

PAPER

Adversarial consistency and the uniqueness of the adversarial bayes classifier

Natalie S. Frank 

Mathematics, Courant Institute, New York, NY, USA
Email: nf1066@nyu.edu

Received: 31 May 2024; **Revised:** 20 October 2024; **Accepted:** 26 January 2025

Keywords: Statistics theory; optimization

2020 Mathematics Subject Classification: 62A99 (Primary); 65K99 (Secondary)

Abstract

Minimizing an adversarial surrogate risk 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 surrogate 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 [36], medical imaging [25] and identifying traffic signs in self-driving cars [13]. Deep learning models exhibit a concerning security risk – small perturbations imperceptible to the human eye can cause a neural net to misclassify an image [9, 32]. 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 that involves minimizing an adversarial surrogate risk.

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 prediction. The proportion of mislabelled 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 nontrivial task [8, 27]. Further studies [15, 18, 28, 33, 35] investigate additional properties of these minimizers, and Frank [15] describes a notion of uniqueness for adversarial Bayes classifiers. The main result in this paper will connect this notion of uniqueness the statistical consistency of a popular defense method.

The empirical adversarial classification error is a discrete notion 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 also minimizes the adversarial classification risk. A loss is simply called *adversarially consistent* if it is adversarially consistent for all possible data distributions. Surprisingly, Meunier et al. [23] show that no convex surrogate is adversarially consistent, in contrast to the standard classification setting where most convex losses are statistically consistent [6, 22, 24, 31, 37].

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.

Prior work [15] further demonstrates several distributions for which the adversarial Bayes classifier is unique, and thus a typical convex loss would be consistent. Understanding general conditions under which uniqueness occurs is an open question.

Paper outline. Section 2 discusses related works, and Section 3 presents the problem background. Section 4 states our main theorem and presents some examples. Subsequently, Section 5 discusses intermediate results necessary for proving our main theorem. Next, our consistency results are proved in Sections 6 and 7. Appendices A and B present deferred proofs from Section 5, while Appendix C presents deferred proofs on surrogate risks from Sections 5 and 6. Finally, Appendix D presents deferred proofs from Section 7.

2. Related works

Our results are inspired by prior work that showed that no convex loss is adversarially consistent [3, 23] yet a wide class of adversarial losses is adversarially consistent [16]. These consistency results rely on the theory of surrogate losses, studied by Bartlett et al. [6], Lin [22] in the standard classification setting; and by Frank and Niles-Weed [17], Li and Telgarsky [21] in the adversarial setting. Furthermore, [2, 5, 31] study a property of related to consistency called *calibration*, which [23] relate to consistency. Complementing this analysis, another line of research studies $\mathcal{H}$ -consistency, which refines the concept of consistency to specific function classes [3, 26]. Our proof combines results on losses with minimax theorems for various adversarial risks, as studied by [16, 17, 28, 34].

Furthermore, our result leverages recent results on the adversarial Bayes classifier, which are extensively studied by [4, 8, 10, 16, 27, 28]. Specifically, [4, 8, 10] prove the existence of the adversarial Bayes classifier, while Trillos and Murray [33] derive necessary conditions that describe the boundary of the adversarial Bayes classifier. Frank [15] defines the notion of uniqueness up to degeneracy and proves that in one dimension, under reasonable distributional assumptions, every adversarial Bayes classifier is equivalent up to degeneracy to an adversarial Bayes classifier which satisfies the necessary conditions of [33]. Finally, [8, 27] also calculate the adversarial Bayes classifier for distributions in dimensions higher than one by finding an optimal coupling, but whether this method can calculate the equivalence classes under equivalence up to degeneracy remains an open question.

3. Notation and 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 measure $\mathbb{P}_0$ while class $+1$ is distributed according to measure $\mathbb{P}_1$. A *classifier* is a Borel set A and 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 :

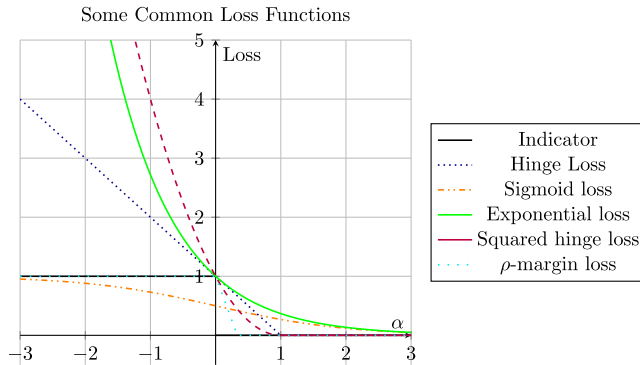


Figure 1. Several common loss functions for classification along with the indicator $\mathbf{1}_{\alpha \leq 0}$.

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

A minimizer to R is called a *Bayes classifier*.

However, minimizing the empirical classification risk is a computationally intractable problem [7]. A common approach is to instead learn a function f and then threshold at zero to obtain the classifier $A = \{\mathbf{x} : f(\mathbf{x}) > 0\}$. We define the classification risk of a function f by

$$R(f) = R(\{f > 0\}) = \int \mathbf{1}_{f \leq 0} d\mathbb{P}_1 + \int \mathbf{1}_{-f < 0} d\mathbb{P}_0 \quad (1)$$

In order to learn f , machine learning algorithms typically minimize a better-behaved alternative to the classification risk called a *surrogate risk*. To obtain this risk, we replace the indicator functions in (1) with the loss ϕ , resulting in:

$$R_\phi(f) = \int \phi(f) d\mathbb{P}_1 + \int \phi(-f) d\mathbb{P}_0. \quad (2)$$

We restrict to losses with similar properties to the indicator functions in (1) yet are easier to optimize. In particular we require:

Assumption 1. The loss ϕ is nonincreasing, continuous, and $\lim_{\alpha \rightarrow \infty} \phi(\alpha) = 0$.

See Figure 1 for a comparison of the indicator function and a several common losses. Losses on $\mathbb{R}$ -valued functions in machine learning typically satisfy Assumption 1.

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 $g : \mathbb{R}^d \rightarrow \mathbb{R}$, define

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

where $\overline{B_\epsilon(\mathbf{x})} = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\| \leq \epsilon\}$ 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 (1), we define $R^\epsilon(f) = R^\epsilon(\{f > 0\})$:

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

Again, minimizing an empirical adversarial classification risk is computationally intractable. A surrogate to the adversarial classification risk is formulated as²

$$R_\phi^\epsilon(f) = \int S_\epsilon(\phi \circ f) d\mathbb{P}_1 + \int S_\epsilon(\phi \circ -f) d\mathbb{P}_0. \quad (3)$$

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_ϕ^ϵ 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_ϕ^ϵ 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 ubiquitous in machine learning. In the non-adversarial context, Theorem 2 of [6] shows that a convex loss ϕ is consistent iff ϕ is differentiable at zero and $\phi'(0) < 0$. In contrast, Meunier et al. [23] show that no convex loss is adversarially consistent. Further results of [16] characterize the adversarially consistent losses in terms of the function C_ϕ^* :

Theorem 1. *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,

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

due to convexity. Notice that Theorem 1 does not preclude the adversarial consistency of a loss satisfying $C_\phi^*(1/2) = \phi(0)$ for some particular $\mathbb{P}_0, \mathbb{P}_1$. Prior work [16, 23] 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 3 below). Furthermore, Frank [15] developed a notion of uniqueness for adversarial Bayes classifiers.

¹The functions $S_\epsilon(\mathbf{1}_A)$, $S_\epsilon(\mathbf{1}_{A^c})$ must be measurable in order to define this integral. See [17, Section 3.3] for a treatment of this matter.

²Again, see [17, Section 3.3] for a treatment of measurability.

³This concept is referred to as *calibration* in the non-adversarial machine learning context [6, 31]. We use the term ‘consistent’, as prior work on adversarial learning [1, 23] use ‘calibration’ to refer to a different but related concept.

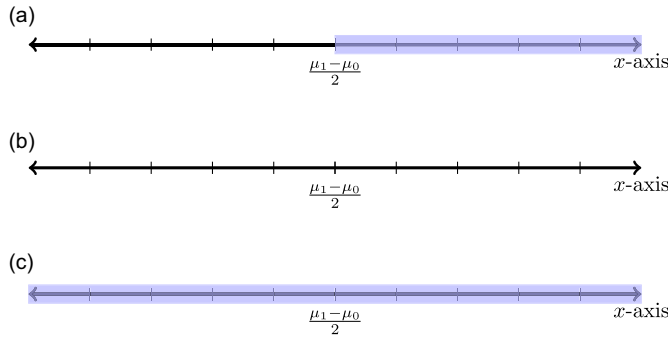


Figure 2. The adversarial Bayes classifier for two Gaussians with equal variances and differing means. We assume in this figure that $\mu_1 > \mu_0$. The shaded blue area depicts the region inside the adversarial Bayes classifier. Figure 2a depicts an adversarial Bayes when $\epsilon \leq (\mu_1 - \mu_0)/2$ and Figure 2b and 2c depict the adversarial Bayes classifier when $\epsilon \geq (\mu_1 - \mu_0)/2$. (See [15, Example 4.1] for a justification of these illustrations.) the adversarial Bayes classifiers in Figure 2b and 2c are not equivalent up to degeneracy.

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 equivalent up to degeneracy.

When the measure

$$\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$$

is absolutely continuous with respect to Lebesgue measure, then equivalence up to degeneracy is an equivalence relation [15, Theorem 3.3]. The central result of this paper relates the consistency of convex losses to the uniqueness of the adversarial Bayes classifier.

Theorem 2. 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.

Prior results of Frank [15] provide 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. In the examples below, the function p_1 will represent the density of $\mathbb{P}_1$ and the function p_0 will represent the density of $\mathbb{P}_0$.

- Consider mean zero Gaussians with different variances: $p_0(x) = \frac{1}{2\sqrt{2\pi}\sigma_0} e^{-x^2/2\sigma_0^2}$ and $p_1(x) = \frac{1}{2\sqrt{2\pi}\sigma_1} e^{-x^2/2\sigma_1^2}$. The adversarial Bayes classifier is unique up to degeneracy for all ϵ [15, Example 4.2].
- 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 $\epsilon < |\mu_1 - \mu_0|/2$ [15, Example 4.1]. See Figure 2 for an illustration of the adversarial Bayes classifiers for this distribution.

Theorem 2 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 [15, Example 4.5] provides an example of a distribution for which the adversarial Bayes classifier is not unique up to degeneracy for all $\epsilon > 0$, even though the Bayes classifier is unique. At the same time, one would hope that if the Bayes classifier is unique and $\mathbb{P}_0, \mathbb{P}_1$ are sufficiently regular, then the adversarial

Bayes classifier would be unique up to degeneracy for sufficiently small ϵ . In general, understanding when the adversarial Bayes classifier is unique up to degeneracy for well-behaved distributions is an open problem.

The examples above rely on the techniques of [15] for calculating the equivalence classes under uniqueness up to degeneracy. Frank [15] proves that in one dimension, if $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, every adversarial Bayes classifier is equivalent up to degeneracy to an adversarial Bayes classifier whose boundary points are strictly more than 2ϵ apart [15, Theorem 3.5]. Therefore, to find all adversarial Bayes classifiers under equivalence under degeneracy, it suffices to consider all sets whose boundary points satisfy the first order necessary conditions obtained by differentiating the adversarial classification risk of a set A with respect to its boundary points [15, Theorem 3.7]. The corresponding statement in higher dimensions is false – there exist distributions for which no adversarial Bayes classifier has enough regularity to allow for an equivalent statement. For instance, Bungert et al. [10] demonstrate a distribution for which there is no adversarial Bayes classifier with two-sided tangent balls at all points in the boundary. Developing a general method for calculating these equivalence classes in dimensions higher than one remains an open problem.

Proposition 2 in Section 6 presents a condition under which one can conclude consistency without the absolute continuity assumption. This result proves consistency whenever the optimal adversarial classification risk is zero, see the discussion after Proposition 2 for details. Consequently, if $\text{supp}\mathbb{P}_0$ and $\text{supp}\mathbb{P}_1$ are separated by more than 2ϵ , then consistent losses are always adversarially consistent for such distributions. On the other hand, our analysis of the reverse direction of Theorem 2 requires the absolute continuity assumption. Using Proposition 2 to further understand consistency is an open question.

5. Preliminary results

5.1 Minimizers of standard risks

Minimizers to the classification risk can be expressed in terms of the measure $\mathbb{P}$ 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 $C : [0, 1] \times \{0, 1\} \rightarrow \mathbb{R}$ and $C^* : [0, 1] \rightarrow \mathbb{R}$ are defined by

$$C(\eta, b) = \eta b + (1 - \eta)(1 - b), \quad C^*(\eta) = \inf_{b \in \{0, 1\}} C(\eta, b) = \min(\eta, 1 - \eta). \quad (4)$$

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

$$\{\mathbf{x} : \eta(\mathbf{x}) > 1/2\} \quad \text{and} \quad \{\mathbf{x} : \eta(\mathbf{x}) \geq 1/2\} \quad (5)$$

are both Bayes classifiers.

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

$$R_\phi(f) = \int C_\phi(\eta(\mathbf{x}), f(\mathbf{x})) d\mathbb{P} \quad (6)$$

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

$$C_\phi(\eta, \alpha) = \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha), \quad C_\phi^*(\eta) = \inf_{\alpha} C_\phi(\eta, \alpha) \quad (7)$$

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 $\mathbb{P}$. Because ϕ is continuous, the function

$$\alpha_\phi(\eta) = \inf\{\alpha \in \overline{\mathbb{R}} : \alpha \text{ is a minimizer of } C_\phi(\eta, \cdot)\} \quad (8)$$

maps each η to the smallest minimizer of $C_\phi(\eta, \cdot)$. Consequently, the function

$$\alpha_\phi(\eta(\mathbf{x})) \quad (9)$$

minimizes $C_\phi(\eta(\mathbf{x}), \cdot)$ at each point $\mathbf{x}$. Next, we will argue this function is measurable, and therefore is a minimizer of the risk R_ϕ .

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

The proof of this result is presented below Lemma 7 in Appendix C. Because α_ϕ is monotonic, the composition in (9) is always measurable, and thus this function is a minimizer of R_ϕ . Allowing for minimizers in extended real numbers $\mathbb{R} = \{-\infty, +\infty\} \cup \mathbb{R}$ 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 $\mathbb{R}$.

5.2 Dual problems for the adversarial risks

The proof Theorem 2 relies on a dual formulation of the adversarial classification problem involving the Wasserstein- ∞ metric. 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- ∞ metric involves 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 the two measures is then

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

Theorem 2.6 of [19] proves that this infimum is always assumed. Equivalently, $W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon$ iff there is a coupling between $\mathbb{Q}$ and $\mathbb{Q}'$ supported on

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

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*

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

Proof. 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. □

Lemma 2 implies:

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

Does equality hold and can one swap the infimum and the supremum? Frank and Niles-Weed [16], Pydi and Jog [28] answer this question in the affirmative:

Theorem 3. *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 (4). 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 [16] for a proof. Theorems 6, 8, and 9 of [17] show an analogous minimax theorem for surrogate risks.

Theorem 4. *Let $\mathbb{P}_0, \mathbb{P}_1$ be finite Borel measures. Define*

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

with the function C_ϕ^* is defined in (7). Then

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

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 [16] states that

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

5.3 Minimizers of adversarial risks

A formula analogous to (9) defines minimizers to adversarial risks. Let I_ϵ denote the infimum of a function over an ϵ ball:

$$I_\epsilon(g) = \inf_{\mathbf{x}' \in \bar{B}_\epsilon(\mathbf{x})} g(\mathbf{x}') \quad (10)$$

Lemma 24 of [17] and Theorem 9 of [17] prove the following result:

Theorem 5. *There exists a function $\hat{\eta} : \mathbb{R}^d \rightarrow [0, 1]$ 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) $I_\epsilon(\hat{\eta})(\mathbf{x}) = \hat{\eta}(\mathbf{x}') \gamma_0^*$ -a.e. and $S_\epsilon(\hat{\eta})(\mathbf{x}) = \hat{\eta}(\mathbf{x}') \gamma_1^*$ -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 .
- III) The function $\alpha_\phi(\hat{\eta}(\mathbf{x}))$ is a minimizer of R_ϕ^ϵ for any loss ϕ , where α_ϕ is the function defined in (8).

The function $\hat{\eta}$ can be viewed as the conditional probability of label +1 under an ‘optimal’ adversarial attack [17]. Just as in the standard learning scenario, the function $\alpha(\hat{\eta}(\mathbf{x}))$ may be $\mathbb{R}$ -valued. Item 3 is actually a consequence of Item 1 and Item 2: Item 1 and Item 2 imply that $R_\phi^\epsilon(\alpha_\phi(\hat{\eta})) = \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*)$ and Theorem 4 then implies that $\alpha_\phi(\hat{\eta})$ minimizes R_ϕ^ϵ and $\mathbb{P}_0^*, \mathbb{P}_1^*$ maximize $\bar{R}_\phi$. (A similar argument is provided later in this paper in Lemma 6 of Appendix D.1.) Furthermore, the relation $R_\phi^\epsilon(\alpha_\phi(\hat{\eta})) = \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*)$ also implies

Lemma 3. The $\mathbb{P}_0^*, \mathbb{P}_1^*$ of Theorem 5 maximize $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ for every ϕ .

We emphasize that a formal proof of Theorems 5 and 3 is not included in this paper, and refer to Lemma 26 and Theorem 9 of [17] for full arguments.

Next, we derive some further results about the function $\hat{\eta}$. Recall that Bayes classifiers can be constructed by thresholding the conditional probability η at $1/2$, see Equation 5). The function $\hat{\eta}$ plays an analogous role for adversarial learning.

Theorem 6. Let $\hat{\eta}$ be the function described by Theorem 5. Then the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are adversarial Bayes classifiers. Furthermore, any adversarial Bayes classifier A satisfies

$$\int S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}^c}) d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}^c}) d\mathbb{P}_1 \quad (11)$$

and

$$\int S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}}) d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}}) d\mathbb{P}_0 \quad (12)$$

See Appendix A for a formal proof; these properties follow direction from Item 1 and Item 2 of Theorem 5. Equations (11) and (12) imply that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ can be viewed as ‘minimal’ and ‘maximal’ adversarial Bayes classifiers.

Theorem 6 is proved in Appendix A— Item 1 and Item 2 imply that $R^\epsilon(\{\hat{\eta} > 1/2\}) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*) = R^\epsilon(\{\hat{\eta} \geq 1/2\})$ and consequently Theorem 3 implies that $\{\hat{\eta} > 1/2\}$, $\{\hat{\eta} \geq 1/2\}$ minimize R^ϵ and $\mathbb{P}_0^*, \mathbb{P}_1^*$ maximize $\bar{R}$. This proof technique is analogous to the approach employed by [17] to establish Theorem 5. Lastly, uniqueness up to degeneracy can be characterized in terms of these $\mathbb{P}_0^*, \mathbb{P}_1^*$.

Theorem 7. 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) $\mathbb{P}^*(\eta^* = 1/2) = 0$, 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 5.

See Appendix B for a proof of Theorem 7. In relation to prior work – the proof of [15, Theorem 3.4] shows Theorem 7 but a full proof of Theorem 7 is included in this paper for clarity as [15] did not discuss the role of the function $\hat{\eta}$.

6. Uniqueness up to degeneracy implies consistency

Before presenting the full proof of consistency, we provide an overview of the strategy of this argument. First, a minimizing sequence of R_ϕ^ϵ must satisfy the approximate complementary slackness conditions derived in [16, Proposition 4].

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

$$\lim_{n \rightarrow \infty} \int C_\phi(\eta^*, f_n) d\mathbb{P}^* = \int C_\phi^*(\eta^*) d\mathbb{P}^* \quad (13)$$

$$\lim_{n \rightarrow \infty} \int S_\epsilon(\phi \circ f_n) d\mathbb{P}_1 - \int \phi \circ f_n d\mathbb{P}_1^* = 0, \quad \lim_{n \rightarrow \infty} \int S_\epsilon(\phi \circ -f_n) d\mathbb{P}_0 - \int \phi \circ -f_n d\mathbb{P}_0^* = 0, \quad (14)$$

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 (13) and (14) must minimize R^ϵ . Specifically, we will prove that every minimizing sequence f_n of R_ϕ^ϵ must satisfy

$$\limsup_{n \rightarrow \infty} \int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \int \mathbf{1}_{\eta^* \leq \frac{1}{2}} d\mathbb{P}_1^*, \quad (15)$$

$$\limsup_{n \rightarrow \infty} \int S_\epsilon(\mathbf{1}_{f_n \geq 0}) d\mathbb{P}_0 \leq \int \mathbf{1}_{\eta^* \geq \frac{1}{2}} d\mathbb{P}_0^* \quad (16)$$

for the measures $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ in Theorem 5. Consequently, $\mathbb{P}^*(\eta^* = 1/2) = 0$ would imply that $\limsup_{n \rightarrow \infty} R^\epsilon(f_n) \leq \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)$ and the strong duality relation in Theorem 3 implies that f_n must in fact be a minimizing sequence of R^ϵ .

Next, we summarize the argument establishing Equation 15. We make several simplifying assumptions in the following discussion. First, 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 (13) 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 α 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 states that $S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi \circ f_n(\mathbf{x}') \gamma_1^*$ -a.e., Equation 14 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 $S_\epsilon(\phi \circ f_n)(\mathbf{x})$ must converge to $\phi(\alpha_\phi(\eta^*(\mathbf{x}')))$ in $L^1(\gamma_1^*)$. As L^1 convergence implies convergence in measure [14, Proposition 2.29], one can conclude that

$$\lim_{n \rightarrow \infty} \gamma_1^*(S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi \circ (\alpha_\phi(\hat{\eta}(\mathbf{x}')))) > c) = 0 \quad (17)$$

for any $c > 0$. The lower semi-continuity of $\alpha \mapsto \mathbf{1}_{\alpha \leq 0}$ implies that $\int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \int \mathbf{1}_{S_\epsilon(\phi(f_n))(\mathbf{x}) \geq \phi(0)} d\mathbb{P}_1$ and furthermore (17) implies

$$\limsup_{n \rightarrow \infty} \int \mathbf{1}_{S_\epsilon(\phi(f_n))(\mathbf{x}) \geq \phi(0)} d\gamma_1^* \leq \int \mathbf{1}_{\phi(\alpha_\phi(\eta^*(\mathbf{x}')))) < \phi(0) - c} d\gamma_1^* = \int \mathbf{1}_{\eta^* \geq \alpha_\phi^{-1} \circ \phi^{-1}(\phi(0) - c)} d\mathbb{P}_1^*. \quad (18)$$

Next, we will also assume that α_ϕ^{-1} is continuous and $\alpha_\phi(1/2) = 0$. (The exponential loss satisfies these requirements as well.) Due to our assumptions on ϕ and α_ϕ , the quantity $\phi^{-1}(\phi(0) - c)$ is strictly smaller than 0, and consequently, $\alpha_\phi^{-1} \circ \phi^{-1}(\phi(0) - c)$ is strictly smaller than $1/2$. However, if α_ϕ^{-1} is continuous, one can choose c small enough so that $\mathbb{P}^*(|\eta - 1/2| < 1/2 - \alpha_\phi^{-1} \circ \phi^{-1}(\phi(0) - c)) < \delta$ for any $\delta > 0$ when $\mathbb{P}^*(\eta^* = 1/2) = 0$. This choice of c along with (18) proves (15).

To avoid the prior assumptions on ϕ and α , 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 $\text{sign}(\eta - 1/2)\alpha \leq \alpha_r$ then $C_\phi(\eta, \alpha) - C_\phi^*(\eta) \geq k_r$, and this α_r satisfies $\phi(\alpha_r) < \phi(0)$.*

See Appendix C for a proof. A minor modification of the argument above our main result:

Proposition 2. *Assume there exist $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ 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 assumptions of this proposition due to Theorem 7. However, this result applies even to distributions which are not absolutely continuous with respect to Lebesgue measure. For instance, if the optimal classification risk is zero, the $\mathbb{P}^*(\eta^* = 1/2) = 0$. To show this statement, notice that if $\inf_A R^\epsilon(A) = 0$, then Theorem 3 implies that for any measures $\mathbb{P}'_0 \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}'_1 \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$, one can conclude that $\mathbb{P}'(\eta' = 1/2) = 0$, where $\mathbb{P}' = \mathbb{P}'_0 + \mathbb{P}'_1$ and $\eta' = d\mathbb{P}'_1/d\mathbb{P}'$.

Proof of Proposition 2. We will show that every minimizing sequence of R_ϕ^ϵ must satisfy (15) and (16). These equations together with the assumption $\mathbb{P}^*(\eta^* = 1/2) = 0$ imply that

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

The strong duality result of Theorem 3 then implies that f_n must be a minimizing sequence of R^ϵ .

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

$$\mathbb{P}^*(|\eta^* - 1/2| < r) < \delta. \quad (19)$$

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 $S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi \circ f_n(\mathbf{x}') \gamma_1^*$ -a.e., and thus (14) 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 [14, Proposition 2.29], the quantity $S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi \circ f_n(\mathbf{x}')$ 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. Analogously, as $C_\phi^*(\eta^*, f_n) \geq C_\phi^*(\eta^*)$, Equation 13 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

$$\gamma_1^*(S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi \circ f_n(\mathbf{x}') \geq \phi(0) - \phi(\alpha_r)) < \delta, \quad (20)$$

$$\gamma_0^*(S_\epsilon(\phi \circ -f_n)(\mathbf{x}) - \phi \circ -f_n(\mathbf{x}') \geq \phi(0) - \phi(\alpha_r)) < \delta, \quad (21)$$

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

$$\mathbb{P}^*(|\eta^* - 1/2| \geq r, f_n \text{sign}(\eta^* - 1/2) \leq \alpha_r) < \delta \quad (22)$$

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

$$\int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \gamma_1^*(\phi \circ f_n(\mathbf{x}') + \phi(0) - \phi(\alpha_r) > \phi(0)) + \delta \leq \mathbb{P}_1^*(\phi \circ f_n > \phi(\alpha_r)) + \delta \quad (23)$$

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

$$\int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \mathbb{P}_1^*(f_n < \alpha_r) + \delta \leq \mathbb{P}_1^*(f_n < \alpha_r, |\eta^* - 1/2| \geq r) + 2\delta. \quad (24)$$

by (19). Next, (22) implies $\mathbb{P}_1^*(\eta^* \geq 1/2 + r, f_n \leq \alpha_r) < \delta$ and consequently

$$\int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \mathbb{P}_1^*(f_n < \alpha_r, \eta^* \leq 1/2 - r) + 3\delta \leq \mathbb{P}_1^*(\eta^* \leq 1/2) + 3\delta.$$

Because δ is arbitrary, this relation implies (15). Next, to prove (16), observe that $\mathbf{1}_{f \geq 0} = \mathbf{1}_{-f \leq 0}$, and thus the inequalities (23) and (24) hold with $-f_n$ in place of f_n , $\mathbb{P}_0, \mathbb{P}_0^*, \gamma_0^*$ in place of $\mathbb{P}_1, \mathbb{P}_1^*, \gamma_1^*$, and (21) in place of (20). Equation 22 further implies $\mathbb{P}_1^*(\eta^* \leq 1/2 - r, f_n \geq -\alpha_r) < \delta$, and these relations imply (16). $\square$

7. Consistency requires uniqueness up to degeneracy

We prove the reverse direction of Theorem 2 by constructing a sequence 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 $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure and that the adversarial Bayes classifier is not unique up to degeneracy. Then any consistent loss ϕ satisfying $C_\phi^*(1/2) = \phi(0)$ is not adversarially consistent.

We will outline the proof in this section below, and the full argument is presented in Appendix D. Before discussing the components of this argument, we illustrate this construction using the adversarial Bayes classifiers depicted in Figure 2. In this example, when $\epsilon \leq (\mu_1 - \mu_0)/2$, two adversarial Bayes classifiers are $A_1 = \emptyset$ and $A_2 = \mathbb{R}$, as depicted in Figure 2b and 2c. These sets are not equivalent up to degeneracy, and in particular the set $\{1\}$ is not an adversarial Bayes classifier. However, the fact that both $\emptyset$ and $\mathbb{R}$ are adversarial Bayes classifiers suggests that $\hat{\eta}(x) \equiv 1/2$ on all of $\mathbb{R}$, and thus $f(x) \equiv 0$ is a minimizer of R_ϕ^ϵ . Consequently, the function sequence

$$f_n(x) = \begin{cases} \frac{1}{n} & \text{if } x \in \{1\} \\ -\frac{1}{n} & \text{otherwise} \end{cases}$$

is a minimizing sequence of R_ϕ^ϵ for which $R^\epsilon(f_n)$ is constant in n and not equal to the adversarial Bayes risk.

To generalize this construction to an arbitrary distribution, one must find a set $\tilde{A}$ contained between two adversarial Bayes classifiers that is not itself an adversarial Bayes classifier. The absolute continuity assumption is key in this step of the argument. Theorem 6 together with a result of [15] imply that when $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, the adversarial Bayes classifier is unique iff $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are equivalent up to degeneracy, see Appendix D for a 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.*

Consequently, if the adversarial Bayes classifier is not unique up to degeneracy, then there is a set $\tilde{A}$ that is not an adversarial Bayes classifier but $\{\hat{\eta} > 1/2\} \subset \tilde{A} \subset \{\hat{\eta} \geq 1/2\}$. Next, we prove that one can replace the value of $\alpha_\phi(1/2)$ by 0 in the minimizer $\alpha_\phi(\hat{\eta}(\mathbf{x}))$ and still retain a minimizer of R_ϕ^ϵ . Formally:

Lemma 6. *Let $\alpha_\phi : [0, 1] \rightarrow \mathbb{R}$ be as in Lemma 1 and define a function $\tilde{\alpha}_\phi : [0, 1] \rightarrow \mathbb{R}$ by*

$$\tilde{\alpha}_\phi(\eta) = \begin{cases} \alpha_\phi(\eta) & \text{if } \eta \neq 1/2 \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

Let $\hat{\eta} : \mathbb{R}^d \rightarrow [0, 1]$ be the function described in Theorem 5. 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 D.1 for a proof of this result. 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

$$f_n(\mathbf{x}) = \begin{cases} \tilde{\alpha}_\phi(\hat{\eta}(\mathbf{x})) & \hat{\eta}(\mathbf{x}) \neq 1/2 \\ \frac{1}{n} & \hat{\eta}(\mathbf{x}) = 1/2, \mathbf{x} \in \tilde{A} \\ -\frac{1}{n} & \hat{\eta}(\mathbf{x}) = 1/2, \mathbf{x} \notin \tilde{A} \end{cases} \quad (26)$$

Then $R^\epsilon(f_n) = R^\epsilon(\tilde{A}) > \inf_A R^\epsilon(A)$ 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 $\mathbb{R}$ -valued. Truncating 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 C for details.

8. Conclusion

In summary, we prove that under a reasonable distributional assumption, a convex loss is adversarially consistent if and only if 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. Analyzing adversarial consistency in the multiclass setting remains an open problem. Prior work [29] has proposed an alternative loss function to address the issue of robust overfitting. Furthermore, many recent papers suggest that classification with a rejection option is a better framework for robust

learning – a non-exhaustive list includes [11, 12, 20, 30]. Analyzing the behavior of these alternative risks using the tools developed in this paper is an open problem as well. Hopefully, our results will aid in the analysis and development of further algorithms for adversarial learning.

Acknowledgements. Special thanks to Jonathan Niles-Weed for helpful conversations.

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.

Competing interests. Author declares no competing interests.

References

- [1] Awasthi, P., Frank, N. S., Mao, A., Mohri, M. & Zhong, Y. (2021). Calibration and consistency of adversarial surrogate losses. *NeurIPS*.
- [2] Awasthi, P., Mao, A., Mohri, M. & Zhong, Y. (2021). A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv: 2105.01550*.
- [3] Awasthi, P., Mao, A., Mohri, M. & Zhong, Y. (2022). H-consistency bounds for surrogate loss minimizers, In ChaudhuriK., Jegelka S., Song L., Szepesvari C., Niu G. & Sabato S. (eds.), *Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research*. PMLR.
- [4] Awasthi, P., Frank, N. S. & Mohri, M. (2023). On the existence of the adversarial bayes classifier (extended version). *arXiv preprint arXiv: 2112.01694*.
- [5] Bao, H., Scott, C. & Sugiyama, M. (2021). Calibrated surrogate losses for adversarially robust classification, *arXiv preprint arXiv: 2005.13748*.
- [6] Bartlett, P. L., Jordan, M. I. & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *J. Am. Stat. Assoc.* **101**(473), 138–156.
- [7] Ben-David, S., Eiron, N. & Long, P. M. (2003). On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.* **66**(3), 496–514.
- [8] Bhagoji, A. N., Cullina, D. & Mittal, P. (2019). Lower bounds on adversarial robustness from optimal transport. *NeurIPS*.
- [9] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G. & Roli, F. (2013). Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, Springer, pp. 387–402.
- [10] Bungert, L., Trillos, N. G. & Murray, R. (2021). The geometry of adversarial training in binary classification. *Inf. Inference. J. IMA*.
- [11] Chen, J., Raghuram, J., Choi, J., Wu, X., Liang, Y. & Jha, S. (2022). Revisiting adversarial robustness of classifiers with a reject option. *Association for the Advancement of Artificial Intelligence*.
- [12] Chen, J., Raghuram, J., Choi, J., Wu, X., Liang, Y. & Jha, S. (2023). Stratified adversarial robustness with rejection. *International Conference of Machine Learning*.
- [13] Deng, Y., Zheng, X., Zhang, T., Chen, C., Lou, G. & Kim, M. (2020). An analysis of adversarial attacks and defenses on autonomous driving models. *IEEE International Conference on Pervasive Computing and Communications*.
- [14] Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*, Vol. **40**, John Wiley & Sons.
- [15] Frank, N. S. (2024). A notion of uniqueness for the adversarial bayes classifier, *arXiv preprint arXiv: 2404.16956*.
- [16] Frank, N. S. & Niles-Weed, J. (2024). The adversarial consistency of surrogate risks for binary classification, *NeurIPS*.
- [17] Frank, N. S. & Niles-Weed, J. (2024). Existence and minimax theorems for adversarial surrogate risks in binary classification. *JMLR*, **25**, 1–41.
- [18] Gnecco-Heredia, L., Chevalyre, Y., Negrevergne, B., Meunier, L. & Pydi, M. S. (2023). On the role of randomization in adversarially robust classification. *NeurIPS*.
- [19] Jylhä, H. (2014). The ℓ^∞ optimal transport: Infinite cyclical monotonicity and the existence of optimal transport maps. *Calc. Var. Partial Dif.* **52**, 303–326.
- [20] Kato, M., Cui, Z. & Fukuhara, Y. (2020). ATRO: Adversarial training with a rejection option. *CoRR*. arxiv: [2010.12905](https://arxiv.org/abs/2010.12905).
- [21] Li, J. D. & Telgarsky, M. (2023). On achieving optimal adversarial test error. *ICLR*.
- [22] Lin, Y. (2004). A note on margin-based loss functions in classification. *Stat. Probabil. Lett.* **68**(1), 73–82.
- [23] Meunier, L., Ettedgui, R., Pinot, R., Chevalyre, Y. & Atif, J. (2022). Towards consistency in adversarial classification. *NeurIPS*.
- [24] Zhang, S. A. M. (2020). Consistency vs. h-consistency: The interplay between surrogate loss functions and the scoring function class. *NeurIPS*.
- [25] Paschali, M., Conjeti, S., Navarro, F. & Navab, N. (2018). *Generalizability vs. Robustness: Adversarial Examples for Medical Imaging*, Springer.
- [26] Long, R. A. S. P. M. (2013). Consistency versus realizable h-consistency for multiclass classification, *ICML*.
- [27] Pydi, M. S. & Jog, V. (2020). Adversarial risk via optimal transport and optimal couplings. *ICML*.
- [28] Pydi, M. S. & Jog, V. (2021). The many faces of adversarial risk. *Neural Information Processing Systems*.

- [29] Robey, A., Latorre, F., Pappas, G. J., Hassani, H. & Cevher, V. (2024). Adversarial training should be cast as a non-zero-sum game. *ICLR*.
- [30] Shah, V., Chaudhari, T. & Manwani, N. (2024). Towards calibrated losses for adversarial robust reject option classification. *ACML*.
- [31] Steinwart, I. (2007). How to compare different loss functions and their risks. *Constr. Approx.* **26**(2), 225–287.
- [32] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Dumitru E., Goodfellow, I. & Fergus, R. (2013). Intriguing properties of neural networks, arXiv preprint arXiv: [1312.6199](https://arxiv.org/abs/1312.6199).
- [33] Trillos, N. G. & Murray, R. (2022). Adversarial classification: Necessary conditions and geometric flows. *JMLR*. **23**, 1–38.
- [34] Trillos, N. G., Jacobs, M. & Kim, J. (2022). The multimarginal optimal transport formulation of adversarial multiclass classification. *JMLR*. **24**, 1–56.
- [35] Trillos, N. G., Jacobs, M. & Kim, J. (2023). On the existence of solutions to adversarial training in multiclass classification, arXiv preprint arXiv: [2305.00075](https://arxiv.org/abs/2305.00075).
- [36] Xu, Y., Raja, K., Ramachandra, R. & Busch, C. (2022). *Adversarial Attacks On Face Recognition Systems*, Springer International Publishing, Cham, pp. 139–161.
- [37] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.* **32**(1). pp. 56–85.

Appendix A. Further Results on the Function $\hat{\eta}$ — Proof of Theorem 6

We prove that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ minimize R_ϕ^ϵ by showing that their adversarial classification risks equal $\bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)$, for the measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ in Theorem 5.

Proposition 4. *Let $\hat{\eta}$ be the function in Theorem 5. Then the sets $\{\hat{\eta} > 1/2\}$, $\{\hat{\eta} \geq 1/2\}$ are both adversarial 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 5 and set $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$, $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. 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 2 implies that the function $\hat{\eta}(\mathbf{x})$ assumes its infimum on an ball $\overline{B_\epsilon(\mathbf{x})}$ γ_1^* -a.e. and therefore $S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}^c})(\mathbf{x}) = \mathbf{1}_{\{I_\epsilon(\hat{\eta}(\mathbf{x}) > 1/2\}^c} \gamma_1^*$ -a.e. (Recall the notation I_ϵ was defined in (10). Item 2 further implies that $\mathbf{1}_{\{I_\epsilon(\hat{\eta}(\mathbf{x}) > 1/2\}^c} = \mathbf{1}_{\{\hat{\eta}(\mathbf{x}) > 1/2\}^c} \gamma_1^*$ -a.e. and consequently,

$$S_\epsilon(\mathbf{1}_{\{\hat{\eta}(\mathbf{x}) > 1/2\}^c})(\mathbf{x}) = \mathbf{1}_{\{\hat{\eta}(\mathbf{x}') > 1/2\}^c} \gamma_1^* \text{-a.e.} \quad (27)$$

An analogous argument shows

$$S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}})(\mathbf{x}) = \mathbf{1}_{\{\hat{\eta}(\mathbf{x}') > 1/2\}} \gamma_0^* \text{-a.e.} \quad (28)$$

Equations (27) and (28) 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 1 of Theorem 5 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 3 implies that $\{\hat{\eta} > 1/2\}$ must minimize R^ϵ . $\square$

Finally, the complementary slackness conditions from [15, Theorem 2.4] characterize minimizers of R^ϵ and maximizers of $\bar{R}$, and this characterization proves Equations (11) and (12). Verifying these conditions would be another method of proving Proposition 4.

Theorem 8. *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 $W_\infty(\mathbb{P}_0^*, \mathbb{P}_0) \leq \epsilon$, $W_\infty(\mathbb{P}_1^*, \mathbb{P}_1) \leq \epsilon$, and*

1)

$$\int S_\epsilon(\mathbf{1}_{A^c})d\mathbb{P}_1 = \int \mathbf{1}_{A^c}d\mathbb{P}_1^* \quad \text{and} \quad \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 = \int \mathbf{1}_Ad\mathbb{P}_0^* \quad (29)$$

2)

$$C(\eta^*, \mathbf{1}_A(\mathbf{x}')) = C^*(\eta^*(\mathbf{x}')) \quad \mathbb{P}^*\text{-a.e.} \quad (30)$$

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 $S_\epsilon(\mathbf{1}_{A^c})(\mathbf{x}) \geq \mathbf{1}_{A^c}(\mathbf{x}') \gamma_1^*\text{-a.e.}$ and $S_\epsilon(\mathbf{1}_A)(\mathbf{x}) \geq \mathbf{1}_A(\mathbf{x}') \gamma_0^*\text{-a.e.}$, the complementary slackness condition in (29) is equivalent to

$$S_\epsilon(\mathbf{1}_{A^c})(\mathbf{x}) = \mathbf{1}_{A^c}(\mathbf{x}') \quad \gamma_1^*\text{-a.e.} \quad \text{and} \quad S_\epsilon(\mathbf{1}_A)(\mathbf{x}) = \mathbf{1}_A(\mathbf{x}') \quad \gamma_0^*\text{-a.e.} \quad (31)$$

This observation completes the proof of Theorem 6.

Proof of Theorem 6. Let $\hat{\eta}, \mathbb{P}_0^*, \mathbb{P}_1^*$ be the function and measures of Theorem 5, and set $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$, $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. 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 .

Proposition 4 proves that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are in fact adversarial Bayes classifiers. If A is any adversarial Bayes classifier, the complementary slackness condition (30) implies that $\mathbf{1}_{\eta^* > 1/2} \leq \mathbf{1}_A \leq \mathbf{1}_{\eta^* \geq 1/2}$ $\mathbb{P}^*\text{-a.e.}$ Thus Item 1 implies that

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

and

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

The complementary slackness condition (31) then implies Equations (11) and (12). $\square$

Appendix B. Uniqueness up to Degeneracy and $\hat{\eta}(\mathbf{x})$ —Proof of Theorem 7

Theorem 3.4 of [15] proves the following result:

Theorem 9. 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 , the value of $\int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0$ is unique or the value of $\int S_\epsilon(\mathbf{1}_{A^c})d\mathbb{P}_1$ is unique

Thus it remains to show that Item 2 of Theorem 9 is equivalent to Item 2 of Theorem 7. We will apply the complementary slackness conditions of Theorem 8.

Proof of Theorem 7. Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be the measures of Theorem 5.

First, we show that Item 2 implies Item 2. Assume that Item 2 holds. Notice that for an adversarial Bayes classifier A ,

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

where R^* is the minimal value of R^ϵ . Thus amongst all adversarial Bayes classifiers A , the value of $\int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0$ is unique iff the value of $\int S_\epsilon(\mathbf{1}_{A^c})d\mathbb{P}_1$ is unique. Thus Item 2 implies both $\int S_\epsilon(\mathbf{1}_{A_1})d\mathbb{P}_0 = \int S_\epsilon(\mathbf{1}_{A_2})d\mathbb{P}_0$ and $\int S_\epsilon(\mathbf{1}_{A_1^c})d\mathbb{P}_1 = \int S_\epsilon(\mathbf{1}_{A_2^c})d\mathbb{P}_1$ for any two adversarial Bayes classifiers A_1 and A_2 .

Applying this statement to the adversarial Bayes classifiers $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ produces

$$\int S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}^c})d\mathbb{P}_1 = \int S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}^c})d\mathbb{P}_1 \quad \text{and} \quad \int S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}})d\mathbb{P}_0 = \int S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}})d\mathbb{P}_0$$

The complementary slackness condition (29) 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 1 of Theorem 5 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, $\mathbb{P}^*(\eta^* = 1/2) = 0$.

To show the other direction, we apply the inequalities in Theorem 6. The complementary slackness conditions in Theorem 8 and the first inequality in Theorem 6 imply that for any adversarial Bayes classifier A ,

$$\int \mathbf{1}_{\{\eta^* < 1/2\}} d\mathbb{P}_1^* \leq \int S_\epsilon(\mathbf{1}_{A^c}) d\mathbb{P}_1 \leq \int \mathbf{1}_{\{\hat{\eta}^* \leq 1/2\}} d\mathbb{P}_1^*$$

Consequently, if $\mathbb{P}^*(\eta^* = 1/2) = 0$, then $\int \mathbf{1}_{\{\eta^* < 1/2\}} d\mathbb{P}_1^* = \int S_\epsilon(\mathbf{1}_{A^c}) d\mathbb{P}_1$, which implies that $\int S_\epsilon(\mathbf{1}_{A^c}) d\mathbb{P}_1$ assumes a unique value over all possible adversarial Bayes classifiers. $\square$

Appendix C. Technical loss function proofs—Proof of Lemmas 1 and 4

To begin, we prove a result that compares minimizers of $C_\phi(\eta, \cdot)$ for differing values of η .

This result is then used to prove Lemma 1.

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 $C_\phi(\eta_1, \cdot)$, if $\alpha < \alpha_\phi(\eta_1)$ then $C_\phi(\eta_1, \alpha) > C_\phi^*(\eta_1)$ and thus $C_\phi(\eta_2, \alpha) > C_\phi(\eta_2, \alpha_\phi(\eta_1))$. Consequently, every minimizer of $C_\phi(\eta_2, \cdot)$ must be greater than or equal to $\alpha_\phi(\eta_1)$. $\square$

Next we use this result to prove Lemma 1.

Proof of Lemma 1. For $\eta_2 > \eta_1$, apply Lemma 7 with the choice $\alpha_2^* = \alpha_\phi(\eta_2)$. $\square$

Next, 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. 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, $\eta = 1/2$. $\square$

Combining these two results proves Lemma 4.

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 $k_r > 0$.

Lastly, $\phi(\alpha_r) < \phi(0)$ implies $\alpha_r > 0$. $\square$

Appendix D. Deferred arguments from Section 7— Proof of Proposition 3

To start, 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 [15, Proposition 5.1].

Theorem 10. Assume that $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure 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 $S_\epsilon(\mathbf{1}_{A_1^c}) = S_\epsilon(\mathbf{1}_{A_2^c})$ - $\mathbb{P}_1$ -a.e.

Notice that when there is a single equivalence class, the equivalence between Item 1 and Item 2 of Theorem 10 is simply the equivalence between Item 1 and Item 2 in Theorem 9. This result together with Theorem 6 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 6 and Item 2 of Theorem 10 imply that $S_\epsilon(\mathbf{1}_A) = S_\epsilon(\mathbf{1}_{\{\hat{\eta} > 1/2\}})$ - $\mathbb{P}_0$ -a.e. Item 2 of Theorem 10 again implies that A and $\{\hat{\eta} > 1/2\}$ must be equivalent up to degeneracy, and consequently the adversarial Bayes classifier must be unique up to degeneracy.

Conversely, if the adversarial Bayes classifier is unique up to degeneracy, then the adversarial Bayes classifiers $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ must be equivalent up to degeneracy. $\square$

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 (26). 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 a sequence of $\mathbb{R}$ -valued functions f_n 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 (26) and let $\tilde{\alpha}_\phi$ be the function in Lemma 6. Lemma 8 implies that $\tilde{\alpha}_\phi(\eta) \neq 0$ whenever $\eta \neq 1/2$ and thus $\{f_n > 0\} = \tilde{A}$ for all n . We will show that in the limit $n \rightarrow \infty$, 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 due to Lemma 6.

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

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

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

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

$$S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta}))(\mathbf{x}) = \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta}))(\mathbf{x}), \phi(0)) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta}))(\mathbf{x}) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \quad (33)$$

and similarly

$$S_\epsilon(\phi \circ f_n)(\mathbf{x}) \leq \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ \alpha_\phi(\hat{\eta}))(\mathbf{x}), \phi(-\frac{1}{n})) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ \alpha_\phi(\hat{\eta}))(\mathbf{x}) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \quad (34)$$

$$S_\epsilon(\phi \circ -f_n)(\mathbf{x}) \leq \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta}))(\mathbf{x}), \phi(-\frac{1}{n})) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta}))(\mathbf{x}) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \quad (35)$$

Therefore, by comparing (34) with (32) and (35) with (33), one can conclude that

$$\limsup_{n \rightarrow \infty} S_\epsilon(\phi \circ f_n) \leq S_\epsilon(\phi \circ \tilde{\alpha}_\phi(\hat{\eta})) \quad \text{and} \quad \limsup_{n \rightarrow \infty} S_\epsilon(\phi \circ -f_n) \leq S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta})). \quad (36)$$

Furthermore, the right-hand of (34) is bounded above by $S_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})) + \phi(-1)$ while the right-hand of (35) is bounded above by $S_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})) + \phi(-1)$. Thus the dominated convergence theorem and (36) implies that

$$\limsup_{n \rightarrow \infty} R_\phi^\epsilon(f_n) \leq R_\phi^\epsilon(\tilde{\alpha}_\phi(\hat{\eta}))$$

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 threshold a subsequence f_{n_j} of f_n at an appropriate value T_j . If g is an $\overline{\mathbb{R}}$ -valued function and $g^{(N)}$ is the function g thresholded at N , then $\lim_{N \rightarrow \infty} R_\phi^\epsilon(g^{(N)}) = R_\phi^\epsilon(g)$.

Lemma 9. *Let g be an $\overline{\mathbb{R}}$ -valued function and let $g^{(N)} = \min(\max(g, -N), N)$. Then $\lim_{N \rightarrow \infty} R_\phi^\epsilon(g^{(N)}) = R_\phi^\epsilon(g)$.*

See Appendix D.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 $R_\phi^\epsilon(f_{n_j}) - \inf_f R_\phi^\epsilon(f) < 1/j$. Next, Lemma 9 implies that for each j one can pick a threshold 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 threshold 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 . □

D.1 Proof of Lemma 6

The proof of Lemma 6 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 5, and then Theorem 4 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 10. *If $C_\phi^*(1/2) = \phi(0)$, then the function $\tilde{\alpha}_\phi : [0, 1] \rightarrow \overline{\mathbb{R}}$ defined in (25) is non-decreasing and maps each η to a minimizer of $C_\phi(\eta, \cdot)$.*

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 $\eta \in (1/2, 1]$. Consequently, $\tilde{\alpha}_\phi$ is non-decreasing on all of $[0, 1]$. □

This result together with the properties of $\mathbb{P}_0^*, \mathbb{P}_1^*$ in Theorem 5 suffice to prove Lemma 6.

Proof of Lemma 6. Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be the measures of Theorem 5 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 4 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 2 of Theorem 5 and Lemma 10 imply that

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

and

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

(Recall the notation I_ϵ was introduced in (10).) 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 1 of Theorem 5 implies that $\hat{\eta}(\mathbf{x}') = \eta^*(\mathbf{x}')$ and consequently

$$R_\phi^\epsilon(\tilde{\alpha}_\phi(\hat{\eta})) = \int C_\phi(\eta^*, \tilde{\alpha}_\phi(\hat{\eta})) d\mathbb{P}^* = \int C_\phi(\eta^*, \tilde{\alpha}_\phi(\eta^*)) d\mathbb{P}^* = \int C_\phi^*(\eta^*) d\mathbb{P}^* = \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*)$$

Therefore, the strong duality result in Theorem 4 implies that $\tilde{\alpha}_\phi(\hat{\eta})$ must minimize R_ϕ^ϵ . $\square$

D.2 Proof of Lemma 9

This argument is taken from the proof of Lemma 8 in [16].

Proof of Lemma 9. 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 $N \rightarrow \infty$. 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

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