Adversarial Consistency and the Uniqueness of the Adversarial Bayes Classifier

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

Abstract

Adversarial training is a common technique for learning robust classifiers. Prior work showed that convex surrogate losses are not statistically consistent in the adversarial context— or in other words, a minimizing sequence of the adversarial surrogate risk will not necessarily minimize the adversarial classification error. We connect the consistency of adversarial surrogate losses to properties of minimizers to the adversarial classification risk, known as *adversarial Bayes classifiers*. Specifically, under reasonable distributional assumptions, a convex loss is statistically consistent for adversarial learning iff the adversarial Bayes classifier satisfies a certain notion of uniqueness.

1 Introduction

Robustness is a core concern in machine learning, as models are deployed in classification tasks such as facial recognition [Xu et al., 2022], medical imaging [Paschali et al., 2018], and identifying traffic signs in self-driving cars [Deng et al., 2020]. Deep learning models exhibit a concerning security risk— small perturbations imperceptible to the human eye can cause a neural net to misclassify an image [Biggio et al., 2013, Szegedy et al., 2013]. The machine learning literature has proposed many defenses, but many of these techniques remain poorly understood. This paper analyzes the statistical consistency of a popular defense method called *adversarial training*.

The central goal in a classification task is minimizing the proportion of mislabeled data-points— also known as the *classification risk*. Minimizers to the classification risk are easy to compute analytically, and are known as *Bayes classifiers*. In the adversarial setting, each point is perturbed by a malicious adversary before the classifier makes a prediciton. The proportion of mislabeled data under such an attack is called the *adversarial classification risk*, and minimizers to this risk are called *adversarial Bayes classifiers*. Unlike the standard classification setting, computing minimizers to the adversarial classification risk is a non-trivial task [Bhagoji et al., 2019, Pydi and Jog, 2020]. Further studies [Frank, 2024, Trillos and Murray, 2022, Trillos et al., 2023, Pydi and Jog, 2021, Gnecco-Heredia et al., 2023] investigate additional properties of these minimizers, and Frank [2024] in particular describes a notion of uniqueness for adversarial Bayes classifiers. The main result in this paper will connect this notion of uniqueness to the behavior of adversarial training.

The empirical adversarial classification error is a discrete object and minimizing this quantity is computationally intractable. Instead, typical machine learning algorithms minimize a *surrogate risk* in place of the classification error. In the robust setting, the adversarial training algorithm uses a surrogate risk that computes the supremum of loss over the adversary's possible attacks, which we refer to as *adversarial surrogate risks*. However, one must verify that minimizing this adversarial surrogate will also minimize the classification risk. A loss function is *adversarially consistent* for a particular data distribution if every minimizing sequence of the associated adversarial surrogate risk

Preprint. Under review.

also minimizes the adversarial classification risk. A loss is simply called *consistent* if it is consistent for all possible data distributions. Meunier et al. [2022] show that no convex adversarial surrogate is consistent, in contrast to the standard classification setting where most convex losses are statistically consistent [Bartlett et al., 2006, Lin, 2004, Steinwart, 2007, Mingyuan Zhang, 2020, Zhang, 2004].

Our Contributions: We relate the statistical consistency of losses in the adversarial setting to the uniqueness of the adversarial Bayes classifier. Specifically, under reasonable assumptions, a convex loss is adversarially consistent for a specific data distribution iff the adversarial Bayes classifier is unique.

Frank [2024] further demonstrates several distributions for which the adversarial Bayes classifier is unique, and thus a convex loss would be consistent. Understanding general conditions under which uniqueness occurs is an open question.

2 Related Works

Our results are inspired by prior work which showed that no convex loss is adversarially consistent [Meunier et al., 2022, Awasthi et al., 2022] yet a wide class of adversarial losses is adversarially consistent [Frank and Niles-Weed, 2024a]. These consistency results rely on the theory of surrogate losses, studied by Bartlett et al. [2006], Lin [2004] in the standard classification setting and by Frank and Niles-Weed [2024b], Li and Telgarsky [2023] in the adversarial setting. Furthermore, [Bao et al., 2021, Awasthi et al., 2021b, Steinwart, 2007] study a property of related to consistency called *calibration*, whcih [Meunier et al., 2022] relate to consistency. Complimenting this analysis, another line of research studies $\mathcal{H}$-consistency, which refines the concept of consistency to specific function classes [Philip M. Long, 2013, Awasthi et al., 2022]. Our proof combines results on losses with minimax theorems for various adversarial risks, as studied by [Frank and Niles-Weed, 2024b, Trillos et al., 2022, Pydi and Jog, 2021, Frank and Niles-Weed, 2024a]. Lastly, this work leverages recent results on the adversarial Bayes classifier, which are extensively studied by [Bhagoji et al., 2019, Trillos et al., 2022, Pydi and Jog, 2020, Frank, 2024].

3 Background

3.1 Surrogate Risks

This paper investigates binary classification on $\mathbb{R}^d$ with labels $\{-1,+1\}$. Class -1 is distributed according to a a measure $\mathbb{P}_0$ and while class $+1$ is distributed according to measure $\mathbb{P}_1$. A *classifier* is a Borel set A that specifies the region predicted to have label $+1$. The *classification risk* of a set A is the expected proportion of errors when label $+1$ is predicted on A and label -1 is predicted on A^C:

$$R(A)=\int \mathbf{1}_{A^C}d\mathbb{P}_1+\int \mathbf{1}_A d\mathbb{P}_0.$$

A minimizer to R is a *Bayes classifier*. These minimizers can be expressed in terms of the measure $\mathbb{P}=\mathbb{P}_0+\mathbb{P}_1$ and the function $\eta=d\mathbb{P}_1/d\mathbb{P}$. The risk R in terms of these quantities is

$$R(A)=\int C(\eta,\mathbf{1}_A)d\mathbb{P}.$$

and $\inf_A R(A)=\int C^*(\eta)d\mathbb{P}$ where the functions $C:[0,1]\times\{0,1\}\to\mathbb{R}$ and $C^*:[0,1]\to\mathbb{R}$ are defined by

$$C(\eta,b)=\eta b+(1-\eta)(1-b),\quad C^*(\eta)=\inf_{b\in\{0,1\}}C(\eta,b)=\min(\eta,1-\eta) \tag{1}$$

Thus if A is a minimizer of R, then $\mathbf{1}_A$ must minimize the function $C(\eta,\cdot)$ almost everywhere according to the measure $\mathbb{P}$. Consequently, the sets

$$\{\mathbf{x}:\eta(\mathbf{x})>1/2\}\quad\text{and}\quad\{\mathbf{x}:\eta(\mathbf{x})\geq 1/2\} \tag{2}$$

are both Bayes classifiers.

While the Bayes classifier can be described mathematically, minimizing the empirical classification risk is a computationally intractable problem [Ben-David et al., 2003]. A common approach is to instead minimize a better-behaved alternative called a *surrogate risk*. As a surrogate to R, we consider:

$$R_\phi(f) = \int \phi(f) d\mathbb{P}_1 + \int \phi(-f) d\mathbb{P}_0 \,. \tag{3}$$

The loss ϕ is selected so that the resulting risk is easy to optimize. We assume

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

A classifier is obtained by minimizing R_ϕ over all measurable functions and then thresholding f at 0: explicitly, the classifier is $A = \{\mathbf{x} : f(\mathbf{x}) > 0\}$. Due to this construction, we define

$$R(f) = R(\{f > 0\}) \tag{4}$$

for a function f.

One can compute the infimum of R_ϕ by expressing the risk in terms of the quantities $\mathbb{P}$ and η:

$$R_\phi(f) = \int C_\phi(\eta(\mathbf{x}), f(\mathbf{x})) d\mathbb{P} \tag{5}$$

and $\inf_f R_\phi(f) = \int C_\phi^*(\eta(\mathbf{x})) d\mathbb{P}(\mathbf{x})$ where the functions $C_\phi(\eta,\alpha)$ and $C_\phi^*(\eta)$ are defined by

$$C_\phi(\eta,\alpha) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha), \quad C_\phi^*(\eta) = \inf_\alpha C_\phi(\eta,\alpha) \tag{6}$$

for $\eta \in [0,1]$. Thus a minimizer f of R_ϕ must minimize $C_\phi(\eta(\mathbf{x}),\cdot)$ almost everywhere according to the probability measure $\mathbb{P}$. Allowing for minimizers in extended real numbers $\overline{\mathbb{R}} = \{-\infty,+\infty\}\cup\mathbb{R}$ is necessary for certain losses— For instance when ϕ is the exponential loss, then $C_\phi(1,\alpha)=e^{-\alpha}$ does not assume its infimum on $\mathbb{R}$. The following lemma describes a method for mapping $\eta(\mathbf{x})$ to a minimizer of $C_\phi(\eta(\mathbf{x}),\cdot)$.

Lemma 1. *The function $\alpha_\phi : [0,1] \to \overline{\mathbb{R}}$ that maps η to the smallest minimizer of $C_\phi(\eta,\cdot)$ is non-decreasing.*

See Appendix A for a proof. Because α_ϕ is monotonic, the composition

$$\alpha_\phi(\eta(\mathbf{x})) \tag{7}$$

is always measurable, and thus this function is a minimizer of R_ϕ.

3.2 Adversarial Surrogate Risks

In the adversarial setting, a malicious adversary corrupts each data point. We model these corruptions as bounded by ϵ in some norm $\|\cdot\|$. The adversary knows both the classifier A and the label of each data point. Thus, a point $(\mathbf{x},+1)$ is misclassified when it can be displaced into the set A^C by a perturbation of size at most ϵ. This statement can be conveniently written in terms of a supremum. For any function $g:\mathbb{R}^d\to\mathbb{R}$, define

$$S_\epsilon(g)(\mathbf{x}) = \sup_{\mathbf{x}'\in\overline{B_\epsilon(\mathbf{x})}} g(\mathbf{x}'),$$

where $\overline{B_\epsilon(\mathbf{x})}$ is the ball of allowed perturbations. The expected error rate of a classifier A under an adversarial attack is then

$$R^\epsilon(A) = \int S_\epsilon(\mathbf{1}_{A^C}) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0,$$

which is known as the *adversarial classification risk*[1]. Minimizers of R^ϵ are called *adversarial Bayes classifiers*.

Just like (4), we define $R^\epsilon(f) = R^\epsilon(\{f>0\})$ or explicitly,

$$R^\epsilon(f) = \int S_\epsilon(\mathbf{1}_{f\leq 0}) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_{f>0}) d\mathbb{P}_0$$

[1]The functions $S_\epsilon(\mathbf{1}_A)$, $S_\epsilon(\mathbf{1}_{A^C})$ must be measurable in order to define this integral. See [Frank and Niles-Weed, 2024b, Section 3.3] for a treatment of this matter.

Again, minimizing an empirical adversarial classification risk is computationally intractable. A surrogate to the adversarial classification risk is formulated as[2]

$$R^\epsilon_\phi(f) = \int S_\epsilon(\phi \circ f) d\mathbb{P}_1 + \int S_\epsilon(\phi \circ -f) d\mathbb{P}_0. \tag{8}$$

Theorem 9 of [Frank and Niles-Weed, 2024b] then extends the construction of a minimizer in (7) to the adversarial setting.

Theorem 1. *Let α_ϕ be the function in Lemma 1. Then for any distribution $\mathbb{P}_0$, $\mathbb{P}_1$, there is a function $\hat{\eta} : \mathbb{R}^d \to [0,1]$ for which $\alpha_\phi(\hat{\eta}(\mathbf{x}))$ is a minimizer of R^ϵ_ϕ for any loss ϕ.*

The function $\hat{\eta}$ can be viewed as the conditional probability of label $+1$ under an 'optimal' adversarial attack [Frank and Niles-Weed, 2024b]. Just as in the standard learning scenario, the function $\alpha(\hat{\eta}(\mathbf{x}))$ may be $\overline{\mathbb{R}}$-valued. Furthermore, recall that Bayes classifiers can be constructed by theshholding the conditional probability η at $1/2$, as in (2). The function $\hat{\eta}$ plays an analogous role for adversarial learning.

Theorem 2. *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be finite measures and let $\hat{\eta}$ be the function described by Theorem 1. Then the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are adversarial Bayes classifiers. Furthermore, any adversarial Bayes classifier A satisfies*

$$\int S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}^C}) d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}^C}) d\mathbb{P}_1 \tag{9}$$

and

$$\int S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}}) d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}}) d\mathbb{P}_0 \tag{10}$$

See Appendix C for a proof and more about the function $\hat{\eta}$. Equations (9) and (10) imply that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ can be viewed as 'minimal' and 'maximal' adversarial Bayes classifiers.

3.3 The Statistical Consistency of Surrogate Risks

Learning algorithms typically minimize a surrogate risk using an iterative procedure, thereby producing a sequence of functions f_n. One would hope that that f_n also minimizes that corresponding classification risk. This property is referred to as *statistical consistency*[3].

Definition 1.

- *If every sequence of functions f_n that minimizes R_ϕ also minimizes R for the distribution $\mathbb{P}_0, \mathbb{P}_1$, then the loss ϕ is* consistent for the distribution $\mathbb{P}_0, \mathbb{P}_1$. *If R_ϕ is consistent for every distribution $\mathbb{P}_0, \mathbb{P}_1$, we say that ϕ is* consistent.
- *If every sequence of functions f_n that minimizes R^ϵ_ϕ also minimizes R^ϵ for the distribution $\mathbb{P}_0, \mathbb{P}_1$, then the loss ϕ is* adversarially consistent for the distribution $\mathbb{P}_0, \mathbb{P}_1$. *If R^ϵ_ϕ is adversarially consistent for every distribution $\mathbb{P}_0, \mathbb{P}_1$, we say that ϕ is* adversarially consistent.

A case of particular interest is convex ϕ, as these losses are particularly ubiquitous in machine learning. In the non-adversarial context, Theorem 2 of [Bartlett et al., 2006] shows that a convex loss ϕ is consistent iff ϕ is differentiable at zero and $\phi'(0) < 0$. In contrast, Meunier et al. [2022] show that no convex loss is adversarially consistent. Further results of [Frank and Niles-Weed, 2024a] characterize the adversarially consistent losses in terms of the function C^*_ϕ:

Theorem 3. *The loss ϕ is adversarially consistent if and only if $C^*_\phi(1/2) < \phi(0)$.*

Notice that all convex losses satisfy $C^*_\phi(1/2) = \phi(0)$: By evaluating at $\alpha = 0$, one can conclude that $C^*_\phi(1/2) = \inf_\alpha C_\phi(1/2, \alpha) \leq C_\phi(1/2, 0) = \phi(0)$. However, due to convexity,

$$C^*_\phi(1/2) = \inf_\alpha \frac{1}{2}\phi(\alpha) + \frac{1}{2}\alpha(-\alpha) \geq \phi(0)$$

[2] Again, see See [Frank and Niles-Weed, 2024b, Section 3.3] for a treatment of measurability.

[3] This concept is referred to as *calibration* in the non-adversarial machine learning context [Bartlett et al., 2006, Steinwart, 2007]. We use the term 'consistent', as in the context of adversarial learning, prior work [Awasthi et al., 2021a, Meunier et al., 2022] use 'calibration' to refer to a different but related concept.

Notice that Theorem 3 does not preclude the adversarial consistency of a loss satisfying $C_\phi^*(1/2) = \phi(0)$ for any particular $\mathbb{P}_0, \mathbb{P}_1$. Prior work [Meunier et al., 2022, Frank and Niles-Weed, 2024a] provides a counterexample to consistency only for a single, atypical distribution. The goal of this paper is characterizing when adversarial consistency fails for losses satisfying $C_\phi^*(1/2) = \phi(0)$.

4 Main Result

Prior work has shown that there always exists minimizers to the adversarial classification risk, which are referred to as *adversarial Bayes classifiers* (see Theorem 5 below). Frank [2024] further develops a notion of uniqueness for the adversarial Bayes classifier.

Definition 2. *The adversarial Bayes classifiers A_1 and A_2 are* equivalent up to degeneracy *if any Borel set A with $A_1 \cap A_2 \subset A \subset A_1 \cup A_2$ is also an adversarial Bayes classifier. The adversarial Bayes classifier is* unique up to degeneracy *if any two adversarial Bayes classifiers are unique up to degeneracy.*

When $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, then equivalence up to degeneracy is an equivalence relation [Frank, 2024]. The central result of this paper relates the consistency of convex losses to the uniqueness of the adversarial Bayes classifier.

Theorem 4. *Assume that $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure and let ϕ be a loss with $C_\phi^*(1/2) = \phi(0)$. Then ϕ is adversarially consistent for the distribution $\mathbb{P}_0$, $\mathbb{P}_1$ iff the adversarial Bayes classifier is unique up to degeneracy.*

Uniqueness up to degeneracy has a number of other useful characterizations (see Theorem 8. The reverse direction of this theorem is proven in Section 6 through a direct application of Definition 2 while the forward direction in Section 5 applies a formulation of uniqueness up to degeneracy in terms of the dual problem.

Frank [2024] provides the tools for verifying when the adversarial Bayes classifier is unique up to degeneracy for a wide class of one dimensional distributions. Below we highlight two interesting examples. Let p_1 be the density of $\mathbb{P}_1$ and p_0 be the density of $\mathbb{P}_0$.

- Consider mean zero gaussians with different variances: $p_0(x) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-x^2/2\sigma_0^2}$ and $p_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-x^2/2\sigma_1^2}$. The adversarial Bayes classifier is unique up to degeneracy for all ϵ for this distribution.
- Consider gaussians with variance σ and means μ_0 and μ_1: $p_0(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_0)^2/2\sigma^2}$ and $p_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_1)^2/2\sigma^2}$. Then the adversarial Bayes classifier is unique up to degeneracy iff $\epsilon < |\mu_1 - \mu_0|/2$.

Theorem 4 implies that a convex loss is always adversarially consistent for the first gaussian mixture above. Furthermore, a convex loss is adversarially consistent for the second gaussian mixture when the perturbation radius ϵ is small compared to the differences between the means. However, Frank [2024] provide an example of a distribution for which the adversarial Bayes classifier is not unique up to degeneracy for all $\epsilon > 0$. Understanding when the adversarial Bayes classifier is unique up to degeneracy for reasonable distributions is an open problem.

5 Uniqueness up to Degeneracy implies Consistency

The proof of the forward direction in Theorem 4 relies on a dual formulation of the adversarial classification problem involving the Wasserstein-∞ metric. This tool is presented in the next section and is then used to prove the forward direction of Theorem 4 in Section 5.2.

5.1 Background— A Dual Problem for the Adversarial Classification Risk

Informally, a measure $\mathbb{Q}'$ is within ϵ of $\mathbb{Q}$ in the Wasserstein-∞ metric if one can produce $\mathbb{Q}'$ by perturbing each point in $\mathbb{R}^d$ by at most ϵ under the measure $\mathbb{Q}$. The formal definition of the Wasserstein$-\infty$ metric relies on couplings between probability measures: a *coupling* between two

Borel measures $\mathbb{Q}$ and $\mathbb{Q}'$ with $\mathbb{Q}(\mathbb{R}^d) = \mathbb{Q}'(\mathbb{R}^d)$ is a measure γ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mathbb{Q}$ and $\mathbb{Q}'$: $\gamma(A \times \mathbb{R}^d) = \mathbb{Q}(A)$ and $\gamma(\mathbb{R}^d \times A) = \mathbb{Q}'(A)$ for any Borel set A. The set of all such couplings is denoted $\Pi(\mathbb{Q}, \mathbb{Q}')$. The ∞-Wasserstein distance between two measures is then

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

[Jylhä, 2014, Theorem 2.6] proves that this infimum is always assumed. Therefore, $W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon$ iff there is a coupling between $\mathbb{Q}$ and $\mathbb{Q}'$ supported on the set

$$\Delta_\epsilon = \{(\mathbf{x}, \mathbf{x}') : \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon\}.$$

Let $\mathcal{B}_\epsilon^\infty(\mathbb{Q}) = \{\mathbb{Q} : W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon\}$ be the set of measures within ϵ of $\mathbb{Q}$ in the W_∞ metric. The minimax relations from prior work leverage a relationship between the Wasserstein-∞ metric and the integral of the supremum function over an ϵ-ball.

Lemma 2. *Let E be a Borel set. Then*

$$\int S_\epsilon(\mathbf{1}_E) d\mathbb{Q} \geq \sup_{\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})} \int \mathbf{1}_E d\mathbb{Q}'$$

See Appendix B for a proof. Consequently,

$$\inf_f R^\epsilon(f) \geq \inf_f \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon(\mathbb{P}_1)}} \int \mathbf{1}_{f \leq 0} d\mathbb{P}_1' + \int \mathbf{1}_{f > 0} d\mathbb{P}_0.$$

Does equality hold and can one swap the infimum and the supremum? [Frank and Niles-Weed, 2024a, Pydi and Jog, 2021] answer this question in the affirmative:

Theorem 5. *Let $\mathbb{P}_0$, $\mathbb{P}_1$ be finite Borel measures. Define*

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

where the function C^* is defined in (1). Then

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

and furthermore equality is attained for some f^*, $\mathbb{P}_0^*$, $\mathbb{P}_1^*$.

See Theorem 1 of [Frank and Niles-Weed, 2024a]. Frank and Niles-Weed [2024b] prove an analogous minimax theorem for surrogate risks.

Theorem 6. *Let $\mathbb{P}_0$, $\mathbb{P}_1$ be finite Borel measures. Define*

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

where the function C_ϕ^* is defined in (6). Then

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

and furthermore equality is attained for some f^*, $\mathbb{P}_0^*$, $\mathbb{P}_1^*$.

Just like R_ϕ, the risk R_ϕ^ϵ may not have an $\mathbb{R}$-valued minimizer. However, Lemma 8 of [Frank and Niles-Weed, 2024a] states that

$$\inf_{\substack{f \text{ Borel} \\ \overline{\mathbb{R}}\text{-valued}}} R_\phi^\epsilon(f) = \inf_{\substack{f \text{ Borel} \\ \mathbb{R}\text{-valued}}} R_\phi^\epsilon(f).$$

Additionally, there exists a maximizer to $\bar{R}_\phi$ with especially nice properties. Let I_ϵ denote the infimum of a function over an ϵ ball:

$$I_\epsilon(g) = \inf_{\mathbf{x}' \in \overline{B_\epsilon(\mathbf{x})}} g(\mathbf{x}') \tag{11}$$

Lemma 24 of [Frank and Niles-Weed, 2024b] proves the following result:

Theorem 7. *There exists a function $\hat{\eta} : \mathbb{R}^d \to [0,1]$ and 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*

I) $\hat{\eta} = \eta^*$ $\mathbb{P}^*$-a.e., where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$

II) $I_\epsilon(\hat{\eta})(\mathbf{x}) = \hat{\eta}(\mathbf{x}')$ γ_0^*-a.e. and $S_\epsilon(\hat{\eta})(\mathbf{x}) = \hat{\eta}(\mathbf{x}')$ γ_1^*-a.e., where γ_0^*, γ_1^* are couplings between $\mathbb{P}_0$, $\mathbb{P}_0^*$ and $\mathbb{P}_1$, $\mathbb{P}_1^*$ supported on Δ_ϵ.

This result implies Theorem 1: Item I) and Item II) imply that $R_\phi^\epsilon(\alpha_\phi(\hat{\eta})) = \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*)$ and Theorem 6 then implies that $\alpha_\phi(\hat{\eta})$ is a minimizer of R_ϕ^ϵ and $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ maximize $\bar{R}_\phi$. (A similar argument is given in the proof of Lemma 9 of Appendix F.1 in this paper.) Thus we obtain

Lemma 3. *The $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ of Theorem 7 maximize $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ for every ϕ.*

See [Frank and Niles-Weed, 2024b, Lemma 26] for more details. Theorem 2 is proves analogously to Theorem 1 in Appendix C– Item I) and Item II) imply that $R^\epsilon(\{\hat{\eta} > 1/2\}) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)$, $R^\epsilon(\hat{\eta} \geq 1/2\}) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)$ and consequently Theorem 5 implies that $\{\hat{\eta} > 1/2\}$, $\{\hat{\eta} \geq 1/2\}$ minimize R^ϵ and $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ maximize $\bar{R}$. Lastly, uniqueness up to degeneracy can be characterized in terms of these measures $\mathbb{P}_0^*$,$\mathbb{P}_1^*$. See Appendix D for a proof of Theorem 8.

Theorem 8. *Assume that $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure. Then the following are equivalent:*

A) The adversarial Bayes classifier is unique up to degeneracy

B) Amongst all adversarial Bayes classifiers A, either the value of $\mathbb{P}_0(A^\epsilon)$ is unique or the value of $\mathbb{P}_1((A^C)^\epsilon)$ is unique

C) $\mathbb{P}^*(\eta^* = 1/2) = 0$, where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$ for the measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ of Theorem 7.

5.2 Proving that Uniqueness implies Consistency

Before presenting the full proof of consistency, we provide an overview the strategy of this argument. [Frank and Niles-Weed, 2024a] derive a conditions that describe minimizing sequences of R_ϕ^ϵ:

Proposition 1. *Assume that the measures $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ maximize $\bar{R}_\phi$. Then any minimizing sequence f_n of R_ϕ^ϵ must satisfy*

$$\lim_{n\to\infty} \int C_\phi(\eta^*, f_n) d\mathbb{P}^* = \int C_\phi^*(\eta^*) d\mathbb{P}^* \tag{12}$$

$$\lim_{n\to\infty} \int S_\epsilon(\phi\circ f_n) d\mathbb{P}_1 - \lim_{n\to\infty} \int \phi\circ f_n d\mathbb{P}_1^* = 0, \quad \lim_{n\to\infty} \int S_\epsilon(\phi\circ -f_n) d\mathbb{P}_0^* - \lim_{n\to\infty} \int \phi\circ -f_n d\mathbb{P}_0^* = 0 \tag{13}$$

where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$.

We will show that when $\mathbb{P}^*(\eta^* = 1/2) = 0$, every sequence of functions satisfying (12) and (13) must minimize R^ϵ. Specifically, we will prove that every minimizing sequence f_n of R_ϕ^ϵ must satisfy

$$\limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \int \mathbf{1}_{\eta^* \leq \frac{1}{2}} d\mathbb{P}_1^* \tag{14}$$

and

$$\limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n \geq 0}) d\mathbb{P}_0 \leq \int \mathbf{1}_{\eta^* \geq \frac{1}{2}} d\mathbb{P}_0^* \tag{15}$$

for the measures $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ in Theorem 7. Consequently, $\mathbb{P}^*(\eta^* = 1/2) = 0$ implies that $\limsup_{n\to\infty} R^\epsilon(f_n) \leq \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)$ and the strong duality relation Theorem 5 implies that f_n must in fact be a minimizing sequence of R^ϵ.

We next describe the proof of (14). To simplify the discussion, we assume that the functions ϕ, α_ϕ are strictly monotonic and that for each η, there is a unique value of α for which $\eta\phi(\alpha) + (1-\eta)\phi(-\alpha) = C_\phi^*(\eta)$. (For instance, the exponential loss $\phi(\alpha) = e^{-\alpha}$ satisfies these requirements.) Let γ_1^* be a coupling between $\mathbb{P}_1$ and $\mathbb{P}_1^*$ supported on Δ_ϵ.

Because $C_\phi(\eta^*, f_n) \geq C_\phi^*(\eta^*)$, the condition (12) implies that $C_\phi(\eta^*, f_n)$ converges to $C_\phi^*(\eta^*)$ in $L^1(\mathbb{P}^*)$, and the assumption that there is a single value of $\phi(\alpha)$ for which $\eta\phi(\alpha)+(1-\eta)\phi(-\alpha) = C_\phi^*(\eta)$ implies that the function $\phi(f_n(\mathbf{x}'))$ must converge to $\phi(\alpha_\phi(\eta^*(\mathbf{x}'))$ in $L^1(\mathbb{P}_1^*)$. Similarly, because Lemma 2 implies that $S_\epsilon(\phi\circ f_n)(\mathbf{x}) \geq \phi\circ f_n(\mathbf{x}')$ γ_1^*-a.e., (13) implies that $S_\epsilon(\phi\circ f_n)(\mathbf{x}) - \phi\circ f_n(\mathbf{x}')$ converges to 0 in $L^1(\gamma_1^*)$. Consequently $S_\phi(\phi\circ f_n)(\mathbf{x})$ must converge to $\phi(\alpha_\phi(\eta^*))$ in $L^1(\gamma_1^*)$. As L^1 convergence implies convergence in measure [Folland, 1999, Proposition 2.29], one can conclude that

$$\lim_{n\to\infty} \gamma_1^*(S_\epsilon(\phi\circ f_n)(\mathbf{x}) - \phi\circ\alpha_\phi(\mathbf{x}') > c) = 0$$

for any $c > 0$. The lower semi-continuity of $\alpha \mapsto \mathbf{1}_{\alpha\leq 0}$ implies that $\int S_\epsilon(\mathbf{1}_{f_n\geq 0})d\mathbb{P}_1 \leq \int \mathbf{1}_{S_\epsilon(f_n)(\mathbf{x})\geq\phi(0)}d\mathbb{P}_1$ and therefore

$$\lim_{n\to\infty}\int \mathbf{1}_{S_\epsilon(f_n)(\mathbf{x})\geq\phi(0)}d\gamma_1^* \leq \lim_{n\to\infty}\int \mathbf{1}_{\phi(\alpha_\phi(\eta^*(\mathbf{x})))<-c}d\gamma_1 = \int \mathbf{1}_{\eta^*\leq\alpha_\phi^{-1}\circ\phi^{-1}(-c)}. \tag{16}$$

Due to our assumptions on ϕ and α_ϕ, the quantity $\alpha_\phi^{-1}\circ\phi^{-1}(-c)$ is strictly larger than 0. However, one can choose c small enough so that $\mathbb{P}^*(|\eta - 1/2| < \alpha_\phi^{-1}\circ\phi^{-1}(-c)) < \delta$ for any $\delta > 0$ when α_ϕ is continuous and $\mathbb{P}^*(\eta^*) = 0$. This choice for c along with (16) proves (14).

To avoid the prior assumptions on ϕ and α, we we prove that when η is bounded away from $1/2$ and α is bounded away from the minimizers of $C_\phi(\eta,\cdot)$, then $C_\phi(\eta,\alpha)$ is bounded away from $C_\phi^*(\eta)$.

Lemma 4. *Let ϕ be a consistent loss. For all $r > 0$, there is a constant $k_r > 0$ and an $\alpha_r > 0$ for which if $|\eta - 1/2| \geq r$ and $\alpha\,\mathrm{sign}(\eta-1/2) \leq \alpha_r\,\mathrm{sign}(\eta - 1/2)$ then $C_\phi(\eta,\alpha_r) - C_\phi^*(\eta) \geq k_r$, and this α_r satisfies $\phi(\alpha_r) < \phi(0)$.*

See Appendix E for a proof. A minor modification of this argument proves our main result:

Proposition 2. *Assume there exist $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$ that maximize $\bar{R}_\phi$ for which $\mathbb{P}^*(\eta^* = 1/2) = 0$. Then any consistent loss is adversarially consistent.*

When $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, uniqueness up to degeneracy of the adversarial Bayes classifier implies the conditions of this propositon.

Proof. We will show that every minimizing sequence of R_ϕ^ϵ must satisfy (14) and (15). These equations together with the assumption $\mathbb{P}^*(\eta^* = 1/2) = 0$ imply that

$$\limsup_{n\to\infty} R^\epsilon(f_n) \leq \int \mathbf{1}_{\eta^*\leq\frac{1}{2}}d\mathbb{P}_1^* + \int \mathbf{1}_{\eta^*\geq\frac{1}{2}}d\mathbb{P}_0^* = \int \eta^*\mathbf{1}_{\eta^*\leq 1/2} + (1-\eta)^*\mathbf{1}_{\eta^*>1/2}d\mathbb{P}^* = \bar{R}(\mathbb{P}_0^*,\mathbb{P}_1^*).$$

The strong duality result of Theorem 5 then implies that f_n must be a minimizing sequence of R^ϵ.

Let δ be arbitrary and due to the assumption $\mathbb{P}^*(\eta^* = 1/2) = 0$, one can pick an r for which

$$\mathbb{P}^*(|\eta^* - 1/2| < r) < \delta. \tag{17}$$

Next, let α_r, k_r be as in Lemma 4.

Let γ_i^* be couplings between $\mathbb{P}_i$ and $\mathbb{P}_i^*$ supported on Δ_ϵ. Lemma 2 implies that $S_\epsilon(\phi\circ f_n)(\mathbf{x}) \geq \phi\circ f_n(\mathbf{x}')$ γ_1^*-a.e., and thus (13) implies that $S_\epsilon(\phi\circ f_n)(\mathbf{x}) - \phi\circ f_n(\mathbf{x}')$ converges to 0 in $L^1(\gamma_1^*)$. Because convergence in L^1 implies convergence in measure [Folland, 1999, Proposition 2.29], $S_\epsilon(\phi\circ f_n)(\mathbf{x}) - \phi\circ f_n(\mathbf{x}')$ converges to 0 in γ_1^*-measure. Similarly, one can conclude that $S_\epsilon(\phi\circ -f_n)(\mathbf{x}) - \phi\circ -f_n(\mathbf{x}')$ converges to zero in γ_0^*-measure. Additionally, as $C_\phi^*(\eta^*, f_n) \geq C_\phi^*(\eta^*)$, (12) implies that $C_\phi^*(\eta^*, f_n)$ converges in $\mathbb{P}^*$-measure to $C_\phi^*(\eta^*)$. Therefore, Proposition 1 implies that one can choose N large enough so that $n > N$ implies that

$$\gamma_1^*\left(S_\epsilon(\phi\circ f_n)(\mathbf{x}) - \phi\circ f_n(\mathbf{x}') \geq \phi(0) - \phi(\alpha_r)\right) < \delta \tag{18}$$

$$\gamma_0^*\left(S_\epsilon(\phi\circ -f_n)(\mathbf{x}) - \phi\circ -f_n(\mathbf{x}') \geq \phi(0) - \phi(\alpha_r)\right) < \delta \tag{19}$$

and $\mathbb{P}^*(C_\phi^*(\eta^*, f_n) > C_\phi^*(\eta^*) + k_r) < \delta$. The relation $\mathbb{P}^*(C_\phi^*(\eta^*, f_n) > C_\phi^*(\eta^*) + k_r) < \delta$ implies that

$$\mathbb{P}^*\big(|\eta^* - 1/2| \geq r, f_n\,\mathrm{sign}(\eta^* - 1/2) \leq \alpha_r\,\mathrm{sign}(\eta^* - 1/2)\big) < \delta \tag{20}$$

due to Lemma 4. Because ϕ is non-increasing, $\mathbf{1}_{f_n\leq 0} \leq \mathbf{1}_{\phi\circ f_n \geq \phi(0)}$ and since the function $z \mapsto \mathbf{1}_{z\geq \phi(0)}$ is upper semi-continuous,

$$\int S_\epsilon(\mathbf{1}_{f_n\leq 0})d\mathbb{P}_1 \leq \int \mathbf{1}_{S_\epsilon(\phi\circ f_n)\geq \phi(0)} d\mathbb{P}_1 = \int \mathbf{1}_{S_\epsilon(\phi\circ f_n)(\mathbf{x})\geq \phi(0)} d\gamma_1^* = \gamma_1^*(S_\epsilon(\phi\circ f_n)(\mathbf{x}) \geq \phi(0)).$$

Now (18) implies that for $n > N$, outside a set of γ_1^*-measure δ, $S_\epsilon(\phi\circ f_n)(\mathbf{x}) < (\phi\circ f_n)(\mathbf{x}') + \phi(0) - \phi(\alpha_r)$ and thus

$$\int S_\epsilon(\mathbf{1}_{f_n\leq 0})d\mathbb{P}_1 \leq \gamma_1^*\big(\phi\circ f_n(\mathbf{x}') + \phi(0) - \phi(\alpha_r) > \phi(0)\big) + \delta \leq \mathbb{P}_1^*\big(\phi\circ f_n > \phi(\alpha_r)\big) + \delta \tag{21}$$

Next, the monotonicity of ϕ implies that $\mathbb{P}_1^*(\phi\circ f_n(\mathbf{x}') > \phi(\alpha_r)) \leq \mathbb{P}_1^*(f_n < \alpha_r)$ and thus (17) implies that

$$\int S_\epsilon(\mathbf{1}_{f_n\leq 0})d\mathbb{P}_1 \leq \mathbb{P}_1^*(f_n < \alpha_r) + \delta \leq \mathbb{P}_1^*(f_n < \alpha_r, |\eta^* - 1/2| \geq r) + 2\delta \tag{22}$$

Next, (20) implies

$$\int S_\epsilon(\mathbf{1}_{f_n\leq 0})d\mathbb{P}_1 \leq \mathbb{P}_1^*(f_n > \alpha_r, \eta^* \geq 1/2 + r) + 3\delta \leq \mathbb{P}_0^*(\eta^* \geq 1/2) + 3\delta \tag{23}$$

Because δ is arbitrary, this last relation implies (14). Observe that $\mathbf{1}_{f\geq 0} = \mathbf{1}_{-f\leq 0}$, and thus the chain of inequalities (21) to (23) hold with $-f_n$ in place f_n, $\mathbb{P}_0$, $\mathbb{P}_0^*$, γ_0^* in place of $\mathbb{P}_1, \mathbb{P}_1^*$, γ_1^*, and (19) in place of (18).

□

6 Consistency Requires Uniqueness up to Degeneracy

We prove the reverse direction of Theorem 4 by constructing a sequence of of functions f_n that minimize R_ϕ^ϵ for which $R^\epsilon(f_n)$ is constant in n and not equal to the minimal adversarial Bayes risk.

Proposition 3. *Assume that $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous with respect to Lebesgue meaure the adversarial Bayes classifier is not unique up to degeneracy for the distribution $\mathbb{P}_0, \mathbb{P}_1$. Then any consistent loss ϕ satisfying $C_\phi^*(1/2) = \phi(0)$ is not adversarially consistent.*

First, Theorem 2 together with a result of [Frank, 2024] imply the adversarial Bayes classifier is unique iff $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are equivalent up to degeneracy, see Appendix F for proof.

Lemma 5. *Assume $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure. Then adversarial Bayes classifier is unique up to degeneracy iff the adversarial Bayes classifiers $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are equivalent up to degeneracy.*

Therefore, if the adversarial Bayes classifier is unique up to degeneracy, then there is a set $\tilde{A}$ that is not an adversarial Bayes classifier but $\{\hat{\eta} > 1/2\} \subset \tilde{A} \subset \{\hat{\eta} \geq 1/2\}$. Theorem 1 suggests that a minimizer of R_ϕ^ϵ can equal zero only when $\hat{\eta} = 1/2$. Thus we select a sequence f_n that is strictly positive on $\tilde{A}$, strictly negative on $\tilde{A}^C$, and approaches 0 on $\{\hat{\eta} = 1/2\}$. Consider the sequence

$$f_n(\mathbf{x}) = \begin{cases} \alpha_\phi(\hat{\eta}(\mathbf{x})) & \hat{\eta}(\mathbf{x}) \neq 1/2 \\ \frac{1}{n} & \hat{\eta}(\mathbf{x}) = 1/2, \mathbf{x} \in \tilde{A} \\ -\frac{1}{n} & \hat{\eta}(\mathbf{x}) = 1/2, \mathbf{x} \notin \tilde{A} \end{cases} \tag{24}$$

Then $R^\epsilon(f_n) = R^\epsilon(\tilde{A})$ for all n and one can show that f_n is a minimizing sequence of R_ϕ^ϵ. However, f_n may assume the values $\pm\infty$ because the function α_ϕ is $\overline{\mathbb{R}}$-valued. A slight modification of these functions produces an $\mathbb{R}$-valued sequence that minimizes R_ϕ^ϵ but $R^\epsilon(f_n) = R^\epsilon(\tilde{A})$ for all n. See Appendix F for a formal proof.

7 Conclusion

In summary, we prove that under a reasonable distributional assumption, a convex loss is adversarially consistent iff the adversarial Bayes classifier is unique up to degeneracy. This result connects an analytical property of the adversarial Bayes classifier to a statistical property of surrogate risks. Hopefully, this connection will aid in the analysis and development of further algorithms for adversarial learning.

Acknowledgments and Disclosure of Funding

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 and NSF grant DMS-2210583.

References

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

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. A finer calibration analysis for adversarial robustness. *arxiv*, 2021b.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds for surrogate loss minimizers. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2022.

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

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.

A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport, 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.

Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim. An analysis of adversarial attacks and defenses on autonomous driving models, 2020.

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

N. S. Frank. A notion of uniqueness for the adversarial bayes classifier. *arxiv*, 2024.

N. S. Frank and J. Niles-Weed. The adversarial consistency of surrogate risks for binary classification. *NeurIps*, 2024a.

N. S. Frank and J. Niles-Weed. Existence and minimax theorems for adversarial surrogate risks in binary classification. *JMLR*, 2024b.

L. Gnecco-Heredia, Y. Chevaleyre, B. Negrevergne, L. Meunier, and M. S. Pydi. On the role of randomization in adversarially robust classification, 2023.

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

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.

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.

M. Paschali, S. Conjeti, F. Navarro, and N. Navab. Generalizability vs. robustness: Adversarial examples for medical imaging. *Springer*, 2018.

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. *ICML*, 2020.

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

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.

N. G. Trillos and R. Murray. Adversarial classification: Necessary conditions and geometric flows. *arxiv*, 2022.

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

N. G. Trillos, M. Jacobs, and J. Kim. On the existence of solutions to adversarial training in multiclass classification, 2023.

Y. Xu, K. Raja, R. Ramachandra, and C. Busch. *Adversarial Attacks on Face Recognition Systems*, pages 139–161. Springer International Publishing, Cham, 2022.

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

A Proof of Lemma 1

Lemma 6. *The smallest minimizer of $C_\phi(\eta, \cdot)$ is well-defined.*

Proof. First, define

$$\alpha_\phi(\eta) = \inf\{\alpha \in \overline{\mathbb{R}} : \alpha \text{ is a minimizer of } C_\phi^*$$

This infimum exists because $\overline{\mathbb{R}}$ is closed. Furthermore, the value $\alpha_\phi(\eta)$ is a minimizer of $C_\phi(\eta, \cdot)$ because the loss ϕ is continuous. ☐

the next result proves that α_ϕ is non-decreasing.

Lemma 7. *If α_2^* is any minimizer of $C_\phi(\eta_2, \cdot)$ and $\eta_2 > \eta_1$, then $\alpha_\phi(\eta_1) \leq \alpha_2^*$.*

Proof. One can express $C_\phi(\eta_2, \alpha)$ as

$$C_\phi(\eta_2, \alpha) = C_\phi(\eta_1, \alpha) + (\eta_2 - \eta_1)(\phi(\alpha) - \phi(-\alpha))$$

Notice that the function $\alpha \mapsto \phi(\alpha) - \phi(-\alpha)$ is non-increasing in α. As $\alpha_\phi(\eta_1)$ is the smallest minimizer of $C_\phi(\eta_1, \alpha)$, if $\alpha < C_\phi(\eta_1, \alpha)$ then $C_\phi(\eta_1, \alpha) > C_\phi^*(\eta_1)$ and thus $C_\phi(\eta_2, \alpha) > C_\phi(\eta_2, \alpha)$. Thus every minimizer of $C_\phi(\eta_2, \cdot)$ must be greater than or equal to $\alpha_\phi(\eta_1)$. ☐

Proof of Lemma 1. Lemma 6 proves that α_ϕ is well-defined while the choice $\alpha_2^* = \alpha_\phi(\eta_2)$ proves that the function α_ϕ is non-decreasing.

☐

B Proof of Lemma 2

Proof of Lemma 2. Let $\mathbb{Q}'$ be a measure in $\mathcal{B}_\epsilon^\infty(\mathbb{Q})$, and let γ^* be a coupling between these two measures supported on Δ_ϵ. Then if $(\mathbf{x}, \mathbf{x}') \in \Delta_\epsilon$, then $\mathbf{x}' \in \overline{B_\epsilon(\mathbf{x})}$ and thus $S_\epsilon(\mathbf{1}_E)(\mathbf{x}) \geq \mathbf{1}_E(\mathbf{x}')$ γ^*-a.e. Consequently,

$$\int S_\epsilon(\mathbf{1}_E)(\mathbf{x}) d\mathbb{Q}_1 = \int S_\epsilon(\mathbf{1}_E)(\mathbf{x}) d\gamma^*(\mathbf{x}, \mathbf{x}') \geq \int \mathbf{1}_E(\mathbf{x}') d\gamma^*(\mathbf{x}, \mathbf{x}') = \int \mathbf{1}_E d\mathbb{Q}'$$

Taking a supremum over all $\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})$ proves the result.

☐

C Proof of Theorem 2

We prove that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ minimize R_ϕ^ϵ by showing that $R_\phi^\epsilon(\{\hat{\eta} > 1/2) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)$ for the measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ in Theorem 7.

Proposition 4. *Let $\hat{\eta}$ be the function in Theorem 1. Then the sets $\{\hat{\eta} > 1/2\}$, $\hat{\eta} \geq 1/2\}$ are both Bayes classifiers.*

Proof. We prove the statement for $\{\hat{\eta} > 1/2\}$, the argument for the set $\{\hat{\eta} \geq 1/2\}$ is analogous.

Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be the measures of Theorem 7 and set $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Furthermore, let γ_0^*, γ_1^* be the couplings between $\mathbb{P}_0, \mathbb{P}_0^*$ and $\mathbb{P}_1, \mathbb{P}_1^*$ supported on Δ_ϵ.

First, Item II) implies that the function $\hat{\eta}(\mathbf{x})$ assumes its infimum on an ball $\overline{B_\epsilon(\mathbf{x})}$ γ_1^*-a.e. and thus and therefore $S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2\}^C})(\mathbf{x}) = \mathbf{1}_{\{I_\epsilon(\hat{\eta}(\mathbf{x}))>1/2\}^C}$ γ_1^*-a.e. (Recall the notation I_ϵ was defined in (11).) Item II) further implies that

$$S_\epsilon(\mathbf{1}_{\{\hat{\eta}(\mathbf{x})>1/2\}^C})(\mathbf{x}) = \mathbf{1}_{\{\hat{\eta}(\mathbf{x}')>1/2\}^C} \quad \gamma_1^*\text{-a.e.} \tag{25}$$

An analogous argument shows

$$S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2\}}) = \mathbf{1}_{\{\hat{\eta}(\mathbf{x}')>1/2} \quad \gamma_0^*\text{-a.e.} \tag{26}$$

Equations (25) and (26) then imply that

$$\begin{aligned}R^\epsilon(\{\hat\eta > 1/2\}) &= \int \mathbf{1}_{\{\hat\eta(\mathbf{x}')>1/2\}^C} d\gamma_1^* + \int \mathbf{1}_{\{\hat\eta(\mathbf{x}')>1/2\}} d\gamma_0^*\\ &= \int \mathbf{1}_{\{\hat\eta(\mathbf{x}')>1/2\}^C} d\mathbb{P}_1^* + \int \mathbf{1}_{\{\hat\eta(\mathbf{x}')>1/2\}} d\mathbb{P}_0^* = \int C(\eta^*, \mathbf{1}_{\{\hat\eta>1/2\}}) d\mathbb{P}^*.\end{aligned}$$

Next Item I) of Theorem 7 implies that $\hat\eta(\mathbf{x}') = \eta^*(\mathbf{x}')$ $\mathbb{P}^*$-a.e. and consequently

$$R^\epsilon(\{\hat\eta > 1/2\}) = \int C(\eta^*, \mathbf{1}_{\{\eta^*>1/2\}}) d\mathbb{P}^* = \bar R(\mathbb{P}_0^*, \mathbb{P}_1^*).$$

Therefore, the strong duality result in Theorem 5 implies that $\{\hat\eta > 1/2\}$ must minimize R^ϵ. □

Finally, the complimentary slackness conditions from [Frank, 2024, Theorem 2.4] characterize minimizers of R^ϵ and maximizers of $\bar R$, and this characterization proves Equations (9) and (10). Verifying these conditions would be another method of proving Proposition 4.

Theorem 9. *The set A is a minimizer of R^ϵ and $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ is a maximizer of $\bar R$ over the W_∞ balls around $\mathbb{P}_0$ and $\mathbb{P}_1$ iff $W_\infty(\mathbb{P}_0^*, \mathbb{P}_0) \le \epsilon$, $W_\infty(\mathbb{P}_1^*, \mathbb{P}_1) \le \epsilon$, and*

1)

$$\int S_\epsilon(\mathbf{1}_{A^C}) d\mathbb{P}_1 = \int \mathbf{1}_{A^C} d\mathbb{P}_1^* \quad \text{and} \quad \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0 = \int \mathbf{1}_A d\mathbb{P}_0^* \tag{27}$$

2)

$$C(\eta^*, \mathbf{1}_A(\mathbf{x}')) = C^*(\eta^*(\mathbf{x}')) \quad \mathbb{P}^*\text{-a.e.} \tag{28}$$

where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$.

Let γ_0^*, γ_1^* be couplings between $\mathbb{P}_0$, $\mathbb{P}_0^*$ and $\mathbb{P}_1$, $\mathbb{P}_1^*$ supported on Δ_ϵ. Notice that because Lemma 2 implies that $\mathbf{1}_A^C(\mathbf{x}') \le S_\epsilon(\mathbf{1}_{A^C})$ γ_1^*-a.e. and $\mathbf{1}_A(\mathbf{x}') \le S_\epsilon(\mathbf{1}_A)(\mathbf{x})$, the complimentary slackness condition in (27) is equivalent to

$$S_\epsilon(\mathbf{1}_{A^C})(\mathbf{x}) = \mathbf{1}_{A^C}(\mathbf{x}') \quad \gamma_1^*\text{-a.e.} \quad \text{and} \quad S_\epsilon(\mathbf{1}_A)(\mathbf{x})\mathbf{1}_A(\mathbf{x}') \quad \gamma_0^*\text{-a.e.} \tag{29}$$

This observation completes the proof of Theorem 2.

Proof of Theorem 2. First, Proposition 4 proves that the sets $\{\hat\eta > 1/2\}$ and $\{\hat\eta \ge 1/2\}$ are in fact adversarial Bayes classifiers.

Next, let $\hat\eta$ be the function in Theorem 7 and let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ be the functions in Theorem 7. Let $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$, $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$, and let γ_1^*, γ_0^* be couplings between $\mathbb{P}_0$, $\mathbb{P}_0^*$ and $\mathbb{P}_1$, $\mathbb{P}_1^*$ supported on Δ_ϵ. If A is any adversarial Bayes classifier, the complimentary slackness condition (28) implies that $\mathbf{1}_{\eta^*>1/2} \le \mathbf{1}_A \le \mathbf{1}_{\eta^*\ge 1/2}$ $\mathbb{P}^*$-a.e. Thus Item I) implies that

$$\mathbf{1}_{\{\hat\eta>1/2\}}(\mathbf{x}') \le \mathbf{1}_A(\mathbf{x}') \le \mathbf{1}_{\{\hat\eta\ge 1/2\}}(\mathbf{x}') \quad \gamma_0^*\text{-a.e.}$$

and

$$\mathbf{1}_{\{\hat\eta>1/2\}^C}(\mathbf{x}') \le \mathbf{1}_{A^C}(\mathbf{x}') \le \mathbf{1}_{\{\hat\eta\ge 1/2\}^C}(\mathbf{x}') \quad \gamma_1^*\text{-a.e.}$$

The complimentary slackness condition (29) then implies Equations (9) and (10). □

D Proof of Theorem 8

Frank [2024, Theorem 3.3] proves the following result:

Theorem 10. *Assume that $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure. Then the following are equivalent:*

1) *The adversarial Bayes classifier is unique up to degeneracy*
2) *Amongst all adversarial Bayes classifiers A, either the value of $\mathbb{P}_0(A^\epsilon)$ is unique or the value of $\mathbb{P}_1((A^C)^\epsilon)$ is unique*

3) There are maximizers $\mathbb{P}_0^*, \mathbb{P}_1^*$ of $\bar{R}$ for which $\mathbb{P}^*(\eta^* = 1/2) = 0$, where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$.

Thus it remains to show that Item 3) of Theorem 10 implies Item C) of Theorem 8. We will apply the complimentary slackness conditions of Theorem 9.

Proof of Theorem 8. Item A) of Theorem 8 is the same as Item 1) of Theorem 10 and Item B) is the same as Item 2). Next, Item C) implies Item 3). Thus it remains to show that Item 3) of Theorem 10 implies Item C) of Theorem 8.

Assume that Theorem 10 holds, and let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ be the measures on Theorem 7. Then Item 2) of Theorem 10 and Theorem 2 imply that

$$S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2\}^C}) = S_\epsilon(\mathbf{1}_{\{\hat{\eta}\geq 1/2\}^C}) \quad \mathbb{P}_1\text{-a.e.} \quad \text{and} \quad S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2\}}) = S_\epsilon(\mathbf{1}_{\{\hat{\eta}\geq 1/2\}}) \quad \mathbb{P}_0\text{-a.e.}$$

The complimentary slackness condition (27) implies that

$$\int \mathbf{1}_{\{\hat{\eta}>1/2\}^C} d\mathbb{P}_1^* = \int \mathbf{1}_{\{\hat{\eta}\geq 1/2\}^C} d\mathbb{P}_1^* \quad \text{and} \quad \int \mathbf{1}_{\{\hat{\eta}>1/2\}} d\mathbb{P}_0^* = \int \mathbf{1}_{\{\hat{\eta}\geq 1/2\}} d\mathbb{P}_0^*$$

and subsequently, Item I) of Theorem 7 implies that

$$\int \mathbf{1}_{\{\eta^*>1/2\}^C} d\mathbb{P}_1^* = \int \mathbf{1}_{\{\eta^*\geq 1/2\}^C} d\mathbb{P}_1^* \quad \text{and} \quad \int \mathbf{1}_{\{\eta^*>1/2\}} d\mathbb{P}_0^* = \int \mathbf{1}_{\{\eta^*\geq 1/2\}} d\mathbb{P}_0^*.$$

Consequently, $\mathbb{P}^*(\eta^* = 1/2) = 0$. □

E Proof of Lemma 4

First, if the loss ϕ is consistent, then 0 can minimize $C_\phi(\eta, \cdot)$ only when $\eta = 1/2$.

Lemma 8. *Let ϕ be a consistent loss. Then if $0 \in \operatorname{argmin} C_\phi(\eta, \cdot)$ then $\eta = 1/2$.*

Proof. Consider a distribution for which $\eta(\mathbf{x}) \equiv \eta$ is constant. Then by the consistency of ϕ, if 0 minimizes $C_\phi(\eta, \cdot)$, then it also must minimize $C(\eta, \cdot)$ and therefore $\eta \leq 1/2$.

However, notice that $C_\phi(\eta, \alpha) = C_\phi(1-\eta, -\alpha)$. Thus if 0 minimizes $C_\phi(\eta, \cdot)$ it must also minimize $C_\phi(1-\eta, \cdot)$. The consistency of ϕ then implies that $1-\eta \leq 1/2$ as well and consequently, $\eta = 1/2$.

□

The proof of Lemma 4 also uses Lemma 7 from Appendix A.

Proof of Lemma 4. Notice that $C_\phi(\eta, \alpha) = C_\phi(1-\eta, -\alpha)$ and thus it suffices to consider $\eta \geq 1/2 + r$.

Lemma 8 implies that $C_\phi(1/2 + r, \alpha_\phi(1/2 + r)) < \phi(0)$. Furthermore, as $\phi(-\alpha) \geq \phi(0) \geq \phi(\alpha)$ when $\alpha \geq 0$, one can conclude that $\phi(\alpha_\phi(1/2 + r)) < \phi(0)$. Now pick an $\alpha_r \in (0, \alpha_\phi(1/2 + r))$ for which $\phi(\alpha_\phi(1/2 + r)) < \phi(\alpha_r) < \phi(0)$. Then by Lemma 7, if $\eta \geq 1/2 + r$, every α less than or equal to α_r does not minimize $C_\phi(\eta, \alpha)$ and thus $C_\phi(\eta, \alpha) - C_\phi^*(\eta) > 0$. Now define

$$k_r = \inf_{\substack{\eta\in[1/2+r,1]\\ \alpha\in[-\infty,\alpha_r]}} C_\phi(\eta, \alpha) - C_\phi^*(\eta)$$

The set $[1/2+r, 1] \times [-\infty, \alpha_r]$ is sequentially compact and the function $(\eta, \alpha) \mapsto C_\phi(\eta, \alpha) - C_\phi^*(\eta)$ is continuous and strictly positive on this set. Therefore, the infimum above is assumed for some η, α and consequently $k_r > 0$. □

F Proof of Proposition 3

First, we show that replacing the value of $\alpha_\phi(1/2)$ with 0 in Theorem 1 results in a minimizer of R_ϕ^ϵ.

Lemma 9. *Let $\alpha_\phi : [0,1] \to \mathbb{R}$ be as in Lemma 1 and define a function $\tilde{\alpha}_\phi : [0,1] \to \overline{\mathbb{R}}$ by*

$$\tilde{\alpha}_\phi(\eta) = \begin{cases} \alpha_\phi(\eta) & \text{if } \eta \neq 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

Let $\hat{\eta} : \mathbb{R}^d \to [0,1]$ be the function described in Theorem 7. If ϕ is consistent and $C_\phi^*(1/2) = \phi(0)$, then $\tilde{\alpha}(\hat{\eta}(\mathbf{x}))$ is a minimizer of R_ϕ^ϵ.

See Appendix F.1 for a proof of this result. Next, we formally prove that if the adversarial Bayes classifier is not unique up to degeneracy, then the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are not equivalent up to degeneracy.

This result in Lemma 5 relies on a characterization of equivalence up to degeneracy from [Frank, 2024] .

Theorem 11. *Assume that $\mathbb{P} \ll \mu$ and let A_1 and A_2 be two adversarial Bayes classifiers. Then the following are equivalent:*

1) The adversarial Bayes classifiers A_1 and A_2 are equivalent up to degeneracy

2) Either $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$-$\mathbb{P}_0$-a.e. or $S_\epsilon(\mathbf{1}_{A_2^C}) = S_\epsilon(\mathbf{1}_{A_1^C})$-$\mathbb{P}_1$-a.e.

Notice that when there is a single equivalence class, the equivalence between Item 1) and Item 2) is simply the equivalence between Item A) and Item B). This result together with Theorem 2 proves Lemma 5:

Proof of Lemma 5. Let A be any adversarial Bayes classifier. If the adversarial Bayes classifiers $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are equivalent up to degeneracy, then Theorem 2 and Item 2) of Theorem 11 imply that $S_\epsilon(\mathbf{1}_A) = S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}})$ $\mathbb{P}_0$-a.e. Item 2) of Theorem 11 again implies that A and $\{\hat{\eta} > 1/2\}$ must be equivalent up to degeneracy. ◻

Thus, if the adversarial Bayes classifier is not unique up to degeneracy, then there is a set $\tilde{A}$ with $\{\hat{\eta} > 1/2\} \subset \tilde{A} \subset \{\hat{\eta} \geq 1/2\}$ that is not an adversarial Bayes classifier, and this set is used to construct the sequence f_n in (24). Next, we show that f_n minimizes R_ϕ^ϵ but not R^ϵ.

Proposition 5. *Assume that $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure and that the adversarial Bayes classifier is not unique up to degeneracy. Then there is an $\overline{\mathbb{R}}$-valued sequence of functions that minimize R_ϕ^ϵ but $R^\epsilon(f_n)$ is constant in n and not equal to the adversarial Bayes risk.*

Proof. By Lemma 5, there is a set $\tilde{A}$ with $\{\hat{\eta} > 1/2\} \subset \tilde{A} \subset \{\hat{\eta} \geq 1/2\}$ which is not an adversarial Bayes classifier. For this set $\tilde{A}$, define the sequence f_n by (24) and let $\tilde{\alpha}_\phi$ be the function in Lemma 9. Lemma 4 implies that $\tilde{\alpha}(\eta) \neq 0$ whenever $\eta \neq 1/2$ and thus $\{f_n > 0\} = \tilde{A}$ for all n. We will show that in the limit $n \to \infty$, the function sequence $S_\epsilon(\phi \circ f_n)$ is bounded above by $S_\epsilon(\phi \circ \tilde{\alpha}_\phi(\hat{\eta}))$ while $S_\epsilon(\phi \circ -f_n)$ is bounded above by $S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta}))$. This result will imply that f_n is a minimizing sequence of R_ϕ^ϵ.

Let $\tilde{S}_\epsilon(g)$ denote the supremum of a function g on an ϵ-ball excluding the set $\hat{\eta}(\mathbf{x}) = 1/2$:

$$\tilde{S}_\epsilon(g) = \begin{cases} \sup_{\substack{\mathbf{x}' \in \overline{B_\epsilon(\mathbf{0})} \\ \hat{\eta}(\mathbf{x}') \neq 1/2}} g(\mathbf{x}') & \text{if } \overline{B_\epsilon(\mathbf{x})} \cap \{\hat{\eta} \neq 1/2\}^C \neq \emptyset \\ -\infty & \text{otherwise} \end{cases}$$

With this notation, because $\tilde{\alpha}_\phi(1/2) = 0$, one can express $S_\epsilon(\phi \circ \tilde{\alpha}_\phi(\hat{\eta}))$, $S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta}))$ as

$$S_\epsilon(\phi \circ \tilde{\alpha}_\phi(\hat{\eta})) = \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})), \phi(0)) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \tag{31}$$

$$S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta})) = \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})), \phi(0)) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \tag{32}$$

and similarly

$$S_\epsilon(\phi\circ f_n)\leq \begin{cases}\max(\tilde S_\epsilon(\phi\circ\alpha_\phi(\hat\eta)),\phi(-\frac{1}{n})) & \mathbf{x}\in\{\hat\eta=1/2\}^\epsilon\\ S_\epsilon(\phi\circ\alpha_\phi(\hat\eta)) & \mathbf{x}\not\in\{\hat\eta=1/2\}^\epsilon\end{cases} \tag{33}$$

$$S_\epsilon(\phi\circ -f_n)\leq \begin{cases}\max(\tilde S_\epsilon(\phi\circ-\alpha_\phi(\hat\eta)),\phi(-\frac{1}{n})) & \mathbf{x}\in\{\hat\eta=1/2\}^\epsilon\\ S_\epsilon(\phi\circ-\alpha_\phi(\hat\eta)) & \mathbf{x}\not\in\{\hat\eta=1/2\}^\epsilon\end{cases} \tag{34}$$

Therefore, by comparing (33) with (31) and (34) with (32), one can conclude that

$$\limsup_{n\to\infty} S_\epsilon(\phi\circ f_n)\leq S_\epsilon(\phi\circ\tilde\alpha_\phi(\hat\eta))\quad\text{and}\quad \limsup_{n\to\infty} S_\epsilon(\phi\circ -f_n)\leq S_\epsilon(\phi\circ-\tilde\alpha_\phi(\hat\eta)). \tag{35}$$

Furthermore, (33) implies that $S_\epsilon(\phi\circ f_n)\leq S_\epsilon(\phi\circ\alpha_\phi(\hat\eta))+\phi(-1)$ and (34) implies that $S_\epsilon(\phi\circ -f_n)\leq S_\epsilon(\phi\circ-\alpha_\phi(\hat\eta))+\phi(-1)$. Thus the dominated convergence theorem and (35) implies that

$$\limsup_{n\to\infty} R_\phi^\epsilon(f_n)\leq R_\phi^\epsilon(\tilde\alpha_\phi(\hat\eta))$$

and thus f_n minimizes R_ϕ^ϵ.

□

Lastly, it remains to construct an $\mathbb{R}$-valued sequence that minimizes R_ϕ^ϵ but not R^ϵ. To construct this sequence, we threshhold a subsequence f_{n_j} of f_n at appropriate T_j. If g is an $\overline{\mathbb{R}}$-valued function and g_N is the function g threshholded at N, then $\lim_{N\to\infty} R_\phi^\epsilon(g^N)=R_\phi^\epsilon(g)$.

Lemma 10. *Let g be an $\overline{\mathbb{R}}$-valued function and let $g^{(N)}=\min(\max(g,-N),N)$. Then $\lim_{N\to\infty}R_\phi^\epsilon(g^{(N)})=R_\phi^\epsilon(g)$.*

See Appendix F.2 for a proof. Proposition 3 then follows from this lemma and Proposition 5:

Proof of Proposition 3. Let f_n be the $\overline{\mathbb{R}}$-valued sequence of functions in Proposition 5, and let f_{n_j} be a subsequence for which $R_\phi^\epsilon(f_{n_j})<1/j$. Next, Lemma 10 implies that for each j one can pick a threshhold N_j for which $|R_\phi^\epsilon(f_{n_j})-R_\phi^\epsilon(f_{n_j}^{(N_j)})|\leq 1/j$. Consequently, $f_{n_j}^{(N_j)}$ is an $\mathbb{R}$-valued sequence of functions that minimizes R_ϕ^ϵ. However, notice that $\{f\leq 0\}=\{f^{(T)}\leq 0\}$ and $\{f>0\}=\{f^{(T)}>0\}$ for any strictly positive threshhold T. Thus $R^\epsilon(f_{n_j}^{(N_j)})=R^\epsilon(f_{n_j})$ and consequently $f_{n_j}^{(N_j)}$ does not minimize R^ϵ. □

F.1 Proof of Lemma 9

The proof of Lemma 9 follows the same outline as the argument for Proposition 4: we show that $R_\phi^\epsilon(\tilde\alpha_\phi(\hat\eta))=\bar R_\phi(\mathbb{P}_0^*,\mathbb{P}_1^*)$ for the measures $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ in Theorem 7, and then Theorem 6 implies that $\tilde\alpha_\phi(\hat\eta)$ must minimize R_ϕ^ϵ. Similar to the proof of Proposition 4, swapping the order of the S_ϵ operation and $\tilde\alpha_\phi$ is a key step. To show that this swap is possible, we first prove that $\tilde\alpha_\phi$ is monotonic.

Lemma 11. *If $C_\phi^*(1/2)=\phi(0)$, then the function $\tilde\alpha_\phi:[0,1]\to\overline{\mathbb{R}}$ defined in (30) is non-decreasing.*

Proof. Lemma 1 implies that $\tilde\alpha_\phi(\eta)$ is a minimizer of $C_\phi(\eta,\cdot)$ for all $\eta\neq 1/2$ and the assumption $C_\phi^*(1/2)=\phi(0)$ implies that $\tilde\alpha_\phi(1/2)$ is a minimizer of $C_\phi(1/2,\cdot)$. Furthermore, Lemma 1 implies that $\tilde\alpha_\phi$ is non-decreasing on $[0,1/2)$ and $(1/2,1]$. However, Lemma 4 implies that $\alpha_\phi(\eta)<0$ when $\eta\in[0,1/2)$ and $\alpha_\phi(\eta)>0$ when $\eta\in(1/2,1]$. □

This result together with the properties of $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ suffice to prove Lemma 9.

Proof of Lemma 9. Let $\mathbb{P}_0^*,\mathbb{P}_1^*$ be the measures of Theorem 7 and set $\mathbb{P}^*=\mathbb{P}_0^*+\mathbb{P}_1^*$, $\eta^*=d\mathbb{P}_1^*/d\mathbb{P}^*$. We will prove that $R_\phi^\epsilon(\tilde\alpha_\phi(\hat\eta))=\bar R_\phi(\mathbb{P}_0^*,\mathbb{P}_1^*)$ and thus Theorem 6 will imply that $\tilde\alpha_\phi(\hat\eta)$ minimizes

R_ϕ^ϵ. Let γ_0^* and γ_1^* be the couplings supported on Δ_ϵ between $\mathbb{P}_0$, $\mathbb{P}_0^*$ and $\mathbb{P}_1$, $\mathbb{P}_1^*$ respectively. Item II) and Lemma 11 imply that

$$S_\epsilon(\phi(\tilde{\alpha}_\phi(\hat{\eta})))(\mathbf{x}) = \phi(\tilde{\alpha}_\phi(I_\epsilon(\hat{\eta}(\mathbf{x})))) = \phi(\tilde{\alpha}_\phi(\hat{\eta}(\mathbf{x}'))) \quad \gamma_1^* - a.e.$$

and

$$S_\epsilon(\phi(-\tilde{\alpha}_\phi(\hat{\eta})))(\mathbf{x}) = \phi(-\tilde{\alpha}_\phi(S_\epsilon(\hat{\eta}(\mathbf{x})))) = \phi(\tilde{\alpha}_\phi(-\hat{\eta}(\mathbf{x}'))) \quad \gamma_0^* - a.e.$$

(Recall the the notation I_ϵ was introduced in (11).) Therefore,

$$\begin{aligned} R_\phi^\epsilon(\tilde{\alpha}_\phi(\hat{\eta})) &= \int \phi(\tilde{\alpha}_\phi(\hat{\eta}(\mathbf{x}'))d\gamma_1^* + \int \phi(-\tilde{\alpha}_\phi(\hat{\eta}(\mathbf{x}')))d\gamma_0^* \\ &= \int \phi(\tilde{\alpha}_\phi(\hat{\eta}(\mathbf{x}')))d\mathbb{P}_1^* + \int \phi(-\tilde{\alpha}_\phi(\hat{\eta}(\mathbf{x}')))d\mathbb{P}_0^* = \int C_\phi(\eta^*, \tilde{\alpha}_\phi(\hat{\eta}))d\mathbb{P}^* \end{aligned}$$

Next, Item I) of Theorem 7 implies that $\hat{\eta}(\mathbf{x}') = \eta^*(\mathbf{x}')$ and consequently

$$R_\phi^\epsilon(\tilde{\alpha}_\phi(\hat{\eta})) = \int C_\phi(\eta^*, \tilde{\alpha}_\phi(\hat{\eta}))d\mathbb{P}^* = \int C_\phi(\eta^*, \tilde{\alpha}_\phi(\eta^*))d\mathbb{P}^*$$

Therefore, the strong duality result in Theorem 6 implies that $\tilde{\alpha}_\phi(\hat{\eta})$ must minimize R_ϕ^ϵ.

□

F.2 Proof of Lemma 10

This argument is taken from the proof of Lemma 8 in Frank and Niles-Weed [2024a].

Proof of Lemma 10. Define

$$\sigma_{[a,b]}(\alpha) = \begin{cases} a & \text{if } \alpha < a \\ \alpha & \text{if } \alpha \in [a,b] \\ b & \text{if } \alpha > b \end{cases}$$

Notice that

$$S_\epsilon(\sigma_{[a,b]}(h)) = \sigma_{[a,b]}(S_\epsilon(h))$$

and

$$\phi(\sigma_{[a,b]}(g)) = \sigma_{[\phi(b),\phi(a)]}(\phi(g))$$

for any functions g and h. Therefore,

$$S_\epsilon(\phi(g^{(N)})) = \sigma_{[\phi(N),\phi(-N)]}(S_\epsilon(\phi \circ g)) \quad \text{and} \quad S_\epsilon(\phi \circ -g^{(N)}) = \sigma_{[\phi(N),\phi(-N)]}(S_\epsilon(\phi \circ -g)),$$

which converge to $S_\epsilon(\phi \circ g)$ and $S_\epsilon(\phi \circ -g)$ pointwise and $N \to \infty$. Furthermore, the functions $S_\epsilon(\phi \circ g^{(N)})$ and $S_\epsilon(\phi \circ -g^{(N)})$ are bounded above by

$$S_\epsilon(\phi \circ g^{(N)}) \leq S_\epsilon(\phi \circ g) + \phi(1) \quad \text{and} \quad S_\epsilon(\phi \circ -g^{(N)}) \leq S_\epsilon(\phi \circ -g) + \phi(1)$$

for $N \geq 1$. As the functions $S_\epsilon(\phi \circ g) + \phi(1)$ and $S_\epsilon(\phi \circ -g) + \phi(1)$ are integrable with respect to $\mathbb{P}_1$ and $\mathbb{P}_0$ respectively, the dominated convergence theorem implies that

$$\lim_{n \to \infty} R_\phi^\epsilon(g^{(N)}) = R_\phi^\epsilon(g).$$

□