Adversarial Surrogate Risk Bounds for Binary Classification

Anonymous authors
Paper under double-blind review

Abstract

A central concern in classification is the vulnerability of machine learning models to adversarial attacks. Adversarial training is one of the most popular techniques for training robust classifiers, which involves minimizing an adversarial surrogate risk. Recent work characterized when a minimizing sequence of an adversarial surrogate risk is also a minimizing sequence of the adversarial classification risk for binary classification— a property known as *adversarial consistency*. However, these results do not address the rate at which the adversarial classification risk converges to its optimal value for such a sequence of functions that minimize the adversarial surrogate. This paper provides surrogate risk bounds that quantify that convergence rate. Additionally, we derive distribution-dependent surrogate risk bounds in the standard (non-adversarial) learning setting, that may be of independent interest.

1 Introduction

A central concern regarding sophisticated machine learning models is their susceptibility to adversarial attacks. Prior work (Biggio et al., 2013; Szegedy et al., 2013) demonstrated that imperceptible perturbations can derail the performance of neural nets. As such models are used in security-critical applications such as facial recognition (Xu et al., 2022) and medical imaging (Paschali et al., 2018), training robust models remains a central concern in machine learning.

In the standard classification setting, the *classification risk* is the proportion of incorrectly classified data. Rather that minimizing this quantity directly, which is a combinatorial optimization problem, typical machine learning algorithms perform gradient descent on a well-behaved alternative *surrogate risk*. If a sequence of functions that minimizes this surrogate risk also minimizes the classification risk, then the surrogate risk is referred to as *consistent* for that specific data distribution. In addition to consistency, one would hope that minimizing the surrogate risk would be an efficient method for minimizing the classification risk. This convergence rate can be bounded by *surrogate risk bounds*, which are functions that provide a bound on the excess classification risk in terms of the excess surrogate risk.

In the standard binary classification setting, consistency and surrogate risk bounds are well-studied topics (Bartlett et al., 2006; Lin, 2004; Steinwart, 2007; Zhang, 2004). On the other hand, fewer results are known about the adversarial setting. The adversarial classification risk incurs a penalty when a point can be perturbed into the opposite class. Similarly, adversarial surrogate risks involve computing the worst-case value (i.e. supremum) of a loss function over an ϵ-ball. Frank & Niles-Weed (2024a) characterized which risks are consistent for all data distributions, and the corresponding losses are referred to as *adversarially consistent*. Unfortunately, no convex loss function can be adversarially consistent for all data distributions (Meunier et al., 2022). On the other hand, Frank (2025) showed that such situations are rather atypical— when the data distribution is absolutely continuous, a surrogate risk is adversarially consistent so long as the adversarial Bayes classifier satisfies a certain notion of uniqueness called *uniqueness up to degeneracy*. While these results characterize consistency, none describe convergence rates.

Our Contributions:

- We prove a linear surrogate risk bound for adversarially consistent losses (Theorem 11).
- If the 'distribution of optimal attacks' satisfies a bounded noise condition, we prove a linear surrogate risk bound, under mild conditions on the loss function (Theorems 11 and 12).
- We prove a distribution dependent surrogate risk bound that applies whenever a loss is adversarially consistent for a data distribution (Theorem 13).

Notably, this last bullet applies to convex loss functions. Due to the consistency results in prior work (Frank, 2025; Frank & Niles-Weed, 2024a; Meunier et al., 2022), one cannot hope for distribution independent surrogate bounds for non-adversarially consistent losses. To the best of the authors' knowledge this paper is the first to prove surrogate risk bounds for the risks most commonly used in adversarial training, see Section 6 for a comparison with prior work. Understanding the optimality of the bounds presented in this paper remains an open problem.

2 Background and Preliminaries

2.1 Surrogate Risks

This paper studies binary classification on $\mathbb{R}^d$ with labels -1 and $+1$. The measures $\mathbb{P}_0$, $\mathbb{P}_1$ describe the probabilities of finding data with labels -1, $+1$, respectively, in subset of $\mathbb{R}^d$. The *classification risk* of a set A is the misclassification rate when points in A are classified as $+1$ and points in A^C are classified as -1:

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

The minimal classification risk over all Borel sets is denoted R_*. As the derivative of an indicator function is zero wherever it is defined, the empirical version of this risk cannot be optimized with first order descent methods. Consequently, common machine learning algorithms minimize a different quantity called a *surrogate risk*. The surrogate risk of a function f is defined as

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

In practice, the *loss function* ϕ selected so that it has well-behaved derivative. In this paper, We assume:

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

The minimal surrogate risk over all Borel measurable functions is denoted $R_{\phi,*}$. After optimizing the surrogate risk, a classifier is obtained by threshholding the resulting f at zero. Consequently, we define the classification error of a function by $R(f) = R(\{f > 0\})$ or equivalently,

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

It remains to verify that minimizing the surrogate risk R_ϕ will also minimize the classification risk R.

Definition 1. *The loss function ϕ is* consistent for the distribution $\mathbb{P}_0$, $\mathbb{P}_1$ *if every minimizing sequence of R_ϕ is also a minimizing sequence of R when the data is distributed according to $\mathbb{P}_0$ and $\mathbb{P}_1$. The loss function ϕ is* consistent *if it is consistent for all data distributions.*

Prior work establishes conditions under which many common loss functions are consistent.

Theorem 1. *A convex loss ϕ is consistent iff it is differentiable at zero and $\phi'(0) < 0$.*

See (Bartlett et al., 2006, Theorem 2). Furthermore, (Frank & Niles-Weed, 2024a, Proposition 3) establishes a condition that applies to non-convex losses:

Theorem 2. *If* $\inf_\alpha 1/2(\phi(\alpha) + \phi(-\alpha)) < \phi(0)$*, then the loss ϕ is consistent.*

The ρ-margin loss $\phi_\rho(\alpha) = \min(1, \max(1-\alpha/\rho, 0))$ and the shifted sigmoid loss $\phi_\tau(\alpha) = 1/(1+\exp(\alpha-\tau))$, $\tau > 0$, both satisfy this criterion. However, a convex loss ϕ cannot satisfy this condition:

$$\frac{1}{2}\left(\phi(\alpha) + \phi(-\alpha)\right) \geq \phi\big(\frac{1}{2}\alpha + \frac{1}{2}\cdot -\alpha\big) = \phi(0). \tag{1}$$

In addition to consistency, understanding convergence *rates* is a key concern. Specifically, prior work (Bartlett et al., 2006; Zhang, 2004) establishes *surrogate risk bounds* of the form $\Psi(R(f) - R_*) \leq R_\phi(f) - R_{\phi,*}$ for some function Ψ. This inequality bounds the convergence rate of $R(f) - R_*$ in terms of the convergence of $R_\phi(f) - R_{\phi,*}$.

The values R_*, $R_{\phi,*}$ can be expressed in terms of the data distribution by re-writing these quantities in terms of the total probability measure $\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$ and the conditional probability of the label $+1$, given by $\eta(x) = d\mathbb{P}_1/d\mathbb{P}$. An equivalent formulation of the classification risk is

$$R(f) = \int C(\eta(\mathbf{x}), f(\mathbf{x}))d\mathbb{P}(\mathbf{x}) \tag{2}$$

with

$$C(\eta, \alpha) = \eta \mathbf{1}_{\alpha \leq 0} + (1-\eta)\mathbf{1}_{\alpha > 0}, \tag{3}$$

and the minimal classification risk is found by minimizing the integrand of (2) at each $\mathbf{x}$. Define

$$C^*(\eta) = \inf_\alpha C(\eta, \alpha) = \min(\eta, 1-\eta), \tag{4}$$

then the minimal classification risk is

$$R_* = \int C^*(\eta(\mathbf{x}))d\mathbb{P}(\mathbf{x}).$$

Analogously, the surrogate risk in terms of η and $\mathbb{P}$ is

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

and the minimal surrogate risk is

$$R_{\phi,*} = \int C_\phi^*(\eta(\mathbf{x}))d\mathbb{P}(\mathbf{x})$$

with the conditional risk $C_\phi(\eta, \alpha)$ and minimal conditional risk $C_\phi^*(\eta)$ defined by

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

Notice that minimizers to R_ϕ may need to be $\overline{\mathbb{R}}$-valued— consider the exponential loss $\phi(\alpha) = e^{-\alpha}$ and a distribution with $\eta(\mathbf{x}) \equiv 1$. Then the only minimizer to R_ϕ would be $+\infty$.

The consistency of ϕ can be fully characterized by the properties of the function $C_\phi^*(\eta)$.

Theorem 3. *A loss ϕ is consistent iff $C_\phi^*(\eta) < \phi(0)$ for all $\eta \neq 1/2$.*

Surprisingly, this criterion has not appeared in prior work. See Appendix A for a proof.

In terms of the function C_ϕ^*, Theorem 2 states that any loss ϕ with $C_\phi^*(1/2) < \phi(0)$ is consistent.

The function C_ϕ^* is a key component of surrogate risk bounds from prior work. Specifically, Bartlett et al. (2006) show:

Theorem 4. *Let ϕ be any loss satisfying Assumption 1 with $C_\phi^*(1/2) = \phi(0)$ and define*

$$\Psi(\theta) = \phi(0) - C_\phi^*\left(\frac{1+\theta}{2}\right).$$

Then

$$\Psi(C(\eta, f) - C^*(\eta)) \leq C_\phi(\eta, f) - C^*_\phi(\eta) \tag{7}$$

and consequently

$$\Psi(R(f) - R_*) \leq R_\phi(f) - R^*_\phi. \tag{8}$$

Equation 8 is a consequence of (7) and Jensen's inequality. Furthermore, a result of (Bartlett et al., 2006) implies a linear bound when $C^*_\phi(1/2) < \phi(0)$:

$$R(f) - R_* \leq \frac{1}{\phi(0) - C^*_\phi(1/2)}(R_\phi(f) - R_{\phi,*}) \tag{9}$$

Furthermore, a distribution with zero classification error R_* has the surrogate risk bound

$$R(f) - R_* \leq \frac{1}{\phi(0)}(R_\phi(f) - R_{\phi,*}) \tag{10}$$

so long as $\phi(0) > 0$. Such distributions are referred to as *realizable*. A proof of this result that transfers directly to the adversarial scenario is provided in Appendix B.1.

A distribution is said to satisfy *Massart's noise condition* (Massart & Nédélec, 2006) if there is an $\alpha \in (0, 1/2]$ such that $|\eta - 1/2| \geq \alpha$ holds $\mathbb{P}$-a.e. Under this condition, Massart & Nédélec (2006) establish improved sample complexity guarantees. Furthermore, such distributions exhibit a linear surrogate loss bound as well. These linear bounds, the realizable bounds from (10), and the linear bounds from (9) are summarized in a single statement below.

Proposition 1. *Let η, $\mathbb{P}$ be a distribution that satisfies $|\eta - 1/2| \geq \alpha$ $\mathbb{P}$-a.e. with a constant $\alpha \in [0, 1/2]$, and let ϕ be a loss with $\phi(0) > C^*_\phi(1/2 - \alpha)$. Then for all $|\eta - 1/2| \geq \alpha$,*

$$C(\eta, f) - C^*(\eta) \leq \frac{1}{\phi(0) - C^*_\phi(\frac{1}{2} - \alpha)}(C_\phi(\eta, f) - C^*_\phi(\eta)) \tag{11}$$

and consequently

$$R(f) - R_* \leq \frac{1}{\phi(0) - C^*_\phi(\frac{1}{2} - \alpha)}(R_\phi(f) - R_{\phi,*}) \tag{12}$$

When $\alpha \neq 0$, this surrogate risk bound proves a linear convergence rate under Massart's noise condition. If $\alpha = 0$ and $C^*_\phi(1/2) < \phi(0)$, then the bound in (12) reduces to (9) while if $\alpha = 1/2$ then this bound reduces to (10). See Appendix B.2 for a proof of this result. One of the main results of this paper is that (12) generalizes to adversarial risks.

Note that the surrogate risk bound of Theorem 4 can be linear even for convex loss functions. For the hinge loss $\phi(\alpha) = \max(1 - \alpha, 0)$, the function ϕ computes to $\phi(\theta) = |\theta|$. Prior work (Frongillo & Waggoner, 2021, Theorem 1) observed a linear surrogate bound for piecewise linear losses: if ϕ is piecewise linear, then $C^*_\phi(\eta)$ is piecewise linear and Jensen's inequality implies a linear surrogate bound so long as ϕ is consistent (due to Theorem 3). On the other hand, (Frongillo & Waggoner, 2021, Theorem 2) show that convex losses which are locally strictly convex and Lipschitz achieve at best a square root surrogate risk rate.

Mahdavi et al. (2014) emphasize the importance of a linear convergence rate in a surrogate risk bound. Their paper studies the sample complexity of estimating a classifier with a surrogate risk. They note that typically convex surrogate losses exhibiting favorable sample complexity do not satisfy favorable surrogate risk bounds, due to the results of (Frongillo & Waggoner, 2021). Consequently, Proposition 1 hints that proving favorable sample complexity guarantees for learning with convex surrogate risks could require distributional assumptions, such as Massart's noise condition.

2.2 Adversarial Risks

This paper extends surrogate risk bounds of Equations (8), (10) and (12) to adversarial risks. The adversarial classification risk incurs a penalty of 1 whenever a point $\mathbf{x}$ can be perturbed into the opposite class. This penalty can be expressed in terms of supremums of indicator functions— the adversarial classification risk incurs a penalty of 1 whenever $\sup_{\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon} \mathbf{1}_A(\mathbf{x}') = 1$ or $\sup_{\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon} \mathbf{1}_{A^C}(\mathbf{x}') = 1$. Define

$$S_\epsilon(g)(\mathbf{x}) = \sup_{\|\mathbf{x}-\mathbf{x}'\|\leq\epsilon} g(\mathbf{x}').$$

The adversarial classification risk is then

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

and the adversarial surrogate risk is[1]

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

A minimizer of the adversarial classification risk is called an *adversarial Bayes classifier*. After optimizing the surrogate risk, a classifier is obtained by threshholding the resulting function f at zero. Consequently, the adversarial classification error of a function f is defined as $R^\epsilon(f) = R^\epsilon(\{f > 0\})$ or equivalently,

$$R^\epsilon(f) = \int S_\epsilon(\mathbf{1}_{f\leq 0})d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_{f>0})d\mathbb{P}_0. \tag{13}$$

Just as in the standard case, one would hope that minimizing the adversarial surrogate risk would minimize the adversarial classification risk.

Definition 2. *We say a loss ϕ is* adversarially consistent for the distribution $\mathbb{P}_0$, $\mathbb{P}_1$ *if any minimizing sequence of R^ϵ_ϕ is also a minimizing sequence of R^ϵ. We say that ϕ is* adversarially consistent *if it is adversarially consistent for every possible $\mathbb{P}_0$, $\mathbb{P}_1$.*

Theorem 2 of (Frank & Niles-Weed, 2024a) characterizes the adversarially consistent losses:

Theorem 5. *The loss ϕ is adversarially consistent iff $C^*_\phi(1/2) < \phi(0)$.*

Theorem 2 implies that every adversarially consistent loss is also consistent. Unfortunately, (1) shows that no convex loss is adversarially consistent. However, the data distribution for which adversarial consistency fails presented in (Meunier et al., 2022) is fairly atypical: Let $\mathbb{P}_0$, $\mathbb{P}_1$ be the uniform distributions on $\overline{B_\epsilon(\mathbf{0})}$. Then one can show that the function sequence

$$f_n = \begin{cases} \frac{1}{n} & \mathbf{x} \neq 0 \\ -\frac{1}{n} & \mathbf{x} = 0 \end{cases} \tag{14}$$

minimizes R^ϵ_ϕ but not R^ϵ whenever $C^*_\phi(1/2) = \phi(0)$ (See Proposition 2 of (Frank & Niles-Weed, 2024a)). A more refined analysis relates adversarial consistency for losses with $C^*_\phi(1/2) = \phi(0)$ to a notion of uniqueness of the adversarial Bayes classifier for losses satisfying $C^*_\phi(1/2) = \phi(0)$.

Definition 3. *Two adversarial Bayes classifiers A_1, A_2 are* equivalent up to degeneracy *if any 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.*

Theorem 3.3 of (Frank, 2024) proves that whenever $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, then equivalence up to degeneracy is in fact an equivalence relation. Next, Theorem 4 of (Frank, 2025) relates this condition to the consistency of ϕ.

Theorem 6. *Let ϕ be a loss with $C^*_\phi(1/2) = \phi(0)$ and assume that $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure. Then ϕ is adversarially consistent for the data distribution given by $\mathbb{P}_0$, $\mathbb{P}_1$ iff the adversarial Bayes classifier is unique up to degeneracy.*

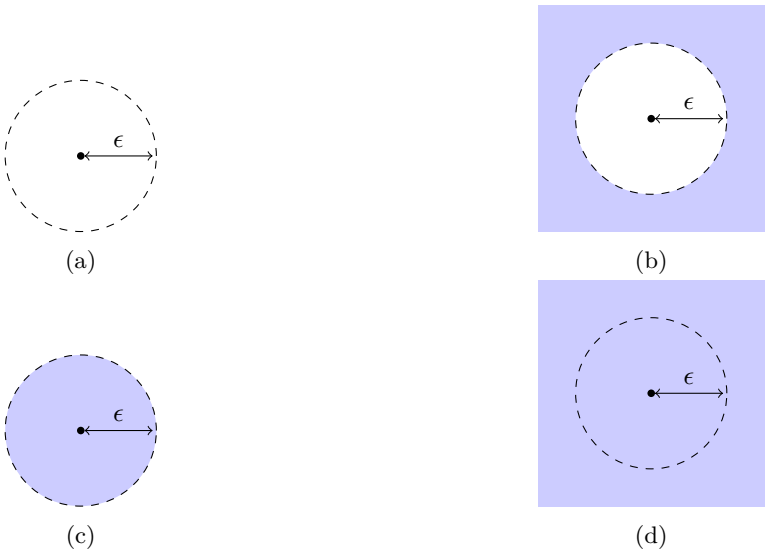


Figure 1: Adversarial Bayes classifiers for the example considered in (14). The adversarial Bayes classifiers in (a) and (b) are equivalent up to degeneracy and the the adversarial Bayes classifiers in (c) and (d) are equivalent up to degeneracy, but the adversarial Bayes classifiers in (a) and (c) are not equivalent up to degeneracy.

The extension of Theorem 4 to the adversarial setting must reflect the consistency results of Theorems 5 and 6.

2.3 Minimax Theorems

A central tool in analyzing the adversarial consistency of surrogate risks is minimax theorems. These results allow one to express adversarial risks in a 'pointwise' manner analogous to (5). We will then combine this 'pointwise' expression together with the proof of Theorem 4 to produce surrogate bounds for adversarial risks.

These minimax theorems utilize the ∞-Wasserstein (W_∞) metric from optimal transport. Let $\mathbb{Q}$ and $\mathbb{Q}'$ be two finite positive measures with the same total mass. Informally, the measure $\mathbb{Q}'$ is within ϵ of $\mathbb{Q}$ in the W_∞ metric if one can achieve the measure $\mathbb{Q}'$ by moving points of $\mathbb{Q}$ by at most ϵ.

The W_∞ metric is formally defined in terms of the set of *couplings* between $\mathbb{Q}$ and $\mathbb{Q}'$. A Borel measure γ on $\mathbb{R}^d \times \mathbb{R}^d$ is a coupling between $\mathbb{Q}$ and $\mathbb{Q}'$ if its first marginal is $\mathbb{Q}$ and its second marginal is $\mathbb{Q}'$, or in other words, $\gamma(A \times \mathbb{R}^d) = \mathbb{Q}(A)$ and $\gamma(\mathbb{R}^d \times A) = \mathbb{Q}'(A)$ for all Borel sets A. Let $\Pi(\mathbb{Q}, \mathbb{Q}')$ be the set of all couplings between the measures $\mathbb{Q}$ and $\mathbb{Q}'$. Then the W_∞ between $\mathbb{Q}$ and $\mathbb{Q}'$ is then

$$W_\infty(\mathbb{Q}, \mathbb{Q}') = \inf_{\gamma \in \Pi(\mathbb{Q}, \mathbb{Q}')} \operatorname*{ess\,sup}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\|. \tag{15}$$

Theorem 2.6 of (Jylhä, 2014) proves that the infimum in (15) is always attained. The ϵ-ball around $\mathbb{Q}$ in the W_∞ metric is

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

The minimax theorem below will relate the adversarial risks R_ϕ^ϵ, R^ϵ to dual problems in which an adversary seeks to maximize some dual quantity over Wasserstein-∞ balls. Specifically, one can show:

[1]In order to define the risks R_ϕ^ϵ and R^ϵ, one must argue that $S_\epsilon(g)$ is measurable. Theorem 1 of (Frank & Niles-Weed, 2024b) proves that whenever g is Borel, $S_\epsilon(g)$ is always measurable with respect to the completion of any Borel measure.

Lemma 1. *Let g be a Borel function. Let γ be a coupling between the measures $\mathbb{Q}$ and $\mathbb{Q}'$ supported on $\Delta_\epsilon = \{(\mathbf{x}, \mathbf{x}') : \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon\}$. Then $S_\epsilon(g)(\mathbf{x}) \geq g(\mathbf{x}')$ γ-a.e. and consequently*

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

See Appendix C for a proof. Thus, applying Lemma 1, the quantity $\inf_A R^\epsilon(A)$ can be lower bounded by an infimum followed by a supremum. Is it possible to swap this infimum and supremum? (Pydi & Jog, 2021) answers this question in the affirmative. Let $C^*(\eta)$ be as defined in (4) and let

$$\bar{R}(\mathbb{P}_0', \mathbb{P}_1') = \inf_{A \text{ Borel}} \int \mathbf{1}_{A^C} d\mathbb{P}_1' + \int \mathbf{1}_A d\mathbb{P}_0' = \int C^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_1' + \mathbb{P}_0')} \right) d(\mathbb{P}_0' + \mathbb{P}_1'). \tag{16}$$

Theorem 7. *Let $\bar{R}$ be as defined in (16). Then*

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

Furthermore, equality is attained at some Borel measurable A, $\mathbb{P}_0^*$, and $\mathbb{P}_1^*$ with $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$ and $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$.

See (Frank & Niles-Weed, 2024a, Theorem 1) for a proof of this statement. The maximizers $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ can be interpreted as optimal adversarial attacks (see discussion following (Frank & Niles-Weed, 2024b, Theorem 7)). Frank (2024, Theorem 3.4) provide a criterion for uniqueness up to degeneracy in terms of dual maximizers.

Theorem 8. *The following are equivalent:*

A) The adversarial Bayes classifier is unique up to degeneracy

B) 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 $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$

In other words, the adversarial Bayes classifier is unique up to degeneracy iff the region where both classes are equally probable has measure zero under some optimal adversarial attack. Theorems 6 and 8 relate adversarial consistency and the dual problem, suggesting that these optimal adversarial attacks $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ may appear in adversarial surrogate bounds.

Frank & Niles-Weed (2024b) prove an minimax principle analogous to Theorem 7 for the adversarial surrogate risk. Let $C_\phi^*(\eta)$ be as defined in (6) and let

$$\bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') = \inf_{f \text{ Borel}} \int \phi(f) d\mathbb{P}_1' + \int \phi(-f) d\mathbb{P}_0' = \int C_\phi^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_1' + \mathbb{P}_0')} \right) d(\mathbb{P}_0' + \mathbb{P}_1') \tag{17}$$

Theorem 9. *Let $\bar{R}_\phi$ be defined as in (17). Then*

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

Furthermore, equality is attained at some Borel measurable f^*, $\mathbb{P}_0^*$, and $\mathbb{P}_1^*$ with $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$ and $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$.

Just as in the non-adversarial scenario, R_ϕ^ϵ may not assume its infimum at an $\mathbb{R}$-valued function. However, (Frank & Niles-Weed, 2024a, Lemma 8) show that

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

Lastly, one can show that maximizers of $\bar{R}_\phi$ are always maximizers of $\bar{R}$ as well. In other words— optimal attacks on minimizers of the adversarial surrogate R^ϵ_ϕ are always optimal attacks on minimizers of the adversarial classification risk R^ϵ as well.

Theorem 10. *Consider maximizing the dual objectives $\bar{R}_\phi$ and $\bar{R}$ over $\mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \times \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$.*

1) *Any maximizer $(\mathbb{P}^*_0, \mathbb{P}^*_1)$ of $\bar{R}_\phi$ over $\mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \times \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$ must maximize $\bar{R}$ as well.*

2) *If the adversarial Bayes classifier is unique up to degeneracy, then there are maximizers $(\mathbb{P}^*_0, \mathbb{P}^*_1)$ of $\bar{R}_\phi$ where $\mathbb{P}^*(\eta^* = 1/2) = 0$, with $\mathbb{P}^* = \mathbb{P}^*_0 + \mathbb{P}^*_1$ and $\eta^* = d\mathbb{P}^*_1/d\mathbb{P}^*$.*

See Appendix D for a proof of Item 1), Item 2) is shown in Theorems 5 and 7 of (Frank, 2025).

3 Main Results

Prior work has characterized when a loss ϕ is adversarially consistent with respect to a distribution $\mathbb{P}_0$, $\mathbb{P}_1$: Theorem 5 proves that a distribution independent surrogate risk bound is only possible when $C^*_\phi(1/2) < \phi(0)$ while Theorem 6 suggests that a surrogate bound must depend on the marginal distribution of η^* under $\mathbb{P}^*$, and furthermore, such a bound is only possible when $\mathbb{P}^*(\eta^* = 1/2) = 0$.

Compare these statements to Proposition 1: Theorems 5 and 6 imply that ϕ is adversarially consistent for $\mathbb{P}_0$, $\mathbb{P}_1$ if $C^*_\phi(1/2) < \phi(0)$ or if there exist some maximizers of $\bar{R}$ that satisfy Massart's noise condition. Alternatively, due to Theorem 10, one can equivalently assume that there are maximizers of $\bar{R}_\phi$ satisfying Massart's noise condition. Our first bound extends Proposition 1 to the adversarial scenario, with the data distribution $\mathbb{P}_0$, $\mathbb{P}_1$ replaced with the distribution of optimal adversarial attacks.

Theorem 11. *Let $\mathbb{P}_0$, $\mathbb{P}_1$ be a distribution for which there are maximizers $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ of the dual problem $\bar{R}_\phi$ that satisfy $|\eta^* - 1/2| \geq \alpha$ $\mathbb{P}^*$-a.e. for some constant $\alpha \in [0, 1/2]$ with $C^*_\phi(1/2 - \alpha) < \phi(0)$, where $\mathbb{P}^* = \mathbb{P}^*_0 + \mathbb{P}^*_1$, $\eta^* = d\mathbb{P}^*_1/d\mathbb{P}^*$. Then*

$$R^\epsilon(f) - R^\epsilon_* \leq \frac{3 + \sqrt{5}}{2} \frac{1}{\phi(0) - C^*_\phi(1/2 - \alpha)} (R^\epsilon_\phi(f) - R^\epsilon_{\phi,*}).$$

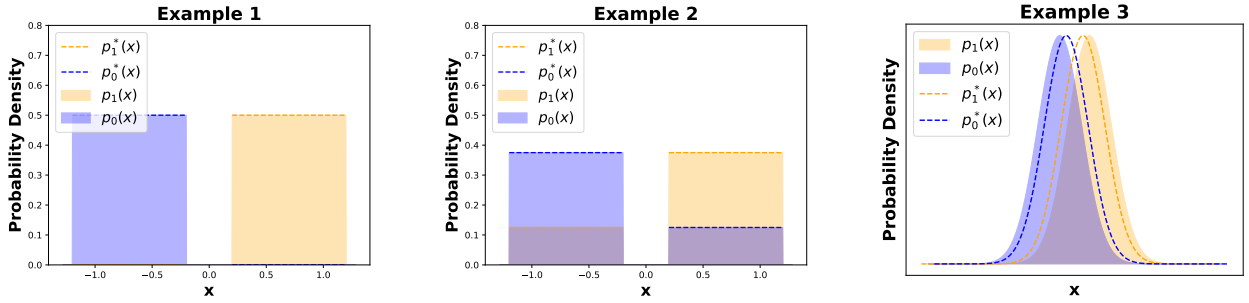
When $C^*_\phi(1/2) < \phi(0)$, one can select $\alpha = 0$ in Theorem 11 to produce a distribution-independent bound. The constant $(3 + \sqrt{5})/2$ may be sub-optimal; in fact Theorem 4 of Frank (2025) proves that $R^\epsilon(f) - R^\epsilon_* \leq 1/(2(\phi(0) - C^*_{\phi_\rho}(1/2)))(R^\epsilon_{\phi_\rho}(f) - R^\epsilon_{\phi_\rho,*})$ where $\phi_\rho(\alpha) = \min(1, \max(0, 1 - \alpha/\rho))$ is the ρ-margin loss. Furthermore, the bound in (10) extends directly to the adversarial setting.

Theorem 12. *Let ϕ be any loss with $\phi(0) > 0$ satisfying Assumption 1. Then if $R^\epsilon_* = 0$,*

$$R^\epsilon(f) - R^\epsilon_* \leq \frac{1}{\phi(0)} \left(R^\epsilon_\phi(f) - R^\epsilon_{\phi,*}\right)$$

A distribution will have zero adversarial risk whenever the supports of $\mathbb{P}_0$ and $\mathbb{P}_1$ are separated by at least 2ϵ, see Example 1 and Figure 2a for and example. Zero adversarial classification risk corresponds to $\alpha = 1/2$ in Massart's noise condition.

In contrast, Theorem 11 states that if some distribution of *optimal adversarial attacks* satisfies Massart's noise condition, then the excess adversarial surrogate risk is at worst a linear upper bound on the excess adversarial classification risk. However, if $C^*_\phi(1/2) = \phi(0)$, this constant approaches infinity as $\alpha \to 0$, reflecting the fact that adversarial consistency fails when the adversarial Bayes classifier is not unique up to degeneracy. When $\alpha \neq 1/2$, understanding what assumptions on $\mathbb{P}_0$, $\mathbb{P}_1$ guarantee Massart's noise condition for $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ is an open question. Example 4.6 of (Frank, 2024) demonstrates a distribution that satisfies Massart's noise condition and yet the adversarial Bayes classifier is not unique up to degeneracy. Thus Massart's noise condition for $\mathbb{P}_0, \mathbb{P}_1$ does not guarantee Massart's noise condition for $\mathbb{P}^*_0$, $\mathbb{P}^*_1$. See Example 2 and Figure 2b for an example where Theorem 11 applies with $\alpha > 0$.



(a) (b) (c)

Figure 2: Distributions from Examples 1 to 3 along with attacks that maximize the dual $\bar{R}_\phi$.

Finally, by averaging bounds of the form Theorem 11 over all values of η^* produces a distribution-dependent surrogate bound, valid whenever the adversarial Bayes classifier is unique up to degeneracy. For a given function f, let the *concave envelope* of f be the smallest concave function larger than f:

$$\operatorname{conc}(f) = \inf\{g :\geq f \text{ on } \operatorname{dom}(h), g \text{ concave and upper semi-continuous}\} \tag{18}$$

Theorem 13. *Assume that $C_\phi^*(1/2) = \phi(0)$ and that the adversarial Bayes classifier is unique up to degeneracy. Let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ be maximizers of $\bar{R}_\phi$ for which $\mathbb{P}^*(\eta^* = 1/2) = 0$, with $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Let $H(z) = \operatorname{conc}(\mathbb{P}^*(|\eta^* - 1/2| \leq z))$. Let Ψ be the function defined by Theorem 4 and let $\tilde{\Lambda}(z) = \Psi^{-1}(\min(\frac{z}{6}, \phi(0)))$. Then*

$$R^\epsilon(f) - R_*^\epsilon \leq \tilde{\Phi}(R_\phi^\epsilon(f) - R_{\phi,*}^\epsilon)$$

with

$$\tilde{\Phi}(z) = 6\left(\mathrm{id} + \min(1, \sqrt{-2eH\ln 2H})\right) \circ \tilde{\Lambda}$$

See Example 3 and Figure 2c for an example of calculating a distribution-dependent surrogate risk bound.

One can prove that the function H is always continuous and satisfies $H(0) = 0$, proving that this bound is non-vacuous (see Lemma 2 below). Further notice that $H \ln H$ approaches zero as $H \to 0$.

The map $\tilde{\Phi}$ combines two components: $\tilde{\Lambda}$, a modified version of Ψ^{-1}, and H, a modification of the cdf of $|\eta^* - 1/2|$. The function $\tilde{\Lambda}$ is a scaled version of Ψ^{-1}, where Ψ is the surrogate risk bound in the non-adversarial case of Theorem 4. The domain of Ψ^{-1} is $[0, \phi(0)]$, and thus the role of the min in the definition of $\tilde{\Lambda}$ is to truncate the argument so that it fits into this domain. The factor of $1/6$ in this function appears to be an artifact of our proof, see Section 5 for further discussion. In contrast, the map H translates the distribution of η^* into a surrogate risk transformation. Compare with Theorem 6, which states that consistency fails if $\mathbb{P}^*(\eta^* = 1/2) > 0$; accordingly, the function H becomes a poorer bound when more mass of η^* is near $1/2$.

Examples

Below are three examples for which each of our three main theorems apply. These examples are all one-dimensional distributions, and we denote the pdfs of $\mathbb{P}_0$, and $\mathbb{P}_1$ by p_0 and p_1.

To start, a distribution for which the supports of $\mathbb{P}_0$, $\mathbb{P}_1$ are more than 2ϵ apart must have zero risk. Furthermore, if $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure and the supports of $\mathbb{P}_0$, $\mathbb{P}_1$ are exactly 2ϵ apart, then the adversarial classification risk will be zero (see for instance (Awasthi et al., 2023a, Lemma 4) or (Pydi & Jog, 2021, Lemma 4.3)).

Example 1 (When $R_*^\epsilon = 0$)**.** Let p_0 and p_1 be defined by

$$p_0(x) = \begin{cases} 1 & \text{if } x \in [-1-\delta, -\delta] \\ 0 & \text{otherwise} \end{cases} \qquad p_1(x) = \begin{cases} 1 & \text{if } x \in [\delta, 1+\delta] \\ 0 & \text{otherwise} \end{cases}$$

for some $\delta > 0$. See Figure 2a for a depiction of p_0 and p_1. This distribution satisfies $R^\epsilon_* = 0$ for all $\epsilon \leq \delta$ and thus the surrogate bound of Theorem 12 applies.

Examples 2 and 3 require computing maximizers to the dual $\bar{R}_\phi$. See Appendices J.1 and J.2 for these calculations. The following example illustrates a distribution for which Massart's noise condition can be verified for a distribution of optimal attacks.

Example 2 (Massart's noise condition)**.** Let $\delta > 0$ and let p be the uniform density on $[-1-\delta, -\delta] \cup [\delta, 1+\delta]$. Define η by

$$\eta(x) = \begin{cases} \frac{1}{4} & \text{if } x \in [-1-\delta, -\delta] \\ \frac{3}{4} & \text{if } x \in [\delta, 1+\delta] \end{cases} \tag{19}$$

see Figure 2b for a depiction of p_0 and p_1. For this distribution and $\epsilon \leq \delta$, the minimal surrogate and adversarial surrogate risks are always equal ($R_{\phi,*} = R^\epsilon_{\phi,*}$). This fact together with Theorem 9 imply that optimal attacks on this distribution are $\mathbb{P}^*_1 = \mathbb{P}_1$ and $\mathbb{P}^*_0 = \mathbb{P}_0$, see Appendix J.1 for details. Consequently: the distribution of optimal attacks $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ satisfies Massart's noise condition with $\alpha = 1/4$ and as a result the bounds of Theorem 11 apply.

Finally, the next example presents a case in which Massart's noise condition fails for the distribution of optimal adversarial attacks, yet the adversarial Bayes classifier remains unique up to degeneracy. Theorem 13 continues to yield a valid surrogate bound.

Example 3 (Gaussian example)**.** Consider an equal Gaussian mixture with equal variances and differing means, with $\mu_1 > \mu_0$:

$$p_0(x) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}, \quad p_1(x) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$$

We further assume $\mu_1 - \mu_0 \leq \sqrt{2}\sigma$; see Figure 2c for a depiction. We will show that when $\mu_1 - \mu_0 < 2\epsilon$, the optimal attacks $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ are gaussians centered at $\mu_0 + \epsilon$ and $\mu_1 - \epsilon$— explicitly the pdfs of $\mathbb{P}^*_0$ and $\mathbb{P}^*_1$ are given by

$$p^*_0(x) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-(\mu_0+\epsilon))^2}{2\sigma^2}}, \quad p^*_1(x) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-(\mu_1-\epsilon))^2}{2\sigma^2}}, \tag{20}$$

see Appendix J.2 for details. We verify that $\mathbb{P}^*_0$ and $\mathbb{P}^*_1$ are in fact optimal by finding a function f^* for which $R^\epsilon_\phi(f) = \bar{R}_\phi(\mathbb{P}^*_0, \mathbb{P}^*_1)$, the strong duality result in Theorem 9 will then imply that $\mathbb{P}^*_0$ and $\mathbb{P}^*_1$ must maximize the dual $\bar{R}_\phi$, see Appendix J.2 for details.

However, when $\mu_1 - \mu_0 \leq \sqrt{2}\sigma$, then the function $h(z) = \mathbb{P}^*(|\eta^* - 1/2| \leq z)$ is concave in z for all $\epsilon < (\mu_1 - \mu_0)/2$ and consequently $h = H$, see Appendix J.3 for details. Unfortunately, h is a rather unweildy function. By comparing to the linear approximation at zero, one can show the following upper bound on H:

$$H(z) \leq \min\left(\frac{16\sigma^2}{\mu_1 - \mu_0 - 2\epsilon} z, 1\right). \tag{21}$$

Again, see Appendix J.3 for details.

When $\epsilon \geq (\mu_1 - \mu_0)/2$, (Frank, 2024, Example 4.1) shows that the adversarial Bayes classifier is not unique up to degeneracy. Notably, the bound in preceding example deteriorates as $(\mu_1 - \mu_0)/2 \to \epsilon$, and then fails entirely when $\epsilon = (\mu_1 - \mu_0)/2$.

4 Linear Surrogate Bounds— Proof of Theorems 11 and 12

The proof of Theorem 12 simply involves bounding the indicator functions $S_\epsilon(\mathbf{1}_{f>0})$, $S_\epsilon(\mathbf{1}_{f\leq 0})$ in terms of the functions $S_\epsilon(\phi \circ f)$ and $S_\epsilon(\phi \circ -f)$. This strategy is entirely analogous to that the argument for the (non-adversarial) surrogate bound (10) in Appendix B.1. A similar argument is also an essential intermediate step of the proof of Theorem 11.

Proof of Theorem 12. If $R^\epsilon_* = 0$, then the duality result Theorem 7 implies that for any measures $\mathbb{P}'_0 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_0)$, $\mathbb{P}'_1 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$ then $\mathbb{P}'(\eta' = 0 \text{ or } 1) = 1$, where $\mathbb{P}' = \mathbb{P}'_0 + \mathbb{P}'_1$ and $\eta' = d\mathbb{P}'_1/d\mathbb{P}'$. Consequently, $\bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) = 0$ for any $(\mathbb{P}'_0, \mathbb{P}'_1) \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \times \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$ and consequently $R^\epsilon_{\phi,*} = 0$. Thus it remains to show that $R^\epsilon(f) \leq \frac{1}{\phi(0)} R^\epsilon_\phi(f)$ for all functions f. We will prove the bound

$$S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) \leq \frac{1}{\phi(0)} S_\epsilon(\phi(f))(\mathbf{x}). \tag{22}$$

The inequality (22) trivially holds when $S_\epsilon(\mathbf{1}_{f\leq 0}) = 0$. Alternatively, the relation $S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) = 1$ implies $f(\mathbf{x}') \leq 0$ for some $\mathbf{x}' \in \overline{B_\epsilon(\mathbf{x})}$ and consequently $S_\epsilon(\phi(f))(\mathbf{x}) \geq \phi(0)$. Thus whenever $S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) = 1$,

$$S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) = \frac{\phi(0)}{\phi(0)} \leq \frac{1}{\phi(0)} S_\epsilon(\phi(f))(\mathbf{x}). \tag{23}$$

An analogous argument implies that whenever $S_\epsilon(\mathbf{1}_{f>0})(\mathbf{x}) = 1$,

$$S_\epsilon(\mathbf{1}_{f>0})(\mathbf{x}) = S_\epsilon(\mathbf{1}_{-f<0})(\mathbf{x}) \leq \frac{1}{\phi(0)} S_\epsilon(\phi(-f))(\mathbf{x}).$$

As a result:

$$\begin{aligned} R^\epsilon(f) &= \int S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_{f>0})(\mathbf{x}) d\mathbb{P}_0 \leq \frac{1}{\phi(0)} \left(\int S_\epsilon(\phi(f))(\mathbf{x}) d\mathbb{P}_1 + \int S_\epsilon(\phi(-f)(\mathbf{x}) d\mathbb{P}_0 \right) \\ &= \frac{1}{\phi(0)} R^\epsilon_\phi(f) \end{aligned}$$

□

In contrast, when the optimal surrogate risk $R^\epsilon_{\phi,*}$ is non-zero, the bound in Theorem 11 necessitates a more sophisticated argument. Below, we decompose both the adversarial classification risk and the adversarial surrogate risk as the sum of four terms positive terms.

Let $\mathbb{P}^*_0, \mathbb{P}^*_1$ be any maximizers of $\bar{R}_\phi$ that also maximize $\bar{R}$ by Theorem 10. Set $\mathbb{P}^* = \mathbb{P}^*_0 + \mathbb{P}^*_1$, $\eta^* = d\mathbb{P}^*_1/d\mathbb{P}^*$. Let γ^*_0, γ^*_1 be the couplings between $\mathbb{P}_0$, $\mathbb{P}^*_0$ and $\mathbb{P}_1$, $\mathbb{P}^*_1$ respectively that achieve the infimum in (15). Then due to the strong duality in Theorem 7, one can decompose the excess classification risk as

$$R^\epsilon(f) - R^\epsilon_* = R^\epsilon(f) - \bar{R}(\mathbb{P}^*_0, \mathbb{P}^*_1) = I_1(f) + I_0(f) \tag{24}$$

with

$$\begin{aligned} I_1(f) &= \left(\int S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f\leq 0}(\mathbf{x}') d\gamma^*_1 \right) + \left(\int C(\eta^*, f) - C^*(\eta^*) d\mathbb{P}^*_1 \right) \\ I_0(f) &= \left(\int S_\epsilon(\mathbf{1}_{f>0})(\mathbf{x}) - \mathbf{1}_{f>0}(\mathbf{x}') d\gamma^*_0 \right) + \left(\int C(\eta^*, f) - C^*(\eta^*) d\mathbb{P}^*_0 \right) \end{aligned}$$

Lemma 1 implies that $S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f\leq 0}(\mathbf{x}')$ must be positive, while the definition of C^* implies that $C(\eta^*, f) - C^*(\eta^*) \geq 0$.

Similarly, one can express the excess surrogate risk as

$$R^\epsilon_\phi(f) - R^\epsilon_{\phi,*} = I^\phi_1(f) + I^\phi_0(f) \tag{25}$$

with

$$\begin{aligned} I^\phi_1(f) &= \left(\int S_\epsilon(\phi(f))(\mathbf{x}) - \phi(f)(\mathbf{x}') d\gamma^*_1 \right) + \left(\int C_\phi(\eta^*, f) - C^*_\phi(\eta^*) d\mathbb{P}^*_1 \right) \\ I^\phi_0(f) &= \left(\int S_\epsilon(\phi(-f))(\mathbf{x}) - \phi(-f)(\mathbf{x}') d\gamma^*_0 \right) + \left(\int C_\phi(\eta^*, f) - C^*_\phi(\eta^*) d\mathbb{P}^*_0 \right) \end{aligned}$$

Define $K_\phi = \frac{3+\sqrt{5}}{2} \cdot \frac{1}{\phi(0) - C_\phi^*(1/2-\alpha)}$. We will argue that:

$$I_0(f) \leq K_\phi I_0^\phi(f). \tag{26}$$

$$I_1(f) \leq K_\phi I_1^\phi(f). \tag{27}$$

Below, we discuss the proof of (27) and an analogous argument will imply (26).

The proof proceeds by splitting the domain $\mathbb{R}^d \times \mathbb{R}^d$ into three different regions D_1, E_1 F_1 and proving the inequality in each case with a slightly different argument. These three cases will also appear in the proof of theorem Theorem 13. Define the sets D_1, E_1, F_1 by

$$D_1 = \{(\mathbf{x}, \mathbf{x}') : S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f(\mathbf{x}')\leq 0} = 0\} \tag{28}$$

$$E_1 = \{(\mathbf{x}, \mathbf{x}') : S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f(\mathbf{x}')\leq 0} = 1, f(\mathbf{x}') \geq \beta\} \tag{29}$$

$$F_1 = \{(\mathbf{x}, \mathbf{x}') : S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f(\mathbf{x}')\leq 0} = 1, f(\mathbf{x}') < \beta\} \tag{30}$$

where $\beta > 0$ is some constant, to be specified later (see Equations (32) and (33)). On the set D_1 the adversarial error matches the non-adversarial error with respect to the distribution $\mathbb{P}_0^*$, $\mathbb{P}_1^*$, and thus the bound in (12) implies a linear surrogate bound. On E_1, the same argument as (23) together with (12) proves a linear surrogate bound for adversarial risks. In short: this argument uses the first term in $I_1^\phi(f)$ to bound the first term in $I_1(f)$ and the second term of $I_1^\phi(f)$ to bound the second term of $I_1(f)$.

In contrast, The counterexample discussed in (14) demonstrates that when f is near 0, the quantity $S_\epsilon(\phi(f))(\mathbf{x}) - \phi(f)(\mathbf{x}')$ can be small even though $S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f\leq 0}(\mathbf{x}')$ can be large. Consequently, a different strategy is required to establish a linear surrogate bound on the set F_1. The key observation is that under the assumptions of Proposition 1, the function f must be bounded away from zero whenever it misclassifies a point. As a result, the excess conditional risk $C_\phi(\eta, f) - C_\phi^*(\eta)$ is bounded below by a positive constant and thus can be used to bound terms comprising $I_1^\phi(f)$. The constant β is then specifically chosen to balance the contribution of the risks over the sets E_1 and F_1.

Proof of Theorem 11. We will will prove (27), the argument for (26) is analogous. Due due to Equations (24) and (25), these inequalities prove the desired result. First, notice that (12) implies that

$$C(\eta^*(\mathbf{x}'), f(\mathbf{x}')) - C^*(\eta^*(\mathbf{x}')) \leq \frac{1}{\phi(0) - C_\phi^*(1/2-\alpha)} \left(C_\phi(\eta^*(\mathbf{x}'), f(\mathbf{x}')) - C_\phi^*(\eta^*(\mathbf{x}'))\right) \quad \mathbb{P}^*\text{-a.e.} \tag{31}$$

Choose the constant β to satisfy

$$\phi(\beta) = tC_\phi^*(1/2-\alpha) + (1-t)\phi(0) \tag{32}$$

with $t = (3-\sqrt{5})/2$. The parameter t is specifically selected to balance the contributions of the errors on E_1 and F_1, specifically it should satisfy

$$\frac{1}{t} = 1 + \frac{1}{1-t} = \frac{3+\sqrt{5}}{2} = K_\phi(\phi(0) - C_\phi^*(1/2-\alpha)) \tag{33}$$

Next, we prove the relation (27) on each of the sets D_1, E_1, F_1 separately.

1. **On the set D_1:**

   Lemma 1 implies that $S_\epsilon(\phi(f))(\mathbf{x}) - \phi(f(\mathbf{x}')) \geq 0$. This fact together with Equation 31 implies (27).

2. **On the set E_1:**

   If $S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f(\mathbf{x}')\leq 0} = 1$ but $f(\mathbf{x}') \geq \beta$, then $S_\epsilon(\phi \circ f)(\mathbf{x}) \geq \phi(0)$ while $\phi(f(\mathbf{x}')) \leq \phi(\beta)$ and thus $S_\epsilon(\phi \circ f)(\mathbf{x}) - \phi(f(\mathbf{x}')) \geq \phi(0) - \phi(\beta) = t(\phi(0) - C_\phi^*(1/2-\alpha)) = 1/K_\phi$ by (33). Thus

   $$S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f\leq 0}(\mathbf{x}') = 1 = \frac{K_\phi}{K_\phi} \leq K_\phi(S_\epsilon(\phi \circ f)(\mathbf{x}) - \phi(f(\mathbf{x}')) \tag{34}$$

   This relation together with (31) implies (27).

3. **On the set F_1:**

   First, $S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f(\mathbf{x}')\leq 0} = 1$ implies that $f(\mathbf{x}') > 0$. If additionally $f(\mathbf{x}') < \beta$, then both $f(\mathbf{x}') < \beta$ and $-f(\mathbf{x}') < \beta$ and consequently $C_\phi(\eta^*, f(\mathbf{x}')) \geq \phi(\beta)$. Furthermore, as C_ϕ^* is increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$ (see Lemma 5 in Appendix B.2), $\sup_{|\eta - 1/2|\geq\alpha} C_\phi^*(\eta) = C_\phi^*(1/2-\alpha)$. Thus due to the choice of β in (32):

$$C_\phi(\eta^*, f(\mathbf{x}')) - C_\phi^*(\eta^*) \geq \phi(\beta) - C_\phi^*(1/2-\alpha) = (1-t)(\phi(0) - C_\phi^*(1/2-\alpha)).$$

   The same argument as (34) then implies

$$S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) - \mathbf{1}_{f\leq 0}(\mathbf{x}') \leq \frac{1}{(1-t)(\phi(0) - C_\phi^*(1/2-\alpha))}(C_\phi^*(\eta^*, f(\mathbf{x}')) - C_\phi^*(\eta^*))$$

   This relation together with Equations (31) and (33) and Lemma 1 imply (27).

□

5 Proof of Theorem 13

Before proving Theorem 13, we will show that this bound is non-vacuous when the adversarial Bayes classifier is unique up to degeneracy. The function $h(z) = \mathbb{P}(|\eta^* - 1/2| \leq z)$ is a cdf, and is thus right-continuous in z. Furthermore, if the adversarial Bayes classifier is unique up to degeneracy, then $h(0) = 0$. The following lemma implies that if $H = \mathrm{conc}(h)$ then H is continuous at 0 and $H(0) = 0$. See Appendix E for a proof.

Lemma 2. *Let $h : [0, 1/2] \to \mathbb{R}$ be a non-decreasing function with $h(0) = 0$ and $h(1/2) = 1$ that is right-continuous at 0. Then $\mathrm{conc}(h)$ is non-decreasing, $\mathrm{conc}(h)(0) = 0$, and continuous on $[0, 1/2]$.*

The first step in proving Theorem 13 is showing an analog of Theorem 11 with $\alpha = 0$ for which the linear function is replaced by an η-dependent concave function.

Proposition 2. *Let Φ be a concave non-decreasing function for which $C(\eta, \alpha) - C^*(\eta) \leq \Phi(C_\phi(\eta, \alpha) - C_\phi^*(\eta))$ for any $\eta \in [0, 1]$ and $\alpha \in \overline{\mathbb{R}}$. Let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ be any two maximzers of $\bar{R}_\phi$ for which $\mathbb{P}^*(\eta^* = 1/2) = 0$ for $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Let $G : [0, \infty) \to \mathbb{R}$ be any non-decreasing concave function for which the quantity*

$$K = \int \frac{1}{G((\phi(0) - C_\phi^*(\eta^*))/2)} d\mathbb{P}^*$$

is finite. Then $R^\epsilon(f) - R_*^\epsilon \leq \tilde{\Phi}(R_\phi^\epsilon(f) - R_{\phi,*}^\epsilon)$, where

$$\tilde{\Phi}(z) = 6\sqrt{KG\left(\frac{1}{6}z\right)} + 2\Phi\left(\frac{1}{2}z\right) \tag{35}$$

The function Ψ^{-1} in Theorem 4 and the surrogate bounds of Zhang (2004) provide examples of candidate functions for Φ. As before, this result is proved by dividing the risks R_ϕ^ϵ, R^ϵ as the sum of four terms as in (24), (25) and then bounding these quantities over the sets D_1, E_1, and F_1 defined in (28),(29), and (30) separately. The key observation is that when f is bounded away from $\operatorname{argmin} C_\phi(\eta, \cdot)$, the excess conditional risk $C_\phi(\eta, f) - C_\phi^*(\eta)$ must be strictly positive. This quantity again serves to bound both components of $I_1^\phi(f)$, even if iti s not uniformly bounded away from zero. As before, the constant β is selected to balance the contributions of the risk on the sets E_1 and F_1. This time, the value β is function of $\eta^*(\mathbf{x}')$, where $\beta : [0, 1] \to \overline{\mathbb{R}}$ is a monotonic function for which

$$\phi(\beta(\eta)) = tC_\phi^*(\eta) + (1-t)\phi(0) = \frac{1}{2}\left(\phi(0) + C_\phi^*(\eta)\right) \tag{36}$$

with $t = 1/2$. In Appendix F, we show that there exists such a function β. An argument like the proof of Theorem 11 mixed with applications of the Cauchy-Schwartz and Jensen's inequality then proves Proposition 2, see Appendix G for details. Again, the function β is chosen to balance the contributions of the upper bounds on the risk on E_1 and F_1.

The factor of 1/6 in (35) arises as an artifact of the proof technique. Specifically, the constant $6 = 2 \cdot 3$ reflects two distinct components of the argument: the factor of 3 results from averaging over three sets D_1, E_1, F_1, (see (68) in Appendix G), the factor of 2 arises from combining the bounds associated with the two integrals $I_1(f)$ and $I_0(f)$ (see Equations (66) and (68) in Appendix G).

We now turn to the problem of identifying functions G for which the constant K in the preceding proposition is guaranteed to be finite. Observe that $\phi(0) - C_\phi^*(\eta^*) \geq \phi(0) - C_\phi^*(1/2)$ and so if $\phi(0) > C_\phi^*(1/2)$, the identity function is a possible choice for G. This option results in

$$\tilde{\Phi}(z) = \frac{2}{\phi(0) - C_\phi^*(1/2)} z + 2\Phi\left(\frac{1}{2}z\right),$$

which may improve the convergence rate relative to the bound in Theorem 11. The results developed here extend the classical analysis of Bartlett et al. (2006) to the adversarial setting. Moreover, Proposition 2 points to a pathway for generalizing the framework of Zhang (2004) to robust classification.

Alternatively, we consider constructing a function G for which the constant K in Proposition 2 is always finite when the adversarial Bayes classifier is unique, but distribution dependent. Observe that if h is the cdf of $|\eta - 1/2|$ and h is continuous, then $\int 1/h^r dh$ is always finite. This calculation suggests $\Phi = h \circ \Psi^{-1}$, with Ψ defined in Theorem 4. To ensure the concavity of G, we instead select $G = H \circ \Psi^{-1}$ with $H = \text{conc}(h)$.

Lemma 3. *Assume $C_\phi^*(1/2) = \phi(0)$. Let $\mathbb{P}_1$, $\mathbb{P}_0$, $\mathbb{P}_1^*$, $\mathbb{P}_0^*$, ϕ, H, and Ψ be as in Theorem 13. Let $\Lambda(z) = \Psi^{-1}(\min(z, \phi(0)))$. Then for any $r \in (0,1)$, Then*

$$R^\epsilon(f) - R_*^\epsilon \leq \tilde{\Phi}(R_\phi^\epsilon(f) - R_{\phi,*}^\epsilon) \tag{37}$$

with

$$\tilde{\Phi}(z) = 6\sqrt{\frac{1}{1-r} 2^r H\left(\Lambda\left(\frac{1}{6}z\right)\right)^r} + 2\Lambda\left(\frac{z}{2}\right).$$

Proof. For convenience, let $\bar{\Lambda}(z) = \frac{1}{2}\Lambda(2z)$. Let $G = (H \circ \bar{\Lambda})^r$, where $h(z) = \mathbb{P}^*(|\eta^* - 1/2| \leq z)$. Then G is concave because it is the composition of an concave function and an increasing concave function. We will verify that

$$K = \int \frac{1}{G((\phi(0) - C_\phi^*(\eta^*))/2)} d\mathbb{P}^* \leq \frac{2^r}{1-r}$$

First,

$$\int \frac{1}{G((\phi(0) - C_\phi^*(\eta^*))/2)} d\mathbb{P}^* = \int \frac{1}{H(|\eta^* - 1/2|)^r} d\mathbb{P}^* = \int_{[0,\frac{1}{2}]} \frac{1}{H(s)^r} d\mathbb{P}^* \sharp s = \int_{(0,\frac{1}{2}]} \frac{1}{H(s)^r} d\mathbb{P}^* \sharp s$$

with $s = |\eta^* - 1/2|$. The assumption $\mathbb{P}^*(|\eta^* = 1/2|) = 0$ allows us to drop 0 from the domain of integration. Because the function H is continuous on $(0,1]$ by Lemma 2, this last expression can actually be evaluated as a Riemann-Stieltjes integral with respect to the function $h(s) = \mathbb{P}(|\eta^* - 1/2| \leq s)$:

$$\int \frac{1}{H(s)^r} d\mathbb{P}^* \sharp s = \int \frac{1}{H(s)^r} dh \tag{38}$$

This result is standard when $\mathbb{P}^*$ is Lebesgue measure, (see for instance Theorem 5.46 of (Wheeden & Zygmund, 1977)). We prove equality in (38) for strictly decreasing functions in Proposition 4 in Appendix H.1.

Finally, the integral in (38) can be bounded as

$$\int \frac{1}{H(s)^r} dh \leq \frac{2^r}{1-r} \tag{39}$$

If h were differentiable, then the chain rule would imply

$$\int \frac{1}{H(s)^r} dh \leq \int \frac{1}{h(s)^r} dh = \int_0^1 \frac{1}{h(s)^r} h'(s) dz = \frac{1}{1-r} h(s)^{1-r} \Big|_0^1 = \frac{1}{1-r}.$$

This calculation is more delicate for non-differentiable H; we formally prove inequality in (39) in Appendix H.2.

This calculation proves the inequality (37) with $\tilde{\Phi}$ as

$$6\sqrt{\frac{\cdot 2^r}{1-r} H\left(\frac{1}{2}\Lambda\left(\frac{2}{6}z\right)\right)^r} + \Lambda(z)$$

The concavity of Λ together with the fact that $\Lambda(0) = 0$ then proves the result. □

Minimizing this bound over r then produces Theorem 13, see Appendix I for details.

6 Related Works

The most similar results to this paper are Li & Telgarsky (2023); Mao et al. (2023a). Li & Telgarsky (2023) prove a surrogate bound for convex losses, when one can minimize over the thresh-holding value in (13) rather than just 0. (Mao et al., 2023a) proves an adversarial surrogate bound for a modified ρ-margin loss.

Many papers study the statistical consistency of surrogate risks in the standard and adversarial context. Bartlett et al. (2006); Zhang (2004) prove surrogate risk bounds that apply to the class of all measurable functions Lin (2004); Steinwart (2007) prove further results on consistency in the standard setting, and Frongillo & Waggoner (2021) study the optimally of such surrogate risk bounds. (Bao, 2023) relies on the modulus of convexity of C_ϕ^* to construct surrogate risk bounds. Philip M. Long (2013); Mingyuan Zhang (2020); Awasthi et al. (2022); Mao et al. (2023a;b); Awasthi et al. (2023b) further study consistency restricted to a specific family of functions; a concepts called $\mathcal{H}$-*consistency*. Prior work Mahdavi et al. (2014)also uses these surrogate risk bounds in conjunction with surrogate generalization bounds to study the generalization of the classification error.

In the adversarial setting, (Meunier et al., 2022; Frank & Niles-Weed, 2024a) identify which losses are adversarially consistent for all data distributions while (Frank, 2025) shows that under reasonable distributional assumptions, a consistent loss is adversarially consistent for a specific distribution iff the adversarial Bayes classifier is unique up to degeneracy. (Awasthi et al., 2021) study adversarial consistency for a well-motivated class of linear functions while some prior work also studies the approximation error caused by learning from a restricted function class $\mathcal{H}$. Liu et al. (2024) study the approximation error of the surrogate risk. Complimenting this result, Awasthi et al. (2023b); Mao et al. (2023a) study $\mathcal{H}$-consistency in the adversarial setting for specific surrogate risks. Standard and adversarial surrogate risk bounds are a tool central tool in the derivation of the $\mathcal{H}$-consistency bounds in this line of research. Whether the adversarial surrogate bounds presented in this paper could result in improved adversarial $\mathcal{H}$-consistency bounds remains an open problem.

Our proofs rely on prior results that study adversarial risks and adversarial Bayes classifiers. Notably, (Bungert et al., 2021; Pydi & Jog, 2021; 2020; Bhagoji et al., 2019; Awasthi et al., 2023a) establish the existence of the adversarial Bayes classifier while (Frank & Niles-Weed, 2024b; Pydi & Jog, 2020; 2021; Bhagoji et al., 2019; Frank, 2025) prove various minimax theorems for the adversarial surrogate and classification risks. Subsequently, (Pydi & Jog, 2020) uses a minimax theorem to study the adversarial Bayes classifier, and (Frank, 2024) uses minimax results to study the notion of uniqueness up to degeneracy.

7 Conclusion

In conclusion, we prove surrogate risk bounds for adversarial risks whenever ϕ is adversarially consistent for the distribution $\mathbb{P}_0$, $\mathbb{P}_1$. When ϕ is adversarially consistent or the distribution of optimal adversarial attacks

satisfies Massart's noise condition, we prove a linear surrogate risk bound. For the general case, we prove a concave distribution-dependent bound. Understanding the optimality of these bounds remains an open problem, as does understanding how these bounds interact with the sample complexity of estimating the surrogate quantity. These questions were partly addressed by (Frongillo & Waggoner, 2021) and (Mahdavi et al., 2014) in the standard setting, but remain unstudied in the adversarial scenario.

References

Luigi Ambrosio, Nicola Fosco, and Diego Pallara. *Functions of Bounded Variation and Free Discontuity Problems*. Oxford Mathematics Monographs. Oxford University Press, 2000.

Tom M. Apostol. Mathematical analysis, 1974.

Pranjal Awasthi, Natalie S. Frank, Anqui Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. *NeurIps*, 2021.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022.

Pranjal Awasthi, Natalie S. Frank, and Mehryar Mohri. On the existence of the adversarial bayes classifier (extended version). *arxiv*, 2023a.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for adversarial robustness. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2023b.

Han Bao. Proper losses, moduli of convexity, and surrogate regret bounds. In *Proceedings of Thirty Sixth Conference on Learning Theory*, Proceedings of Machine Learning Research. PMLR, 2023.

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

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport, 2019.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennslvania, 2005.

Leon Bungert, Nicolás García Trillos, and Ryan Murray. The geometry of adversarial training in binary classification. *arxiv*, 2021.

Natalie S. Frank. A notion of uniqueness for the adversarial bayes classifier, 2024.

Natalie S. Frank. Adversarial consistency and the uniqueness of the adversarial bayes classifier. *European Journal of Applied Mathematics*, 2025.

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

Natalie S. Frank and Jonathan Niles-Weed. Existence and minimax theorems for adversarial surrogate risks in binary classification. *Journal of Machine Learning Research*, 2024b.

Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems*, 2021.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. Fundamentals of convex analysis, 2001.

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

Justin D. Li and Matus Telgarsky. On achieving optimal adversarial test error, 2023.

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

Changyu Liu, Yuling Jiao, Junhui Wang, and Jian Huang. Nonasymptotic bounds for adversarial excess risk under misspecified models. *SIAM Journal on Mathematics of Data Science*, 6(4), 2024. URL `https://doi.org/10.1137/23M1598210`.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Binary excess risk for smooth convex surrogates, 2014.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications, 2023a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Structured prediction with stronger consistency guarantees. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023b.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34, 2006.

Laurent Meunier, Raphaël Ettedgui, Rafael Pinot, Yann Chevaleyre, and Jamal Atif. Towards consistency in adversarial classification. *arXiv*, 2022.

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

Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: Adversarial examples for medical imaging. *Springer*, 2018.

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

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. *ICML*, 2020.

Muni Sreenivas Pydi and Varun Jog. The many faces of adversarial risk. *Neural Information Processing Systems*, 2021.

Adrian Raftery and Raftery Gneiting. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistcal Association*, 2007.

Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009. Association for Computing Machinery.

Walter Rudin. *Principles of Mathematical Analysis*. Mathematics Series. McGraw-Hill International Editions, third edition, 1976.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336), 1971.

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 1989.

Elias Stein and Rami Shakarchi. *Real analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press, 2005.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Richard L. Wheeden and Antoni Zygmund. *Measure and Integral*. Pure and Applied Mathematics. Marcel Dekker Inc., 1977.

Ying Xu, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. *Adversarial Attacks on Face Recognition Systems*, pp. 139–161. Springer International Publishing, Cham, 2022.

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