary, i.e., for a particular choice

of the perturbed pair of distributions p_0' and p_1', the best that a classifier can attain is,

$$\inf_{A\in\mathcal{B}(\mathcal{X})} \frac{1}{2}p_0'(A) + \frac{1}{2}p_1'((A^c)) = \frac{1}{2}\left[1 - \sup_{A\in\mathcal{B}(\mathcal{X})} \left(p_1'(A) - p_0'(A)\right)\right] = \frac{1}{2}\left[1 - D_{TV}(p_0', p_1')\right].$$

We say that (p_0^*, p_1^*) are the least favorable pair of distributions (LFDs) for the robust hypothesis testing problem if the optimal risk for testing between (p_0^*, p_1^*) is equal to the minimax risk shown in equation 4.1. Proving the existence of LFDs is a standard problem in robust hypothesis testing literature, as it allows for reducing the robust hypothesis testing problem to a standard hypothesis testing problem involving the LFDs[38,40].

From Theorem 7 and using the fact that $R_{\Gamma_\varepsilon}(\ell_{0/1}, A) = R_{\oplus\varepsilon}(\ell_{0/1}, A)$ for all $A \in \mathcal{B}(\mathcal{X})$ from Theorem 4, we have the following.

$$R^*_{\oplus\varepsilon} = R^*_{\Gamma_\varepsilon} = \inf_{A\in\mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_1,p_1'),W_\infty(p_0,p_0')\leq\varepsilon} \frac{1}{2}p_0'(A) + \frac{1}{2}p_1'((A^c)) = \frac{1}{2}\left[1 - D_\varepsilon(p_0, p_1)\right].$$

Hence, if there exists a pair of distributions (p_0^*, p_1^*) such that $W_\infty(p_0, p_0*) \leq \varepsilon$, $W_\infty(p_1, p_1*) \leq \varepsilon$ and the following holds:

$$D_{TV}(p_0^*, p_1^*) = D_\varepsilon(p_0, p_1),$$

then (p_0^*, p_1^*) must be the LFDs for the robust hypothesis testing problem with ∞-Wasserstein uncertainty sets centered around p_0 and p_1. In this section, we show that there indeed exist LFDs that satisfy the above condition. We prove this result for general Polish spaces with a *mid-point property* stated below.

Definition 14 (Midpoint property)**.** A metric space $(\mathcal{X}, d)$ is said to have the midpoint property if for every $x_1, x_2 \in \mathcal{X}$, there exists $x \in \mathcal{X}$ such that, $d(x_1, x) = d(x, x_2) = d(x_1, x_2)/2$.

Any normed vector space with distance defined as $d(x, x') = \|x - x'\|$ satisfies the midpoint property. An example of a metric space without this property is the discrete metric space where $d(x, x') = 1\{x \neq x'\}$. The midpoint property plays a crucial role in proving the following theorem, which shows that the D_ε transport cost between two distributions is the shortest total variation distance between their ε-neighborhoods in W_∞ metric. A similar result was also presented in Dohmatob[24].

Theorem 22 (D_ε as shortest D_{TV} between W_∞ balls)**.** *Let $(\mathcal{X}, d)$ have the midpoint property. Let $\mu, \nu \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \geq 0$. Then,*

$$D_\varepsilon(\mu, \nu) = \inf_{W_\infty(\mu,\mu'), W_\infty(\nu,\nu') \leq \varepsilon} D_{TV}(\mu', \nu'). \tag{4.2}$$

Moreover, the infimum over D_{TV} in the above equation is attained.

The proof of Theorem 22 is in Appendix D.1.

As a result of the above theorem, we have the following corrolary on the existence of LFDs for the case of equal priors in binary classification.

Corollary 4.1.1. *Consider the binary classification setup with equal priors as in Theorem 7. Let $\mathcal{X}$ have the mid-point property. Then there exists LFDs (p_0^*, p_1^*) such that the following holds.*

$$\inf_{A \in \mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_1,p_1'), W_\infty(p_0,p_0') \leq \varepsilon} \frac{1}{2} p_0'(A) + \frac{1}{2} p_1'((A^c)) = \inf_{A \in \mathcal{B}(\mathcal{X})} \frac{1}{2} p_0^*(A) + \frac{1}{2} p_1^*((A^c)).$$

Proof. From Theorem 22, there exists (p_0^*, p_1^*) satisfying $W_\infty(p_1, p_1')$, $W_\infty(p_0, p_0') \leq \varepsilon$ such that $D_{TV}(p_0^*, p_1^*) = D_\varepsilon(p_0, p_1)$. Hence,

$$\begin{aligned}
\inf_{A \in \mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_1,p_1'), W_\infty(p_0,p_0') \leq \varepsilon} \frac{1}{2} p_0'(A) + \frac{1}{2} p_1'((A^c)) &= \frac{1}{2} \left[1 - D_\varepsilon(p_0, p_1)\right] \\
&= \frac{1}{2} \left[1 - D_{TV}(p_0^*, p_1^*)\right] \\
&= \frac{1}{2} \left[1 - \sup_{A \in \mathcal{B}(\mathcal{X})} (p_1^*(A) - p_0^*(A))\right] \\
&= \inf_{A \in \mathcal{B}(\mathcal{X})} \frac{1}{2} p_0^*(A) + \frac{1}{2} p_1^*((A^c)).
\end{aligned}$$

□

For the case of $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$ (which satisfies the mid-point property), we can use the powerful theory of Huber and Strassen's 2-alternating capacities[46] to extend the above theorem to the case of unequal priors.

Theorem 23. *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$. Then for any $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and $\varepsilon \geq 0$, there exist LFDs (p_0^*, p_1^*) such*

that the following holds.

$$\inf_{A\in\mathcal{B}(\mathcal{X})}\sup_{W_\infty(p_1,p_1')\le\varepsilon,W_\infty(p_0,p_0')\le\varepsilon}\frac{T}{T+1}p_0'(A)+\frac{1}{T+1}p_1'((A^c))=\inf_{A\in\mathcal{B}(\mathcal{X})}\frac{T}{T+1}p_0^*(A)+\frac{1}{T+1}p_1^*((A^c)).$$

The proof of Theorem 23 is given in Appendix D.1. Crucial to the proof of Theorem 24 is Lemma 2.4.3, which shows that the set-valued maps $A\mapsto p_0(A^{\oplus\varepsilon})$ and $A^c\mapsto p_1((A^c)^{\oplus\varepsilon})$ are 2-alternating capacities. The same proof technique is not applicable in general Polish spaces because the map $A\mapsto\mu(A^{\oplus\varepsilon})$ is not a capacity for a general $\mu\in\overline{\mathcal{P}}(\mathcal{X})$. This is because $A^{\oplus\varepsilon}$ is not measurable for all $A\in\overline{\mathcal{B}}(\mathcal{X})$.

4.2 Game Theoretic Perspective

Several works[67,62,13] propose a game-theoretic formulation for adversarial risk. Consider a game between two players:

1. The adversary whose action space is pairs of distributions $p_0',p_1'\in\overline{\mathcal{P}}(\mathcal{X})$.
2. The classifier whose action space is the space of decision regions of the form $A\in\mathcal{B}(\mathcal{X})\}$.

For $T>0$, define the payoff function, $r:\mathcal{B}(\mathcal{X})\times\overline{\mathcal{P}}(\mathcal{X})\times\overline{\mathcal{P}}(\mathcal{X})\to[0,1]$ as,

$$r(A,\mu,\nu)=\frac{T}{T+1}\mu(A)+\frac{1}{T+1}\nu((A^c)).\tag{4.3}$$

Given $A\in\mathcal{A}$, the adversary chooses $p_0',p_1'\in\mathcal{P}(\mathcal{X})$ to maximize $r(A,p_0',p_1')$. Similarly, given $p_0',p_1'\in\mathcal{P}(\mathcal{X})$, the algorithm chooses $A\in\mathcal{A}$ to minimize $r(A,p_0',p_1')$. This defines a zero-sum game between both players with $r(A,p_0',p_1')$ as the pay-off function.

The max-min inequality stated below shows that the pay-off for any player is better (i.e. more for adversary and less for classifier) when they make the first move.

$$\sup_{W_\infty(p_0,p_0')\le\varepsilon,W_\infty(p_1,p_1')\le\varepsilon}\inf_{A\in\mathcal{A}}r(A,p_0',p_1')\le\inf_{A\in\mathcal{B}(\mathcal{X})}\sup_{W_\infty(p_0,p_0')\le\varepsilon,W_\infty(p_1,p_1')\le\varepsilon}r(A,p_0',p_1').\tag{4.4}$$

The game is said to have a value if the following result holds:

$$\inf_{A\in\mathcal{A}} \sup_{\substack{W_\infty^\varepsilon(p_0,p_0')\le\varepsilon \\ W_\infty^\varepsilon(p_1,p_1')\le\varepsilon}} r(A,p_0',p_1') = \sup_{\substack{W_\infty^\varepsilon(p_0,p_0')\le\varepsilon \\ W_\infty^\varepsilon(p_1,p_1')\le\varepsilon}} \inf_{A\in\mathcal{A}} r(A,p_0',p_1'). \tag{4.5}$$

If the inequality in (4.4) is an equality, we say that the game has zero duality gap, and admits a value equal to either expression in (4.4). In the equality setting, there is no advantage to a player making the first move. Our minimax theorems establish such an equality. If, in addition to having an equality in (4.4), there exist $p_0^*, p_1^* \in \mathcal{P}(\mathcal{X})$ that achieve the supremum on the left-hand side and $A^* \in \mathcal{B}(\mathcal{X})$ that achieves the infimum on the right-hand side, we say that $((p_0^*, p_1^*), A^*)$ is a pure Nash equilibrium of the game. On the other hand, we say that $((p_0^*, p_1^*), A^*)$ is a δ-approximate pure Nash equilibrium of the game if the following inequality holds.

$$\sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} r(A^*,p_0',p_1') - \delta \le r(A^*,p_0^*,p_1^*) \le \inf_{A\in\mathcal{A}} r(A,p_0^*,p_1^*) + \delta.$$

The following theorem proves the minimax equality and the existence of a Nash equilibrium for the adversarial robustness game in $\mathbb{R}^d$.

Theorem 24 (Minimax theorem in $\mathbb{R}^d$)**.** *Let $(\mathcal{X}, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \ge 0$. Define r as in* (4.3)*. Then,*

$$\sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} \inf_{A\in\mathcal{L}(\mathcal{X})} r(A,p_0',p_1') = \inf_{A\in\mathcal{L}(\mathcal{X})} \sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} r(A,p_0',p_1'). \tag{4.6}$$

Moreover, there exist $p_0^*, p_1^* \in \overline{\mathcal{P}}(\mathcal{X})$ and $A^* \in \mathcal{L}(\mathcal{X})$ that achieve the supremum and infimum on the left and right hand sides of the above equation.

Proof. From Theorem 23, there exist LFDs (p_0^*, p_1^*) such that the following holds.

$$\inf_{A\in\mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_1,p_1'),W_\infty(p_0,p_0')\le\varepsilon} r(A,p_0',p_1') = \inf_{A\in\mathcal{B}(\mathcal{X})} r(A,p_0^*,p_1^*). \tag{4.7}$$

Since (p_0^*, p_1^*) satisfy the constraints $W_\infty(p_0,p_0'),\ W_\infty(p_1,p_1') \le \varepsilon$, we have

$$\inf_{A\in\mathcal{B}(\mathcal{X})} r(A,p_0^*,p_1^*) \le \sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} \inf_{A\in\mathcal{L}(\mathcal{X})} r(A,p_0',p_1'). \tag{4.8}$$

Combining equations (4.7) and (4.7), we get the following inequality.

$$\inf_{A\in\mathcal{B}(\mathcal{X})}\sup_{W_\infty(p_1,p_1'),W_\infty(p_0,p_0')\le\varepsilon} r(A,p_0',p_1') \le \sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon}\inf_{A\in\mathcal{L}(\mathcal{X})} r(A,p_0',p_1').$$

The desired result follows from combining the above inequality with the max-min inequality (4.4).

The existence of $A^*\in\mathcal{L}(\mathcal{X})$ that attains the infimum on the right in (4.6) follows from Lemma 3.1 in Huber and Strassen[46] and the equality $R_{\oplus\varepsilon}(\ell_{0/1},A)=R_{\Gamma_\varepsilon}(\ell_{0/1},A)$ proved in Theorem 6.

□

As noted before, the same proof technique is not applicable in general Polish spaces because the map $A\mapsto\mu(A^{\oplus\varepsilon})$ is not a capacity for a general $\mu\in\overline{\mathcal{P}}(\mathcal{X})$. This is because $A^{\oplus\varepsilon}$ is not measurable for all $A\in\overline{\mathcal{B}}(\mathcal{X})$.

In general Polish spaces, we can use the characterization of D_ε as the shortest total variation distance between W_∞ ball from Theorem 22 to prove the minimax theorem for the case of equal priors.

Theorem 25 (Minimax theorem for equal priors)**.** *Let $(\mathcal{X},d)$ have the midpoint property. Let $p_0,p_1\in\overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon\ge 0$. Define r as in (4.3) with $T=1$. Then*

$$\sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon}\inf_{A\in\mathcal{B}(\mathcal{X})} r(A,p_0',p_1') = \inf_{A\in\mathcal{B}(\mathcal{X})}\sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} r(A,p_0',p_1'). \tag{4.9}$$

Moreover, there exist $p_0^*,p_1^*\in\mathcal{P}(\mathcal{X})$ that achieve the supremum on the left hand side of the above equation.

Proof. We have the following series of equalities.

$$\begin{aligned}\inf_{A\in\mathcal{B}(\mathcal{X})}\sup_{\substack{W_\infty(p_0,p_0')\le\varepsilon\\ W_\infty(p_1,p_1')\le\varepsilon}} r(A,p_0',p_1') &= \inf_{A\in\mathcal{B}(\mathcal{X})} R_{\Gamma_\varepsilon}(\ell_{0/1},A)\\ &\overset{(i)}{=} \inf_{A\in\mathcal{B}(\mathcal{X})} R_{\oplus\varepsilon}(\ell_{0/1},A)\\ &\overset{(ii)}{=} \frac{1}{2}\left[1-D_\varepsilon(p_0,p_1)\right],\end{aligned}$$

and

$$\sup_{\substack{W_\infty(p_0,p_0')\le\varepsilon\\ W_\infty(p_1,p_1')\le\varepsilon}}\inf_{A\in\mathcal{B}(\mathcal{X})} r(A,p_0',p_1') \overset{(iii)}{=} \sup_{\substack{W_\infty(p_0,p_0')\le\varepsilon\\ W_\infty(p_1,p_1')\le\varepsilon}} \frac{1}{2}\left[1-D_{TV}(p_0',p_1')\right]$$

$$= \frac{1}{2}\left[1 - \inf_{\substack{W_\infty(p_0,p_0')\le\varepsilon \\ W_\infty(p_1,p_1')\le\varepsilon}} D_{TV}(p_0', p_1')\right],$$

where (i) follows from Theorem 4, (ii) from Theorem 10, and (*iii*) again from Theorem 10 with $\varepsilon = 0$. The expressions on the right extremes of the above equations are equal by Theorem 22. The existence of $p_0^*, p_1^* \in \overline{\mathcal{P}}(\mathcal{X})$ follows Theorem 22. □

To prove the minimax theorem for unequal priors, we need the following generalization of Theorem 22 to finite measures of unequal mass.

Lemma 4.2.1. *Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \ge 0$. Then for $T \ge 1$,*

$$\begin{aligned}\inf_{q\in\overline{\mathcal{P}}(\mathcal{X}):q\preceq Tp_0} D_\varepsilon(q, p_1) &= \inf_{q\in\overline{\mathcal{P}}(\mathcal{X}):q\preceq Tp_0} \inf_{W_\infty(q,q'),W_\infty(p_1,p_1')\le\varepsilon} D_{TV}(q', p_1') \\ &= \inf_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} \inf_{q'\in\overline{\mathcal{P}}(\mathcal{X}):q'\preceq Tp_0'} D_{TV}(q', p_1') \qquad (4.10)\end{aligned}$$

The proof of Lemma 4.2.1 is contained in Appendix D.

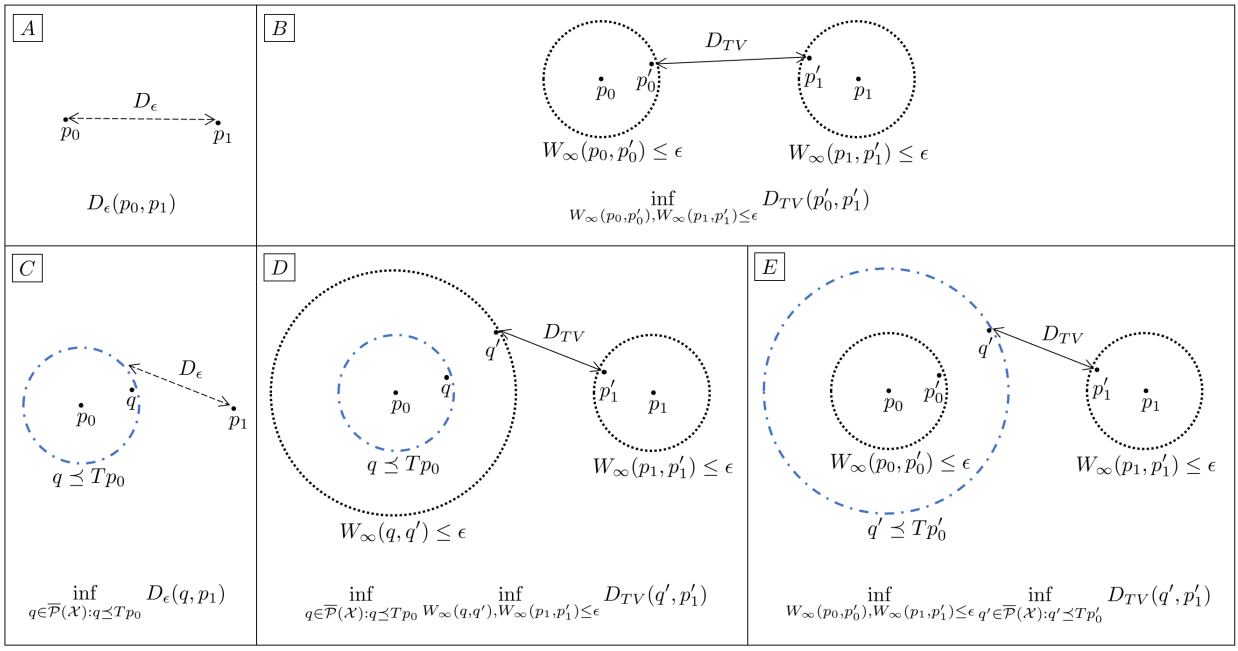
Now, we prove the minimax equality for unequal priors.

Theorem 26 (Minimax theorem for unequal priors)**.** *Let $(\mathcal{X}, d)$ have the midpoint property. Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and let $\varepsilon \ge 0$. For $T > 0$, define r as in* (4.3)*. Then*

$$\sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} \inf_{A\in\mathcal{B}(\mathcal{X})} r(A, p_0', p_1') = \inf_{A\in\mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\le\varepsilon} r(A, p_0', p_1'). \qquad (4.11)$$

Proof. Without loss of generality, we assume $T \ge 1$. (If $T < 1$, we simply repeat the proof with labels 0 and 1 swapped.) We have

$$\begin{aligned}\inf_{A\in\mathcal{B}(\mathcal{X})} \sup_{\substack{W_\infty(p_0,p_0')\le\varepsilon \\ W_\infty(p_1,p_1')\le\varepsilon}} r(A, p_0', p_1') &= \inf_{A\in\mathcal{B}(\mathcal{X})} R_{\Gamma_\varepsilon}(\ell_{0/1}, A) \\ &\overset{(i)}{=} \inf_{A\in\mathcal{B}(\mathcal{X})} R_{\oplus\varepsilon}(\ell_{0/1}, A) \\ &\overset{(ii)}{=} \frac{1}{T+1}\left[1 - \inf_{\substack{q\in\overline{\mathcal{P}}(\mathcal{X}): \\ q\preceq Tp_0}} D_\varepsilon(p_0, p_1)\right]\end{aligned}$$



$$T > 1 : R^*_{\oplus\epsilon} = \frac{1}{T+1}\left[1 - \boxed{C}\right], \qquad \boxed{C} = \boxed{D} = \boxed{E}$$
$$T = 1 : R^*_{\oplus\epsilon} = \frac{1}{2}\left[1 - \boxed{A}\right], \qquad \boxed{A} = \boxed{B} = \boxed{C} = \boxed{D} = \boxed{E}$$

Figure 4.1: Illustration of various equivalent formulations of the optimal adversarial risk. The equalities summarize the results of Section 3.1 and Chapter 4. For equal priors ($T = 1$), $\boxed{A}$ and $\boxed{B}$ denote two ways of obtaining the optimal adversarial risk, $R^*_{\oplus\varepsilon}$: 1) $\boxed{A}$, which denotes the D_ε cost between the true label distributions p_0 and p_1, and 2) $\boxed{B}$, which denotes the shortest total variation distance between ∞-Wasserstein balls of radius ε around p_0 and p_1. For unequal priors ($T > 1$), $\boxed{C}$, $\boxed{D}$ and $\boxed{E}$ denote three equivalent ways of obtaining $R^*_{\oplus\varepsilon}$. The black dotted balls denote ∞-Wasserstein balls and the blue dashed balls denote sets defined using stochastic domination. The order in which the two types of balls appear around p_0 is reversed between $\boxed{D}$ and $\boxed{E}$.

$$\overset{(iii)}{=} \frac{1}{T+1}\left[1 - \inf_{\substack{W_\infty(p_0,p_0')\le\varepsilon \\ W_\infty(p_1,p_1')\le\varepsilon}} \inf_{\substack{q'\in\overline{\mathcal{P}}(\mathcal{X}): \\ q'\preceq_T p_0'}} D_{TV}(q',p_1')\right]$$

$$= \sup_{\substack{W_\infty(p_0,p_0')\le\varepsilon \\ W_\infty(p_1,p_1')\le\varepsilon}} \frac{1}{T+1}\left[1 - \inf_{\substack{q'\in\overline{\mathcal{P}}(\mathcal{X}): \\ q'\preceq_T p_0'}} D_{TV}(q',p_1')\right]$$

$$\overset{(iv)}{=} \sup_{\substack{W_\infty(p_0,p_0')\le\varepsilon \\ W_\infty(p_1,p_1')\le\varepsilon}} \inf_{A\in\mathcal{B}(\mathcal{X})} r(A,p_0',p_1'),$$

where (i) follows from Theorem 4, (ii) from Theorem 10, (iii) from Lemma 4.2.1 and (iv) follows again from Theorem 10 with $\varepsilon = 0$. □

Remark 3. *Unlike Theorem 24, Theorems 25 and 26 do not guarantee the existence of an optimal decision region A^*. While Theorem 25 guarantees the existence of worst-case pair of perturbed distributions p_0^*, p_1^*, Theorem 26 does not do so. Nevertheless, a δ-approximate pure Nash equilibrium exists in all the cases. This is in sharp contrast with the non-existence of Nash equilibrium proven in Pinot et al.[67] The result of Pinot et al.[67] is valid for a "regularized" adversary, where the point-wise budget constraint $d(x,x') \le \varepsilon$ is replaced with a regularization term added to the adversarial risk formulation. Our Nash equilibrium result holds for the standard formulation of adversarial risk as in[58,73], without the need for a regularization term.*

Remark 4. *A recent work[62] shows the existence of mixed Nash equilibrium for randomized classifiers parametrized by points in a Polish space. Other works[67,13] consider a similar setup, but with a "regularized" adversary. The equilibrium analysis in these works uses Fan's minimax theorem with concave-convex condition. Since we consider non-parametric classifiers represented by arbitrary decision regions, Fan's theorem is inapplicable in our setting. Instead, we use tools from Huber's 2-alternating capacities for $\mathbb{R}^d$, and the generalized Strassen's duality theorem for general Polish spaces. The connection with Huber's capacities (which we prove in Lemma 2.4.3) and the generalization of Strassen's theorem (Theorem 9) are both novel to the best of our knowledge.*

Run, rabbit, run

Dig that hole, forget the sun

When, at last, the work is done

Don't sit down, it's time to dig another one

Pink Floyd, Breathe

5

Conclusion

We examined different notions of adversarial risk and laid down the conditions under which these definitions are equivalent. By verifying the conditions in Sections 2.3 and 2.4, researchers may use different definitions interchangeably.

We introduced the D_ε optimal transport cost 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 $R^*_{\oplus\varepsilon}$ for the case of binary classification under 0-1 loss function. For general loss functions, we give lower and upper bounds on $R^*_{\oplus\varepsilon}$ in terms of the Lipschitz and strong convexity parameters of the loss function.

Using a formulation of optimal transport between finite measures of unequal mass, we extended the D_ε based characterization of adversarial risk to unequal priors by generalizing Strassen's theorem. This may find applications in the study of excess cost optimal transport[97,96]. A recent work[84] obtains a different characterization of optimal adversarial risk using optimal transport on the product space $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{Y}$ is the label space. Further, they show the evolution of the optimal classifier A^* as ε grows, in terms of a mean curvature flow. This raises an interesting question on the evolution of the optimal adversarial distributions $p_0^*, p_1^* \in \overline{\mathcal{P}}(\mathcal{X})$ with ε.

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.

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.

We proved a minimax theorem for adversarial robustness game and the existence of a Nash equilibrium. We constructed the worst-case pair of distributions $p^*_0, p^*_1 \in \overline{\mathcal{P}}(\mathcal{X})$ in terms of true data distributions and showed that their total variation distance gives the optimal adversarial risk. Identifying worst case distributions could lead to a new approach to developing robust algorithms.

We used Choquet capacities for results in $\mathbb{R}^d$ and measurable selections in Polish spaces. Specifically, we showed that the measure of ε-Minkowski expansion is a 2-alternating capacity. This connection could help generalize our results to total variation and Prokhorov distance based contaminations.

We largely focused on the binary classification setup with 0-1 loss function. While we extended our results on measurability and relation to ∞-Wasserstein distributional robustness to more general loss functions and a multi-class setup, it is unclear how our results on generalized Strassen's theorem and Nash equilibria can be extended further.

Our results on various equivalent formulations of optimal adversarial risk are specific to adversarial perturbations (or equivalently, ∞-Wasserstein distributional perturbations). An interesting open question is whether these results hold for more general perturbation models.

A

Preliminary Lemmas

Lemma A.0.1. *Let $A_n \in \mathcal{B}(\mathcal{X})$ for $n \in \{1, 2, \ldots\}$. Then,*

$$(\cup_n A_n)^{\oplus\varepsilon} = \cup_n A_n^{\oplus\varepsilon},$$
$$(\cap_n A_n)^{\oplus\varepsilon} \subseteq \cap_n A_n^{\oplus\varepsilon}.$$

Proof. Suppose $a \in (\cup_n A_n)^{\oplus\varepsilon}$. Then there exists $a_i \in A_i$ for some $i \in \{1, 2, \ldots\}$ such that $d(a, a_i) \leq \varepsilon$. Hence, $a \in A_i^{\oplus\varepsilon} \subseteq \cup_n A_n^{\oplus\varepsilon}$. Therefore, $(\cup_n A_n)^{\oplus\varepsilon} \subseteq \cup_n A_n^{\oplus\varepsilon}$.

Suppose $b \in \cup_n A_n^{\oplus\varepsilon}$. Then $b \in A_j^{\oplus\varepsilon}$ for some $j \in \mathbb{N}$. So there must exist $b' \in A_j$ such that $d(b, b') \leq \varepsilon$. Since $b' \in \cup_n A_n$, we get that $b \in (\cup_n A_n)^{\oplus\varepsilon}$. Therefore, $\cup_n A_n^{\oplus\varepsilon} \subseteq (\cup_n A_n)^{\oplus\varepsilon}$.

Suppose $c \in (\cap_n A_n)^{\oplus\varepsilon}$. Then there exists $c' \in \cap_n A_n$ such that $d(c, c') \leq \varepsilon$. Since $c' \in A_n$ for all $n \in \{1, 2, \ldots\}$, $c \in A_n^{\oplus\varepsilon}$ for all $n \in \{1, 2, \ldots\}$. Hence, $c \in \cap_n A_n^{\oplus\varepsilon}$. Therefore, $(\cap_n A_n)^{\oplus\varepsilon} \subseteq \cap_n A_n^{\oplus\varepsilon}$. □

Lemma A.0.2. *Let (F_n) be a sequence of closed sets in $\mathcal{X}$ such that $F_k \supseteq F_{k+1}$ for $k \in \mathcal{N}$. Then,*

$$(\cap_n F_n)^{\oplus\varepsilon} = \cap_n F_n^{\oplus\varepsilon}.$$

Proof. Suppose $x \in (\cap_n F_n)^{\oplus\varepsilon}$. Then there exists $x' \in \cap_n F_n$ such that $d(x, x') \leq \varepsilon$. Since $x' \in F_n$ for all $n \in \mathbb{N}$, $x \in F_n^{\oplus\varepsilon}$ for all $n \in \mathbb{N}$. Hence, $x \in \cap_n F_n^{\oplus\varepsilon}$ and therefore $(\cap_n F_n)^{\oplus\varepsilon} \subseteq \cap_n F_n^{\oplus\varepsilon}$. We will now show the set inclusion in the opposite direction.

Let $x \in \cap_n F_n^{\oplus\varepsilon}$. Then $x \in F_n^{\oplus\varepsilon}$ for all $n \in \mathbb{N}$. Hence, there exists $x_n \in F_n$ such that $d(x, x_n) \leq \varepsilon$ for all $n \in \mathbb{N}$. Since (x_n) is a bounded sequence, it has a subsequence (x_{n_k}) that converges to some x^*. We claim that $x^* \in F := \cap_n F_n$. Indeed, for any $m \in \mathbb{N}$, the tail of the subsequence (x_{n_k}) with indices greater than m is contained in F_m. Since F_m is closed, x^* must be in F_m. Since the choice of m was arbitrary, $x^* \in \cap_m F_m = F$. Hence, $x \in F^{\oplus\varepsilon}$ because $d(x, x^*) \leq \varepsilon$. Therefore, $\cap_n F_n^{\oplus\varepsilon} \subseteq F^{\oplus\varepsilon}$. □

Lemma A.0.3. *Let $A \in \mathcal{B}(\mathcal{X})$. Let $(\gamma_n)_{n=1}^{\infty}$ be a non-negative, monotonically decreasing sequence converging to 0. Let $\overline{A}$ denote the closure of A in $\mathcal{X}$. Then, $A^{\gamma_n} \downarrow \overline{A}$.*

Proof. We know $\overline{A} \subseteq \overline{A}^{\gamma_n} = A^{\gamma_n}$ for all n. Hence $\overline{A} \subseteq \lim_{n\to\infty} \bigcap_{k=1}^{n} A^{\gamma_k}$.

Suppose $x \in \lim_{n\to\infty} \bigcap_{k=1}^{n} A^{\gamma_k}$. Then it must be that $d(x, \overline{A}) = 0$ because otherwise x would not lie in A^{γ_n} for all large enough n. Since $d(x, \overline{A}) = 0$, we can find a sequence of points in $\overline{A}$ that tend to x. But since $\overline{A}$ is closed, we must have $x \in \overline{A}$. Hence, $\lim_{n\to\infty} \bigcap_{k=1}^{n} A^{\gamma_k} \subseteq \overline{A}$. □

Lemma A.0.4. *Let $\varepsilon > 0$. If A is a closed set, then $A^{\oplus\varepsilon}$ and $A^{\ominus\varepsilon}$ are also closed sets.*

Proof. 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 $A^{\oplus\varepsilon}$ converging to a limit z. Assume without loss of generality that $d(z_i, z) < \delta/2$. We shall show that $z \in A^{\oplus\varepsilon}$ 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

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

Thus $\tilde{a}_i \in B(z, \varepsilon + \delta) \cap A$, 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 $A^{\oplus\varepsilon}$, which shows that $A^{\oplus\varepsilon}$ is closed.

Recall that $A^{\ominus\varepsilon} = ((A^c)^{\oplus\varepsilon})^c$. Since A^c is an open set, it is enough to show that $C^{\oplus\varepsilon}$ is open if C is open. Let $z \in C^{\oplus\varepsilon}$, 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 $N_\delta(z) \subseteq C^{\oplus\varepsilon}$, since $N_\delta(z) = N_\delta(c) + b$. This shows that every point $z \in C^{\oplus\varepsilon}$ admits a small ball around it that is contained in $C^{\oplus\varepsilon}$, or equivalently, $C^{\oplus\varepsilon}$ is open. This completes the proof. □

Lemma A.0.5. *For a closed set A, we have $A^\varepsilon = A^{\oplus\varepsilon}$.*

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

To prove the reverse direction, suppose $x \in A^\varepsilon$. This means we can a sequence of points $\{a_i\}$ such that $a_i \in A$ and $\liminf_i d(x, a_i) \leq \varepsilon$. Fix a $\delta > 0$ and assume without loss of generality that $d(x, a_i) \leq \varepsilon + \delta$ for all $i > 0$. Then $a_i \in B(x, \varepsilon + \delta) \cap A$ for all $i > 0$. As A is closed, the set $B(x_i, \varepsilon + \delta) \cap A$ 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

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

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

Lemma A.0.6. *Let A be a closed set. Then $(A^{\ominus\varepsilon})^{\oplus\varepsilon} \subseteq A$ and $A \subseteq (A^{\oplus\varepsilon})^{\ominus\varepsilon}$.*

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

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

Lemma A.0.7. *Let $\varepsilon_1 > \varepsilon_2 > 0$ and $A \in \mathcal{B}(\mathcal{X})$. Then for any $\delta \in (0, \varepsilon_1 - \varepsilon_2)$, $A^{\varepsilon_1 - \varepsilon_2 - \delta} \subseteq (A^{\varepsilon_1})^{-\varepsilon_2}$.*

Proof. Recall that for $\varepsilon > 0$, $A^{-\varepsilon} = ((A^c)^{\varepsilon})^c$. From the definition, $x \in A^{-\varepsilon}$ if and only if $d(x, A^c) > \varepsilon$.

Let $\delta \in (0, \varepsilon_1 - \varepsilon_2)$ and $x \in A^{\varepsilon_1 - \varepsilon_2 - \delta}$. Then, $d(x, A) \leq \varepsilon_1 - \varepsilon_2 - \delta$. Consider any $y \in (A^{\varepsilon_1})^c$. Then, $d(y, A) > \varepsilon_1$. By the triangle inequality,

$$d(x, y) \geq d(y, A) - d(x, A) > \varepsilon_1 - (\varepsilon_1 - \varepsilon_2 - \delta) = \varepsilon_2 + \delta.$$

Hence,

$$d(x, (A^{\varepsilon_1})^c) = \inf_{y \in (A^{\varepsilon_1})^c} d(x, y) \geq \varepsilon_2 + \delta > \varepsilon_2.$$

Therefore, $x \in (A^{\varepsilon_1})^{-\varepsilon_2}$. □

Lemma A.0.8. *Let $A \in \mathcal{B}(\mathcal{X})$. Then, $1\{x \in A^{\oplus\varepsilon}\} = \sup_{x' \in B_\varepsilon(x)} 1\{x' \in A\}$.*

Proof. Suppose $x \in A^{\oplus\varepsilon}$. Then there exists $x' \in A$ such that $x' \in B_\varepsilon(x)$. Hence, $\sup_{x' \in B_\varepsilon(x)} 1\{x' \in A\} = 1$.

Suppose $x \in \mathcal{X}$ is such that $\sup_{x' \in B_\varepsilon(x)} 1\{x' \in A\} = 1$. Then there is a sequence $(x_n)_{n=1}^{\infty}$ such that $d(x, x_n) \leq \varepsilon$ and $x_n \in A$ for all n. Since (x_n) is a bounded sequence in a closed set, $B_\varepsilon(x)$, it has a subsequence that converges to some x^* such that $d(x, x^*) \leq \varepsilon$ and $x^* \in A$. Hence, $x \in B_\varepsilon(x^*) \subseteq A^{\oplus\varepsilon}$. □

Lemma A.0.9. *For any real-valued function $f : \mathcal{X} \to \mathbb{R}$ and any $t \in \mathbb{R}$,*

$$\left\{ x \in \mathcal{X} : \sup_{d(x,x') \leq \varepsilon} f(x') > t \right\} = \{x \in \mathcal{X} : f(x) > t\}^{\oplus\varepsilon}.$$

Proof. Suppose $a \in \{x \in \mathcal{X} : f(x) > t\}^{\oplus\varepsilon}$. Then there exists $a' \in \mathcal{X}$ such that $f(a') > t$ and $d(a, a') \leq \varepsilon$. Hence, $\sup_{d(a,x') \leq \varepsilon} f(x') \geq f(a') > t$. Therefore, $a \in \left\{ x \in \mathcal{X} : \sup_{d(x,x') \leq \varepsilon} f(x') > t \right\}$.

Suppose $b \in \left\{ x \in \mathcal{X} : \sup_{d(x,x') \leq \varepsilon} f(x') > t \right\}$. Then there exists $b' \in \mathcal{X}$ such that $f(b') > t$ and $d(b, b') \leq \varepsilon$. Hence, $b \in \{x \in \mathcal{X} : f(x) > t\}^{\oplus\varepsilon}$. □

Lemma A.0.10 (Max-min Inequality)**.** *Let $p_0, p_1 \in \mathcal{P}(\mathcal{X})$ and let $\varepsilon \geq 0$. For $T > 0$, define $r : \mathcal{B}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, 1]$ as in* (4.3)*. Then,*

$$\sup_{\substack{W_\infty(p_0, p_0') \leq \varepsilon \\ W_\infty(p_1, p_1') \leq \varepsilon}} \inf_{A \in \mathcal{B}(\mathcal{X})} r(A, p_0', p_1') \leq \inf_{A \in \mathcal{B}(\mathcal{X})} \sup_{\substack{W_\infty(p_0, p_0') \leq \varepsilon \\ W_\infty(p_1, p_1') \leq \varepsilon}} r(A, p_0', p_1'). \tag{A.1}$$

Proof. For any $A \in \mathcal{B}(\mathcal{X})$ and p_0', p_1' such that $W_\infty(p_i, p_i') \leq \varepsilon\,(i = 0, 1)$, we have

$$\inf_{A \in \mathcal{B}(\mathcal{X})} r(A, p_0', p_1') \leq r(A, p_0', p_1').$$

Taking supremum over p_0' and p_1' such that $W_\infty(p_i, p_i') \leq \varepsilon$ for $i \in \{0, 1\}$ on both sides of the above inequality, we get the following for any $A \in \mathcal{B}(\mathcal{X})$.

$$\sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \varepsilon} \inf_{A \in \mathcal{B}(\mathcal{X})} r(A, p_0', p_1') \leq \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \varepsilon} r(A, p_0', p_1').$$

Since the above inequality holds for any $A \in \mathcal{B}(\mathcal{X})$, we have,

$$\sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \varepsilon} \inf_{A \in \mathcal{B}(\mathcal{X})} r(A, p_0', p_1') \leq \inf_{A \in \mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \varepsilon} r(A, p_0', p_1').$$

□

B

Proofs from Chapter 2

B.1 Proofs from Section 2.3

B.1.1 Proofs from Section 2.3.1

Proof of Lemma 2.3.1. We prove the above statement by using a counterexample motivated from Example 2.4 in Luiro et al.[56] For any $\varepsilon > 0$, there exists a Borel measurable set $S \subseteq [-\varepsilon, \varepsilon]^2$ such that its projection onto the first coordinate is not Borel measurable (see Luiro et al.,[56] Theorem 6.7.2 and Theorem 6.7.11 in Bogachev[12]). That is, $S \in \mathcal{B}(\mathbb{R}^2)$ but $S_1 := \{x_1 \in \mathbb{R} : (x_1, x_2) \in S\} \notin \mathcal{B}(\mathbb{R})$.

Define a homeomorphism $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ as $\varphi(x_1, x_2, x_3) := (x_1, x_2, \sqrt{\varepsilon^2 - x_2^2})$. φ maps the plane $[-\varepsilon, \varepsilon]^2 \times \{0\}$ onto the half-cylinder, $\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 \in [-\varepsilon, \varepsilon], x_2^2 + x_3^2 = \varepsilon^2, x_3 \geq 0\}$, of radius ε. Let $A := \varphi(S \times \{0\})$. Then $A \in \mathcal{B}(\mathbb{R}^3)$ because $S \times \{0\} \in \mathcal{B}(\mathbb{R}^3)$. We have the following equality.

$$A^{\oplus \varepsilon} \cap (\mathbb{R} \times \{0\}^2) = S_1 \times \{0\}^2$$

Suppose $A^{\oplus \varepsilon} \in \mathcal{B}(\mathbb{R}^3)$. Then the above equality implies that $S_1 \in \mathcal{B}(\mathbb{R})$ contradicting our choice of S. Hence,

$A^{\oplus\varepsilon} \notin \mathcal{B}(\mathbb{R}^3)$.

□

Proof of Lemma 2.3.2. Recall that an analytic set is a continuous image of a Borel set in a Polish space. Although an analytic set need not be Borel measurable, it is always universally measurable, i.e., measurable with respect to any measure defined on a complete measure space [5].

We will now show that if $A \in \mathcal{B}(\mathcal{X})$, then $A^{\oplus\varepsilon}$ is an analytic set, thus showing that it is measurable in the complete measure space $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$.

Define $D = \{(x, x') \in \mathcal{X}^2 : d(x, x') \leq \varepsilon\}$. D is Borel measurable because it is the preimage of the Borel set $(-\infty, \varepsilon]$ under the Borel measurable function d. Define $f : D \to \mathbb{R}$ as $f(x, x') = -1\{x' \in A\}$. For $c \in \mathbb{R}$, we have the following.

$$\{(x, x') \in \mathcal{X}^2 : f(x, x') < c\} = \begin{cases} \varnothing & c \leq -1, \\ (\mathcal{X} \times A) \cap D & c \in (-1, 0], \\ \mathcal{X}^2 & c > 0. \end{cases}$$

Since $A \in \mathcal{B}(\mathcal{X})$ and $D \in \mathcal{B}(\mathcal{X}^2)$, $(\mathcal{X} \times A) \cap D \in \mathcal{B}(\mathcal{X}^2)$. Hence, by Definition 7.21 in Bertsekas and Shreve [5], f is a lower semianalytic function. By Proposition 7.47 in Bertsekas and Shreve [5], the function $f^* : \mathcal{X} \to \mathbb{R}$ defined as $f^*(x) := \inf_{x' \in B_\varepsilon(x)} f(x, x')$ is lower semianalytic. By Lemma A.0.8, we have

$$f^*(x) = \inf_{x' \in B_\varepsilon(x)} -1\{x' \in A\} = -\sup_{x' \in B_\varepsilon(x)} 1\{x' \in A\} = -1\{x \in A^{\oplus\varepsilon}\}.$$

By Definition 7.21 in Bertsekas and Shreve [5], it follows that $A^{\oplus\varepsilon}$ is an analytic set. By Corollary 7.42.1 in Bertsekas and Shreve [5], $A^{\oplus\varepsilon} \in \overline{\mathcal{B}}(\mathcal{X})$.

□

Proof of Lemma 2.3.3. Let $\beta = 1/4$. Take any $e \in E$. Since $E = A^{\varepsilon} \backslash A^{\varepsilon)}$, we have the following two implications: 1) $E \subseteq A^{\varepsilon}$ which implies that $d(e, A) \leq \varepsilon$, and 2) $E \cap A^{\varepsilon)} = \varnothing$ which implies that $d(e, A) > \varepsilon$. Combining the two implications, we get that $d(e, A) = \varepsilon$. Hence, for every $r \in (0, \varepsilon]$, there must exist an $a_r \in A$ such that

$\varepsilon \leq \|e - a_r\| < \varepsilon + r/4$. We pick an $x' \in \mathcal{X}$ on the line segment joining a_r and x as follows.

$$t := \frac{r}{2\|e - a_r\|},$$
$$x' := ta_r + (1-t)e.$$

Since $\|e - a_r\| \in [\varepsilon, \varepsilon + r/4)$ and $r \in (0, \varepsilon]$, it is clear that $t \in (0, 1/2)$. From the definition of x', it follows that $\|x' - e\| = t\|e - a_r\| = r/2$. We will now show that $B_{\beta r}(x') \subseteq B_r(e)\backslash E$. For any $y \in B_{\beta r}(x')$, we have the following.

$$\|y - e\| \leq \|y - x'\| + \|x' - e\| \leq \beta r + r/2 < r.$$

Hence, $y \in B_r(e)$. Moreover,

$$\|y - a_r\| \leq \|y - x'\| + \|x' - a_r\| \leq \beta r + (\|e - a_r\| - r/2) < \varepsilon.$$

Hence, $y \in A^{\varepsilon)}$ and so $y \notin E$. Therefore, $B_{\beta r}(x') \subseteq B_r(e)\backslash E$. Hence, we have the following property (call it $(*)$): For any $e \in E$ and any $r \in (0, \varepsilon]$, there is an $x' \in \mathcal{X}$ such that $B_{\beta r}(x') \subseteq B_r(e)\backslash E$. The property $(*)$ is depicted in Figure B.1.

Let $\alpha = \beta(1-\beta)$. Take any $x \in \mathcal{X}$ and $r \in (0, \varepsilon]$. We will now show that there exists $x' \in \mathcal{X}$ such that $B_{\alpha r}(x') \subseteq B_r(x)\backslash E$.

Suppose $x \in E$. Then by the property $(*)$, there exists $x' \in \mathcal{X}$ such that $B_{\alpha r}(x') \subseteq B_{\beta r}(x') \subseteq B_r(x)\backslash E$. Suppose on the other hand $x \notin E$. If $B_{\beta r}(x) \cap E = \varnothing$, then choosing $x' = x$ we have $B_{\alpha r}(x') \subseteq B_{\beta r}(x') \subseteq B_r(x)\backslash E$. If not, then there exists $e \in B_{\beta r}(x) \cap E$. We claim that $B_{(1-\beta)r}(e) \subseteq B_{\beta r}(x)$. Indeed, for any $y \in B_{(1-\beta)r}(e)$ we have

$$\|y - x\| \leq \|y - e\| + \|e - x\| \leq (1-\beta)r + \beta r = r.$$

Since $(1-\beta)r \in (0, \varepsilon]$, by the property $(*)$, there exists $x' \in \mathcal{X}$ such that $B_{\alpha r}(x') = B_{\beta(1-\beta)r}(x') \subseteq B_{(1-\beta)r}(x)\backslash E \subseteq B_r(x)\backslash E$.

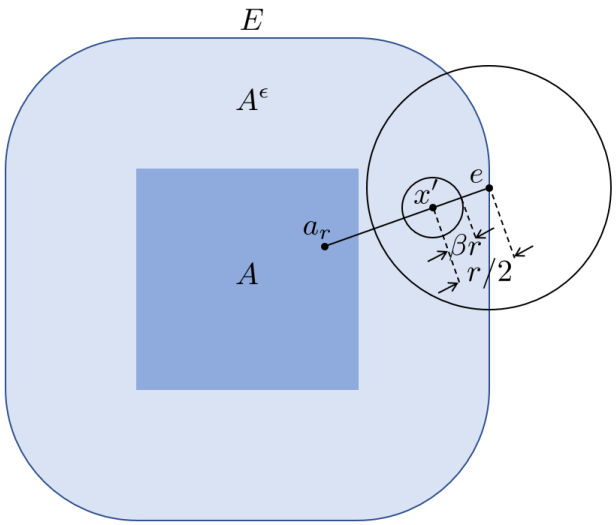


Figure B.1: A depiction of the property $(*)$ in the proof of Lemma 2.3.3. e is an arbitrary point in $E = A^\varepsilon \backslash A^{(\varepsilon)}$. For some $r \in (0, \varepsilon]$, $a_r \in A$ is picked so that $\|e - a_r\| \in [\varepsilon, \varepsilon + r/4)$. x' is a point on the line segment joining a_r and e such that $\|x' - e\| = r/2$. Then, $B_{ar}(x') \subseteq B_r(e) \backslash E$.

□

B.1.2 Proofs from Section 2.3.2

Proof of Lemma 2.3.4. Fix $y \in \mathcal{Y}$ and $w \in \mathcal{W}$. Consider the function $f : D \to \mathbb{R}$ defined as $f(x, x') = -\ell((x', y), w)$, where $D = \{(x, x') \in \mathcal{X}^2 : d(x, x') \leq \varepsilon\}$. Define $f^* : \mathcal{X} \to \mathbb{R}$ as $f^*(x) := \inf_{x' \in B_\varepsilon(x)} f(x, x') = -\sup_{x' \in B_\varepsilon(x)} \ell((x', y), w)$. By Proposition 7.47 in Bertsekas and Shreve[5], f^* is upper semi-analytic. Therefore, the worst-case loss function $\sup_{d(x,x') \leq \varepsilon} \ell((x', y), w)$ is upper semi-analytic and hence universally measurable. Consequently, $R_{\oplus\varepsilon}(\ell, w)$ is well-defined on the measure space $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$. □

Proof of Lemma 2.3.5. Since $\ell((\cdot, y), w)$ is Lebesgue measurable, the set $\{x \in \mathcal{X} : \ell((x, y), w) > t\}^{\oplus\varepsilon}$ is Lebesgue measurable. By Lemma A.0.9, all the level sets of the worst-case loss function $\sup_{d(x,x') \leq \varepsilon} \ell((x', y), w)$ are Lebesgue measurable. Therefore,

$$\begin{aligned} R_{\oplus\varepsilon}(\ell, w) &= \mathbb{E}_{(x,y)\sim\rho} \left[\sup_{d(x,x') \leq \varepsilon} \ell((x', y), w) \right] \\ &= \mathbb{E}_{y\sim\rho_y} \mathbb{E}_{x\sim\rho_{x|y}} \left[\sup_{d(x,x') \leq \varepsilon} \ell((x', y), w) \right], \end{aligned}$$

is well-defined. □

B.2 Proofs from Section 2.4

B.2.1 Proofs from Section 2.4.1

Proof of Lemma 2.4.1. Let $\mu' \in \mathcal{P}(\mathcal{X})$ be such that $W_\infty(\mu, \mu') \leq \varepsilon$. Then there exists a coupling $\lambda \in \Pi(\mu', \mu)$ such that for $(x, x') \sim \lambda$, $d(x, x') \leq \varepsilon$ λ-a.e. Hence,

$$\mu'(A) = \lambda(A \times \mathcal{X}) = \lambda(A \times A^{\oplus\varepsilon}) \leq \lambda(\mathcal{X} \times A^{\oplus\varepsilon}) = \mu(A^{\oplus\varepsilon}).$$

Since the choice of μ' was arbitrary in the set $\{\nu \in \mathcal{P}(\mathcal{X}) : W_\infty(\mu, \nu) \leq \varepsilon\}$, we have,

$$\sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mu'(A) \leq \mu(A^{\oplus\varepsilon}).$$

Now we show the inequality in the opposite direction. Like in the proof of Lemma 2.3.2, consider the function $f : D \to \mathbb{R}$ defined as $f(x, x') = -1\{x' \in A\}$, where $D = \{(x, x') \in \mathcal{X}^2 : d(x, x') \leq \varepsilon\}$. Define $f^* : \mathcal{X} \to \mathbb{R}$ as $f^*(x) := \inf_{x' \in B_\varepsilon(x)} f(x, x')$. As shown in the proof of Lemma 2.3.2, $f^*(x) = -1\{x \in A^{\oplus\varepsilon}\}$. By Proposition 7.50(a) in Bertsekas and Shreve [5], there exists a measurable function $\varphi : \mathcal{X} \to \mathcal{X}$ such that $|f^*(x) - f(x, \varphi(x))| < \delta$ for any $\delta > 0$. Since f and f^* are both 0-1 valued functions, we get $f^*(x) = f(x, \varphi(x))$ for all $x \in \mathcal{X}$ by choosing $\delta = 1/2$. Moreover, by Proposition 7.50(a) in Bertsekas and Shreve [5], $Gr(\varphi) \subseteq D$ i.e., $d(x, \varphi(x)) \leq \varepsilon$ for all $x \in \mathcal{X}$. Therefore,

$$\sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mu'(A) \geq \varphi_{\sharp\mu}(A) = \mu(\varphi^{-1}(A)) = \mu(A^{\oplus\varepsilon}).$$

Hence, $\sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mu'(A) = \varphi_{\sharp\mu}(A) = \mu(A^{\oplus\varepsilon})$ for any set $A \in \mathcal{B}(\mathcal{X})$. □

Proof of Lemma 2.4.2. Let $\mu' \in \overline{\mathcal{P}}(\mathcal{X})$ be such that $W_\infty(\mu, \mu') \leq \varepsilon$. Then there exists $\lambda \in \Pi(\mu', \mu)$ such that $\lambda(\{(x, x') \in \mathcal{X}^2 : d(x, x') > \varepsilon\}) = 0$. Then,

$$\begin{aligned}
\mathbb{E}_{x \sim \mu'}[\varphi(x)] &= \mathbb{E}_{(x,x') \sim \lambda}[\varphi(x)] \\
&= \mathbb{E}_{(x,x') \sim \lambda}\left[\sup_{x \in B_\varepsilon(x')} \varphi(x')\right]
\end{aligned}$$

$$= \mathbb{E}_{x' \sim \mu} \left[\sup_{x \in B_\varepsilon(x')} \varphi(x') \right].$$

Since the above inequality is true for any $\mu' \in \overline{\mathcal{P}}(\mathcal{X})$ satisfying $W_\infty(\mu, \mu') \leq \varepsilon$, we have,

$$\sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mathbb{E}_{x \sim \mu'}[\varphi(x)] \leq \mathbb{E}_{x' \sim \mu} \left[\sup_{x \in B_\varepsilon(x')} \varphi(x') \right].$$

Now we will show the inequality in the opposite direction. Consider the function $f : D \to \mathbb{R}$ defined as $f(x, x') = -\varphi(x')$, where $D = \{(x, x') \in \mathcal{X}^2 : d(x, x') \leq \varepsilon\}$. Define $f^* : \mathcal{X} \to \mathbb{R}$ as $f^*(x) := \inf_{x' \in B_\varepsilon(x)} f(x, x') = -\sup_{x' \in B_\varepsilon(x)} \varphi(x')$. Choose a $\delta > 0$. By Proposition 7.50(a) in Bertsekas and Shreve [5], there exists a universally measurable function $m_\delta : \mathcal{X} \to \mathcal{X}$ such that $|f^*(x) - f(x, m_\delta(x))| \leq \delta$ and $d(x, m_\delta(x) \leq \varepsilon)$ for all $x \in \mathcal{X}$. Hence,

$$\begin{aligned}
\mathbb{E}_{x \sim \mu} \left[\sup_{d(x,x') \leq \varepsilon} \varphi(x') \right] &= \mathbb{E}_{x \sim \mu}[-f^*(x)] \\
&\leq \mathbb{E}_{x \sim \mu}[-f(x, m_\delta(x))] + \delta \\
&= \mathbb{E}_{x \sim \mu}[\varphi(m_\delta(x))] + \delta \\
&= \mathbb{E}_{x \sim m_{\delta \sharp \mu}}[\varphi(x)] + \delta \\
&\leq \sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mathbb{E}_{x \sim \mu'}[\varphi(x)] + \delta,
\end{aligned}$$

where the last inequality follows because $W_\infty(\mu, m_{\delta \sharp \mu}) \leq \varepsilon$ because $d(x, m_\delta(x) \leq \varepsilon)$ for all $x \in \mathcal{X}$. Taking $\delta \to 0$, we get the following inequality.

$$\mathbb{E}_{x \sim \mu} \left[\sup_{d(x,x') \leq \varepsilon} \varphi(x') \right] \leq \sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mathbb{E}_{x \sim \mu'}[\varphi(x)].$$

Combining the above inequality with the reverse inequality shown previously, we obtain (2.18).

Suppose the function φ is upper semi-continuous. Then f is lower semi-continuous. Hence, for every $x \in \mathcal{X}$, there exists x^* in the compact set $B_\varepsilon(x)$ such that $\inf_{x' \in B_\varepsilon(x)} f(x, x') = f(x, x^*)$. By Proposition 7.50(b), there exists a universally measurable function $m : \mathcal{X} \to \mathcal{X}$ such that $f^*(x) = f(x, m(x))$ for all $x \in \mathcal{X}$. Hence, we have

$$\sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mathbb{E}_{x \sim \mu'}[\varphi(x)] = \mathbb{E}_{x \sim \mu} \left[\sup_{d(x,x') \leq \varepsilon} \varphi(x') \right] = \mathbb{E}_{x \sim \mu} \left[\varphi(m(x)) \right] = \mathbb{E}_{x \sim m_{\sharp \mu}} \left[\varphi(x) \right].$$

Therefore, $\mu^* := m_{\sharp\mu}$ attains the supremum on the left side of the above equation.

□

B.2.2 Proofs from Section 2.4.2

Proof of Lemma 2.4.3. The following properties of v are trivially true: $v(\varphi) = 0$, $v(\mathcal{X}) = 1$ and $v(A) \leq v(B)$ for $A \subseteq B$.

Consider a sequence of sets (A_n) in $\mathcal{X}$ such that $A_k \subseteq A_{k+1}$ for $k \in \mathcal{N}$. Let $A = \cup_n A_n$. That is, $A_n \uparrow A$. Then by Lemma A.0.1 we have, $A^{\oplus\varepsilon} = \cup_n A_n^{\oplus\varepsilon}$. Hence, $A_n^{\oplus\varepsilon} \uparrow A^{\oplus\varepsilon}$ and by the continuity of measure, $v(A_n) = \mu(A_n^{\oplus\varepsilon}) \uparrow \mu(A^{\oplus\varepsilon}) = v(A)$.

Consider a sequence of closed sets (F_n) in $\mathcal{X}$ such that $F_k \supseteq F_{k+1}$ for $k \in \mathbb{N}$. Let $F = \cap_n F_n$. That is, $F_n \downarrow F$. By Lemma A.0.2, $F_n^{\oplus\varepsilon} \downarrow F^{\oplus\varepsilon}$. Hence, by the continuity of measure, we have $v(F_n) = \mu(F_n^{\oplus\varepsilon}) \downarrow \mu(F^{\oplus\varepsilon}) = v(F)$.

For any two sets $A, B \in \mathcal{L}(\mathcal{X})$,

$$\begin{aligned} v(A \cup B) &= \mu((A \cup B)^{\oplus\varepsilon}) \\ &\overset{(i)}{=} \mu(A^{\oplus\varepsilon} \cup B^{\oplus\varepsilon}) \\ &= \mu(A^{\oplus\varepsilon}) + \mu(B^{\oplus\varepsilon}) - \mu(A^{\oplus\varepsilon} \cap B^{\oplus\varepsilon}) \\ &\overset{(ii)}{\leq} \mu(A^{\oplus\varepsilon}) + \mu(B^{\oplus\varepsilon}) - \mu((A \cap B)^{\oplus\varepsilon}) \\ &= v(A) + v(B) - v(A \cap B), \end{aligned}$$

where (i) and (ii) follow from Lemma A.0.1. Hence, v is a 2-alternating capacity. □

Proof of Lemma 2.4.4. Let $\mu' \in \mathcal{P}(\mathcal{X})$ be such that $W_\infty(\mu, \mu') \leq \varepsilon$. Then there exists a coupling $\gamma \in \Pi(\mu', \mu)$ such that for $(x, x') \sim \gamma$, $d(x, x') \leq \varepsilon$ γ-a.e. Hence,

$$\mu'(A) = \gamma(A \times \mathcal{X}) = \gamma(A \times A^{\oplus\varepsilon}) \leq \gamma(\mathcal{X} \times A^{\oplus\varepsilon}) = \mu(A^{\oplus\varepsilon}).$$

Since the choice of μ' was arbitrary in the set $\{\nu \in \mathcal{P}(\mathcal{X}) : W_\infty(\mu, \nu) \leq \varepsilon\}$, we have,

$$\sup_{W_\infty(\mu,\mu') \leq \varepsilon} \mu'(A) \leq \mu(A^{\oplus\varepsilon}).$$

We will now show the inequality in the reverse direction. By Lemma 2.4.3, $A \mapsto \mu(A^{\oplus\varepsilon})$ is a 2-alternating capacity. Hence by Lemma 2.5 in Huber and Strassen[46], for any Lebesgue measurable $A \subseteq \mathcal{X}$, there exists a $\nu \in \mathcal{P}(\mathcal{X})$ such that $\nu(A) = \mu(A^{\oplus\varepsilon})$ and $\nu(B) \leq \mu(B^{\oplus\varepsilon})$ for all Lebesgue measurable $B \subseteq \mathcal{X}$. For such a ν, it is clear that $W_\infty(\mu, \nu) \leq \varepsilon$. Hence,

$$\sup_{W_\infty(\mu,\mu')\leq\varepsilon} \mu'(A) \geq \nu(A) = \mu(A^{\oplus\varepsilon}).$$

Hence, $\sup_{W_\infty(\mu,\mu')\leq\varepsilon} \mu'(A) = \nu(A) = \mu(A^{\oplus\varepsilon})$. □

C

Proofs from Chapter 3

C.1 Proofs from Section 3.1

C.1.1 Proofs from Section 3.1.2

Proof of Corollary 3.1.1. From Theorem 7, we have

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

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

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

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

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

□

Proof of Theorem 8. Since $W_\infty(\mu,\nu) = \inf\{\delta > 0 \mid \mu(A) \leq \nu(A^\delta) \text{ for all measurable } A\}$, if $W_\infty(\mu,\nu) \leq 2\varepsilon$, then $\mu(A) \leq \nu(A^{2\varepsilon})$ for all closed sets A. Hence,

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

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

For the reverse direction, suppose that $D_\varepsilon(\mu,\nu) = 0$. This means there exists a sequence of couplings $(\pi_i)_{i\geq 1}$ such that $\mathbb{E}_{\pi_i} c_\varepsilon(x,x') \to 0$ where $\pi_i \in \Pi(\mu,\nu)$. We now show that the sequence of measures (π_i) is tight. Given a $\delta > 0$, let $E \subseteq \mathcal{X}$ be a compact set such that $\min\{\mu(E), \nu(E)\} > 1 - \delta/2$. Then,

$$\pi_i((E \times E)^c) \leq \mu(E^c) + \nu(E^c) < \delta.$$

Hence, by Prokhorov's theorem (for reference, see Theorem 5.1 in Billingsley[10]), there is a subsequence of (π_i) that converges weakly to a coupling $\pi^* \in \Pi(\mu,\nu)$. Since c is a lower semicontinuous cost function, the coupling π^* satisfies $\mathbb{E}_{\pi^*} c_\varepsilon(x,x') = 0$, or equivalently, $\operatorname{ess\,sup}_{(x,x')\sim\pi^*} d(x,x') \leq 2\varepsilon$. Using the definition of W_∞, we conclude that $W_\infty(\mu,\nu) \leq 2\varepsilon$. □

C.1.2 Proofs from Section 3.1.3

The following lemma presents a discrete version of the generalized Strassen's theorem.

Lemma C.1.1. *Let $\mathcal{X}_n = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$. Let $p = (p_i)_{i=1}^n, q = (q_i)_{i=1}^n$ be such that $p_i, q_i \geq 0$ for $i \in [n]$ and $\sum_i p_i \leq \sum_i q_i$. Let $\varepsilon > 0$. For $A \subseteq \mathcal{X}_n$, let $A^\varepsilon := \{x \in \mathcal{X}_n : d(x,x') \leq \varepsilon, \text{for some } x' \in A\}$. For $A \subseteq \mathcal{X}_n$, let*

$p(A) = \sum_{i:x_i \in A} p_i$ *and* $q(A) = \sum_{i:x_i \in A} q_i$. *For* $i, j \in [n]$, *let* $c_{ij} = 1\{d(x_i, x_j) > 2\varepsilon\}$. *Then,*

$$\max_{A \subseteq \mathcal{X}_n} p(A) - q(A^{2\varepsilon}) = \min_{\substack{x_{ij} \geq 0 \\ \sum_j x_{ij} = p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} c_{ij} x_{ij}. \tag{C.1}$$

Proof. For $i, j \in [n]$, define $d_{ij} := 1 - c_{ij}$. Then,

$$\min_{\substack{x_{ij} \geq 0 \\ \sum_j x_{ij} = p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} c_{ij} x_{ij} = \sum_i p_i - \max_{\substack{x_{ij} \geq 0 \\ \sum_j x_{ij} = p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} d_{ij} x_{ij} \tag{C.2}$$

Consider the following modification to the linear program on the right hand side of (C.2), where the constraint $\sum_j x_{ij} = p_i$ is replaced by $\sum_j x_{ij} \leq p_i$.

$$\max_{\substack{x_{ij} \geq 0 \\ \sum_j x_{ij} \leq p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} d_{ij} x_{ij}. \tag{C.3}$$

We will show that the above linear program is equivalent to the linear program on the right hand side of (C.2). Since the above linear program is bounded and feasible, it admits a solution. Let $\{x^*_{ij}\}_{i,j \in [n]}$ be the solution to (C.3). Suppose there exists $m \in [n]$ such that $\sum_j x^*_{mj} < p_m$. Let $s = p_m - \left(\sum_j x^*_{mj}\right) > 0$. For $j \in [n]$, define $s_j := q_j - \sum_i x^*_{ij}$. Then,

$$\sum_j s_j = \sum_j q_j - \sum_{i,j} x_{ij} \geq \sum_i p_i - \left(\left(\sum_{i \neq m} p_i \right) + p_m - s \right) = s.$$

Therefore, $\sum_j s_j \geq s$. Let k be the largest integer for which $\sum_{j=1}^k s_j < s$. Define,

$$y_{ij} = \begin{cases} x^*_{ij} & i \neq m, \\ x^*_{mj} + s_j & i = m, j \leq k, \\ x^*_{mk} + s - \sum_{j=1}^k s_j & i = m, j = k+1, \\ x^*_{mj} & i = m, j \geq k+1. \end{cases} \tag{C.4}$$

By the above definition we have,

$$\sum_j y_{ij} = \begin{cases} \sum_j x^*_{ij} & i \neq m, \\ \sum_j x^*_{ij} + s & i = m. \end{cases}$$

$$\sum_i y_{ij} = \begin{cases} \sum_i x^*_{mj} + s_j & j \leq k, \\ \sum_i x^*_{mk} + s - \sum_{j=1}^{k} s_j & j = k+1, \\ \sum_i x^*_{mj} & j \geq k+1. \end{cases}$$

Combining the above with the definitions of k, s and $\{s_j\}_{j\in[n]}$, we see that $\sum_j y_{ij} \leq p_i$ and $\sum_i y_{ij} \leq q_j$. Moreover, $y_{ij} \geq x_{ij}$ for all $i, j \in [n]$. Hence, $\sum_{ij} d_{ij} y_{ij} \geq \sum_{ij} d_{ij} x_{ij}$. Therefore, any solution $\{x^*_{ij}\}_{i,j\in[n]}$ for which there exists $m \in [n]$ such that $\sum_j x^*_{mj} < p_m$, can be improved to a solution $\{y_{ij}\}_{i,j\in[n]}$ for which $\sum_j y_{mj} = p_m$. Hence,

$$\max_{\substack{x_{ij}\geq 0 \\ \sum_j x_{ij} = p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} d_{ij} x_{ij} = \max_{\substack{x_{ij}\geq 0 \\ \sum_j x_{ij} \leq p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} d_{ij} x_{ij}. \tag{C.5}$$

Since the maximization in (C.5) is a linear program in canonical form, we employ the strong duality theorem (for a reference, see Chapter 6 in Matousek and Gartner[60]) to get the following.

$$\max_{\substack{x_{ij}\geq 0 \\ \sum_j x_{ij} \leq p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} d_{ij} x_{ij} = \min_{\substack{u_i, v_i \geq 0 \\ u_i + v_j \geq d_{ij}}} \sum_i (p_i u_i + q_i v_i). \tag{C.6}$$

Since $d_{ij} \in \{0, 1\}$, we may assume $u_i, v_i \leq 1$ for the minimization in (C.6) without violating other constraints because any decrease of u_i, v_i down to 1 will only decrease the value of $\sum_i (p_i u_i + q_i v_i)$, which we seek to minimize. Defining $w_i := 1 - u_i$, we have the following from (C.2) and (C.6).

$$\min_{\substack{x_{ij}\geq 0 \\ \sum_j x_{ij} = p_i \\ \sum_i x_{ij} \leq q_j}} \sum_{i,j} c_{ij} x_{ij} = \max_{\substack{w_i, v_i \in [0,1] \\ w_i - v_j \leq c_{ij}}} \sum_i (p_i w_i - q_i v_i). \tag{C.7}$$

The optimal w^*_i, v^*_i that achieve the maximum in (C.7) must lie at one of the vertices of the polyhedron supported by the hyperplanes, $w_i = 0, w_i = 1, v_i = 0, v_i = 1$ and $w_i - v_j = c_{ij}$. Hence, $w^*_i, v^*_i \in \{0, 1\}$. Moreover if $c_{ij} = 0$

and $w_i^* = 1$ for some $i, j \in [n]$, then $v_j^* = 1$. On the other hand if $c_{ij} = 1$, then v_j^* can be set to 0 without violating other constraints and without decreasing the maximization objective. Therefore, setting $A := \{x_i \in \mathcal{X}_n : w_i^* = 1\}$, we see that the maximum in (C.7) equals the maximum in (C.1). □

Proof of Theorem 9. Let $(\gamma_n)_{n=1}^{\infty}$ be a non-negative, monotonically decreasing sequence converging to 0. Let $(x_n)_{n=1}^{\infty}$ be a dense sequence in $\mathcal{X}$. Define a function $f : \mathcal{X} \to \{x_n\}_{n=1}^{\infty}$ such that $f(x) = x_k$ for the least integer k with $d(x, x_k) < \gamma_n$. Let $H_n = \{x_1, \ldots, x_n\}$. Let s_n be the least positive integer such that,

$$\mu(f^{-1}(H_{s_n-1})) > \mu(\mathcal{X}) - \gamma_n, \tag{C.8}$$

$$\nu(f^{-1}(H_{s_n-1})) > \nu(\mathcal{X}) - \gamma_n. \tag{C.9}$$

Given n, construct a discrete measure μ_n supported on the finite set H_{s_n} such that $\mu_n(x_k) := \mu(f^{-1}(x_k))$ for $k \in [s_n - 1]$ and $\mu_n(\mathcal{X}) = \mu(\mathcal{X})$. Similarly, construct ν_n supported on H_{s_n} such that $\nu_n(x_k) := \nu(f_n^{-1}(x_k))$ for $k \in [s_n - 1]$ and $\nu_n(\mathcal{X}) = \nu(\mathcal{X})$.

Let $A \in \mathcal{B}(\mathcal{X})$. We have,

$$\begin{aligned}
\mu_n(A) &\overset{(i)}{=} \mu_n(A \cap H_{s_n}) \\
&\overset{(ii)}{<} \mu_n(A \cap H_{s_n-1}) + \gamma_n \\
&\overset{(iii)}{=} \mu(f^{-1}(A \cap H_{s_n-1})) + \gamma_n \\
&\overset{(iv)}{\leq} \mu(A^{\gamma_n}) + \gamma_n,
\end{aligned} \tag{C.10}$$

where (i) follows from the fact that μ_n is supported on H_{s_n}, (ii) follows from (C.8), (iii) follows from the definition of μ_n and (iv) follows because of the following: For any $y \in A \cap H_{s_n-1}$, $f^{-1}(y) \subseteq \{x \in \mathcal{X} : d(x, y) < \gamma_n\} \subseteq A^{\gamma_n}$. Hence, $f^{-1}(A \cap H_{s_n-1}) \subseteq A^{\gamma_n}$. Applying (C.10), with A^c instead of A, we have the following.

$$\mu(A^{-\gamma_n}) - \gamma_n \leq \mu_n(A) \leq \mu(A^{\gamma_n}) + \gamma_n. \tag{C.11}$$

Letting $n \to \infty$ in (C.11) and using Lemma A.0.3, we get that $\limsup_n \mu_n(A) \leq \mu(A)$ for all closed subsets A of $\mathcal{X}$. Hence, by applying the Portmanteau theorem (Theorem 2.1 in Billingsley[9]), we conclude that the sequence of measures $(\mu_n)_{n=1}^{\infty}$ converges weakly to μ. Similarly, $\nu_n \to \nu$ weakly.

For any fixed n, we apply Lemma C.1.1 to the measures μ_n, ν_n on the finite space H_{s_n} to get the following.

$$\max_{A \subseteq H_{s_n}} \mu_n(A) - \nu_n(A^{2\varepsilon+4\gamma_n}) = \min_{\substack{x_{ij} \geq 0 \\ \sum_j x_{ij} = \mu_n(x_i) \\ \sum_i x_{ij} \leq \nu_n(x_j)}} \sum_{i,j} x_{ij} 1\{d(x_i, x_j) > 2\varepsilon + 4\gamma_n\}, \tag{C.12}$$

where the indices i, j run over $[s_n]$. We have that $\mu_n(\mathcal{X}) = \mu(\mathcal{X}) \leq \nu(\mathcal{X}) = \nu(\mathcal{X})$. Define a coupling $\pi_n \in \Pi(\mu_n, \nu_n)$ supported on $H_{s_n} \times H_{s_n}$ using the optimal solution $\{x_{ij}\}_{i,j \in [s_n]}$ to the minimization in (C.12) by setting $\pi_n(i,j) = x^*_{ij}$. Let $T_n \subseteq H_{s_n}$ be the set that achieves the maximum in (C.12).

We will now construct a candidate coupling for the infimum in (3.7). Since μ, ν are finite measures on a Polish space, they are tight (see for example, Theorem 1.3 in Billingsley[9]). Hence, given a $\delta > 0$, there exists a compact set $K \subseteq \mathcal{X}$ such that $\min\{\mu(K^c), \nu(K^c)\} < \delta/3$. Since μ_n and ν_n converge weakly to μ and ν respectively, choose N large enough so that $\min\{\mu_n(K^c), \nu_n(K^c)\} < \delta/2$ for all $n \geq N$. Let ν'_n be the second marginal of the coupling π_n. Then, $\nu'_n \preceq \nu_n$. By union bound, we have the following.

$$\pi_n((K \times K)^c) \leq \mu_n(K^c) + \nu'_n(K^c) \leq \mu_n(K^c) + \nu_n(K^c) < \delta. \tag{C.13}$$

Hence, the sequence $(\pi_n)_{n \geq N}$ is uniformly tight. Hence, by Prokhorov's theorem (for reference, see Theorem 5.1 in Billingsley[9]), there is a subsequence (π_{n_k}) of $(\pi_n)_{n \geq N}$ that converges weakly to some measure $\pi^* \in \mathcal{M}(\mathcal{X} \times \mathcal{X})$. Moreover, $\pi^* \in \Pi(\mu, \nu)$ by virtue of the constraints imposed on the converging subsequence of $(\pi_n)_{n \geq N}$.

Let $\Phi = \sup_{A \in \mathcal{B}(\mathcal{X})} \mu(A) - \nu(A^{2\varepsilon})$ and $\Psi = \mathcal{T}_{c_\varepsilon}(\mu, \nu)$. For any n we have,

$$\begin{aligned}
\pi_n(d(x_i, x_j) > 2\varepsilon + 4\gamma_n) &\overset{(i)}{=} \mu_n(T_n) - \nu_n(T_n^{2\varepsilon+4\gamma_n}) \\
&\overset{(ii)}{\leq} (\mu(T_n^{\gamma_n}) + \gamma_n) - (\nu((T_n^{2\varepsilon+4\gamma_n})^{-\gamma_n}) - \gamma_n) \\
&\overset{(iii)}{\leq} \mu(T_n^{\gamma_n}) - \nu((T_n^{2\varepsilon+4\gamma_n-\gamma_n-\gamma_n/2}) + 2\gamma_n \\
&\leq \mu(T_n^{2\gamma_n}) - \nu((T_n^{2\varepsilon+2\gamma_n}) + 2\gamma_n \\
&\overset{(iv)}{\leq} \Phi + 2\gamma_n,
\end{aligned} \tag{C.14}$$

where (i) follows from the definition of π_n and T_n, (ii) follows from (C.11), (iii) follows from Lemma A.0.7 and

(iv) follows from the definition of Φ. Further,

$$\begin{aligned}\Psi &= \inf_{\pi\in\Pi(\mu,\nu)} \pi[d(x,x') > 2\varepsilon] \\ &\overset{(i)}{\leq} \pi^*[d(x,x') > 2\varepsilon] \\ &\overset{(ii)}{\leq} \liminf_{n_k} \pi_{n_k}[d(x,x') > 2\varepsilon] \\ &\leq \limsup_{n} \pi_n[d(x,x') \geq 2\varepsilon] \\ &\overset{(iii)}{\leq} \Phi, \end{aligned} \tag{C.15}$$

where (i) follows because $\pi^* \in \Pi(\mu,\nu)$, (ii) follows from Portmanteau's theorem because (π_{n_k}) that converges to π^* and the set $\{(x,x') \in \mathcal{X}^2 : d(x,x') > 2\varepsilon\}$ is an open set, and (iii) follows by taking $n \to \infty$ in (C.14).

To show the inequality $\Phi \leq \Psi$, consider a sequence of measures $(\lambda_n)_{n=1}^{\infty}$ such that $\lambda_n \in \Pi(\mu,\nu)$ and $\lim_n \lambda_n[d(x,x') > \varepsilon] = \Psi$. For any $A \in \mathcal{B}(\mathcal{X})$,

$$\begin{aligned}\mu(A) &= \lambda_n[x \in A, x' \in A^\varepsilon] + \lambda_n[x \in A, x' \notin A^\varepsilon] \\ &\leq \nu(A^\varepsilon) + \lambda_n[d(x,x') > \varepsilon].\end{aligned}$$

Letting $n \to \infty$, we have $\mu(A) - \nu(A^\varepsilon) \leq \Psi$ for all $A \in \mathcal{B}(\mathcal{X})$. Hence, $\Phi \leq \Psi$. Combining this with (C.15), we conclude $\Phi = \Psi$. □

C.2 Proofs from Section 3.2

Proof of Theorem 16. Like in the proof for Theorem 14, we prove Theorem 16 by partitioning the real line into several regions for μ and ν, and transporting mass between these regions. Figure C.1 shows the optimal coupling for the case when $I^{2\varepsilon} \subseteq J$.

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

$$D_\varepsilon(\mu,\nu) \geq \mu(A) - \nu(A^{2\varepsilon}) = 1 - \nu(I^{2\varepsilon}). \tag{C.16}$$

To establish the upper bound, we need to find a coupling that transports μ to ν such that the cost of transportation

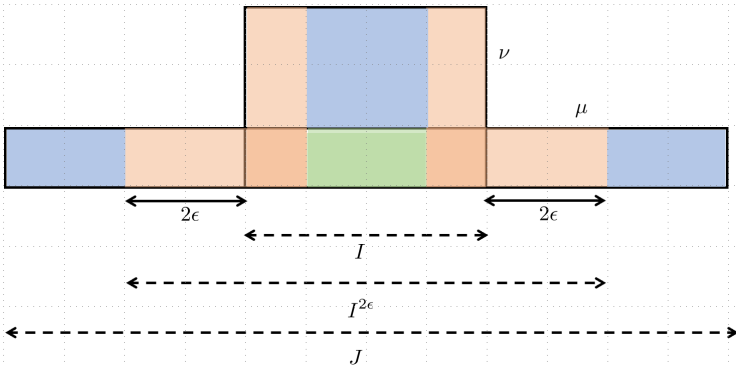


Figure C.1: Optimal coupling for two uniform distributions. The region shaded in green is kept in place (at no cost). The two regions shaded in orange are transported monotonically from either side at a cost not exceeding 2ε per unit mass. The remaining region in blue is moved at the cost of 1 per unit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

is bounded above by $1 - \nu(I^{2\varepsilon})$. 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: $2\varepsilon < w_2 - w_1$**.**

We split the analysis into the following five sub-cases.

Case 1(a): $c \in [w_1 + w_2 + 2\varepsilon, \infty)$**.**

In this case, the intervals I and J are separated by at least 2ε. Hence, $\nu(I^{2\varepsilon}) = 0$, and therefore, $D_\varepsilon(\mu, \nu) \geq 1 - \nu(I^{2\varepsilon}) = 1$. Combining this with the fact that $D_\varepsilon(\mu, \nu) \leq 1$, we get that $D_\varepsilon(\mu, \nu) = 1 = 1 - \nu(I^{2\varepsilon})$.

Case 1(b): $c \in [-w_1 + w_2 - 2\varepsilon, w_1 + w_2 + 2\varepsilon)$**.**

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

$$\frac{w_1 - u}{2w_1} = \mu([u, w_1]) = \nu([c - w_2, w_1 + 2\varepsilon]) = \frac{w_1 + 2\varepsilon - c + w_2}{2w_2}$$
$$\implies u = w_1 - \frac{w_1}{w_2}(w_1 + 2\varepsilon - c + w_2).$$

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

Let μ_0 be the restriction of μ to $[u, w_1]$ and ν_0 be the restriction of ν to $[c - w_2, w_1 + 2\varepsilon]$. Then, by construction, $\mu_0(\mathbb{R}) = \nu_0(\mathbb{R})$. By Lemma 3.2.1, we have a monotone transport map $T : [u, w_1] \to [c - w_2, w_1 + 2\varepsilon]$ that transports μ_0 to ν_0 given by $T(x) = \frac{w_1 + 2\varepsilon - c + w_2}{w_1 - u}(x - u) + (c - w_2)$. Note that T transports u to $c - w_2$ and w_1

to $w_1 + 2\varepsilon$. Also, $T(x) > x$. Since T has a slope greater than 1, $T(x) - x$ is an increasing function. Moreover, $T(w_1) - w_1 = 2\varepsilon$ and $T(u) - u = (c - w_2) - u < 2\varepsilon$. Hence, $|T(x) - x| \leq 2\varepsilon$ for all $x \in [u, w_1]$. Hence, $D_\varepsilon(\mu_0, \nu_0) = 0$. Therefore, $D_\varepsilon(\mu, \nu) \leq 1 - \min(\mu_0, \nu_0) = 1 - \nu([c, w_1 + 2\varepsilon]) = 1 - \nu(I^{2\varepsilon})$. Combining with the lower bound in (C.16), we conclude that $D_\varepsilon(\mu, \nu) = 1 - \nu(I^{2\varepsilon})$.

Case 1(c): $c \in (-w_2 + w_1 + 2\varepsilon, -w_1 + w_2 - 2\varepsilon)$.

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

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

Combining with the lower bound in (C.16), we conclude that $D_\varepsilon(\mu, \nu) = 1 - \nu(I^{2\varepsilon})$.

Case 1(d): $c \in (-w_1 - w_2 - 2\varepsilon, w_1 - w_2 + 2\varepsilon]$.

The geometry of this case is a mirror image of that in case 1b. Hence, just as in case 2, we have $D_\varepsilon(\mu, \nu) = 1 - \nu(I^{2\varepsilon})$.

Case 1(e): $c \in (-\infty, -w_1 - w_2 - 2\varepsilon]$.

Like in case 1, the intervals I and J are separated by at least 2ε. Hence, similar to case 1, we get that $D_\varepsilon(\mu, \nu) = 1 = 1 - \nu(I^{2\varepsilon})$.

Case 2: $2\varepsilon \geq w_2 - w_1$.

In this case, we have the following sub-cases.

Case 2(a): $c \in [w_1 + w_2 + 2\varepsilon, \infty)$.

Like in case 1a, the intervals I and J are separated by at least 2ε. Hence, $D_\varepsilon(\mu, \nu) = 1 = 1 - \nu(I^{2\varepsilon})$.

Case 2(b): $c \in [w_1 - w_2 + 2\varepsilon, w_1 + w_2 + 2\varepsilon)$**.**

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

Case 2(c): $c \in (-w_1 + w_2 - 2\varepsilon, w_1 - w_2 + 2\varepsilon)$**.**

In this case, the supports of μ and ν are within 2ε of each other. More specifically, $J \subseteq I^{2\varepsilon}$. Hence, $\nu(I^{2\varepsilon}) = 1$. Let T denote the monotone transport map from μ and ν as defined in Lemma 3.2.1. 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 $|T(x) - x| \leq 2\varepsilon$. Hence, $D_\varepsilon(\mu, \nu) = 0 = 1 - \nu(I^{2\varepsilon})$.

Case 2(d): $c \in (-w_1 - w_2 - 2\varepsilon, -w_1 + w_2 - 2\varepsilon]$**.**

This case is a mirror image of case 2b and hence the result $D_\varepsilon(\mu, \nu) = 1 - \nu(I^{2\varepsilon})$ remains the same.

Case 2(e): $c \in (-\infty, -w_1 - w_2 - 2\varepsilon]$**.**

Like in case 1a, the intervals I and J are separated by at least 2ε. Hence, $D_\varepsilon(\mu, \nu) = 1 = 1 - \nu(I^{2\varepsilon})$.

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

□

Proof of Theorem 17. We have the following cases:

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

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

$$D_\varepsilon(\mu, \nu) = \sup_{A \text{ closed}} \mu(A^{\ominus\varepsilon}) - \nu(A^{\oplus\varepsilon}) \geq \mu(A^{\ominus\varepsilon}) - \nu(A^{\oplus\varepsilon}) = 1.$$

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

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

In this case,

$$|(m_2 + \varepsilon) - (m_1 - \varepsilon)| = |(m_2 - m_1) + 2\varepsilon| = (m_2 - m_1) + 2\varepsilon < \delta_2 - \delta_1,$$

$$|(m_2 - \varepsilon) - (m_1 + \varepsilon)| = |(m_2 - m_1) - 2\varepsilon| \leq |m_2 - m_1| + 2\varepsilon < \delta_2 - \delta_1.$$

Hence, by Lemma 3.2.6, the equations $f(x+\varepsilon) = g(x-\varepsilon)$ and $f(x-\varepsilon) = g(x+\varepsilon)$ have exactly two solutions each, on the supports of $\Delta(m_1-\varepsilon, \delta_1)$ and $\Delta(m_1+\varepsilon, \delta_1)$ respectively. Hence, l must be the minimum of the two solutions to $f(x+\varepsilon) = g(x-\varepsilon)$ and r must be the maximum of the two solutions to $f(x-\varepsilon) = g(x+\varepsilon)$. As in the proof of Theorem 15, we divide the real line into five regions as shown in Table C.1, where l' is the leftmost point such that $\mu([l+\varepsilon, l']) = \nu([l-\varepsilon, l'])$ and r' is the rightmost point such that $\mu([r', r-\varepsilon]) = \nu([r', r+\varepsilon])$. Observe that by construction, $f(x) \leq g(x+2\varepsilon)$ for $x \in [r-\varepsilon, m_1+\delta_1]$. Hence by Lemma 3.2.3, $D_\varepsilon(\mu_{++}, \nu_{++}) = 0$. Similarly, we also get $D_\varepsilon(\mu_{--}, \nu_{--}) = 0$.

We will now use Lemma 3.2.4 to show that $D_\varepsilon(\mu_{-}, \nu_{-}) = 0$. Let $a = l-\varepsilon, a' = l+\varepsilon, b = l'$ and $\tilde{t} = l'-2\varepsilon$. 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 $[a', b] = [a+2\varepsilon, b]$. (2) $g(x) \geq f(x)$ for $x \in [a, t)$ and $f(x) \geq g(x)$ for $x \in (t, b]$. (3) $g(x) \leq f(x+2\varepsilon)$ for $x \in [a, \tilde{t})$ and the the interval $(\tilde{t}, b-2\varepsilon]$ is empty because $\tilde{t} = b-2\varepsilon$. Hence, $D_\varepsilon(\mu_-, \nu_-) = 0$. Similarly, $D_\varepsilon(\mu_+, \nu_+) = 0$.

Finally, $D_\varepsilon(\mu_0, \nu_0) = 0$. This is because $f(x) \geq g(x)$ 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 12, we can upper bound $D_\varepsilon(\mu, \nu)$ as follows.

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

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

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

In this case,

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

Hence, $|(m_2-\varepsilon) - (m_1+\varepsilon)| < \delta_2 + \delta_1$. By Lemma 3.2.6, the equation $f(x-\varepsilon) = g(x+\varepsilon)$ has either one or two solutions. Therefore, r must be the rightmost solution to $f(x-\varepsilon) = g(x+\varepsilon)$.

We will split the analysis into three sub-cases.

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

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 $[m_1 - \delta_1, r - \varepsilon]$ and $[r - \varepsilon, m_1 + \delta_1]$ respectively. Let ν_- and ν_+ be the restriction of ν to the intervals $[m_2 - \delta_2, r + \varepsilon]$ and $[r + \varepsilon, m_2 + \delta_2]$ respectively. The following inequality shows that the support of ν_- is of a lesser length than that of μ_-.

$$[(r+\varepsilon) - (m_2 - \delta_2)] - [(r-\varepsilon) - (m_1 - \delta_1)] = \delta_2 - \delta_1 + 2\varepsilon - (m_2 - m_1) < 0.$$

It follows that the support of ν_+ is of a greater length than that of μ_+. By construction, $g(x - 2\varepsilon) \leq f(x)$ for $x \in [m_2 - \delta_2, r + \varepsilon]$. Hence, by Lemma 3.2.3, $D_\varepsilon(\mu_-, \nu_-) = 0$. A similar analysis shows that $D_\varepsilon(\mu_+, \nu_+) = 0$. Hence,

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

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

Case 3(b): $\delta_2 - \delta_1 < m_2 - m_1 \leq \delta_2 - \delta_1 + 2\varepsilon$**.**

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_+.

$$[(m_2 + \delta_2) - (r + \varepsilon)] - [(m_1 + \delta_1) - (r - \varepsilon)] = (m_2 - m_1) + \delta_2 - \delta_1 - 2\varepsilon > 0.$$

Moreover, $f(x) \leq g(x + 2\varepsilon)$ for $x \in [r - \varepsilon, m_1 + \delta_1]$. Hence by Lemma 3.2.3, $D_\varepsilon(\mu_+, \nu_+) = 0$.

We will now show that $D_\varepsilon(\mu_-, \nu_-) = 0$ by verifying the conditions of Lemma 3.2.4. Since $2\varepsilon < 2\delta_1$, we have the following.

$$\delta_2 - \delta_1 < m_2 - m_1 \leq \delta_2 - \delta_1 + 2\varepsilon < \delta_2 - \delta_1 + 2\delta_1 = \delta_2 + \delta_1.$$

Hence, by Lemma 3.2.6, 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 $a = m_2 - \delta_2 - 2\varepsilon$, $a' = a + 2\varepsilon$ and $b = r + \varepsilon$. Then, (1) the support of μ_- is $[m_1 - \delta_1, r - \varepsilon]$ which is a subset of $[a, b]$, and the support of ν_- is $[a', b]$. (2) $f(x) \geq g(x)$ for $x \in (a, t]$ and $f(x) \leq g(x)$ for $x \in (t, b]$. Hence, the first two conditions of Lemma 3.2.4 are verified. To verify, the third condition, we note the following.

$$(m_2 - 2\varepsilon) - m_1 = m_2 - m_1 - 2\varepsilon < \delta_2 - \delta_1,$$
$$m_1 - (m_2 - 2\varepsilon) = m_1 - m_2 + 2\varepsilon < 2\varepsilon < \delta_2 - \delta_1.$$

Hence, by Lemma 3.2.6, $f(x) - g(x + 2\varepsilon) = 0$ exactly twice on the support of μ. The greater of the two will be $r - \varepsilon$. Let $\tilde{t}$ be the lesser of the two. Then, $\tilde{t} < r - \varepsilon = b - 2\varepsilon$. Further, $f(x) \leq g(x + 2\varepsilon)$ for $x \in [a, \tilde{t})$ and $f(x) \geq g(x + 2\varepsilon)$ for $x \in (\tilde{t}, b - 2\varepsilon]$. Hence, $D_\varepsilon(\mu_-, \nu_-) = 0$ by Lemma 3.2.4. Therefore, the optimal risk and optimal classifier remain the same as in case 3(a).

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

We will partition the real line into four regions as shown in Table C.2, 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 $D_\varepsilon(\mu_+, \nu_+) = D_\varepsilon(\mu_{++}, \nu_{++}) = 0$.

We will now show $D_\varepsilon(\mu_{--}, \nu_{--}) = 0$ using Lemma 3.2.4. Let $a = m_1 - \delta_1 - 2\varepsilon$, $a' = a + 2\varepsilon$, $b = l'$ and $\tilde{t} = b - 2\varepsilon$. Since $m_2 - m_1 \leq \delta_2 - \delta_1$, by Lemma 3.2.6, $f(x) - g(x) = 0$ 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 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) $g(x) \leq f(x + 2\varepsilon)$ for $x \in [a, \tilde{t})$ and the the interval $(\tilde{t}, b - 2\varepsilon]$ is empty because $\tilde{t} = b - 2\varepsilon$. Hence, $D_\varepsilon(\mu_-, \nu_-) = 0$.

Finally, $D_\varepsilon(\mu_-, \nu_-) = 0$ 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.

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

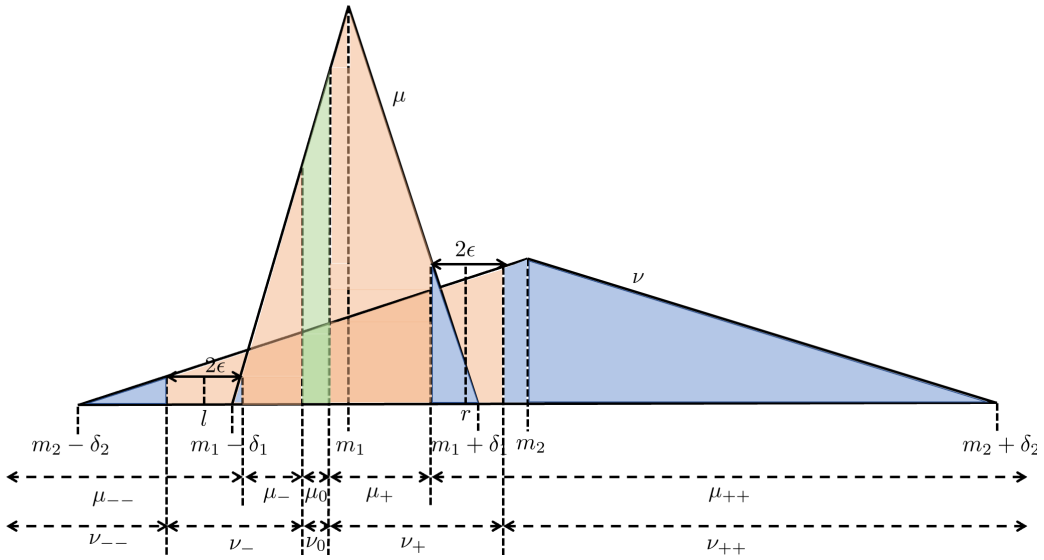


Figure C.2: Optimal transport coupling for triangular distributions μ and ν. As in the proof of Theorem 15, we divide the real line into five regions. The transport plan from μ to ν consists of five maps transporting $\mu_{--} \to \nu_{--}$ (blue regions to the left), $\mu_- \to \nu_-$ (orange regions 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).

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

μ_{--}	$(m_1 - \delta_1, l + \varepsilon]$
μ_-	$(l + \varepsilon, l']$
μ_0	(l', r')
μ_+	$[r', r - \varepsilon)$
μ_{++}	$[r - \varepsilon, m_1 + \delta_1)$

ν_{--}	$(m_2 - \delta_2, l - \varepsilon]$
ν_-	$(l - \varepsilon, l']$
ν_0	(l', r')
ν_+	$[r', r + \varepsilon)$
ν_{++}	$[r + \varepsilon, m_2 + \delta_2)$

Table C.1: The real line is partitioned into five regions for μ and ν for Case 2.

μ_{--}	$(m_1 - \delta_1, l']$
μ_-	(l', r')
μ_+	$[r', r - \varepsilon)$
μ_{++}	$[r - \varepsilon, m_1 + \delta_1)$

ν_{--}	$(m_2 - \delta_2, l']$
ν_-	(l', r')
ν_+	$[r', r + \varepsilon)$
ν_{++}	$[r + \varepsilon, m_2 + \delta_2)$

Table C.2: The real line is partitioned into four regions for μ and ν for Case 3(c).

□

D

Proofs from Chapter 4

D.1 Proofs from Section 4.1

Proof of Theorem 22. Consider any μ' and ν' such that $W_\infty(\mu, \mu') \leq \varepsilon$ and $W_\infty(\nu, \nu') \leq \varepsilon$. Then there exist $\gamma_\mu \in \Pi(\mu, \mu')$ and $\gamma_\nu \in \Pi(\nu, \nu')$ such that

$$\mathbb{P}_{(x,x')\sim\gamma_\mu}(d(x, x') > \varepsilon) = 0,$$
$$\mathbb{P}_{(x,x')\sim\gamma_\nu}(d(x, x') > \varepsilon) = 0.$$

Let $\gamma' \in \Pi(\mu', \nu')$ be the coupling that achieves the optimal transport cost $D_{TV}(\mu', \nu')$. Construct a coupling $\gamma_0 \in \Pi(\mu, \nu)$ as $\gamma_0 = \gamma_\mu \circ \gamma' \circ \gamma_\nu$. Then,

$$\begin{aligned} D_\varepsilon(\mu, \nu) &\leq \int_{\mathcal{X}^2} 1\{d(x, x') > 2\varepsilon\} d\gamma_0 \\ &\leq \int_{\mathcal{X}^2} 1\{d(x, x') > 0\} d\gamma' \\ &= D_{TV}(\mu', \nu'). \end{aligned}$$

Since the above inequalty is true for any μ' and ν' such that $W_\infty(\mu, \mu') \leq \varepsilon$ and $W_\infty(\nu, \nu') \leq \varepsilon$, we have the following inequality.

$$D_\varepsilon(\mu, \nu) \leq \inf_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \varepsilon} D_{TV}(\mu', \nu').$$

Now we will show the above inequality in the reverse direction. Let $\gamma \in \Pi(\mu, \nu)$ be the coupling that achieves the optimal transport cost for $D_\varepsilon(\mu, \nu)$. Let $M : \mathcal{X}^2 \to \mathcal{X}$ be a measurable midpoint map. (See Dohmatob[24] for why such a map exists.) That is, for all $(x, x') \in \mathcal{X}^2$ we have

$$d(x, M(x, x')) = d(x', M(x, x')) = \frac{1}{2} d(x, x').$$

Consider a transport map $T : \mathcal{X}^2 \to \mathcal{X}^2$ defined as

$$T(x, x') = \begin{cases} (M(x, x'), M(x, x')) & d(x, x') \leq 2\varepsilon, \\ (x, x') & \text{otherwise.} \end{cases}$$

T is measurable because it is piece-wise measurable on measurable sets. Further, it follows from the definition of M that each coordinate of a point (x, x') is transported by T by a distance no further than ε. Let μ_0 ad ν_0 be the probability measures corresponding to the first and second marginals of $T_{\sharp\gamma}$ respectively. Then, $W_\infty(\mu, \mu_0) \leq \varepsilon$ and $W_\infty(\nu, \nu_0) \leq \varepsilon$. Hence,

$$\begin{aligned} D_\varepsilon(\mu, \nu) &= \int_{\mathcal{X}^2} 1\{d(x, x') > 2\varepsilon\} d\gamma \\ &= \int_{\mathcal{X}^2} 1\{d(x, x') > 0\} d\gamma_{\sharp T} \\ &\geq D_{TV}(\mu_0, \nu_0) \\ &\geq \inf_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \varepsilon} D_{TV}(\mu', \nu'). \end{aligned}$$

Combining with the reverse inequality that we proved above, it is clear that the infimum over D_{TV} is attained by μ_0 and ν_0. □

Proof of Theorem 23. By Lemma 2.4.3, the set-valued maps $A \mapsto p_0(A^{\oplus\varepsilon})$ and $A^c \mapsto p_1((A^c)^{\oplus\varepsilon})$ are 2-alternating

capacities. Hence, by Theorem 4.1 in Huber and Strassen [46], there exist $p_0^*, p_1^* \in \mathcal{P}(\mathcal{X})$ such that $W_\infty(p_i, q_i) \leq \varepsilon$ for $i = 0, 1$ and,

$$\inf_{A\in\mathcal{L}(\mathcal{X})} \sup_{W_\infty(p_0,p_0'),W_\infty(p_1,p_1')\leq\varepsilon} \frac{T}{T+1}p_0'(A) + \frac{1}{T+1}p_1'((A^c)) = \inf_{A\in\mathcal{L}(\mathcal{X})} \frac{T}{T+1}p_0^*(A) + \frac{1}{T+1}p_1^*((A^c)).$$

□

D.2 Proofs from Section 4.2

Proof of Lemma 4.2.1. The first equality in (4.10) follows from Theorem 22. For the second equality, we have the following.

$$\begin{aligned}
\inf_{\substack{q\in\overline{\mathcal{P}}(\mathcal{X}):\\ q\preceq Tp_0}} D_\varepsilon(q, p_1) &\overset{(i)}{=} 1 - (T+1) \inf_{A\in\mathcal{B}(\mathcal{X})} R_{\oplus\varepsilon}(\ell_{0/1}, A) \\
&\overset{(ii)}{=} 1 - (T+1) \inf_{A\in\mathcal{B}(\mathcal{X})} R_{\Gamma_\varepsilon}(\ell_{0/1}, A) \\
&= 1 - (T+1) \inf_{A\in\mathcal{B}(\mathcal{X})} \sup_{\substack{W_\infty(p_0,p_0')\leq\varepsilon\\ W_\infty(p_1,p_1')\leq\varepsilon}} r(A, p_0', p_1') \\
&\overset{(iii)}{\leq} 1 - (T+1) \sup_{\substack{W_\infty(p_0,p_0')\leq\varepsilon\\ W_\infty(p_1,p_1')\leq\varepsilon}} \inf_{A\in\mathcal{B}(\mathcal{X})} r(A, p_0', p_1') \\
&= \inf_{\substack{W_\infty(p_0,p_0')\leq\varepsilon\\ W_\infty(p_1,p_1')\leq\varepsilon}} [1 - (T+1) \inf_{A\in\mathcal{B}(\mathcal{X})} r(A, p_0', p_1')] \\
&\overset{(iv)}{\leq} \inf_{\substack{W_\infty(p_0,p_0')\leq\varepsilon\\ W_\infty(p_1,p_1')\leq\varepsilon}} \inf_{\substack{q'\in\overline{\mathcal{P}}(\mathcal{X}):\\ q'\preceq Tp_0'}} D_{TV}(q', p_1'),
\end{aligned}$$

where (i) follows from Theorem 10, (ii) from Theorem 4, (iii) from Lemma A.0.10, and (iv) again from Theorem 10 with $\varepsilon = 0$.

We will now show the inequality in the opposite direction. That is, we will show the following.

$$\inf_{\substack{q\in\overline{\mathcal{P}}(\mathcal{X}):\\ q\preceq Tp_0}} \inf_{\substack{W_\infty(q,q')\leq\varepsilon\\ W_\infty(p_1,p_1')\leq\varepsilon}} D_{TV}(q', p_1') \geq \inf_{\substack{W_\infty(p_0,p_0')\leq\varepsilon\\ W_\infty(p_1,p_1')\leq\varepsilon}} \inf_{\substack{q'\in\overline{\mathcal{P}}(\mathcal{X}):\\ q'\preceq Tp_0'}} D_{TV}(q', p_1') \tag{D.1}$$

Consider arbitrary probability measures $q', p_1' \in \overline{\mathcal{P}}(\mathcal{X})$ generated in accordance with the constraints over the

infimum terms on the left hand side of the above inequality. That is, let q' and p_1' be such that $W_\infty(q, q') \leq \varepsilon$ and $W_\infty(p_1, p_1') \leq \varepsilon$ where $q \preceq Tp_0$. We will now construct $p_0' \in \overline{\mathcal{P}}(\mathcal{X})$ such that $q' \preceq Tp_0'$ and $W_\infty(p_0, p_0') \leq \varepsilon$. This will show that the set of $q', p_1' \in \overline{\mathcal{P}}(\mathcal{X})$ satisfying the constraints over the infimum terms on the right hand side is a superset of the corresponding set on the right hand side, and hence prove the above inequality.

Define a probability measure $p_0' \in \overline{\mathcal{P}}(\mathcal{X})$ as $p_0'(A) = p_0(A) + \frac{1}{T}q'(A) - \frac{1}{T}q(A)$ for $A \in \mathcal{B}(\mathcal{X})$. To show that p_0' is a valid probability measure, we have the following.

$$
\begin{aligned}
p_0'(\mathcal{X}) &= p_0(\mathcal{X}) + \frac{1}{T}q'(\mathcal{X}) - \frac{1}{T}q(\mathcal{X}) = 1 \\
p_0'(A) &= \frac{1}{T}(Tp_0(A) - q(A)) + \frac{1}{T}q'(A) \geq \frac{1}{T}q'(A) \geq 0.
\end{aligned}
$$

The above equality also shows that $q' \preceq Tp_0'$. We will now show that $W_\infty(p_0, p_0') \leq \varepsilon$. Since $W_\infty(q, q') \leq \varepsilon$, there exists $\gamma \in \Pi(q, q')$ such that $\gamma(\{(x, x') \in \mathcal{X}^2 : d(x, x') \leq 2\varepsilon\}) = 1$. Define $\gamma' \in \Pi(p_0, p_0')$ as follows for $A \in \mathcal{B}(\mathcal{X}^2)$.

$$
\gamma'(A) = p_0(\{x \in \mathcal{X} : (x, x) \in A\}) + \frac{1}{T}\gamma(A) - \frac{1}{T}q(\{x \in \mathcal{X} : (x, x) \in A\}).
$$

To see that $\gamma' \in \Pi(p_0, p_0')$, we have the following for $A_1, A_2 \in \mathcal{B}(\mathcal{X})$.

$$
\begin{aligned}
\gamma'(A_1 \times \mathcal{X}) &= p_0(A_1) + \frac{1}{T}q(A_1) - \frac{1}{T}q(A_1) = p_0(A_1), \\
\gamma'(\mathcal{X} \times A_2) &= p_0(A_2) + \frac{1}{T}q'(A_2) - \frac{1}{T}q(A_2) = p_0'(A_2).
\end{aligned}
$$

Moreover,

$$
\gamma'(\{(x, x') \in \mathcal{X}^2 : d(x, x') \leq 2\varepsilon\}) = p_0(\mathcal{X}) + \frac{1}{T}\gamma(\{(x, x') \in \mathcal{X}^2 : d(x, x') \leq 2\varepsilon\}) - \frac{1}{T}q(\mathcal{X}) = 1.
$$

Therefore, $W_\infty(p_0, p_0') \leq \varepsilon$. □

References

[1] Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*.

[2] Bao, H., Scott, C., & Sugiyama, M. (2020). Calibrated surrogate losses for adversarially robust classification. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 (pp. 408–451).

[3] Baraud, Y. & Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B), 3767–3804.

[4] Baraud, Y., Birgé, L., & Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 207(2), 425–517.

[5] Bertsekas, D. P. & Shreve, S. E. (1996). *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific.

[6] Bertsimas, D., Shtern, S., & Sturt, B. (2020). A data-driven approach to multi-stage stochastic linear optimization. *Available at Optimization-Online*.

[7] Bhagoji, A. N., Cullina, D., & Mittal, P. (2019). Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 7498–7510).

[8] Bhattacharjee, R. & Chaudhuri, K. (2020). When are non-parametric methods robust? In *International Conference on Machine Learning (ICML)*.

[9] Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley Series in Probability and Statistics.

[10] Billingsley, P. (2013). *Convergence of Probability Measures*. John Wiley & Sons.

[11] Blanchet, J. & Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2), 565–600.

[12] Bogachev, V. I. (2007). *Measure Theory*, volume 2. Springer Science & Business Media.

[13] Bose, A., Gidel, G., Berrard, H., Cianflone, A., Vincent, P., Lacoste-Julien, S., & Hamilton, W. (2020). Adversarial example games. *Advances in Neural Information Processing Systems*.

[14] Bulò, S., Biggio, B., Pillai, I., Pelillo, M., & Roli, F. (2016). Randomized prediction games for adversarial machine learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11), 2466–2478.

[15] Bungert, L., Trillos, N. G., & Murray, R. (2021). The geometry of adversarial training in binary classification. *arXiv preprint arXiv:2111.13613*.

[16] Carlini, N. & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy* (pp. 39–57).: IEEE.

[17] Chalkidis, I. & Kampas, D. (2019). Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2), 171–198.

[18] Chen, R. & Paschalidis, I. C. (2018). A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1), 517–564.

[19] Choquet, G. (1954). Theory of capacities. In *Annales de l'institut Fourier*, volume 5 (pp. 131–295).

[20] Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*.

[21] Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*.

[22] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186).: Association for Computational Linguistics.

[23] Diochnos, D. I., Mahloujifar, S., & Mahmoody, M. (2018). Adversarial risk and robustness: General definitions and implications for the uniform distribution. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 10359–10368.

[24] Dohmatob, E. (2020). Universal lower-bounds on classification error under adversarial attacks and random corruption. *arXiv preprint arXiv:2006.09989*.

[25] Dudley, R. (2010). Distances of probability measures and random variables. In *Selected Works of RM Dudley* (pp. 28–37). Springer.

[26] Erdős, P. & Stone, A. (1970). On the sum of two Borel sets. *Proceedings of the American Mathematical Society*, 25(2), 304–306.

[27] Esfahani, P. M. & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2), 115–166.

[28] Gao, R. & Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *INFORMS Annual Meeting*.

[29] Gardner, R. (2002). The Brunn-Minkowski inequality. *Bulletin of the American Mathematical Society*, 39(3), 355–405.

[30] Givens, C. R. & Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2), 231–240.

[31] Goh, J. & Sim, M. (2010). Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1), 902–917.

[32] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[33] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*.

[34] Gourdeau, P., Kanade, V., Kwiatkowska, M., & Worrell, J. (2019). On the hardness of robust classification. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 7446–7455).

[35] Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649).: IEEE.

[36] Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5), 1153–1159.

[37] Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3), 362–386.

[38] Gül, G. (2017). *Robust and Distributed Hypothesis Testing*. Springer.

[39] Gül, G. & Zoubir, A. M. (2013). Robust hypothesis testing for modeling errors. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5514–5518).: IEEE.

[40] Gül, G. & Zoubir, A. M. (2017). Minimax robust hypothesis testing. *IEEE Transactions on Information Theory*, 63(9), 5572–5587.

[41] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

[42] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 82–97.

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

[44] Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, (pp. 1753–1758).

[45] Huber, P. J. (2004). *Robust Statistics*, volume 523. John Wiley & Sons.

[46] Huber, P. J. & Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, (pp. 251–263).

[47] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 125–136).

[48] Khim, J. & Loh, P.-L. (2018). Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*.

[49] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 84–90.

[50] Kumar, R., O'Brien, D., Albert, K., & Vilojen, S. (2018). Law and adversarial machine learning. *NeurIPS Workshop on Security in Machine Learning*.

[51] Lanckriet, G. R., Ghaoui, L. E., Bhattacharyya, C., & Jordan, M. I. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, 3(Dec), 555–582.

[52] LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1), 38–53.

[53] Lee, J. & Raginsky, M. (2018). Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2687–2696.

[54] Levy, B. C. (2008). Robust hypothesis testing with a relative entropy tolerance. *IEEE Transactions on Information Theory*, 55(1), 413–421.

[55] Liu, Z., Zhang, J., Jog, V., Loh, P., & McMillan, A. (2019). Robustifying deep networks for image segmentation. *arXiv preprint arXiv:1908.00656*.

[56] Luiro, H., Parviainen, M., & Saksman, E. (2014). On the existence and uniqueness of p-harmonious functions. *Differential and Integral Equations*, 27(3/4), 201 – 216.

[57] Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110, 107332.

[58] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*.

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

[60] Matousek, J. & Gärtner, B. (2007). *Understanding and Using Linear Programming*. Springer Science & Business Media.

[61] Mazuelas, S., Zanoni, A., & Pérez, A. (2020). Minimax classification with 0-1 loss and performance guarantees. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

[62] Meunier, L., Scetbon, M., Pinot, R., Atif, J., & Chevaleyre, Y. (2021). Mixed Nash equilibria in the adversarial examples game. *International Conference on Machine Learning*.

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

[64] Muhammad, K., Ullah, A., Lloret, J., Del Ser, J., & de Albuquerque, V. (2020). Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*.

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

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

[67] Pinot, R., Ettedgui, R., Rizk, G., Chevaleyre, Y., & Atif, J. (2020). Randomization matters. how to defend against strong adversarial attacks. *International Conference on Machine Learning (ICML)*.

[68] Pydi, M. S. & Jog, V. (2021). Adversarial risk via optimal transport and optimal couplings. *IEEE Transactions on Information Theory*, 67(9), 6031–6052.

[69] Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63), 94.

[70] Schay, G. (1974). Nearest random variables with given distributions. *The Annals of Probability*, (pp. 163–166).

[71] Shafahi, A., Huang, W. R., Studer, S., Feizi, S., & Goldstein, T. (2019). Are adversarial examples inevitable? *International Conference on Learning Representations (ICLR)*.

[72] Shafieezadeh Abadeh, S., Esfahani, M., M., P., & Kuhn, D. (2015). Distributionally robust logistic regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 1576–1584.

[73] Shaham, U., Yamada, Y., & Negahban, S. (2018). Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307, 195–204.

[74] Shivaswamy, P. K., Bhattacharyya, C., & Smola, A. J. (2006). Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7(Jul), 1283–1314.

[75] Sierpiński, W. (1920). Sur la question de la mesurabilité de la base de M. Hamel. *Fundamenta Mathematicae*, 1(1), 105–111.

[76] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.

[77] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419), 1140–1144.

[78] Sinha, A., Namkoong, H., & Duchi, J. C. (2017). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*.

[79] Staib, M. & Jegelka, S. (2017). Distributionally robust deep learning as a generalization of adversarial training. In *NeurIPS workshop on Machine Learning and Computer Security*.

[80] Strassen, V. (1965). The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2), 423–439.

[81] Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., & Gao, Y. (2018). Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 631–648).

[82] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*.

[83] Tovar-Lopez, A. & Jog, V. (2018). Generalization error bounds using Wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)* (pp. 1–5).: IEEE.

[84] Trillos, N. & Murray, R. (2020). Adversarial classification: Necessary conditions and geometric flows. *arXiv preprint arXiv:2011.10797*.

[85] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. *International Conference on Learning Representations (ICLR)*.

[86] Tu, Z., Zhang, J., & Tao, D. (2019). Theoretical analysis of adversarial learning: A minimax approach. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 12280–12290).

[87] Uesato, J., Oadonoghue, B., Kohli, P., & Oord, A. (2018). Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning* (pp. 5025–5034).: PMLR.

[88] Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

[89] Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., & Silver, D. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782), 350–354.

[90] Wagner, D. H. (1977). Survey of measurable selection theorems. *SIAM Journal on Control and Optimization*, 15(5), 859–903.

[91] Wiesemann, W., Kuhn, D., & Sim, M. (2014). Distributionally robust convex optimization. *Operations Research*, 62(6), 1358–1376.

[92] Xu, H., Caramanis, C., & Mannor, S. (2009). Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(7).

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

[94] Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., & Chaudhuri, K. (2020b). A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

[95] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13, 55–75.

[96] Yu, L. (2019). Asymptotics for strassen's optimal transport problem. *arXiv preprint arXiv:1912.02051*.

[97] Yu, L. & Tan, V. (2018). Asymptotic coupling and its applications in information theory. *IEEE Transactions on Information Theory*, 65(3), 1321–1344.

[98] Yue, M., Kuhn, D., & Wiesemann, W. (2021). On linear optimization over Wasserstein balls. *Mathematical Programming*, (pp. 1–16).

[99] Zajíček, L. (2005). On σ-porous sets in abstract spaces. In *Abstract and Applied Analysis*, volume 2005 (pp. 509–534).: Hindawi.