Journal of Machine Learning Research 25 (2024) 1-41 Submitted 4/23; Revised 12/23; Published 1/24

Existence and Minimax Theorems for Adversarial Surrogate Risks in Binary Classification

Natalie S. Frank nf1066@nyu.edu
Courant Institute
New York University
New York, NY 10012, USA

Jonathan Niles-Weed jnw@cims.nyu.edu
Courant Institute and Center for Data Science
New York University
New York, NY 10012, USA

Editor: Gabor Lugosi

Abstract

We prove existence, minimax, and complementary slackness theorems for adversarial surrogate risks in binary classification. These results extend recent work of Pydi and Jog (2021), who established analogous minimax and existence theorems for the adversarial classification risk. We show that their conclusions continue to hold for a very general class of surrogate losses; moreover, we remove some of the technical restrictions present in prior work. Our results provide an explanation for the phenomenon of transfer attacks and inform new directions in algorithm development.

Keywords: Adversarial Learning, Minimax Theorems, Optimal Transport, Adversarial Bayes risk, Convex Relaxation

1. Introduction

Neural networks are state-of-the-art methods for a variety of machine learning tasks including image classification and speech recognition. However, a concerning problem with these models is their susceptibility to *adversarial attacks*: small perturbations to inputs can cause incorrect classification by the network (Szegedy et al., 2013; Biggio et al., 2013). This issue has security implications; for instance, Gu et al. (2017) show that a yellow sticker can cause a neural net to misclassify a stop sign. Furthermore, one can find adversarial examples that generalize to other neural nets; these sort of attacks are called *transfer attacks*. In other words, an adversarial example generated for one neural net will sometimes be an adversarial example for a different neural net trained for the same classification problem (Tramèr et al., 2017; Demontis et al., 2018; Kurakin et al., 2017; Rozsa et al., 2016; Papernot et al., 2016). This phenomenon shows that access to a specific neural net is not necessary for generating adversarial examples. One method for defending against such adversarial perturbations is *adversarial training*, in which a neural net is trained on adversarially perturbed data points (Kurakin et al., 2017; Madry et al., 2019; Wang et al., 2021). However, adversarial training is not well understood from a theoretical perspective.

From a theoretical standpoint, the most fundamental question is whether it is possible to design models which are robust to such attacks, and what the properties of such robust models might be. In contrast to the classical, non-adversarial setting, much is still unknown about the basic properties of optimal robust models. In the context of binary classification, several prior works study properties of the *adversarial classification risk*—the expected number of classification errors under adversarial perturbations. Recently, Awasthi et al. (2023), Bungert et al. (2021), and Pydi and Jog (2021) all showed existence of a minimizer to the adversarial classification risk under suitable assumptions, and characterized some of its properties. A crucial observation, emphasized by Pydi and Jog (2021), is that minimizing the adversarial classification risk is equivalent to a *dual* robust classification problem involving uncertainty sets with respect to the ∞-Wasserstein metric. This observation gives rise to a game-theoretic interpretation of robustness, which takes the form of a zero-sum game between a classifier and an adversary who is allowed to perturb the data by a certain amount. As noted by Pydi and Jog (2021), this interpretation has implications for algorithm design by suggesting that robust classifiers can be constructed by jointly optimizing over classification rules and adversarial perturbations.

This recent progress on adversarial binary classification lays the groundwork for a theoretical understanding of adversarial robustness, but it is limited insofar as it focuses only on minimizers of the adversarial classification risk. In practice, minimizing the empirical adversarial classification risk is computationally intractable; as a consequence; the adversarial training procedure typically minimizes an objective called a *surrogate* risk, which is chosen to be easier to optimize. In the non-adversarial setting, the properties of surrogate risks are well known (see, e.g. Bartlett et al., 2006), but in the adversarial scenario, existing results for the adversarial classification risk fail to carry over to surrogate risks. In particular, the existence and minimax results described above are not known to hold. We close this gap by developing an analogous theory for adversarial surrogate risks. Our main theorems (Theorems 7–9) establish that strong duality holds for the adversarial surrogate risk minimization problem, that solutions to the primal and dual problems exist, and that these optimizers satisfy a complementary slackness condition.

These results suggest explanations for empirical observations, such as the existence of transfer attacks. Specifically, our analysis suggests that adversarial examples are a property of the data distribution rather than a specific model. In fact, the complimentary slackness theorem presented in this paper states that certain attacks are the strongest possible adversary against *any* minimizer of the adversarial surrogate risk, which might explain why adversarial examples tend to transfer between trained neural nets. Furthermore, our theorems suggest that a training algorithm should optimize over neural nets and adversarial perturbations simultaneously. Adversarial training, the current state of the art method for finding adversarially robust networks, does not follow this procedure. The adversarial training algorithm tracks an estimate of the optimal function $\tilde{f}$. To update $\tilde{f}$, the algorithm first finds *optimal* adversarial examples at the current estimate $\tilde{f}$, and then performs a gradient descent step. See the papers (Kurakin et al., 2017; Madry et al., 2019; Goodfellow et al., 2014) for more details on adversarial training. Finding these adversarial examples is a computationally intensive procedure. On the other hand, algorithms for optimizing minimax problems in the finite dimensional setting alternate between primal and dual steps (Mokhtari et al., 2019). This observation suggests that designing an algorithm that optimizes over model parameters and adversarial perturbations simultaneously is a promising research direction. Trillos and Trillos (2023); Wang and Chizat (2023); Domingo-Enrich et al. (2021) adopt this approach, and one can view the minimax results of this paper as a mathematical justification for the use of surrogate losses in such algorithms.

Lastly, our theorems are an important first step in understating statistical properties of surrogate losses. Recall that one minimizes a surrogate risk because minimizing the original risk is computationally intractable. If a sequence of functions which minimizes the surrogate risk also minimizes the classification risk, then that surrogate is referred to as a *consistent risk*. Similarly, if a sequence of functions which minimizes the adversarial surrogate risk also minimizes the adversarial classification risk, then that surrogate is referred to as an *adversarially consistent risk*. Much prior work studies the consistency of surrogate risks (Bartlett et al., 2006; Lin, 2004; Steinwart, 2007; Philip M. Long, 2013; Mingyuan Zhang, 2020; Zhang, 2004). Alarmingly, Meunier et al. (2022) show that a family of surrogates used in applications is not adversarially consistent. In follow-up work, we show that our results can be used to characterize adversarially consistent supremum-based risks for binary classification (Frank and Niles-Weed, 2023), strengthening results on calibration in the adversarial setting Bao et al. (2021); Meunier et al. (2022); Awasthi et al. (2021).

2. Related Works

This paper extends prior work on the adversarial Bayes classifier. Pydi and Jog (2021) first proved multiple minimax theorems for the adversarial classification risk using optimal transport and Choquet capacities, showing an intimate connection between adversarial learning and optimal transport. Later, follow-up work used optimal transport minimax reformulations of the adversarial learning problem to derive new algorithms for adversarial learning. Trillos et al. (2022) reformulate adversarial learning in terms of a multi-marginal optimal transport problem and then apply existing techniques from optimal transport to find a new algorithm. Trillos and Trillos (2023); Wang and Chizat (2023); Domingo-Enrich et al. (2021) propose ascent-descent algorithms based on optimal transport and use mean-field dynamics to analyze convergence. These approaches leverage the minimax view of the adversarial training problem to optimize over model parameters and optimal attacks simultaneously. Gao et al. (2022) use an optimal transport reformulation to find regularizers that encourage robustness. Wong et al. (2019); Wu et al. (2020) use Wasserstein metrics to formulate adversarial attacks on neural networks.

Further work analyzes properties of the adversarial Bayes classifier. Awasthi et al. (2023), Bhagoji et al. (2019), and Bungert et al. (2021) all prove the existence of the adversarial Bayes classifier, using different techniques. Yang et al. (2020) studied the adversarial Bayes classifier in the context of non-parametric methods. Pydi and Jog (2019) and Bhagoji et al. (2019) further introduced methods from optimal transport to study adversarial learning. Lastly, (Trillos and Murray, 2020) give necessary and sufficient conditions describing the boundary of the adversarial Bayes classifier. Simultaneous work (Li and Telgarsky, 2023) also proves the existence of minimizers to adversarial surrogate risks using prior results on the adversarial Bayes classifier.

The adversarial training algorithm is also well studied from an empirical perspective. Demontis et al. (2018) discussed an explanation of transfer attacks on neural nets trained using standard methods, but did not extend their analysis to the adversarial training setting. (Wang et al., 2021; Kurakin et al., 2017; Madry et al., 2019) study the adversarial training algorithm and improving the runtime. Two particularly popular attacks used in adversarial training are the FGSM attack (Goodfellow et al., 2014) and the PGD attack (Madry et al., 2019). More recent popular variants of this algorithm include (Shafahi et al., 2019; Xie et al., 2018; Kannan et al., 2018; Wong et al., 2020).

3. Background and Notation

3.1 Adversarial Classification

This paper studies binary classification on $\mathbb{R}^d$ with two classes encoded as -1 and $+1$. Data is distributed according to a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \{-1,+1\}$. We denote the marginals according to the class labels as $\mathbb{P}_0(S) = \mathcal{D}(S\times\{-1\})$ and $\mathbb{P}_1(S) = \mathcal{D}(S\times\{+1\})$. Throughout the paper, we assume $\mathbb{P}_0(\mathbb{R}^d)$ and $\mathbb{P}_1(\mathbb{R}^d)$ are finite but not necessarily that $\mathbb{P}_0(\mathbb{R}^d) + \mathbb{P}_1(\mathbb{R}^d) = 1$.

To classify points in $\mathbb{R}^d$, algorithms typically learn a real-valued function f and then classify points $\mathbf{x}$ according to the sign of f (arbitrarily assigning the sign of 0 to be -1). The *classification error*, also known as the *classification risk*, of a function f is

$$R(f) = \int \mathbf{1}_{f(\mathbf{x})\leq 0} d\mathbb{P}_1 + \int \mathbf{1}_{f(\mathbf{x})>0} d\mathbb{P}_0. \tag{1}$$

Notice that finding minimizers to R is straightforward: define the measure $\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$ and let $\eta = d\mathbb{P}_1/d\mathbb{P}$. Then the risk R can be re-written as

$$R(f) = \int \eta(\mathbf{x})\mathbf{1}_{f(\mathbf{x})\leq 0} + (1-\eta(\mathbf{x}))\mathbf{1}_{f(\mathbf{x})>0} d\mathbb{P}.$$

Hence a minimizer of R must minimize the function $C(\eta(\mathbf{x}),\alpha) = \eta(\mathbf{x})\mathbf{1}_{\alpha\leq 0} + (1-\eta(\mathbf{x}))\mathbf{1}_{\alpha>0}$ at each $\mathbf{x}$ $\mathbb{P}$-a.e. The optimal Bayes risk is then

$$\inf_f R(f) = \int C^*(\eta) d\mathbb{P}$$

where $C^*(\eta) = \inf_\alpha C(\eta,\alpha) = \min(\eta, 1-\eta)$.

This paper analyzes the *evasion attack*, in which an adversary knows both the function f and the true label of the data point, and can perturb each input before it is evaluated by the function f. To constrain the adversary, we assume that perturbations are bounded by ϵ in a norm $\|\cdot\|$. Thus a point $\mathbf{x}$ with label $+1$ is misclassified if there is a perturbation $\mathbf{h}$ with $\|\mathbf{h}\| \leq \epsilon$ for which $f(\mathbf{x}+\mathbf{h}) \leq 0$ and a point $\mathbf{x}$ with label -1 is misclassified if there is a perturbation $\mathbf{h}$ with $\|\mathbf{h}\| \leq \epsilon$ for which $f(\mathbf{x}+\mathbf{h}) > 0$. Therefore, the *adversarial classification risk* is

$$R^\epsilon(f) = \int \sup_{\|\mathbf{h}\|\leq\epsilon} \mathbf{1}_{f(\mathbf{x}+\mathbf{h})\leq 0} d\mathbb{P}_1 + \int \sup_{\|\mathbf{h}\|\leq\epsilon} \mathbf{1}_{f(\mathbf{x}+\mathbf{h})>0} d\mathbb{P}_0 \tag{2}$$

which is the expected proportion of errors subject to the adversarial evasion attack. The expression $\sup_{\|\mathbf{h}\|\leq\epsilon} \mathbf{1}_{f(\mathbf{x}+\mathbf{h})\leq 0}$ evaluates to 1 at a point $\mathbf{x}$ iff $\mathbf{x}$ can be moved into the set

$[f \leq 0]$ by a perturbation of size at most ϵ. Equivalently, this set is the Minkowski sum $\oplus$ of $[f \leq 0]$ and $\overline{B_\epsilon(\mathbf{0})}$. For any set A, let A^ϵ denote

$$A^\epsilon = \{\mathbf{x}\colon \exists \mathbf{h} \text{ with } \|\mathbf{h}\| \leq \epsilon \text{ and } \mathbf{x}+\mathbf{h} \in A\} = A \oplus \overline{B_\epsilon(\mathbf{0})} = \bigcup_{\mathbf{a}\in A} \overline{B_\epsilon(\mathbf{a})}. \tag{3}$$

This operation 'thickens' the boundary of a set by ϵ. With this notation, (2) can be expressed as $R^\epsilon(f) = \int \mathbf{1}_{\{f\leq 0\}^\epsilon} d\mathbb{P}_1 + \int \mathbf{1}_{\{f>0\}^\epsilon} d\mathbb{P}_0$.

Unlike the classification risk R, finding minimizers to R^ϵ is nontrivial. One can re-write R^ϵ in terms of $\mathbb{P}$ and η but the resulting integrand cannot be minimized in a pointwise fashion. Nevertheless, it can be shown that minimizers of R^ϵ exist (Awasthi et al., 2023; Bungert et al., 2021; Pydi and Jog, 2021; Frank and Niles-Weed, 2023).

3.2 Surrogate Risks

As minimizing the empirical version of risk in (1) is computationally intractable, typical machine learning algorithms minimize a proxy to the classification risk called a *surrogate risk*. In fact, Ben-David et al. (2003) show that minimizing the empirical classification risk is NP-hard in general. A popular surrogate is

$$R_\phi(f) = \int \phi(f) d\mathbb{P}_1 + \int \phi(-f) d\mathbb{P}_0 \tag{4}$$

where ϕ is a decreasing function.[1] To define a classifier, one then threshholds f at zero. There are many reasonable choices for ϕ—one typically chooses an upper bound on the zero-one loss which is easy to optimize. We make the following assumption on ϕ:

Assumption 1 *The loss ϕ is non-increasing, non-negative, lower semi-continuous, and $\lim_{\alpha\to\infty} \phi(\alpha) = 0$.*

A particularly important example, which plays a large role in our proofs, is the exponential loss $\psi(\alpha) = e^{-\alpha}$, which will be denoted by ψ in the remainder of this paper. Assumption 1 includes many but not all all surrogate risks used in practice. Notably, some multiclass surrogate risks with two classes are of a somewhat different form, see for instance (Tewari and Bartlett, 2007) for more details.

In order to find minimizers of R_ϕ, we rewrite the risk in terms of $\mathbb{P}$ and η as

$$R_\phi(f) = \int \eta(\mathbf{x})\phi(f(\mathbf{x})) + (1-\eta(\mathbf{x}))\phi(-f(\mathbf{x})) d\mathbb{P} \tag{5}$$

Hence the minimizer of R_ϕ must minimize $C_\phi(\eta, \cdot)$ pointwise $\mathbb{P}$-a.e., where

$$C_\phi(\eta, \alpha) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha).$$

In other words, if one defines $C_\phi^*(\eta) = \inf_\alpha C_\phi(\eta, \alpha)$, then a function f^* is optimal if and only if

$$\eta(\mathbf{x})\phi(f^*(\mathbf{x})) + (1-\eta(\mathbf{x}))\phi(-f^*(\mathbf{x})) = C_\phi^*(\eta(\mathbf{x})) \quad \mathbb{P}\text{-a.e.} \tag{6}$$

1. Notice that due to the asymmetry of the sign function at 0 in (1), R_ϕ is not quite a generalization of R.

Thus one can write the minimum value of R_ϕ as

$$\inf_f R_\phi(f) = \int C_\phi^*(\eta)d\mathbb{P}. \tag{7}$$

To guarantee the existence of a function satisfying (6), we must allow our functions to take values in the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. Allowing the value $\alpha = +\infty$ is necessary, for instance, for the exponential loss $\psi(\alpha) = e^{-\alpha}$: when $\eta = 1$, the minimum of $C_\psi(1, \alpha) = e^{-\alpha}$ is achieved at $\alpha = +\infty$. In fact, one can express a minimizer as a function of the conditional probability $\eta(\mathbf{x})$ using (6). For a loss ϕ, define $\alpha_\phi : [0, 1] \to \overline{R}$ by

$$\alpha_\phi(\eta) = \inf\{\alpha : \alpha \text{ is a minimizer of } C_\phi(\eta, \cdot)\}. \tag{8}$$

Lemma 25 in Appendix C shows that the function α_ϕ is montonic and $\alpha_\phi(\eta)$ is in fact a minimizer of $C_\phi(\eta, \cdot)$. Thus

$$f^*(\mathbf{x}) = \alpha_\phi(\eta(\mathbf{x})) \tag{9}$$

is measurable and satisfies (6). Therefore, the function f^* must be a minimizer of the risk R_ϕ.

Similarly, rather directly minimizing the adversarial classification risk, practical algorithms minimize an *adversarial surrogate*. The adversarial counterpart to (4) is

$$R_\phi^\epsilon(f) = \int \sup_{\|\mathbf{h}\| \leq \epsilon} \phi(f(\mathbf{x} + \mathbf{h}))d\mathbb{P}_1 + \int \sup_{\|\mathbf{h}\| \leq \epsilon} \phi(-f(\mathbf{x} + \mathbf{h}))d\mathbb{P}_0. \tag{10}$$

Due to the definitions of the adversarial risks (2) and (10), the operation of finding the supremum of a function over ϵ-balls is central to our subsequent analysis. For a function g, we define

$$S_\epsilon(g)(\mathbf{x}) = \sup_{\|\mathbf{h}\| \leq \epsilon} g(\mathbf{x} + \mathbf{h}) \tag{11}$$

Using this notation, one can re-write the risk R_ϕ^ϵ as

$$R_\phi^\epsilon(f) = \int S_\epsilon(\phi \circ f)d\mathbb{P}_1 + \int S_\epsilon(\phi \circ -f)d\mathbb{P}_0$$

By analogy to (5), we equivalently write the risk R_ϕ^ϵ in terms of $\mathbb{P}$ and η:

$$R_\phi^\epsilon(f) = \int \eta(\mathbf{x})S_\epsilon(\phi \circ f)(\mathbf{x}) + (1 - \eta(\mathbf{x}))S_\epsilon(\phi \circ -f)(\mathbf{x})d\mathbb{P}. \tag{12}$$

However, unlike (5), because the integrand of R_ϕ^ϵ cannot be minimized in a pointwise manner, proving the existence of minimizers to R_ϕ^ϵ is non-trivial. In fact, unlike the adversarial classification risk R^ϵ, there is little theoretical understanding of the properties of R_ϕ^ϵ.

3.3 Measurability

In order to define the adversarial risks R^ϵ and R_ϕ^ϵ, one must show that $S_\epsilon(\mathbf{1}_A), S_\epsilon(\phi \circ f)$ are measurable. To illustrate this concern, Pydi and Jog (2021) show that for every $\epsilon > 0$ and $d > 1$, there is a Borel set C for which the function $S_\epsilon(\mathbf{1}_C)(\mathbf{x})$ is not Borel measurable. However, if g is Borel, then $S_\epsilon(g)$ is always measurable with respect to a larger σ-algebra called the *universal σ-algebra* $\mathscr{U}(\mathbb{R}^d)$. Such a function is called *universally measurable*. We prove the following theorem and formally define the universal σ-algebra in Appendix A.

Theorem 1 *If f is universally measurable, then $S_\epsilon(f)$ is also universally measurable.*

In fact, in Appendix A, we show that a function defined by a supremum of a universally measurable function over a compact set is universally measurable—a result of independent interest. The universal σ-algebra is smaller than the completion of $\mathcal{B}(\mathbb{R}^d)$ with respect to any Borel measure. Thus, in the remainder of the paper, unless otherwise noted, all measures will be Borel measures and the expression $\int S_\epsilon(f)d\mathbb{Q}$ will be interpreted as the integral of $S_\epsilon(f)$ with respect to the completion of $\mathbb{Q}$.

3.4 The W_∞ Metric

In this section, we explain how the integral of a supremum $\int S_\epsilon(f)d\mathbb{Q}$ can be expressed in terms of a supremum of integrals. We start by defining the Wasserstein-∞ metric.

Definition 2 *Let $\mathbb{P}, \mathbb{Q}$ be two finite measures with $\mathbb{P}(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$. A* coupling *is a positive measure on the product space $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mathbb{P}, \mathbb{Q}$. We denote the set of all couplings with marginals $\mathbb{P}$, $\mathbb{Q}$ by $\Pi(\mathbb{P}, \mathbb{Q})$. The ∞-Wasserstein distance with respect to a norm $\|\cdot\|$ is defined as*

$$W_\infty(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \operatorname*{ess\,sup}_{(\mathbf{x}, \mathbf{x}') \sim \gamma} \|\mathbf{x} - \mathbf{x}'\|$$

Jylhä (2014, Theorem 2.6) proves that the infimum is always attained. Therefore, $\mathbb{P}$, $\mathbb{Q}$ are within a Wasserstein-∞ distance of ϵ if there is a coupling γ for $\mathbb{P}$ and $\mathbb{Q}$ for which $\operatorname{supp} \gamma$ is contained in the set $\Delta_\epsilon = \{(\mathbf{x}, \mathbf{x}') \colon \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon\}$. This optimal coupling will be a useful tool in proving theorems throughout this paper.

The ∞-Wasserstein metric is closely related to the to the operation S_ϵ. First, we show that S_ϵ can be expressed as a supremum of integrals over a Wasserstein-∞ ball. For a measure $\mathbb{Q}$, we write

$$\mathcal{B}^\infty_\epsilon(\mathbb{Q}) = \{\mathbb{Q}' \text{ Borel} : W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon\}.$$

Lemma 3 *Let $\mathbb{Q}$ be a finite positive Borel measure and let $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a Borel measurable function. Then*

$$\int S_\epsilon(f) d\mathbb{Q} = \sup_{\mathbb{Q}' \in \mathcal{B}^\infty_\epsilon(\mathbb{Q})} \int f d\mathbb{Q}' \tag{13}$$

Lemma 5.1 of Pydi and Jog (2021) proves an analogous statement for sets, namely that $\mathbb{Q}(A^\epsilon) = \sup_{\mathbb{Q}' \in \mathcal{B}^\infty_\epsilon(\mathbb{Q})} \mathbb{Q}(A)$, under suitable assumptions on $\mathbb{Q}$ and $\mathbb{Q}'$.

Conversely, the W_∞ distance between two probability measures can be expressed in terms of the integrals of f and $S_\epsilon(f)$. Let $C_b(X)$ be the set of continuous bounded functions on the topological space X.

Lemma 4 *Let $\mathbb{P}, \mathbb{Q}$ be two finite positive Borel measures with $\mathbb{P}(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$. Then*

$$W_\infty(\mathbb{P}, \mathbb{Q}) = \inf_\epsilon \{\epsilon \geq 0 \colon \int h d\mathbb{Q} \leq \int S_\epsilon(h) d\mathbb{P} \quad \forall h \in C_b(\mathbb{R}^d)\}$$

This observation will be central to proving a duality result. See Appendix B for proofs of Lemmas 3 and 4.

4. Main Results and Outline of the Paper

4.1 Summary of Main Results

Our goal in this paper is to understand properties of the surrogate risk minimization problem $\inf_f R_\phi^\epsilon$. The starting point for our results is Lemma 3, which implies that $\inf_f R_\phi^\epsilon$ actually involves a inf followed by a sup:

$$\inf_{f \text{ Borel}} R_\phi^\epsilon(f) = \inf_{f \text{ Borel}} \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \int \phi \circ f d\mathbb{P}_1' + \int \phi \circ -f d\mathbb{P}_0'.$$

We therefore obtain a lower bound on $\inf_f R_\phi^\epsilon$ by swapping the sup and inf and recalling the definition of $C_\phi^*(\eta) = \inf_\alpha C_\phi(\eta, \alpha)$:

$$\begin{aligned}
\inf_{f \text{ Borel}} R_\phi^\epsilon(f) &\geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{f \text{ Borel}} \int \phi \circ f d\mathbb{P}_1' + \int \phi \circ -f d\mathbb{P}_0' \\
&= \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{f \text{ Borel}} \int \frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')} \phi(f) + \left(1 - \frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')}\right) \phi(-f) d(\mathbb{P}_0' + \mathbb{P}_1') \\
&\geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \int C_\phi^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')}\right) d(\mathbb{P}_0' + \mathbb{P}_1'). \qquad (14)
\end{aligned}$$

If we define

$$\bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') = \int C_\phi^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')}\right) d(\mathbb{P}_0' + \mathbb{P}_1'), \qquad (15)$$

then we have shown

$$\inf_{f \text{ Borel}} R_\phi^\epsilon(f) \geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1'). \qquad (16)$$

This statement is a form of weak duality.

When the surrogate adversarial risk is replaced by the standard adversarial classification risk, Pydi and Jog (2021) proved that the analogue of (16) is actually an equality, so that strong duality holds for the adversarial classification problem. Concretely, by analogy to (15), consider

$$\bar{R}(\mathbb{P}_0', \mathbb{P}_1') = \int C^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')}\right) d(\mathbb{P}_0' + \mathbb{P}_1').$$

Let μ be the Lebesgue measure and let $\mathcal{L}_\mu(\mathbb{R}^d)$ be the Lebesgue σ-algebra. Then define

$$\tilde{\mathcal{B}}_\epsilon^\infty(\mathbb{Q}) = \{\mathbb{Q}' \colon W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon, \mathbb{Q}' \text{ a measure on } (\mathbb{R}^d, \mathcal{L}_\mu(\mathbb{R}^d))\}. \qquad (17)$$

Pydi and Jog (2021) show the following.

Theorem 5 (Pydi and Jog, 2021, Theorem 7.1) *Assume that $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous with respect to the Lebesgue measure μ. Then*

$$\inf_{f \text{ Lebesgue}} R^\epsilon(f) = \sup_{\substack{\mathbb{P}_0' \in \tilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \tilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') \qquad (18)$$

and furthermore equality is attained at some Lebesgue measurable $\hat{f}$ *and* $\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_0$.

Additionally, $\hat{\mathbb{P}}_i = \mathbb{P}_i \circ \varphi_i^{-1}$ *for some universally measurable* φ_i *with* $\|\varphi_i(\mathbf{x}) - \mathbf{x}\| \leq \epsilon$, $\sup_{\|\mathbf{y}-\mathbf{x}\|\leq\epsilon} \mathbf{1}_{\hat{f}(\mathbf{y})\leq 0} = \mathbf{1}_{\hat{f}(\varphi_1(\mathbf{x}))\leq 0}$ $\mathbb{P}_1$*-a.e., and* $\sup_{\|\mathbf{y}-\mathbf{x}\|\leq\epsilon} \mathbf{1}_{\hat{f}(\mathbf{y})>0} = \mathbf{1}_{\hat{f}(\varphi_0(\mathbf{x}))>0}$ $\mathbb{P}_0$*-a.e.*

This is a foundational result in the theory of adversarial classification, but it leaves an open question crucial in applications: Does the strong duality relation extend to surrogate risks and to general measures? In this work, we answer this question in the affirmative.

We start by proving the following:

Theorem 6 (Strong Duality) *Let* $\mathbb{P}_0, \mathbb{P}_1$ *be finite Borel measures. Then*

$$\inf_{f \text{ Borel}} R_\phi^\epsilon(f) = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1'). \tag{19}$$

When $\epsilon = 0$, we recover the fundamental characterization of the minimum risk for standard (non-adversarial) classification in (7). Theorem 6 can be rephrased as

$$\inf_{f \text{ Borel}} \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \hat{R}_\phi(f, \mathbb{P}_0', \mathbb{P}_1') = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{f \text{ Borel}} \hat{R}_\phi(f, \mathbb{P}_0', \mathbb{P}_1') \tag{20}$$

where

$$\hat{R}_\phi(f, \mathbb{P}_0', \mathbb{P}_1') = \int \phi(f) d\mathbb{P}_1' + \int \phi(-f) d\mathbb{P}_0'$$

As discussed in Pydi and Jog (2021), this result has an appealing game theoretic interpretation: adversarial learning with a surrogate risk can be though of as a zero-sum game between the learner who selects a function f and the adversary who chooses perturbations through $\mathbb{P}_0'$ and $\mathbb{P}_1'$. Furthermore, the player to pick first does not have an advantage.

Additionally, (20) suggest that training adversarially robust classifiers could be accomplished by optimizing over primal functions f and dual distributions $\mathbb{P}_0', \mathbb{P}_1'$ *simultaneously.*

A consequence of Theorem 6 is the following complementary slackness conditions for optimizers $f^*, \mathbb{P}_0^*, \mathbb{P}_1^*$:

Theorem 7 (Complimentary Slackness) *The function* f^* *is a minimizer of* R_ϕ^ϵ *and* $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ *is a maximizer of* $\bar{R}_\phi$ *over the* W_∞ *balls around* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *iff the following hold:*

1)

$$\int \phi \circ f^* d\mathbb{P}_1^* = \int S_\epsilon(\phi(f^*)) d\mathbb{P}_1 \quad \text{and} \quad \int \phi \circ -f^* d\mathbb{P}_0^* = \int S_\epsilon(\phi(f^*)) d\mathbb{P}_0 \tag{21}$$

2) If we define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ *and* $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$*, then*

$$\eta^*(\mathbf{x})\phi(f^*(\mathbf{x})) + (1 - \eta^*(\mathbf{x}))\phi(-f^*(\mathbf{x})) = C_\phi^*(\eta^*(\mathbf{x})) \quad \mathbb{P}^*\text{-a.e.} \tag{22}$$

This theorem implies that every minimizer f^* of R_ϕ^ϵ and every maximizer $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ of $\bar{R}_\phi$ forms a primal-dual pair. The condition (21) states that every maximizer of $\bar{R}_\phi$ is an optimal adversarial attack on f^* while the condition (22) states that every minimizer

f^* of R^ϵ_ϕ also minimizes the conditional risk $C_\phi(\eta^*, \cdot)$ under the distribution of optimal adversarial attacks. Explicitly: (22) implies that every minimizer f^* minimizes the loss $\hat{R}_\phi(f, \mathbb{P}_0^*, \mathbb{P}_1^*) = \int C(\eta^*(\mathbf{x}), f(\mathbf{x}))d\mathbb{P}^*$ in a pointwise manner $\mathbb{P}^*$-a.e., or n other words, the function f^* minimizes the *standard* surrogate risk with respect to the optimal adversarially perturbed distributions. This fact is the relation (6) with respect to the measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ that maximize the dual $\bar{R}_\phi$.

This interpretation of Theorems 6 and 7 shed light on the phenomenon of transfer attacks. These theorems suggests that adversarial examples are a property of the data distribution rather than a specific model. Later results in the paper even show that there are maximizers of $\bar{R}_\phi$ that are independent of the choice of loss function ϕ (see Lemma 26). Theorem 7 specifically states that every maximizer of $\bar{R}_\phi$ is actually an optimal adversarial attack on *every* minimizer of R^ϵ_ϕ. Notably, this statement is *indepent of the choice of minimizer of* R^ϵ_ϕ. Because neural networks are highly expressive model classes, one would hope that some neural net could achieve adversarial error close to $\inf_f R^\epsilon_\phi(f)$. If f^* is a minimizer of R^ϵ_ϕ and g is a neural net with $R^\epsilon_\phi(g) \approx R^\epsilon_\phi(f^*)$, one would expect that an optimal adversarial attack against f^* would be a successful attack on g as well. Notice that in this discussion, the adversarial attack is independent of the neural net g. A method for calculating these optimal adversarial attacks is an open problem.

Finally, to demonstrate that Theorem 7 and the preceding discussion is non-vacuous, we prove the existence of primal and dual optimizers along with results that elaborate on their structure.

Theorem 8 *Let ϕ be a lower-semicontinuous loss function. Then there exists a maximizer $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ to $\bar{R}_\phi$ over the set $\mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \times \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$.*

Theorem 3.5 of (Jylhä, 2014) implies that when the norm $\|\cdot\|$ is strictly convex and $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous with respect to Lebesgue measure, the optimal $\mathbb{P}_0^*, \mathbb{P}_1^*$ of Theorem 8 are induced by a transport map. Corollary 3.11 of (Jylhä, 2014) further implies that these transport maps are continuous a.e. with respect to the Lebesgue measure μ. As the ℓ_∞ metric is commonly used in practice, whether there exist maximizers of the dual of this type for non-strictly convex norms remains an attractive open problem.

In analogy with (6) and (9) one would hope that due to the complimentary slackness condition (22), one could define a minimizer in terms of the conditional $\eta^*(\mathbf{x})$. Notice, however, that as this quantity is only defined $\mathbb{P}^*$-a.e., verifying the other complimentary slackness condition (21) would be a challenge. To circumvent this issue, we construct a function $\hat{\eta} : \mathbb{R}^d \to [0, 1]$, defined on all of $\mathbb{R}^d$, to which we can apply (9). Concretely, we show that $\alpha_\phi(\hat{\eta}(\mathbf{x}))$ is always a minimizer of R^ϵ_ϕ, with α_ϕ as defined in (8).

Theorem 9 *There exists a Borel function $\hat{\eta} : (\operatorname{supp} \mathbb{P})^\epsilon \to [0, 1]$ for which $f^*(\mathbf{x}) = \alpha_\phi(\hat{\eta}(\mathbf{x}))$ is a minimizer of R^ϵ_ϕ for any ϕ with α_ϕ as in defined in* (8)*. In particular, there exists a Borel minimizer of R^ϵ_ϕ.*

In fact, we show that $\hat{\eta}$ is a version of the conditional derivative $d\mathbb{P}_1^*/d\mathbb{P}^*$, where $\mathbb{P}_0^*, \mathbb{P}_1^*$ are the measures which maximize $\bar{R}_\phi$ independently of the choice of ϕ (see Lemma 24), as described in the discussion preceding Theorem 8. The fact that the function $\hat{\eta}$ is independent of the choice of loss ϕ suggests that the minimizer of R^ϵ_ϕ encodes some fundamental quality of the distribution $\mathbb{P}_0, \mathbb{P}_1$.

Simultaneous work (Li and Telgarsky, 2023) also proves the existence of a minimizer to the primal R_ϕ^ϵ along with a statement on the structure of this minimizer. Their approach leverages prior results on the adversarial Bayes classifier to construct a minimizer to the adversarial surrogate risk.

4.2 Outline of Main Argument

The central proof strategy of this paper is to apply the Fenchel-Rockafellar duality theorem. This classical result relates the infimum of a convex functional with the supremum of a concave functional. One can argue that $\bar{R}_\phi$ is concave (Lemma 12 below); however, the primal R_ϕ^ϵ is not convex for non-convex ϕ. Thus the Fenchel-Rockafellar theorem is applied to a convex relaxation Θ of the primal R_ϕ^ϵ.

The remainder of the paper then argues that minimizing Θ is equivalent to minimizing R_ϕ^ϵ. Notice that the Fenchel-Rockafellar theorem actually implies the existence of dual maximizers. We show that that dual maximizers of $\bar{R}_\psi$ for $\psi(\alpha) = e^{-\alpha}$ satisfy certain nice properties that are *independent* of the loss ψ. These properties then allow us to construct the function $\hat{\eta}$ present in Theorem 9 in addition to minimizers of Θ from the dual maximizers of $\bar{R}_\psi$, for any loss ϕ. The construction of these minimizers guarantee that they minimize R_ϕ^ϵ in addition to Θ.

4.3 Paper Outline

Section 5 proves strong duality and complimentary slackness theorems for $\bar{R}_\phi$ and Θ, the convex relaxation of R_ϕ^ϵ. Next, in Section 6, a version of the complimentary slackness result is used to prove the existence of minimizers to Θ. Subsequently, Section 7 shows the equivalence between Θ and R_ϕ^ϵ.

Appendix A proves Theorem 1 and further discusses universal measurability. Next, Appendix B proves all the results about the W_∞-norm used in this paper. Appendix C then defines the function α_ϕ which is later used in the proof of several results. Appendices D, E, F, and G.3 contain technical deferred proofs.

5. A Duality Result for Θ and $\bar{R}_\phi$

5.1 Strong Duality

The fundamental duality relation of this paper follows from employing the Fenchel-Rockafellar theorem in conjunction with the Riesz representation theorem, stated below for reference. See e.g. (Villani, 2003) for more on this result.

Theorem 10 (Fenchel-Rockafellar Duality Theorem) *Let E be a normed vector space E^* its topological dual and Θ, Ξ two convex functionals on E with values in $\mathbb{R} \cup \{\infty\}$. Let Θ^*, Ξ^* be the Legendre-Fenchel transforms of Θ, Ξ respectively. Assume that there exists $z_0 \in E$ such that*

$$\Theta(z_0) < \infty, \Xi(z_0) < \infty$$

and that Θ is continuous at z_0. Then

$$\inf_{z \in E} [\Theta(z) + \Xi(z)] = \sup_{z^* \in E^*} [-\Theta^*(z^*) - \Xi^*(-z^*)] \tag{23}$$

and furthermore, the supremum on the right hand side is attained.

Let $\mathcal{M}(X)$ be the set of finite signed Borel measures on a space X and recall that $C_b(X)$ is the set of bounded continuous functions on the space X.

Theorem 11 (Riesz Representation Theorem) *Let K be any compact subset of $\mathbb{R}^d$. Then the dual of $C_b(K)$ is $\mathcal{M}(K)$.*

See Theorem 1.9 of (Villani, 2003) and result 7.17 of (Folland, 1999) for more details.

Notice that in the Fenchel-Rockafellar theorem, the left-hand side of (23) is convex while the right-hand side is concave. However, when ϕ is non-convex, R_ϕ^ϵ is not convex. In order to apply the Fenchel-Rockafellar theorem, we will relax the primal R_ϕ^ϵ will to a convex functional Θ.

We define Θ as

$$\Theta(h_0, h_1) = \int S_\epsilon(h_1) d\mathbb{P}_1 + \int S_\epsilon(h_0) d\mathbb{P}_0 \tag{24}$$

which is convex in h_0, h_1 due to the sub-additivity of the supremum operation. Notice that one obtains Θ from R_ϕ^ϵ by replacing $\phi \circ f$ with h_1 and $\phi \circ -f$ with h_0.

The functional Ξ will be chosen to restrict h_0, h_1 in the hope that at the optimal value, $h_1 = \phi(f)$ and $h_0 = \phi(-f)$ for some f. Notice that if $h_1 = \phi(f)$, $h_0 = \phi(-f)$ then for all $\eta \in [0,1]$,

$$\eta h_1(\mathbf{x}) + (1-\eta) h_0 = \eta\phi(f)) + (1-\eta)\phi(-f) \geq C_\phi^*(\eta).$$

Thus we will optimize Θ over the set of functions S_ϕ defined by

$$S_\phi = \begin{Bmatrix} (h_0, h_1) \colon h_0, h_1 \colon K^\epsilon \to \overline{\mathbb{R}} \text{ Borel, } 0 \leq h_0, h_1 \text{ and for} \\ \text{all } \mathbf{x} \in \mathbb{R}^d,\ \eta \in [0,1],\ \eta h_0(\mathbf{x}) + (1-\eta) h_1(\mathbf{x}) \geq C_\phi^*(\eta) \end{Bmatrix} \tag{25}$$

where $K = \operatorname{supp}(\mathbb{P}_0 \cup \mathbb{P}_1)$. (Notice that the definition of $S_\epsilon(g)$ in (11) assumes that the domain of g must include $\overline{B_\epsilon(\mathbf{x})}$. Thus in order to define the integral $\int S_\epsilon(h) d\mathbb{Q}$, the domain of h must include $(\operatorname{supp}\mathbb{Q})^\epsilon$.) Thus we define Ξ as

$$\Xi(h_0, h_1) = \begin{cases} 0 & \text{if } (h_0, h_1) \in S_\phi \\ +\infty & \text{otherwise} \end{cases} \tag{26}$$

The following result expresses $\bar{R}_\phi$ as an infimum of linear functionals continuous with respect to the weak topology on probability measures. This lemma will assist in the computation of Ξ^*. In the remainder of this section, $\mathcal{M}_+(S)$ will denote the set of positive finite Borel measures on a set S.

Lemma 12 *Let $K \subset \mathbb{R}^d$ be compact, $E = C_b(K^\epsilon) \times C_b(K^\epsilon)$, and $\mathbb{P}_0', \mathbb{P}_1' \in \mathcal{M}_+(K^\epsilon)$. Then*

$$\inf_{(h_0, h_1) \in S_\phi \cap E} \int h_1 d\mathbb{P}_1' + \int h_0 d\mathbb{P}_0' = \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') \tag{27}$$

Therefore, $\bar{R}_\phi$ is concave and upper semi-continuous on $\mathcal{M}_+(K^\epsilon) \times \mathcal{M}_+(K^\epsilon)$ with respect to the weak topology on probability measures.

We sketch the proof and formally fill in the details in Appendix D. Let $\mathbb{P}' = \mathbb{P}_0' + \mathbb{P}_1'$, $\eta' = d\mathbb{P}_1'/d\mathbb{P}'$. Then

$$\int h_1 d\mathbb{P}_1' + \int h_0 d\mathbb{P}_0' = \int \eta' h_1 + (1-\eta')h_0 d\mathbb{P}'$$

Clearly, the inequality $\geq$ holds because $\eta' h_1 + (1-\eta')h_0 \geq C_\phi^*(\eta')$ for all $(h_0, h_1) \in S_\phi$. Equality is achieved at $h_1 = \phi(\alpha_\phi(\eta')), h_0 = \phi(-\alpha_\phi(\eta'))$, with α_ϕ as in (8). However, these functions may not be continuous. In Appendix D, we show that h_0, h_1 can be approximated arbitrarily well by elements of $S_\phi \cap E$.

Lemma 13 *Let ϕ be a non-increasing, lower semi-continuous loss function and let $\mathbb{P}_0, \mathbb{P}_1$ be compactly supported finite Borel measures. Let S_ϕ be as in* (25)*.*

Then

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0,h_1) = \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') \tag{28}$$

Furthermore, there exist $\mathbb{P}_0^*, \mathbb{P}_1^*$ which attain the supremum.

Proof We will show a version of (28) with the infimum taken over $S_\phi \cap E$, and then argue that the same claim holds when the infimum is taken over S_ϕ.

Notice that if h_0, h_1 are continuous, then $S_\epsilon(h_0)$, $S_\epsilon(h_1)$ are also continuous and $\int S_\epsilon(h_0)d\mathbb{Q}$, $\int S_\epsilon(h_1)d\mathbb{Q}$ are well-defined for every *Borel* $\mathbb{Q}$. Hence we assume that $\mathbb{P}_0$, $\mathbb{P}_1$ are Borel measures rather than their completion.

Let $K = \text{supp}(\mathbb{P}_0 + \mathbb{P}_1)$. We will apply the Fenchel-Rockafellar Duality Theorem to the functionals given by (24) and (26) on the vector space $E = C_b(K^\epsilon) \times C_b(K^\epsilon)$ equipped with the sup norm. By the Riesz representation theorem, dual of the space E is $E^* = \mathcal{M}(K^\epsilon) \times \mathcal{M}(K^\epsilon)$.

To start, we argue that the Fenchel-Rockafellar duality theorem applies to these functionals. First, notice that if $(h_0, h_1) \in E$, then both h_0, h_1 are bounded so $\Theta(h_0, h_1) < \infty$. Furthermore, Θ is convex due to the subadditivity of supremum and Ξ is convex because the constraint $h_0(\mathbf{x}) + (1-\eta)h_1(\mathbf{x}) \geq C_\phi^*(\eta)$ is linear in h_0 and h_1. Furthermore, Θ is continuous on E because Θ is convex and bounded and E is open, see Theorem 2.14 of (Barbu and Precupanu, 2012).

Because the constant function $(C_\phi^*(1/2), C_\phi^*(1/2))$ is in S_ϕ, Ξ is not identically ∞ and therefore the Fenchel-Rockafellar theorem applies.

Clearly $\inf_E \Theta(h_0, h_1) + \Xi(h_0, h_1)$ reduces to the left-hand side of (28).

We now compute the dual of Ξ. Lemma 12 implies that

$$-\Xi^*(-\mathbb{P}_0', -\mathbb{P}_1') = -\sup_{(h_0,h_1)\in S_\phi\cap E} -\int h_0 d\mathbb{P}_0' - \int h_1 d\mathbb{P}_1' \tag{29}$$

$$= \begin{cases} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') & \text{if } \mathbb{P}_i' \geq 0 \\ -\infty & \text{otherwise} \end{cases}$$

This computation implies that the term $-\Xi^*(-\mathbb{P}_0', -\mathbb{P}_1')$ present in the Fenchel-Rockafellar Theorem is not $-\infty$ iff $\mathbb{P}_0', \mathbb{P}_1'$ are positive measures. Next, notice that because $\Theta(h_0, h_1) <$

$+\infty$ for all $(h_0,h_1)\in E$, $-\Theta^*(\mathbb{P}_0',\mathbb{P}_1')$ is never $+\infty$. Therefore, it suffices to compute Θ^* for positive measures $\mathbb{P}_0',\mathbb{P}_1'$. Lemma 4 implies that for positive measures $\mathbb{P}_0',\mathbb{P}_1'$,

$$\Theta^*(\mathbb{P}_0',\mathbb{P}_1') = \sup_{h_0,h_1\in C_0(K^\epsilon)} \int h_1 d\mathbb{P}_1' + \int h_0 d\mathbb{P}_0' - \left(\int S_\epsilon(h_0)d\mathbb{P}_0 + \int S_\epsilon(h_1)d\mathbb{P}_1\right)$$

$$= \sup_{h_1\in C_0(K^\epsilon)} \left(\int h_1 d\mathbb{P}_1' - \int S_\epsilon(h_1)d\mathbb{P}_1\right) + \sup_{h_0\in C_0(K^\epsilon)} \left(\int h_0 d\mathbb{P}_0' - \int S_\epsilon(h_0)d\mathbb{P}_0 +\right)$$

$$= \begin{cases} 0 & \mathbb{P}_0',\mathbb{P}_1' \text{ positive measures, with } W_\infty(\mathbb{P}_0',\mathbb{P}_0)\leq\epsilon \text{ and } W_\infty(\mathbb{P}_1',\mathbb{P}_1)\leq\epsilon \\ +\infty & \mathbb{P}_0',\mathbb{P}_1' \text{ positive measures, with either } W_\infty(\mathbb{P}_0',\mathbb{P}_0)>\epsilon \text{ or } W_\infty(\mathbb{P}_1',\mathbb{P}_1)>\epsilon \end{cases}$$

Therefore

$$\sup_{\mathbb{P}_0',\mathbb{P}_1'\in\mathcal{M}(K^\epsilon)} -\Theta(\mathbb{P}_0',\mathbb{P}_1') - \Xi(-\mathbb{P}_0',-\mathbb{P}_1') = \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0',\mathbb{P}_1')$$

and furthermore there exist measures $\mathbb{P}_0^*,\mathbb{P}_1^*$ maximizing the dual problem. Therefore the Fenchel-Rockafellar Theorem implies that

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0,h_1) \leq \inf_{(h_0,h_1)\in S_\phi\cap E} \Theta(h_0,h_1) = \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0',\mathbb{P}_1')$$

The opposite inequality follows from the weak duality argument presented in (16) in Section 4.1. See Lemma 45 of Appendix E for a full proof. ■

Note that this proof does not easily extend to an unbounded domain X: for a non-compact space, the dual of $C_b(X)$ is much larger than $\mathcal{M}(X)$, and thus a naive application of the Fenchel-Rockafellar Theorem would result in a different right-hand side than (28). On the other hand, the Reisz representation theorem for an unbounded domain X states that the dual of $C_0(X)$ is $\mathcal{M}(X)$, where $C_0(X)$ is the set of continuous bounded functions vanishing at ∞. At the same time, if $h_0,h_1\in C_0(X)$, then $\eta h_1(\mathbf{x})+(1-\eta)h_0(\mathbf{x})$ becomes arbitrarily small for large $\mathbf{x}$ so the constraint $\eta h_1(\mathbf{x})+(1-\eta)h_0(\mathbf{x})\geq C_\phi^*(\eta)$ cannot hold for all η. Thus if K is unbounded, $S_\phi\cap C_0(X)=\emptyset$ and the functional Ξ would be $+\infty$ everywhere on $C_0(X)$, precluding and application of the Fenchel-Rockafellar Theorem.

However, Lemma 13 can be extended to distributions with arbitrary support via a simple approximation argument. By Lemma 13, the strong duality result holds for the distributions defined by $\mathbb{P}_0^n=\mathbb{P}_0\big|_{\overline{B_n(\mathbf{0})}}, \mathbb{P}_1^n=\mathbb{P}_1\big|_{\overline{B_n(\mathbf{0})}}$. One then shows strong duality by computing the limit of the primal and dual problems as $n\to\infty$. We therefore obtain the following Lemma, which is proved formally in Appendix E.

Lemma 14 *Let ϕ be a non-increasing, lower semi-continuous loss function and let $\mathbb{P}_0,\mathbb{P}_1$ be finite Borel measures supported on $\mathbb{R}^d$. Let S_ϕ be as in* (25)*. Then*

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0,h_1) = \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0',\mathbb{P}_1')$$

Furthermore, there exist $\mathbb{P}_0^*,\mathbb{P}_1^*$ which attain the supremum.

5.2 Complimentary Slackness

Using a standard argument, strong duality (Lemma 14) allows us to prove a version of the complimentary slackness theorem.

Lemma 15 *Assume that $\mathbb{P}_0, \mathbb{P}_1$ are compactly supported. The functions h_0^*, h_1^* minimize Θ over S_ϕ and $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ maximize $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ iff the following hold:*

1)

$$\int h_1^* d\mathbb{P}_1^* = \int S_\epsilon(h_1^*) d\mathbb{P}_1 \quad \text{and} \quad \int h_0^* d\mathbb{P}_0^* = \int S_\epsilon(h_0^*) d\mathbb{P}_0 \tag{30}$$

2) If we define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$, then

$$\eta^*(\mathbf{x}) h_1^*(\mathbf{x}) + (1 - \eta^*(\mathbf{x})) h_0^*(\mathbf{x}) = C_\phi^*(\eta^*(\mathbf{x})) \quad \mathbb{P}^*\text{-a.e.} \tag{31}$$

This lemma is proved in Appendix F. Theorem 7 will later follow from this result.

To show that Lemma 15 is non-vacuous, one must prove that there exist minimizers to Θ over S_ϕ, which we delay to Sections 6 and 7. Notice that the application of the Fenchel-Rockafellar Theorem in Lemma 13 actually implies the existence of dual maximizers in the case of compactly supported $\mathbb{P}_0, \mathbb{P}_1$.

In fact, the complimentary slackness conditions hold approximately for any maximizer of $\bar{R}_\phi$ and any minimizing sequence of Θ. This result is essential for proving the existence of minimizers to Θ.

Lemma 16 *Let (h_0^n, h_1^n) be a minimizing sequence for Θ over S_ϕ: $\lim_{n\to\infty} \Theta(h_0^n, h_1^n) = \inf_{(h_0,h_1)\in S_\phi} \Theta(h_0, h_1)$. Then for any maximizer of the dual problem $(\mathbb{P}_0^*, \mathbb{P}_1^*)$, the following hold:*

1)

$$\lim_{n\to\infty} \int S_\epsilon(h_0^n) d\mathbb{P}_0 - \int h_0^n d\mathbb{P}_0^* = 0, \quad \lim_{n\to\infty} \int S_\epsilon(h_1^n) d\mathbb{P}_1 - \int h_1^n d\mathbb{P}_1^* = 0 \tag{32}$$

2) If we define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$

$$\lim_{n\to\infty} \int \eta^* h_1^n + (1-\eta^*) h_0^n - C_\phi^*(\eta^*) d\mathbb{P}^* = 0 \tag{33}$$

Proof Let

$$m = \inf_{(h_0,h_1)\in S_\phi} \Theta(h_0, h_1).$$

Then the fact that $(h_0^n, h_1^n) \in S_\phi$ and the duality result (Lemma 14) implies

$$\int h_1^n d\mathbb{P}_1^* + \int h_0^n d\mathbb{P}_0^* = \int \eta^* h_1^n + (1-\eta^*) h_0^n d\mathbb{P}^* \geq \int C_\phi^*(\eta^*) d\mathbb{P}^* = m \tag{34}$$

Now pick $\delta > 0$ and an N for which $n \geq N$ implies that $\Theta(h_0^n, h_1^n) \leq m + \delta$. Then

$$m + \delta \geq \int S_\epsilon(h_1^n) d\mathbb{P}_1 + \int S_\epsilon(h_0^n) d\mathbb{P}_0 \geq \int \eta^* h_1^n + (1-\eta^*) h_0^n d\mathbb{P}^* \geq m.$$

Subtracting $m = \int C_\phi^*(\eta^*) d\mathbb{P}^*$ from this inequality results in

$$\delta \geq \int \eta^* h_1^n + (1-\eta^*) h_0^n d\mathbb{P}^* - \int C_\phi^*(\eta^*) d\mathbb{P}^* \geq 0 \tag{35}$$

which implies (33). Next, (34) further implies

$$m - \int h_1^n d\mathbb{P}_1^* + \int h_0^n d\mathbb{P}_0^* \leq 0 \tag{36}$$

Now this inequality implies

$$\begin{aligned} \delta &\geq \delta + m - \left(\int h_1^n d\mathbb{P}_1^* + \int h_0^n d\mathbb{P}_0^* \right) \geq \Theta(h_1^n, h_0^n) - \left(\int h_1^n d\mathbb{P}_1^* + \int h_0^n d\mathbb{P}_0^* \right) \\ &\geq \left(\int S_\epsilon(h_1^n) d\mathbb{P}_1 + \int S_\epsilon(h_0^n) d\mathbb{P}_0 \right) - \left(\int h_1^n d\mathbb{P}_1^* + \int h_0^n d\mathbb{P}_0^* \right) \end{aligned}$$

However, Lemma 3 implies that both $\int S_\epsilon(h_1^n) d\mathbb{P}_1 - \int h_1^n d\mathbb{P}_1^*, \int S_\epsilon(h_0^n) d\mathbb{P}_0 - \int h_0^n d\mathbb{P}_0^*$ are positive quantities. Therefore, we have shown that for any $\delta > 0$, there is an N for which $n \geq N$ implies that

$$\delta > \int S_\epsilon(h_1^n) d\mathbb{P}_1 - \int h_1^n d\mathbb{P}_1^* \geq 0 \quad \text{and} \quad \delta > \int S_\epsilon(h_0^n) d\mathbb{P}_0 - \int h_0^n d\mathbb{P}_0^* \geq 0$$

which implies (32).

■

An analogous approximate complimentary slackness result typically holds in other applications of the Fenchel-Rockafellar theorem. Consider a convex optimization problem which can be written as $\inf_x \Theta(x) + \Xi(x)$ in such a way that the Fenchel-Rockafellar theorem applies and for which Ξ and Θ^* are indicator functions of the convex sets C_P, C_D respectively. Then the Fenchel-Rockafellar Theorem states that

$$\inf_{x \in C_P} \Theta(x) = \inf_{x \in C_p} \sup_{y \in C_D} \langle y, x \rangle = \sup_{y \in C_D} \inf_{x \in C_P} \langle y, x \rangle = \sup_{y \in C_D} \Xi^*(y) \tag{37}$$

Let y^* be a maximizer of the dual problem and let m be the minimal value of Θ over C_P. If x_k is a minimizing sequence of Θ, then for $\delta > 0$ and sufficiently large k, $\delta + m > \Theta(x_k)$ and thus by (37),

$$m + \delta > \Theta(x_k) = \sup_{y \in C_p} \langle y, x_k \rangle \geq \langle y^*, x_k \rangle \geq \inf_{x \in C_D} \langle y^*, x \rangle = \inf_{x \in C_D} \Xi^*(x) = m \tag{38}$$

and therefore $\lim_{k \to \infty} \langle y^*, x_k \rangle = m$. Condition (31) is this statement adapted to the adversarial learning problem. Furthermore, subtracting $\Theta(x_k)$ from (38) and taking the limit $k \to \infty$ results in $\lim_{k \to \infty} \Theta(x_k) - \langle y^*, x_k \rangle = 0$. In our adversarial learning scenario, this statement is equivalent to the conditions in (32) due to Lemma 3. Furthermore, just like the standard complimentary slackness theorems, the relations $\lim_{k \to \infty} \langle y^*, x_k \rangle = m$, $\lim_{k \to \infty} \Theta(x_k) - \langle y^*, x_k \rangle = 0$ imply that x_k is a minimizing sequence for Θ.

In the classical proof of the Kantorovich duality, one can choose Θ, Ξ of a form similar to the discussion above, see for instance Theorem 1.3 of Villani (2003). Using an argument similar to (38), one can prove approximate complimentary slackness for the Kantorovich problem called the quantitative Knott-Smith criteria, see Theorems 2.15, 2.16 of Villani (2003) for further discussion.

6. Existence of Minimizers of Θ over S_ψ

After proving the existence of maximizers to the dual problem, we can now use the approximate complimentary slackness conditions to prove the existence of minimizers to the primal. The exponential loss ψ has certain properties that make it particularly easy to study:

Lemma 17 *Let $\psi(\alpha) = e^{-\alpha}$. Then $C_\psi^*(\eta) = 2\sqrt{\eta(1-\eta)}$ and $\alpha_\psi(\eta) = 1/2\log(\eta/1-\eta)$ is the unique minimizer of $C_\psi(\eta,\cdot)$, with $\alpha_\psi(0), \alpha_\psi(1)$ interpreted as $-\infty$, $+\infty$ respectively. Furthermore, $\partial C_\psi^*(\eta)$ is the singleton $\partial C_\psi^*(\eta) = \{\psi(\alpha_\psi(\eta)) - \psi(-\alpha_\psi(\eta))\}$.*

See Appendix G.1 for a proof. The existence of minimizers of Θ for the exponential loss then follows from properties of C_ψ. Let (h_0^n, h_1^n) be a minimizing sequene of $\bar{R}_\phi$. Because the function C_ψ is strictly concave, one can use the condition (33) to show that there is a subsequence $\{n_k\}$ along which $h_0^{n_k}(\mathbf{x}')$, $h_1^{n_k}(\mathbf{x}')$ converge $\mathbb{P}_0^*, \mathbb{P}_1^*$-a.e. respectively. Due to (32), $S_\epsilon(h_0^{n_k})(\mathbf{x})$, $S_\epsilon(h_1^{n_k})$ also converge $\mathbb{P}_0, \mathbb{P}_1$-a.e. respectively along this subsequence. This observation suffices to show the existence of functions that minimize Θ over S_ψ.

The first step of this proof is to formalize this argument for sequences in $\overline{\mathbb{R}}$.

Lemma 18 *Let (a_n, b_n) be a sequence for which $a_n, b_n \geq 0$ and*

$$\eta a_n + (1-\eta) b_n \geq C_\psi^*(\eta) \text{ for all } \eta \in [0,1] \tag{39}$$

and

$$\lim_{n\to\infty} \eta_0 a_n + (1-\eta_0) b_n = C_\psi^*(\eta_0) \tag{40}$$

for some η_0. Then $\lim_{n\to\infty} a_n = \psi(\alpha_\psi(\eta_0))$ and $\lim_{n\to\infty} b_n = \psi(-\alpha_\psi(\eta_0))$.

Notice that if $\eta a + (1-\eta) b \geq C_\psi^*(\eta)$ and $\eta_0 a + (1-\eta_0) b = C_\psi^*(\eta_0)$, then this lemma implies that $a = \psi(\alpha_\psi(\eta_0))$ and $b = \psi(-\alpha_\psi(\eta_0))$.

To prove Lemma 18, we show that all convergent subsequences of $\{a_n\}$ and $\{b_n\}$ must converge to a and b that satisfy $\eta_0 a + (1-\eta_0) b = C_\phi^*(\eta_0)$ and $a - b \in \partial C_\psi^*(\eta_0)$. As the set $\partial C_\psi^*(\eta_0)$ is a singleton, the values $a = \psi(\alpha_\psi(\eta_0))$ and $b = \psi(\alpha_\psi(\eta_0))$ uniquely solve these equations for a and b. Therefore the sequences $\{a_n\}$ and $\{b_n\}$ must converge to a and b as well. See Appendix G.2 for a formal proof. This result applied to a minimizing sequence of Θ allows one to find a subsequence with certain convergence properties.

Lemma 19 *Let (h_0^n, h_1^n) be a minimizing sequence of Θ over S_ψ. Then there exists a subsequence n_k for which $S_\epsilon(h_1^{n_k})$, $S_\epsilon(h_0^{n_k})$ converge $\mathbb{P}_1$, $\mathbb{P}_0$-a.e. respectively.*

Proof Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be maximizers of the dual problem. Let γ_i be the coupling between $\mathbb{P}_i, \mathbb{P}_i^*$ with $\operatorname{supp}\gamma_i \subset \Delta_\epsilon$.

Then (33) of Lemma 16 implies that

$$\lim_{n\to\infty} \int \eta^*(\mathbf{x}') h_1^n(\mathbf{x}') + (1-\eta^*(\mathbf{x}')) h_0^n(\mathbf{x}') - C_\psi(\eta^*(\mathbf{x}')) d(\gamma_1+\gamma_0)(\mathbf{x},\mathbf{x}') = 0$$

and (32) implies that

$$\lim_{n\to\infty} \int S_\epsilon(h_1^n)(\mathbf{x}) - h_1^n(\mathbf{x}') d\gamma_1(\mathbf{x},\mathbf{x}') = 0, \quad \lim_{n\to\infty} \int S_\epsilon(h_0^n)(\mathbf{x}) - h_0^n(\mathbf{x}') d\gamma_0(\mathbf{x},\mathbf{x}') = 0$$

Recall that on a bounded measure space, L^1 convergence implies a.e. convergence along a subsequence (see Corollary 2.32 of (Folland, 1999)). Thus one can pick a subsequence n_k along which

$$\lim_{k\to\infty} \eta^*(\mathbf{x}')h_1^{n_k}(\mathbf{x}') + (1-\eta^*(\mathbf{x}'))h_0^{n_k}(\mathbf{x}') - C_\psi(\eta^*(\mathbf{x}')) = 0 \tag{41}$$

$\gamma_1 + \gamma_0$-a.e. and

$$\lim_{k\to\infty} S_\epsilon(h_1^{n_k})(\mathbf{x}) - h_1^{n_k}(\mathbf{x}') = 0, \quad \lim_{k\to\infty} S_\epsilon(h_0^{n_k})(\mathbf{x}) - h_0^{n_k}(\mathbf{x}') = 0 \tag{42}$$

γ_1, γ_0-a.e. respectively.

Furthermore, $\eta h_1^n + (1-\eta)h_0^n \geq C_\psi^*(\eta)$ for all $\eta \in [0,1]$. Thus (41) and Lemma 18 imply that $h_{n_k}^1$ converges to $\psi(\alpha_\psi(\eta^*))$ and $h_{n_k}^0$ converges to $\psi(-\alpha_\psi(\eta^*))$ $\gamma_0+\gamma_1$-a.e. Equation 42 then implies that $S_\epsilon(h_1^{n_k})(\mathbf{x})$, $S_\epsilon(h_0^{n_k})(\mathbf{x})$ converge γ_1, γ_0 -a.e. respectively. Because $\mathbb{P}_1, \mathbb{P}_0$ are marginals of γ_1, γ_0, this statement implies the result. ■

The existence of a minimizer then follows from the fact that $S_\epsilon(h_1^{n_k})$ converges. The next lemma describes how the S_ϵ operation interacts with lim infs and lim sups.

Lemma 20 *Let h_n be any sequence of functions. Then the sequence h_n satisfies*

$$\liminf_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\liminf_{n\to\infty} h_n) \tag{43}$$

and

$$\limsup_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\limsup_{n\to\infty} h_n) \tag{44}$$

See Appendix G.3 for a proof.

Finally, we prove that there exists a minimizer to Θ over S_ψ.

Lemma 21 *There exists a minimizer (h_0^*, h_1^*) to Θ over the set S_ψ.*

Proof Let (h_0^n, h_1^n) be a sequence minimizing Θ over S_ψ.

Lemma 19 implies that there is a subsequence $\{n_k\}$ for which $\lim_{k\to\infty} S_\epsilon(h_0^{n_k})$ exists $\mathbb{P}_0$-a.e.

Thus

$$\limsup_{k\to\infty} S_\epsilon(h_0^{n_k}) = \liminf_{k\to\infty} S_\epsilon(h_0^{n_k}) \quad \mathbb{P}_0\text{-a.e.} \tag{45}$$

Next, we will argue that the pair $(\limsup_k h_0^{n_k}, \liminf_k h_1^{n_k})$ is in S_ψ. Because

$$C_\psi^*(\eta) \leq \eta h_1^{n_k} + (1-\eta)h_0^{n_k},$$

one can conclude that

$$C_\psi^*(\eta) \leq \eta \liminf_{k\to\infty}(h_1^{n_k} + (1-\eta)h_0^{n_k}) \leq \eta \liminf_{k\to\infty} h_1^{n_k} + (1-\eta)\limsup_{k\to\infty} h_0^{n_k}.$$

Define

$$h_1^* = \liminf_k h_1^{n_k}, \quad h_0^* = \limsup_k h_0^{n_k}$$

Now Fatou's lemma, Lemma 20, and Equation 45 imply that

$$\lim_{k\to\infty} \Theta(h_0^{n_k}, h_1^{n_k}) \geq \int \liminf_{k\to\infty} S_\epsilon(h_1^{n_k}) d\mathbb{P}_1 + \int \liminf_{k\to\infty} S_\epsilon(h_0^{n_k}) d\mathbb{P}_0 \qquad \text{(Fatou's Lemma)}$$
$$= \int \liminf_{k\to\infty} S_\epsilon(h_1^{n_k}) d\mathbb{P}_1 + \int \limsup_{k\to\infty} S_\epsilon(h_0^{n_k}) d\mathbb{P}_0 \qquad \text{(Equation 45)}$$
$$\geq \int S_\epsilon(\liminf_{k\to\infty} h_1^{n_k}) d\mathbb{P}_1 + \int S_\epsilon(\limsup_{k\to\infty} h_0^{n_k}) d\mathbb{P}_0 \qquad \text{(Lemma 20)}$$
$$= \int S_\epsilon(h_1^*) d\mathbb{P}_1 + \int S_\epsilon(h_0^*) d\mathbb{P}_0$$

Therefore, (h_0^*, h_1^*) must be a minimizer.

■

7. Reducing Θ to R_ϕ^ϵ

Using the properties of $C_\psi^*(\eta)$, we showed in the previous section that there exist minimizers to Θ over the set S_ψ. The inequality $\eta h_1^* + (1-\eta^*)h_0^* \geq C_\psi^*(\eta)$ together with (31) imply that $h_1^*(\mathbf{x}) - h_0^*(\mathbf{x})$ is a supergradient of $C_\psi^*(\eta^*(\mathbf{x}))$ and thus $h_1^* - h_0^* = (C_\psi^*)'(\eta)$. This relation together with (31) provides two equations in two variables that can be uniquely solved for h_0^*, h_1^*, resulting in $h_0^* = \psi \circ -\alpha_\psi(\eta^*), h_1^* = \psi \circ \alpha_\psi(\eta^*)$.

Next, primal minimizers of Θ over S_ϕ for *any* ϕ will be constructed from the dual maximizers $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ of $\bar{R}_\psi$. Because $\alpha_\psi(\eta) = 1/2\log(\eta/1-\eta)$ is a strictly increasing function, the compositions $\psi \circ \alpha_\psi, \psi \circ -\alpha_\psi$ are strictly monotonic. Thus the complimentary slackness condition (30) applied to $h_1^* = \psi(\alpha_\psi(\eta^*)), h_0^* = \psi(-\alpha_\psi(\eta^*))$ implies that $\operatorname{supp}\mathbb{P}_1^*$ is contained in the set of points $\mathbf{x}'$ for which η^* assumes its infimum over some ϵ-ball at $\mathbf{x}'$ and $\operatorname{supp}\mathbb{P}_0^*$ is contained in the set of points $\mathbf{x}'$ where η^* assumes its supremum over some ϵ-ball at $\mathbf{x}'$. Thus, the functions $\phi \circ \alpha_\phi(\eta^*), \phi \circ -\alpha_\phi(\eta^*)$ satisfy (30) for the loss ϕ. The definition of α_ϕ further implies they satisfy (31). Therefore, Lemma 15 implies that for *any* ϕ, $h_1^* = \phi \circ \alpha_\phi(\eta^*)$, $h_0^* = \phi \circ \alpha_\phi(\eta^*)$ are primal optimal and $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ are dual optimal!

This reasoning about η^* is technically wrong but correct in spirit. Because η^* is a Raydon-Nikodym derivative, it is only defined $\mathbb{P}^*$-a.e. As a result, the supremum over an ϵ-ball of the function $\phi(\alpha_\psi(\eta^*))$ is not well-defined. The solution is to replace η^* in the discussion above by a function $\hat{\eta}$ that is defined everywhere. The function $\hat{\eta}$ is actually a version of the Raydon-Nikodym derivative $d\mathbb{P}_1^*/d\mathbb{P}^*$. The next two lemmas describe how one constructs this function $\hat{\eta}$.

The next two lemmas discuss the analog of the ^{c} transform for the Kantorovich problem in optimal transport (see for instance Chapter 1 of (Santambrogio, 2015) or Section 2.5 of (Villani, 2003)).

Lemma 22 *Assume that $h_0, h_1 \geq 0$ and $(h_0(\mathbf{x}), h_1(\mathbf{x}))$ satisfy $\eta h_1 + (1-\eta)h_0 \geq C_\phi^*(\eta)$ for all η. Then if we define $h_0^{C_\phi^*}$ via*

$$h_0^{C_\phi^*} = \sup_{\eta\in[0,1)} \frac{C_\phi^*(\eta) - \eta h_1}{1-\eta} \tag{46}$$

then $h_0^{C_\phi^*} \leq h_0$ *while and* $h_1+(1-\eta)h_0^{C_\phi^*} \geq C_\phi^*(\eta)$ *for all* η*, and* $h_0^{C_\phi^*}$ *is the smallest function* h_0 *for which* $(h_0,h_1) \in S_\phi$*. Furthermore, the function* $h_0^{C_\phi^*}$ *is Borel and there exists a function* $\bar\eta\colon \mathbb{R}^d \to [0,1]$ *for which* $\bar\eta(\mathbf{x})h_1(\mathbf{x}) + (1-\bar\eta(\mathbf{x}))h_1^{C_\phi^*}(\mathbf{x}) = C_\phi^*(\bar\eta(\mathbf{x}))$*.*

Proof For convenience, set $\tilde h_0 = h_1^{C_\phi^*}$. Notice that $\tilde h_0 \geq 0$ because the right-hand side of (46) evaluates to 0 at $\eta=0$. We will show that $\tilde h_0$ is Borel and that $(\tilde h_0, h_1)$ is a feasible pair.

Next, Notice that the map

$$G(\eta,\alpha) = \begin{cases} \frac{C_\phi^*(\eta)-\eta\alpha}{1-\eta} & \text{if } \eta<1 \\ \lim_{\eta\to 1} \frac{C_\phi^*(\eta)-\eta\alpha}{1-\eta} & \text{if } \eta = 1 \end{cases} \tag{47}$$

is continuous in η. Thus, the supremum in (46) can be taken over the countable set $\mathbb{Q}\cap[0,1]$ and hence the function $\tilde h_0(\mathbf{x}) = \sup_{\eta\in[0,1)\cap\mathbb{Q}} G(\eta, h_1(\mathbf{x}))$ is Borel measurable. Because $G(\eta, h_1(\mathbf{x}))$ is continuous in η for each fixed $\mathbf{x}$, $G(\cdot, h_1(\mathbf{x}))$ assumes its maximum on $\eta\in[0,1]$ for each fixed $\mathbf{x}$. Thus there exists a function $\bar\eta(\mathbf{x})$ that maps $\mathbf{x}$ to a maximizer of $G(\cdot, h_1(\mathbf{x}))$. For this function $\bar\eta(\mathbf{x})$, one can conclude that $\tilde h_0(\mathbf{x}) = G(\bar\eta(\mathbf{x}), \mathbf{x})$ and hence

$$\bar\eta(\mathbf{x})h_1(\mathbf{x}) + (1-\bar\eta(\mathbf{x}))\tilde h_0(\mathbf{x}) = C_\phi^*(\bar\eta(\mathbf{x})). \tag{48}$$

Equation 48 implies that if $f(\mathbf{x}) < \tilde h_0(\mathbf{x})$ at any $\mathbf{x}$, then $\eta h_1(\mathbf{x}) + (1-\eta)f(\mathbf{x}) < C_\phi^*(\eta(\mathbf{x}))$ so (f,h_1) is not in the feasible set S_ϕ. Therefore, $\tilde h_0$ is the smallest function f for which $(f,h_1)\in S_\phi$. ■

Next we use this result to define an extension of η^* to all of $\mathbb{R}^d$.

Lemma 23 *There exist a Borel minimizer* (h_0^*, h_1^*) *to* Θ *over* S_ψ *for which*

$$\hat\eta(\mathbf{x})h_1^*(\mathbf{x}) + (1-\hat\eta(\mathbf{x}))h_0^*(\mathbf{x}) = C_\psi^*(\hat\eta(\mathbf{x})) \tag{49}$$

for all $\mathbf{x}$ *and some Borel measurable function* $\hat\eta\colon (\operatorname{supp}\mathbb{P})^\epsilon \to [0,1]$*.*

Proof Let (h_0,h_1), be an arbitrary Borel minimizer to the primal (Lemma 21 implies that such a minimizer exists). Set $h_1^* = h_1$ and $h_0^* = h_1^{C_\psi^*}$. Then Lemma 22 implies that $h_0^* \leq h_0$, so (h_0^*, h_1^*) is also optimal and $\eta h_1^* + (1-\eta)h_0^* \geq C_\psi^*(\eta)$ for all η. Furthermore, Lemma 22 implies that there is a function $\hat\eta$ for which $\hat\eta(\mathbf{x})h_1^*(\mathbf{x}) + (1-\hat\eta(\mathbf{x}))h_0^*(\mathbf{x}) = C_\psi^*(\hat\eta(\mathbf{x}))$.

It remains to show that $\hat\eta$ is Borel measurable. We will express $\hat\eta(\mathbf{x})$ in terms of $h_1^*(\mathbf{x})$, and because $h_1^*(\mathbf{x})$ is Borel measurable, it will follow that $\hat\eta$ is Borel measurable as well. Because $\eta h_1^*(\mathbf{x}) + (1-\eta)h_0^*(\mathbf{x}) \geq C_\psi^*(\eta)$ with equality at $\eta = \hat\eta(\mathbf{x})$, it follows that $h_1^*(\mathbf{x}) - h_0^*(\mathbf{x})$ is a supergradient of C_ψ^* at $\eta = \hat\eta(\mathbf{x})$. Thus Lemma 17 implies that $h_1^* - h_0^* = (1-2\hat\eta)/\sqrt{\hat\eta(1-\hat\eta)} \Leftrightarrow h_1^* = h_0^* + (1-2\hat\eta)/\sqrt{\hat\eta(1-\hat\eta)}$. Plugging this expression and the formula $C_\psi^*(\eta) = 2\sqrt{\eta(1-\eta)}$ into the relation $\hat\eta h_1^* + (1-\hat\eta)h_0^* = C_\psi^*(\hat\eta)$ results in the equation $h_0^* + \hat\eta(1-2\hat\eta)/\sqrt{\hat\eta(1-\hat\eta)} = 2\sqrt{\hat\eta(1-\hat\eta)}$. Solving for $\hat\eta$ then results in $\hat\eta = (h_0^*)^2/(1+(h_0^*)^2)$. Because h_0^* is Borel measurable, $\hat\eta$ is measurable as well. ■

Notice that this result together with Lemma 18 immediately implies that $h_1^*=\psi(\alpha_\psi(\hat\eta))$ and $h_1^*=\psi(-\alpha_\psi(\hat\eta))$, immediately proving that minimizing Θ over S_ψ is equivalent to minimizing R_ψ. Next, this observation is extended to arbitrary losses using properties of $\hat\eta$. Because both ψ and α_ψ are strictly monotonic, $\hat\eta$ interacts in a particularly nice way with maximizers of the dual problem:

Lemma 24 *Let $\mathbb{P}_0^*,\mathbb{P}_1^*$ be any maximizer of $\bar R_\psi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\times\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$. Set $\mathbb{P}^*=\mathbb{P}_0^*+\mathbb{P}_1^*$, $\eta^*=d\mathbb{P}_1^*/d\mathbb{P}^*$. Let $\hat\eta$ be defined as in Lemma 23. Then $\hat\eta=\eta^*$ $\mathbb{P}^*$-a.e.*

Furthermore, let γ_i be a coupling between $\mathbb{P}_i,\mathbb{P}_i^*$ with $\operatorname{supp}\gamma_i\subset\Delta_\epsilon$. Then

$$\operatorname{supp}\gamma_1\subset\{(\mathbf{x},\mathbf{x}')\colon \inf_{\|\mathbf{y}-\mathbf{x}\|\le\epsilon}\hat\eta(\mathbf{y})=\hat\eta(\mathbf{x}')\}\tag{50}$$

$$\operatorname{supp}\gamma_0\subset\{(\mathbf{x},\mathbf{x}')\colon \sup_{\|\mathbf{y}-\mathbf{x}\|\le\epsilon}\hat\eta(\mathbf{y})=\hat\eta(\mathbf{x}')\}\tag{51}$$

The statement $\hat\eta=\eta^*$ $\mathbb{P}^*$-a.e. implies that $\hat\eta$ is in fact a version of the Raydon-Nikodym derivative $d\mathbb{P}_1^*/d\mathbb{P}^*$.

For convenience, in this proof, we introduce the notation

$$I_\epsilon(f)(\mathbf{x})=\inf_{\|\mathbf{y}-\mathbf{x}\|\le\epsilon}f(\mathbf{y}).$$

Proof Let h_0^*,h_1^* be the minimizer described by Lemma 23. Then Lemma 18 implies that $h_1^*=\psi(\alpha_\psi(\hat\eta))$ and $h_0^*=\psi(-\alpha_\psi(\hat\eta))$.

The complimentary slackness condition (31) implies that $\eta^*h_1^*+(1-\eta^*)h_0^*=C_\psi^*(\eta^*)$ $\mathbb{P}^*$-a.e., and thus Lemma 18 implies that $h_1^*=\psi(\alpha_\psi(\eta^*))$ and $h_0^*=\psi(\alpha_\psi(\eta^*))$ $\mathbb{P}^*$-a.e. Therefore, $\psi(\alpha_\psi(\eta^*))=\psi(\alpha_\psi(\hat\eta))$ $\mathbb{P}^*$-a.e. Now because the functions ψ,α_ψ are strictly monotonic, they are invertible. Thus it follows that $\hat\eta=\eta^*$ $\mathbb{P}^*$-a.e.

The complimentary slackness condition (30) states that

$$\int S_\epsilon(h_i)(\mathbf{x})-h_i^*(\mathbf{x}')d\gamma_i=0.$$

Therefore,

$$S_\epsilon(\psi(\alpha_\psi(\hat\eta)))(\mathbf{x})=\psi(\alpha_\psi(\hat\eta(\mathbf{x}'))\quad\gamma_1\text{-a.e.}\quad\text{and}\quad S_\epsilon(\psi(-\alpha_\psi(\hat\eta)))(\mathbf{x})=\psi(-\alpha_\psi(\hat\eta(\mathbf{x}'))\quad\gamma_0\text{-a.e.}$$

which implies

$$\psi(\alpha_\psi(I_\epsilon(\hat\eta)(\mathbf{x})))=\psi(\alpha_\psi(\hat\eta(\mathbf{x}'))\quad\gamma_1\text{-a.e.}\quad\text{and}\quad\psi(-\alpha_\psi(S_\epsilon(\hat\eta)(\mathbf{x})))=\psi(-\alpha_\psi(\hat\eta(\mathbf{x}'))\quad\gamma_0\text{-a.e.}$$

Now ψ,α_ψ are both strictly monotonic and thus invertible. Therefore

$$I_\epsilon(\hat\eta)(\mathbf{x})=\hat\eta(\mathbf{x}')\quad\gamma_1\text{-a.e.}\quad\text{and}\quad S_\epsilon(\hat\eta)(\mathbf{x})=\hat\eta(\mathbf{x}')\quad\gamma_0\text{-a.e.}$$

■

Next, the relation (49) suggests that $h_1^*=\phi\circ f^*$, $h_0^*=\phi\circ -f^*$, where f^* is a function satisfying $C_\psi(\hat\eta(\mathbf{x}),f^*(\mathbf{x}))=C_\psi^*(\hat\eta(\mathbf{x}))$. In fact, Lemma 24 implies that this relation holds for *all* loss functions, and not just the exponential loss ψ. To formalize this idea, we prove the following result about minimizers of $C_\psi(\eta,\cdot)$ in Appendix C:

Lemma 25 *Fix a loss function ϕ and let $\alpha_\phi(\eta)$ be as in (8). Then α_ϕ maps η to the smallest minimizer of $C_\phi(\eta,\cdot)$. Furthermore, the function $\alpha_\phi(\eta)$ non-decreasing in η.*

Finally, we use the complimentary slackness conditions of Lemma 15 to construct a minimizer (h_0^*, h_1^*) to Θ over S_ϕ for which $h_1^* = \phi\circ f^*$, $h_0^* = \phi\circ -f^*$ for some function f^*.

Lemma 26 *Let $\psi = e^{-\alpha}$ be the exponential loss and let ϕ be any arbitrary loss. Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be any maximizer of $\bar{R}_\psi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\times\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$. Define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Let $\hat{\eta}$ be defined as in Lemma 23.*

Then $h_0^* = \phi(-\alpha_\phi(\hat{\eta}))$, $h_1^* = \phi(\alpha_\phi(\hat{\eta}))$ minimize Θ over S_ϕ and $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ maximize $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\times\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$.

Thus there exists a Borel minimizer to R_ϕ^ϵ and $\inf_f R_\phi^\epsilon(f) = \inf_{(h_0,h_1)\in S_\phi}\Theta(h_0,h_1)$.

Proof We will verify the complimentary slackness conditions of Lemma 15.

Lemma 24 implies that $\hat{\eta} = \eta^*$ $\mathbb{P}^*$-a.e. Therefore, $\mathbb{P}^*$-a.e.,

$$C_\phi^*(\eta^*) = C_\phi^*(\hat{\eta}) = \hat{\eta}h_1 + (1-\hat{\eta})h_0 = \eta^* h_1 + (1-\eta^*)h_0$$

This calculation verifies the complimentary slackness condition (31).

We next verify the other complimentary slackness condition (30). Let γ_i be a coupling between $\mathbb{P}_i, \mathbb{P}_i^*$ with $\operatorname{supp}\gamma_i \subset \Delta_\epsilon$. Next, because $\phi\circ\alpha_\phi$, $\phi\circ -\alpha_\phi$ are monotonic, the conditions (50) and (51) imply that

$$\int \phi(\alpha_\phi(\hat{\eta}))d\mathbb{P}_1^* = \int \phi(\alpha_\phi(\hat{\eta}(\mathbf{x}')))d\gamma_1(\mathbf{x},\mathbf{x}') = \int S_\epsilon(\phi(\alpha_\phi(\hat{\eta})))(\mathbf{x})d\gamma_1(\mathbf{x},\mathbf{x}') = \int S_\epsilon(\phi(\alpha_\phi(\hat{\eta})))d\mathbb{P}_1$$

$$\int \phi(-\alpha_\phi(\hat{\eta}))d\mathbb{P}_0^* = \int \phi(-\alpha_\phi(\hat{\eta}(\mathbf{x}')))d\gamma_0(\mathbf{x},\mathbf{x}') = \int S_\epsilon(\phi(-\alpha_\phi(\hat{\eta})))(\mathbf{x})d\gamma_0(\mathbf{x},\mathbf{x}') = \int S_\epsilon(\phi(-\alpha_\phi(\hat{\eta})))d\mathbb{P}_0$$

This calculation verifies the complimentary slackness condition (30). ■

Theorems 6 and 9 immediately follow from Lemmas 14 and 26.

8. Conclusion

We initiated the study of minimizers and minimax relations for adversarial surrogate risks. Specifically, we proved that there always exists a minimizer to the adversarial surrogate risk when perturbing in a closed ϵ-ball and a maximizer to the dual problem. Just like the results of (Pydi and Jog, 2021), our minimax theorem provides an interpretation of the adversarial learning problem as a game between two players. This theory helps explain the phenomenon of transfer attacks. We hope the insights gained from this research will assist in the development of algorithms for training classifiers robust to adversarial perturbations.

Acknowledgments

Natalie Frank was supported in part by the Research Training Group in Modeling and Simulation funded by the National Science Foundation via grant RTG/DMS – 1646339. Jonathan Niles-Weed was supported in part by a Sloan Research Fellowship.

Appendix A. The Universal σ-Algebra and a Generalization of Theorem 1

A.1 Definition of the Universal σ-Algebra and Main Measurability Result

In this Appendix, we prove results for supremums over an arbitrary compact set, not necessarily a unit ball. For a function $g\colon \mathbb{R}^d \to \mathbb{R}^d$, we will abuse notation and denote the supremum of g over the compact set B by

$$S_B(g)(\mathbf{x}) = \sup_{\mathbf{h}\in B} g(\mathbf{x}+\mathbf{h}).$$

Let X be a separable metric space and let $\mathcal{B}(X)$ be the Borel σ-algebra on X. Denote the completion of $\mathcal{B}(X)$ with respect to a Borel measure ν by $\mathcal{L}_\nu(X)$. Let $\mathcal{M}_+(X)$ be the set of all finite[2] positive Borel measures on X. Then the universal σ-algebra on X, $\mathscr{U}(X)$ is

$$\mathscr{U}(X) = \bigcap_{\nu\in\mathcal{M}_+(X)} \mathcal{L}_\nu(X). \tag{52}$$

In other words, the universal σ-algebra is the sigma-algebra of sets which are measurable with respect to the completion of every Borel measure. Thus $\mathscr{U}(X)$ is contained in $\mathcal{L}_\nu(X)$ for every Borel measure ν. The goal of this appendix is to prove

Theorem 27 *If f is universally measurable and B is a compact set, then $S_B(f)$ is universally measurable.*

Thus, if $\mathbb{P}_0, \mathbb{P}_1$, and g are Borel, integrals of the form $\int S_\epsilon(g)d\mathbb{P}_i$ in (10) can be interpreted as the integral of $S_\epsilon(g)$ with respect to the completion of $\mathbb{P}_i$.

A.2 Proof Outline

To prove Theorem 27, we analyze the level sets of $S_B(g)$. One can compute the level set $[S_B(g)(\mathbf{x}) > a]$ using a direct sum.

Lemma 28 *Let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ be any function. For a set B, define $-B = \{-\mathbf{b}\colon \mathbf{b}\in B\}$. Then*

$$[S_B(g) > a] = [g > a] \oplus -B$$

Proof To start, notice that $S_B(g)(\mathbf{x}) > a$ iff there is some $\mathbf{h}\in B$ for which $g(\mathbf{x}+\mathbf{h}) > a$. Thus

$$\mathbf{x}\in [S_B(g) > a] \Leftrightarrow \mathbf{x}+\mathbf{h} \in [g > a] \text{ for some } \mathbf{h}\in B \Leftrightarrow \mathbf{x}\in [g>a]\oplus -B$$

■

Therefore, to show that $S_B(g)$ is measurable for measurable g, it suffices to show that the direct sum of a measurable set and the compact set B is measurable. Thus, to prove Theorem 27, it suffices to demonstrate the following result:

Theorem 29 *Let $A\in \mathscr{U}(\mathbb{R}^d)$ and let B be a compact set. Then $A\oplus B \in \mathscr{U}(\mathbb{R}^d)$.*

2. Alternatively, one could compute the intersection in (52) over all σ-finite measures. These two approaches are equivalent because for every σ-finite measure λ and compact set K, the restriction $\lambda \llcorner K$ is a finite measure with $\mathcal{L}_{\lambda\llcorner K}(X) \supset \mathcal{L}_\lambda(X)$.

The proof of Theorem 29 follows from fundamental concepts of measure theory. A classical measure theory result states that if $f : X \to Y$ is a continuous function, f^{-1} maps Borel sets in Y to Borel sets in X. Consider now the function $w\colon B \times \mathbb{R}^d \to B \times \mathbb{R}^d$ given by $w(\mathbf{h}, \mathbf{x}) = (\mathbf{h}, \mathbf{x} - \mathbf{h})$. Then w is invertible and the inverse of w is $w^{-1}(\mathbf{h}, \mathbf{x} + \mathbf{h})$. Furthermore, w^{-1} maps the set $B \times A$ to $B \times A \oplus B$. Therefore, if $A \in \mathcal{B}(\mathbb{R}^d)$, then $B \times A \oplus B$ is Borel in $\mathcal{B}(B \times \mathbb{R}^d)$. However, from this statement, *one cannot conclude that $A \oplus B$ is Borel in $\mathbb{R}^d$!* On the otherhand, one can use regularity of measures to conclude that $A \oplus B$ is in $\mathscr{U}(\mathbb{R}^d)$. Therefore, to prove Theorem 29, we prove the following two results:

Lemma 30 *Let $B \subset \mathbb{R}^d$ be a compact set. Then $B \times A \in \mathscr{U}(B \times \mathbb{R}^d)$ iff $A \in \mathscr{U}(\mathbb{R}^d)$.*

In this document, we say a function $f\colon X \to Y$ is *universally measurable* if $f^{-1}(E) \in \mathscr{U}(X)$ whenever $E \in \mathscr{U}(Y)$.

Lemma 31 *Let $f : X \to Y$ be a Borel measurable function. Then f is universally measurable as well.*

This result is stated on page 171 of (Bertsekas and Shreve, 1996), but we include a proof below for completeness.

Lemma 31 applied to w implies that the set $B \times A \oplus B$ is universally measurable while Lemma 30 implies that $A \oplus B$ is universally measurable.

A.3 Proof of Theorem 29

We begin by proving Lemma 31.

Proof [Proof of Lemma 31] Let A be a Borel set in Y. We will show that for any finite measure ν on X, $f^{-1}(A) \in \mathcal{L}_\nu(X)$. As ν is arbitrary, this statement will impy that $f^{-1}(A) \in \mathscr{U}(X)$.

Consider the pushforward measure $\mu = f\sharp\nu$. This measure is a finite measure on Y, so by the definition of $\mathscr{U}(Y)$, $A \in \mathcal{L}_\mu(Y)$. Therefore, there are Borel sets $B_1 \subset A \subset B_2$ in Y for which $\mu(B_1) = \mu(B_2)$. Thus, $f^{-1}(B_1), f^{-1}(B_2)$ are Borel sets in X for which $f^{-1}(B_1) \subset f^{-1}(A) \subset f^{-1}(B_2)$ and $\nu(f^{-1}(B_1)) = \nu(f^{-1}(B_2))$. Therefore, $f^{-1}(A) \in \mathcal{L}_\nu(X)$. ∎

On the other hand, the proof of Lemma 30 relies on the definition of a regular space X:

Definition 32 *A measure ν is* inner regular *if for every Borel set A,*

$$\nu(A) = \sup_{\substack{K \text{ compact} \\ K \subset A}} \nu(K).$$

The topological space X is regular *if every finite Borel measure on X is inner regular.*

The following result implies that most topological spaces encountered in applications are regular.

Theorem 33 *A σ-compact locally compact Hausdorff space is regular.*

This theorem is is a consequence of Theorem 7.8 of (Folland, 1999).

The notion of regularity extends to complete measures.

Lemma 34 *Let $\overline{\nu}$ be the completion of a measure ν on a regular space X. Then for any $A \in \mathcal{L}_\nu(X)$,*

$$\overline{\nu}(A) = \sup_{\substack{K \text{ compact} \\ K \subset A}} \nu(K).$$

The proof of this result is left as a exercise to the reader.

Now using the concept of regularity, we prove Lemma 30.

Proof [Proof of Lemma 30] We first prove the forward direction. Consider the projection function $\Pi_2 \colon B \times \mathbb{R}^d \to \mathbb{R}^d$ given by $\Pi_2(\mathbf{x}, \mathbf{y}) = \mathbf{y}$. Then Π_2 is a continuous function and $\Pi_2^{-1}(A) = B \times A$. Therefore Lemma 31 implies that if A is universally measurable in $\mathbb{R}^d$, then $B \times A$ is universally measurable in $B \times \mathbb{R}^d$.

To prove the other direction, assume that $B \times A$ is universally measurable in $B \times \mathbb{R}^d$. Let ν be any finite Borel measure on $\mathbb{R}^d$. We will find Borel sets B_1, B_2 with $B_1 \subset A \subset B_2$ for which $\nu(B_1) = \nu(B_2)$, and thus $A \in \mathcal{L}_\nu(\mathbb{R}^d)$. As ν was arbitrary, it follows that A is universally measurable.

Theorem 33 implies that $B \times \mathbb{R}^d$ is a regular space. Fix a Borel probability measure λ on B. Then $\lambda \times \nu$ is a finite Borel measure on $B \times \mathbb{R}^d$, so it is inner regular. Let $\overline{\lambda \times \nu}$ be the completion of $\lambda \times \nu$. Then by Lemma 34,

$$\overline{\lambda \times \nu}(B \times A) = \sup_{\substack{K \text{ compact} \\ K \subset B \times A}} \lambda \times \nu(K)$$

We will now argue that

$$\sup_{\substack{K \text{ compact} \\ K \subset B \times A}} \lambda \times \nu(K) = \sup_{\substack{K \text{ compact} \\ K \subset A}} \nu(K) \tag{53}$$

Let $K \subset B \times A$ and let Π_2 be projection onto the second coordinate. Because the continuous image of a compact set is compact, K'=$\Pi_2(K)$ is compact and contained in A. Thus $B \times A \supset B \times K' \supset K$, which implies (53). Now (53) applied to A^C implies that

$$\overline{\lambda \times \nu}(X \times A) = \inf_{\substack{U^C \text{ compact} \\ U \supset B \times A}} \lambda \times \nu(U) = \inf_{\substack{U^C \text{ compact} \\ U \supset A}} \nu(U).$$

Thus

$$\sup_{\substack{K \text{ compact} \\ K \subset A}} \nu(K) = \inf_{\substack{U^C \text{ compact} \\ U \supset A}} \nu(U) := m$$

Let K_n be a sequence of compact sets contained in A for which $\lim_{n \to \infty} \nu(K_n) = m$ and U_n a sequence of sets containing A for which U_n^C is compact and $\lim_{n \to \infty} \nu(U_n) = m$. Because a finite union of compact sets is compact, one can choose such sequences that satisfy $K_{n+1} \supset K_n$ and $U_{n+1} \subset U_n$. Then $B_1 = \bigcup K_n$, $B_2 = \bigcap U_n$ are Borel sets that satisfy $B_1 \subset A \subset B_2$ and $\nu(B_1) = \nu(B_2)$ so $A \in \mathcal{L}_\nu(\mathbb{R}^d)$.

■

Lastly, we formally prove Theorem 29.

Proof [Proof of Theorem 29] Consider the function $w\colon B\times\mathbb{R}^d \to B\times\mathbb{R}^d$ given by $w(\mathbf{h},\mathbf{x}) = (\mathbf{h},\mathbf{x}-\mathbf{h})$. Then w is continuous, invertible, and $w^{-1}(\mathbf{h},\mathbf{x}) = (\mathbf{x},\mathbf{x}+\mathbf{h})$.

Now let $A \in \mathscr{U}(\mathbb{R}^d)$. Then Lemma 30 implies that $B\times\mathbb{R}^d$ is universally measurable in $B\times A$. Lemma 31 then implies that $w^{-1}(B\times A) = B\times A\oplus B$ is universally measurable as well. Finally, Lemma 30 implies that $A\oplus B \in \mathscr{U}(\mathbb{R}^d)$ as well. ∎

Appendix B. Alternative Characterizations of the W_∞ Metric

We start with proving Lemma 3 using a measurable selection theorem.

Theorem 35 *Let X, Y be Borel sets and assume that $D \subset X\times Y$ is also Borel. Let D_x denote*

$$D_x = \{y\colon (x,y)\in D\}$$

and

$$\mathrm{Proj}_X(D)\colon = \{x\colon (x,y)\in D\}$$

Let $f\colon D\to\overline{\mathbb{R}}$ be a Borel function mapping D to $\overline{\mathbb{R}}$ and define

$$f^*(x) = \inf_{y\in D_x} f(x,y)$$

Assume that $f^*(x) > -\infty$ for all x. Then for any $\delta > 0$, there is a universally measurable $\varphi\colon \mathrm{Proj}_X(D)\to Y$ for which

$$f(x,\varphi(x)) \le f(x)+\delta$$

This statement is a consequence of Proposition 7.50 from (Bertsekas and Shreve, 1996).

We use the following results about universally measurable functions, see Lemma 7.27 of (Bertsekas and Shreve, 1996).

Lemma 36 *Let $g:\mathbb{R}^d\to\mathbb{R}$ be a universally measurable function and let $\mathbb{Q}$ be a Borel measure. Then there is a Borel measurable function φ for which $\varphi = g$ $\mathbb{Q}$-a.e.*

This result can be extended to $\mathbb{R}^d$-valued functions:

Lemma 37 *Let $g:\mathbb{R}^d\to\mathbb{R}^d$ be a universally measurable function and let $\mathbb{Q}$ be a Borel measure. Then there is a Borel measurable function φ for which $\varphi = g$ $\mathbb{Q}$-a.e.*

Proof Let $\mathbf{e}_i$ denote the ith basis vector. Then $g_i := \mathbf{e}_i\cdot g$ is a universally measurable function from $\mathbb{R}^d$ to $\mathbb{R}$, so by Lemma 36, there is a Borel function φ_i for which $\varphi_i = g_i$ $\mathbb{Q}$-a.e. Then if we define $\varphi = (\varphi_1,\varphi_2,\ldots,\varphi_d)$, this function is equal to g $\mathbb{Q}$-a.e. ∎

Finally, we prove Lemma 3. Due to Lemmas 36 and 37, this lemma heavily relies on the fact that the domain of our functions is $\mathbb{R}^d$ rather than an arbitrary metric space.

Lemma 3 *Let $\mathbb{Q}$ be a finite positive Borel measure and let $f\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ be a Borel measurable function. Then*

$$\int S_\epsilon(f)d\mathbb{Q} = \sup_{\mathbb{Q}'\in\mathcal{B}^\infty_\epsilon(\mathbb{Q})} \int fd\mathbb{Q}' \tag{13}$$

Recall that this paper defines the left-left hand side of (13) as the integral of $S_\epsilon(f)$ with respect to the completion of $\mathbb{Q}$.

Proof

To start, let $\mathbb{Q}'$ be a Borel measure satisfying $W_\infty(\mathbb{Q}',\mathbb{Q})\leq \epsilon$. Let γ be a coupling with marginals $\mathbb{Q}$ and $\mathbb{Q}'$ supported on Δ_ϵ. Then

$$\int fd\mathbb{Q}' = \int f(\mathbf{x}')d\gamma(\mathbf{x},\mathbf{x}') = \int f(\mathbf{x}')\mathbf{1}_{\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon}d\gamma(\mathbf{x},\mathbf{x}')$$
$$\leq \int S_\epsilon(f)(\mathbf{x})\mathbf{1}_{\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon}d\gamma(\mathbf{x},\mathbf{x}') = \int S_\epsilon(f)(\mathbf{x})d\gamma(\mathbf{x},\mathbf{x}') = \int S_\epsilon(f)d\mathbb{Q}$$

Therefore, we can conclude that

$$\sup_{\mathbb{Q}'\in\mathcal{B}^\infty_\epsilon(\mathbb{Q})} \int fd\mathbb{Q}' \leq \int S_\epsilon(f)d\mathbb{Q}.$$

We will show the opposite inequality by applying the measurable selection theorem. Theorem 35 implies for each $\delta>0$, one can find a universally measurable function $\varphi\colon\mathbb{R}^d\to\overline{B_\epsilon(\mathbf{x})}$ for which $f(\varphi(\mathbf{x}))+\delta\geq S_\epsilon(f)(\mathbf{x})$. By Lemma 37, one can find a Borel measurable function T for which $T=\varphi$ $\mathbb{Q}$-a.e.

Let $\mathbb{Q}'=\mathbb{Q}\circ T^{-1}$. Because T is Borel measurable, $\mathbb{Q}'$ and $f\circ T$ are Borel. We will now argue that $\int fd\mathbb{Q}'+\delta\geq\int S_\epsilon(f)d\mathbb{Q}$. Recall that φ is always measurable with respect to the completion of $\mathbb{Q}$, and by convention $\int gd\mathbb{Q}$ means integration with respect to the completion of $\mathbb{Q}$. Then if we define $M=\mathbb{Q}(\mathbb{R}^d)$,

$$\int fd\mathbb{Q}' = \int fd\mathbb{Q}\circ T^{-1} = \int f(T(\mathbf{x}))d\mathbb{Q} = \int f(\varphi(\mathbf{x}))d\mathbb{Q} \geq \int S_\epsilon(f)-\delta d\mathbb{Q} = \int S_\epsilon(f)d\mathbb{Q}-\delta M$$

Because $\delta>0$ was arbitrary and $\mathbb{Q}'\in\mathcal{B}^\infty_\epsilon(\mathbb{Q})$,

$$\int S_\epsilon(f)d\mathbb{Q} \leq \sup_{\mathbb{Q}'\in\mathcal{B}^\infty_\epsilon(\mathbb{Q})} \int fd\mathbb{Q}'$$

It remains to show that $W_\infty(\mathbb{Q},\mathbb{Q}')\leq\epsilon$. Define a function $G\colon\mathbb{R}^d\to\mathbb{R}^d\times\mathbb{R}^d$, $G(\mathbf{x})=(\mathbf{x},T(\mathbf{x}))$ and a coupling γ by $\gamma=G\sharp\mathbb{Q}$. Then $\gamma(\Delta_\epsilon)=G\sharp(\mathbb{Q})(\Delta_\epsilon)=\mathbb{Q}(G^{-1}(\Delta_\epsilon))=1$, so $\text{supp}(\gamma)\subseteq\Delta_\epsilon$. ∎

Next we prove Lemma 4. We begin by presenting Strassen's theorem, see Corollary 1.28 of (Villani, 2003) for more details

Theorem 38 (Strassen's Theorem) *Let $\mathbb{P},\mathbb{Q}$ be positive finite measures with the same mass and let $\epsilon\geq 0$. Let $\Pi(\mathbb{P},\mathbb{Q})$ denote the set couplings of $\mathbb{P}$ and $\mathbb{Q}$. Then*

$$\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\pi(\{\|\mathbf{x}-\mathbf{y}\|>\epsilon) = \sup_{A \textit{ closed}} \mathbb{Q}(A)-\mathbb{P}(A^\epsilon) \tag{54}$$

Strassen's theorem is usually written with A^ϵ in (54) replaced by $A^{\epsilon]} = \{\mathbf{x}\colon \text{dist}(\mathbf{x}, A) \leq \epsilon\}$—however, for closed sets $A^{\epsilon]} = A^\epsilon$. Strassen's theorem together with Urysohn's lemma then immediately proves Lemma 4.

Lemma 39 (Urysohn's Lemma) *Let A and B be two closed and disjoint subsets of $\mathbb{R}^d$. Then there exists a function $f\colon \mathbb{R}^d \to [0,1]$ for which $f = 0$ on A and $f = 1$ on B.*

See for instance result 4.15 of (Folland, 1999).

Lemma 4 *Let $\mathbb{P}, \mathbb{Q}$ be two finite positive Borel measures with $\mathbb{P}(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$. Then*

$$W_\infty(\mathbb{P}, \mathbb{Q}) = \inf_\epsilon \{\epsilon \geq 0\colon \int h d\mathbb{Q} \leq \int S_\epsilon(h) d\mathbb{P} \quad \forall h \in C_b(\mathbb{R}^d)\}$$

Proof First, notice that Lemma 3 implies that if $\mathbb{Q} \in \mathcal{B}_\epsilon^\infty(\mathbb{P})$, then $\int S_\epsilon(h) d\mathbb{P} \geq \int h d\mathbb{Q}$ for all $h \in C_b(\mathbb{R}^d)$, proving the inequality $\geq$ in the statement of the lemma.

We will now argue the other inequality: specifically, we will show that

$$\sup_{A \text{ closed}} \mathbb{Q}(A) - \mathbb{P}(A^\epsilon) \leq \sup_{h \in C_b(\mathbb{R}^d)} \int h d\mathbb{Q} - \int S_\epsilon(h) d\mathbb{P} \tag{55}$$

Strassen's theorem will then imply that $W_\infty(\mathbb{P}, \mathbb{Q}) \leq \epsilon$. Let δ be arbitrary and let A be a closed set that satisfies $\sup_{A \text{ closed}} \mathbb{Q}(A) - \mathbb{P}(A^\epsilon) \leq \mathbb{Q}(A) - \mathbb{P}(A^\epsilon) + \delta$. Now because A is closed, $A_n = A \oplus B_{1/n}(\mathbf{0})$ is a series of open sets decreasing to A and $A_n^\epsilon = A^\epsilon \oplus B_{1/n}(\mathbf{0})$ is a sequence of open sets decreasing to A^ϵ. Thus pick n sufficiently large so that $\mathbb{P}(A_n^\epsilon - \mathbb{P}(A^\epsilon) \leq \delta$. By Urysohn's lemma, one can choose a function h which is 1 on A, 0 on A_n^C, and between 0 and 1 on $A_n - A^C$. Then $S_\epsilon(h)$ is 1 on A^ϵ, 0 on $(A_n^\epsilon)^C$ and between 0 and 1 on $A_n^\epsilon - A^\epsilon$. Then $\int h d\mathbb{Q} - \mathbb{Q}(A) \geq 0$ and thus

$$\left(\int h d\mathbb{Q} - \int S_\epsilon(h) d\mathbb{P} \right) - (\mathbb{Q}(A) - \mathbb{P}(A^\epsilon)) \geq \mathbb{P}(A^\epsilon) - \mathbb{P}(A_n^\epsilon) \geq -\delta.$$

Because δ was arbitrary, (55) follows. ■

Appendix C. Minimizers of $C_\phi(\eta, \cdot)$: Proof of Lemma 25

Lemma 25 *Fix a loss function ϕ and let $\alpha_\phi(\eta)$ be as in (8). Then α_ϕ maps η to the smallest minimizer of $C_\phi(\eta, \cdot)$. Furthermore, the function $\alpha_\phi(\eta)$ non-decreasing in η.*

Proof To start, we will show that $\alpha_\phi(\eta)$ as defined in (8) is a minimizer of $C_\phi(\eta, \cdot)$. Let S be the set of minimizers of $C_\phi^*(\eta, \cdot)$, which is non-empty due to the lower semi-continuity of ϕ. Let $a = \inf S = \alpha_\phi(\eta)$ and let $s_i \in S$ be a sequence converging to a. Then because ϕ is lower semi-continuous,

$$C_\phi^*(\eta) = \liminf_{i \to \infty} \eta\phi(s_i) + (1-\eta)\phi(-s_i) \geq \eta\phi(a) + (1-\eta)\phi(-a)$$

Then a is in fact a minimizer of $C_\phi^*(\eta, \cdot)$, so it is the smallest minimizer of $C_\phi^*(\eta, \cdot)$.

We will now show that the function α_ϕ is non-decreasing.

One can write

$$\begin{aligned}C_\phi(\eta_2,\alpha) &= \eta_2\phi(\alpha)+(1-\eta_2)\phi(-\alpha)\\ &= \eta_1\phi(\alpha)+(1-\eta_1)\phi(-\alpha)+(\eta_2-\eta_1)(\phi(\alpha)-\phi(-\alpha))\\ &= C_\phi(\eta_1,\alpha)+(\eta_2-\eta_1)(\phi(\alpha)-\phi(-\alpha))\end{aligned} \tag{56}$$

Notice that the function $\alpha\mapsto\phi(\alpha)-\phi(-\alpha)$ is non-increasing. Then because $\alpha_\phi(\eta_1)$ is the smallest minimizer of $C_\phi(\eta_1,\alpha)$, if $\alpha<\alpha_\phi(\eta_1)$, then $C_\phi(\eta_1,\alpha)>C_\phi(\eta_1,\alpha_\phi(\eta_1))$. Furthermore, $\phi(\alpha)-\phi(-\alpha)\geq\phi(\alpha_\phi(\eta_1))-\phi(-\alpha_\phi(\eta_1))$. Therefore, (56) implies that $C_\phi(\eta_2,\alpha)>C_\phi(\eta_2,\alpha_\phi(\eta_1))$, and thus α cannot be a minimizer of $C_\phi(\eta_2,\cdot)$. Therefore, $\alpha_\phi(\eta_2)\geq\alpha_\phi(\eta_1)$.

■

Appendix D. Continuity Properties of $\bar{R}_\phi$—Proof of Lemma 12

Recall the function $G(\eta,\alpha)$ defined by (47). With this notation, one can write the C_ϕ^* transform as $h_1^{C_\phi^*}=\sup_{\eta\in[0,1]}G(\eta,h_1)$.

Lemma 40 *Let $c>0$ and consider $\alpha\geq c$. Let $a(\alpha)=\alpha^{C_\phi^*}$, where the C_ϕ^* transform is as in Lemma 22. Then there is a constant $k<1$ for which*

$$a(\alpha)=\sup_{\eta\in[0,k]}\frac{C_\phi^*(\eta)-\eta\alpha}{1-\eta} \tag{57}$$

The constants k depends only on c.

Proof Recall that the function $G(\eta,\alpha)$ is decreasing in α for fixed η and continuous on $[1,0)$. Let $k=\sup\{\eta\colon G(\eta,c)>0\}$. As c is strictly positive, one can conclude that $\lim_{\eta\to1}G(\eta,c)=-\infty$ and as a result $k<1$. Because G is decreasing in α, one can conclude that $G(\eta,\alpha)\leq 0$ for all $\eta>k$ and $\alpha\geq c$. However, $\sup_{\eta\in[0,1]}G(\eta,\alpha)\geq 0$ because $G(0,\alpha)=0$ for all α. Thus (57) holds.

■

Lemma 41 *Let $\{f_\alpha\}$ be a set of L-Lipschitz functions. Then $\sup_\alpha f_\alpha$ is also L-Lipschitz.*

This statement is proved in Box 1.8 of (Santambrogio, 2015).

Lemma 42 *Let $\mathbb{Q}$ be any finite measure and assume that g is a non-negative function in $L^1(\mathbb{Q})$. Let $\delta>0$. Then there is a lower semi-continuous function $\tilde g$ for which $\int|g-\tilde g|<\delta$ and $g\geq 0$.*

See Proposition 7.14 of Folland.

Lemma 43 *Let g be a lower semi-continuous function bounded from below. Then there is a sequence of Lipschitz functions that approaches g from below.*

This statement appears in Box 1.5 of (Santambrogio, 2015).

Corollary 44 *Let h be an $L^1(\mathbb{Q})$ function with $h \geq 0$. Then for any δ, there exists a Lipschitz $\tilde{h}$ for which $\int |h - \tilde{h}| d\mathbb{Q} < \delta$.*

Proof By Lemma 42, one can pick a lower semi-continuous $\tilde{g}$ for which $\tilde{g} \geq 0$ and $\int |h - \tilde{g}| d\mathbb{Q} < \delta/2$. Next, by Lemma 43, one can pick a Lipschitz $\tilde{h}$ for which $\int |\tilde{g} - \tilde{h}| d\mathbb{Q} \leq \delta/2$. Thus $\int |h - \tilde{h}| d\mathbb{Q} < \delta$. ■

Lemma 12 *Let $K \subset \mathbb{R}^d$ be compact, $E = C_b(K^\epsilon) \times C_b(K^\epsilon)$, and $\mathbb{P}_0', \mathbb{P}_1' \in \mathcal{M}_+(K^\epsilon)$. Then*

$$\inf_{(h_0,h_1) \in S_\phi \cap E} \int h_1 d\mathbb{P}_1' + \int h_0 d\mathbb{P}_0' = \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') \tag{27}$$

Therefore, $\bar{R}_\phi$ is concave and upper semi-continuous on $\mathcal{M}_+(K^\epsilon) \times \mathcal{M}_+(K^\epsilon)$ with respect to the weak topology on probability measures.

Proof Let $\mathbb{P}' = \mathbb{P}_0' + \mathbb{P}_1'$ and $\eta' = d\mathbb{P}_1'/d\mathbb{P}'$. Then for any $(h_0, h_1) \in S_\phi \cap E$,

$$\int h_1 d\mathbb{P}_1' + \int h_0 d\mathbb{P}_0' = \int \eta' h_1 + (1 - \eta') h_0 d\mathbb{P}' \geq \int C_\phi^*(\eta') d\mathbb{P}' = \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1').$$

We will now focus on showing the other inequality. Define a function f by

$$f(\mathbf{x}) = \begin{cases} \alpha_\phi(\eta'(\mathbf{x})) & \mathbf{x} \in \operatorname{supp} \mathbb{P}' \\ 0 & \mathbf{x} \notin \operatorname{supp} \mathbb{P}' \end{cases}$$

Let $h_1 = \phi \circ f$, $h_0 = \phi \circ -f$. Then h_1, h_0 satisfy the inequality $\eta h_1 + (1 - \eta) h_0 \geq C_\phi^*(\eta)$ for all η while on $\operatorname{supp} \mathbb{P}'$, $\eta'(\mathbf{x}) h_1(\mathbf{x}) + (1 - \eta'(\mathbf{x})) h_0(\mathbf{x}) = C_\phi^*(\eta')$ and therefore

$$\int h_1 d\mathbb{P}_1' + \int h_0 d\mathbb{P}_0' = \int \eta' h_1 + (1 - \eta') h_0 d\mathbb{P}' = \int C_\phi^*(\eta') d\mathbb{P}'.$$

However, $(h_0, h_1) \notin E$. We will now approximate h_0, h_1 by bounded continuous functions contained in S_ϕ. Let $\delta > 0$ be arbitrary. Pick a constant $c > 0$ for which $\int c d\mathbb{P}' < \delta$ and set $\tilde{h}_1 = \max(h_1, c)$. The pair $(h_0, \tilde{h}_1)$ are feasible pair for the set S_ϕ, and thus

$$C_\phi^*(\eta) - \eta \tilde{h}_1 - (1 - \eta) h_0 \leq 0 \tag{58}$$

Furthermore,

$$\int \tilde{h}_1 d\mathbb{P}_1' + \int h_0 d\mathbb{P}_0' < \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') + \delta. \tag{59}$$

Let k be the constant described by Lemma 40 corresponding to c. Now by Corollary 44, there is a Lipschitz function g for which $\int |h_1 - g| d\mathbb{P}' < \min((1 - k)/k, 1)\delta$. Let $\hat{h}_1 =$

$\max(g, c)$. Then Lemma 41 implies that $\hat{h}_1$ has the same Lipschitz constant as g, and the fact that $\tilde{h}_1 \geq c$ implies that

$$\int |\tilde{h}_1 - \hat{h}_1| d\mathbb{P}' \leq \int |\tilde{h}_1 - g| d\mathbb{P}' < \min\left(\frac{1-k}{k}, 1\right)\delta \tag{60}$$

Now let $\hat{h}_0 = \hat{h}_1^{C_\phi^*}$. By Lemma 40, the supremum in the $^{C_\phi^*}$ transform for computing $\hat{h}_0$ can be taken over $[0, k]$. Therefore, if L is the Lipschitz constant of $\hat{h}_1$, Lemma 41 implies that the Lipschitz constant of $\hat{h}_0$ is at most $kL/(1-k)$. Furthermore, $\hat{h}_0, \hat{h}_1$ are bounded on K^ϵ because Lipschitz functions are bounded over compact sets. Thus $(\hat{h}_0, \hat{h}_1)$ is in $S_\phi \cap E$. Next, we will show that $\int \hat{h}_0$ is close to $\int h_0$.

$$\int \hat{h}_0 - h_0 d\mathbb{P}_0' = \int \sup_{[0,k]} \frac{C_\phi^*(\eta) - \eta \hat{h}_1}{1-\eta} - h_0 d\mathbb{P}_0' = \int \sup_{[0,k]} \frac{C_\phi^*(\eta) - \eta \hat{h}_1 - (1-\eta) h_0}{1-\eta} d\mathbb{P}_0'$$
$$= \int \sup_{[0,k]} \left(\frac{C_\phi^*(\eta) - \eta \tilde{h}_1 - (1-\eta) h_0}{1-\eta} + \frac{\eta}{1-\eta}(\tilde{h}_1 - \hat{h}_1) \right) d\mathbb{P}_0'$$
$$\leq \int \sup_{[0,k]} \frac{C_\phi^*(\eta) - \eta \tilde{h}_1 - (1-\eta) h_0}{1-\eta} + \sup_{[0,k]} \frac{\eta}{1-\eta}(\tilde{h}_1 - \hat{h}_1) d\mathbb{P}_0' \leq \int \sup_{[0,k]} \frac{\eta}{1-\eta}(\tilde{h}_1 - \hat{h}_1) d\mathbb{P}_0' \quad \text{(Equation 58)}$$
$$= \frac{k}{1-k} \int \tilde{h}_1 - \hat{h}_1 d\mathbb{P}_0' \leq \delta \quad \text{(Equation 60)}$$

Therefore, by (59), (60), and the computation above,

$$\int \hat{h}_1 d\mathbb{P}_1' + \int \hat{h}_0 d\mathbb{P}_0' \leq \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') + 3\delta$$

AS $\delta > 0$ is arbitrary, this inequality implies (27). Because K^ϵ is compact, the upper semi-continuity and concavity of $\bar{R}_\phi$ then follows from (27) together with the Reisz representation theorem. ∎

Appendix E. Duality for Distributions with Arbitrary Support—Proof of Lemma 14

We begin with the simple observation that weak duality holds for measures supported on $\mathbb{R}^d$. This argument is essentially swapping the order of an infimum and a supremum as presented in Section 4.1.

Lemma 45 (Weak Duality) *Let ϕ be a non-increasing and lower semi-continuous loss function. Let S_ϕ be the set of pairs of functions defined in* (25) *for $K = \mathbb{R}^d$.*

Then

$$\inf_{(h_0,h_1) \in S_\phi} \Theta(h_0, h_1) \geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1')$$

Proof By Lemma 3,

$$\inf_{(h_0,h_1)\in S_\phi} \int S_\epsilon(h_0) d\mathbb{P}_0 + \int S_\epsilon(h_1) d\mathbb{P}_1 = \inf_{(h_0,h_1)\in S_\phi} \sup_{\substack{\mathbb{P}_0'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \int h_0 d\mathbb{P}_0' + \int h_1 d\mathbb{P}_1'.$$

Thus by swapping the inf and the sup,

$$\inf_{(h_0,h_1)\in S_\phi} \int S_\epsilon(h_0) d\mathbb{P}_0 + \int S_\epsilon(h_1) d\mathbb{P}_1 \geq \sup_{\substack{\mathbb{P}_0'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{(h_0,h_1)\in S_\phi} \int h_0 d\mathbb{P}_0' + \int h_1 d\mathbb{P}_1'$$

$$= \sup_{\substack{\mathbb{P}_0'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{(h_0,h_1)\in S_\phi} \int \frac{d\mathbb{P}_1'}{d(\mathbb{P}_0'+\mathbb{P}_1')} h_1 + \left(1 - \frac{d\mathbb{P}_1'}{d(\mathbb{P}_0'+\mathbb{P}_1')}\right) h_0 d(\mathbb{P}_0'+\mathbb{P}_1') \geq \sup_{\substack{\mathbb{P}_0'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1')$$

■

The main strategy in this section is approximating measures with unbounded support by measures with bounded support. To this end, we define the *restriction* of a measure $\mathbb{P}$ to a set K by $\mathbb{P}|_K(A) = \mathbb{P}(K \cap A)$.

The Portmaneau theorem then allows us to draw some conclusions about weakly convergent sequences of measures.

Theorem 46 (Portmanteau Theorem) *The following are equivalent:*

1) *The sequence $\mathbb{Q}^n \in \mathcal{M}_+(\mathbb{R}^d)$ converges weakly to $\mathbb{Q}$*
2) *For all closed sets C, $\limsup_{n\to\infty} \mathbb{Q}^n(C) \leq \mathbb{Q}(C)$ and $\lim_{n\to\infty} \mathbb{Q}^n(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$*
3) *For all open sets U, $\liminf_{n\to\infty} \mathbb{Q}^n(U) \geq \mathbb{Q}(U)$ and $\lim_{n\to\infty} \mathbb{Q}^n(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$*

See Theorem 8.2.3 of (Bogachev, 2007). This result allows us to draw conclusions about restrictions of weakly convergent sequences.

Lemma 47 *Let $\mathbb{Q}^n, \mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^d)$ and assume that $\mathbb{Q}^n$ converges weakly to $\mathbb{Q}$. Let K be a compact set with $\mathbb{Q}(\partial K) = 0$. Then $\mathbb{Q}^n|_K$ converges weakly to $\mathbb{Q}|_K$.*

Proof We will verify 2) of Theorem 46 for the measures $\mathbb{Q}^n|_K$, $\mathbb{Q}$.

First, because $\mathbb{Q}(K) = \mathbb{Q}(\operatorname{int} K)$, Theorem 46 implies that

$$\limsup_{n\to\infty} \mathbb{Q}^n(K) \leq \mathbb{Q}(K) = \mathbb{Q}(\operatorname{int} K) \leq \liminf_{n\to\infty} \mathbb{Q}^n(\operatorname{int} K) \leq \liminf_{n\to\infty} \mathbb{Q}^n(K).$$

Therefore, $\lim_{n\to\infty} \mathbb{Q}^n|_K(\mathbb{R}^d) = \lim_{n\to\infty} \mathbb{Q}^n(K) = \mathbb{Q}(K)$. Next, for any closed set C, the set $C \cap K$ is also closed so the fact that $\mathbb{Q}^n$ weakly converges to $\mathbb{Q}$ implies that

$$\limsup_{n\to\infty} \mathbb{Q}^n|_K(C) = \limsup_{n\to\infty} \mathbb{Q}^n(K \cap C) \leq \mathbb{Q}(K \cap C) = \mathbb{Q}|_K(C).$$

■

Next, Prokhorov's theorem allows us to identify weakly convergent subsequences.

Theorem 48 *Let $\mathbb{Q}^n$ be a sequence of measures for which $\sup_n \mathbb{Q}^n(\mathbb{R}^d) < \infty$ and for all δ, there exists a compact K for which $\mathbb{Q}^n(K^C) < \delta$ for all n. Then $\mathbb{Q}^n$ has a weakly convergent subsequence.*

See Theorem 8.6.2 of (Bogachev, 2007). These results imply that $\bar{R}_\phi$ is upper semi-continuous on $\mathcal{M}_+(\mathbb{R}^d) \times \mathcal{M}_+(\mathbb{R}^d)$.

Lemma 49 *The functional $\bar{R}_\phi$ is upper semi-continuous with respect to the weak topology on probability measures (in duality with $C_0(\mathbb{R}^d)$).*

Notice that Lemma 12 implies that $\bar{R}_\phi$ is upper semi-continuous on the space $\mathcal{M}_+(K^\epsilon) \times \mathcal{M}_+(K^\epsilon)$ for a *compact* set K. However, on $\mathbb{R}^d$, weak convergence of measures is defined with respect to the dual of $C_0(\mathbb{R}^d)$, the set of continuous functions vanishing at ∞. This set is strictly smaller than $C_b(\mathbb{R}^d)$, and thus the relation (27) would not immediately imply the the upper semi-continuity of R_ϕ^ϵ.

Proof Let $\mathbb{Q}_0^n, \mathbb{Q}_1^n$ be sequences of measures converging to $\mathbb{Q}_0, \mathbb{Q}_1$ respectively. Set $\mathbb{Q} = \mathbb{Q}_0 + \mathbb{Q}_1$.

Define a function $F(R) = \mathbb{Q}(\overline{B_R(\mathbf{0})}^C)$. Then because this function is non-increasing, it has finitely many points of discontinuity.

Let $\delta > 0$ be arbitrary and choose R large enough so that $F(R) < \delta/C_\phi^*(1/2)$ and F is continuous at R. Then notice that $\mathbb{P}(\partial B_R(\mathbf{0})) = 0$ and thus one can apply Lemma 47 with the set $\overline{B_R(\mathbf{0})}$.

Now let ν_0, ν_1 be arbitrary measures. Consider ν_i^R defined by $\nu_i^R = \nu_i|_{\overline{B_R(\mathbf{0})}}$. Set $\nu = \nu_0 + \nu_1$, $\eta = d\nu_1/d\nu$, $\nu^R = \nu_0^R + \nu_1^R$, $\eta^R = d\nu_1^R/d\nu^R$. Then on $\overline{B_R(\mathbf{0})}$, $\eta^R = \eta$ a.e. Thus

$$|\bar{R}_\phi(\nu_0^R, \nu_1^R) - \bar{R}_\phi(\nu_0, \nu_1)| = \left| \int C_\phi^*(\eta) \mathbf{1}_{\overline{B_R(\mathbf{0})}} d\nu - \int C_\phi^*(\eta) d\nu \right| \leq C_\phi^* \left(\frac{1}{2}\right) \nu(\overline{B_R(\mathbf{0})^C}) \quad (61)$$

If we define $\mathbb{Q}_{i,R}, \mathbb{Q}_{i,R}^n$ via $\mathbb{Q}_{i,R} = \mathbb{Q}_i|_{\overline{B_R(\mathbf{0})}}$, $\mathbb{Q}_{i,R}^n = \mathbb{Q}_i^n = \mathbb{Q}_i^n|_{\overline{B_R(\mathbf{0})}}$, Lemma 47 implies that $\mathbb{Q}_{i,R}^n$ converges weakly to $\mathbb{Q}_{i,R}$ and $\lim_{n\to\infty} \mathbb{Q}^n(\overline{B_R(\mathbf{0})}^C) = \mathbb{Q}(B_R(\mathbf{0})^C) < \delta$. Therefore, for sufficiently large n, $\mathbb{Q}^n(\overline{B_R(\mathbf{0})}^C) < 2\delta/C_\phi^*(1/2)$. By Lemma 12 and (61),

$$\limsup_{n\to\infty} \bar{R}_\phi(\mathbb{Q}_0^n, \mathbb{Q}_1^n) \leq \limsup_{n\to\infty} \bar{R}_\phi(\mathbb{Q}_{0,R}^n, \mathbb{Q}_{1,R}^n) + 2\delta \leq \bar{R}_\phi(\mathbb{Q}_{0,R}, \mathbb{Q}_{1,R}) + 2\delta \leq \bar{R}_\phi(\mathbb{Q}_0, \mathbb{Q}_1) + 3\delta$$

Because δ was arbitrary, the result follows.

■

Next we consider an approximation of $\mathbb{P}_0$, $\mathbb{P}_1$ by compactly supported measures.

Lemma 50 *Let $\mathbb{P}_0, \mathbb{P}_1$ be finite measures. Define $\mathbb{P}_i^n = \mathbb{P}_i|_{\overline{B_n(\mathbf{0})}}$ for $n \in \mathbb{N}$. Then $\mathbb{P}_0^n, \mathbb{P}_1^n$ converge weakly to $\mathbb{P}_0$, $\mathbb{P}_1$ respectively. Furthermore, there are measures $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0), \mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ for which*

$$\limsup_{n\to\infty} \sup_{\substack{\mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1^n) \\ \mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)^n}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') \leq \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*) \quad (62)$$

Proof Set $\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$, $\mathbb{P}^n = \mathbb{P}_0^n + \mathbb{P}_1^n$. Notice that 2) of Theorem 46 implies that $\mathbb{P}_i^n$ converges weakly to $\mathbb{P}_i$. Let $\mathbb{P}_0^{*,n}, \mathbb{P}_1^{*,n}$ be maximizers of $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0^n) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1^n)$. Next, by Strassen's theorem (Theorem 38), $\mathbb{P}_i^n(\overline{B_r(\mathbf{0})}) \leq \mathbb{P}_i^{n,*}(\overline{B_{r+\epsilon}(\mathbf{0})})$ and thus $\mathbb{P}_i(\overline{B_r(\mathbf{0})}^C) \geq \mathbb{P}_i^n(\overline{B_r(\mathbf{0})}^C) \geq \mathbb{P}_i^{n,*}(\overline{B_{r+\epsilon}(\mathbf{0})})$. Therefore, one can apply Prokhorov's theorem (Thereom 48) to conclude that $\mathbb{P}_0^{n,*}$, $\mathbb{P}_1^{n,*}$ have subsequences $\mathbb{P}_0^{n_k,*}$, $\mathbb{P}_1^{n_k,*}$ that converge to measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ respectively. The upper semi-continuity of R_ϕ (Lemma 49) then implies that $\mathbb{P}_0^*, \mathbb{P}_1^*$ satisfy (62).

It remains to show that $\mathbb{P}_i^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_i)$. We will apply Lemma 4. Because $\mathbb{P}_i^{n_k,*} \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_i^{n_k})$ for all n_k, Lemma 4 implies that for every $f \in C_b(\mathbb{R}^d)$, $\int S_\epsilon(f) d\mathbb{P}_i^{n_k} \geq \int f d\mathbb{P}_i^{*,n_k}$. Because $\mathbb{P}_i^{n_k}$ converges weakly to $\mathbb{P}_i$ and $\mathbb{P}_i^{*,n_k}$ converges weakly to $\mathbb{P}_i^*$, one can take the limit $k \to \infty$ to conclude $\int S_\epsilon(f) d\mathbb{P}_i \geq \int f d\mathbb{P}_i^*$ for all $f \in C_b(\mathbb{R}^d)$. Lemma 4 then implies $\mathbb{P}_i^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_i)$. ■

Lemma 14 *Let ϕ be a non-increasing, lower semi-continuous loss function and let $\mathbb{P}_0, \mathbb{P}_1$ be finite Borel measures supported on $\mathbb{R}^d$. Let S_ϕ be as in* (25)*. Then*

$$\inf_{(h_0,h_1) \in S_\phi} \Theta(h_0, h_1) = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1')$$

Furthermore, there exist $\mathbb{P}_0^*, \mathbb{P}_1^*$ which attain the supremum.

Proof Let $\mathbb{P}_0^n$, $\mathbb{P}_1^n$, $\mathbb{P}_0^*, \mathbb{P}_1^*$ be the the measures described in Lemma 50. Notice that because $\mathbb{P}_0^n$, $\mathbb{P}_1^n$ are compactly supported, Lemma 13 applies. Define

$$\Theta^n(h_0, h_1) = \int S_\epsilon(h_1) d\mathbb{P}_1^n + \int S_\epsilon(h_0) d\mathbb{P}_0^n.$$

Thus Lemmas 13 and Lemma 50 imply that

$$\limsup_{n\to\infty} \inf_{(h_0,h_1)\in S_\phi} \Theta^n(h_0,h_1) = \limsup_{n\to\infty} \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0^n) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1^n)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') \leq \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*) \leq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1'). \tag{63}$$

We will show

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0,h_1) \leq \limsup_{n\to\infty} \inf_{(h_0,h_1)\in S_\phi} \Theta_n(h_0,h_1). \tag{64}$$

Equations 63 and 64 imply that

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0,h_1) \leq \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*) \leq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1'). \tag{65}$$

This relation together with weak duality (Lemma 45) imply that the inequalities in (65) are actually equalities. Therefore strong duality holds and $\mathbb{P}_0^*, \mathbb{P}_1^*$ maximizes the dual.

Next, we prove the inequality in (64). Let $\delta > 0$ be arbitrary and choose an $n \in \mathbb{N}$ for which $n > 2\epsilon$ and

$$\mathbb{P}_1(\overline{B_{n-2\epsilon}(\mathbf{0})}^C) + \mathbb{P}_0(\overline{B_{n-2\epsilon}(\mathbf{0})}^C) \leq \delta \tag{66}$$

Let $(h_0^n, h_1^n) \in S_\phi$ be functions for which

$$\Theta^n(h_0^n, h_1^n) \leq \inf_{(h_0,h_1)\in S_\phi} \Theta^n(h_0, h_1) + \delta \tag{67}$$

Define

$$\tilde{h}_0^n = \begin{cases} h_0^n & \mathbf{x} \in \overline{B_{n-\epsilon}(\mathbf{0})} \\ C_\phi^*\left(\frac{1}{2}\right) & \mathbf{x} \not\in \overline{B_{n-\epsilon}(\mathbf{0})} \end{cases} \quad \tilde{h}_1^n = \begin{cases} h_1^n & \mathbf{x} \in \overline{B_{n-\epsilon}(\mathbf{0})} \\ C_\phi^*\left(\frac{1}{2}\right) & \mathbf{x} \not\in \overline{B_{n-\epsilon}(\mathbf{0})} \end{cases}$$

Because $\eta h_0^n + (1-\eta)h_1^n \geq C_\phi^*(\eta)$ $\forall \eta \in [0,1]$ on $B_{n-\epsilon}(\mathbf{0})$ and $(C_\phi^*(1/2), C_\phi^*(1/2)) \in S_\phi$, one can conclude that $(\tilde{h}_0^n, \tilde{h}_1^n) \in S_\phi$.

Now because $n > 2\epsilon$, the regions $\overline{B_{n-\epsilon}(\mathbf{0})}, \overline{B_{n-2\epsilon}(\mathbf{0})}$ are non-empty. One can bound $S_\epsilon(\tilde{h}_i)$ in terms of $S_\epsilon(h_i)$ and $C_\phi^*(1/2)$:

$$\begin{aligned} &S_\epsilon(\tilde{h}_i)(\mathbf{x}) = S_\epsilon(h_i)(\mathbf{x}) && \text{for } \mathbf{x} \in \overline{B_{n-2\epsilon}(\mathbf{0})} \\ &S_\epsilon(\tilde{h}_i)(\mathbf{x}) \leq \max(S_\epsilon(h_i)(\mathbf{x}), C_\phi^*(1/2)) \leq S_\epsilon(h_i) + C_\phi^*(1/2) && \text{for } \mathbf{x} \in \overline{B_n(\mathbf{0})} \\ &S_\epsilon(\tilde{h}_i) = C_\phi^*(1/2) && \text{for } \mathbf{x} \in \overline{B_n(\mathbf{0})}^C \end{aligned}$$

Now for each i, these bounds imply that

$$\int S_\epsilon(\tilde{h}_i^n)d\mathbb{P}_i \leq \int_{\overline{B_{n-2\epsilon}(\mathbf{0})}} S_\epsilon(h_i^n)d\mathbb{P}_i + \int_{\overline{B_n(\mathbf{0})} - \overline{B_{n-2\epsilon}(\mathbf{0})}} S_\epsilon(h_i^n) + C_\phi^*\left(\frac{1}{2}\right)d\mathbb{P}_i + \int_{\overline{B_n(\mathbf{0})}^C} C_\phi^*\left(\frac{1}{2}\right)d\mathbb{P}_i$$
$$= \int_{\overline{B_n(\mathbf{0})}} S_\epsilon(h_i^n)d\mathbb{P}_i + \int_{\overline{B_{n-2\epsilon}(\mathbf{0})}^C} C_\phi^*\left(\frac{1}{2}\right)d\mathbb{P}_i$$

Then, applying this bound for each i,

$$\Theta(\tilde{h}_0^n, \tilde{h}_1^n) = \int S_\epsilon(\tilde{h}_1^n)d\mathbb{P}_1 + \int S_\epsilon(\tilde{h}_0^n)d\mathbb{P}_0$$
$$\leq \left(\int_{\overline{B_n(\mathbf{0})}} S_\epsilon(h_1^n)d\mathbb{P}_1 + \int_{\overline{B_n(\mathbf{0})}} S_\epsilon(h_0^n)d\mathbb{P}_0\right) + \left(\int_{\overline{B_{n-2\epsilon}(\mathbf{0})}^C} C_\phi^*\left(\frac{1}{2}\right)d\mathbb{P}_1 + \int_{\overline{B_{n-2\epsilon}(\mathbf{0})}^C} C_\phi^*\left(\frac{1}{2}\right)d\mathbb{P}_0\right)$$
$$= \Theta^n(h_0^n, h_1^n) + C_\phi^*\left(\frac{1}{2}\right)\left(\mathbb{P}_0(\overline{B_{n-2\epsilon}(\mathbf{0})}^C) + \mathbb{P}_1(\overline{B_{n-2\epsilon}(\mathbf{0})}^C)\right) \leq \left(\inf_{(h_0,h_1)\in S_\phi} \Theta^n(h_0,h_1) + \delta\right) + \delta C_\phi^*\left(\frac{1}{2}\right)$$

The last inequality follows from Equations 66 and 67. Because δ arbitrary, (64) holds. ■

Appendix F. Complimentary Slackness

Lemma 15 *Assume that $\mathbb{P}_0, \mathbb{P}_1$ are compactly supported. The functions h_0^*, h_1^* minimize Θ over S_ϕ and $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ maximize $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ iff the following hold:*

1)

$$\int h_1^* d\mathbb{P}_1^* = \int S_\epsilon(h_1^*) d\mathbb{P}_1 \quad \textit{and} \quad \int h_0^* d\mathbb{P}_0^* = \int S_\epsilon(h_0^*) d\mathbb{P}_0 \tag{30}$$

2) If we define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ *and* $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$, *then*

$$\eta^*(\mathbf{x}) h_1^*(\mathbf{x}) + (1 - \eta^*(\mathbf{x})) h_0^*(\mathbf{x}) = C_\phi^*(\eta^*(\mathbf{x})) \quad \mathbb{P}^*\text{-}a.e. \tag{31}$$

Notice that the forward direction of this lemma is actually a consequence of the approximate complimentary slackness result in Lemma 16, but we provide a separate self-contained proof below.

Proof

First assume that $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ maximizes $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ and (h_0^*, h_1^*) minimizes Θ over S_ϕ. Because $\mathbb{P}_i^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_i)$ and $(h_0^*, h_1^*) \in S_\phi$, by Lemma 3

$$\Theta(h_0^*, h_1^*) = \int S_\epsilon(h_1^*) d\mathbb{P}_1 + \int S_\epsilon(h_0^*) d\mathbb{P}_0 \geq \int h_1^* d\mathbb{P}_1^* + \int h_0^* d\mathbb{P}_0^* \tag{68}$$

$$= \int \eta^* h_1^* + (1 - \eta^*)\, h_0^* d\mathbb{P}^* \geq \int C_\phi^*(\eta^*) d\mathbb{P}^* = \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*) \tag{69}$$

By Lemma 14, both the first expression of (68) and the last expression of (69) are equal. Thus all the inequalities above must be equalities which implies (31). Next, because (69) implies that

$$\int S_\epsilon(h_1^*) d\mathbb{P}_1 + \int S_\epsilon(h_0^*) d\mathbb{P}_0 = \int h_1^* d\mathbb{P}_1^* + \int h_0^* d\mathbb{P}_0^*$$

and Lemma 3 implies that $\int S_\epsilon(h_0^*) d\mathbb{P}_0 \geq \int h_0^* d\mathbb{P}_0^*$ and $\int S_\epsilon(h_1^*) d\mathbb{P}_1 \geq \int h_1^* d\mathbb{P}_1^*$ we can conclude (30).

We will now show the opposite implication. Assume that $h_0^*, h_1^*, \mathbb{P}_0^*, \mathbb{P}_1^*$ satisfy (30) and (31). Then

$$\begin{aligned}
\Theta(h_0^*, h_1^*) &= \int S_\epsilon(h_1^*) d\mathbb{P}_1 + \int S_\epsilon(h_0^*) d\mathbb{P}_0 \\
&= \int h_1^* d\mathbb{P}_1^* + \int h_0^* d\mathbb{P}_0^* && \text{(Equation 30)} \\
&= \int \eta^* h_1^* + (1 - \eta^*) h_0^* d\mathbb{P}^* = \int C_\phi^*(\eta^*) d\mathbb{P}^* && \text{(Equation 31)} \\
&= \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*)
\end{aligned}$$

However, Lemma 14 implies that $\Theta(h_0, h_1) \geq \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1')$ for *any* $h_0, h_1, \mathbb{P}_0', \mathbb{P}_1'$. Therefore, h_0^*, h_1^* must be optimal for Θ and $\mathbb{P}_0^*, \mathbb{P}_1^*$ must be optimal for $\bar{R}_\phi$.

■

Notably, a similar strategy shows that if $(h_0^n, h_1^n) \in S_\phi$ is a sequence that satisfies 1) and 2) of Lemma 16, then (h_0^n, h_1^n) must be a minimizing sequence for Θ.

Appendix G. Technical Lemmas from Section 6

G.1 Proof of Lemma 17

Lemma 17 *Let $\psi(\alpha) = e^{-\alpha}$. Then $C_\psi^*(\eta) = 2\sqrt{\eta(1-\eta)}$ and $\alpha_\psi(\eta) = 1/2 \log(\eta/1-\eta)$ is the unique minimizer of $C_\psi(\eta, \cdot)$, with $\alpha_\psi(0), \alpha_\psi(1)$ interpreted as $-\infty$, $+\infty$ respectively. Furthermore, $\partial C_\psi^*(\eta)$ is the singleton $\partial C_\psi^*(\eta) = \{\psi(\alpha_\psi(\eta)) - \psi(-\alpha_\psi(\eta))\}$.*

Proof First, one can verify that $-\infty$ minimizes $C_\psi(0, \alpha)$ and ∞ minimizes $C_\psi(1, \alpha)$, and that $C_\psi^*(0) = C_\psi^*(1) = 0$. To find minimizers of $C_\psi(\eta, \alpha)$ for $\eta \in (0, 1)$, we solve $\partial_\alpha C_\psi(\eta, \alpha) = -\eta e^{-\alpha} + (1-\eta)e^\alpha = 0$, resulting in $\alpha_\psi(\eta) = 1/2 \log(\eta/1-\eta)$. This formula allows for computation of $C_\psi^*(\eta)$ via $C_\psi^*(\eta) = C_\psi(\eta, \alpha_\psi(\eta))$.

Next, by definition

$$\eta\psi(\alpha_\psi(\eta)) + (1-\eta)(-\psi(\alpha_\psi(\eta))) = C_\psi^*(\eta) \quad \text{and } s\psi(\alpha_\psi(\eta)) + (1-s)(-\psi(\alpha_\psi(\eta))) \geq C_\psi^*(s)$$

for all $s \in [0, 1]$. Therefore, $\psi(\alpha_\psi(\eta)) - \psi(-\alpha_\psi(\eta))$ is a supergradient of $C_\psi^*(\eta)$ at η.

The function C_ψ^* is differentiable on $(0, 1)$, and thus the superdifferential is unique on this set. To show that $\partial C_\psi^*(0)$, $\partial C_\psi^*(1)$ are singletons, it suffices to observe that

$$\lim_{\eta \to 0} \frac{d}{d\eta} C_\psi^*(\eta) = +\infty, \lim_{\eta \to 1} \frac{d}{d\eta} C_\psi^*(\eta) = -\infty.$$

■

G.2 Proof of Lemma 18

Lemma 18 *Let (a_n, b_n) be a sequence for which $a_n, b_n \geq 0$ and*

$$\eta a_n + (1-\eta)b_n \geq C_\psi^*(\eta) \text{ for all } \eta \in [0, 1] \tag{39}$$

and

$$\lim_{n \to \infty} \eta_0 a_n + (1-\eta_0)b_n = C_\psi^*(\eta_0) \tag{40}$$

for some η_0. Then $\lim_{n \to \infty} a_n = \psi(\alpha_\psi(\eta_0))$ and $\lim_{n \to \infty} b_n = \psi(-\alpha_\psi(\eta_0))$.

Proof Recall that on the extended real number line, every subsequence has a convergent subsequence. We will show that $\lim_{n \to \infty} a_n = \psi(\alpha_\psi(\eta_0))$ and $\lim_{n \to \infty} b_n = \psi(-\alpha_\psi(\eta_0))$ by proving that every convergent subsequence of $\{a_n\}$ converges to $\psi(\alpha_\psi(\eta_0))$ and every convergent subsequence of b_n converges to $\psi(\alpha_\psi(\eta_0))$.

Let a_{n_k}, b_{n_k} be a convergent subsequences of $\{a_n\}$, $\{b_n\}$ respectively. (Again, this convergence is in $\overline{\mathbb{R}}$.) Set $a = \lim_{k \to \infty} a_{n_k}$, $b = \lim_{k \to \infty} b_{n_k}$.

Then (39) (40) imply that

$$\eta a + (1-\eta)b \geq C_\psi^*(\eta) \text{ for all } \eta \in [0, 1]$$

$$\eta_0 a + (1-\eta_0)b = C_\psi^*(\eta_0) \tag{70}$$

These equations imply that $a - b \in \partial C_\psi^*(\eta_0)$ and thus

$$a - b = \psi(\alpha_\psi(\eta_0)) - \psi(-\alpha_\psi(\eta_0)) \tag{71}$$

while (70) is equivalent to

$$\eta_0 a + (1 - \eta_0) b = \eta_0 \psi(\alpha_\psi(\eta_0)) + (1 - \eta_0)\psi(-\alpha_\psi(\eta_0)) \tag{72}$$

The equations (71) and (72) comprise a system of equations in two variables with a unique solution for a and b. ■

G.3 Proof of Lemma 20

Lastly, we prove Lemma 20.

Lemma 20 *Let h_n be any sequence of functions. Then the sequence h_n satisfies*

$$\liminf_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\liminf_{n\to\infty} h_n) \tag{43}$$

and

$$\limsup_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\limsup_{n\to\infty} h_n) \tag{44}$$

Proof We start by showing (43).

$$\begin{aligned}
&\liminf_{n\to\infty} S_\epsilon(h_n)(\mathbf{x}) = \liminf_{n\to\infty} \sup_{\|\mathbf{h}\|\leq\epsilon} h_n(\mathbf{x}+\mathbf{h}) = \sup_N \inf_{n\geq N} \sup_{\|\mathbf{h}\|\leq\epsilon} h_n(\mathbf{x}+\mathbf{h}) \\
&\geq \sup_{\|\mathbf{h}\|\leq\epsilon} \sup_N \inf_{n\geq N} h_n(\mathbf{x}+\mathbf{h}) = \sup_{\|\mathbf{h}\|\leq\epsilon} \liminf_{n\to\infty} h_n(\mathbf{x}+\mathbf{h}) = S_\epsilon(\liminf_{n\to\infty} h_n)(\mathbf{x})
\end{aligned}$$

Equation 44 can then be proved by the same argument:

$$\begin{aligned}
&\limsup_{n\to\infty} S_\epsilon(h_n)(\mathbf{x}) = \limsup_{n\to\infty} \sup_{\|\mathbf{h}\|\leq\epsilon} h_n(\mathbf{x}+\mathbf{h}) = \inf_N \sup_{n\geq N} \sup_{\|\mathbf{h}\|\leq\epsilon} h_n(\mathbf{x}+\mathbf{h}) \\
&\geq \sup_{\|\mathbf{h}\|\leq\epsilon} \inf_N \sup_{n\geq N} h_n(\mathbf{x}+\mathbf{h}) = \sup_{\|\mathbf{h}\|\leq\epsilon} \limsup_{n\to\infty} h_n(\mathbf{x}+\mathbf{h}) = S_\epsilon(\limsup_{n\to\infty} h_n)(\mathbf{x})
\end{aligned}$$

■

References

P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. *NeurIps*, 2021.

P. Awasthi, N. S. Frank, and M. Mohri. On the existence of the adversarial bayes classifier (extended version). 2023.

H. Bao, C. Scott, and M. Sugiyama. Calibrated surrogate losses for adversarially robust classification. *arxiv*, 2021.

V. Barbu and T. Precupanu. *Convexity and Optimization in Banach Spaces*. Springer Monographs in Mathematics, 4th edition, 2012.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473), 2006.

S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer System Sciences*, 2003.

D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.

A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pages 7498–7510, 2019.

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

V. I. Bogachev. *Measure Theory*, volume II. Springer, 2007.

L. Bungert, N. G. Trillos, and R. Murray. The geometry of adversarial training in binary classification. *arxiv*, 2021.

A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. *CoRR*, 2018.

C. Domingo-Enrich, S. Jelassi, A. Mensch, G. Rotskoff, and J. Bruna. A mean-field analysis of two-player zero-sum games, 2021.

G. B. Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

N. S. Frank and J. Niles-Weed. The adversarial consistency of surrogate risks for binary classification. 2023.

R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization, 2022.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2014.

T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, 2017.

H. Jylhä. The l^∞ optimal transport: Infinite cyclical monotonicity and the existence of optimal transport maps. *Calculus of Variations and Partial Differential Equations*, 2014.

H. Kannan, A. Kurakin, and I. J. Goodfellow. Adversarial logit pairing. *CoRR*, 2018.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ICLR*, 2017.

J. D. Li and M. Telgarsky. On achieving optimal adversarial test error, 2023.

Y. Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73–82, 2004.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2019.

L. Meunier, R. Ettedgui, R. Pinot, Y. Chevaleyre, and J. Atif. Towards consistency in adversarial classification. *arXiv*, 2022.

S. A. Mingyuan Zhang. Consistency vs. h-consistency: The interplay between surrogate loss functions and the scoring function class. *NeurIps*, 2020.

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach, 2019.

N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples. *CoRR*, abs/1602.02697, 2016. URL `http://arxiv.org/abs/1602.02697`.

R. A. S. Philip M. Long. Consistency versus realizable h-consistency for multiclass classification. *ICML*, 2013.

M. S. Pydi and V. Jog. Adversarial risk via optimal transport and optimal couplings. *arXiv preprint arXiv:1912.02794*, 2019.

M. S. Pydi and V. Jog. The many faces of adversarial risk. *Neural Information Processing Systems*, 2021.

A. Rozsa, M. Günther, and T. E. Boult. Are accuracy and robustness correlated? *CoRR*, 2016.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 1st edition, 2015.

A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. P. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *CoRR*, 2019.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 2007.

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

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(36), 2007.

F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. The space of transferable adversarial examples. *arXiv*, 2017.

C. G. Trillos and N. G. Trillos. On adversarial robustness and the use of wasserstein ascent-descent dynamics to enforce it, 2023.

N. G. Trillos and R. Murray. Adversarial classification: Necessary conditions and geometric flows. *arxiv*, 2020.

N. G. Trillos, M. Jacobs, and J. Kim. The multimarginal optimal transport formulation of adversarial multiclass classification. *arXiv*, 2022.

C. Villani. *Topics in Optimal Transportation.* American Mathematical Society, 2nd edition, 2003.

G. Wang and L. Chizat. An exponentially converging particle method for the mixed nash equilibrium of continuous games, 2023.

Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu. On the convergence and robustness of adversarial training. *ICML*, 2021.

E. Wong, F. Schmidt, and Z. Kolter. Wasserstein adversarial examples via projected Sinkhorn iterations. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2019.

E. Wong, L. Rice, and J. Z. Kolter. Fast is better than free: Revisiting adversarial training. *CoRR*, abs/2001.03994, 2020.

K. Wu, A. H. Wang, and Y. Yu. Stronger and faster wasserstein adversarial attacks, 2020.

C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille, and K. He. Feature denoising for improving adversarial robustness. *CoRR*, 2018.

Y.-Y. Yang, C. Rashtchian, Y. Wang, and K. Chaudhuri. Robustness for non-parametric classification: A generic attack and defense. *Proceedings of Machine Learning Research*, 2020.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 2004.