The Many Faces of Adversarial Risk: An Expanded Study

Muni Sreenivas Pydi and Varun Jog

Abstract—Adversarial risk quantifies the performance of classifiers on adversarially perturbed data. Numerous definitions of adversarial risk—not all mathematically rigorous and differing subtly in the details—have appeared in the literature. In this paper, we revisit these definitions, fix measure theoretic issues, and critically examine their similarities and differences. Our technical tools derive from optimal transport, robust statistics, functional analysis, and game theory. Our contributions include the following: generalizing Strassen’s theorem to the unbalanced optimal transport setting with applications to adversarial classification with unequal priors; showing an equivalence between adversarial robustness and robust hypothesis testing with \( \infty \)-Wasserstein uncertainty sets; proving the existence of a pure Nash equilibrium in the two-player game between the adversary and the algorithm; and characterizing adversarial risk by the minimum Bayes error between a pair of distributions belonging to the \( \infty \)-Wasserstein uncertainty sets. Our results generalize and deepen recently discovered connections between optimal transport and adversarial robustness and reveal new connections to Choquet capacities and game theory.

Index Terms—Adversarial robustness, optimal transport, robust hypothesis testing, Strassen’s theorem, \( \infty \)-Wasserstein distances.

I. INTRODUCTION

NEURAL networks are known to be vulnerable to adversarial attacks, which are imperceptible perturbations to input data that maximize loss [1], [2], [3]. Developing algorithms resistant to such attacks has received considerable attention in recent years [4], [5], [6], [7], motivated by safety-critical applications such as autonomous driving [8], [9], medical imaging [10], [11], [12] and law [13], [14].

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 robust. Procedures for finding them have been effective in practice [5], [6], [15], spurring numerous theoretical investigations into adversarial risk and its minimizers.

There is no universally agreed upon definition of adversarial risk. Even the simplest setting of binary classification in \( \mathbb{R}^d \) with an \( \ell_2 \) adversary admits various definitions involving set expansions [16], [17], transport maps [18], Markov kernels [19], and couplings [20]. 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.

Optimal adversarial risk is most commonly defined as the minimax risk under adversarial contamination [6], [21]. Other notable characterizations include an optimal transport cost between data generating distributions in [22], [23], [24], and [25], the optimal value of a distributionally robust optimization problem [26], [27], [28], and the value of a two-player zero-sum game [18], [20], [29], [30].

The diversity of definitions for adversarial risk makes it challenging to compare approaches. Moreover, not all approaches 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.

A mathematically rigorous foundation for adversarial risk is essential for future research.

In this paper, we examine various notions of adversarial risk in two settings: (1) binary classification in a non-parametric setting under 0-1 loss function, where the decision boundary (or decision region) of a classifier is an arbitrary subset of the input space, and (2) multi-class classification in a parametric setting under a general loss function, where a classifier is parametrized by a \( w \) in a hypothesis set \( \mathcal{W} \). We present rigorous definitions of adversarial risk and identify conditions under which these definitions are equivalent. We consider the general setting of Polish spaces (complete, separable metric spaces), and present stronger results for the Euclidean space \( \mathbb{R}^d \). Our contributions are as follows:

- **Well-definedness of adversarial risk:** We examine the definition of adversarial risk based on set expansions. For Polish spaces, we observe that adversarial risk is not Borel measurable, and hence, not well-defined when the input space is not complete or separable. We present a modified definition of adversarial risk that is well-defined and satisfies additional desirable properties.

Manuscript received 4 October 2022; revised 4 May 2023; accepted 28 July 2023. Date of publication 7 August 2023; date of current version 26 December 2023. This work was supported in part by the NSF under Grant CCF-1841190 and Grant CCF-1907786, and in part by the University of Cambridge. An earlier version of this paper was presented in part at the Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS) [38]. (Corresponding author: Muni Sreenivas Pydi.)

Muni Sreenivas Pydi is with the LAMSADe Research Center, Université Paris Dauphine-PSL, 75775 Paris, France (e-mail: muni.pydi@lamsade.dauphine.fr).

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

Communicated by E. Gassiat, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3303221.

Digital Object Identifier 10.1109/TIT.2023.3303221
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 5) shows that the Lebesgue σ-algebra is preferred over the Borel σ-algebra because set expansions are Lebesgue measurable but not necessarily Borel measurable. These results are contained in Section IV.

- **Equivalence between various notions of adversarial risk:** 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 [31], [32]. In $\mathbb{R}^d$, we are able to use the powerful theory of Choquet capacities [33] (in particular, Huber and Strassen's 2-alternating capacities [34]) to establish results of a similar nature. In addition, we derive the conditions under which this notion of adversarial risk is equivalent to alternative notions defined using transport maps and Markov kernels. These results are contained in Section V.

- **Optimal transport characterization of optimal adversarial risk:** We consider the binary classification setup with unequal priors and show (under suitable assumptions) that the optimal adversarial risk from the above definitions is characterized by an unbalanced optimal transport cost between data-generating distributions. For both Polish spaces and $\mathbb{R}^d$, the main tool we use is Theorem 20 in which we generalize a classical result of Strassen on excess-cost optimal transport [35],[36] from probability measures to finite measures with possibly unequal mass. This generalizes results of [19] and [23] on binary classification, which were only for equal priors. These results are contained in Section VI.

- **Game-theoretic view on adversarial risk and existence of Nash equilibria:** We consider the setup of a zero-sum game between the adversary and the algorithm. 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 the results of [18], [20], and [29] to non-parametric classifiers. These results are contained in Section VII.

The paper is organized as follows: In Section II, we present preliminary definitions from optimal transport and metric space topology. In Section III, we discuss various definitions of adversarial risk and present more related work. Sections IV, V, VI and VII contain our main contributions summarized above. We conclude the paper in Section VIII and discuss future research directions.

We emphasize that rectifying measure theoretic issues with existing formulations of adversarial risk is one of our contributions, but not the main focus of our paper. We start our presentation by addressing measurability and well-definedness (in Section IV) because otherwise we will not be able to rigorously present our main results in the subsequent sections, namely: relation to robust hypothesis testing and Choquet capacities in Section V, generalizing the results of [23] and [22] in Section VI proving minimax theorems and existence of Nash equilibria and extending the results of [18], [20], and [29] in Section VII.

Two distinct formulations for adversarial risk appear in the literature: corrupted-instance risk and error-region risk [16], [17], [37]. For corrupted-instance risk, the optimal classifier in the adversarial setting can be different from that of the standard setting, giving rise to a robustness-accuracy trade-off. For error-region risk, the Bayes classifier remains optimal even in the adversarial setting. We discuss the differences between both types of risk and the equivalences between the two in Subsection III-B.1. We investigate the well-definedness of both types of risks and their relation to ∞-Wasserstein distributional robustness in Section IV and Section V, respectively. Sections VI and VII provide results that are specific to the corrupted-instance risk, as the optimal adversarial risk and classifier remain unchanged from the standard setting to the adversarial setting for the error-region risk.

A shorter version of this paper was presented at NeurIPS 2021 [38], and focused exclusively on binary classification in a non-parametric setting under the 0-1 loss function. In this extended version, we present new results for multi-class classification in a parametric setting under a general loss function. Further, we include new results pertaining to “error-region” family of adversarial risks.

**Notation:** Throughout the paper, 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 $r \geq 0$, 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 $\nu$ dominates $\mu$ 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 $\lambda$ to denote the $d$-dimensional Lebesgue measure. For a positive integer $n$, we use $[n]$ to denote the finite set $\{1, \ldots, n\}$. 

**II. Preliminaries**

**A. Metric Space Topology**

We introduce three different notions of set expansions. For $\epsilon \geq 0$ and $A \in \mathcal{B}(\mathcal{X})$, the $\epsilon$-Minkowski expansion of $A$ is given by $A^{\epsilon} := \bigcup_{a \in A} B_r(a)$. The $\epsilon$-closed expansion of $A$ is defined as $A^\epsilon := \{ x \in \mathcal{X} : d(x,A) \leq \epsilon \}$, where $d(x,A) = \inf_{a \in A} d(x,a)$. The $\epsilon$-open expansion of $A$ is defined as $A^{\epsilon} := \{ x \in \mathcal{X} : d(x,A) < \epsilon \}$. We use the notation $A^{-\epsilon}$ to denote $((A^\epsilon)^c)^c$. Similarly, $A^{\epsilon\epsilon} := ((A^\epsilon)^{\epsilon\epsilon})^c$. For
example, consider the set \( A = (0, 1] \) in the space \( (X, d) = (\mathbb{R}, |\cdot|) \) and \( \epsilon > 0 \). Then \( A_{\text{dir}} = (-\epsilon, 1 + \epsilon) \) and \( A_\epsilon = [-\epsilon, 1 + \epsilon] \) and \( A^1 = (-\epsilon, 1 + \epsilon) \). For any \( A \in B(X) \), \( A^1 \) is closed and \( A^0 \) is open. Hence, \( A^1, A^0 \subseteq B(X) \). Moreover, \( A^0 \) may not be in \( B(X) \) (see Lemma 2).

In general, the Minkowski sum of two Borel sets need not be Borel [39], and that of two Lebesgue measurable sets need not be Lebesgue measurable [40].

### B. Optimal Transport

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

\[
\mathcal{T}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int c(x, y) \, d\pi(x, y).
\]

For a positive integer \( p \), the \( p \)-Wasserstein distance between \( \mu \) and \( \nu \) is defined as \( W_p(\mu, \nu) \). 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 [41].

### III. DEFINITIONS AND RELATED WORK

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, which non-parametric classifiers of the form \( f_A(x) = 1\{x \in A\} \) correspond to decision regions \( A \subseteq X \).

#### A. General Loss Setting

Let \( X \) be the feature space, a Polish space equipped with a distance metric \( d \). Let \( \mathcal{Y} \) be a finite set of labels. Let \( p \) be the true data distribution of labeled data points \((x, y) \in X \times \mathcal{Y} \), 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 \( x \in X \) given the label \( y \). Let \( \mathcal{W} \) denote the hypothesis class. Let \( \ell : (X \times \mathcal{Y}) \times \mathcal{W} \to [0, \infty] \) denote a loss function that is measurable with respect to \( B(X) \) for all \( w \in \mathcal{W} \).

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

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

If the loss function \( \ell((\cdot), w) \) is upper semi-continuous and bounded above for all \( w \in \mathcal{W} \), Meunier et al. [20] show that \( R_{\ell, w} \) 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 \( F := \{ f_y : X \to X, f_y \text{ is } \rho_y \text{-measurable}\} \) denote a collection of measurable maps for each label \( y \in \mathcal{Y} \). We say that \( F \) is of budget \( \epsilon \) (denoted by \( F \in F_\epsilon \)) if \( d(x, f_y(x)) \leq \epsilon \) with probability 1 for \((x, y) \sim \rho \). Under such an adversary, the adversarial risk may be defined as follows.

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

The above definition was used for the binary classification setting in [18]. A more general definition for adversarial risk was proposed in [19] using Markov kernels. Let \( \kappa \) denote a set of Markov kernels \( \kappa_y \) for \( y \in \mathcal{Y} \). Let \( \rho^e(x, y, x', y) \) denote the joint distribution of \((x, y, x') \) induced by \( \kappa \). We say that the Markov kernel adversary \( \kappa \) has a budget \( \epsilon \) (denoted by \( \kappa \in \mathcal{K}_\epsilon \)) if \( d(x, x') \leq \epsilon \), \( \rho^e(x, y, x', y) \)-a.s. where \( \rho^e(x, y, x', y) \) 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 \( \kappa_y \). Under such a Markov kernel adversary, adversarial risk is defined as the following in [19].

\[
R_{\mathcal{K}_\epsilon}(\ell, w) = \sup_{\kappa \in \mathcal{K}_\epsilon} \mathbb{E}_{(x, y, x') \sim \rho^e(x, y, x') \ell((x', y), w)}. \tag{3}
\]

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 \( \infty \)-Wasserstein metric (denoted by \( W_\infty \)) has been used to define such perturbations [19], [20]. Let an adversary \( \gamma \) be defined as a collection of perturbed probability distributions for each label i.e., \( \gamma := \{ \rho^\gamma_{x|y} \in \mathcal{P}(X) | y \in \mathcal{Y} \} \). We say that the adversary \( \gamma \) has a budget \( \epsilon \) (denoted by \( \gamma \in \mathcal{G}_\epsilon \)) if \( W_\infty(\rho^\gamma_{x|y}, \rho^\gamma_{x'|y}) \leq \epsilon \) for all \( y \in \mathcal{Y} \). Under such a distribution-perturbing adversary, the adversarial risk is defined as,

\[
R_{\mathcal{G}_\epsilon}(\ell, w) = \sup_{\gamma \in \mathcal{G}_\epsilon} \mathbb{E}_{(x', y) \sim \rho^{\gamma}_{x'|y}} \ell((x', y), w). \tag{4}
\]

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

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

\[
R_{F_\epsilon}(\ell, w) \leq R_{\mathcal{K}_\epsilon}(\ell, w) \leq R_{\mathcal{G}_\epsilon}(\ell, w). \tag{5}
\]

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

### B. Binary Classification With 0-1 Loss Setting

In this subsection, we consider a binary classification setting where \( \mathcal{Y} = \{0, 1\} \). Let \( p_0, p_1 \in \mathcal{P}(X) \) be the data 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 \( A \in B(X) \), 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, $R_{\delta_0}(\ell_{0/1}, A) = \frac{T}{T+1} p_0(A) + \frac{1}{T+1} p_1(A^c)$. An adversary of budget $\epsilon > 0$ can perturb any $x \in \mathcal{X}$ to $x' \in B_\epsilon(x)$. It follows that any $x \in A$ can be perturbed to $x' \in \bigcup_{A} B_\epsilon(a) = A^{\oplus \epsilon}$. Hence, adversarial risk can be defined as

$$R_{\delta_0}(\ell_{0/1}, A) = \frac{T}{T+1} p_0(A^{\oplus \epsilon}) + \frac{1}{T+1} p_1((A^c)^{\oplus \epsilon}).$$

(6)

The above formulation is a special case of (1) for the 0-1 loss function. Indeed, for $x \in \mathcal{X}$ and $y \in \{0, 1\}$, $\ell_{0/1}(x, y, A) = \mathbb{1}\{x \in A, y = 0\} + \mathbb{1}\{x \in A^c, y = 1\}$. Hence,

$$R_{\delta_0}(\ell_{0/1}, A) = \frac{T}{T+1} \mathbb{E}_{p_0}\left[\sup_{d(x, x') \leq \epsilon} \mathbb{1}\{x' \in A\}\right] + \frac{1}{T+1} \mathbb{E}_{p_1}\left[\sup_{d(x, x') \leq \epsilon} \mathbb{1}\{x' \in A^c\}\right].$$

$$= \frac{T}{T+1} p_0(A^{\oplus \epsilon}) + \frac{1}{T+1} p_1((A^c)^{\oplus \epsilon}).$$

(7)

A problem with the definition in equation (6) is the ambiguity over the measurability of the sets $A^{\oplus \epsilon}$ and $(A^c)^{\oplus \epsilon}$. Even when $A \in \mathcal{B}(\mathcal{X})$, it is not guaranteed that $A^{\oplus \epsilon}, (A^c)^{\oplus \epsilon} \in \mathcal{B}(\mathcal{X})$ (see Appendix C-A for an example). Hence, $R_{\delta_0}(\ell_{0/1}, A)$ is not well-defined for all $A \in \mathcal{B}(\mathcal{X})$. It is shown in [19] that $R_{\delta_0}(\ell_{0/1}, A)$ 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 [37]. This leads to the following formulation.

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

(8)

The above definition is well-defined for any $A \in \mathcal{B}(\mathcal{X})$ because $A^c$ and $(A^c)^c$ are both closed and hence, measurable. However, under the above definition, a point $x \in A$ may be perturbed to $x' \in A'$ such that $d(x, x') > \epsilon$. For example, when $A = (0, 1)$, we have $A' = [-\epsilon, 1+\epsilon]$ and an adversary may transport $x = \delta > 0$ to $x' = -\epsilon$, violating the budget constraint at $x$.

Another approach to defining adversarial risk is by explicitly defining the class of adversaries of budget $\epsilon$ as measurable transport maps $f : \mathcal{X} \to \mathcal{X}$ that push-forward the true data distribution such that no point is transported more than a distance of $\epsilon$; i.e., $d(x, f(x)) \leq \epsilon$. Let $\mathcal{F}$ denote the set of such budget-constrained transport maps. The transport map-based adversarial risk [18] is formally defined as follows:

$$R_{F_\epsilon}(\ell_{0/1}, A) = \sup_{f_0, f_1 \in \mathcal{F}} \frac{T}{T+1} f_0 p_0(A) + \frac{1}{T+1} f_1 p_1(A^c).$$

(8)

It is easy to see that the above definition is a special case of the definition in equation (2) for the 0-1 loss function. Yet another definition uses the robust hypothesis framework with $W_\infty$ uncertainty sets. In this approach, an adversary perturbs the true distribution $p_t$ to a corrupted distribution $p'_t$ such that $W_\infty(p_t, p'_t) \leq \epsilon$. This is equivalent to the existence of a coupling $\pi \in \Pi(p_t, p'_t)$ such that $\text{ess sup}_{(x, x') \in \pi} d(x, x') \leq \epsilon$. The adversarial risk with such an adversary is given by

$$R_{F_\epsilon}(\ell_{0/1}, A) = \sup_{W_\infty(p_t, p'_t), W_\infty(p_t, p'_t) \leq \epsilon} \frac{T}{T+1} p_0(A) + \frac{1}{T+1} p_1'(A^c).$$

(9)

Clearly, $R_{F_\epsilon}(\ell_{0/1}, A) \leq R_{\delta_0}(\ell_{0/1}, A)$, but conditions for equality were not studied in prior work. Moreover, their relation to set expansion-based definitions in (6) and (7) was also unknown.

Next we discuss some characterizations of optimal adversarial risk, defined as,

$$R_{\delta_0}^* := \inf_{A \in \mathcal{B}(\mathcal{X})} R_{\delta_0}(\ell_{0/1}, A).$$

(10)

In [22] and [23], it is shown that $R_{\delta_0}^* = \frac{1}{2}[1 - D_e(p_0, p_1)]$ for equal priors ($T = 1$), where $D_e$ is an optimal transport cost defined as follows.

**Definition 1 (D_e Cost):** Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Let $\epsilon \geq 0$. Let $c_\epsilon : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be such that $c_\epsilon(x, x') = 1 \{d(x, x') > \epsilon\}$. Then for $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $\epsilon \geq 0$, $D_e(\mu, \nu) = T_{\epsilon}(\mu, \nu)$.

For $\epsilon = 0$, $D_e$ reduces to the total variation distance. While $D_0$ is a metric on $\mathcal{P}(\mathcal{X})$, $D_e$ (for $\epsilon > 0$) is neither a metric nor a pseudometric [19]. Other formulations of optimal adversarial risk are inspired from game theory [18], [20], [29]. Consider a game between two players: (1) The adversary whose action space is pairs of distributions $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$, and (2) The algorithm 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,

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

The payoff when the algorithm plays first is given by $\inf_{A \in \mathcal{B}(\mathcal{X})} \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0, p_1')$, and this quantity is interpreted as the optimal adversarial risk in this setup.

**1) Error-Region Setting:** The formulations in equations (1), (6) and (7) 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 [17] and “corrupted-instance” risk in [16] and [37]. 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 $2\epsilon$. This is not the case with other formulations of adversarial risk such as “exact-in-the-ball” risk [17], “prediction-change” risk and “error-region” risk [16], [37]. Although the main focus of our work is the “corrupted-instance” family of risks, our tools also help clarify the well-definedness and equivalences among the “error-region” family of risks. In this subsection, we introduce the “error-region” definitions of adversarial risk.

Let $B \in \mathcal{B}(\mathcal{X})$ be the decision region corresponding to a base classifier $f_B(x) = \mathbb{1}\{x \in B\}$. The base classifier is typically chosen to be the Bayes optimal classifier i.e. $R_0^* = R_0(\ell_{0/1}, B)$. Given a classifier with decision region
For $A \in \mathcal{B}(X)$, the error-region adversarial risk is defined to be the measure of the $\epsilon$-expansion of the error region $A \Delta B := (A \cap B^{\complement}) \cup (B \cap A^{\complement})$ with respect to the true data distribution, $p := \frac{T}{T+1}p_0 + \frac{1}{T+1}p_1$. Analogous to $R_{\epsilon|\alpha}$ and $R_\epsilon$, we have the following definitions of error-region adversarial risk.

$$R_{\epsilon|\alpha}'(\ell_{0/1}, A \mid B) = p((A \Delta B)^{\epsilon}), \quad (11)$$

$$R_\epsilon'(\ell_{0/1}, A \mid B) = p((A \Delta B)'^{\epsilon}), \quad (12)$$

Similarly, we have the following error-region definitions that are analogous to $R_{\epsilon|\alpha}$ and $R_\epsilon$.

$$R_{\epsilon|\alpha}'(\ell_{0/1}, A \mid B) = \sup_{f_0, f_1 \in \mathcal{F}} \frac{T}{T+1}f_{0\alpha p_0}(A \Delta B) + \frac{1}{T+1}f_{1\beta p_1}(A \Delta B), \quad (13)$$

$$R_\epsilon'(\ell_{0/1}, A \mid B) = \sup_{W_{\infty}(p_0, p_1) \leq \epsilon} \frac{T}{T+1}\ell_0'(A \Delta B) - \frac{1}{T+1}\ell_1'(A \Delta B). \quad (14)$$

With the definition in $(11)$, an error is made on $x$ if $x$ is $\epsilon$-close to the error region $A \Delta B$. This corresponds to the setting of [16], [17], [37], and [42]. However, as discussed previously, $R_{\epsilon|\alpha}'$ is not well-defined because the Minkowski set expansion $(A \Delta B)^{\epsilon}$ may not be measurable. Hence, the closed set expansion based definition in $(12)$ is typically used instead, or both the definitions are erroneously conflated as one. The transport map based definition in $(13)$ and the $W_{\infty}$ uncertainty set based definition in $(14)$ present alternative ways to fix the measurability issues with $R_{\epsilon|\alpha}'$, but we are unaware of any works using these formulations for error-region adversarial risk.

With all the definitions $(11)$, $(12)$, $(13)$ and $(14)$, the optimal classifier for any adversarial budget $\epsilon > 0$ is simply the base classifier $f_0(x) = 1(x \in B)$ and the optimal adversarial risk is 0. Consequently, there is no need for a trade-off between standard risk and adversarial risk, which is one of the main reasons why these definitions are considered in place of the “corrupted-instance” family of risks.

In general, the adversarial risk in the “corrupted-instance” definitions $(11)$, $(12)$, $(13)$ and $(14)$ is not equal to the corresponding notions of adversarial risk in “error-region” definitions $(6)$, $(7)$, $(8)$ and $(9)$. However, the following lemma provides upper and lower bounds on the “corrupted-instance” risks in terms of the corresponding “error-region” risks. For the purpose of the following lemma, we assume that $R_{\epsilon|\alpha}'(\ell_{0/1}, A \mid B)$ is well-defined, but we will state the precise conditions for well-definedness and equivalences among “error-region” risks in subsection IV-A.1.

**Lemma I:** Let $p_0, p_1 \in \mathcal{P}(X)$. Let $A, B \in \mathcal{B}(X)$. Then inequalities of the following type hold.

$$R_{\square}(\ell_{0/1}, A) - R_{\square}(\ell_{0/1}, B) \leq R_{\epsilon|\alpha}'(\ell_{0/1}, A \mid B) \leq R_{\epsilon}(\ell_{0/1}, A) - R_{\epsilon}(\ell_{0/1}, B),$$

where $R_{\square}$ can be substituted by $R_\epsilon, R_{\epsilon|\alpha},$ or $R_{\epsilon|\alpha}'$ (and correspondingly, $R_{\epsilon|\alpha}'$ by $R_\epsilon, R_{\epsilon|\alpha},$ or $R_{\epsilon|\alpha}'$). If $A, B$ are such that $R_{\epsilon|\alpha}'(\ell_{0/1}, A \mid B), R_{\epsilon|\alpha}(\ell_{0/1}, A)$ and $R_{\epsilon|\alpha}(\ell_{0/1}, B)$ are well-defined, then $R_{\square}$ can also be substituted by $R_{\epsilon|\alpha}'$ (and correspondingly, $R_{\epsilon|\alpha}'$ by $R_{\epsilon|\alpha}$).

The proof of Lemma 1 is in Appendix B. From Lemma 1 it is clear that a sufficient condition for the equivalence between the “error-region” risks and the corresponding “corrupted-instance” risks is that the base classifier $B$ achieves 0 “corrupted-instance” risk. This holds when the supports of the distributions $p_0$ and $p_1$ are at least 2$\epsilon$ distance apart. This condition is equivalent to the margin condition in Theorem 6 of [42] and the $\epsilon$-separation condition of [43].

**Remark:** Unless otherwise specified, throughout the remainder of the paper, the term “adversarial risk” refers to the $R_{\epsilon|\alpha}'$ definition.

**Example:** Let $(X, d) = (\mathbb{R}, \cdot | \cdot )$. Let $\epsilon \in (0, 1/2]$. Let $p_0 = (1-\alpha_0)\delta_{-1} + \alpha_0 \delta_{-\epsilon}$ and $p_1 = (1-\alpha_1)\delta_1 + \alpha_1 \delta_{\epsilon}$, where $\delta_x$ denotes the Dirac measure at $x$ and $\alpha_0, \alpha_1 \in [0, 1]$. Consider equal priors i.e., $T = 1$. Let $A = \{0, 2\epsilon\}$ and $B = \{0, 1 + 2\epsilon\}$. It is easy to see that $R_0(\ell_{0/1}, A) = \frac{\alpha_0}{2}$ and $R_0(\ell_{0/1}, B) = 0$. We have the following computation for various definitions of adversarial risk. (For the sake of brevity, we omit the argument $\ell_{0/1}$ from the adversarial risk expressions.)

$$R_{\epsilon|\alpha}(A) = \frac{1}{2}p_0([-\epsilon, 3\epsilon]) + \frac{1}{2}p_1([\epsilon, 3\epsilon]) = \frac{\alpha_0 + 1 - \alpha_1}{2},$$

$$R_{\epsilon|\alpha}(B) = \frac{1}{2}p_0([-\epsilon, 1 + 3\epsilon]) + \frac{1}{2}p_1([\epsilon, 1 + 3\epsilon]) = \frac{\alpha_0}{2},$$

$$R_{\epsilon}(A) = \frac{1}{2}p_0([-\epsilon, 3\epsilon]) + \frac{1}{2}p_1([\epsilon, 3\epsilon]) = \frac{\alpha_0 - \alpha_1}{2},$$

$$R_{\epsilon}(B) = \frac{1}{2}p_0([-\epsilon, 1 + 3\epsilon]) + \frac{1}{2}p_1([\epsilon, 1 + 3\epsilon]) = \frac{\alpha_0 - \alpha_1}{2},$$

Observe that for $\alpha_0 = 0$, we have $R_{\epsilon|\alpha}(\ell_{0/1}, B) = 0$ and as expected from Lemma 1, we have $R_{\epsilon|\alpha}(\ell_{0/1}, A \mid B) = R_{\epsilon|\alpha}(\ell_{0/1}, A)$. Similarly for $\alpha_0 = \alpha_1 = 0$ we have $R_{\epsilon}(\ell_{0/1}, B) = 0$ and hence $R_{\epsilon}(\ell_{0/1}, A \mid B) = R_{\epsilon}(\ell_{0/1}, A)$.

**IV. WELL-DEFINEDNESS OF ADVERSARIAL RISK**

In this section, we discuss the conditions under which the definitions for adversarial risk presented in Section III are well-defined. In subsection IV-A we present the results for the binary classification setting under 0-1 loss and in subsection IV-B we discuss more general loss functions.

**A. Binary Classification With 0-1 Loss Setting**

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

**Lemma 2:** For any $\epsilon > 0$, there exists $A \in \mathcal{B}(X)$ such that $A^{\epsilon} \notin \mathcal{B}(X)$.

The proof of Lemma 2 is in Appendix C-A.

In this section, we lay down the conditions under which the ambiguity on the measurability of $A^{\epsilon}$ can be resolved. We begin by presenting a lemma that shows that $A^{\epsilon}$ 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 $\mathcal{B}(\mathcal{X})$, the universal completion of the Borel $\sigma$-algebra $\mathcal{B}(\mathcal{X})$, and are measurable with respect to any finite measure defined on the complete measure space, $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Lemma 3: Let $A \in \mathcal{B}(\mathcal{X})$. Then, $A^{\mathcal{B}}$ is an analytic set. Consequently, $A^{\mathcal{B}} \in \mathcal{B}(\mathcal{X})$.

The proof of Lemma 3 is in Appendix C-A. By virtue of Lemma 3, we have the following.

Theorem 4: Let $p_0, p_1 \in \mathcal{P}(\mathcal{X})$. Then for any $A \in \mathcal{B}(\mathcal{X})$, $R_{\mathcal{B}}(\ell_{0/1}, A)$ is well-defined.

For the special case of $\mathcal{X} = \mathbb{R}^d$, we can further strengthen Theorem 4 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 2 (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 $x \in E$, there is an $x' \in \mathcal{X}$ such that $B_{\alpha r}(x') \subseteq B_r(x) \setminus E$.

Porous sets are a subclass of nowhere dense sets. Importantly, $\lambda(E) = 0$ for any porous set $E \subset \mathbb{R}^d$ [44]. By the following lemma, the set difference between the closed/open set expansions is porous.

Lemma 5: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$ and $A \in \mathcal{L}(\mathcal{X})$. Then $E = A\setminus A^\delta$ is porous.

The proof of Lemma 5 is in Appendix C-A. Lemma 5 plays a crucial role in proving that $A^{\mathcal{B}} \in \mathcal{L}(\mathcal{X})$ whenever $A \in \mathcal{L}(\mathcal{X})$. We recall that $A^{\mathcal{B}}$ is the Minkowski sum of $A$ with the closed $c$-ball. In general, the Minkowski sum of two Lebesgue measurable sets is not always Lebesgue measurable [40], [45]. So the fact that one of them is a closed ball in case of $A^{\mathcal{B}}$ is important.

In the following theorem, we use Lemma 5 to prove the measurability of $A^{\mathcal{B}}$ and in turn prove that $R_{\mathcal{B}}(\ell_{0/1}, A)$ is well-defined.

Theorem 6: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_0, p_1 \in \mathcal{P}(\mathcal{X})$ and let $\epsilon \geq 0$. Then for any $A \in \mathcal{L}(\mathcal{X})$, $R_{\mathcal{B}}(\ell_{0/1}, A)$ is well-defined. If, in addition, $p_0$ and $p_1$ are absolutely continuous with respect to the Lebesgue measure, then $R_{\mathcal{B}}(\ell_{0/1}, A | B) = R_{\mathcal{B}}(\ell_{0/1}, A | B)$.

Proof: By Lemma 5 $A\setminus A^\epsilon$ is porous, and so $\lambda(A\setminus A^\epsilon) = 0$. Hence, $\lambda(A^\epsilon) = \lambda(A)$. Using the fact that $A^\epsilon \subseteq A^{\mathcal{B}} \subseteq A^\epsilon$, we have $\lambda(A^{\mathcal{B}} \setminus A^\epsilon) = 0$. Therefore, $A^{\mathcal{B}} \in \mathcal{L}(\mathcal{X})$ and $\lambda(A^{\mathcal{B}}) = \lambda(A)$.

Since $A^{\mathcal{B}} \in \mathcal{L}(\mathcal{X})$, $R_{\mathcal{B}}(\ell_{0/1}, A)$ is well-defined. If $p_0$ and $p_1$ are absolutely continuous with respect to the Lebesgue measure, the equation $R_{\mathcal{B}}(\ell_{0/1}, A) = R_{\mathcal{B}}(\ell_{0/1}, A)$ follows from the previous conclusion.

1) Error-Region Setting: The following theorem on the well-definedness of the error-region adversarial risk is an immediate consequence of Lemma 3.

Theorem 7: Let $p_0, p_1 \in \mathcal{P}(\mathcal{X})$. Then for any $A \in \mathcal{B}(\mathcal{X})$, $R_{\mathcal{B}}(\ell_{0/1}, A)$ is well-defined.

Similar to Theorem 6, we get the following strengthening of Theorem 7 for the special case of $\mathcal{X} = \mathbb{R}^d$. The proof strategy is identical to that of Theorem 6.

Theorem 8: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_0, p_1 \in \mathcal{P}(\mathcal{X})$ and let $\epsilon \geq 0$. Then for any $A, B \in \mathcal{L}(\mathcal{X})$, $R_{\mathcal{B}}(\ell_{0/1}, A | B)$ is well-defined. If, in addition, $p_0$ and $p_1$ are absolutely continuous with respect to the Lebesgue measure, then $R_{\mathcal{B}}(\ell_{0/1}, A | B) = R_{\mathcal{B}}(\ell_{0/1}, A | B)$.

Example (Continued): Observe that for $A = [0, 2\epsilon]$, $E = A \setminus A^\epsilon = [-\epsilon, 3\epsilon] \setminus (-\epsilon, 3\epsilon) = \{\epsilon, 3\epsilon\}$ is indeed porous. However, $p_0 = (1 - \alpha_0)\delta_{-\epsilon} + \alpha_0\delta_\epsilon$ and $p_1 = (1 - \alpha_1)\delta_{-\epsilon} + \alpha_1\delta_\epsilon$ are not absolutely continuous with respect to the Lebesgue measure. Hence, $R_{\mathcal{B}}(\ell_{0/1}, A) = R_{\mathcal{B}}(\ell_{0/1}, A)$ and $R_{\mathcal{B}}(\ell_{0/1}, A | B) = R_{\mathcal{B}}(\ell_{0/1}, A | B)$.

B. General Loss Setting

In the expected-supremum formulation of adversarial risk shown in (1), the worst-case loss function $\sup_{\mathcal{P}_d(x,x')} \ell((x', y), w)$ 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 $\phi : \mathcal{X} \to \mathbb{R}$ is called upper semi-analytic if the set $\{x \in \mathcal{X} : \phi(x) > t\}$ 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 $\sup_{\mathcal{P}_d(x,x')} \ell((x', y), w)$ is universally measurable if $\ell((x', y), w)$ is upper semi-analytic for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$.

Lemma 9: If the loss function $\ell((x', y), w)$ is upper semi-analytic for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, then the worst-case loss function $\sup_{\mathcal{P}_d(x,x')} \ell((x', y), w)$ is also upper semi-analytic and hence universally measurable. Therefore, $R_{\mathcal{B}}(\ell, w)$ is well-defined on the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

The proof of Lemma 9 is in Appendix C-B. 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 10: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$. Then, $R_{\mathcal{B}}(\ell, w)$ is well-defined for any loss function $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{W} \to [0, \infty]$ for which $\ell((x, y), w)$ is Lebesgue measurable for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$.

The proof of Lemma 10 is in Appendix C-B.

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

V. EQUIVALENCE WITH $\infty$-WASSERSTEIN ROBUSTNESS

In this section, we show the conditions under which $R_{\mathcal{B}}(\ell, w)$ is equivalent to other notions of adversarial risk based on transport maps and $W_\infty$ robustness. The equivalences established in this section are summarized in Tables I and II. In Subsection V-A, we consider general Polish spaces and in Subsection V-B, we consider the Euclidean space.
TABLE I

| Equivalences in Adversarial Risk | Conditions |
|---------------------------------|------------|
| $R_{\tilde{W}}(A) = R_{\tilde{R}}(A)$ | $\mathbb{R}^d; A \in \mathcal{L}(\mathcal{X})$ or $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$: $A \in \mathcal{B}(\mathcal{X})$ |
| $R_{\hat{W}}(A) = R_{\hat{R}}(A)$ | $\mathbb{R}^d; A \in \mathcal{L}(\mathcal{X})$ and $p_0, p_1$ have densities |

TABLE II

| Equivalences in Adversarial Risk | Conditions |
|---------------------------------|------------|
| $R_{\tilde{W}}(w) = R_{\tilde{R}}(w)$ | $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ denotes the universal completion of the Borel measure space, $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ |
| $R_{\hat{W}}(w) = R_{\hat{R}}(w) = R_{\hat{R}}(w)$ | $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ denotes the universal completion of the Borel measure space |

A. $W_\infty$ Robustness in Polish Spaces via Measurable Selections

We begin by presenting a lemma that links the measure of $\epsilon$-Minkowski set expansion to the worst case measure over a $W_\infty$ probability ball of radius $\epsilon$.

**Lemma 11:** Let $\mu \in \overline{\mathcal{P}}(\mathcal{X})$ and $A \in \mathcal{B}(\mathcal{X})$. Then
\[
\sup_{W_\infty(\mu, \mu') \leq \epsilon} \mu'(A) = \mu(A^{\#\epsilon}).
\]

Moreover, the supremum in the previous equation is achieved by a $\mu^* \in \overline{\mathcal{P}}(\mathcal{X})$ that is induced from $\mu$ via a measurable transport map $\phi : \mathcal{X} \to \mathcal{X}$ (i.e. $\mu^* = \phi_\mu$) satisfying $d(x, \phi(x)) \leq \epsilon$ for all $x \in \mathcal{X}$.

The proof of Lemma 11 is in Appendix D-A. A crucial step in the proof of Lemma 11 is finding a measurable transport map $\phi$ such that $\phi^{-1}(A) = A^{\#\epsilon}$ and $d(x, \phi(x)) \leq \epsilon$ for all $x \in \mathcal{X}$. In the following theorem, we use Lemma 11 to establish the equivalence between the three different notions of adversarial risk introduced in section III.

**Theorem 12:** Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and $A \in \mathcal{B}(\mathcal{X})$. Then
\[
R_{\tilde{W}}(\ell_0/1, A) = R_{\tilde{F}}(\ell_0/1, A) = R_{\hat{F}}(\ell_0/1, A).
\]
In addition, the supremum over $f_0$ and $f_1$ in $R_{\tilde{F}}(\ell_0/1, A)$ is attained. Similarly, the supremum over $p_0'$ and $p_1'$ in $R_{\hat{G}}(\ell_0/1, A)$ is attained.

**Proof:** Since $A \in \mathcal{B}(\mathcal{X})$, $A^c \in \mathcal{B}(\mathcal{X})$ and by Lemma 3, $A^{\#\epsilon} (A^{\#\epsilon}) \in \mathcal{B}(\mathcal{X})$. Therefore $R_{\tilde{W}}(\ell_0/1, A)$ is well-defined. By Lemma 11, we have
\[
R_{\tilde{G}}(\ell_0/1, A) = \sup_{W_\infty(p_0, p_1) \leq \epsilon} \left[ \frac{T}{T+1} p_0'(A) + \frac{1}{T+1} p_1'(A^{\#\epsilon}) \right].
\]
By Lemma 11 again, the supremum over $p_0'$ and $p_1'$ in $R_{\hat{G}}(\ell_0/1, A)$ 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.

Similar to Theorem 12, we have the following theorem for error-region based definitions of adversarial risk.

**Theorem 13:** Let $p_0, p_1 \in \overline{\mathcal{P}}(\mathcal{X})$ and $A \in \mathcal{B}(\mathcal{X})$. Then
\[
R_{\tilde{W}}(\ell_0/1, A) = R_{\tilde{F}}(\ell_0/1, A) = R_{\hat{F}}(\ell_0/1, A).
\]
In addition, the supremum over $f_0$ and $f_1$ in $R_{\tilde{F}}(\ell_0/1, A)$ is attained. Similarly, the supremum over $p_0'$ and $p_1'$ in $R_{\hat{G}}(\ell_0/1, A)$ is attained.

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

**Lemma 14:** Let $\mu \in \overline{\mathcal{P}}(\mathcal{X})$. Then for any real-valued upper semi-continuous function $\phi : \mathcal{X} \to [0, \infty)$,
\[
\sup_{W_\infty(\mu, \mu') \leq \epsilon} \mathbb{E}_{x \sim \mu} \phi(x) = \mathbb{E}_{x \sim \mu} \left[ \sup_{d(x, x') \leq \epsilon} \phi(x') \right].
\]
Moreover, if the function $\phi$ 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 $\mu$ via a universally measurable transport map $m : \mathcal{X} \to \mathcal{X}$ (i.e. $\mu^* = m_\mu$) satisfying $d(x, m(x)) \leq \epsilon$ for all $x \in \mathcal{X}$.

The proof of Lemma 14 is in Appendix D-A. We note that Lemma 1 of [28] is similar in spirit to Lemma 14, but ignores the measurability issues.

Using Lemma 14, we prove the following theorem, which generalizes Theorem 12 to more general loss functions.

**Theorem 15:** If the loss function $\ell((x, y), w)$ is upper semi-continuous for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, then $R_{\tilde{W}}(\ell, w) = R_{\tilde{G}}(\ell, w)$. In addition, $\ell((x, y), w)$ is upper semi-continuous for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, then $R_{\tilde{W}}(\ell, w) = R_{\hat{G}}(\ell, w) = R_{\hat{R}}(\ell, w) = R_{\hat{R}}(\ell, w)$.

**Proof:**
\[
R_{\tilde{G}}(\ell, w) = \sup_{\gamma \in \Gamma} \mathbb{E}_{x \sim \gamma} \phi_{\gamma} \mathbb{E}_{x \sim \gamma} \left[ \ell((x', y), w) \right] = \mathbb{E}_{(x, y) \sim \rho_{\gamma}} \sup_{d(x, x') \leq \epsilon} \ell((x', y), w)
\]
where the second inequality follows from Lemma 14 because of the assumption that $\ell((\cdot, y), w)$ is upper semi-anchytic for all $y \in Y$ and $w \in W$.

With the stronger assumption that $\ell((\cdot, y), w)$ is upper semi-continuous for all $y \in Y$ and $w \in W$, Lemma 14 shows that for every $y \in Y$, there exists a universally measurable transport map $m_y : X \to X$ satisfying $d(x, m(x)) \leq \epsilon$ for all $x \in X$ such that the following holds.

$$R_{\Gamma_\epsilon}(\ell, w) = \sup_{\gamma \in \Gamma_\epsilon} \mathbb{E}(x, y) \sim p_{\gamma}^0 \mathbb{E}(x', y) \left[ \ell((x', y), w) \right]$$

$$= \sup_{F \in F_\epsilon} \mathbb{E}(x, y) \sim p_{\gamma}^0 \mathbb{E}(f_\gamma(x), y) \left[ \ell(f_\gamma(x), y), w) \right]$$

$$\leq \sup_{F \in F_\epsilon} \mathbb{E}(x, y) \sim p_{\gamma}^0 \left[ \ell(f_\gamma(x), y), w) \right]$$

$$= R_{\Gamma_\epsilon}(\ell, w).$$

Combining the above inequality with (5), we have $R_{\epsilon}(\ell, w) = R_{\epsilon}(\ell, w) = R_{\Gamma_\epsilon}(\ell, w)$. $\Box$

Example (Continued): Continuing our example with $p_\epsilon = (1 - \alpha_0)\delta_{-1} + \alpha_0\delta_{-1}$, $p_1 = (1 - \alpha_1)\delta_0 + \alpha_1\delta_0$, and $A = [0, 2\epsilon]$, by applying Theorem 12, we get the following.

$$R_{\epsilon}(\ell_{0, 1}, A) = R_{\Gamma_\epsilon}(\ell_{0, 1}, A) = R_{\epsilon}(\ell_{0, 1}, A) = \frac{\alpha_0 + 1 - \alpha_1}{2}.$$

The supremum in $R_{\epsilon}(\ell_{0, 1}, A)$ is attained for $f_0, f_1 : \mathbb{R} \to \mathbb{R}$ defined as $f_0(x) = x + \epsilon$ and $f_1(x) = x - \epsilon$. The supremum in $R_{\Gamma_\epsilon}(\ell_{0, 1}, A)$ is attained for $p_\epsilon'(0) = (1 - \alpha_0)\delta_{-1} + \alpha_0\delta_0$, $p_\epsilon'(1) = (1 - \alpha_1)\delta_1 + \alpha_1\delta_0$. We note that the choices of $f_0, f_1$ for attaining the supremum in $R_{\epsilon}(\ell_{0, 1}, A)$ are not unique.

Similarly, we get the following by applying Theorem 13.

$$R_{\epsilon}'(\ell_{0, 1}, A) = R_{\epsilon}'(\ell_{0, 1}, A) = R_{\epsilon}'(\ell_{0, 1}, A) = \frac{1 - \alpha_1}{2}.$$

The supremum in $R_{\epsilon}'(\ell_{0, 1}, A)$ is attained for any arbitrary $f_0, f_1 : \mathbb{R} \to \mathbb{R}$ satisfying the perturbation budget constraint. Similarly, the supremum in $R_{\epsilon}'(\ell_{0, 1}, A)$ is attained for any arbitrary $p_\epsilon'$, $p_\epsilon'$ satisfying the $W_\infty$ perturbation budget constraint.

B. $\infty$ Robustness in $\mathbb{R}^d$ via 2-Alternating Capacities

In this subsection, we establish a connection between adversarial risk and Choquet capacities [33] in $\mathbb{R}^d$. This allows us to extend Theorem 12 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 Section VII-A. We begin with the following definitions.

Definition 3 (Capacity): A set function $v : \mathcal{B}(X) \to [0, 1]$ is a capacity if it satisfies the following conditions: (1) $v(\emptyset) = 0$ and $v(X) = 1$; (2) For $A, B \in \mathcal{B}(X)$, $A \subseteq B \implies v(A) \leq v(B)$; (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 4 (2-Alternating Capacity): A capacity $v$ defined on the measure space $(X, \mathcal{B}(X))$ is called 2-alternating if $v(A \cup B) + v(A \cap B) \leq v(A) + v(B)$ for all $A, B \in \mathcal{B}(X)$.

For any compact set of probability measures $\Xi \subseteq \mathcal{P}(X)$, the upper probability defined as $v(A) = \sup_{\mu \in \Xi} \mu(A)$ is a capacity [34]. The upper probability of $\epsilon$-neighborhoods of a $\mu \in \mathcal{P}(X)$ defined using either the total variation metric or the Levy-Prokhorov metric can be shown to be a 2-alternating capacity [34]. The following lemma shows that $A \mapsto \mu(A^{\infty})$ is a 2-alternating capacity under some conditions.

Lemma 16: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $\mu \in \mathcal{P}(X)$ and let $\epsilon \geq 0$. Define a set function $v$ on $X$ such that for any $A \in \mathcal{L}(X)$, $v(A) := \mu(A^\infty)$. Then $v$ is a 2-alternating capacity.

The proof of Lemma 16 is included in Appendix D-B.

Now we relate the capacity defined in Lemma 16 to the $W_\infty$ metric. Since the $\epsilon$-neighborhood of a $\mu \in \mathcal{P}(X)$ in $W_\infty$ metric is a compact set of probability measures [46], the upper probability over this $W_\infty$ $\epsilon$-ball is a capacity. The following lemma shows that it is a 2-alternating capacity, and identifies it with the capacity defined in Lemma 16.

Lemma 17: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $\mu \in \mathcal{P}(X)$. Then for any $A \in \mathcal{L}(X)$, $\sup_{W_\infty (p_\epsilon, p_\epsilon') \leq \epsilon} \mu(A^{\infty}) = \mu(A^{\infty})$. Moreover, the supremum in the previous equation is attained.

The proof of Lemma 17 is included in Appendix D-B. Lemma 17 plays a similar role to Lemma 11 in proving the following equivalence between adversarial robustness and $W_\infty$ robustness.

Theorem 18: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_\epsilon, p_1 \in \mathcal{P}(X)$ and let $\epsilon \geq 0$. Then for any $A \in \mathcal{L}(X)$, $R_{\epsilon}(\ell_{0, 1}, A) = R_{\epsilon'}(\ell_{0, 1}, A)$, and the supremum over $p_\epsilon'$ and $p_1'$ in $R_{\epsilon'}(\ell_{0, 1}, A)$ is attained.

Proof: Observe that

$$R_{\epsilon'}(\ell_{0, 1}, A) = \frac{T}{T + 1} \left( \sup_{W_\infty (p_\epsilon, p_\epsilon') \leq \epsilon} p_\epsilon'(A) + \frac{1}{T + 1} p_1'(A^{\infty}) \right)$$

where $(*)$ follows from Lemma 17. By Lemma 17 again, the supremum over $p_\epsilon'$ and $p_1'$ in $R_{\epsilon'}(\ell_{0, 1}, A)$ is attained.

Unlike Theorem 12, Theorem 18 does not show the equivalence of $R_{\epsilon'}(\ell_{0, 1}, A)$ with the other definitions under the relaxed assumption of $A \in \mathcal{L}(X)$. This is because Lemma 17 does not provide a push-forward map $\phi$ such that $\mu^* = \phi_{2\mu^*}$ with $\mu^*$ attaining the supremum over the $W_\infty$ ball.

Similar to Theorem 18, we have the following theorem for error-region based definitions of adversarial risk.

Theorem 19: Let $(X, d) = (\mathbb{R}^d, \|\cdot\|)$. Let $p_\epsilon, p_1 \in \mathcal{P}(X)$ and let $\epsilon \geq 0$. Then for any $A \in \mathcal{L}(X)$, $R_{\epsilon'}(\ell_{0, 1}, A) = R_{\epsilon'}(\ell_{0, 1}, A)$, and the supremum over $p_\epsilon'$ in $R_{\epsilon'}(\ell_{0, 1}, A)$ is attained.

VI. OPTIMAL ADVERSARIAL RISK VIA GENERALIZED STRASSEN’S THEOREM

In Section V, we analyzed adversarial risk for a specific decision region $A \in \mathcal{B}(X)$. In this section, we analyze infimum of adversarial risk over all possible decision regions; i.e., the optimal adversarial risk. We show that optimal adversarial risk in binary classification with unequal priors is characterized by an unbalanced optimal transport cost.
between data-generating distributions. Our main technical lemma generalizes Strassen’s theorem to unbalanced optimal transport. We present this result in Subsection VI-A and present our characterization of optimal adversarial risk in Subsection VI-B.

Remark: We note that we only consider “corrupted-instance” risks in this section. For error-region risks, the base classifier $B$ is trivially the optimal classifier.

A. Unbalanced Optimal Transport and Generalized Strassen’s Theorem

Recall from Section III that the optimal transport cost $D_{c}$ characterizes the optimal adversarial risk in binary classification for equal priors. The following result gives an alternative characterization of $D_{c}$.

**Proposition 1 (Strassen’s theorem):** (Corollary 1.28 in [36]) Let $\mu, \nu \in P(\mathcal{X})$. Let $\epsilon \geq 0$. Then

$$\sup_{A \in B(\mathcal{X})} \mu(A) - \nu(A^{\epsilon}) = D_{c}(\mu, \nu).$$

**Proof:**

Proposition 1 is a special case of Kantorovich-Rubinstein duality [36] applied to $\{0, 1\}$-valued cost functions. We now generalize this result to measures with unequal masses. We begin with some definitions that generalize the concepts we introduced in Subsection II-B.

Let $\mu, \nu \in M(\mathcal{X})$ be such that $\mu(\mathcal{X}) \leq \nu(\mathcal{X})$. A coupling between $\mu$ and $\nu$ is a measure $\pi \in M(\mathcal{X}^{2})$ such that for any $A \in B(\mathcal{X})$, $\pi(A \times \mathcal{X}) = \mu(A)$ and $\pi(\mathcal{X} \times A) \leq \nu(A)$. The set $\Pi(\mu, \nu)$ is defined to be the set of all couplings between $\mu$ and $\nu$. For a cost function $c : \mathcal{X}^{2} \rightarrow [0, \infty)$, the optimal transport cost between $\mu$ and $\nu$ under $c$ is defined as $D_{c}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X}^{2}} c(x, x') d\pi(x, x').$

**Theorem 20 (Generalized Strassen’s Theorem):** Let $\mu, \nu \in M(\mathcal{X})$ be such that $0 < M = \mu(\mathcal{X}) \leq \nu(\mathcal{X})$. Let $\epsilon > 0$. Define $c_{\epsilon} : \mathcal{X}^{2} \rightarrow [0, 1]$ as $c_{\epsilon}(x, x') = 1\{x, x' \in \mathcal{X}^{2} : d(x, x') \geq 2\epsilon\}$. Then

$$\sup_{A \in B(\mathcal{X})} \mu(A) - \nu(A^{\epsilon}) = T_{c_{\epsilon}}(\mu, \nu).$$

Moreover, the infimum on the right hand side is attained. (Equivalently, there is a coupling $\pi \in \Pi(\mu, \nu)$ that attains the unbalanced optimal transport cost $T_{c_{\epsilon}}(\mu, \nu)$.)

The proof of Theorem 20 is contained in Appendix E-A. The leverages strong duality in linear programming. We first establish (17) 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 $T_{c_{\epsilon}}(\mu, \nu)$ in (17). We then show that the constructed coupling satisfies (17). This proof strategy is adapted from the works of [47] and [48].

B. Optimal Adversarial Risk for Unequal Priors

Generalized Strassen’s theorem involves closed set expansions. The following lemma allows us to switch to Minkowski set expansions.

**Lemma 21:** Let $\mu, \nu \in \overline{M}(\mathcal{X})$ and let $\epsilon \geq 0$. Then,

$$\sup_{A \in B(\mathcal{X})} \mu(A) - \nu(A^{\epsilon}) = \sup_{A \in B(\mathcal{X})} \mu(A^{\epsilon}) - \nu(A^{\epsilon}).$$

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 21 is contained in Appendix E-B. Using Lemma 21 and the generalized Strassen’s theorem, we show the following result on optimal adversarial risk for unequal priors, generalizing the result of [22] and [23].

**Theorem 22:** Let $p_{0}, p_{1} \in P(\mathcal{X})$ and let $\epsilon \geq 0$. Then,

$$\inf_{A \in B(\mathcal{X})} R_{\overline{\mathcal{X}}}(\ell_{0/1}, A) = \frac{1}{T + 1} \left[ 1 - \inf_{q \in P(\mathcal{X})} D_{c}(q, p_{1}) \right].$$

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

**Proof:**

Theorem 22 extends the result of [19] in two ways: (1) the infimum is taken over all sets for which $R_{\overline{\mathcal{X}}}(\ell_{0/1}, A)$ is well-defined, instead of restricting to closed sets, and (2) the priors on both labels can be unequal. We also note that for $(\mathcal{X}, d) = ([0, 1], || \cdot ||)$, (18) holds with the infimum on the left hand side taken over all $A \in L(\mathcal{X})$.

**Example (Continued):** We continue our example with $p_{0} = (1 - \alpha_{0})\delta_{-1} + \alpha_{0}\delta_{+} \quad \text{and} \quad p_{1} = (1 - \alpha_{1})\delta_{+} + \alpha_{1}\delta_{-}$. Consider unequal priors, i.e., $T > 1$. Let $q = (1 - \alpha_{q})\delta_{-1} + \alpha_{q}\delta_{+}$ for some $\alpha_{q} \in [0, 1]$. Then, $D_{c}(q, p_{1}) = 1 - \min\{\alpha_{q}, \alpha_{1}\}$.

Further,

$$\inf_{q \in P(\mathcal{X}) : q \not\subset T_{p_{0}}} D_{c}(q, p_{1}) = \inf_{\alpha_{q} \in [0, 1]} \{ 1 - \min\{\alpha_{q}, \alpha_{1}\} \} = 1 - \min\{T_{\alpha_{0}}, \alpha_{1}\}.$$
VII. Minimax Theorems and Nash Equilibria

In this section, we revisit the zero-sum game between the adversary and the algorithm introduced in Section III. Recall that for $A \in \mathcal{B}(X)$ and $p_0', p_1' \in \mathcal{P}(X)$, the payoff function is given by,

$$r(A, p_0', p_1') = \frac{T}{T + 1} p_0'(A) + \frac{1}{T + 1} p_1'((A^*)').$$

(19)

The max-min inequality gives the following.

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

(20)

If the inequality in (20) is an equality, we say that the game has zero duality gap, and admits a value equal to either expression in (20). In the equality setting, there is no advantage to a player making the first move. Our minmax theorems establish such an equality. If, in addition to having an equality in (20), there exist $p_0', p_1', A^* \in \mathcal{P}(X)$ that achieve the supremum on the left-hand side and $A^* \in \mathcal{B}(X)$ 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 $\delta$-approximate pure Nash equilibrium of the game if the following inequality holds.

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

In Section VII-A, we prove the minimax theorem and the existence of a pure Nash equilibrium in $\mathbb{R}^d$ using the theory of 2-alternating capacities [34] and the relation to adversarial risk from Section V-B. Section VII-B extends these results to more general Polish spaces with a “midpoint property.”

A. Minimax Theorem in $\mathbb{R}^d$ via 2-Alternating Capacities

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

**Theorem 23 (Minimax theorem in $\mathbb{R}^d$):** Let $(X, d) = (\mathbb{R}^d, \| \cdot \|)$. Let $p_0, p_1 \in \mathcal{P}(X)$ and let $\epsilon \geq 0$. Consider the zero-sum game with the payoff function $r$ as in (19). Then,

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

(21)

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

The proof of Theorem 23 is in Appendix F-A. Crucial to the proof of Theorem 23 is Lemma 16, which shows that the set-valued maps $A \mapsto p_0(A^{|\epsilon})$ and $A^c \mapsto p_1((A^c)^{|\epsilon})$ are 2-alternating capacities. The same proof technique is not applicable in general Polish spaces because the map $A \mapsto \mu(A^{|\epsilon})$ is not a capacity for a general $\mu \in \mathcal{P}(X)$. This is because $A^{|\epsilon}$ is not measurable for all $A \in \mathcal{B}(X)$.

B. Minimax Theorem in Polish Spaces via Optimal Transport

We now extend the minimax theorem from $\mathbb{R}^d$ to general Polish spaces with the following property.

**Definition 5 (Midpoint Property):** A metric space $(X, d)$ is said to have the midpoint property if for every $x_1, x_2 \in X$, there exists $x \in X$ 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 $d(x, x') = \mathbb{1}(x \neq x')$. The midpoint property plays a crucial role in proving the following theorem, which shows that the $D_e$ transport cost between two distributions is the shortest total variation distance between their $\epsilon$-neighborhoods in $W_{\infty}$ metric. A similar result was also presented in [25].

**Theorem 24 ($D_e$ as shortest $D_{TV}$ between $W_{\infty}$ balls):** Let $(X, d)$ have the midpoint property. Let $\mu, \nu \in \mathcal{P}(X)$ and let $\epsilon \geq 0$. Then $D_e(\mu, \nu) = \inf_{W_{\infty}(\mu, \nu')} \sup_{D_{TV}(\mu', \nu')} D_{TV}(\mu', \nu')$. Moreover, the infimum over $D_{TV}$ in the above equation is attained.

The proof of Theorem 24 is in Appendix F-B. The following theorem uses Theorem 24 to prove the minimax equality and the existence of a Nash equilibrium for any Polish space with the midpoint property for the case of equal priors.

**Theorem 25 (Minimax theorem for equal priors):** Let $(X, d)$ have the midpoint property. Let $p_0, p_1 \in \mathcal{P}(X)$ and let $\epsilon \geq 0$. Consider the zero-sum game with the payoff function $r$ as in (19) with $T = 1$. Then,

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

(22)

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

**Proof:** We have the following series of equalities.

$$\inf_{A \in \mathcal{B}(X)} \sup_{W_{\infty}(p_0, p_0')} r(A, p_0', p_1') = \inf_{A \in \mathcal{B}(X)} \sup_{W_{\infty}(p_0, p_0')} R_{\epsilon}(\ell_{0/1}, A),$$

and,

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

where (i) follows from Theorem 12, (ii) from Theorem 22, and (iii) again from Theorem 22 with $\epsilon = 0$. The expressions on
the right extremes of the above equations are equal by Theorem 24. The existence of \( p_0^*, p_1^* \in \mathcal{P}(X) \) follows Theorem 24.

To prove the minimax theorem for unequal priors, we need the following generalization of Theorem 24 to finite measures of unequal mass.

**Lemma 26:** Let \( p_0, p_1 \in \mathcal{P}(X) \) and let \( \epsilon \geq 0 \). Then for \( T \geq 1 \),

\[
\begin{align*}
\inf_{q \in \mathcal{P}(X): q \preceq T p_0} D_\epsilon(q, p_1) &= \inf_{q \in \mathcal{P}(X): q \preceq T p_0} W_\infty(q, q'), W_\infty(p_1, p_1') \leq \epsilon \inf_{q' \in \mathcal{P}(X): q' \preceq T p_0'} D_{TV}(q', p_1') \\
&= \inf_{A \in \mathcal{B}(X)} \inf_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0', p_1').
\end{align*}
\]

The proof of Lemma 26 is contained in Appendix F-B.

Now, we prove the minimax equality for unequal priors.

**Theorem 27 (Minimax theorem for unequal priors):** Let \((X, d)\) have the midpoint property. Let \( p_0, p_1 \in \mathcal{P}(X) \) and let \( \epsilon \geq 0 \). Consider the zero-sum game with the payoff function \( r \) as in (19) with \( T > 0 \). Then

\[
\begin{align*}
\sup_{A \in \mathcal{B}(X)} \inf_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0', p_1') &= \inf_{A \in \mathcal{B}(X)} \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0', p_1').
\end{align*}
\]

**Proof:** Without loss of generality, we assume \( T \geq 1 \).

(If \( T < 1 \), we simply repeat the proof with labels 0 and 1 swapped.) We have the following.

\[
\begin{align*}
&\sup_{A \in \mathcal{B}(X)} \inf_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0', p_1') \\
&= \inf_{A \in \mathcal{B}(X)} R_{\epsilon}(\ell_0/1, A) \\
&= \inf_{A \in \mathcal{B}(X)} R_{\epsilon}(\ell_0 / 1, A) \\
&= \frac{1}{T + 1} \left[ 1 - \inf_{q \in \mathcal{P}(X): q \preceq T p_0} D_\epsilon(p_0, p_1) \right] \\
&= \frac{1}{T + 1} \left[ 1 - \inf_{W_\infty(p_0, p_0') \leq \epsilon} D_{TV}(q, p_1') \right] \\
&= \sup_{W_\infty(p_0, p_0') \leq \epsilon} \frac{1}{T + 1} \left[ 1 - \inf_{q \in \mathcal{P}(X): q \preceq T p_0} D_{TV}(q, p_1') \right] \\
&= \sup_{W_\infty(p_0, p_0') \leq \epsilon} \inf_{A \in \mathcal{B}(X)} r(A, p_0', p_1'),
\end{align*}
\]

where (i) follows from Theorem 12, (ii) from Theorem 22, (iii) from Lemma 26 and (iv) follows again from Theorem 22 with \( \epsilon = 0 \).\(\square\)

**Remark:** Unlike Theorem 23, Theorems 25 and 27 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 27 does not do so. Nevertheless, a \( \delta \)-approximate pure Nash equilibrium exists in all the cases. This is in sharp contrast with the non-existence of Nash equilibrium proven in [18]. The result of [18] is valid for a “regularized” adversary, where the point-wise budget constraint \( d(x, x') \leq \epsilon \) 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 [6] and [21], without the need for a regularization term.

**Remark:** A recent work [20] shows the existence of mixed Nash equilibrium for randomized classifiers parametrized by points in a Polish space. Other works [18], [29] 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 (Lemma 16) and the generalization of Strassen’s theorem (Theorem 20) are both novel to the best of our knowledge.

**Example (Continued):** We continue our example with \( p_0 = (1 - \alpha_0)\delta_{-1} + \alpha_0 \delta_0 \) and \( p_1 = (1 - \alpha_1)\delta_1 + \alpha_1 \delta_0 \). We compute the terms appearing in various equivalent formulations of the optimal adversarial risk depicted in Figure 1.

\[
\begin{align*}
A &\quad D_\epsilon(p_0, p_1) = 1 - \min\{\alpha_0, \alpha_1\}, \\
B &\quad \text{The infimum over } p_0', p_1' \text{ in } B \text{ is attained for } p_0' = (1 - \alpha_0)\delta_{-1} + \alpha_0 \delta_0 \text{ and } p_1' = (1 - \alpha_1)\delta_1 + \alpha_1 \delta_0. \quad D_{TV}(p_0', p_1') = 1 - \min\{\alpha_0, \alpha_1\} = D_\epsilon(p_0, p_1). \quad \text{Also, } \quad W_\infty(p_0, p_0') = W_\infty(p_1, p_1') = \epsilon. \\
C &\quad \text{As shown before, the infimum over } q \text{ is attained for } q^* = (1 - \alpha_q^*)\delta_{-1} + \alpha_q^* \delta_{-1} \epsilon \text{ where } \alpha_q^* = \min\{T \alpha_0, \alpha_1\}. \quad D_\epsilon(q^*, p_1') = \min\{T \alpha_0, \alpha_1\}. \\
D &\quad \text{The infimum over } q' \text{ is attained for the same } q^* \text{ as in the case of } C. \quad \text{The infimum over } q' \text{ is attained for } q'^* = (1 - \alpha_q')\delta_{-1} + \alpha_q' \delta_0, \text{ where } \alpha_q' \text{ is defined in } C. \quad \text{Moreover, } \quad W_\infty(q^*, q'^*) = \epsilon. \quad \text{The infimum over } p_1' \text{ is attained for } p_1^* \text{ as defined in } B. \\
E &\quad \text{The infimum over } p_0', p_1' \text{ is attained for } p_0', p_1' \text{ as defined in } B. \quad \text{The infimum over } q' \text{ is attained for } q'^* \text{ as defined in } D.
\end{align*}
\]

**VIII. DISCUSSION**

We examined different notions of adversarial risk and laid down the conditions under which these definitions are equivalent. By verifying the conditions in Sections IV and V, researchers may use different definitions interchangeably. In the following, we We largely focused on the binary classification setup with 0-1 loss function. While we extended our results on measurability and relation to \( \infty \)-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, \( \infty \)-Wasserstein distributional perturbations). An interesting open question is whether these results hold for more general perturbation models. A starting point for approaching this question is to consider the Choquet capacities.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
in $\mathbb{R}^d$. In Lemma 16 we show a key property that the measure of $\epsilon$-Minkowski expansion is a 2-alternating capacity and then exploit this connection to establish equivalences between adversarial robustness and $W_\infty$ distributional robustness, and later on, the existence of a Nash equilibrium in the game between adversary and algorithm. One can in fact generalize this lemma further as shown below.

**Lemma 28:** Let $(\mathcal{X}, d) = (\mathbb{R}^d, \| \cdot \|)$. Let $\mu \in \mathcal{P}(\mathcal{X})$ and let $U \in \mathcal{B}(\mathcal{X})$ be a compact set. Define a set function $v$ on $\mathcal{X}$ such that for any $A \in \mathcal{L}(\mathcal{X})$, $v(A) := \mu(A \ominus U)$. Then $v$ is a 2-alternating capacity.

Similarly, we have the following generalization of Lemma 17

**Lemma 29:** Let $(\mathcal{X}, d) = (\mathbb{R}^d, \| \cdot \|)$. Let $\mu \in \mathcal{P}(\mathcal{X})$. Define $\mathcal{U}(\mu) := \{ \mu' \in \mathcal{P}(\mathcal{X}) : \exists \gamma \in \Pi(\mu, \mu') \text{ with } \mathcal{P}(x, x') \sim_{\gamma} [x' \in x + U] = 1 \}$. Then for any $A \in \mathcal{L}(\mathcal{X})$, $\sup_{\mu' \in \mathcal{U}(\mu)} \mu'(A) = \mu(A^{\oplus \epsilon})$. Moreover, the supremum in the previous equation is attained.

Lemma 29 facilitates the study of a much broader notion of adversarial risk wherein, rather than perturbing $x$ to $x' \in B_\epsilon(x)$, one may consider the perturbation $x' \in x + U$, where $U$ represents an arbitrary closed uncertainty set, a specific instance of which is $B_\epsilon(0)$. A general perturbation model such as this has been considered in prior works before (see for example [23]), but an equivalent formulation in terms of a distributional perturbation does not appear in the literature. Lemma 29 shows that a data perturbation of the form $x' \in x + U$, corresponds to a distributional perturbation of the form $\mu' \in \mathcal{U}(\mu)$. Perhaps, it is possible to exploit this connection to extend some of the findings of sections V, VI and VII to this general perturbation model. We leave this for a future work.

We analyzed optimal adversarial risk for (non-parametric) decision region-based classifiers. Using a formulation of optimal transport between finite measures of unequal mass, we extended the optimal transport based characterization of adversarial risk of [22] and [23] to unequal priors by generalizing Strassen’s theorem. This may find applications in the study of excess cost optimal transport [49], [50]. A recent work [51] 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 $A^\ast$ as $\epsilon$ grows, in terms of a mean curvature flow. This raises an interesting question on the evolution of the optimal adversarial distributions $p_0^\ast, p_1^\ast \in \mathcal{P}(\mathcal{X})$ with $\epsilon$.

We proved a minimax theorem for adversarial robustness game and the existence of a Nash equilibrium. We constructed the worst-case pair of distributions $p_0^\ast, p_1^\ast \in \mathcal{P}(\mathcal{X})$ 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. Specifically, the case of equal priors in $\mathbb{R}^d$ admits a simple construction of $p_0^\ast, p_1^\ast$, as shown in the proof of Theorem 24. We restate it below for convenience.
Corollary 30: Let \((\mathcal{X}, d) = (\mathbb{R}^d, \| \cdot \|)\). Let \(p_0, p_1 \in \mathcal{P}(\mathcal{X})\) and let \(\epsilon \geq 0\). Let \(\gamma \in \Pi(p_0, p_1)\) be the coupling that achieves the optimal transport cost for \(D_\epsilon(p_0, p_1)\). Consider a “mid-point” transport map \(t : \mathcal{X}^2 \to \mathcal{X}\) defined as follows.

\[
t(x, x') = \begin{cases} \frac{(x + x')}{2}, & \|x - x'\| \leq 2\epsilon, \\ (x, x'), & \text{otherwise.} \end{cases}
\]

Then the optimal classifier \(A^* \in \mathcal{B}(\mathcal{X})\) minimizing the adversarial risk \(R_{\epsilon, A}(\ell_{0/1}, A)\) for the case of equal priors (if one exists), is the Bayes classifier for the worst-case pair of distributions \(p_0^*, p_1^* \in \mathcal{P}(\mathcal{X})\) such that \(\gamma = \Pi(p_0^*, p_1^*)\), i.e. \(p_0^*\) and \(p_1^*\) are the first and second marginals of \(\gamma\), respectively. We note that the worst-case distributions \(p_0^*\) and \(p_1^*\) may not admit densities even when \(p_0\) and \(p_1\) do. Moreover, estimating \(p_0^*\) and \(p_1^*\) may be a harder task estimating the optimal classifier \(A^*\). However, approaches based on obtaining an adversarially robust classifier by first learning to generate adversarial examples (which is akin to sampling from \(p_0^*\) and \(p_1^*\)) have been successful to some degree in practice [52], [53], [54], [55].

**APPENDIX A**

**PRELIMINARY LEMMAS**

**Lemma 31:** Let \(A_n \in \mathcal{B}(\mathcal{X})\) for \(n \in \{1, 2, \ldots\}\). Then,

\[
(\bigcap_n A_n)^{\oplus \epsilon} = \bigcup_n A_n^{\oplus \epsilon},
\]

\[
(\bigcup_n A_n)^{\oplus \epsilon} \subseteq \bigcap_n A_n^{\oplus \epsilon}.
\]

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

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

**Lemma 32:** Let \(F_n\) be a sequence of closed sets in \(\mathcal{X}\) such that \(F_k \supseteq F_{k+1}\) for \(k \in \mathbb{N}\). Then,

\[
(\bigcap_n F_n)^{\oplus \epsilon} = \bigcap_n F_n^{\oplus \epsilon}.
\]

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

Let \(x \in \bigcap_n F_n^{\oplus \epsilon}\). Then \(x \in F_n^{\oplus \epsilon}\) for all \(n \in \mathbb{N}\). Hence, there exists \(x_n \in F_n\) such that \(d(x, x_n) \leq \epsilon\) 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 \(x^* \in F := \bigcap_n F_n\). Indeed, for any \(m \in \mathbb{N}\), \(x^* \in F_n\) for all \(n \in \mathbb{N}\). Hence, \(x \in \bigcap_n F_n^{\oplus \epsilon}\) and therefore \(\bigcap_n F_n^{\oplus \epsilon} \subseteq \bigcap_n F_n^{\oplus \epsilon}\). We will now show the set inclusion in the opposite direction.

Let \(x \in \bigcap_n F_n^{\oplus \epsilon}\). Then \(x \in F_n^{\oplus \epsilon}\) for all \(n \in \mathbb{N}\). Hence, there exists \(x_n \in F_n\) such that \(d(x, x_n) \leq \epsilon\) 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 \(x^* \in F := \bigcap_n F_n\). Indeed, for any \(m \in \mathbb{N}\), \(x^* \in F_n\) for all \(n \in \mathbb{N}\). Hence, \(x \in \bigcap_n F_n^{\oplus \epsilon}\) and therefore \(\bigcap_n F_n^{\oplus \epsilon} \subseteq \bigcap_n F_n^{\oplus \epsilon}\). We will now show the set inclusion in the opposite direction.

Let \(x \in \bigcap_n F_n^{\oplus \epsilon}\). Then \(x \in F_n^{\oplus \epsilon}\) for all \(n \in \mathbb{N}\). Hence, there exists \(x_n \in F_n\) such that \(d(x, x_n) \leq \epsilon\) 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 \(x^* \in F := \bigcap_n F_n\). Indeed, for any \(m \in \mathbb{N}\), \(x^* \in F_n\) for all \(n \in \mathbb{N}\). Hence, \(x \in \bigcap_n F_n^{\oplus \epsilon}\) and therefore \(\bigcap_n F_n^{\oplus \epsilon} \subseteq \bigcap_n F_n^{\oplus \epsilon}\).
\[ + \sup_{f_0, f_1 \in F_\epsilon} \left( \frac{T}{T + 1} f_{0p_0}(B) + \frac{1}{T + 1} f_{1p_1}(B^c) \right) \geq \sup_{f_0, f_1 \in F_\epsilon} \left[ \left( \frac{T}{T + 1} f_{0p_0}(A \cup B) + \frac{1}{T + 1} f_{1p_1}(A^c \cup B^c) \right) \right] \]

For the upper bound, we have the following.

\[ R_{F_\epsilon}(\ell_{0/1}, A) + R_{F_\epsilon}(\ell_{0/1}, B) \]

\[ = \sup_{f_0, f_1 \in F_\epsilon} \left( \frac{T}{T + 1} f_{0p_0}(A) + \frac{1}{T + 1} f_{1p_1}(A^c) \right) \geq \sup_{f_0, f_1 \in F_\epsilon} \left( \frac{T}{T + 1} f_{0p_0}(A \cup B) + \frac{1}{T + 1} f_{1p_1}(A^c \cup B^c) \right) \geq R_{F_\epsilon}(\ell_{0/1}, A \mid B), \]

The bounds for \( R_{F_\epsilon}(\ell_{0/1}, A) \) can be proved by following a similar outline as that for \( R_{F_\epsilon}(\ell_{0/1}, A \mid B) \) above.

Suppose \( A, B \) are such that \( R_{F_\epsilon}(\ell_{0/1}, A \mid B), R_{F_\epsilon}(\ell_{0/1}, A) \) and \( R_{F_\epsilon}(\ell_{0/1}, B) \) are well-defined. The following gives the desired lower bound on \( R_{F_\epsilon}(\ell_{0/1}, A) \).

\[ R_{F_\epsilon}(\ell_{0/1}, A) \]

where (a) follows from Lemma 31 and (b) follows from the monotonicity property of the Minkowski set expansion. Similarly, the following gives the desired upper bound on \( R_{F_\epsilon}(\ell_{0/1}, A) \).

\[ R_{\bar{F}_\epsilon}(\ell_{0/1}, A) + R_{\bar{F}_\epsilon}(\ell_{0/1}, B) \]

\[ = \frac{T}{T + 1} p_0(A \bar{\epsilon}^c) + \frac{1}{T + 1} p_1((A \bar{\epsilon}^c) \bar{\epsilon}^c) + \frac{T}{T + 1} p_0(B \bar{\epsilon}^c) + \frac{1}{T + 1} p_1((B \bar{\epsilon}^c) \bar{\epsilon}^c) \geq \frac{T}{T + 1} p_0((A \cup B) \bar{\epsilon}^c) + \frac{1}{T + 1} p_1((A \cup B) \bar{\epsilon}^c) \geq R_{\bar{F}_\epsilon}(\ell_{0/1}, A \mid B), \]

where (c) follows from the fact that \( A \Delta B \subseteq A \cup B \) and \( A \Delta B \subseteq A^c \cup B^c \).

The bounds for \( R_{F_\epsilon}(\ell_{0/1}, A \mid B) \) can be proved by following the same outline as that for \( R_{F_\epsilon}(\ell_{0/1}, A \mid B) \), using the monotonicity of the closed set expansion and the fact that \( (A_1 \cup A_2)^c \subseteq A_1^c \cup A_2^c \) for any \( A_1, A_2 \in B(X) \). \( \square \)

## Appendix C

**Proofs from Section IV**

### A. Proofs From Section IV-A

**Proof of Lemma 2**: We prove the above statement by using a counterexample motivated from Example 2.4 in [56].

For any \( \epsilon > 0 \), there exists a Borel measurable set \( S \subseteq [-\epsilon, \epsilon]^2 \) such that its projection onto the first coordinate is not Borel measurable ([56], Theorem 6.7.2 and Theorem 6.7.11 in [57]). 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 \( \phi : \mathbb{R}^3 \rightarrow \mathbb{R}^3 \) as \( \phi(x_1, x_2, x_3) := (x_1, x_2, \sqrt{\epsilon^2 - x_1^2}) \). \( \phi \) maps the plane \([0, \epsilon)^2 \times \{ 0 \}\) onto the half-cylinder, \( \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 \in [-\epsilon, \epsilon], x_2^2 + x_3^2 = \epsilon^2, x_3 \geq 0 \} \), of radius \( \epsilon \). Let \( A := \phi(S \times \{ 0 \}) \). Then \( A \in \mathcal{B}(\mathbb{R}^3) \) because \( S \times \{ 0 \} \in \mathcal{B}(\mathbb{R}^2) \). We have the following equality.

\[ A \bar{\epsilon}^c \cap (\mathbb{R} \times \{ 0 \}^2) = S_1 \times \{ 0 \}^2 \]

Suppose \( A \bar{\epsilon}^c \in \mathcal{B}(\mathbb{R}^3) \). Then the above equality implies that \( S_1 \in \mathcal{B}(\mathbb{R}) \) contradicting our choice of \( S \). Hence, \( A \bar{\epsilon}^c \notin \mathcal{B}(\mathbb{R}^3) \). \( \square \)

**Proof of Lemma 3**: 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 [31].

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

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

\[ \{(x, x') \in \mathcal{X}^2 : f(x, x') < c\} \]
implies that $d \in A$. By Definition 7.21 in [31], it follows that

$$Y \rightarrow f \subseteq H$$

Hence, by Definition 7.21 in [31], we have

$$f^*(x) = \inf_{x' \in B_r(x)} -1 \{x' \in A\} = -\sup_{x' \in B_r(x)} \{x' \in A\} = -1 \{x \in A^{\infty}e\}.$$

By Definition 7.21 in [31], it follows that $A^{\infty}e$ is an analytic set. By Corollary 7.42.1 in [31], $A^{\infty}e \subseteq B(X)$.

**Proof of Lemma 5:** Let $\beta = 1/4$. Take any $e \in E$. Since $E = A^e \setminus A^2$, we have the following two implications: 1) $E \subseteq A^\ast$ which implies that $d(e, A) \leq e$, and 2) $E \cap A^2 = \emptyset$ which implies that $d(e, A) > e$. Combining the two implications, we get that $d(e, A) = e$. Hence, for every $r \in (0, e]$, there must exist an $a_r \in A$ such that $e \leq ||e - a_r|| < e + r/4$. We pick an $x' \in X$ on the line segment joining $a_r$ and $x$ as follows.

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

Since $||e - a_r|| \in [e, e + r/4)$ and $r \in (0, e]$, 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 $B_{\beta^2}(x') \subseteq B_r(e)\setminus E$.

For any $y \in B_{\beta^2}(x')$, we have the following.

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

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

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

Hence, $y \in A^2$ and so $y \notin E$. Therefore, $B_{\beta^2}(x') \subseteq B_r(e)\setminus E$. Hence, we have the following property (call it (*)): For any $e \in E$ and any $r \in (0, e]$, there is an $x' \in X$ such that $B_{\beta^2}(x') \subseteq B_r(e)\setminus E$. The property (*) is depicted in Figure 2.

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

Suppose $x \in E$. Then by the property (*), there exists $x' \in X$ such that $B_{\alpha r}(x') \subseteq B_{\beta^2}(x') \subseteq B_r(x)\setminus E$. Suppose on the other hand $x \notin E$. If $B_{\beta^2}(x) \cap E = \emptyset$, then choosing $x' = x$ we have $B_{\alpha r}(x') \subseteq B_{\beta^2}(x') \subseteq B_r(x)\setminus E$. If not, then there exists $e \in B_{\beta^2}(x) \cap E$. We claim that $B_{(1 - \beta)r}(e) \subseteq B_{\beta^2}(x)$. 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 $(1 - \beta)r \in (0, e]$, by the property (*), there exists $x' \in X$ such that $B_{\alpha r}(x') = B_{(1 - \beta)r}(x') \subseteq B_{(1 - \beta)r}(x)\setminus E \subseteq B_r(x)\setminus E$.

**Proof of Lemma 11:** Let $\mu' \in \mathcal{P}(X)$ be such that $W_\infty(\mu, \mu') \leq e$. Then there exists a coupling $\lambda \in \Pi(\mu', \mu)$ such that for $(x, x') \sim \lambda$, $d(x, x') \leq e \lambda a.e.$ Hence, 

$$\mu'(A) = \lambda(A \times X) = \lambda(A \times A^{\infty}e) \leq \lambda(X \times A^{\infty}e) = \mu(A^{\infty}e).$$

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

$$\sup_{W_\infty(\mu, \nu) \leq e} \mu'(A) \leq \mu(A^{\infty}e).$$
Now we show the inequality in the opposite direction. Like in the proof of Lemma 3, consider the function \( f : D \to \mathbb{R} \) defined as \( f(x, x') = -1 \{ x' \in A \} \), where \( D = \{ (x, x') \in \mathcal{X}^2 : d(x, x') \leq \epsilon \} \). Define \( f^* : \mathcal{X} \to \mathbb{R} \) as \( f^*(x) = \inf_{x' \in B, (x)} f(x, x') \). As shown in the proof of Lemma 3, \( f^*(x) = -1 \{ x \in A^\oplus \} \). By Proposition 7.50(a) in [31], there exists a measurable function \( \phi : \mathcal{X} \to \mathcal{X} \) such that \( |f^*(x) - f(x, \phi(x))| \leq \delta \) for any \( \delta > 0 \). Since \( f \) and \( f^* \) are both 0-1 valued functions, we get \( f^*(x) = f(x, \phi(x)) \) for all \( x \in \mathcal{X} \) by choosing \( \delta = 1/2 \). Moreover, by Proposition 7.50(a) in [31], \( Gr(\phi) \subseteq D \) i.e., \( d(x, \phi(x)) \leq \epsilon \) for all \( x \in \mathcal{X} \). Therefore,

\[
\sup_{W_{\epsilon}(\mu, \mu') \leq \epsilon} \mu'(A) \geq \phi_{2}(A) = \mu(\phi^{-1}(A)) = \mu(A^\oplus).
\]

Hence, \( \sup_{W_{\epsilon}(\mu, \mu') \leq \epsilon} \mu'(A) = \phi_{1}(A) = \mu(A) \) for any set \( A \in \mathcal{B}(\mathcal{X}) \).

**Proof of Lemma 14:** Let \( \mu' \in \mathcal{P}(\mathcal{X}) \) be such that \( W_{\epsilon}(\mu, \mu') \leq \epsilon \). Then there exists \( \lambda \in \Pi(\mu, \mu') \) such that \( \lambda(\{(x, x') \in \mathcal{X}^2 : d(x, x') > \epsilon \}) = 0 \). Then,

\[
E_{x \sim \mu}[\phi(x)] = E_{x \sim \mu}[\lambda(\phi(x))]
\]

\[
= E_{x \sim \mu}\left[ \sup_{x \in B, (x)} \phi(x) \right]
\]

\[
= E_{x \sim \mu}[\phi(x)].
\]

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

\[
\sup_{W_{\epsilon}(\mu, \mu') \leq \epsilon} E_{x \sim \mu}[\phi(x)] \leq \sup_{W_{\epsilon}(\mu, \mu') \leq \epsilon} \left[ \sup_{x \in B, (x)} \phi(x) \right].
\]

Now we will show the inequality in the opposite direction. Consider the function \( f : D \to \mathbb{R} \) defined as \( f(x, x') = -\phi(x') \), where \( D = \{ (x, x') \in \mathcal{X}^2 : d(x, x') \leq \epsilon \} \). Define \( f^* : \mathcal{X} \to \mathbb{R} \) as \( f^*(x) = \inf_{x' \in B, (x)} f(x, x') = -\sup_{x' \in B, (x)} \phi(x') \). Choose a \( \delta > 0 \). By Proposition 7.50(a) in [31], 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 \( d(x, m_{\delta}(x)) \leq \epsilon \) for all \( x \in \mathcal{X} \). Hence,

\[
E_{x \sim \mu}\left[ \sup_{d(x, x') \leq \epsilon} \phi(x') \right] = E_{x \sim \mu}[\phi(x)]
\]

\[
\leq E_{x \sim \mu}[\phi(m_{\delta}(x))] + \delta
\]

\[
= E_{x \sim \mu}[\phi(x)] + \delta
\]

\[
\leq \sup_{W_{\epsilon}(\mu, \mu') \leq \epsilon} E_{x \sim \mu}[\phi(x)] + \delta,
\]

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

\[
E_{x \sim \mu}\left[ \sup_{d(x, x') \leq \epsilon} \phi(x') \right] \leq \sup_{W_{\epsilon}(\mu, \mu') \leq \epsilon} E_{x \sim \mu}[\phi(x)].
\]

Combining the above inequality with the reverse inequality shown previously, we obtain (15).

Suppose the function \( \phi \) is upper semi-continuous. Then \( f \) is lower semi-continuous. Hence, for every \( x \in \mathcal{X} \), there exists \( x^* \) in the compact set \( B_{\epsilon}(x) \) such that \( \inf_{x' \in B_{\epsilon}(x)} f(x, x') = f(x, x^*) \). 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

\[
\sup_{W_{\epsilon}(\mu, \mu') \leq \epsilon} E_{x \sim \mu}[\phi(x)] = E_{x \sim \mu}\left[ \sup_{d(x, x') \leq \epsilon} \phi(x') \right]
\]

\[
= E_{x \sim \mu}[\phi(m(x))]
\]

\[
= E_{x \sim \mu}[\phi(x)].
\]

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

**B. Proofs From Section V-B**

**Proof of Lemma 16:** The following properties of \( v \) are trivially true: \( v(\phi) = 0 \), \( v(\mathcal{X}) = 1 \) 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 \( A = \bigcup_{n} A_n \). That is, \( A_n \uparrow A \). Then by Lemma 31 we have, \( A^\oplus = \bigcup_{n} A_n^\oplus \). Hence, \( A^\oplus \uparrow A^\oplus \) and by the continuity of measure, \( v(A_n) = \mu(A^\oplus) \) and \( v(A_t) = v(A) \).

Consider a sequence of closed sets \( (F_n) \) in \( \mathcal{X} \) such that \( F_k \subseteq F_{k+1} \) for \( k \in \mathcal{N} \). For \( F = \bigcap_{n} F_n \), that is, \( F_n \uparrow F \). By Lemma 32, \( F_n^\oplus \subseteq F^\oplus \). Hence, by the continuity of measure, we have \( v(F_n) = \mu(F_n^\oplus) \subseteq v(F) \).

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

\[
v(A \cup B) = \mu((A \cup B)^\oplus)
\]

\[
\sup_{W_{\epsilon}(\mu, \nu) \leq \epsilon} \mu'(A) \leq \mu(A^\oplus).
\]

We will now show the inequality in the reverse direction. By Lemma 16, \( A \mapsto \mu(A^\oplus) \) is a 2-alternating capacity. Hence by Lemma 2.5 in [34], for any Lebesgue measurable \( A \subseteq \mathcal{X} \), there exists a \( \nu \in \mathcal{P}(\mathcal{X}) \) such that \( v(A) = \mu(A^\oplus) \) and \( v(B) \leq \mu(B^\oplus) \) for all Lebesgue measurable \( B \subseteq \mathcal{X} \). For such a \( \nu \), it is clear that \( W_{\epsilon}(\mu, \nu) \leq \epsilon \). Hence,

\[
\sup_{W_{\epsilon}(\mu, \nu) \leq \epsilon} \mu'(A) \geq v(A) = \mu(A^\oplus).
\]
Hence, $\sup_{\rho, \mu} \rho \leq \rho(A) = \mu(A^{\text{dir}})$.

\section*{Appendix E
Proofs From Section VI-A

We first prove a discrete version of Theorem 20 on a finite space.

\textbf{Lemma 37:} Let $\mathcal{X}_n = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$. Let $p = (p_1, \ldots, p_n)$ be such that $p_i, q_j \geq 0$ for $i \in [n]$ and $\sum_i p_i \leq \sum_j q_j$. Let $\epsilon > 0$. For $A \subseteq \mathcal{X}_n$, let $A^* := \{x \in \mathcal{X}_n : d(x, x') \leq \epsilon$, for some $x' \in A\}$. For $A \subseteq \mathcal{X}_n$, let $p(A) = \sum_{x \in A} p_i$ and $q(A) = \sum_{x \in A} q_i$. For $i, j \in [n]$, let $c_{ij} = \mathbb{I}\{d(x_i, x_j) > 2\epsilon\}$. Then,

$$\max_{A \subseteq \mathcal{X}_n} p(A) - q(A^{\text{dir}}) = \min_{\sum_{x_i \in p_i}, \sum_{x_j \leq q_j}} \sum c_{ij} x_{ij}. \tag{25}$$

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

$$\min_{\sum_{x_i \leq p_i}, \sum_{x_j \leq q_j}} \sum c_{ij} x_{ij} = \max_{\sum_{x_i \geq p_i}, \sum_{x_j \leq q_j}} \sum_{i,j} d_{ij} x_{ij}. \tag{26}$$

Consider the following modification to the linear program on the right hand side of (26), where the constraint $\sum_j x_{ij} = p_i$ is replaced by $\sum_j x_{ij} \leq p_i$.

$$\max_{\sum_{x_i \leq p_i}, \sum_{x_j \leq q_j}} \sum d_{ij} x_{ij}. \tag{27}$$

We will show that the above linear program is equivalent to the linear program on the right hand side of (26). Since the above linear program is bounded and feasible, it admits a solution. Let $x_{ij}^*_{i,j \in [n]}$ be the solution to (27). Suppose there exists $m \in [n]$ such that $\sum_j x_{ij}^* < p_m$. Let $s = p_m - \sum_j x_{ij}^*$.

$$\sum_j s_j = \sum_j q_j - \sum_j x_{ij} \geq \sum_i p_i - \left(\left(\sum_{i \neq m} p_i\right) + p_m - s\right) = s. \tag{28}$$

Therefore, $\sum_j s_j \geq s$. Let $k$ be the largest integer for which $\sum_{j=1}^k s_j < s$. Define,$$y_{ij} = \begin{cases} x_{ij}^* & i \neq m, \\ x_{mj}^* + s_j & i = m, j \leq k, \\ x_{mk}^* + s - \sum_{j=1}^k s_j & i = m, j = k + 1, \\ x_{mj}^* & i = m, j \geq k + 1. \end{cases} \tag{29}$$

By the above definition we have,$$
\sum_i y_{ij} = \begin{cases} \sum_i x_{ij}^* & i \neq m, \\ \sum_i x_{ij}^* + s & i = m. \end{cases} \tag{30}$$

Combining the above with the definitions of $k, s$ and $s_j$, 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, $\sum_i d_{ij} y_{ij} \geq \sum_i d_{ij} x_{ij}$, which is the required bound.

The optimal $w^*_i, v^*_i$ that achieve the maximum in (31) must lie at one of the vertices of the polyhedron supported by the hyperplanes, $w_i = 0, w_{i+1} = 0, v_{i+1} = 0, v_i = 1$ and $w_{i+1} - v_i = c_{ij}$.

Hence, $w^*_i, v^*_i \in \{0, 1\}$. Moreover if $c_{ij} = 0$ and $w^*_i = 1$ for some $i,j \in [n]$ then $v^*_i = 1$. On the other hand if $c_{ij} = 1$, then $v^*_i$ 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 (31) equals the maximum in (25).

\textbf{Proof of Theorem 20:} Let $x_1, \ldots, x_n$ be a non-negative, monotonically decreasing sequence converging to 0. Let $\mathcal{N}_n \subseteq \mathcal{X}_n$ be a dense sequence in $\mathcal{X}$. Define a function $f : \mathcal{X} \rightarrow [x_1, \ldots, x_n]$ 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,$$
\mu(f^{-1}(H_{s_n-1})) > \mu(\mathcal{X} - \gamma_n), \tag{32}$$
$$\nu(f^{-1}(H_{s_n-1})) > \nu(\mathcal{X} - \gamma_n). \tag{33}$$

Let $\mu_n$ be a discrete measure supported on $H_{s_n}$ such that $\mu_n(x_k) := \mu(f^{-1}(x_k))$ for $k \in [s_n - 1]$ and $\mu_n(x_1) = \mu(x_1)$. Similarly, construct $\nu_n$ supported on $H_{s_n}$ such that $\nu_n(x_k) := \nu(f^{-1}(x_k))$ for $k \in [s_n - 1]$ and $\nu_n(x_1) = \nu(x_1)$.

Let $A \in \mathcal{B}(\mathcal{X})$. We have,$$
\mu_n(A) \leq \mu_n(A \cap H_{s_n}) \tag{i}\leq \mu_n(A \cap H_{s_n}) + \gamma_n \tag{ii} \leq \mu_n(A \cap H_{s_n+1}) + \gamma_n \tag{iii} \leq \mu(f^{-1}(A \cap H_{s_n+1})) + \gamma_n \tag{iv}.$$
where (i) follows from the fact that $\mu_n$ is supported on $H_{s_n}$, (ii) follows from (32), (iii) follows from the definition of $\mu_n$ and (iv) follows because of the following: For any $y \in A \cap H_{s_n-1}$, $f^{-1}(y) \subseteq \{ x \in X : d(x, y) < \gamma_n \} \subseteq A^{2\gamma_n}$. Hence, $f^{-1}(A \cap H_{s_n-1}) \subseteq A^{2\gamma_n}$. Applying (34), with $A^{\gamma}$ instead of $A$, we have the following.

$$\mu(A^{\gamma_n}) - \gamma_n \leq \mu_n(A) \leq \mu(A^{\gamma_n}) + \gamma_n. \quad (35)$$

Letting $n \to \infty$ in (35) and using Lemma 33, we get that $\limsup_n \mu_n(A) \leq \mu(A)$ for all closed subsets $A$ of $X$. Hence, by applying the Portmanteau theorem (Theorem 2.1 in [59]), we conclude that the sequence of measures $(\mu_n)_{n=1}^{\infty}$ converges weakly to $\mu$. Similarly, $\nu_n \to \nu$ weakly.

For any fixed $n$, we apply Lemma 37 to the measures $\mu_n, \nu_n$ on the finite space $H_{s_n}$ to get the following.

$$\max_{A \subseteq H_{s_n}} \mu_n(A) - \nu_n(A^{2\gamma_n+4\gamma_n}) = \min_{\sum_{ij} x_{ij} \geq \mu_n(x_{ij})} \sum_{i,j} x_{ij} \mathbb{I}\{d(x_i, x_j) > 2\epsilon + 4\gamma_n\}, \quad (36)$$

where the indices $i, j$ run over $[s_n]$. We have that $\mu_n(A) = \mu(A) \leq \nu(A) \leq \nu(A)$. 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 (36) by setting $\pi_n(i,j) = x_{ij}$. Let $T_n \subseteq H_{s_n}$ be the set that achieves the maximum in (36).

We will now construct a candidate coupling for the infimum in (17). Since $\mu, \nu$ are finite measures on a Polish space, they are tight (see for example, Theorem 1.3 in [59]). Hence, given a $\delta > 0$, there exists a compact set $K \subseteq X$ such that $\min\{\mu(K^c), \nu(K^c)\} < \delta/3$. Since $\mu_n$ and $\nu_n$ converge weakly to $\mu$ and $\nu$ respectively, choose $N$ large enough so that $\min\{\mu_n(K^c), \nu_n(K^c)\} < \delta/2$ for all $n \geq N$. Let $\nu' \subseteq K$ be the second marginal of the coupling $\pi_n$. Then, $\nu_n \leq \nu'$. By union bound, we have the following.

$$\pi_n(K \times K^c) \leq \mu_n(K^c) + \nu'(K^c) \leq \mu_n(K^c) + \nu_n(K^c) < \delta. \quad (37)$$

Hence, the sequence $(\pi_n)_{n \geq N}$ is uniformly tight. Hence, by Prokhorov’s theorem (for reference, see Theorem 5.1 in [59]), there is a subsequence $(\pi_{n_k})_k$ of $(\pi_n)_{n \geq N}$ that converges weakly to some measure $\pi^* \in \mathcal{M}(X \times X)$. Moreover, $\pi^* \in \Pi(\mu, \nu)$ by virtue of the constraints imposed on the converging subsequence of $(\pi_n)_{n \geq N}$.

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

$$\pi_n(d(x_i, x_j) > 2\epsilon + 4\gamma_n) \leq \mu_n(T_n) - \nu_n(T_n^{2\epsilon+4\gamma_n}) \leq (\mu_n(T_n) + \gamma_n) - (\nu_n(T_n^{2\epsilon+4\gamma_n} - \gamma_n) - \gamma_n) \leq \mu_n(T_n^{2\epsilon+4\gamma_n} - \gamma_n - \gamma_n/2) + 2\gamma_n \leq \mu(T_n^{2\epsilon+4\gamma_n}) - \nu(T_n^{2\epsilon+2\gamma_n} + 2\gamma_n) \leq \Phi + 2\gamma_n, \quad (38)$$

where (i) follows from the definition of $\pi_n$ and $T_n$, (ii) follows from (35), (iii) follows from Lemma 34 and (iv) follows from the definition of $\Phi$. Further,

$$\Psi = \inf_{\pi \in \Pi(\mu, \nu)} \pi(d(x, x') > 2\epsilon) \leq \pi^*[d(x, x') > 2\epsilon] \leq \liminf_{n_k} \pi_n[d(x, x') > 2\epsilon] \leq \limsup_n \pi_n[d(x, x') > 2\epsilon] \leq \Phi, \quad (39)$$

where (i) follows because $\pi^* \in \Pi(\mu, \nu)$, (ii) follows from Portmanteau’s theorem because $(\pi_{n_k})_k$ that converges to $\pi^*$ and the set $\{ (x, x') : d(x, x') > 2\epsilon \}$ is an open set, and (iii) follows by taking $n \to \infty$ in (38).

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 $\liminf_n \lambda_n[x \in A, x' \notin A'] > \epsilon = \Psi$. For any $A \in B(X)$,

$$\mu(A) = \lambda_n[x \in A, x' \notin A'] + \lambda_n[x \in A, x' \notin A' \subseteq A^{2\epsilon+4\gamma_n}] \leq \nu(A') + \lambda_n[d(x, x') > \epsilon].$$

Letting $n \to \infty$, $(\mu(A) - \nu(A')) \leq \Psi$ for all $A \in B(X)$. Hence, $\Phi \leq \Psi$. Combining this with (39), we conclude $\Phi = \Psi$. \hfill \Box

B. Proof of Section VI-B

Proof of Lemma 21: We have,

$$\sup_{A \in B(X)} \mu(A) - \nu(A^{2\epsilon}) \leq \mu(A) - \nu(A^{2\epsilon}) \leq \mu(A^{2\epsilon}) - \nu(A^{2\epsilon}) \leq \mu(A^{2\epsilon}) - \nu(A^{2\epsilon}),$$

where (i) follows because we may assume that the supremum of $\mu(A) - \nu(A^{2\epsilon})$ is achieved by a closed set. Indeed, $\mu(A) - \nu(A^{2\epsilon}) \geq \mu(A) - \nu(A^{2\epsilon}) \geq \mu(A) - \nu(A^{2\epsilon}) \geq \mu(A) - \nu(A^{2\epsilon})$ (see Lemma 3.3 in [19]), and (ii) follows from the following two facts: 1) $A \subseteq A^{2\epsilon+4\gamma}$ (see Lemma 3.3 in [19]), and 2) $A' = A^{2\epsilon}$ for closed sets $A$ (see Lemma 3.2 in [19]). (iii) follows from Lemma 3 because $\mu, \nu \in \mathcal{P}(X)$ and $A^{2\epsilon} \subseteq B(X)$ whenever $A \in B(X)$.

Now, we show that the above inequality also holds in the opposite direction. Let $\mu' = \mu/t \in \mathcal{P}(X)$ for some fixed $t > 0$. For $x, y \in X$, define the cost function $c(x, y) = 1[d(x, y) < 2\epsilon]$. For any $\nu' \in \mathcal{P}(X)$, we have the following from Kantorovich duality theorem.

$$D_t(\mu', \nu') = \sup_{\phi \in \phi(\mu' + \psi(\mu'))} \int \phi d\mu' + \int \psi d\nu',$$

For any $A \in B(X)$, define $\phi(x) = 1[x \in A^{2\epsilon}]$ and $\psi(y) = -1[y \in A^{2\epsilon}]$. We will now show that $\phi'(x) + \psi'(y) \leq c(x, y)$. If $x, y$ are such that $c(x, y) = 1$, the inequality holds trivially. Suppose on the other hand, $x, y$ are such that $c(x, y) = 0$. Then $d(x, y) \leq 2\epsilon$. Hence, for any $x \in A^{2\epsilon}$, we have $y \in (A^{2\epsilon})^{2\epsilon+4\gamma_n} = ((A^{2\epsilon})^{2\epsilon+4\gamma_n})^{2\epsilon+4\gamma_n} \subseteq A^{2\epsilon}$ (the set inclusion here follows from Lemma 3.3 in [19]). Therefore,

$$\phi'(x) + \psi'(y) = 1[x \in A^{2\epsilon}] - 1[y \in A^{2\epsilon}] = 0.$$
Hence,
\[
D_\epsilon(\mu', \nu') \geq \int \phi' d\mu' + \int \psi' d\nu' = \mu'(A^{2\epsilon}) - \nu'(A^{2\epsilon}).
\]
Now,
\[
\begin{align*}
sup_{A \in B(X)} \mu(A) - \nu(A^{2\epsilon}) &= \sup_{A \in B(X)} \mu(A) - \nu(A^{2\epsilon}) \\
&= \sup_{A \in B(X)} \sup_{\nu' \in \mathcal{P}(X)} \mu'(A^{2\epsilon}) - \nu'(A^{2\epsilon}) \\
&= \sup_{A \in B(X)} \sup_{\nu' \in \mathcal{P}(X)} \mu'(A^{2\epsilon}) - \nu'(A^{2\epsilon}) \\
&\leq \mu(A^{2\epsilon}) - \nu(A^{2\epsilon}) \\
&= \mu(A^{2\epsilon}) - \nu(A^{2\epsilon})
\end{align*}
\]
where (\ast) follows from Theorem 20. Since the above inequality is valid for any \( A \in B(X) \), we get the following.
\[
\sup_{A \in B(X)} \mu(A) - \nu(A^{2\epsilon}) \geq \sup_{A \in B(X)} \mu(A^{2\epsilon}) - \nu(A^{2\epsilon}).
\]

**APPENDIX F**

**Proofs From Section VII**

**A. Proofs From Section VII-A**

Proof of Theorem 23: By Lemma 16, the set-valued maps \( A \rightarrow p_0(A^{2\epsilon}) \) and \( A \rightarrow p_1(A^{2\epsilon}) \) are 2-alternating capacities. Hence, the existence of \( A^* \in \mathcal{L}(X) \) that attains the infimum on the right in (21) follows from Lemma 1 in [34] and the equality \( R_{\mathcal{B}}(X, A) = R_{\mathcal{T}}(X, A) \) proved in Theorem 18. By Theorem 4.1 in [34], there exist \( q_0, q_1 \in \mathcal{P}(X) \) such that \( W_\infty(p_1, q_i) \leq \epsilon \) for \( i = 0, 1 \) and,
\[
\inf_{A \in \mathcal{L}(X)} \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0, p_1) = \inf_{A \in \mathcal{L}(X)} r(A, q_0, q_1).
\]
Hence,
\[
\begin{align*}
\inf_{A \in \mathcal{L}(X)} \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} & r(A, p_0, p_1) \\
= & \inf_{A \in \mathcal{L}(X)} r(A, q_0, q_1) \\
\leq & \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} \inf_{A \in \mathcal{L}(X)} r(A, p_0, p_1).
\end{align*}
\]
The desired result follows by combining the above inequality with (20). Clearly, \( q_0 = p_0' \) and \( q_1 = p_1' \).

**B. Proofs From Section VII-B**

Lemma 38 (Max-min Inequality): Let \( p_0, p_1 \in \mathcal{P}(X) \) and let \( \epsilon \geq 0 \). For \( T > 0 \), define \( r : B(X') \times \mathcal{P}(X') \times \mathcal{P}(X') \rightarrow [0, 1] \) as in (19). Then,
\[
\begin{align*}
\sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} & \inf_{A \in B(X)} r(A, p_0, p_1') \\
\leq & \inf_{A \in B(X)} \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0, p_1').
\end{align*}
\]

Proof: For any \( A \in \mathcal{B}(X) \) and \( p_0', p_1' \) such that \( W_\infty(p_0, p_0') \leq \epsilon \) (\( i = 0, 1 \)), we have
\[
\inf_{A \in \mathcal{B}(X)} r(A, p_0', p_1') \leq r(A, p_0', p_1').
\]
Taking supremum over \( p_0' \) and \( p_1' \) such that \( W_\infty(p_1, p_1') \leq \epsilon \) for \( i \in \{0, 1\} \) on both sides of the above inequality, we get the following for any \( A \in \mathcal{B}(X) \).
\[
\begin{align*}
\sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} & \inf_{A \in \mathcal{B}(X)} r(A, p_0', p_1') \\
\leq & \sup_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} r(A, p_0', p_1').
\end{align*}
\]
Since the above inequality holds for any \( A \in \mathcal{B}(X) \), we get the desired result.

Proof of Theorem 24: Consider any \( \mu' \) and \( \nu' \) such that \( W_\infty(\mu, \mu') \leq \epsilon \) and \( W_\infty(\nu, \nu') \leq \epsilon \). Then there exist \( \gamma_\mu \in \Pi(\mu, \mu') \) and \( \gamma_\nu \in \Pi(\nu, \nu') \) such that
\[
\begin{align*}
\mathbb{P}_{(x, x') \sim \gamma_\mu}(d(x, x') > \epsilon) &= 0, \\
\mathbb{P}_{(x, x') \sim \gamma_\nu}(d(x, x') > \epsilon) &= 0.
\end{align*}
\]
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,
\[
D_\epsilon(\mu, \nu) \leq \int_{X^2} \mathbb{1}\{d(x, x') > 2\epsilon\} d\gamma_0
\leq \int_{X^2} \mathbb{1}\{d(x, x') > 0\} d\gamma'
\leq D_{TV}(\mu', \nu').
\]
Since the above inequality is true for any \( \mu' \) and \( \nu' \) such that \( W_\infty(\mu, \mu') \leq \epsilon \) and \( W_\infty(\nu, \nu') \leq \epsilon \), we have the following.
\[
D_\epsilon(\mu, \nu) \leq \inf_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} D_{TV}(\mu', \nu').
\]

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 \( D_\epsilon(\mu, \nu) \). Let \( M : \mathcal{X}^2 \rightarrow \mathcal{X}^2 \) be a measurable midpoint map. (See [25] 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 \( T : \mathcal{X}^2 \rightarrow \mathcal{X}^2 \) defined as
\[
T(x, x') = \begin{cases} (M(x, x'), M(x, x')) & d(x, x') \leq 2\epsilon, \\
(x, x') & \text{otherwise.}
\end{cases}
\]
\( T \) is measurable because it is piece-wise measurable on measurable sets. Further, each coordinate of \( (x, x') \) is transported by \( T \) by a distance no further than \( \epsilon \). Let \( \mu_{\gamma} \) and \( \nu_{\gamma} \) be the first and second marginals of \( T_\gamma \), respectively. Then, \( W_\infty(\mu, \mu_\gamma) \leq \epsilon \) and \( W_\infty(\nu, \nu_\gamma) \leq \epsilon \). Hence,
\[
D_\epsilon(\mu, \nu) = \int_{X^2} \mathbb{1}\{d(x, x') > 2\epsilon\} d\gamma
\leq \int_{X^2} \mathbb{1}\{d(x, x') > 0\} d\gamma_{\gamma_T}
\geq \inf_{W_\infty(p_0, p_0'), W_\infty(p_1, p_1') \leq \epsilon} D_{TV}(\mu', \nu').
\]
Combining this with (41), it is clear that the infimum over $D_{TV}$ is attained by $p_0$ and $\nu_0$.

**Proof of Lemma 26**: The first equality in (23) follows from Theorem 24. For the second equality, we have the following.

$$
\inf_{q \in \mathcal{P}'} D_{\epsilon}(q, p_1)
$$

(i) $= 1 - (T + 1) \inf_{A \in \mathcal{B}(\mathcal{X})} R_{\epsilon}(\ell_{0,1}, A)
$

(ii) $\geq 1 - (T + 1) \inf_{A \in \mathcal{B}(\mathcal{X})} R_{\epsilon}(\ell_{0,1}, A)
$

(iii) $= 1 - (T + 1) \inf_{A \in \mathcal{B}(\mathcal{X})} \sup_{R_{\epsilon}(p_0, p_1)} r(A, p_0, p_1)
$

(iv) $\leq 1 - (T + 1) \inf_{W_{\infty}(p_0, p_1)} \sup_{A \in \mathcal{B}(\mathcal{X})} \inf_{q \in \mathcal{P}'} D_{TV}(q', p_1)
$

Consider arbitrary probability measures $q', p_1 \in \mathcal{P}$ 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 $W_{\infty}(q, q') \leq \epsilon$ and $W_{\infty}(p_1, p_1') \leq \epsilon$ where $q \in T p_0$. We will now construct $p_0 \in \mathcal{P}'$ such that $q' \leq T p_0$ and $W_{\infty}(p_0, p_0) \leq \epsilon$. This will show that the set of $q', p_1 \in \mathcal{P}'$ 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 $p_0' \in \mathcal{P}$ as $p_0'(A) = p_0(A) + 1/2 q'(A) - 1/2 q(A)$ for $A \in \mathcal{B}(\mathcal{X})$. To show that $p_0'$ is a valid probability measure, we have the following.

$$
p_0'(A) = p_0'(X) = p_0(A) + \frac{1}{2} q'(A) - \frac{1}{2} q(A) = 1
$$

$$
p_0'(A) = \frac{1}{T} (T p_0(A) - q(A)) + \frac{1}{T} q'(A) \geq \frac{1}{T} q'(A) \geq 0.
$$

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

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

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

$$
\gamma'(A_1 \times A_2) = p_0(A_1 + p_0 A_2) = p_0(A_1) + p_0(A_2)
$$

Therefore, $W_{\infty}(p_0, p_0') \leq \epsilon$. □

**ACKNOWLEDGMENT**

The authors thank anonymous reviewers for their insightful comments on the journal version of this article.

**References**

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

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[3] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[4] J. Cohen, E. Rosenfeld, and I. Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1310–1320.

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

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

[7] A. Jalal, A. Ilyas, C. Daskalakis, and A. G. Dimakis, “The robust manifold defense: Adversarial training using generative models,” 2017, arXiv:1712.09196.

[8] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” *J. Field Robot.,* vol. 37, no. 3, pp. 362–386, Apr. 2020.

[9] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. de Albuquerque, “Deep learning for safe autonomous driving: Current challenges and future directions,” *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4316–4336, Jul. 2021.

[10] H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest editorial: Deep learning in medical image analysis systems,” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[11] X. Ma et al., “Understanding adversarial attacks on deep learning based medical image analysis systems,” *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107332.

[12] Z. Liu, J. Zhang, V. Jog, P.-L. Loh, and A. B. Mcmillan, “Robustifying deep networks for medical image segmentation,” *J. Digit. Imag.*, vol. 34, no. 5, pp. 1279–1293, Oct. 2021.

[13] R. Kumar, D. O'Brien, K. Albert, and S. Vilojen, “Law and adversarial machine learning,” in *Proc. NIPS*, 2018, pp. 1–6.

[14] I. Chalkidis and D. Kampas, “Deep learning in law: Early adaptation and legal word embeddings trained on large corpora,” *Artif. Intell. Law.*, vol. 27, no. 2, pp. 171–198, Jun. 2019.

[15] J. Uesato, B. Oadonoghue, P. Kohli, and A. Oord, “Adversarial risk and the dangers of evaluating against weak attacks,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5025–5034.
D. I. Diochnos, S. Mahloujifar, and M. Mahmoody, “Adversarial risk and robustness: General definitions and implications for the uniform distribution,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 31, 2018, pp. 10359–10368.

P. Gourreau, V. Kanade, M. Kwiatkowska, and J. Worrell, “On the hardness of robust classification,” J. Mach. Learn. Res., vol. 22, no. 1, pp. 12521–12549, 2021.

R. Pinot, R. Ettegdui, G. Rizk, Y. Chevaleyre, and J. Atif, “Randomization matters how to defend against strong adversarial attacks,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 7717–7727.

M. S. Pydi and V. Jog, “Adversarial risk via optimal transport and optimal couplings,” IEEE Trans. Inf. Theory, vol. 67, no. 9, pp. 6031–6052, Sep. 2021.

L. Meunier, M. Scetbon, R. Pinot, J. Atif, and Y. Chevaleyre, “Mixed Nash equilibria in the adversarial examples game,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 7496–7508.

A. Shaham, Y. Yamada, and S. Negahban, “Understanding adversarial training: Increasing local stability of supervised models through robust optimization,” Neurocomputing, vol. 307, pp. 195–204, Sep. 2018.

M. S. Pydi and V. Jog, “Adversarial risk via optimal transport and optimal couplings,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 7814–7823.

A. Bhagoji, D. Cullina, and P. Mittal, “Lower bounds on adversarial robustness from optimal transport,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 7496–7508.

E. Dohmatob, “Generalized no free lunch theorem for adversarial robustness,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 1646–1654.

E. Dohmatob, “Classifier-independent lower-bounds for adversarial robustness,” 2020, arXiv:2006.09989.

M. Staib and S. Jegelka, “Distributionally robust deep learning as a generalization of adversarial training,” in Proc. NIPS, 2017, pp. 1–4.

A. Sinha, H. Namkoong, and J. C. Duchi, “Certifying some distributional robustness with principled adversarial training,” in Proc. Int. Conf. Learn. Represent., 2017, pp. 1–49.

Z. Tu, J. Zhang, and D. Tao, “Theoretical analysis of adversarial learning: A minimax approach,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–11.

A. Bose et al., “Adversarial example games,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 8921–8934.

S. R. Bulo, B. Biggio, I. Pillai, M. Pelillo, and F. Rolli, “Randomized prediction games for adversarial machine learning,” IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 11, pp. 4666–4678, Nov. 2017.

D. P. Bertsekas and S. E. Shreve, Stochastic Optimal Control: The Discrete-Time Case, vol. 5. Nashua, NH, USA: Athena Scientific, 1996.

D. H. Wagner, “Survey of measurable selection theorems,” SIAM J. Control Optim., vol. 15, no. 5, pp. 859–903, Aug. 1977.

G. Choquet, “Theory of capacities,” Annales de L’Institut Fourier, vol. 5, pp. 131–295, Jan. 1954.

P. J. Huber and V. Strassen, “Minimax tests and the Neyman–Pearson Lemma for capacities,” Ann. Statist., vol. 1, no. 2, pp. 251–263, Mar. 1973.

V. Strassen, “The existence of probability measures with given marginals,” Ann. Math. Statist., vol. 36, no. 2, pp. 423–439, Apr. 1965.

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

S. Mahloujifar, D. I. Diochnos, and M. Mahmoody, “The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure,” in Proc. AAAI, 2019, vol. 33, no. 1, pp. 4536–4543.

M. S. Pydi and V. Jog, “The many faces of adversarial risk,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 10000–10012.

P. Erdos and A. Stone, “On the sum of two Borel sets,” Proc. Amer. Math. Soc., vol. 25, no. 2, pp. 304–306, 1970.

W. Sierpiński, “Sur la question de la mesurabilité de la base de M. Hamel,” Fundamenta Mathematicae, vol. 1, no. 1, pp. 105–111, 1920.

C. R. Givens and R. M. Shortt, “A class of Wasserstein metrics for probability distributions,” Michigan Math. J., vol. 31, no. 2, pp. 231–240, Jan. 1984.

A. S. Suggala, A. Prasad, V. Nagarajan, and P. Ravikumar, “Revisiting adversarial risk,” in Proc. 22nd Int. Conf. Artif. Intell. Statist., 2019, pp. 2331–2339.

R. Bhattacharjee and K. Chaudhuri, “When are non-parametric methods robust?” in Proc. Int. Conf. Mach. Learn., 2020, pp. 832–841.

L. Zajicek, “On σ-porous sets in abstract spaces,” in Abstract and Applied Analysis. London, U.K.: Hindawi, 2005, pp. 509–534.

R. J. Gardner, “The brunn-minkowski inequality,” Bull. Amer. Math. Soc., vol. 39, no. 3, pp. 355–405, 2002.

M. Yue, D. Kuhn, and W. Wiesemann, “On linear optimization over Wasserstein balls,” Math. Program., vol. 195, pp. 1107–1122, Sep. 2021.

R. Dudley, “Distances of probability measures and random variables,” in Selected Works of RM Dudley. Berlin, Germany: Springer, 2010, pp. 28–37.

G. Schay, “Nearest random variables with given distributions,” Ann. Probab., vol. 2, no. 1, pp. 163–166, Feb. 1974.

L. Yu and V. Y. F. Tan, “Asymptotic coupling and its applications in information theory,” IEEE Trans. Inf. Theory, vol. 65, no. 3, pp. 1321–1344, Mar. 2019.

L. Yu, “Asymptotics for Strassen’s optimal transport problem,” 2019, arXiv:1912.02051.

N. S. G. Trillos and R. Murray, “Adversarial classification: Necessary conditions and geometric flows,” J. Mach. Learn. Res., vol. 23, no. 1, pp. 8486–8523, 2022.

Y. Jang, T. Zhao, S. Hong, and H. Lee, “Adversarial defense via learning to generate diverse attacks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2740–2749.

H. Jiang, Z. Chen, Y. Shi, B. Dai, and T. Zhao, “Learning to defend by learning to attack,” in Proc. Int. Conf. Artif. Intell. Statist., 2021, pp. 577–585.

A. Jeddi, M. J. Shaifee, M. Karg, C. Scharfenberger, and A. Wong, “Learn2Perturb: An end-to-end feature perturbation learning to improve adversarial robustness,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 1238–1247.

I. M. Baytas and D. Deb, “Robustness via-synthesis: Robust training with generative adversarial perturbations,” Neurocomputing, vol. 516, pp. 49–60, Jan. 2023.

H. Luiro, M. Parviainen, and E. Saksman, “On the existence and uniqueness of p-harmonic functions,” Differ. Integral Equ., vol. 27, nos. 3–4, pp. 201–216, 2014.

V. I. Bogachev, Measure Theory, vol. 2. Berlin, Germany: Springer, 2007.

J. Matousek and B. Gärtner, Understanding and Using Linear Programming. Berlin, Germany: Springer, 2007.

P. Billingsley, Convergence of Probability Measures (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 1999.

Muni Sreenivas Pydi received the B.Tech. degree in electrical engineering from IIT Madras in 2014, and the M.S. degree in electrical engineering from the University of Wisconsin–Madison, WI, USA, in 2019. He is currently an inaugural Junior Fellow in artificial intelligence at the Université PSL (Paris Sciences & Lettres), affiliated with the LAMSADE (Laboratoire d’Analyse et de Modélisation de Systèmes pour l’Aide à la Décision) Research Center, Université Dauphine-PSL, France. His research interests include information theory, machine learning, and optimal transport.

Varun Jog received the B.Tech. degree in electrical engineering from IIT Bombay in 2010, and the Ph.D. degree in electrical engineering and computer sciences (EECS) from UC Berkeley in 2015. He was an Assistant Professor with the Electrical and Computer Engineering Department, UW-Madison, from 2016 to 2020. Since January 2021, he has been with the Department of Pure Mathematics and Mathematical Statistics (DPMMS), University of Cambridge. His research interests include information theory, machine learning, and statistics. He was a recipient of the NSF-CAREER Award in 2020, the Eli Jury Award from the EECS Department at UC Berkeley in 2015, and the Jack Keil Wolf Student Paper Award at ISIT 2015.