A Theory of Adversarial Risks for Binary Classification

by

Natalie Frank

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

May, 2024

Dr. Jonathan Niles-Weed

Acknowledgments

I am grateful to the many people who helped me through Courant's doctoral program. Thank you to my advisor Jonathan Niles-Weed for overseeing this project with patience, flexibility, and excellent advice. I learned a lot about fascinating mathematics and the research process along the way. I am grateful the supportive and collaborative environment at Courant that has expanded my horizons. Thank you to my friends Tanya and Rivki for their encouragement, optimism, and heartfelt friendship. Lastly, I owe my parents a deep thanks for their unwavering support. Their perspectives have taught me open-mindedness and resilience.

I'd like to thank the NSF for supporting this research. The Research Training Group in Modeling and Simulation funded by the National Science Foundation via grant RTG/DMS – 1646339 partly supported this research. This research was also supported by and NSF grants DMS-2210583,CCF-1535987, and IIS-1618662.

Abstract

Experiments demonstrate that in a variety of machine learning models, small but carefully chosen perturbations to data at test time can significantly increase the classification error. As a result, robustness to adversarial attacks is an increasingly important criterion in security-critical applications. To improve robustness, one would hope to minimize the classification error under an attack, known as the *adversarial classification risk*. The literature proposes a plethora of tools for improving the robustness of machine learning models, but many of these methods are poorly understood. One of the most popular defenses is *adversarial training*, in which one aims to minimize an *adversarial surrogate risk* that computes the worst-case loss over some allowed set of perturbations. The theory of risks in the non-adversarial setting is well understood, and includes results such as formulas for minimizers and a characterization of their statistical consistency. We extend some of these results to the adversarial setting. Lastly our results provide an explanation for the phenomenon of transfer attacks.

Contents

List of Figures

List of Appendices

1 — Introduction

Neural nets are state-of-the-art models for a variety of classification tasks. However, these models exhibit a concerning phenomenon— imperceptible perturbations to data at test-time can derail their accuracy [14, 58]. Such attacks are a concern in security sensitive applications such as medical imaging [47], facial recognition [72], and identifying traffic signs in self-driving cars [37]. The error rate of a classifier under an adversarial attack is referred to as the *adversarial classification risk*. One of the most popular defense algorithms is *adversarial training* which performs gradient descent on an *adversarial surrogate risk* that averages the value of some loss function over the worst possible attack at each point. The theory of such risks in the non-adversarial setting is well understood [8, 38, 57], and this thesis extends some of these results to the adversarial setting.

A better understanding of the theoretical underpinnings of adversarial learning could motivate new directions in algorithm development and explain empirical observations. Specifically, the results proved in this thesis explain the empirical phenomenon of transfer attacks, describe the structure of minimizers to adversarial risks, and characterize the statistical consistency of adversarial surrogate risks. An overview of each of these results is provided in Sections 1.2, 1.3, and 1.4.

Chapter 2 of this thesis was published in JMLR [25] while chapter 4 was published in NeurIps [26].

1.1 Notation and Background

This section describes prior work on risks in the standard classification setting. We consider the problem of binary classification on $\mathbb{R}^d$ with labels $\{-1, +1\}$. The measures $\mathbb{P}_0$, $\mathbb{P}_1$, respectively, describe the probability of data with labels -1, $+1$ occurring in a region of $\mathbb{R}^d$. Data with label -1 is distributed according to the finite measure $\mathbb{P}_0$ and data with label $+1$ is distributed according to the finite measure $\mathbb{P}_0$. The *classification risk* of a set A is then the proportion of errors if the set A is labeled $+1$ and the set A^C is labeled -1:

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

Re-writing this quantity in terms of $\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$ and the conditional probability of label $+1$, $\eta = d\mathbb{P}_1/d\mathbb{P}$, assists in finding the infimum of this risk:

$$R(A) = \int \eta \mathbf{1}_{A^C} + (1-\eta)\mathbf{1}_A d\mathbb{P} = \int C(\eta(\mathbf{x}), \mathbf{1}_A(\mathbf{x})) d\mathbb{P}(\mathbf{x})$$

with the conditional risk $C : [0,1] \times \{0,1\} \to [0,1]$ as

$$C(\eta, b) = (1-\eta)b + \eta(1-b).$$

This function represents the classification error when the conditional probability of class $+1$ is the constant η. Consequently, minimizing R is equivalent to minimizing $C(\eta(\mathbf{x}), \cdot)$ pointwise. As a result, the sets

$$\{\eta(\mathbf{x}) > 1/2\} \quad \text{and} \quad \{\eta(\mathbf{x}) \geq 1/2\} \tag{1.1}$$

are both Bayes classifiers. The Bayes classifier is *unique* if $\mathbb{P}(\eta = 1/2) = 0$, or alter-

natively, amongst all Bayes classifiers, either the value of $\mathbb{P}_0(A)$ is unique or the value of $\mathbb{P}_1(A^C)$ is unique.

However, minimizing the empirical classification risk is computationally difficult and consequently machine learning algorithms typically minimize a different quantity called a *surrogate risk*. We consider the margin-based surrogate,

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

The function f is then threshholded at 0 to obtain a classifier; we define the classification error of f as $R(f) = R(\{f > 0\})$. The *loss function* ϕ is non-increasing, so that the quantity $\phi(yf(\mathbf{x}))$ can be interpreted as a confidence level, with a high value of $yf(\mathbf{x})$ implying high confidence.

Again, one can compute minimizers to R_ϕ by writing this risk in terms of the quantities $\mathbb{P}$ and η:

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

with the conditional risk as the function

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

Again, the conditional risk is the surrogate risk when the conditional probability of class $+1$ is the constant η. Thus minimizing the integrand $C_\phi(\eta(\mathbf{x}), \cdot)$ pointwise will produce a minimizer of R_ϕ. However, minimizers may not exist on $\mathbb{R}$: consider for instance a distribution for which $\eta(\mathbf{x}) \equiv 1$ and $\phi = e^{-\alpha}$ is the exponential loss, so that $C_\phi(\eta(\mathbf{x}), \alpha) = e^{-\alpha}$. However, minimizers will exist over the extended real numbers $\overline{\mathbb{R}}$:

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

Consequently, the function

$$\alpha_\phi(\eta(\mathbf{x})) \tag{1.2}$$

is a minimizer of R_ϕ. The minimal value of $R_\phi(f)$ is then $\int C_\phi^*(\eta)d\mathbb{P}$ with

$$C_\phi^*(\eta) = \inf_\alpha C_\phi(\eta, \alpha). \tag{1.3}$$

However, minimizing the surrogate R_ϕ may not minimize the classification risk R. If every minimizing sequence of R_ϕ is also a minimizing sequence of R, then the loss ϕ is *consistent*. The consistency of surrogate risks is a well studied problem. In particular, [8] show

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

This thesis extends these results to adversarial risks. In the adversarial setting, a point is misclassified if a malicious adversary can perturb the point into the opposite class. The adversary's possible attacks are modeled by an ϵ ball in some norm $\|\cdot\|$. Thus, a point $\mathbf{x} \in A$ is missclassified when there is some $\mathbf{h} \in \overline{B_\epsilon(\mathbf{0})}$ for which $\mathbf{x} + \mathbf{h} \in A^C$, or in other words, $\sup_{\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon} \mathbf{1}_A(\mathbf{x}') = 1$. The operation of computing a supremum over a closed ϵ-ball is denoted by

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

Thus, a point $\mathbf{x}$ in A is misclassified iff $S_\epsilon(\mathbf{1}_A)(\mathbf{x}) = 1$ while a point in A^C is misclassified iff $S_\epsilon(\mathbf{1}_{A^C})(\mathbf{x}) = 1$. The *adversarial classification risk* is the proportion of errors when the set A is labeled $+1$ and the set A^C is labeled -1:

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

Just as in the setting of standard learning, minimizing an empirical version of the adversarial classification risk is computationally intractable. Instead, one typically minimizes the

surrogate

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

Notice that in order to define these adversarial risks, one must show that $S_\epsilon(g)$ is measurable when g is measurable. This topic is addressed in Chapter 2. A loss is *adversarially consistent* if every minimizing sequence of R_ϕ^ϵ is also a minimizing sequence of R^ϵ.

This thesis will study the form of minimizers to the adversarial risks R^ϵ, R_ϕ^ϵ and analyze the adversarial consistency of surrogate losses.

1.2 Transfer Attacks

Prior experimental work shows that adversarial examples tend to transfer between deep networks trained for the same task— in other words, if both f_1 and f_2 are trained for the same classification task, then an adversarial example that fools f_1 will frequently fool f_2. Such attacks are referred to as *transfer attacks*, and they provide a method for attacking a machine learning model without access to the model parameters. These methods are referred to as *black box attacks*, in contrast to a *white box attacks* in which the adversary has full access to the model. Transfer attacks have a lower success rate than white box attacks. For instance, [18] train a neural net on handwritten 8 and 9 digits. In their experiments on neural nets with adversarial perturbations of size at most 1 in the ℓ_2 norm, transfer attacks have a success rate of $10\% - 20\%$ while their white box attack succeeds $20\% - 30\%$ of the time (see Figure 7). Furthermore, they show that the phenomenon of transfer attacks extends to other models in addition to neural nets such as random forests, logistic regression, and kernel SVMs.

Our results provide an explanation of transfer attacks in terms of complimentary slackness conditions. The minimax theorem above models an attack as a measure of distance at most ϵ from the original data distribution in the *Wasserstein ∞-metric.* Informally, a measure is

within ϵ of $\mathbb{Q}$ if one can obtain the measure $\mathbb{Q}'$ by moving each point in $\mathbb{R}^d$ by at most ϵ under the measure $\mathbb{Q}$. Consequently, the Wasserstein-∞ metric is well-suited for modeling a norm-bounded adversary. This metric, also denoted W_∞, is formally defined in Chapters 2, 3, 4, and 5. Let $\mathcal{B}_\epsilon^\infty(\mathbb{Q}) = \{\mathbb{Q}' : W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon\}$ denote the ∞-Wasserstein ball of measures around $\mathbb{Q}$. Chapter 2 relates the risk R_ϕ^ϵ to a dual quantity.

Theorem. *Let $\mathbb{P}_0$, $\mathbb{P}_1$ be finite Borel measures and let C_ϕ^* be the function defined by (1.3). Define*

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

Then

$$\inf_{\substack{f \text{ Borel} \\ \bar{\mathbb{R}}\text{-valued}}} R_\phi^\epsilon(f) = \sup_{\substack{\mathbb{P}'_0 \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}'_1 \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) \tag{1.4}$$

Furthermore, both the infimum is attained by a function f^* and the supremum is attained by measures $\mathbb{P}_0^*$, $\mathbb{P}_1^*$.

Earlier work [52] proved a similar minimax theorem for R^ϵ that replaced C_ϕ^* in the definition of $\bar{R}_\phi$ with C^*. This theorem directly leads to complimentary slackness conditions that characterize the minimizers of R_ϕ^ϵ and the maximizers of $\bar{R}_\phi$.

Theorem. *The function f^* is a minimizer of R_ϕ^ϵ and $\mathbb{P}_0^*, \mathbb{P}_1^*$ maximize $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$ and $\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ iff:*

1)

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

2) Let $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Then

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

Item 1) states that the measures $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ must be optimal adversarial attacks against f^* while item 2) states that f^* must minimize the conditional risk $C_\phi^*(\eta^*(\mathbf{x}), \cdot)$ of optimal adversarial attacks $\mathbb{P}^*$-a.e. These conditions apply to *any* minimizer of R_ϕ^ϵ and *any* maximizers of $\bar{R}_\phi$. Thus an optimal adversarial attack must preform equally well against any two minimizers of R_ϕ^ϵ! The fact that transfer attacks have a significantly lower success rate that white box attacks suggests that either the state-of-the-art attacks or state-of-the-art defenses are far from optimal.

Minimizers of risks in machine learning are typically found via some optimization algorithm, and these procedures can only achieve approximate optimally. Do approximate minimizers and maximizers exhibit the effect of transfer attacks? Below, we answer this question in the affirmative: specifically, if f is an approximate minimizer of R_ϕ^ϵ and $\mathbb{P}'_0$,$\mathbb{P}'_1$ are an approximate maximizer of the dual problem $\bar{R}_\phi$, then $\int \phi \circ f d\mathbb{P}_1 + \int \phi \circ -f d\mathbb{P}_0 \approx R_{\phi,*}^\epsilon$, where $R_{\phi,*}^\epsilon$ is the optimal value of the optimization problem. Assume that $R_\phi^\epsilon(f) \leq R_{\phi,*}^\epsilon + \delta$ and $\bar{R}(\mathbb{P}'_0, \mathbb{P}'_1) \geq R_{\phi,*}^\epsilon - \delta$ for some $\delta > 0$. Then moving each point under $\mathbb{Q}$ by the worst-case amount in an ϵ-ball will results in the function $S_\epsilon(g)$, and consequently $\int S_\epsilon(g) d\mathbb{Q} \geq \int S_\epsilon(g) d\mathbb{Q}'$ for any $\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})$ (see Chapter 2 for a formal statement and proof). Similarly, the definition of the function C_ϕ^* in (1.2) implies that $C_\phi^*(\eta^*) \leq C_\phi(\eta^*, f)$ for any function f. Therefore,

$$\begin{aligned} R_{\phi,*}^\epsilon + \delta \geq R_\phi^\epsilon(f) &\geq \int \phi \circ f d\mathbb{P}'_1 + \int \phi \circ -f d\mathbb{P}'_0 \\ &= \int C_\phi(\eta', f) d\mathbb{P}' \geq \int C_\phi^*(\eta') d\mathbb{P}' = \bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) \geq R_{\phi,*}^\epsilon - \delta \end{aligned}$$

where $\mathbb{P}' = \mathbb{P}'_0 + \mathbb{P}'_1$ and $\eta' = d\mathbb{P}'_1 / d\mathbb{P}'$. Subtracting $R_{\phi,*}^\epsilon$ from both sides of this inequality shows that

$$\left| \left(\int \phi \circ f d\mathbb{P}'_1 + \int \phi \circ -f d\mathbb{P}'_0 \right) - R_{\phi,*}^\epsilon \right| \leq \delta$$

Therefore, as any almost-optimal attack against any almost-optimal minimizer of R_ϕ^ϵ will

achieve almost the same risk, one would expect to find transfer attacks even for approximate optimizers found by machine learning models.

1.3 The Structure of Minimizers to R^ϵ and R^ϵ_ϕ

Chapters 2 and 5 extend Equations 1.1 and 1.2 to the adversarial setting. Specifically, Chapter 2 proves that there is a function $\hat{\eta} : \mathbb{R}^d \to [0,1]$ that reflects the conditional probability of class $+1$ under an optimal adversarial attack, and specifically connects this function to the quantity $d\mathbb{P}_1^*/d(\mathbb{P}_1^* + d\mathbb{P}_0^*)$ for some optimal 'attack measures' $\mathbb{P}_0^*$ and $\mathbb{P}_1^*$. One can show that minimizers to R^ϵ and R^ϵ_ϕ can be constructed just like those in Equations 1.1 and 1.2.

Theorem. *For every distribution $\mathbb{P}_0$, $\mathbb{P}_1$, there is a function $\hat{\eta} : \mathbb{R}^d \to [0,1]$ for which*

I) The function $\alpha_\phi(\hat{\eta}(\mathbf{x}))$ minimizes R^ϵ_ϕ for every ϕ, where α_ϕ is as defined in Section 1.1

II) The sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ minimize R^ϵ

The sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are 'minimal' and 'maximal' adversarial Bayes classifiers in the sense that

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

and

$$S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}}) \leq S_\epsilon(\mathbf{1}_A) \leq S_\epsilon(\mathbf{1}_{\{\hat{\eta} \geq 1/2\}}) \quad \mathbb{P}_0\text{-a.e.}$$

for any other adversarial Bayes classifier A. Chapter 3 also defines uniqueness for the adversarial Bayes classifier by constructing an equivalence relation on such sets. Two adversarial Bayes classifiers A_1 and A_2 are *equivalent up to degeneracy* if for any set A with

$A_1 \cap A_2 \subset A \subset A_1 \cup A_2$ is also an adversarial Bayes classifier. When $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, equivalence up to degeneracy defines an equivalence relation. The adversarial Bayes classifier is *unique up to degeneracy* if there is a single equivalence class. There are a few other useful characterizations of uniqueness up to degeneracy.

Informal Theorem. *Assume that $\mathbb{P}_0$ and $\mathbb{P}_1$ are absolutely continuous with respect to Lebesgue measure. Then the following are equivalent:*

A) The adversarial Bayes classifier is unique up to degeneracy

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

C) There are measures representing 'optimal adversarial attacks' $\mathbb{P}_0^*, \mathbb{P}_1^*$ for which $\mathbb{P}^*(\eta^* = 1/2) = 0$, where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$

Again, the measures of 'optimal adversarial attacks' $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ are defined using the Wasserstein-∞ distance. Item C) generalizes the criterion $\mathbb{P}(\eta = 1/2) = 0$ for Bayes classifiers. Similarly, Item B) generalizes the criterion that amongst all Bayes classifiers, either the value of $\mathbb{P}_0(A)$ is unique or the value of $\mathbb{P}_1(A^C)$ is unique.

Furthermore, Chapter 3 provides the tools for computing a representative of each equivalence class of adversarial Bayes classifiers under equivalence up to degeneracy. An example from Chapter 3 shows that uniqueness up to degeneracy can fail for all $\epsilon > 0$ even when the Bayes classifier is unique. However, the densities of this example distribution were discontinuous while the other examples in this section were better behaved.

Conjecture. *If the densities of $\mathbb{P}_0$ and $\mathbb{P}_1$ are sufficiently smooth with non-zero derivatives on the boundary of the Bayes classifier, then the adversarial Bayes classifier is unique up to degeneracy.*

Proving or refuting this conjecture remains an open problem.

1.4 The Consistency of Adversarial Surrogate Risks

Lastly, the results above provide the tools for analyzing the statistical consistency of adversarial surrogate risks. Prior work [42] provides an example which proves that no convex loss is adversarially consistent: Let $\mathbb{P}_0$, $\mathbb{P}_1$ be uniform distributions of equal mass the ball $\overline{B_{\epsilon/2}(\mathbf{0})}$: then every point in the support can be reached from every other point by a perturbation of size at most ϵ and thus if $\mathbf{1}_A$ is non-constant on $\overline{B_\epsilon(\mathbf{0})}$ then $R^\epsilon(A) = 1$. On the other hand, the constant classifiers $\mathbb{R}^d$, $\emptyset$ each achieve the risk $1/2$, and therefore must be optimal. Next, assume that our loss function satisfies $C^*_\phi(1/2) = \phi(0)$. This property is satisfied by ever convex ϕ because a convex function must satisfy $C_\phi(1/2, \alpha) = \frac{1}{2}\phi(\alpha) + \frac{1}{2}\phi(-\alpha) \geq \phi(0)$. Consider the constant function $f \equiv 0$. Then $R^\epsilon_\phi(f) = \phi(0)$, which also equals the minimum standard risk for this problem $R_{\phi,*}$. The optimal standard risk is always a lower bound on the optimal adversarial risk and as a result $f \equiv 0$ is a minimizer of R^ϵ_ϕ.

Consider the sequence of functions

$$f_n = \begin{cases} \frac{1}{n} & \text{if } \mathbf{x} = \mathbf{0} \\ -\frac{1}{n} & \text{otherwise} \end{cases}$$

Then $R^\epsilon_\phi(f_n) = \phi(-1/n)$ which approaches $\phi(0)$, while $R^\epsilon(f_n) = 1$. Thus f_n is a minimizing sequence of R^ϵ_ϕ that is not a minimizing sequence of R^ϵ (see Chapter 4 for a rigorous exposition of this example.)

The example above demonstrates that the obstacle to adversarial consistency for convex losses is the discontinuity of the indicator functions $\mathbf{1}_{\alpha \leq 0}$, $\mathbf{1}_{\alpha > 0}$ at zero.

We propose two methods for circumventing this difficulty: first, one can use a loss function

for which minimizers to $C_\phi(\eta, \cdot)$ are bounded away from zero for all η. Losses with $C_\phi^*(1/2) < \phi(0)$ satisfy this requirement.

Lemma. *Let ϕ be a loss with $C_\phi^*(1/2) < \phi(0)$. Then there exists some $\alpha_* > 0$ for which every minimizer α of $C_\phi(\eta, \cdot)$ must satisfy $|\alpha| \geq \alpha_*$.*

Losses satisfying this requirement are in fact adversarially consistent:

Theorem. *Any loss with $C_\phi^*(1/2) < \phi(0)$ is both consistent and adversarially consistent.*

However, if a loss is consistent, every minimizer of $C_\phi(\eta, \cdot)$ must satisfy $|\alpha| > 0$ so long as $\eta \neq 1/2$. Thus, another method of circumventing the discontinuity at zero is considering distributions for which the conditional probability of $1/2$ is measure zero, according to an appropriate measure. The appropriate measure in this case is the measure of optimal adversarial attacks $\mathbb{P}_0^*, \mathbb{P}_1^*$ discussed in the prior two sections. However, the condition $\mathbb{P}^*(\eta^* = 1/2) = 0$ for $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$, $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$ is equivalent to the uniqueness of the adversarial Bayes classifier under reasonable conditions. Consequently, assuming that the adversarial Bayes classifier is unique up to degeneracy will also avoid the discontinuity at zero.

Informal Theorem. *Let ϕ be a consistent loss with $C_\phi^*(1/2) = \phi(0)$. Then ϕ is consistent for the distribution $\mathbb{P}_0$, $\mathbb{P}_1$ iff the adversarial Bayes classifier is unique up to degeneracy.*

2 — A Minimax Theorem for Adversarial Surrogate Risks

2.1 Introduction

Neural networks are state-of-the-art methods for a variety of machine learning tasks including image classification and speech recognition. However, a concerning problem with these models is their susceptibility to *adversarial attacks*: small perturbations to inputs can cause incorrect classification by the network [14, 58]. This issue has security implications; for instance, Gu, Dolan-Gavitt, and Garg [31] show that a yellow sticker can cause a neural net to misclassify a stop sign. Furthermore, one can find adversarial examples that generalize to other neural nets; these sort of attacks are called *transfer attacks*. In other words, an adversarial example generated for one neural net will sometimes be an adversarial example for a different neural net trained for the same classification problem [18, 35, 46, 54, 60]. This phenomenon shows that access to a specific neural net is not necessary for generating adversarial examples. One method for defending against such adversarial perturbations is *adversarial training*, in which a neural net is trained on adversarially perturbed data points [35, 39, 67]. However, adversarial training is not well understood from a theoretical perspective.

From a theoretical standpoint, the most fundamental question is whether it is possible

to design models which are robust to such attacks, and what the properties of such robust models might be. In contrast to the classical, non-adversarial setting, much is still unknown about the basic properties of optimal robust models. In the context of binary classification, several prior works study properties of the *adversarial classification risk*—the expected number of classification errors under adversarial perturbations. Recently, Awasthi, Frank, and Mohri [2], Bungert, Trillos, and Murray [16], and Pydi and Jog [52] all showed existence of a minimizer to the adversarial classification risk under suitable assumptions, and characterized some of its properties. A crucial observation, emphasized by Pydi and Jog [52], is that minimizing the adversarial classification risk is equivalent to a *dual* robust classification problem involving uncertainty sets with respect to the ∞-Wasserstein metric. This observation gives rise to a game-theoretic interpretation of robustness, which takes the form of a zero-sum game between a classifier and an adversary who is allowed to perturb the data by a certain amount. As noted by Pydi and Jog [52], this interpretation has implications for algorithm design by suggesting that robust classifiers can be constructed by jointly optimizing over classification rules and adversarial perturbations.

This recent progress on adversarial binary classification lays the groundwork for a theoretical understanding of adversarial robustness, but it is limited insofar as it focuses only on minimizers of the adversarial classification risk. In practice, minimizing the empirical adversarial classification risk is computationally intractable; as a consequence; the adversarial training procedure typically minimizes an objective called a *surrogate* risk, which is chosen to be easier to optimize. In the non-adversarial setting, the properties of surrogate risks are well known [see, e.g. 8], but in the adversarial scenario, existing results for the adversarial classification risk fail to carry over to surrogate risks. In particular, the existence and minimax results described above are not known to hold. We close this gap by developing an analogous theory for adversarial surrogate risks. Our main theorems (Theorems 7–9) establish that strong duality holds for the adversarial surrogate risk minimization problem,

that solutions to the primal and dual problems exist, and that these optimizers satisfy a complementary slackness condition.

These results suggest explanations for empirical observations, such as the existence of transfer attacks. Specifically, our analysis suggests that adversarial examples are a property of the data distribution rather than a specific model. In fact, the complementary slackness theorem presented in this paper states that certain attacks are the strongest possible adversary against *any* minimizer of the adversarial surrogate risk, which might explain why adversarial examples tend to transfer between trained neural nets. Furthermore, our theorems suggest that a training algorithm should optimize over neural nets and adversarial perturbations simultaneously. Adversarial training, the current state of the art method for finding adversarially robust networks, does not follow this procedure. The adversarial training algorithm tracks an estimate of the optimal function $\tilde{f}$. To update $\tilde{f}$, the algorithm first finds *optimal* adversarial examples at the current estimate $\tilde{f}$, and then performs a gradient descent step. See the papers [30, 35, 39] for more details on adversarial training. Finding these adversarial examples is a computationally intensive procedure. On the other hand, algorithms for optimizing minimax problems in the finite dimensional setting alternate between primal and dual steps [44]. This observation suggests that designing an algorithm that optimizes over model parameters and adversarial perturbations simultaneously is a promising research direction. Domingo-Enrich et al. [20], Trillos and Trillos [61], and Wang and Chizat [66] adopt this approach, and one can view the minimax results of this paper as a mathematical justification for the use of surrogate losses in such algorithms.

Lastly, our theorems are an important first step in understating statistical properties of surrogate losses. Recall that one minimizes a surrogate risk because minimizing the original risk is computationally intractable. If a sequence of functions which minimizes the surrogate risk also minimizes the classification risk, then that surrogate is referred to as a *consistent risk*. Similarly, if a sequence of functions which minimizes the adversarial surrogate risk

also minimizes the adversarial classification risk, then that surrogate is referred to as an *adversarially consistent risk*. Much prior work studies the consistency of surrogate risks [8, 38, 43, 49, 57, 75]. Alarmingly, [42] show that a family of surrogates used in applications is not adversarially consistent. In follow-up work, we show that our results can be used to characterize adversarially consistent supremum-based risks for binary classification [26], strengthening results on calibration in the adversarial setting [4, 6, 42].

2.2 Related Works

This paper extends prior work on the adversarial Bayes classifier. Pydi and Jog [52] first proved multiple minimax theorems for the adversarial classification risk using optimal transport and Choquet capacities, showing an intimate connection between adversarial learning and optimal transport. Later, follow-up work used optimal transport minimax reformulations of the adversarial learning problem to derive new algorithms for adversarial learning. Trillos, Jacobs, and Kim [63] reformulate adversarial learning in terms of a multi-marginal optimal transport problem and then apply existing techniques from optimal transport to find a new algorithm. Domingo-Enrich et al. [20], Trillos and Trillos [61], and Wang and Chizat [66] propose ascent-descent algorithms based on optimal transport and use mean-field dynamics to analyze convergence. These approaches leverage the minimax view of the adversarial training problem to optimize over model parameters and optimal attacks simultaneously. Gao, Chen, and Kleywegt [27] use an optimal transport reformulation to find regularizers that encourage robustness. Wong, Schmidt, and Kolter [69] and Wu, Wang, and Yu [70] use Wasserstein metrics to formulate adversarial attacks on neural networks.

Further work analyzes properties of the adversarial Bayes classifier. Awasthi, Frank, and Mohri [2], Bhagoji, Cullina, and Mittal [11], and Bungert, Trillos, and Murray [16] all prove the existence of the adversarial Bayes classifier, using different techniques. Yang et al. [73]

studied the adversarial Bayes classifier in the context of non-parametric methods. Pydi and Jog [50] and Bhagoji, Cullina, and Mittal [11] further introduced methods from optimal transport to study adversarial learning. Lastly, [64] give necessary and sufficient conditions describing the boundary of the adversarial Bayes classifier. Simultaneous work [36] also proves the existence of minimizers to adversarial surrogate risks using prior results on the adversarial Bayes classifier.

The adversarial training algorithm is also well studied from an empirical perspective. Demontis et al. [18] discussed an explanation of transfer attacks on neural nets trained using standard methods, but did not extend their analysis to the adversarial training setting. [35, 39, 67] study the adversarial training algorithm and improving the runtime. Two particularly popular attacks used in adversarial training are the FGSM attack [30] and the PGD attack [39]. More recent popular variants of this algorithm include [34, 56, 68, 71].

2.3 Background and Notation

2.3.1 Adversarial Classification

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

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

$$R(f) = \int \mathbf{1}_{f(\mathbf{x}) \leq 0} d\mathbb{P}_1 + \int \mathbf{1}_{f(\mathbf{x}) > 0} d\mathbb{P}_0. \tag{2.1}$$

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

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

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

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

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

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

$$R^\epsilon(f) = \int \sup_{\|\mathbf{h}\|\leq\epsilon} \mathbf{1}_{f(\mathbf{x}+\mathbf{h})\leq 0} d\mathbb{P}_1 + \int \sup_{\|\mathbf{h}\|\leq\epsilon} \mathbf{1}_{f(\mathbf{x}+\mathbf{h})>0} d\mathbb{P}_0 \tag{2.2}$$

which is the expected proportion of errors subject to the adversarial evasion attack. The expression $\sup_{\|\mathbf{h}\|\leq\epsilon} \mathbf{1}_{f(\mathbf{x}+\mathbf{h})\leq 0}$ evaluates to 1 at a point $\mathbf{x}$ iff $\mathbf{x}$ can be moved into the set $[f \leq 0]$ by a perturbation of size at most ϵ. Equivalently, this set is the Minkowski sum $\oplus$ of $[f \leq 0]$ and $\overline{B_\epsilon(\mathbf{0})}$. For any set A, let A^ϵ denote

$$A^\epsilon = \{\mathbf{x}\colon \exists \mathbf{h} \text{ with } \|\mathbf{h}\| \leq \epsilon \text{ and } \mathbf{x}+\mathbf{h} \in A\} = A \oplus \overline{B_\epsilon(\mathbf{0})} = \bigcup_{\mathbf{a}\in A} \overline{B_\epsilon(\mathbf{a})}. \tag{2.3}$$

This operation 'thickens' the boundary of a set by ϵ. With this notation, (2.2) can be

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

Unlike the classification risk R, finding minimizers to R^ϵ is nontrivial. One can re-write R^ϵ in terms of $\mathbb{P}$ and η but the resulting integrand cannot be minimized in a pointwise fashion. Nevertheless, it can be shown that minimizers of R^ϵ exist [2, 16, 26, 52].

2.3.2 Surrogate Risks

As minimizing the empirical version of risk in (2.1) is computationally intractable, typical machine learning algorithms minimize a proxy to the classification risk called a *surrogate risk*. In fact, Ben-David, Eiron, and Long [9] show that minimizing the empirical classification risk is NP-hard in general. A popular surrogate is

$$R_\phi(f) = \int \phi(f) d\mathbb{P}_1 + \int \phi(-f) d\mathbb{P}_0 \tag{2.4}$$

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

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

A particularly important example, which plays a large role in our proofs, is the exponential loss $\psi(\alpha) = e^{-\alpha}$, which will be denoted by ψ in the remainder of this paper. Assumption 1 includes many but not all all surrogate risks used in practice. Notably, some multiclass surrogate risks with two classes are of a somewhat different form, see for instance [59] for more details.

[1] Notice that due to the asymmetry of the sign function at 0 in (2.1), R_ϕ is not quite a generalization of R.

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

$$R_\phi(f) = \int \eta(\mathbf{x})\phi(f(\mathbf{x})) + (1-\eta(\mathbf{x}))\phi(-f(\mathbf{x}))d\mathbb{P} \tag{2.5}$$

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

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

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

$$\eta(\mathbf{x})\phi(f^*(\mathbf{x})) + (1-\eta(\mathbf{x}))\phi(-f^*(\mathbf{x})) = C_\phi^*(\eta(\mathbf{x})) \quad \mathbb{P}\text{-a.e.} \tag{2.6}$$

Thus one can write the minimum value of R_ϕ as

$$\inf_f R_\phi(f) = \int C_\phi^*(\eta)d\mathbb{P}. \tag{2.7}$$

To guarantee the existence of a function satisfying (2.6), we must allow our functions to take values in the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. Allowing the value $\alpha = +\infty$ is necessary, for instance, for the exponential loss $\psi(\alpha) = e^{-\alpha}$: when $\eta = 1$, the minimum of $C_\psi(1, \alpha) = e^{-\alpha}$ is achieved at $\alpha = +\infty$. In fact, one can express a minimizer as a function of the conditional probability $\eta(\mathbf{x})$ using (2.6). For a loss ϕ, define $\alpha_\phi : [0, 1] \to \overline{R}$ by

$$\alpha_\phi(\eta) = \inf\{\alpha : \alpha \text{ is a minimizer of } C_\phi(\eta, \cdot)\}. \tag{2.8}$$

Lemma 25 in Appendix A.3 shows that the function α_ϕ is monotonic and $\alpha_\phi(\eta)$ is in fact a minimizer of $C_\phi(\eta, \cdot)$. Thus

$$f^*(\mathbf{x}) = \alpha_\phi(\eta(\mathbf{x})) \tag{2.9}$$

is measurable and satisfies (2.6). Therefore, the function f^* must be a minimizer of the risk

R_ϕ.

Similarly, rather directly minimizing the adversarial classification risk, practical algorithms minimize an *adversarial surrogate*. The adversarial counterpart to (2.4) is

$$R_\phi^\epsilon(f) = \int \sup_{\|\mathbf{h}\| \leq \epsilon} \phi(f(\mathbf{x}+\mathbf{h}))d\mathbb{P}_1 + \int \sup_{\|\mathbf{h}\| \leq \epsilon} \phi(-f(\mathbf{x}+\mathbf{h}))d\mathbb{P}_0. \tag{2.10}$$

Due to the definitions of the adversarial risks (2.2) and (2.10), the operation of finding the supremum of a function over ϵ-balls is central to our subsequent analysis. For a function g, we define

$$S_\epsilon(g)(\mathbf{x}) = \sup_{\|\mathbf{h}\| \leq \epsilon} g(\mathbf{x}+\mathbf{h}) \tag{2.11}$$

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

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

By analogy to (2.5), we equivalently write the risk R_ϕ^ϵ in terms of $\mathbb{P}$ and η:

$$R_\phi^\epsilon(f) = \int \eta(\mathbf{x})S_\epsilon(\phi \circ f)(\mathbf{x}) + (1-\eta(\mathbf{x}))S_\epsilon(\phi \circ -f)(\mathbf{x})d\mathbb{P}. \tag{2.12}$$

However, unlike (2.5), because the integrand of R_ϕ^ϵ cannot be minimized in a pointwise manner, proving the existence of minimizers to R_ϕ^ϵ is non-trivial. In fact, unlike the adversarial classification risk R^ϵ, there is little theoretical understanding of the properties of R_ϕ^ϵ.

2.3.3 Measurability

In order to define the adversarial risks R^ϵ and R_ϕ^ϵ, one must show that $S_\epsilon(\mathbf{1}_A), S_\epsilon(\phi \circ f)$ are measurable. To illustrate this concern, Pydi and Jog [52] show that for every $\epsilon > 0$ and $d > 1$, there is a Borel set C for which the function $S_\epsilon(\mathbf{1}_C)(\mathbf{x})$ is not Borel measurable.

However, if g is Borel, then $S_\epsilon(g)$ is always measurable with respect to a larger σ-algebra called the *universal σ-algebra* $\mathscr{U}(\mathbb{R}^d)$. Such a function is called *universally measurable*. We prove the following theorem and formally define the universal σ-algebra in Appendix A.1.

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

In fact, in Appendix A.1, we show that a function defined by a supremum of a universally measurable function over a compact set is universally measurable—a result of independent interest. The universal σ-algebra is smaller than the completion of $\mathcal{B}(\mathbb{R}^d)$ with respect to any Borel measure. Thus, in the remainder of the paper, unless otherwise noted, all measures will be Borel measures and the expression $\int S_\epsilon(f)d\mathbb{Q}$ will be interpreted as the integral of $S_\epsilon(f)$ with respect to the completion of $\mathbb{Q}$.

2.3.4 The W_∞ Metric

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

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

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

Jylhä [33, Theorem 2.6] proves that the infimum is always attained. Therefore, $\mathbb{P}$, $\mathbb{Q}$ are within a Wasserstein-∞ distance of ϵ if there is a coupling γ for $\mathbb{P}$ and $\mathbb{Q}$ for which $\operatorname{supp}\gamma$ is contained in the set $\Delta_\epsilon = \{(\mathbf{x}, \mathbf{x}') \colon \|\mathbf{x} - \mathbf{x}'\| \le \epsilon\}$. This optimal coupling will be a useful tool in proving theorems throughout this paper.

The ∞-Wasserstein metric is closely related to the to the operation S_ϵ. First, we show that S_ϵ can be expressed as a supremum of integrals over a Wasserstein-∞ ball. For a

measure $\mathbb{Q}$, we write

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

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

$$\int S_\epsilon(f) d\mathbb{Q} = \sup_{\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})} \int f d\mathbb{Q}' \tag{2.13}$$

Lemma 5.1 of Pydi and Jog [52] proves an analogous statement for sets, namely that $\mathbb{Q}(A^\epsilon) = \sup_{\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})} \mathbb{Q}(A)$, under suitable assumptions on $\mathbb{Q}$ and $\mathbb{Q}'$.

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

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

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

This observation will be central to proving a duality result. See Appendix A.2 for proofs of Lemmas 3 and 4.

2.4 Main Results and Outline of the Paper

2.4.1 Summary of Main Results

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

involves a inf followed by a sup:

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

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

$$\begin{aligned}
\inf_{f \text{ Borel}} R_\phi^\epsilon(f) &\geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{f \text{ Borel}} \int \phi \circ f d\mathbb{P}_1' + \int \phi \circ -f d\mathbb{P}_0' \\
&= \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{f \text{ Borel}} \int \frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')} \phi(f) + \left(1 - \frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')}\right) \phi(-f) d(\mathbb{P}_0' + \mathbb{P}_1') \\
&\geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \int C_\phi^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')}\right) d(\mathbb{P}_0' + \mathbb{P}_1'). \qquad (2.14)
\end{aligned}$$

If we define

$$\bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') = \int C_\phi^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')}\right) d(\mathbb{P}_0' + \mathbb{P}_1'), \qquad (2.15)$$

then we have shown

$$\inf_{f \text{ Borel}} R_\phi^\epsilon(f) \geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1'). \qquad (2.16)$$

This statement is a form of weak duality.

When the surrogate adversarial risk is replaced by the standard adversarial classification risk, Pydi and Jog [52] proved that the analogue of (2.16) is actually an equality, so that strong duality holds for the adversarial classification problem. Concretely, by analogy to (2.15), consider

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

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

$$\tilde{\mathcal{B}}_\epsilon^\infty(\mathbb{Q}) = \{\mathbb{Q}' \colon W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon, \mathbb{Q}' \text{ a measure on } (\mathbb{R}^d, \mathcal{L}_\mu(\mathbb{R}^d))\}. \tag{2.17}$$

Pydi and Jog [52] show the following.

Theorem 5 ([52, Theorem 7.1])**.** *Assume that $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous with respect to the Lebesgue measure μ. Then*

$$\inf_{f \text{ Lebesgue}} R^\epsilon(f) = \sup_{\substack{\mathbb{P}_0' \in \tilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \tilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') \tag{2.18}$$

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

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

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

We start by proving the following:

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

$$\inf_{f \text{ Borel}} R_\phi^\epsilon(f) = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1'). \tag{2.19}$$

When $\epsilon = 0$, we recover the fundamental characterization of the minimum risk for standard (non-adversarial) classification in (2.7). Theorem 6 can be rephrased as

$$\inf_{f \text{ Borel}} \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \hat{R}_\phi(f, \mathbb{P}_0', \mathbb{P}_1') = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{f \text{ Borel}} \hat{R}_\phi(f, \mathbb{P}_0', \mathbb{P}_1') \tag{2.20}$$

where

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

As discussed in Pydi and Jog [52], this result has an appealing game theoretic interpretation: adversarial learning with a surrogate risk can be though of as a zero-sum game between the learner who selects a function f and the adversary who chooses perturbations through $\mathbb{P}'_0$ and $\mathbb{P}'_1$. Furthermore, the player to pick first does not have an advantage.

Additionally, (2.20) suggest that training adversarially robust classifiers could be accomplished by optimizing over primal functions f and dual distributions $\mathbb{P}'_0, \mathbb{P}'_1$ *simultaneously*.

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

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

1)

$$\int \phi \circ f^* d\mathbb{P}^*_1 = \int S_\epsilon(\phi(f^*)) d\mathbb{P}_1 \quad \textit{and} \quad \int \phi \circ -f^* d\mathbb{P}^*_0 = \int S_\epsilon(\phi(-f^*)) d\mathbb{P}_0 \tag{2.21}$$

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

$$\eta^*(\mathbf{x})\phi(f^*(\mathbf{x})) + (1 - \eta^*(\mathbf{x}))\phi(-f^*(\mathbf{x})) = C^*_\phi(\eta^*(\mathbf{x})) \quad \mathbb{P}^*\textit{-a.e.} \tag{2.22}$$

This theorem implies that every minimizer f^* of R^ϵ_ϕ and every maximizer $(\mathbb{P}^*_0, \mathbb{P}^*_1)$ of $\bar{R}_\phi$ forms a primal-dual pair. The condition (2.21) states that every maximizer of $\bar{R}_\phi$ is an optimal adversarial attack on f^* while the condition (2.22) states that every minimizer f^* of R^ϵ_ϕ also minimizes the conditional risk $C_\phi(\eta^*, \cdot)$ under the distribution of optimal adversarial attacks. Explicitly: (2.22) implies that every minimizer f^* minimizes the loss $\hat{R}_\phi(f, \mathbb{P}^*_0, \mathbb{P}^*_1) = \int C(\eta^*(\mathbf{x}), f(\mathbf{x})) d\mathbb{P}^*$ in a pointwise manner $\mathbb{P}^*$-a.e., or in other words, the

function f^* minimizes the *standard* surrogate risk with respect to the optimal adversarially perturbed distributions. This fact is the relation (2.6) with respect to the measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ that maximize the dual $\bar{R}_\phi$.

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

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

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

Theorem 3.5 of [33] implies that when the norm $\|\cdot\|$ is strictly convex and $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous with respect to Lebesgue measure, the optimal $\mathbb{P}_0^*, \mathbb{P}_1^*$ of Theorem 8 are induced by a transport map. Corollary 3.11 of [33] further implies that these transport maps are continuous a.e. with respect to the Lebesgue measure μ. As the ℓ_∞ metric is commonly used in practice, whether there exist maximizers of the dual of this type for non-strictly convex norms remains an attractive open problem.

In analogy with (2.6) and (2.9) one would hope that due to the complementary slackness condition (2.22), one could define a minimizer in terms of the conditional $\eta^*(\mathbf{x})$. Notice, however, that as this quantity is only defined $\mathbb{P}^*$-a.e., verifying the other complementary slackness condition (2.21) would be a challenge. To circumvent this issue, we construct a function $\hat{\eta} : \mathbb{R}^d \to [0,1]$, defined on all of $\mathbb{R}^d$, to which we can apply (2.9). Concretely, we show that $\alpha_\phi(\hat{\eta}(\mathbf{x}))$ is always a minimizer of R^ϵ_ϕ, with α_ϕ as defined in (2.8).

Theorem 9. *There exists a Borel function $\hat{\eta} : (\operatorname{supp} \mathbb{P})^\epsilon \to [0,1]$ for which $f^*(\mathbf{x}) = \alpha_\phi(\hat{\eta}(\mathbf{x}))$ is a minimizer of R^ϵ_ϕ for any ϕ with α_ϕ as in defined in (2.8). In particular, there exists a Borel minimizer of R^ϵ_ϕ.*

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

Simultaneous work [36] also proves the existence of a minimizer to the primal R^ϵ_ϕ along with a statement on the structure of this minimizer. Their approach leverages prior results on the adversarial Bayes classifier to construct a minimizer to the adversarial surrogate risk.

2.4.2 Outline of Main Argument

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

The remainder of the paper then argues that minimizing Θ is equivalent to minimizing R^ϵ_ϕ. Notice that the Fenchel-Rockafellar theorem actually implies the existence of dual

maximizers. We show that that dual maximizers of $\bar{R}_\psi$ for $\psi(\alpha) = e^{-\alpha}$ satisfy certain nice properties that are *independent* of the loss ψ. These properties then allow us to construct the function $\hat{\eta}$ present in Theorem 9 in addition to minimizers of Θ from the dual maximizers of $\bar{R}_\psi$, for any loss ϕ. The construction of these minimizers guarantee that they minimize R^ϵ_ϕ in addition to Θ.

2.4.3 Paper Outline

Section 2.5 proves strong duality and complementary slackness theorems for $\bar{R}_\phi$ and Θ, the convex relaxation of R^ϵ_ϕ. Next, in Section 2.6, a version of the complementary slackness result is used to prove the existence of minimizers to Θ. Subsequently, Section 2.7 shows the equivalence between Θ and R^ϵ_ϕ.

Appendix A.1 proves Theorem 1 and further discusses universal measurability. Next, Appendix A.2 proves all the results about the W_∞-norm used in this paper. Appendix A.3 then defines the function α_ϕ which is later used in the proof of several results. Appendices A.4, A.5, A.6, and A.7.3 contain technical deferred proofs.

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

2.5.1 Strong Duality

The fundamental duality relation of this paper follows from employing the Fenchel-Rockafellar theorem in conjunction with the Riesz representation theorem, stated below for reference. See e.g. [65] for more on this result.

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

such that

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

and that Θ *is continuous at* z_0*. Then*

$$\inf_{z \in E}[\Theta(z) + \Xi(z)] = \sup_{z^* \in E^*}[-\Theta^*(z^*) - \Xi^*(-z^*)] \tag{2.23}$$

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

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

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

See Theorem 1.9 of [65] and result 7.17 of [22] for more details.

Notice that in the Fenchel-Rockafellar theorem, the left-hand side of (2.23) is convex while the right-hand side is concave. However, when ϕ is non-convex, R^ϵ_ϕ is not convex. In order to apply the Fenchel-Rockafellar theorem, we will relax the primal R^ϵ_ϕ will to a convex functional Θ.

We define Θ as

$$\Theta(h_0, h_1) = \int S_\epsilon(h_1) d\mathbb{P}_1 + \int S_\epsilon(h_0) d\mathbb{P}_0 \tag{2.24}$$

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

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

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

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

$$S_\phi = \left\{ \begin{array}{l} (h_0, h_1) \colon h_0, h_1 \colon K^\epsilon \to \overline{\mathbb{R}} \text{ Borel, } 0 \leq h_0, h_1 \text{ and for} \\ \text{all } \mathbf{x} \in \mathbb{R}^d,\ \eta \in [0,1],\ \eta h_0(\mathbf{x}) + (1-\eta) h_1(\mathbf{x}) \geq C_\phi^*(\eta) \end{array} \right\} \tag{2.25}$$

where $K = \operatorname{supp}(\mathbb{P}_0 \cup \mathbb{P}_1)$. (Notice that the definition of $S_\epsilon(g)$ in (2.11) assumes that the domain of g must include $\overline{B_\epsilon(\mathbf{x})}$. Thus in order to define the integral $\int S_\epsilon(h) d\mathbb{Q}$, the domain of h must include $(\operatorname{supp} \mathbb{Q})^\epsilon$.) Thus we define Ξ as

$$\Xi(h_0, h_1) = \begin{cases} 0 & \text{if } (h_0, h_1) \in S_\phi \\ +\infty & \text{otherwise} \end{cases} \tag{2.26}$$

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

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

$$\inf_{(h_0,h_1) \in S_\phi \cap E} \int h_1 d\mathbb{P}'_1 + \int h_0 d\mathbb{P}'_0 = \bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) \tag{2.27}$$

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

We sketch the proof and formally fill in the details in Appendix A.4. Let $\mathbb{P}' = \mathbb{P}'_0 + \mathbb{P}'_1$, $\eta' = d\mathbb{P}'_1 / d\mathbb{P}'$. Then

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

Clearly, the inequality $\geq$ holds because $\eta' h_1 + (1-\eta')h_0 \geq C_\phi^*(\eta')$ for all $(h_0, h_1) \in S_\phi$. Equality is achieved at $h_1 = \phi(\alpha_\phi(\eta')), h_0 = \phi(-\alpha_\phi(\eta'))$, with α_ϕ as in (2.8). However, these functions may not be continuous. In Appendix A.4, we show that h_0, h_1 can be approximated arbitrarily well by elements of $S_\phi \cap E$.

Lemma 13. *Let ϕ be a non-increasing, lower semi-continuous loss function and let $\mathbb{P}_0, \mathbb{P}_1$ be compactly supported finite Borel measures. Let S_ϕ be as in (2.25).*

Then

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0, h_1) = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') \tag{2.28}$$

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

Proof. We will show a version of (2.28) with the infimum taken over $S_\phi \cap E$, and then argue that the same claim holds when the infimum is taken over S_ϕ.

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

Let $K = \text{supp}(\mathbb{P}_0 + \mathbb{P}_1)$. We will apply the Fenchel-Rockafellar Duality Theorem to the functionals given by (2.24) and (2.26) on the vector space $E = C_b(K^\epsilon) \times C_b(K^\epsilon)$ equipped with the sup norm. By the Riesz representation theorem, dual of the space E is $E^* = \mathcal{M}(K^\epsilon) \times \mathcal{M}(K^\epsilon)$.

To start, we argue that the Fenchel-Rockafellar duality theorem applies to these functionals. First, notice that if $(h_0, h_1) \in E$, then both h_0, h_1 are bounded so $\Theta(h_0, h_1) < \infty$. Furthermore, Θ is convex due to the subadditivity of supremum and Ξ is convex because the constraint $h_0(\mathbf{x}) + (1-\eta)h_1(\mathbf{x}) \geq C_\phi^*(\eta)$ is linear in h_0 and h_1. Furthermore, Θ is continuous on E because Θ is convex and bounded and E is open, see Theorem 2.14 of [7].

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

Clearly $\inf_E \Theta(h_0, h_1) + \Xi(h_0, h_1)$ reduces to the left-hand side of (2.28).

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

$$
\begin{aligned}
-\Xi^*(-\mathbb{P}'_0, -\mathbb{P}'_1) &= - \sup_{(h_0,h_1)\in S_\phi \cap E} - \int h_0 d\mathbb{P}'_0 - \int h_1 d\mathbb{P}'_1 \qquad (2.29)\\
&= \begin{cases} \bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) & \text{if } \mathbb{P}'_i \geq 0 \\ -\infty & \text{otherwise} \end{cases}
\end{aligned}
$$

This computation implies that the term $-\Xi^*(-\mathbb{P}'_0, -\mathbb{P}'_1)$ present in the Fenchel-Rockafellar Theorem is not $-\infty$ iff $\mathbb{P}'_0, \mathbb{P}'_1$ are positive measures. Next, notice that because $\Theta(h_0, h_1) < +\infty$ for all $(h_0, h_1) \in E$, $-\Theta^*(\mathbb{P}'_0, \mathbb{P}'_1)$ is never $+\infty$. Therefore, it suffices to compute Θ^* for positive measures $\mathbb{P}'_0, \mathbb{P}'_1$. Lemma 4 implies that for positive measures $\mathbb{P}'_0, \mathbb{P}'_1$,

$$
\begin{aligned}
\Theta^*(\mathbb{P}'_0, \mathbb{P}'_1) &= \sup_{h_0,h_1\in C_0(K^\epsilon)} \int h_1 d\mathbb{P}'_1 + \int h_0 d\mathbb{P}'_0 - \left(\int S_\epsilon(h_0) d\mathbb{P}_0 + \int S_\epsilon(h_1) d\mathbb{P}_1 \right) \\
&= \sup_{h_1\in C_0(K^\epsilon)} \left(\int h_1 d\mathbb{P}'_1 - \int S_\epsilon(h_1) d\mathbb{P}_1 \right) + \sup_{h_0\in C_0(K^\epsilon)} \left(\int h_0 d\mathbb{P}'_0 - \int S_\epsilon(h_0) d\mathbb{P}_0 + \right) \\
&= \begin{cases} 0 & \mathbb{P}'_0, \mathbb{P}'_1 \text{ positive measures, with } W_\infty(\mathbb{P}'_0, \mathbb{P}_0) \leq \epsilon \text{ and } W_\infty(\mathbb{P}'_1, \mathbb{P}_1) \leq \epsilon \\ +\infty & \mathbb{P}'_0, \mathbb{P}'_1 \text{ positive measures, with either } W_\infty(\mathbb{P}'_0, \mathbb{P}_0) > \epsilon \text{ or } W_\infty(\mathbb{P}'_1, \mathbb{P}_1) > \epsilon \end{cases}
\end{aligned}
$$

Therefore

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

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

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

The opposite inequality follows from the weak duality argument presented in (2.16) in Section 2.4.1. See Lemma 125 of Appendix A.5 for a full proof. □

Note that this proof does not easily extend to an unbounded domain X: for a non-compact space, the dual of $C_b(X)$ is much larger than $\mathcal{M}(X)$, and thus a naive application of the Fenchel-Rockafellar Theorem would result in a different right-hand side than (2.28). On the other hand, the Reisz representation theorem for an unbounded domain X states that the dual of $C_0(X)$ is $\mathcal{M}(X)$, where $C_0(X)$ is the set of continuous bounded functions vanishing at ∞. At the same time, if $h_0, h_1 \in C_0(X)$, then $\eta h_1(\mathbf{x}) + (1-\eta)h_0(\mathbf{x})$ becomes arbitrarily small for large $\mathbf{x}$ so the constraint $\eta h_1(\mathbf{x}) + (1-\eta)h_0(\mathbf{x}) \geq C^*_\phi(\eta)$ cannot hold for all η. Thus if K is unbounded, $S_\phi \cap C_0(X) = \emptyset$ and the functional Ξ would be $+\infty$ everywhere on $C_0(X)$, precluding an application of the Fenchel-Rockafellar Theorem.

However, Lemma 13 can be extended to distributions with arbitrary support via a simple approximation argument. By Lemma 13, the strong duality result holds for the distributions defined by $\mathbb{P}_0^n = \mathbb{P}_0\big|_{\overline{B_n(\mathbf{0})}}, \mathbb{P}_1^n = \mathbb{P}_1\big|_{\overline{B_n(\mathbf{0})}}$. One then shows strong duality by computing the limit of the primal and dual problems as $n \to \infty$. We therefore obtain the following Lemma, which is proved formally in Appendix A.5.

Lemma 14. *Let ϕ be a non-increasing, lower semi-continuous loss function and let $\mathbb{P}_0, \mathbb{P}_1$ be finite Borel measures supported on $\mathbb{R}^d$. Let S_ϕ be as in (2.25). Then*

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

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

2.5.2 Complementary Slackness

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

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

1)

$$\int h_1^* d\mathbb{P}_1^* = \int S_\epsilon(h_1^*) d\mathbb{P}_1 \quad \text{and} \quad \int h_0^* d\mathbb{P}_0^* = \int S_\epsilon(h_0^*) d\mathbb{P}_0 \tag{2.30}$$

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

$$\eta^*(\mathbf{x}) h_1^*(\mathbf{x}) + (1 - \eta^*(\mathbf{x})) h_0^*(\mathbf{x}) = C_\phi^*(\eta^*(\mathbf{x})) \quad \mathbb{P}^*\text{-a.e.} \tag{2.31}$$

This lemma is proved in Appendix A.6. Theorem 7 will later follow from this result.

To show that Lemma 15 is non-vacuous, one must prove that there exist minimizers to Θ over S_ϕ, which we delay to Sections 2.6 and 2.7. Notice that the application of the Fenchel-Rockafellar Theorem in Lemma 13 actually implies the existence of dual maximizers in the case of compactly supported $\mathbb{P}_0, \mathbb{P}_1$.

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

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

1)

$$\lim_{n\to\infty} \int S_\epsilon(h_0^n) d\mathbb{P}_0 - \int h_0^n d\mathbb{P}_0^* = 0, \quad \lim_{n\to\infty} \int S_\epsilon(h_1^n) d\mathbb{P}_1 - \int h_1^n d\mathbb{P}_1^* = 0 \tag{2.32}$$

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

$$\lim_{n\to\infty} \int \eta^* h_1^n + (1-\eta^*)h_0^n - C_\phi^*(\eta^*)d\mathbb{P}^* = 0 \tag{2.33}$$

Proof. Let

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

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

$$\int h_1^n d\mathbb{P}_1^* + \int h_0^n d\mathbb{P}_0^* = \int \eta^* h_1^n + (1-\eta^*)h_0^n d\mathbb{P}^* \geq \int C_\phi^*(\eta^*)d\mathbb{P}^* = m \tag{2.34}$$

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

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

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

$$\delta \geq \int \eta^* h_1^n + (1-\eta^*)h_0^n d\mathbb{P}^* - \int C_\phi^*(\eta^*)d\mathbb{P}^* \geq 0 \tag{2.35}$$

which implies (2.33). Next, (2.34) further implies

$$m - \int h_1^n d\mathbb{P}_1^* + \int h_0^n d\mathbb{P}_0^* \leq 0 \tag{2.36}$$

Now this inequality implies

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

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

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

which implies (2.32).

□

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

$$\inf_{x\in C_P} \Theta(x) = \inf_{x\in C_p} \sup_{y\in C_D} \langle y, x\rangle = \sup_{y\in C_D} \inf_{x\in C_P} \langle y, x\rangle = \sup_{y\in C_D} \Xi^*(y) \tag{2.37}$$

Let y^* be a maximizer of the dual problem and let m be the minimal value of Θ over C_P. If x_k is a minimizing sequence of Θ, then for $\delta > 0$ and sufficiently large k, $\delta + m > \Theta(x_k)$ and thus by (2.37),

$$m + \delta > \Theta(x_k) = \sup_{y\in C_p} \langle y, x_k\rangle \geq \langle y^*, x_k\rangle \geq \inf_{x\in C_D} \langle y^*, x\rangle = \inf_{x\in C_D} \Xi^*(x) = m \tag{2.38}$$

and therefore $\lim_{k\to\infty}\langle y^*, x_k\rangle = m$. Condition (2.31) is this statement adapted to the adversarial learning problem. Furthermore, subtracting $\Theta(x_k)$ from (2.38) and taking the limit $k \to \infty$ results in $\lim_{k\to\infty} \Theta(x_k) - \langle y^*, x_k\rangle = 0$. In our adversarial learning scenario, this statement is equivalent to the conditions in (2.32) due to Lemma 3. Furthermore, just like the standard complementary slackness theorems, the relations $\lim_{k\to\infty}\langle y^*, x_k\rangle = m$,

$\lim_{k\to\infty} \Theta(x_k) - \langle y^*, x_k\rangle = 0$ imply that x_k is a minimizing sequence for Θ.

In the classical proof of the Kantorovich duality, one can choose Θ, Ξ of a form similar to the discussion above, see for instance Theorem 1.3 of [65]. Using an argument similar to (2.38), one can prove approximate complementary slackness for the Kantorovich problem called the quantitative Knott-Smith criteria, see Theorems 2.15, 2.16 of [65] for further discussion.

2.6 Existence of Minimizers to Θ over S_ψ

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

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

See Appendix A.7.1 for a proof. The existence of minimizers of Θ for the exponential loss then follows from properties of C_ψ. Let (h_0^n, h_1^n) be a minimizing sequene of $\bar{R}_\phi$. Because the function C_ψ is strictly concave, one can use the condition (2.33) to show that there is a subsequence $\{n_k\}$ along which $h_0^{n_k}(\mathbf{x}')$, $h_1^{n_k}(\mathbf{x}')$ converge $\mathbb{P}_0^*, \mathbb{P}_1^*$-a.e. respectively. Due to (2.32), $S_\epsilon(h_0^{n_k})(\mathbf{x})$, $S_\epsilon(h_1^{n_k})$ also converge $\mathbb{P}_0, \mathbb{P}_1$-a.e. respectively along this subsequence. This observation suffices to show the existence of functions that minimize Θ over S_ψ.

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

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

$$\eta a_n + (1-\eta) b_n \geq C^*_\psi(\eta) \text{ for all } \eta \in [0,1] \tag{2.39}$$

and

$$\lim_{n\to\infty} \eta_0 a_n + (1-\eta_0) b_n = C_\psi^*(\eta_0) \tag{2.40}$$

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

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

To prove Lemma 18, we show that all convergent subsequences of $\{a_n\}$ and $\{b_n\}$ must converge to a and b that satisfy $\eta_0 a + (1-\eta_0) b = C_\phi^*(\eta_0)$ and $a - b \in \partial C_\psi^*(\eta_0)$. As the set $\partial C_\psi^*(\eta_0)$ is a singleton, the values $a = \psi(\alpha_\psi(\eta_0))$ and $b = \psi(\alpha_\psi(\eta_0))$ uniquely solve these equations for a and b. Therefore the sequences $\{a_n\}$ and $\{b_n\}$ must converge to a and b as well. See Appendix A.7.2 for a formal proof. This result applied to a minimizing sequence of Θ allows one to find a subsequence with certain convergence properties.

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

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

Then (2.33) of Lemma 16 implies that

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

and (2.32) implies that

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

Recall that on a bounded measure space, L^1 convergence implies a.e. convergence along a

subsequence (see Corollary 2.32 of [22]). Thus one can pick a subsequence n_k along which

$$\lim_{k\to\infty} \eta^*(\mathbf{x}')h_1^{n_k}(\mathbf{x}') + (1-\eta^*(\mathbf{x}'))h_0^{n_k}(\mathbf{x}') - C_\psi(\eta^*(\mathbf{x}')) = 0 \tag{2.41}$$

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

$$\lim_{k\to\infty} S_\epsilon(h_1^{n_k})(\mathbf{x}) - h_1^{n_k}(\mathbf{x}') = 0, \quad \lim_{k\to\infty} S_\epsilon(h_0^{n_k})(\mathbf{x}) - h_0^{n_k}(\mathbf{x}') = 0 \tag{2.42}$$

γ_1, γ_0-a.e. respectively.

Furthermore, $\eta h_1^n + (1-\eta)h_0^n \geq C_\psi^*(\eta)$ for all $\eta \in [0,1]$. Thus (2.41) and Lemma 18 imply that $h_{n_k}^1$ converges to $\psi(\alpha_\psi(\eta^*))$ and $h_{n_k}^0$ converges to $\psi(-\alpha_\psi(\eta^*))$ $\gamma_0+\gamma_1$-a.e. Equation 2.42 then implies that $S_\epsilon(h_1^{n_k})(\mathbf{x})$, $S_\epsilon(h_0^{n_k})(\mathbf{x})$ converge γ_1, γ_0 -a.e. respectively. Because $\mathbb{P}_1, \mathbb{P}_0$ are marginals of γ_1, γ_0, this statement implies the result. □

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

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

$$\liminf_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\liminf_{n\to\infty} h_n) \tag{2.43}$$

and

$$\limsup_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\limsup_{n\to\infty} h_n) \tag{2.44}$$

See Appendix A.7.3 for a proof.

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

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

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

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

Thus

$$\limsup_{k\to\infty} S_\epsilon(h_0^{n_k}) = \liminf_{k\to\infty} S_\epsilon(h_0^{n_k}) \quad \mathbb{P}_0\text{-a.e.} \tag{2.45}$$

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

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

one can conclude that

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

Define

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

Now Fatou's lemma, Lemma 20, and Equation 2.45 imply that

$$\begin{aligned}
\lim_{k\to\infty} \Theta(h_0^{n_k}, h_1^{n_k}) &\ge \int \liminf_{k\to\infty} S_\epsilon(h_1^{n_k}) d\mathbb{P}_1 + \int \liminf_{k\to\infty} S_\epsilon(h_0^{n_k}) d\mathbb{P}_0 && \text{(Fatou's Lemma)} \\
&= \int \liminf_{k\to\infty} S_\epsilon(h_1^{n_k}) d\mathbb{P}_1 + \int \limsup_{k\to\infty} S_\epsilon(h_0^{n_k}) d\mathbb{P}_0 && \text{(Equation 2.45)} \\
&\ge \int S_\epsilon(\liminf_{k\to\infty} h_1^{n_k}) d\mathbb{P}_1 + \int S_\epsilon(\limsup_{k\to\infty} h_0^{n_k}) d\mathbb{P}_0 && \text{(Lemma 20)} \\
&= \int S_\epsilon(h_1^*) d\mathbb{P}_1 + \int S_\epsilon(h_0^*) d\mathbb{P}_0
\end{aligned}$$

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

□

2.7 Reducing Θ to R^{ϵ}_{ϕ}

Using the properties of $C^*_\psi(\eta)$, we showed in the previous section that there exist minimizers to Θ over the set S_ψ. The inequality $\eta h_1^* + (1-\eta^*)h_0^* \geq C^*_\psi(\eta)$ together with (2.31) imply that $h_1^*(\mathbf{x}) - h_0^*(\mathbf{x})$ is a supergradient of $C^*_\psi(\eta^*(\mathbf{x}))$ and thus $h_1^* - h_0^* = (C^*_\psi)'(\eta)$. This relation together with (2.31) provides two equations in two variables that can be uniquely solved for h_0^*, h_1^*, resulting in $h_0^* = \psi \circ -\alpha_\psi(\eta^*), h_1^* = \psi \circ \alpha_\psi(\eta^*)$.

Next, primal minimizers of Θ over S_ϕ for *any* ϕ will be constructed from the dual maximizers $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ of $\bar{R}_\psi$. Because $\alpha_\psi(\eta) = 1/2 \log(\eta/1-\eta)$ is a strictly increasing function, the compositions $\psi \circ \alpha_\psi, \psi \circ -\alpha_\psi$ are strictly monotonic. Thus the complementary slackness condition (2.30) applied to $h_1^* = \psi(\alpha_\psi(\eta^*)), h_0^* = \psi(-\alpha_\psi(\eta^*))$ implies that $\operatorname{supp} \mathbb{P}_1^*$ is contained in the set of points $\mathbf{x}'$ for which η^* assumes its infimum over some ϵ-ball at $\mathbf{x}'$ and $\operatorname{supp} \mathbb{P}_0^*$ is contained in the set of points $\mathbf{x}'$ where η^* assumes its supremum over some ϵ-ball at $\mathbf{x}'$. Thus, the functions $\phi \circ \alpha_\phi(\eta^*), \phi \circ -\alpha_\phi(\eta^*)$ satisfy (2.30) for the loss ϕ. The definition of α_ϕ further implies they satisfy (2.31). Therefore, Lemma 15 implies that for *any* ϕ, $h_1^* = \phi \circ \alpha_\phi(\eta^*)$, $h_0^* = \phi \circ \alpha_\phi(\eta^*)$ are primal optimal and $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ are dual optimal!

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

The next two lemmas discuss the analog of the ^{c} transform for the Kantorovich problem in optimal transport (see for instance Chapter 1 of [55] or Section 2.5 of [65]).

Lemma 22. *Assume that* $h_0, h_1 \geq 0$ *and* $(h_0(\mathbf{x}), h_1(\mathbf{x}))$ *satisfy* $\eta h_1 + (1-\eta)h_0 \geq C^*_\phi(\eta)$ *for*

all η. Then if we define $h_0^{C_\phi^*}$ via

$$h_0^{C_\phi^*} = \sup_{\eta \in [0,1)} \frac{C_\phi^*(\eta) - \eta h_1}{1 - \eta} \tag{2.46}$$

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

Proof. For convenience, set $\tilde{h}_0 = h_1^{C_\phi^*}$. Notice that $\tilde{h}_0 \geq 0$ because the right-hand side of (2.46) evaluates to 0 at $\eta = 0$. We will show that $\tilde{h}_0$ is Borel and that $(\tilde{h}_0, h_1)$ is a feasible pair.

Next, Notice that the map

$$G(\eta, \alpha) = \begin{cases} \frac{C_\phi^*(\eta) - \eta\alpha}{1-\eta} & \text{if } \eta < 1 \\ \lim_{\eta \to 1} \frac{C_\phi^*(\eta) - \eta\alpha}{1-\eta} & \text{if } \eta = 1 \end{cases} \tag{2.47}$$

is continuous in η. Thus, the supremum in (2.46) can be taken over the countable set $\mathbb{Q} \cap [0,1]$ and hence the function $\tilde{h}_0(\mathbf{x}) = \sup_{\eta \in [0,1) \cap \mathbb{Q}} G(\eta, h_1(\mathbf{x}))$ is Borel measurable. Because $G(\eta, h_1(\mathbf{x}))$ is continuous in η for each fixed $\mathbf{x}$, $G(\cdot, h_1(\mathbf{x}))$ assumes its maximum on $\eta \in [0,1]$ for each fixed $\mathbf{x}$. Thus there exists a function $\bar{\eta}(\mathbf{x})$ that maps $\mathbf{x}$ to a maximizer of $G(\cdot, h_1(\mathbf{x}))$. For this function $\bar{\eta}(\mathbf{x})$, one can conclude that $\tilde{h}_0(\mathbf{x}) = G(\bar{\eta}(\mathbf{x}), \mathbf{x})$ and hence

$$\bar{\eta}(\mathbf{x}) h_1(\mathbf{x}) + (1 - \bar{\eta}(\mathbf{x})) \tilde{h}_0(\mathbf{x}) = C_\phi^*(\bar{\eta}(\mathbf{x})). \tag{2.48}$$

Equation 2.48 implies that if $f(\mathbf{x}) < \tilde{h}_0(\mathbf{x})$ at any $\mathbf{x}$, then $\eta h_1(\mathbf{x}) + (1-\eta) f(\mathbf{x}) < C_\phi^*(\eta(\mathbf{x}))$ so (f, h_1) is not in the feasible set S_ϕ. Therefore, $\tilde{h}_0$ is the smallest function f for which $(f, h_1) \in S_\phi$. □

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

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

$$\hat{\eta}(\mathbf{x})h_1^*(\mathbf{x}) + (1 - \hat{\eta}(\mathbf{x}))h_0^*(\mathbf{x}) = C_\psi^*(\hat{\eta}(\mathbf{x})) \tag{2.49}$$

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

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

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

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

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

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

$$\operatorname{supp}\gamma_1 \subset \{(\mathbf{x}, \mathbf{x}') \colon \inf_{\|\mathbf{y}-\mathbf{x}\|\leq\epsilon} \hat{\eta}(\mathbf{y}) = \hat{\eta}(\mathbf{x}')\} \tag{2.50}$$

$$\operatorname{supp}\gamma_0 \subset \{(\mathbf{x}, \mathbf{x}') \colon \sup_{\|\mathbf{y}-\mathbf{x}\|\leq\epsilon} \hat{\eta}(\mathbf{y}) = \hat{\eta}(\mathbf{x}')\} \tag{2.51}$$

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

For convenience, in this proof, we introduce the notation

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

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

The complementary slackness condition (2.31) implies that $\eta^* h_1^* + (1-\eta^*)h_0^* = C_\psi^*(\eta^*)$ $\mathbb{P}^*$-a.e., and thus Lemma 18 implies that $h_1^* = \psi(\alpha_\psi(\eta^*))$ and $h_0^* = \psi(\alpha_\psi(\eta^*))$ $\mathbb{P}^*$-a.e. Therefore, $\psi(\alpha_\psi(\eta^*)) = \psi(\alpha_\psi(\hat{\eta}))$ $\mathbb{P}^*$-a.e. Now because the functions ψ, α_ψ are strictly monotonic, they are invertible. Thus it follows that $\hat{\eta} = \eta^*$ $\mathbb{P}^*$-a.e.

The complementary slackness condition (2.30) states that

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

Therefore,

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

which implies

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

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

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

□

Next, the relation (2.49) suggests that $h_1^* = \phi \circ f^*$, $h_0^* = \phi \circ -f^*$, where f^* is a function satisfying $C_\psi(\hat{\eta}(\mathbf{x}), f^*(\mathbf{x})) = C_\psi^*(\hat{\eta}(\mathbf{x}))$. In fact, Lemma 24 implies that this relation holds for *all* loss functions, and not just the exponential loss ψ. To formalize this idea, we prove the following result about minimizers of $C_\psi(\eta, \cdot)$ in Appendix A.3:

Lemma 25. *Fix a loss function ϕ and let $\alpha_\phi(\eta)$ be as in (2.8). Then α_ϕ maps η to the smallest minimizer of $C_\phi(\eta, \cdot)$. Furthermore, the function $\alpha_\phi(\eta)$ non-decreasing in η.*

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

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

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

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

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

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

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

This calculation verifies the complementary slackness condition (2.31).

We next verify the other complementary slackness condition (2.30). Let γ_i be a coupling between $\mathbb{P}_i, \mathbb{P}_i^*$ with $\operatorname{supp}\gamma_i \subset \Delta_\epsilon$. Next, because $\phi\circ\alpha_\phi$, $\phi\circ-\alpha_\phi$ are monotonic, the conditions (2.50) and (2.51) imply that

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

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

This calculation verifies the complementary slackness condition (2.30). □

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

2.8 Conclusion

We initiated the study of minimizers and minimax relations for adversarial surrogate risks. Specifically, we proved that there always exists a minimizer to the adversarial surrogate risk when perturbing in a closed ϵ-ball and a maximizer to the dual problem. Just like the results of [52], our minimax theorem provides an interpretation of the adversarial learning problem as a game between two players. This theory helps explain the phenomenon of transfer

attacks. We hope the insights gained from this research will assist in the development of algorithms for training classifiers robust to adversarial perturbations.

Acknowledgements

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

3 — The Uniqueness of the Adversarial Bayes Classifier

3.1 Introduction

A crucial reliability concern for machine learning models is their susceptibility to adversarial attacks. Neural nets are particularly sensitive to small perturbations to data. For instance, [14, 58] show that perturbations imperceptible to the human eye can cause a neural net to misclassify an image. In order to reduce the susceptibility of neural nets to such attacks, several methods have been proposed to minimize the *adversarial classification risk*, which incurs a penalty when a data point can be perturbed into the opposite class. However, state-of-the-art methods for minimizing this risk still achieve significantly lower accuracy than standard neural net training on simple datasets, even for small perturbations. For example, on the CIFAR10 dataset, [48] achieves 71% robust accuracy for ℓ_∞ perturbations size 8/255 while [21] achieves over 99% accuracy without an adversary.

In the setting of standard (non-adversarial) classification, a *Bayes classifier* is defined as a minimizer of the classification risk. This classifier simply predicts the most probable class at each point. If multiple classes have the same probability, then the Bayes classifier may not be unique. The Bayes classifier has been a helpful tool in the development of machine learning classification algorithms [32, Chapter 2.4]. On the other hand, in the adversarial setting,

computing minimizers of the adversarial classification risk in terms of the data distribution is a challenging problem. These minimizers are referred to as *adversarial Bayes classifiers*. Prior work [1, 11, 50] calculates these classifiers by first proving a minimax principle relating the adversarial risk with a dual problem, and then showing that the adversarial risk of a proposed set matches the dual risk of a point in the dual space.

In this paper, we propose a new notions of 'uniqueness' and 'equivalence' for adversarial Bayes classifiers in the setting of binary classification under the evasion attack. In the non-adversarial setting, two classifiers are *equivalent* if they are equal a.e. with respect to the data distribution, and one can show that any two equivalent classifiers have the same classification risk. The Bayes classifier is unique if any two minimizers of the classification risk are equivalent. However, under this notion of equivalence, two equivalent sets can have different adversarial classification risks. This discrepancy necessitates a new definition of equivalence for adversarial Bayes classifiers.

Further analyzing these new notions of uniqueness and equivalence in one dimension results in a method for calculating all possible adversarial Bayes classifiers for a well-motivated family of distributions. We apply this characterization to demonstrate that certain forms of regularity in adversarial Bayes classifiers improve as ϵ increases. Subsequent examples show that different adversarial Bayes classifiers achieve varying levels of (standard) classification risk. These examples illustrate that the accuracy-robustness tradeoff could be mitigated by a careful selection of an adversarial Bayes classifier (see [74] for a further discussion of this phenomenon). Followup work [24] demonstrates that the concepts presented in this paper have algorithmic implications— when the data distribution is absolutely continuous with respect to Lebesgue measure, adversarial training with a convex loss is adversarially consistent iff the adversarial Bayes classifier is unique, according to the new notion of uniqueness defined in this paper. Hopefully, a better understanding of adversarial Bayes classifiers will aid the design of algorithms for robust classification.

3.2 Background

3.2.1 Adversarial Bayes Classifiers

We study binary classification on the space $\mathbb{R}^d$ with labels $\{-1, +1\}$. The measure $\mathbb{P}_0$ describes the probability of data with label -1 occurring in regions of $\mathbb{R}^d$ while the measure $\mathbb{P}_1$ describes the probability of data with label $+1$ occurring in regions of $\mathbb{R}^d$. Most of our results will assume that $\mathbb{P}_0$ and $\mathbb{P}_1$ are absolutely continuous with respect to the Lebesgue measure μ. Vectors in $\mathbb{R}^d$ will be denoted in boldface ($\mathbf{x}$). Many of the results in this paper focus on the case $d = 1$ for which we will use non-bold letters (x). The functions p_0 and p_1 will denote the densities of $\mathbb{P}_0$, $\mathbb{P}_1$ respectively. A classifier is represented as the set of points A with label $+1$. The *classification risk* of the set A is then the proportion of incorrectly classified data:

$$R(A) = \int \mathbf{1}_{A^C} d\mathbb{P}_1 + \int \mathbf{1}_A d\mathbb{P}_0. \tag{3.1}$$

A minimizer of the classification risk is called a *Bayes classifier*. Analytically finding the minimal classification risk and Bayes classifiers is a straightforward calculation: Let $\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$, representing the total probability of a region, and let η be the the Radon-Nikodym derivative $\eta = d\mathbb{P}_1/d\mathbb{P}$, the conditional probability of the label $+1$ at a point $\mathbf{x}$. Thus one can re-write the classification risk is

$$R(f) = \int C(\eta(\mathbf{x}), f(\mathbf{x}))d\mathbb{P}(\mathbf{x}). \tag{3.2}$$

and the minimum classification risk as $\inf_f R(f) = \int C^*(\eta)d\mathbb{P}$ with

$$C(\eta, \alpha) = \eta\mathbf{1}_{\alpha\leq 0} + (1-\eta)\mathbf{1}_{\alpha>0}, \quad C^*(\eta) = \inf_\alpha C(\eta, \alpha). \tag{3.3}$$

The set $B = \{\mathbf{x}\colon \eta(\mathbf{x}) > 1 - \eta(\mathbf{x})\}$ is then a Bayes classifier. Note that the set of points

with $\eta(\mathbf{x}) = 1/2$ can be arbitrarily split between B and B^C. The Bayes classifier is *unique* if this ambiguous set has $\mathbb{P}$-measure zero. Equivalently, the Bayes classifier is unique if the value of $\mathbb{P}_0(B)$ or $\mathbb{P}_1(B^C)$ are the same for each Bayes classifier. When p_0 and p_1 are continuous, points in the boundary of the Bayes classifier must satisfy

$$p_1(\mathbf{x}) - p_0(\mathbf{x}) = 0 \tag{3.4}$$

A central goal of this paper is extending Equation (3.4) and a notion of uniqueness to adversarial classification.

In the adversarial scenario an adversary tries to perturb the data point $\mathbf{x}$ into the opposite class of a classifier A. We assume that perturbations are in a closed ϵ-ball $\overline{B_\epsilon(\mathbf{0})}$ in some norm $\|\cdot\|$. The proportion of incorrectly classified data under an adversarial attack is the *adversarial classification risk*,[1]

$$R^\epsilon(A) = \int S_\epsilon(\mathbf{1}_{A^C}) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0 \tag{3.5}$$

where the S_ϵ operation on a function g is defined as

$$S_\epsilon(g)(\mathbf{x}) = \sup_{\|\mathbf{h}\| \leq \epsilon} g(\mathbf{x} + \mathbf{h}). \tag{3.6}$$

Under this model, a set A incurs a penalty wherever $\mathbf{x} \in A \oplus \overline{B_\epsilon(\mathbf{0})}$, and thus we define the ϵ-*expansion* of a set A as

$$A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}.$$

[1] In order to define the adversarial classification risk, one must show that $S_\epsilon(\mathbf{1}_A)$ is measurable for measurable A. A full discussion of this issue is delayed to Section 3.5.2.

Hence the adversarial risk can also be written as

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

Prior work shows that there always exists minimizers to Equation (3.5), referred to as *adversarial Bayes classifiers* [2, 11, 26, 52], see [26, Theorem 1] for an existence theorem that matches the setup of this paper. Finding minimizers to Equation (3.5) is difficult because unlike the standard classification problem, one cannot write the integrand of Equation (3.5) so that it can be minimized in a pointwise manner. Furthermore, prior research [16] on the structure of minimizers to R^ϵ proves:

Lemma 27. *If A_1, A_2 are two adversarial Bayes classifiers, then so are $A_1 \cup A_2$ and $A_2 \cap A_1$.*

See Appendix B.1 for a proof.

Next, we focus on classifiers in one dimension as this case is simple to analyze yet still yields non-trivial behavior. Prior work shows that when $\mathbb{P}_0, \mathbb{P}_1 \ll \mu$ and p_0, p_1 are continuous, if the adversarial Bayes classifier is sufficiently 'regular', one can find necessary conditions describing the boundary of the adversarial Bayes classifier [64] . Assume that an adversarial Bayes classifier A can be expressed as a union of disjoint intervals $A = \bigcup_{i=m}^{M} (a_i, b_i)$, where the m, M, a_i, and b_i can be $\pm\infty$. Notice that one can arbitrarily include/exclude the endpoints $\{a_i\}$, $\{b_i\}$ without changing the value of the adversarial risk R^ϵ. If $b_i - a_i > 2\epsilon$ and $a_{i+1} - b_i > 2\epsilon$, the adversarial classification risk can then be expressed as:

$$R^\epsilon(A) = \cdots + \int_{b_{i-1}-\epsilon}^{a_i+\epsilon} p_1(x)dx + \int_{a_i-\epsilon}^{b_i+\epsilon} p_0(x)dx + \int_{b_i-\epsilon}^{a_{i+1}+\epsilon} p_1(x)dx + \cdots \tag{3.7}$$

When the densities p_0 and p_1 are continuous, differentiating this expression in a_i and b_i produces necessary conditions describing the adversarial Bayes classifier:

$$p_1(a_i + \epsilon) - p_0(a_i - \epsilon) = 0 \tag{3.8a}$$

$$p_0(b_i + \epsilon) - p_1(b_i - \epsilon) = 0 \tag{3.8b}$$

When $\epsilon = 0$, these equations reduce to the condition describing the boundary of the Bayes classifer in Equation (3.4). Prior work shows that when p_0, p_1 are well-behaved, this necessary condition holds for sufficiently small ϵ.

Theorem 28 ([64]). *Assume that p_0, p_1 are C^1, the relation $p_0(x) = p_1(x)$ is satisfied at finitely many points $x \in \operatorname{supp} \mathbb{P}$, and that at these points, $p_0'(x) \neq p_1'(x)$. Then for sufficiently small ϵ, there exists an adversarial Bayes classifier for which the a_i and b_i satisfy the necessary conditions Equation (3.8).*

For a proof, see the discussion of Equation (4.1) and Theorem 5.4 in [64]. A central goal of this paper is producing necessary conditions analogous to Equation (3.8) that hold for all ϵ.

3.2.2 Minimax Theorems for the Adversarial Classification Risk

We analyze the properties of adversarial Bayes classifiers by expressing the minimal R^ϵ risk in a 'pointwise' manner analogous to Equation (3.2). The Wasserstein-∞ metric from optimal transport and the minimax theorems in [26, 52] are essential tools for expressing R^ϵ in this manner.

Informally, the measure $\mathbb{Q}'$ is in the Wasserstein-∞ ball of radius ϵ around $\mathbb{Q}$ if one can produce the measure $\mathbb{Q}'$ by moving points in $\mathbb{R}^d$ by at most ϵ under the measure $\mathbb{Q}$. Formally, the W_∞ metric is defined in terms the set of couplings $\Pi(\mathbb{Q}, \mathbb{Q}')$ between two positive measures $\mathbb{Q}, \mathbb{Q}'$:

$$\Pi(\mathbb{Q}, \mathbb{Q}') = \{\gamma \text{ positive measure on } \mathbb{R}^d \times \mathbb{R}^d : \gamma(A \times \mathbb{R}^d) = \mathbb{Q}(A), \gamma(\mathbb{R}^d \times A) = \mathbb{Q}'(A)\}.$$

The Wasserstein-∞ distance between two positive finite measures $\mathbb{Q}'$ and $\mathbb{Q}$ with $\mathbb{Q}(\mathbb{R}^d) =$

$\mathbb{Q}'(\mathbb{R}^d)$ is then defined as

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

The W_∞ metric is in fact a metric on the space of measures, as it is a limit of the Wasserstein-p metrics as $p \to \infty$, see [17, 33] for details. We denote the ϵ-ball in the W_∞ metric around a measure $\mathbb{Q}$ by

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

Prior work [52, 63] applies properties of the W_∞ metric to find a dual problem to the minimization of R^ϵ: let $\mathbb{P}_0', \mathbb{P}_1'$ be finite Borel measures and define

$$\bar{R}(\mathbb{P}_0', \mathbb{P}_1') = \int C^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')} \right) d(\mathbb{P}_0' + \mathbb{P}_1') \tag{3.9}$$

where C^* is defined by (3.3). Prior results [26, 52] relate this risk to R^ϵ.

Theorem 29. *Let $\bar{R}$ be defined by (3.9). Then*

$$\inf_{A \text{ Borel}} R^\epsilon(A) = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') \tag{3.10}$$

and furthermore equality is attained for some Borel measurable A and $\mathbb{P}_1^*, \mathbb{P}_0^*$ with $W_\infty(\mathbb{P}_0^*, \mathbb{P}_0) \le \epsilon$ and $W_\infty(\mathbb{P}_1^*, \mathbb{P}_1) \le \epsilon$.

See Theorem 1 of [26] for the statement above. This minimax theorem then implies complementary slackness conditions that characterize optimal A and $\mathbb{P}_0^*, \mathbb{P}_1^*$. See Appendix B.2 for a proof.

Theorem 30. *The set A is a minimizer of R^ϵ and $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ is a maximizer of $\bar{R}$ over the W_∞ balls around $\mathbb{P}_0$ and $\mathbb{P}_1$ iff $W_\infty(\mathbb{P}_0^*, \mathbb{P}_0) \le \epsilon$, $W_\infty(\mathbb{P}_1^*, \mathbb{P}_1) \le \epsilon$, and*

1)

$$\int S_\epsilon(\mathbf{1}_{A^C})d\mathbb{P}_1 = \int \mathbf{1}_{A^C} d\mathbb{P}_1^* \quad \textit{and} \quad \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 = \int \mathbf{1}_A d\mathbb{P}_0^* \tag{3.11}$$

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

$$\eta^*(\mathbf{y})\mathbf{1}_{A^C} + (1 - \eta^*(\mathbf{y}))\mathbf{1}_A = C^*(\eta^*(\mathbf{y})) \quad \mathbb{P}^*\text{-}a.e. \tag{3.12}$$

3.3 Main Results

Definitions

As discussed in Section 3.2.1, a central goal of this paper is describing the regularity of adversarial Bayes classifiers and finding necessary conditions that hold for every ϵ in one dimension.

As an example of non-regularity, consider a data distribution defined by $p_0(x) = 1/5$, for $|x| \leq 1/4$ and zero elsewhere; and $p_1(x) = 3/5$ for $1 \geq |x| > 1/4$ and zero elsewhere (see Figure 3.2c for a depiction of p_0 and p_1). If $\epsilon = 1/8$, an adversarial Bayes classifier is $A = \mathbb{R}$. However, *any* subset S of $[-1/4 + \epsilon, 1/4 - \epsilon]$ satisfies $R^\epsilon(S^C) = R^\epsilon(\mathbb{R})$, and thus S^C is an adversarial Bayes classifier as well. (These claims are rigorously justified in Example 46.) Consequently there are many adversarial Bayes classifiers lacking regularity, but they all seem to be morally equivalent to the regular set $A = \mathbb{R}$. The notion of *equivalence up to degeneracy* encapsulates this behavior.

Definition 31. *Two adversarial Bayes classifiers* A_1 *and* A_2 *are* equivalent up to degeneracy *if for any Borel set* E *with* $A_1 \cap A_2 \subset E \subset A_1 \cup A_2$*, the set* E *is also an adversarial Bayes classifier. We say that that the adversarial Bayes classifier is* unique up to degeneracy *if any two adversarial Bayes classifiers are equivalent up to degeneracy.*

Due to Lemma 27, to verify that an adversarial Bayes classifier is unique up to degeneracy,

it suffices to show that if A_1 and A_3 are any two adversarial Bayes classifiers with $A_1 \subset A_3$, then any set satisfying $A_1 \subset E \subset A_3$ is an adversarial Bayes classifier as well. In the example presented above, the non-regular portion of the adversarial Bayes classifier could only be some subset of $D = [-1/4 + \epsilon, 1/4 - \epsilon]$. The notion of 'degenerate sets' formalizes this behavior.

Definition 32. *A set D is* degenerate *for an adversarial Bayes classifier A if for all Borel E with $A - D \subset E \subset A \cup D$, the set E is also an adversarial Bayes classifier.*

Equivalently, a set D is degenerate for A if for all disjoint subsets $D_1, D_2 \subset D$, the set $A \cup D_1 - D_2$ is also an adversarial Bayes classifier. In terms of this definition: the adversarial Bayes classifiers A_1 and A_2 are equivalent up to degeneracy iff the set $A_1 \triangle A_2$ is degenerate for either A_1 or A_2.

This paper first studies properties of these new notions, and then uses the resulting insights to characterize adversarial Bayes classifiers in one dimension. To start, we show that when $\mathbb{P} \ll \mu$, equivalence up to degeneracy is in fact an equivalence relation (Theorem 33) and furthermore, every adversarial Bayes classifier has a 'regular' representative when $d = 1$ (Theorem 35). The differentiation argument in Section 3.2.1 then produces necessary conditions characterizing regular adversarial Bayes classifiers in one dimension (Theorem 37). These conditions provide a tool for understanding how the adversarial Bayes classifier depends on ϵ; see Theorem 39 and Propositions 47 to 50. Identifying all adversarial Bayes classifiers then requires characterizing degenerate sets, and we provide such a criterion under specific assumptions. Lastly, Theorem 34 provides alternative criteria for equivalence up to degeneracy.

Theorem Statements

First, equivalence up to degeneracy is in fact an equivalence relation for many common distributions.

Theorem 33. *If $\mathbb{P} \ll \mu$, then equivalence up to degeneracy is an equivalence relation.*

Example 54 shows that the assumption $\mathbb{P} \ll \mu$ is necessary for this result. Additionally, uniqueness up to degeneracy generalizes certain notions of uniqueness for the Bayes classifier.

Theorem 34. *Assume that $\mathbb{P} \ll \mu$ and $\epsilon > 0$. Then the following are equivalent:*

A) The adversarial Bayes classifier is unique up to degeneracy

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

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

When $\epsilon = 0$, Item B) and Item C) are equivalent notions of uniqueness of the Bayes classifier (see Section 3.2.1). However, if B_1 and B_2 are Bayes classifiers, any set E satisfying $B_1 \cap B_2 \subset E \subset B_1 \cup B_2$ is always a Bayes classifier. Thus Item A) is not necessarily equivalent to Items B) and C) when $\epsilon = 0$. When $\mathbb{P} \not\ll \mu$, Theorem 33 is false although Item B) and Item C) are still equivalent (see Example 54 and Lemma 144). This equivalence suggests a different notion of uniqueness for such distributions, see the Section 3.5.1 for more details.

A central result of this paper is that degenerate sets are the only form of non-regularity possible in the adversarial Bayes classifier in one dimension.

Theorem 35. *Assume that $d = 1$ and $\mathbb{P}_0, \mathbb{P}_1 \ll \mu$. Then any adversarial Bayes classifier is equivalent up to degeneracy to an adversarial Bayes classifier $A' = \bigsqcup_{i=m}^{M}(a_i, b_i)$ with $b_i - a_i > 2\epsilon$ and $a_{i+1} - b_i > 2\epsilon$.*

This result motivates the definition of *regularity* in one dimension.

Definition 36. *We say $E \subset \mathbb{R}$ is a* regular set of radius ϵ *if one can write both E and E^C as a disjoint union of intervals of length strictly greater than 2ϵ.*

We will drop 'of radius ϵ' when clear from the context.

When p_0, p_1 are continuous, the necessary conditions Equation (3.8) always hold for a regular adversarial Bayes classifier.

Theorem 37. *Let $d = 1$ and assume that $\mathbb{P} \ll \mu$. Let $A = \bigcup_{i=m}^{M}(a_i, b_i)$ be a regular adversarial Bayes classifier.*

If p_0 is continuous at $a_i - \epsilon$ (resp. $b_i + \epsilon$) and p_1 is continuous at $a_i + \epsilon$ (resp. $b_i - \epsilon$), then a_i (resp. b_i) must satisfy the first order necessary conditions Equation (3.8a) (resp. Equation (3.8b)). Similarly, if p_0 is differentiable at $a_i - \epsilon$ (resp. $b_i + \epsilon$) and p_1 is differentiable at $a_i + \epsilon$ (resp. $b_i - \epsilon$), then a_i (resp. b_i) must satisfy the second order necessary conditions Equation (3.13a) (resp. Equation (3.13b)).

$$p_1'(a_i + \epsilon) - p_0'(a_i - \epsilon) \geq 0 \qquad (3.13a) \qquad p_0'(b_i + \epsilon) - p_1'(b_i - \epsilon) \geq 0 \qquad (3.13b)$$

This theorem provides a method for identifying a representative of every equivalence class of adversarial Bayes classifiers under equivalence up to degeneracy.

1) Let $\mathfrak{a}$, $\mathfrak{b}$ be the set of points that satisfy the necessary conditions for a_i, b_i respectively
2) Form all possible open regular sets $\bigcup_{i=m}^{M}(a_i, b_i)$ with $a_i \in \mathfrak{a}$ and $b_i \in \mathfrak{b}$.
3) Identify which of these sets would be be equivalent up to degeneracy, if they were adversarial Bayes classifiers.
4) Compare the risks of all non-equivalent sets from step 2) to identify which are adversarial Bayes classifiers.

One only need to consider open sets in step 2) because the boundary of a regular adversarial Bayes classifier is always a degenerate set when $\mathbb{P} \ll \mu$, as noted in Section 3.2.1 (see Lemma 160 for a formal proof). Section 3.4 applies this procedure above to several example distributions, see Example 41 for a crisp example. This analysis reveals interesting

patterns. First, boundary points of the adversarial Bayes classifier are frequently within ϵ of boundary points of the Bayes classifier. Proposition 49 and Proposition 50 prove that this phenomenon occurs when either $\mathbb{P}$ is a uniform distribution on an interval or $\eta \in \{0, 1\}$, and Proposition 47 shows that this occurrence can reduce the accuracy-robustness tradeoff. Second, uniqueness up to degeneracy often fails only for a small number of values of ϵ when $\mathbb{P}_0(\mathbb{R}) \neq \mathbb{P}_1(\mathbb{R})$. Understanding both of these occurrences in more detail is an open problem.

Theorem 37 is a tool for identifying a representative of each equivalence class of adversarial Bayes classifiers under equivalence up to degeneracy. Can one characterize all the members of a specific equivalence class? Answering this question requires understanding properties of degenerate sets.

Theorem 38. *Assume that $d = 1$, $\mathbb{P} \ll \mu$, and let A be an adversarial Bayes classifier.*

- *If some interval I is degenerate for A and I is contained in* $\operatorname{supp} \mathbb{P}$*, then $|I| \leq 2\epsilon$.*
- *Conversely, the connected components of A and A^C of length less than or equal to 2ϵ are contained in a degenerate set.*
- *A countable union of degenerate sets is degenerate.*
- *Assume that* $\operatorname{supp} \mathbb{P}$ *is an interval and $\mathbb{P}(\eta \in \{0, 1\}) = 0$. If D is a degenerate set for A, then D must be contained in the degenerate set $\overline{(\operatorname{supp} \mathbb{P}^\epsilon)^C} \cup \partial A$.*

The first two bullets state that within the support of $\mathbb{P}$, degenerate intervals must have length at most 2ϵ, and conversely a component of size at most 2ϵ must be degenerate. The last bullet implies that when $\operatorname{supp} \mathbb{P}$ is an interval and $\mathbb{P}(\eta \in \{0, 1\}) = 0$, the equivalence class of an adversarial Bayes classifier A consists of all Borel sets that differ from A by a measurable subset of $\overline{(\operatorname{supp} \mathbb{P}^\epsilon)^C} \cup \partial A$. Specifically, under these conditions, A cannot have a degenerate interval contained in $\operatorname{supp} \mathbb{P}^\epsilon$. This result is a helpful tool for identifying sets which are equivalent up to degeneracy in step 3) of the procedure above. Both of the assumptions

present in this fourth bullet are necessary— Example 46 presents a counterexample where $\operatorname{supp}\mathbb{P}$ is an interval and $\mathbb{P}(\eta \in \{0,1\}) > 0$ while Example 69 presents a counterexample for which $\mathbb{P}(\eta \in \{0,1\}) = 0$ but $\operatorname{supp}\mathbb{P}$ is not an interval.

Prior work [2, 16] shows that a certain form of regularity for adversarial Bayes classifiers improves as ϵ increases. Theorem 35 is an expression of this principle: this theorem states that each adversarial Bayes classifier A is equivalent to a regular set of radius ϵ, and thus the regularity guarantee improves as ϵ increases. Another form of regularity also improves as ϵ increases—the number of components of A and A^C must decrease for well-behaved distributions. Let $\operatorname{comp}(A) \in \mathbb{N} \cup \{\infty\}$ be the number of connected components of a set A.

Theorem 39. *Assume that $d = 1$, $\mathbb{P} \ll \mu$, $\operatorname{supp}\mathbb{P}$ is an interval I, and $\mathbb{P}(\eta \in \{0,1\}) = 0$. Let $\epsilon_2 > \epsilon_1$ and let A_1, A_2 be regular adversarial Bayes classifiers corresponding to perturbation radiuses ϵ_1 and ϵ_2 respectively. Then $\operatorname{comp}(A_1 \cap I^{\epsilon_1}) \geq \operatorname{comp}(A_2 \cap I^{\epsilon_2})$ and $\operatorname{comp}(A_1^C \cap I^{\epsilon_1}) \geq \operatorname{comp}(A_2^C \cap I^{\epsilon_2})$.*

Section 3.6.3 actually proves a stronger statement: typically, no component of $A_1 \cap I^{\epsilon_1}$ can contain a connected component of A_2^C and no component of $A_1^C \cap I^{\epsilon_1}$ can contain a connected component of A_2. Due to the fourth bullet of Theorem 38, the assumptions of Theorem 39 imply that there is no degenerate interval within $\operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon$, and hence every adversarial Bayes classifier is regular. When computing adversarial Bayes classifiers, Theorem 39 and the stronger version in Section 3.6.3 are useful tools in ruling out some of the sets in step 2) of the procudure above without explicitly computing their risk.

When $d > 1$, we show:

Theorem 40. *Let A be an adversarial Bayes classifier. Then A is equivalent up to degeneracy to a classifier A_1 for which $A_1 = C^\epsilon$ and a classifier A_2 for which $A_2^C = E^\epsilon$, for some sets C, E.*

Further understanding uniqueness up to degeneracy in higher dimension is an open ques-

tion.

Paper Outline

Section 3.4 applies the tools presented above to compute adversarial Bayes classifiers for a variety of distributions. Subsequently, Section 3.5 presents properties of equivalence up to degeneracy, including proofs of Theorems 33, 34 and 40. Sections 3.5.2 and 3.5.3 further study degenerate sets, and these results are later applied in Section 3.6.1 to prove Theorems 35 and 37. Section 3.6.2 further studies degenerate sets in one dimension to prove Theorem 38. Lastly, Section 3.6.3 proves Theorem 39. Technical proofs and calculations appear in the appendix, which is organized so that it can be read sequentially.

3.4 Examples

The examples below find the equivalence classes under equivalence up to degeneracy for any $\epsilon > 0$. Examples 42 and 46 demonstrate distributions for which the adversarial Bayes classifier is unique up to degeneracy for all ϵ while Example 45 demonstrates a distribution for which the adversarial Bayes classifier is not unique up to degeneracy for any $\epsilon > 0$, even though the Bayes classifier is unique. Example 41 and Example 44 describe intermediate situations— uniqueness up to degeneracy fails only for a single value of ϵ in Example 44 and only for sufficiently large ϵ in Example 41. Lastly, Example 46 presents an example with a degenerate set.

Examples 45 and 46 exhibit situations where different adversarial Bayes classifiers have varying levels of (standard) classification risk, for all ϵ contained in some interval. For such distributions, a deliberate selection of the adversarial Bayes classifier would mitigate the tradeoff between robustness and accuracy.

Furthermore, all the examples below except Example 42 exhibit a curious occurrence—

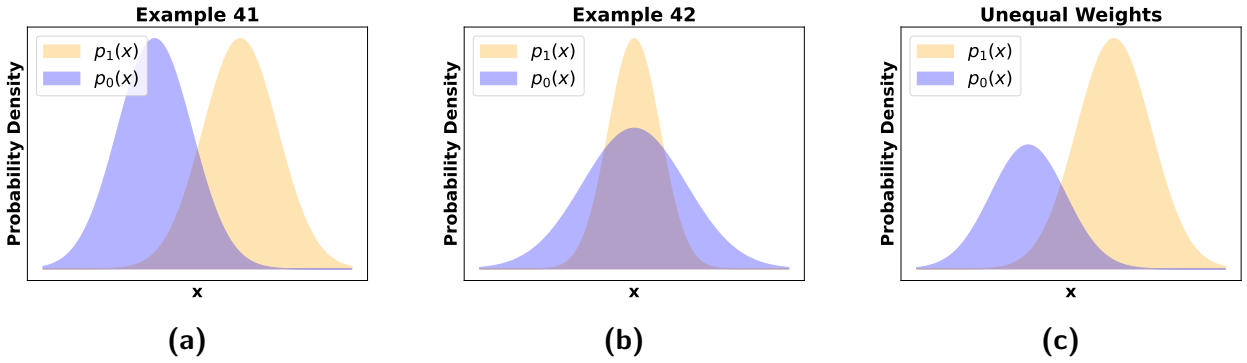

Figure 3.1: (a) Gaussian Mixture with equal means and unequal variances as a in Example 42. (b) Gaussian Mixture with equal weights, unequal means, and equal variances as in Example 41. (c) Gaussian Mixture with unequal weights, unequal means, and equal variances.

the boundary of the adversarial Bayes classifier is within ϵ of the boundary of the Bayes classifier. Propositions 49 and 50 state conditions under which this phenomenon must occur. Next, Proposition 47 shows that if furthermore the Bayes and adversarial Bayes have the same number of components, then one can bound the (standard) classification risk of the adversarial Bayes classifier in terms of the Bayes risk and ϵ, suggesting a reduced robustness-accuracy tradeoff.

The first two examples study Gaussian mixtures: $p_0 = (1-\lambda)g_{\mu_0,\sigma_0}(x)$, $p_1 = \lambda g_{\mu_1,\sigma_1}(x)$, where $\lambda \in (0,1)$ and $g_{\mu,\sigma}$ is the density of a gaussian with mean μ and variance σ^2. Prior work [50] calculates a single adversarial Bayes classifier for $\lambda = 1/2$ and any value of μ_i and σ_i. Below, our goal is to find *all* adversarial Bayes classifiers.

Example 41 (Gaussian Mixtures— equal variances, equal weights)**.** Consider a gaussian mixture with $p_0(x) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_0)^2/2\sigma^2}$, $p_1(x) = \frac{1}{2} \cdot \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_1)^2/\sigma^2}$, and $\mu_1 > \mu_0$, as depicted in Figure 3.1a. The solutions to the first order necessary conditions $p_1(b-\epsilon) - p_0(b+\epsilon) = 0$ and $p_1(a+\epsilon) - p_0(a-\epsilon) = 0$ from Equation (3.8) are

$$a(\epsilon) = b(\epsilon) = \frac{\mu_0 + \mu_1}{2}$$

However, one can show that $b(\epsilon)$ does not satisfy the second order necessary condition Equa-

tion (3.13b) (see Appendix B.11.1). Thus the candidate sets for the Bayes classifier are $\mathbb{R}$, $\emptyset$, and $(a(\epsilon), +\infty)$. The fourth bullet of Theorem 38 implies that none of these sets could be equivalent up to degeneracy. By comparing the adversarial risks of these three sets, one can show that the set $(a(\epsilon), +\infty)$ is an adversarial Bayes classifier iff $\epsilon \leq \frac{\mu_1 - \mu_0}{2}$ and $\mathbb{R}$, $\emptyset$ are adversarial Bayes classifiers iff $\epsilon \geq \frac{\mu_1 - \mu_0}{2}$ (see Appendix B.11.1 for details). Thus the adversarial Bayes classifier is unique up to degeneracy only when $\epsilon < \frac{\mu_1 - \mu_0}{2}$.

When $\epsilon \leq \frac{\mu_1 - \mu_0}{2}$, the set $(a(\epsilon), +\infty)$ is both a Bayes classifier and an adversarial Bayes classifier, and thus there is no accuracy-robustness tradeoff. In this example, uniqueness up to degeneracy fails for all sufficiently large ϵ. In contrast, the example below demonstrates a distribution for which the adversarial Bayes classifier is unique up to degeneracy for all ϵ.

Example 42 (Gaussian Mixtures— equal means)**.** Consider a Gaussian mixture with $p_0(x) = \frac{1-\lambda}{\sqrt{2\pi}\sigma_0} e^{-x^2/2\sigma_0^2}$ and $p_1(x) = \frac{\lambda}{\sqrt{2\pi}\sigma_1} e^{-x^2/2\sigma_1^2}$. Assume that p_0 has a larger variance than p_1 but that the peak of p_0 is below the peak of p_1, or other words, $\sigma_0 > \sigma_1$ but $\frac{\lambda}{\sigma_1} > \frac{1-\lambda}{\sigma_0}$, see Figure 3.1b for a depiction. Calculations similar to Example 41 show that the adversarial Bayes classifier is unique up to degeneracy for every ϵ, and is given by $(-b(\epsilon), b(\epsilon))$ where

$$b(\epsilon) = \frac{\epsilon\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) + \sqrt{\frac{4\epsilon^2}{\sigma_0^2\sigma_1^2} - 2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\ln\frac{(1-\lambda)\sigma_1}{\lambda\sigma_0}}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}}. \tag{3.14}$$

The computational details are similar to those of Example 41, and thus are delayed to Appendix B.11.2.

Unlike Example 41, the Bayes and adversarial Bayes classifiers can differ substantially.

The next three examples are distributions for which $\operatorname{supp}\mathbb{P}$ is a finite interval. In such situations, it is often helpful to assume that a_i, b_i are not near $\partial \operatorname{supp}\mathbb{P}$.

Lemma 43. *Consider a distribution for which $\mathbb{P} \ll \mu$ and $\operatorname{supp}\mathbb{P}$ is an interval. Then every adversarial Bayes classifier is equivalent up to degeneracy to a regular adversarial Bayes*

classifier $A = \bigcup_{i=m}^{M}(a_i, b_i)$ *for which the finite* a_i, b_i *are contained in* $\operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$.

See Appendix B.11.3 for a proof.

Example 44 (Uniqueness fails for a single value of ϵ)**.** Consider a distribution for which

$$p_0 = \begin{cases} \frac{1}{6}(1+x) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad p_1 = \begin{cases} \frac{1}{3}(1-x) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The only solutions to the first order necessary conditions $p_1(a+\epsilon) - p_0(a-\epsilon) = 0$ and $p_0(b+\epsilon) - p_1(b-\epsilon) = 0$ within $\operatorname{supp}\mathbb{P}^{\epsilon}$ are

$$a(\epsilon) = \frac{1}{3}(1-\epsilon) \quad \text{and} \quad b(\epsilon) = \frac{1}{3}(1+\epsilon)$$

We first consider ϵ small enough so that both of these points lie in $\operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$, or in other words, $\epsilon < 1/2$. Then $p_0'(a(\epsilon)-\epsilon) = p_0'(b(\epsilon)+\epsilon) = 1/6$ and $p_1'(a(\epsilon)+\epsilon) = p_1'(b(\epsilon)-\epsilon) = -1/3$. Consequently, the point $a(\epsilon)$ fails to satisfy the second order necessary condition Equation (3.13a). To identify all adversarial Bayes classifiers under uniqueness up to degeneracy for $\epsilon < 1/2$, Lemma 43 imply it remains to compare the adversarial risks of $\emptyset$, $\mathbb{R}$, and $(-\infty, b(\epsilon))$. Theorem 38 implies that none of these sets could be equivalent up to degeneracy. The risks of these sets compute to $R^{\epsilon}(\emptyset) = 2/3$, $R^{\epsilon}(\mathbb{R}) = 1/3$, and $R^{\epsilon}((-\infty, b(\epsilon))) = \frac{2}{9}(1+\epsilon)^2$. Therefore, for all $\epsilon < 1/2$, the set $(-\infty, b(\epsilon))$ is an adversarial Bayes classifier iff $\epsilon \leq \sqrt{3/2} - 1$ while $\mathbb{R}$ is an adversarial Bayes classifier iff $\epsilon \geq \sqrt{3/2} - 1$. Theorem 39 then implies that this last statement holds without the restriction $\epsilon < 1/2$.

Uniqueness up to degeneracy fails for only a single value of ϵ in the example above. In contrast, uniqueness up to degeneracy fails for all ϵ in the distribution below.

Example 45 (Non-uniqueness for all $\epsilon > 0$)**.** Let p be the uniform density on the interval

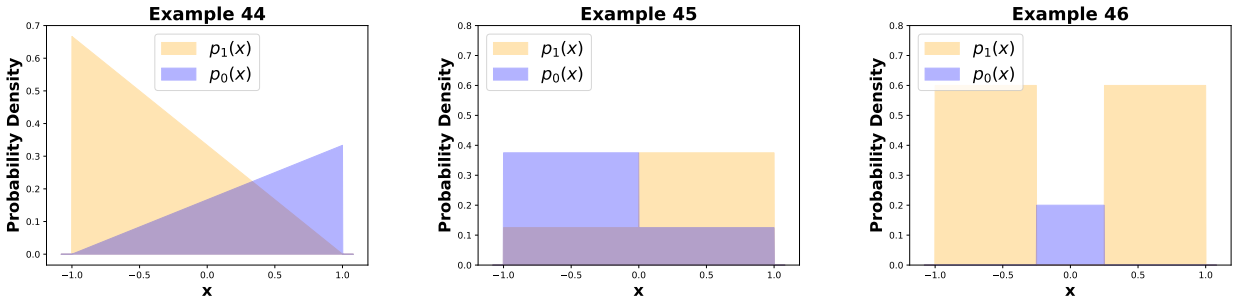

Figure 3.2: (a) The distribution of Example 44. (b) The distribution of Example 45. (c) The distribution of Example 46.

$[-1, 1]$ and let

$$\eta(x) = \begin{cases} \frac{1}{4} & \text{if } x \leq 0 \\ \frac{3}{4} & \text{if } x > 0 \end{cases}$$

Calculations for this example are similar to those in Example 44, so we delay the details to Appendix B.11.4. For this distribution, the set (y, ∞) is an adversarial Bayes classifier for any $y \in [-\epsilon, \epsilon]$ iff $\epsilon \leq 1/3$ and $\emptyset, \mathbb{R}$ are adversarial Bayes classifiers iff $\epsilon \geq 1/3$. Theorem 38 implies that none of these sets could be equivalent up to degeneracy. Therefore, the adversarial Bayes classifier is not unique up to degeneracy for all $\epsilon > 0$ even though the Bayes classifier is unique.

Again, the adversarial Bayes classifier $(0, \infty)$ is also a Bayes classifier when $\epsilon \leq 1/3$, and thus there is no accuracy-robustness tradeoff for this distribution.

A distribution is said to satisfy *Massart's noise condition* if $|\eta(\mathbf{x}) - 1/2| \geq \delta$ $\mathbb{P}$-a.e. for some $\delta > 0$. Prior work [41] relates this condition to the sample complexity of learning from a function class. For the example above, Theorem 34 implies that Massart's noise condition cannot hold for any maximizer of $\bar{R}$ even though $|\eta - 1/2| \geq 1/4$ $\mathbb{P}$-a.e.

The next example exhibits a degenerate set that has positive measure under $\mathbb{P}$.

Example 46 (Example of a degenerate set)**.** Consider a distribution on $[-1, 1]$ with

$$p_0(x) = \begin{cases} \frac{1}{5} & \text{if } |x| \leq 1/4 \\ 0 & \text{otherwise} \end{cases} \qquad p_1(x) = \begin{cases} \frac{3}{5} & \text{if } 1 \geq |x| > 1/4 \\ 0 & \text{otherwise} \end{cases}$$

Theorem 37 and Lemma 43 imply that one only need consider $a_i, b_i \in \{-\frac{1}{4} \pm \epsilon, \frac{1}{4} \pm \epsilon\}$ when identifying a regular representative of each equivalence class of adversarial Bayes classifiers. By comparing the adversarial risks of the regular sets satisfying this crieterion, one can show that when $\epsilon \leq 1/8$, every adversarial Bayes classifier is equivalent up to degeneracy to the regular set $(-\infty, -1/4 + \epsilon) \cup (1/4 - \epsilon, \infty)$ but when $\epsilon \geq 1/8$ then every adversarial Bayes classifier is equivalent up to degeneracy to the regular set $\mathbb{R}$ (see Appendix B.11.5 for details.)

Next consider $\epsilon \in [1/8, 1/4]$. If S is an arbitrary subset of $[-1/4 + \epsilon, 1/4 - \epsilon]$, then $R^\epsilon(\mathbb{R}) = R^\epsilon(S^C)$. Thus the interval $[-1/4 + \epsilon, 1/4 - \epsilon]$ is a degenerate set.

When $\epsilon \in [1/8, 1/4]$, the (standard) classification error of $\mathbb{R}$ and $(-\infty, -1/4+\epsilon) \cup (1/4-\epsilon)$ differ by $\frac{2}{5}(1 - 4\epsilon)$, which is close to $1/5$ for ϵ near $1/8$. Thus a careful selection of the adversarial Bayes classifier can mitigate the accuracy-robustness tradeoff.

The last three propositions in this section specify conditions under which one could hope that the boundary of the adversarial Bayes classifier would be within ϵ of the boundary of the Bayes classifier. If in addition the Bayes and adversarial Bayes classifiers have the same number of components, one can bound the minimal adversarial Bayes error in terms of the Bayes error rate and ϵ.

Proposition 47. *Let $B = \bigcup_{i=1}^M (c_i, d_i)$, $A = \bigcup_{i=1}^M (a_i, b_i)$ be Bayes and adversarial Bayes classifiers respectively. Assume that p_0, p_1 are bounded above by K. Then if $|a_i - c_i| \leq \epsilon$ and $|b_i - d_i| \leq \epsilon$, then $R(A) - R(B) \leq 4\epsilon MK$.*

Thus there will be a minimal robustness-accuracy tradeoff so long as $\epsilon \ll 1/MK$.

Proof.

$$R(A) - R(B) \leq \sum_{i=1}^{M} \int_{\min(a_i,c_i)}^{\max(a_i,c_i)} |p_1(x) - p_0(x)|dx + \int_{\min(b_i,d_i)}^{\max(b_i,d_i)} |p_1(x) - p_0(x)|dx \leq 4\epsilon MK$$

□

The next proposition stipulates a widely applicable criterion under which there is always a solution to the necessary conditions $p_1(a+\epsilon) - p_0(a-\epsilon) = 0$ and $p_1(b-\epsilon) - p_0(b+\epsilon) = 0$ within ϵ of solutions to $p_1(x) = p_0(x)$ (which specifies the boundary of the Bayes classifier).

Proposition 48. *Let z be a point with $p_1(z) - p_0(z) = 0$ and assume that p_0 and p_1 are continuous on $[z-r, z+r]$ for some $r > 0$. Furthermore, assume that one of p_0, p_1 is non-increasing and the other is non-decreasing on $[z-r, z+r]$. Then for all $\epsilon \leq r/2$ there is a solution to the first order necessary conditions Equation (3.8a) and Equation (3.8b) within ϵ of z.*

Proof. Without loss of generality, assume that p_1 is non-increasing and p_0 is non-decreasing on $[z-r, z+r]$. The applying the relation $p_1(z) = p_0(z)$, one can conclude that

$$p_1((z-\epsilon)+\epsilon) - p_0((z-\epsilon)-\epsilon) = p_1(z) - p_0(z-2\epsilon) = p_0(z) - p_0(z-2\epsilon) \geq 0.$$

An analogous argument shows that $p_1((z+\epsilon)+\epsilon) - p_0((z+\epsilon)-\epsilon) \leq 0$. Thus the intermediate value theorem implies that there is a solution to Equation (3.8a) within ϵ of z. Analogous reasoning shows that there is a solution to Equation (3.8b) within ϵ of z. □

However, this proposition does not guarantee that the solution to the necessary conditions within ϵ of z *must* be a boundary point of the adversarial Bayes classifier. To illustrate the utility of this result, consider a gaussian mixture with $p_1(x) = \frac{\lambda}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$, $p_0(x) = \frac{1-\lambda}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}$ for which $p_1(\mu_1) > p_0(\mu_1)$ and $p_0(\mu_0) > p_1(\mu_0)$, see Figure 3.1c) for an illustration. Just as in Example 41, the necessary conditions Equation (3.8) reduce to linear

equations and so there is at most one $a(\epsilon)$ solving Equation (3.8a) and at most one $b(\epsilon)$ solving Equation (3.8b). Thus Proposition 48 implies that the solutions to the first order necessary conditions Equation (3.8) must be within ϵ of the boundary of the Bayes classifier.

Next, if $\mathbb{P}$ is the uniform distribution on an interval, an argument similar to the proof of Proposition 48 implies that solutions to the first order necessary conditions Equation (3.8) are within ϵ of solutions to $p_0(z) = p_1(z)$.

Proposition 49. *Assume that $\mathbb{P}$ is the uniform distribution on an finite interval, p and η are continuous on $\operatorname{supp}\mathbb{P}$, and $\eta(x) = 1/2$ only at finitely many points within $\operatorname{supp}\mathbb{P}$. Then any adversarial Bayes classifier is equivalent up to degeneracy to an adversarial Bayes classifier $A = \bigcup_{i=m}^{M}(a_i, b_i)$ for which each a_i, b_i is at most ϵ from some point z satisfying $\eta(z) = 1/2$.*

The proof is very similar to that of Proposition 48, see Appendix B.11.6 for details.

Finally, under fairly general conditions, when $\eta \in \{0, 1\}$, the boundary of the adversarial Bayes classifier must be within ϵ of the boundary of the Bayes classifier.

Proposition 50. *Assume that $\operatorname{supp}\mathbb{P}$ is an interval $\mathbb{P} \ll \mu$, $\eta \in \{0, 1\}$, and p is continuous on $\operatorname{supp}\mathbb{P}$ and strictly positive. Then any adversarial Bayes classifier is equivalent up to degeneracy to a regular adversarial Bayes classifier $A = \bigcup_{i=m}^{M}(a_i, b_i)$ for which each a_i, b_i is at most ϵ from $\partial\{\eta = 1\}$.*

Again, the proof is very similar to that of Proposition 48, see Appendix B.11.7 for details.

3.5 Equivalence up to Degeneracy

3.5.1 Equivalence up to Degeneracy as an Equivalence Relation

When $\mathbb{P} \ll \mu$, there are several useful characterizations of equivalence up to degeneracy.

Proposition 51. *Let $\mathbb{P} \ll \mu$ and $\epsilon > 0$. Let $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ be a maximizer of $\bar{R}$ and set $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$. Let A_1 and A_2 be adversarial Bayes classifiers. Then the following are equivalent:*

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

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

3) $\mathbb{P}^*(A_2 \triangle A_1) = 0$

Item 2) states that A_1, A_2 are equivalent up to degeneracy if the 'attacked' classifiers $A_1^\epsilon, A_2^\epsilon$ are equal $\mathbb{P}_0$-a.e. Item 3) further states that the adversarial Bayes classifiers A_1, A_2 are unique up to degeneracy if they are equal under the measure of optimal adversarial attacks.

Proposition 51 is proved in Appendix B.3.2, and we provide an overview of this argument below. In this proof, we show that Item 3) is equivalent to Item 2), Item 2) implies Item 1), and Item 1) implies Item 3). First, the complementary slackness conditions of Theorem 30 implies that Item 2) and Item 3) equivalent, (see the proof of Lemma 144 in Appendix B.3). To show that Item 2) implies Item 1), we prove that Item 2) implies $S_\epsilon(\mathbf{1}_{A_1 \cup A_2}) = S_\epsilon(\mathbf{1}_{A_1 \cap A_2})$ $\mathbb{P}_0$-a.e. and $S_\epsilon(\mathbf{1}_{(A_1 \cup A_2)^C}) = S_\epsilon(\mathbf{1}_{(A_1 \cap A_2)^C}$ $\mathbb{P}_1$-a.e. (Lemma 142). Consequently, any two sets between $A_1 \cap A_2$ and $A_1 \cup A_2$ must have the same adversarial risk.

Lastly, to show that Item 1) implies Item 3), we apply the complementary slackness condition of Equation (3.11) of Theorem 30 to argue that $D = A_1 \triangle A_2$ has $\mathbb{P}^*$-measure zero. First, we show that if $D_1 = \operatorname{int} D \cap \mathbb{Q}^d$, $D_2 = \operatorname{int} D \cap (\mathbb{Q}^d)^C$ and $\epsilon > 0$, then $D_1^\epsilon = D_2^\epsilon = (\operatorname{int} D)^\epsilon$ (see Lemma 145). Thus $\int \mathbf{1}_{(A_1 \cap A_2 \cup D_1)^\epsilon} d\mathbb{P}_0 = \int \mathbf{1}_{(A_1 \cap A_2 \cup D_2)^\epsilon} d\mathbb{P}_0 = \int \mathbf{1}_{(A_1 \cap A_2 \cup \operatorname{int} D)^\epsilon} d\mathbb{P}_0$ and the complementary slackness condition Equation (3.11) implies that $\mathbb{P}_0^*(\operatorname{int} D) = 0$. Similarly, one can argue that $\mathbb{P}_1^*(\operatorname{int} D) = 0$. The assumption $\epsilon > 0$ is essential for this step of the proof. Next, to prove $\mathbb{P}^*(D \cap \partial D) = 0$, we prove the boundary ∂A is always a degenerate set for an adversarial Bayes classifier A when $\mathbb{P} \ll \mu$ and $\epsilon > 0$. Consequently:

Lemma 52. *Let A be an adversarial Bayes classifier. If $\mathbb{P} \ll \mu$ and $\epsilon > 0$, then A, $\overline{A}$, and* $\operatorname{int} A$ *are all equivalent up to degeneracy.*

See Appendix B.3.1 for a proof. Again, the assumption $\epsilon > 0$ is essential for this step of the proof.

Proposition 51 has several useful consequences for understanding degenerate sets, which we explore in Section 3.5.3. Specifically, when $\mathbb{P} \ll \mu$, equivalence up to degeneracy is in fact an equivalence relation.

Proof of Theorem 33. Item 3) of Proposition 51 states that two adversarial Bayes classifiers A_1, A_2 are equivalent up to degeneracy iff $\mathbf{1}_{A_1} = \mathbf{1}_{A_2}$ $\mathbb{P}^*$-a.e. Equality of functions $\mathbb{P}^*$-a.e. is an equivalence relation and consequently equivalence up to degeneracy is an equivalence relation. □

Furthermore, Proposition 51 implies Theorem 34. Item 2) of Proposition 51 is equivalent to Item B) of Theorem 34 when the adversarial Bayes classifier is unique up to degeneracy. In the following discussion, we assume that the adversarial Bayes classifier is unique up to degeneracy and show that Item 3) of Proposition 51 is equivalent to Item C) of Theorem 34.

First, to show Item C) $\Rightarrow$ Item 3), notice that the complementary slackness condition in Equation (3.12) implies that

$$\mathbf{1}_{\eta^* > 1/2} \leq \mathbf{1}_A \leq \mathbf{1}_{\eta^* \geq 1/2} \quad \mathbb{P}^*\text{-a.e.} \tag{3.15}$$

for any adversarial Bayes classifier A and any maximizer $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ of $\bar{R}$. Thus, if $\mathbb{P}^*(\eta^* = 1/2) = 0$ then every adversarial Bayes classifier must satisfy $\mathbf{1}_A = \mathbf{1}_{\eta^* > 1/2}$ $\mathbb{P}^*$-a.e. and thus $\mathbb{P}^*(A_1 \triangle A_2) = 0$ for any two adversarial Bayes classifiers A_1 and A_2.

It remains to show that Item 3) implies Item C). To relate the quantity $\mathbb{P}^*(A_1 \triangle A_2)$ to η^*, we show that there are adversarial Bayes classifiers $\hat{A}_1$, $\hat{A}_2$ which match $\{\eta^* > 1/2\}$ and $\{\eta^* \geq 1/2\}$ $\mathbb{P}^*$-a.e.

Lemma 53. *There exists* $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ *which maximize* $\bar{R}$ *and adversarial Bayes classifiers* $\hat{A}_1$, $\hat{A}_2$ *for which* $\mathbf{1}_{\eta^*>1/2} = \mathbf{1}_{\hat{A}_1}$ $\mathbb{P}^*$*-a.e. and* $\mathbf{1}_{\eta^*\geq 1/2} = \mathbf{1}_{\hat{A}_2}$ $\mathbb{P}^*$*-a.e., where* $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ *and* $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$.

Item B) in conjunction with this lemma implies that $0 = \mathbb{P}^*(\hat{A}_2 \triangle \hat{A}_1) = \mathbb{P}^*(\eta^* = 1/2)$ for the $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ in Lemma 53. See Appendix B.4 for proofs of Theorem 34 and Lemma 53. The classifiers $\hat{A}_1$ and $\hat{A}_2$ can be interpreted as *minimal* and *maximal* adversarial Bayes classifiers, in the sense that $\int S_\epsilon(\mathbf{1}_{\hat{A}_1})d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_{\hat{A}_2})d\mathbb{P}_0$ and $\int S_\epsilon(\mathbf{1}_{\hat{A}_1^C})d\mathbb{P}_1 \geq \int S_\epsilon(\mathbf{1}_{A^C})d\mathbb{P}_1 \geq \int S_\epsilon(\mathbf{1}_{\hat{A}_2^C})d\mathbb{P}_1$ for any adversarial Bayes classifier A (see Lemma 148 in Appendix B.4.1).

Theorem 33 is false when $\mathbb{P}$ is not absolutely continuous with respect to μ:

Example 54. Consider a distribution defined by $\mathbb{P}_0 = \frac{1}{2}\delta_{-\epsilon}$ and $\mathbb{P}_1 = \frac{1}{2}\delta_\epsilon$. If $0 \in A$ then $S_\epsilon(\mathbf{1}_A)(\epsilon) = 1$ and if $0 \notin A$ then $S_\epsilon(\mathbf{1}_{A^C})(-\epsilon) = 1$. In either case, $R^\epsilon(A) \geq 1/2$. The classifier $A = \mathbb{R}$ achieves adversarial classification error $1/2$ and therefore the adversarial Bayes risk is $R_*^\epsilon = 1/2$. The sets $\mathbb{R}^{\geq 0}$ and $\mathbb{R}^{>0}$ also achieve error $1/2$ and thus are also adversarial Bayes classifiers. These two classifiers are equivalent up to degeneracy because they differ by one point. Furthermore, the classifiers $\mathbb{R}$ and $\mathbb{R}^{\geq 0}$ are equivalent up to degeneracy: if $D \subset \mathbb{R}^{<0}$, then $S_\epsilon(\mathbf{1}_{\mathbb{R}^{\geq 0}\cup D})(-\epsilon) = 1$ while $S_\epsilon(\mathbf{1}_{(\mathbb{R}^{\geq 0}\cup D)^C})(\epsilon) = 0$ and hence $R^\epsilon(\mathbb{R}^{\geq 0}\cup D) = 1/2$. However, if $D \subset (-2\epsilon, 0)$ then $R^\epsilon(\mathbb{R}^{>0} \cup D) = 1$ and thus $\mathbb{R}$ and $\mathbb{R}^{>0}$ cannot be equivalent up to degeneracy.

In short— the classifiers $\mathbb{R}^{>0}$ and $\mathbb{R}^{\geq 0}$ are equivalent up to degeneracy, the classifiers $\mathbb{R}^{\geq 0}$ and $\mathbb{R}$ are equivalent up to degeneracy, but $\mathbb{R}^{>0}$ and $\mathbb{R}$ are not equivalent up to degeneracy. Thus equivalence up to degeneracy cannot be an equivalence relation for this distribution.

However, Item 2) and Item 3) of Proposition 51 are still equivalent when $\mathbb{P} \not\ll \mu$, as are Item B) and Item C) of Theorem 34 (see Lemma 144 and Proposition 149 in Appendices B.3.2 and B.4.2 respectively.). As the proof of Theorem 33 relies only on Item 3), one could use

Item 2) and Item 3) to define a notion of equivalence for adversarial Bayes classifiers even when $\mathbb{P} \not\ll \mu$.

3.5.2 The Universal σ-algebra, Measurability, and Fundamental Regularity Results

We introduce another piece of notation to state our regularity results. Define $A^{-\epsilon} = ((A^C)^\epsilon)^C$. The set A^ϵ represents all points in $\mathbb{R}^d$ that can be moved into A by a perturbation of size at most ϵ and $A^{-\epsilon}$ is the set of of points inside A that cannot be perturbed outside of A:

$$A^\epsilon = \{\mathbf{x} \colon \overline{B_\epsilon(\mathbf{x})} \text{ intersects } A\} \qquad (3.16) \qquad A^{-\epsilon} = \{\mathbf{x} : \overline{B_\epsilon(\mathbf{x})} \subset A\} \qquad (3.17)$$

See Appendix B.5 for a proof. Prior works [2, 16] note that applying the ϵ, $-\epsilon$ operations in succession can improve the regularity of an adversarial Bayes classifier and reduce the adversarial Bayes risk. Specifically:

Lemma 55. *For any set A, $R^\epsilon((A^{-\epsilon})^\epsilon) \leq R^\epsilon(A)$ and $R^\epsilon((A^\epsilon)^{-\epsilon}) \leq R^\epsilon(A)$.*

See Appendix B.5 for a proof. Thus applying the $^\epsilon$ and $^{-\epsilon}$ operations in succession can only reduce the adversarial risk of a set. In order to perform these regularizing operations, one must minimize R^ϵ over a σ-algebra Σ that is preserved by the $^\epsilon$ operation: in other words, one would wish that $A \in \Sigma$ implies $A^\epsilon \in \Sigma$.

To illustrate this concern, [52] demonstrate a Borel set C for which C^ϵ is not Borel measurable. However, prior work shows that if A is Borel, then A^ϵ is measurable with respect to a larger σ-algebra called the *universal σ-algebra* $\mathscr{U}(\mathbb{R}^d)$. A set in the universal σ-algebra is referred to as *universally measurable*. Theorem 29 of [25] states that

Theorem 56. *If A is universally measurable, then A^ϵ is as well.*

See Appendix B.6 for the definition of the universal σ-algebra $\mathscr{U}(\mathbb{R}^d)$.

Thus, in order to guarantee the existence of minimizers to R^ϵ with improved regularity properties, one could minimize R^ϵ over the universal σ-algebra $\mathscr{U}(\mathbb{R}^d)$. However, many prior papers such as [26, 50, 52] study this minimization problem over the Borel σ-algebra. We show that these two approaches are equivalent:

Theorem 57. *Let $\mathcal{B}(\mathbb{R}^d)$ denote the Borel σ algebra on $\mathbb{R}^d$. Then*

$$\inf_{A\in\mathcal{B}(\mathbb{R}^d)} R^\epsilon(A) = \inf_{A\in\mathscr{U}(\mathbb{R}^d)} R^\epsilon(A)$$

See Appendix B.6 for a proof. Due to this result, in the remainder of the paper, we treat the minimization of R^ϵ over $\mathscr{U}(\mathbb{R}^d)$ and $\mathcal{B}(\mathbb{R}^d)$ as interchangable.

3.5.3 Describing Degenerate Sets and Proof of Theorem 40

Proposition 51 together with fundamental properties of the $^\epsilon$ and $^{-\epsilon}$ operations imply several results on degenerate sets.

First, Lemma 27 implies that countable unions and intersections of adversarial Bayes classifiers are adversarial Bayes classifiers. Item 3) of Proposition 51 then necessitates that countable unions and intersections preserve equivalence up to degeneracy. As a result:

Lemma 58. *Let $\mathbb{P} \ll \mu$. Then a countable union of degenerate sets is degenerate.*

See Appendix B.7.1 for a formal proof.

Next, using the regularizing $^\epsilon$ and $^{-\epsilon}$ operations, we study the relation between uniqueness up to degeneracy and regular adversarial Bayes classifiers. First notice that $(A^{-\epsilon})^\epsilon$ is smaller than A while $(A^\epsilon)^{-\epsilon}$ is larger than A:

Lemma 59. *Let A be any set. Then $(A^{-\epsilon})^\epsilon \subset A \subset (A^\epsilon)^{-\epsilon}$.*

Furthermore, one can compare $S_\epsilon(\mathbf{1}_A)$ with $S_\epsilon(\mathbf{1}_{(A^{-\epsilon})^\epsilon})$ and $S_\epsilon(\mathbf{1}_{A^C})$ with $S_\epsilon(\mathbf{1}_{(A^\epsilon)^{-\epsilon}})$:

Lemma 60. *For any set $A \subset \mathbb{R}^d$, the following set relations hold:* $((A^{\epsilon})^{-\epsilon})^{\epsilon} = A^{\epsilon}$, $((A^{\epsilon})^{-\epsilon})^{-\epsilon} \supset A^{-\epsilon}$, $((A^{-\epsilon})^{\epsilon})^{-\epsilon} = A^{-\epsilon}$, $((A^{-\epsilon})^{\epsilon})^{\epsilon} \subset A^{\epsilon}$.

See Appendix B.5 for proofs of Lemma 59 and Lemma 60. Lemma 60 then implies:

Corollary 61. *Assume $\mathbb{P} \ll \mu$ and let A be an adversarial Bayes classifier. Then A, $(A^{\epsilon})^{-\epsilon}$, and $(A^{-\epsilon})^{\epsilon}$ are all equivalent up to degeneracy.*

Proof. Lemma 60 implies that $(A^{-\epsilon})^{\epsilon}$, $(A^{\epsilon})^{-\epsilon}$ are both adversarial Bayes classifiers satisfying $S_{\epsilon}(\mathbf{1}_A) = S_{\epsilon}(\mathbf{1}_{(A^{\epsilon})^{-\epsilon}})$ and $S_{\epsilon}(\mathbf{1}_{A^C}) = S_{\epsilon}(\mathbf{1}_{((A^{-\epsilon})^{\epsilon})^C})$. Therefore, when $\mathbb{P} \ll \mu$, Item 2) of Proposition 51 implies that A, $(A^{-\epsilon})^{\epsilon})$, and $(A^{\epsilon})^{-\epsilon}$ are all equivalent up to degeneracy. □

Theorem 40 is an immediate consequence of Corollary 61. Furthermore, Corollary 61 implies that "small" components of A and A^C are degenerate sets. Specifically, one can show that if C is a component with $C^{-\epsilon} = \emptyset$, then C is contained in $(A^{\epsilon})^{-\epsilon} - (A^{-\epsilon})^{\epsilon}$.

Proposition 62. *Let A be an adversarial Bayes classifier and let C be a connected component of A or A^C with $C^{-\epsilon} = \emptyset$. Then C is contained in $(A^{\epsilon})^{-\epsilon} - (A^{-\epsilon})^{\epsilon}$, and thus the set*

$$\bigcup \left\{ C : \textit{connected components of } A \textit{ or } A^C \textit{ with } C^{-\epsilon} = \emptyset \right\} \tag{3.18}$$

is contained in a degenerate set of A.

See Appendix B.7.2 for a proof. This result has a sort of converse: A degenerate set D must satisfy $\mathbf{1}_{D^{-\epsilon}} = \mathbf{1}_{\emptyset}$ $\mathbb{P}$-a.e:

Lemma 63. *Assume that $\mathbb{P} \ll \mu$ and let D be a degenerate set for an adversarial Bayes classifier A. Then $\mathbb{P}(D^{-\epsilon}) = 0$.*

See Appendix B.7.3 for a proof.

The adversarial classification risk heavily penalizes the boundary of a classifier. This observation suggests that if two connected components of a degenerate set are close together,

then they must actually be included in a larger degenerate set. The $^\epsilon$ and $^{-\epsilon}$ operations combine to form this enlarging operation.

Lemma 64. *Assume that $\mathbb{P} \ll \mu$. If D is a degenerate set for an adversarial Bayes classifier A, then $(D^\epsilon)^{-\epsilon}$ is as well.*

Proof. Let $A_2 = A \cup (D^\epsilon)^{-\epsilon}$. Then $S_\epsilon(\mathbf{1}_{A^C}) \geq S_\epsilon(\mathbf{1}_{A_2^C})$. We will show that $S_\epsilon(\mathbf{1}_{A_2}) = S_\epsilon(\mathbf{1}_A)$ $\mathbb{P}_0$-a.e., which will then imply that A_2 is an adversarial Bayes classifier, and furthermore A and A_2 are equivalent up to degeneracy by Proposition 51. Notice that the set A_2 satisfies

$$A \subset A_2 \subset ((A \cup D)^\epsilon)^{-\epsilon}$$

and then Lemma 60 implies that $A^\epsilon \subset (A \cup (D^\epsilon)^{-\epsilon})^\epsilon \subset (A \cup D)^\epsilon$. Because D is a degenerate set, $A_3 = A \cup D$ is an adversarial Bayes classifier and thus Proposition 51 implies that $\mathbf{1}_{A^\epsilon} = \mathbf{1}_{(A\cup D)^\epsilon}$-$\mathbb{P}_0$-a.e. which in turn implies $\mathbf{1}_{A^\epsilon} = \mathbf{1}_{A_2^\epsilon}$-$\mathbb{P}_0$-a.e. □

3.6 The Adversarial Bayes Classifier in One Dimension

In this section, we assume that $d = 1$ and the length of an interval I will be denoted $|I|$. Recall that connected subsets of $\mathbb{R}$ are either intervals or single point sets.

3.6.1 Regular adversarial Bayes classifiers—Proof of Theorem 35 and Theorem 37

Notice that if I is a connected component of A and A^C and $|I| < 2\epsilon$, then $I^{-\epsilon} = \emptyset$. Thus the set of connected components of A or A^C of length strictly less than 2ϵ is contained in a degenerate set by Proposition 62.

However, if $|I| = 2\epsilon$, then $I^{-\epsilon}$ contains at most one point: if $I = [x - \epsilon, x + \epsilon]$ then $I^{-\epsilon} = \{x\}$ while $I^{-\epsilon} = \emptyset$ if I is not closed. Due to this observation, the set of connected components of A and A^C of length 2ϵ is actually degenerate set as well. Thus one can argue:

Lemma 65. *Let $\mathbb{P}_0, \mathbb{P}_1 \ll \mu$, A be an adversarial Bayes classifier. Then there are adversarial Bayes classifiers $\tilde{A}_1, \tilde{A}_2$ satisfying $\tilde{A}_1 \subset A \subset \tilde{A}_2$ which are equivalent to A up to degeneracy and*

$$\tilde{A}_1 = \bigcup_{i=m}^{M} (\tilde{a}_i, \tilde{b}_i), \quad \tilde{A}_2^C = \bigcup_{j=n}^{M} (\tilde{e}_j, \tilde{f}_j)$$

where the intervals $(\tilde{a}_i, \tilde{b}_i)$, $(\tilde{e}_i, \tilde{f}_i)$ satisfy $\tilde{b}_i - \tilde{a}_i > 2\epsilon$ and $\tilde{f}_i - \tilde{e}_i > 2\epsilon$.

This statement is a consequence of Proposition 62 and Lemma 52.

Proof of Lemma 65. Lemma 52 implies that $\operatorname{int} A$ and $\overline{A}$ are both adversarial Bayes classifiers equivalent to A, and thus Corollary 61 implies that $D_1 = ((\operatorname{int} A)^\epsilon)^{-\epsilon} - ((\operatorname{int} A)^{-\epsilon})^\epsilon$ and $D_2 = ((\overline{A})^\epsilon)^{-\epsilon} - ((\overline{A})^{-\epsilon})^\epsilon$ are degenerate sets. Thus Lemma 52 and Corollary 61 imply that $\tilde{A}_1 = \operatorname{int} A - D_1$, $\tilde{A}_2 = \overline{A} \cup D_2$, and A are all equivalent up to degeneracy.

The adversarial Bayes classifier $\operatorname{int} A$ is an open set, and thus every connected component of $\operatorname{int} A$ is open. Therefore, if I is a connected component of $\operatorname{int} A$ of length less than or equal 2ϵ, then $I^{-\epsilon} = \emptyset$ and Proposition 62 implies that $I \subset D_1$. Hence every connected component of $\tilde{A}_1$ has length strictly larger than 2ϵ.

As $(\overline{A})^C$ is an open set and $\tilde{A}_2^C = (\overline{A})^C - D_2$, the same argument implies that every connected component of $\tilde{A}_2^C$ has length strictly larger than 2ϵ. □

These classifiers $\tilde{A}_1$ and $\tilde{A}_2$ have "one-sided" regularity— the connected components of $\tilde{A}_1$ and $\tilde{A}_2^C$ have length strictly greater than 2ϵ. Next, we use these classifiers with one-sided regularity to construct a classifier A' for which both A' and $(A')^C$ have components larger than 2ϵ.

This result suffices to prove Theorem 35, which is detailed in Appendix B.8, and we discuss an overview of this proof below. As $\tilde{A}_1 \subset \tilde{A}_2$, the sets $\tilde{A}_1$ and $\tilde{A}_2^C$ are disjoint. Therefore, one can express $\mathbb{R}$ as a disjoint union

$$\mathbb{R} = \tilde{A}_1 \sqcup \tilde{A}_2^C \sqcup D.$$

Both $\tilde{A}_1$ and $\tilde{A}_2^C$ are a disjoint union of intervals of length greater than 2ϵ, and thus $D = \tilde{A}_1^C \cap \tilde{A}_2$ must be a disjoint union of countably many intervals and isolated points. Notice that because D is degenerate, the union of $\tilde{A}_1$ and an arbitrary measurable portion of D is an adversarial Bayes classifier as well. To construct a regular adversarial Bayes classifier, we let D_1 be the connected components of D that are adjacent to some open interval of $\tilde{A}_1$. The remaining components of D, $D_2 = D - D_1$, must be adjacent to $\tilde{A}_2$. Therefore, if $A' = \tilde{A}_1 \cup D_1$ the connected components of $A' = \tilde{A}_1 \cup D_1$ and $(A')^C = \tilde{A}_2 \cup D_2$ must have length strictly greater than 2ϵ.

Next, Theorem 37 is a consequence of the fact that the adversarial risk of $A = \bigcup_{i=m}^{M}(a_i, b_i)$ equals (3.7) when A is regular.

Proof of Theorem 37. Because $b_i - a_i > 2\epsilon$ and $a_i - b_{i-1} > 2\epsilon$, we can treat $R^\epsilon(A)$ as a differentiable function of a_i on a small open interval around a_i as described by Equation (3.7). The first order necessary conditions for a minimizer then imply the first relation of (3.8) and the second order necessary conditions for a minimizer then imply the first relation of (3.13). The argument for b_i is analogous. □

3.6.2 Degenerate Sets in One Dimension—Proof of Theorem 38

First, every component of A or A^C with length less than equal to 2ϵ must be degenerate. In comparison, notice that this statement is strictly stronger than Proposition 62.

Lemma 66. *If a connected component C of A or A^C has length less than or equal to 2ϵ, then C is degenerate.*

Proof of Lemma 66. Let A be an adversarial Bayes classifier and let $\tilde{A}_1$ and $\tilde{A}_2$ be the two equivalent adversarial Bayes classifiers of Lemma 65. Because every connected component of component of $\tilde{A}_1$ has length strictly larger than 2ϵ, the connected components of A of length less than or equal to 2ϵ must be included in the degenerate set $A - \tilde{A}_1$. Similarly, the connected components of A^C of length less than or equal to 2ϵ are included in $\tilde{A}_2^C - A^C$, which is a degenerate set. □

Conversely, the length of a degenerate interval contained in $\operatorname{supp} \mathbb{P}$ is at most 2ϵ.

Corollary 67. *Let $\mathbb{P} \ll \mu$. Assume that $I \subset \operatorname{supp} \mathbb{P}$ is a degenerate interval for an adversarial Bayes classifier A. Then $|I| \leq 2\epsilon$.*

Proof. Lemma 63 implies that if I is a degenerate interval then $\mathbb{P}(I^{-\epsilon}) = 0$. Because I is an interval, the set $I^{-\epsilon}$ is either empty, a single point, or an interval. As $I \subset \operatorname{supp} \mathbb{P}$ and every interval larger than a single point has positive measure under μ, it follows that $I^{-\epsilon}$ is at most a single point and thus $|I| \leq 2\epsilon$. □

This result is then sufficient to prove the fourth bullet of Theorem 38. To start:

Lemma 68. *Let $\mathbb{P} \ll \mu$ and let A be an adversarial Bayes classifier. If $\operatorname{supp} \mathbb{P}$ is an interval and the adversarial Bayes classifier A has a degenerate interval I contained in $\operatorname{supp} \mathbb{P}^\epsilon$, then $\eta(x) \in \{0, 1\}$ on a set of positive measure.*

A formal proof is provided in Appendix B.9.1, we sketch the main ideas below. Let I be a degenerate interval in $\operatorname{supp} \mathbb{P}$. One can then find a 'maximal' degenerate interval $J = [d_3, d_4]$ containing I inside $\operatorname{supp} \mathbb{P}$, in the sense that if J' is a degenerate interval and $J \subset J'$ then $J' = J$. Corollary 67 implies that $|J| \leq 2\epsilon$ and Lemma 64 implies that J is of distance strictly more than 2ϵ from any other degenerate set. Thus the intervals $[d_3 - \epsilon, d_3)$, $(d_4, d_4 + \epsilon]$ do

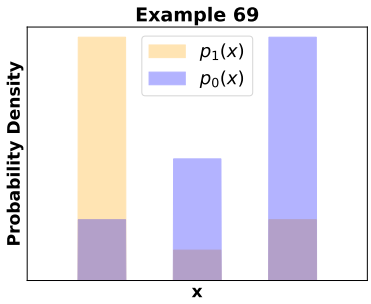

Figure 3.3: The distribution of Example 69.

not intersect a degenerate subset of A, and these intervals must be entirely contained in A or A^C due to Lemma 66. Thus one can compute the difference $R^\epsilon(A \cup J) - R^\epsilon(A - J)$ under four cases: 1) $[d_3 - \epsilon, d_3) \subset A$, $(d_4, d_4 + \epsilon] \subset A$; 2) $[d_3 - \epsilon, d_3) \subset A$, $(d_4, d_4 + \epsilon] \subset A^C$; 3) $[d_3 - \epsilon, d_3) \subset A^C$, $(d_4, d_4 + \epsilon] \subset A$; 4) $[d_3 - \epsilon, d_3) \subset A^C$, $(d_4, d_4 + \epsilon] \subset A^C$.

In each case, this difference results in $\int_{I'} p_1(x)dx = 0$ or $\int_{I'} p_0(x)dx = 0$ on some interval $I' \subset \operatorname{supp} \mathbb{P}$, which implies either $\eta = 1$ or $\eta = 0$, respectively, on a set of positive measure.

Lemma 68 and Lemma 64 together imply the fourth bullet of Theorem 38. The argument is outlined below, with a formal proof in Appendix B.9.2. If $D \subset \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon$ is a degenerate set which contains two points $x \leq y$ at most 2ϵ apart, then Lemma 64 implies that $[x, y] \subset (D^\epsilon)^{-\epsilon}$ is degenerate, which would contradict Lemma 68. Thus $D \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon$ must be comprised of points that are strictly more than 2ϵ apart. However, if $x \in D$ is more than 2ϵ from any point in ∂A, then one can argue that $R^\epsilon(A - \{x\}) - R^\epsilon(A) > 0$ if $x \in A$ and $R^\epsilon(A \cup \{x\}) - R^\epsilon(A) > 0$ if $x \notin A$. Thus if D is a degenerate set is disjoint from $\overline{(\operatorname{supp} \mathbb{P}^\epsilon)^C}$, then D must be contained in ∂A.

Combining previous results then proves Theorem 38— The first bullet of Theorem 38 is Lemma 66, the second bullet is Corollary 67, the third bullet is Lemma 58, and the fourth bullet is shown in Appendix B.9.2.

Lemma 68 and the fourth bullet of Theorem 38 are false when $\operatorname{supp} \mathbb{P}$ is not an interval.

Example 69. Consider a probability distribution for which

$$p_1(x) = \begin{cases} \frac{8}{25\epsilon} & \text{if } -\frac{5}{2}\epsilon \le x \le -\frac{3}{2}\epsilon \\ \frac{1}{25\epsilon} & \text{if } -\frac{1}{2}\epsilon \le x \le +\frac{1}{2}\epsilon \\ \frac{2}{25\epsilon} & \text{if } +\frac{3}{2}\epsilon \le x \le +\frac{5}{2}\epsilon \\ 0 & \text{otherwise} \end{cases} \quad p_0(x) = \begin{cases} \frac{2}{25\epsilon} & \text{if } -\frac{5}{2}\epsilon \le x \le -\frac{3}{2}\epsilon \\ \frac{4}{25\epsilon} & \text{if } -\frac{1}{2}\epsilon \le x \le +\frac{1}{2}\epsilon \\ \frac{8}{25\epsilon} & \text{if } +\frac{3}{2}\epsilon \le x \le +\frac{5}{2}\epsilon \\ 0 & \text{otherwise} \end{cases}$$

See Figure 3.3 for an illustration. Then there are no solutions x to the necessary conditions Equation (3.8) within $\operatorname{supp} \mathbb{P}^\epsilon$ at which p_0 is continuous at $x \pm \epsilon$ and p_1 continuous at $x \mp \epsilon$. Thus the only possible values for the a_is and b_is within $\operatorname{supp} \mathbb{P}^\epsilon$ are $\{-\frac{7}{2}, -\frac{5}{2}\epsilon, -\frac{3}{2}\epsilon, -\frac{1}{2}\epsilon, +\frac{1}{2}\epsilon, +\frac{3}{2}\epsilon, +\frac{5}{2}\epsilon, +\frac{7}{2}\epsilon\}$. By comparing the risks of all adversarial Bayes classifiers with endpoints in this set, one can show that $(-\infty, -\frac{1}{2}\epsilon)$ is an adversarial Bayes classifier. At the same time, $R^\epsilon((-\infty, -\frac{1}{2}\epsilon) \cup S) = R^\epsilon((-\infty, -\frac{1}{2}\epsilon))$ for any subset S of $[-\frac{1}{2}\epsilon, +\frac{1}{2}\epsilon]$. Thus $[-\frac{1}{2}\epsilon, +\frac{1}{2}\epsilon]$ is a degenerate set, but $\eta(x) = \frac{1}{5}$ on $[-\frac{1}{2}\epsilon, +\frac{1}{2}\epsilon]$. See Appendix B.11.8 for details.

3.6.3 Regularity as ϵ Increases—Proof of Theorem 39

Let A_1 and A_2 be two regular adversarial Bayes classifiers corresponding to perturbation radiuses ϵ_1 and ϵ_2 respectively. Notice that the adversarial classification risk in Equation (3.5) pays a penalty of 1 within ϵ of each a_i and b_i. This consideration suggests that as ϵ increases, there should be fewer transitions between the two classes in the adversarial Bayes classifier. The key observation is that so long as A_1 is non-trivial, no connected component of A_2 should contain a connected component of A_1^C and no connected component of A_2^C should contain a connected component of A_1.

We adopt additional notation to formally state this principle. When $\bigcup_{i=m}^{M}(a_i, b_i)$ is a regular adversarial Bayes classifier and M is finite, define a_{M+1} to be $+\infty$. Similarly, if m is

finite, define b_{m-1} as $-\infty$.

Lemma 70. *Assume that $\mathbb{P} \ll \mu$ is a measure for which $\operatorname{supp}\mathbb{P}$ is an interval I and $\mathbb{P}(\eta(x) = 0 \text{ or } 1) = 0$. Let $A_1 = \bigcup_{i=m}^{M}(a_i^1, b_i^1)$ and $A_2 = \bigcup_{j=n}^{N}(a_j^2, b_j^2)$ be two regular adversarial Bayes classifiers corresponding to perturbation sizes $\epsilon_1 < \epsilon_2$.*

- *If both $\mathbb{R}$ and $\emptyset$ are adversarial Bayes classifiers for perturbation radius ϵ_1, then both $\mathbb{R}$ and $\emptyset$ are adversarial Bayes classifiers for perturbation radius ϵ_2.*
- *Assume that $\mathbb{R}$ and $\emptyset$ are not both adversarial Bayes classifiers for perturbation radius ϵ_1. Then for each interval (a_i^1, b_i^1), the set $(a_i^1, b_i^1) \cap I^{\epsilon_1}$ cannot contain any non-empty $(b_j^2, a_{j+1}^2) \cap I^{\epsilon_1}$ and for each interval (b_i^1, a_{i+1}^1), the set $(b_i^1, a_{i+1}^1) \cap I^{\epsilon_1}$ cannot contain any non-empty $(a_j^2, b_j^2) \cap I^{\epsilon_1}$.*

Example 41 demonstrates the the exception to the second bullet— when $\epsilon \geq (\mu_1 - \mu_0)/2$, both $\mathbb{R}$ and $\emptyset$ are adversarial Bayes classifiers.

To show Lemma 70, notice that if $A_2 = \bigcup_{i=1}^{M}(a_i^2, b_i^2)$ is a regular adversarial Bayes classifier and $(a_j^2, b_j^2) \subset I^{-\epsilon_2}$, then $R^{\epsilon_2}(A_2 - (a_j^2, b_j^2)) \geq R^{\epsilon_2}(A_2)$ which is equivalent to

$$0 \leq \int_{a_j^2-\epsilon_2}^{b_j^2+\epsilon_2} p_1 dx - \left(\int_{a_j^2-\epsilon_2}^{a_j^2+\epsilon_2} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx + \int_{b_j^2-\epsilon_2}^{b_j^2+\epsilon_2} p dx \right)$$
$$= \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_1(x) dx - \int_{a_j^2-\epsilon_2}^{b_j^2+\epsilon_2} p_0(x) dx$$

As p_0, p_1 are non-zero on $\operatorname{supp}\mathbb{P}$, replacing ϵ_2 with ϵ_1 in this last expression would increase the first integral and decrease the second, thereby increasing the entire expression.

Thus, if $(a_j^2 - \epsilon_1, b_j^2 + \epsilon_1) \subset A_1^C$, this calculation would imply that $R^{\epsilon_1}(A_1 \cup (a_i^2, b_i^2)) < R^{\epsilon_1}(A_1)$, which would contradict the fact that A_1 is an adversarial Bayes classifier. Similar but more technical calculations performed in Appendix B.10 show that if $(a_i^2, b_i^2) \subset A_1^C \cap I^{\epsilon_1}$ then $R^{\epsilon_1}(A_1 \cup (a_i^2, b_i^2)) < R^{\epsilon_1}(A_1)$ and so A_1 cannot be an adversarial Bayes classifier.

3.7 Related Works

Prior work analyzes several variations of our setup, such as perturbations in open balls [16], alternative perturbation sets [11], attacks using general Wasserstein p-metrics [63, 64], minimizing R^ϵ over Lebesgue measurable sets [52], the multiclass setting [63], and randomized classifiers [28, 63]. Due to the plethora of attacks present in the literature, this paper contains proofs of all intermediate results that appear in prior work (such as Lemma 27 from [16]). Understanding the uniqueness of the adversarial Bayes classifier in these contexts remains an open question. Establishing a notion of uniqueness for randomized classifiers in the adversarial context is particularly interesting, as randomized classifiers are essential in calculating the minimal possible error in adversarial multiclass classification [63] but not binary classification [28].

Prior work [1, 11, 50] adopts a different method for identifying adversarial Bayes classifiers for various distributions. To prove a set is an adversarial Bayes classifier, [11] first show a strong duality result $\inf_A R^\epsilon(A) = \sup_\gamma \tilde{D}(\gamma)$ for some dual risk $\tilde{D}$ on the set of couplings between two measures. Subsequently, [1, 11, 50] exhibit a set A and a coupling γ for which the adversarial risk of A matches the dual risk of γ, and thus A must minimize the adversarial classification risk. This approach involves solving the first order necessary conditions Equation (3.8), and [1] relies on a result of [64] which states that these relations hold for sufficiently small ϵ under reasonable assumptions. In contrast, this paper uses equivalence up to degeneracy to show that it suffices to consider sets with enough regularity for the first order necessary conditions to hold; and the solutions to these conditions typically reduce the possibilities for the adversarial Bayes classifier to a finite number of sets.

Prior work on regularity [2, 16] prove the existence of adversarial Bayes classifiers with one sided tangent balls. Theorem 40 states that each equivalence class under equivalence up to degeneracy has a representative with this type of regularity. Furthermore, results of [1]

imply that under reasonable assumptions, one can choose adversarial Bayes classifiers $A(\epsilon)$ for which $\mathrm{comp}(A(\epsilon)) + \mathrm{comp}(A(\epsilon)^C)$ is always decreasing in ϵ. Specifically, they show that for increasing ϵ, the only possible discontinuous changes in $A(\epsilon)$ are merged components, deleted components, or a endpoint of a component changing discontinuously in ϵ. This statement does not imply Lemma 70, and Lemma 70 does not imply this result of [1].

3.8 Conclusion

We defined a new notion of uniqueness for the adversarial Bayes classifier, which we call *uniqueness up to degeneracy*. This concept generalizes uniqueness for the Bayes classifier. The concept of uniqueness up to degeneracy produces a method for calculating the adversarial Bayes classifier for a reasonable family of distributions in one dimension, and assists in understanding their regularity properties. We hope that the theoretical insights in this paper will assist in the development of algorithms for robust learning.

Acknowledgments

Natalie Frank was supported in part by the Research Training Group in Modeling and Simulation funded by the National Science Foundation via grant RTG/DMS – 1646339 NSF grant DMS-2210583 and grants DMS-2210583, CCF-1535987, IIS-1618662.

4 — Adversarial Consistency

4.1 Introduction

A central issue in the study of neural nets is their susceptibility to adversarial perturbations—perturbations imperceptible to the human eye can cause a neural net to misclassify an image [14, 58]. The same phenomenon appears in other types of data such as speech and text. As deep nets are used in applications such as self-driving cars and medical imaging [37, 47], training classifiers robust to adversarial perturbations is a central question in machine learning.

The foundational theory of surrogates for classication in well understood. In the standard classification setting, one seeks to minimize the *classification* risk— the proportion of incorrectly classified data. Since minimizing the classification risk is typically computationally intractable [9], a common approach is to instead minimize a better-behaved alternative called the *surrogate risk*. However, one must verify that classifiers with low surrogate risk also achieve low classification risk. If for every data distribution, a sequence of functions minimizing the surrogate also minimizes the classification risk, the surrogate risk is called *consistent*. Many classic papers study the consistency of surrogate risks in the standard classification setting [8, 38, 43, 49, 57].

Unlike the standard case, however, little is known about the consistency of surrogate risks in the context of adversarial training, which involves risks that compute the supremum

of a surrogate loss function over an ϵ-ball. Though this question has been partially studied in the literature [3, 4, 42], a general theory is lacking. Existing results reveal, however, that the situation is substantially different from the standard case: for instance, [42] show that no *convex* surrogate can be adversarially consistent. To our knowledge, no adversarially consistent risks are known.

In this work, we give a complete characterization of adversarial consistency for surrogate losses.

Our Contributions:

- In Section 4.4 we give a surprisingly simple necessary and sufficient condition for adversarial consistency:

  Informal Theorem. *Under reasonable assumptions on the surrogate loss ϕ, the supremum-based ϕ-risk is adversarially consistent if and only if* $\inf_\alpha \phi(\alpha)/2 + \phi(-\alpha)/2 < \phi(0)$.

  In particular, this result proves consistency for any loss function that is *not* midpoint convex at the origin.

- In Section 4.5, we specialize to the case of the ρ-margin loss, where we obtain a quantitative proof of adversarial consistency by explicitly bounding the excess adversarial risk.

To the best of the authors' knowledge, this paper is the first to prove that a loss-based learning procedure is consistent for a wide range of distributions in the adversarial setting. As mentioned above, the ρ-margin loss $\phi_\rho(\alpha) = \min(1, \max(1 - \alpha/\rho, 0))$ satisfies the conditions of Informal Theorem above, as does the shifted sigmoid loss $\phi_\tau(\alpha) = 1/(1 + \exp(\alpha - \tau))$ with $\tau > 0$, which confirms a conjecture of Meunier et al. [42]. By contrast, all convex losses satisfy $\inf_\alpha \phi(\alpha)/2 + \phi(-\alpha)/2 = \phi(0)$, and are therefore not adversarially consistent.

In addition to consistency, one would hope to obtain a quantitative comparison between the adversarial surrogate risk and the adversarial classification risk. Our bound in Section 4.5

shows that the excess error of the adversarial ρ-margin loss is a linear upper bound on the adversarial classification error, which implies that minimizing the adversarial ρ-margin loss is an effective procedure for minimizing the adversarial classification error. Extending the bound in Section 4.5 to further losses remains an open question.

4.2 Related Works

Many previous works have studied the consistency of surrogate risks [8, 38, 43, 49, 57]. The classic papers by [8, 38, 75] explore the consistency of surrogate risks over all measurable functions. The works [5, 43, 49] study $\mathcal{H}$-consistency, which is consistency restricted to a smaller set of functions. Steinwart [57] generalizes some of these results into a framework referred to as *calibration*. Awasthi et al. [3, 4], Bao, Scott, and Sugiyama [6], and Meunier et al. [42] then use this framework to analyze the calibration of adversarial surrogate losses. Furthermore Meunier et al. [42] relate calibration to consistency for adversarial losses in certain cases — they show that no convex loss is adversarially consistent. They also conjecture that a class of surrogate losses called the *odd shifted* losses are adversarially consistent. Meunier et al. [42] also show that in a restricted setting, surrogates are consistent for 'optimal attacks'. The proof of our result formalizes this intuition. Simultaneous work [40] shows that the ρ-margin loss is adversarially $\mathcal{H}$-consistent for typical function classes. Lastly, Bhattacharjee and Chaudhuri [12, 13] use a different set of techniques to study the consistency of non-parametric methods in adversarial scenarios.

Our results rely on recent works establishing the properties of minimizers to surrogate adversarial risks. [2, 16, 52] all proved the existence of minimizers to the adversarial risk and [52] proved a minimax theorem for the zero-one loss. Building on the work of [52], [25] later proved similar existence and minimax statements for arbitrary surrogate losses. Trillos, Jacobs, and Kim [62, 63] extend some of these results to the multiclass case. Lastly, [64]

study further properties of the minimizers to the adversarial classification loss.

4.3 Problem Setup

This section contains the necessary background for our results. Section 4.3.1 gives precise definitions for the main concepts, and Section 4.3.2 describes the minimax theorems that are at the heart of our proof.

4.3.1 Surrogate Risks

This paper studies binary classification on $\mathbb{R}^d$. Explicitly, labels are $\{-1, +1\}$ and the data is distributed according to a distribution $\mathcal{D}$ on the set $\mathbb{R}^d \times \{-1, +1\}$. The measures $\mathbb{P}_1$, $\mathbb{P}_0$ define the relative probabilities of finding points with a given label in a region of $\mathbb{R}^d$. Formally, define measures on $\mathbb{R}^d$ by

$$\mathbb{P}_1(A) = \mathcal{D}(A \times \{+1\}), \mathbb{P}_0(A) = \mathcal{D}(A \times \{-1\}).$$

The *classification risk* $R(f)$ is then the probability of misclassifying a point under $\mathcal{D}$:

$$R(f) = \int \mathbf{1}_{f(\mathbf{x}) \leq 0} d\mathbb{P}_1 + \int \mathbf{1}_{f(\mathbf{x}) > 0} d\mathbb{P}_0. \tag{4.1}$$

The surrogate to R is

$$R_\phi(f) = \int \phi(f) d\mathbb{P}_1 + \int \phi(-f) d\mathbb{P}_0 \,. \tag{4.2}$$

A classifier can be obtained by minimizing either R or R_ϕ over the set of all measurable functions. A point $\mathbf{x}$ is then classified according to sign f. There are many possible choices for ϕ—typically one chooses a loss that is easy to optimize. In this paper, we assume that

Assumption 2. *ϕ is non-increasing, non-negative, continuous, and* $\lim_{\alpha\to\infty}\phi(\alpha)=0$.

Most surrogate losses in machine learning satisfy this assumption. Learning algorithms typically optimize the risk in (4.2) using an iterative procedure, which produces a sequence of functions that minimizes R_ϕ. We call R_ϕ a *consistent risk* and ϕ a *consistent loss* if for all distributions, every minimizing sequence of R_ϕ is also a minimizing sequence of R.[1] Alternatively, the risks R, R_ϕ can be expressed in terms of the quantities $\mathbb{P}=\mathbb{P}_0+\mathbb{P}_1$ and $\eta=d\mathbb{P}_1/d\mathbb{P}$. For all $\eta\in[0,1]$, define

$$C(\eta,\alpha)=\eta\mathbf{1}_{\alpha\leq 0}+(1-\eta)\mathbf{1}_{\alpha>0},\quad C^*(\eta)=\inf_\alpha C(\eta,\alpha), \tag{4.3}$$

$$C_\phi(\eta,\alpha)=\eta\phi(\alpha)+(1-\eta)\phi(-\alpha),\quad C_\phi^*(\eta)=\inf_\alpha C_\phi(\eta,\alpha) \tag{4.4}$$

For more on the definitions of R, R_ϕ, C, C_ϕ, see [8] or Sections 3.1 and 3.2 of [25]. Using these definitions, $R(f)=\int C(\eta(\mathbf{x}),f(\mathbf{x}))d\mathbb{P}$ and

$$R_\phi(f)=\int C_\phi(\eta(\mathbf{x}),f(\mathbf{x}))d\mathbb{P} \tag{4.5}$$

This alternative view of the risks R and R_ϕ provides a 'pointwise' criterion for consistency—if the function $f(\mathbf{x})$ minimizes $C_\phi(\eta(\mathbf{x}),\cdot)$ at each point, then it also minimizes R_ϕ. However, minimizers to $C_\phi(\eta,\cdot)$ over $\mathbb{R}$ do not always exist— consider for instance $\eta=1$ for the exponential loss $\phi(\alpha)=e^{-\alpha}$. In general, for minimizers of $C_\phi(\eta,\cdot)$ to exist, one must work over the extended real numbers $\overline{\mathbb{R}}=\mathbb{R}\cup\{-\infty,+\infty\}$. The following proposition proved in Appendix C.1 implies that 'pointwise' considerations also extends to minimizing sequences of functions.

Proposition 71. *The following are equivalent:*

[1] In the context of standard (non-adversarial) learning, the concept we defined as consistency is often referred to as *calibration*, see for instance [8, 57]. We opt for the term 'consistency' as the prior works [3, 4, 42] use calibration to refer to a different but related concept in the adversarial setting.

1) ϕ is consistent

2) Every minimizing sequence of $C_\phi(\eta, \cdot)$ is also a minimizing sequence of $C(\eta, \cdot)$

3) Every $\overline{\mathbb{R}}$-valued minimizer of R_ϕ is a minimizer of R

This result is well-known in prior literature; in particular the equivalence between 2) and 3) is closely related to the equivalence between calibration and consistency in the non-adversarial setting [57]. Most importantly, the equivalence between 1) and 3) reduces studying minimizing sequences of functionals to studying minimizers of functions. We will show that the equivalence between 1) and 2) has an analog in the adversarial scenario, but the equivalence between 1) and 3) does not.

In the adversarial classification setting, every x-value is perturbed by a malicious adversary before undergoing classification by f. We assume that these perturbations are bounded by ϵ in some norm $\|\cdot\|$ and furthermore, the adversary knows both our classifier f and the true label of the point $\mathbf{x}$. In other words, f misclassifies $(\mathbf{x}, y)$ when there is a point $\mathbf{x}' \in \overline{B_\epsilon(\mathbf{x})}$ for which $\mathbf{1}_{f(\mathbf{x}')\leq 0} = 1$ for $y = +1$ and $\mathbf{1}_{f(\mathbf{x}')>0} = 1$ for $y = -1$. Conveniently, this criterion can be expressed in terms of suprema. For any function g, we define

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

A point $\mathbf{x}$ with label $+1$ is misclassified when $S_\epsilon(\mathbf{1}_{f\leq 0})(\mathbf{x}) = 1$ and a point $\mathbf{x}$ with label -1 is misclassified when $S_\epsilon(\mathbf{1}_{f>0})(\mathbf{x}) = 1$. Hence the expected fraction of errors under the adversarial attack is

$$R^\epsilon(f) = \int S_\epsilon(\mathbf{1}_{f\leq 0}) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_{f>0}) d\mathbb{P}_0, \tag{4.6}$$

which is called the *adversarial classification risk* [2]. Again, optimizing the empirical version

[2]Defining this integral requires some care because for a Borel function g, $S_\epsilon(g)$ may not be measurable;

of (4.6) is computationally intractable so instead one minimizes a surrogate of the form

$$R_\phi^\epsilon(f) = \int S_\epsilon(\phi \circ f) d\mathbb{P}_1 + \int S_\epsilon(\phi \circ -f) d\mathbb{P}_0 \tag{4.7}$$

Due to the supremum in this expression, we refer to such a risk as a *supremum-based surrogate*. We define adversarial consistency as

Definition 72. *The risk R_ϕ^ϵ is* adversarially consistent *if for every data distribution, every sequence f_n which minimizes R_ϕ^ϵ over all Borel measurable functions also minimizes R^ϵ. We say that the loss ϕ is* adversarially consistent *if the risk R_ϕ^ϵ is adversarially consistent.*

Many convex and non-convex losses are consistent in standard classification [8, 38, 53, 57, 75]. By contrast, adversarial consistency often fails. For instance, Meunier et al. [42] show that convex losses are not adversarially consistent. Furthermore, their example shows that the equivalence between 1) and 3) in Proposition 71 does *not* hold in the adversarial context. Thus, to understand adversarial consistency, it does not suffice to compare minimizers of R_ϕ^ϵ and R^ϵ. To illustrate this distinction, we show the following result, adapted from [42].

Proposition 73. *Assume that* $\inf_\alpha \phi(\alpha)/2 + \phi(-\alpha)/2 = \phi(0)$. *Then ϕ is not adversarially consistent.*

Proof. Let $\mathbb{P}_0 = \mathbb{P}_1$ be the the uniform distribution on the ball $\overline{B_R(\mathbf{0})}$ and let $\epsilon = 2R$. Let ϕ be a loss function for which $\inf_\alpha \phi(\alpha)/2 + \phi(-\alpha)/2 = C_\phi^*(1/2) = \phi(0)$. Notice that $\inf_f R^\epsilon(f) \geq \inf_f R(f)$ and $\inf_f R_\phi^\epsilon(f) \geq \inf_f R_\phi(f)$. Since $\mathbb{P}_0 = \mathbb{P}_1$, the optimal non-adversarial risk is $\inf_f R(f) = 1/2$. Moreover, as $C_\phi^*(1/2) = \phi(0)$, the optimal non-adversarial surrogate risk is $\inf_f R_\phi(f) = C_\phi^*(1/2) = \phi(0)$. Thus, for the function $f^* \equiv 0$, $R^\epsilon(f^*) = \inf_f R(f) = 1/2$ and $R_\phi^\epsilon(f^*) = \inf_f R_\phi(f) = \phi(0)$. Therefore f^* minimizes both R_ϕ^ϵ and R^ϵ. Now consider

see Section 3.3 and Appendix A of [25] for details.

the sequence of functions

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

Because $\epsilon = 2R$, every point in the support of the distribution can be perturbed to every other point. Thus $S_\epsilon(\phi \circ f_n)(\mathbf{x}) = \phi(-1/n)$ and $S_\epsilon(\phi \circ -f_n)(\mathbf{x}) = \phi(-1/n)$. However, $S_\epsilon(\mathbf{1}_{f \leq 0}) = 1$ and $S_\epsilon(\mathbf{1}_{f>0}) = 1$. Therefore, $R^\epsilon_\phi(f_n) = \phi(-1/n)$ while $R^\epsilon(f_n) = 1$ for all n. As ϕ is continuous, $\lim_{n\to\infty} R^\epsilon_\phi(f_n) = \phi(0)$. Thus f_n is a minimizing sequence of R^ϵ_ϕ but not of R^ϵ, so ϕ is not adversarially consistent. □

This example shows that if $C^*_\phi(1/2) = \phi(0)$, then ϕ is not adversarially consistent. The main result of this paper is that this is the *only* obstruction to adversarial consistency: ϕ is adversarially consistent if and only if $C^*_\phi(1/2) < \phi(0)$.

We begin by showing that this condition suffices for consistency in the *non-adversarial* setting. Surprisingly, despite the wealth of work on this topic, this condition does not appear to be known.

Proposition 74. *If $C^*_\phi(1/2) < \phi(0)$, then ϕ is consistent.*

See Appendix C.3 for a proof.

Again, some losses that satisfy this property are the ρ-margin loss $\phi_\rho(\alpha) = \min(1, \max(1 - \alpha/\rho, 0))$ and the the shifted sigmoid loss proposed by Meunier et al. [42], $\phi(\alpha) = 1/(1 + \exp(\alpha - \tau))$, $\tau > 0$. (In fact, one can show that the class of shifted odd losses proposed by Meunier et al. [42] satisfy $C^*_\phi(1/2) < \phi(0)$.)

Notice that all convex losses satisfy $C^*_\phi(1/2) = \phi(0)$:

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

The opposite inequality follows from the observation that $C^*_\phi(1/2) \leq C_\phi(1/2, 0) = \phi(0)$. In

contrast, recall that a convex loss ϕ with $\phi'(0) < 0$ is consistent [8].

As conjectured by prior work [6, 42], the fundamental reason losses with $C^*_\phi(1/2) < \phi(0)$ are adversarially consistent is that minimizers of $C_\phi(\eta, \cdot)$ are uniformly bounded away from 0 for all η:

Lemma 75. *The loss ϕ satisfies $C^*_\phi(1/2) < \phi(0)$ iff there is an $a > 0$ for which any minimizer α^* of $C_\phi(\eta, \cdot)$ satisfies $|\alpha| \geq a$.*

See C.3 for a proof. Concretely, one can show that for the ρ-margin loss ϕ_ρ, a minimizer α^* of $C_{\phi_\rho}(\eta, \cdot)$ must satisfy $|\alpha^*| \geq \rho$. Similarly, a minimizer α^* of $C_{\phi_\tau}(\eta, \cdot)$ of the shifted sigmoid loss $\phi_\tau = 1/(1 + \exp(\alpha - \tau))$, $\tau > 0$ is always either $-\infty$ or $+\infty$. In 4.4, we use this property to show that minimizing sequences of R^ϵ_ϕ must be uniformly bounded away from zero, thus ruling out the counterexample presented in Proposition 73.

4.3.2 Minimax Theorems for Adversarial Risks

We study the consistency of ϕ by by comparing minimizing sequences of R^ϵ_ϕ with those of R^ϵ. In the next section, in order to compare these minimizing sequences, we will attempt to re-write the adversarial loss in a 'pointwise' manner similar to Proposition 71. In order to achieve this representation of the adversarial loss, we apply minimax and complementary slackness theorems from [25, 52].

Before presenting these results, we introduce the ∞-Wasserstein metric from optimal transport. For two finite probability measures $\mathbb{Q}, \mathbb{Q}'$ satisfying $\mathbb{Q}(\mathbb{R}^d) = \mathbb{Q}'(\mathbb{R}^d)$, let $\Pi(\mathbb{Q}, \mathbb{Q}')$ be the set of *couplings* between $\mathbb{Q}$ and $\mathbb{Q}'$:

$$\Pi(\mathbb{Q}, \mathbb{Q}') = \{\gamma : \text{measure on } \mathbb{R}^d \times \mathbb{R}^d \text{ with } \gamma(A \times \mathbb{R}^d) = \mathbb{Q}(A), \gamma(\mathbb{R}^d \times A) = \mathbb{Q}'(A)\}$$

The distance between $\mathbb{Q}'$ and $\mathbb{Q}$ in the Wasserstein ∞-metric W_∞ is defined as

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

The W_∞ distance is in fact a metric on the space of measures. We denote the ∞-Wasserstein ball around a measure $\mathbb{Q}$ by

$$\mathcal{B}_\epsilon^\infty(\mathbb{Q}) = \{\mathbb{Q}' \colon \mathbb{Q}' \text{ Borel}, \quad W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon\}$$

Informally, the measure $\mathbb{Q}'$ is in $\mathcal{B}_\epsilon^\infty(\mathbb{Q})$ if perturbing points by at most ϵ under the measure $\mathbb{Q}$ can produce $\mathbb{Q}'$. As a result, Wasserstein ∞-balls are fairly useful for modeling adversarial attacks. Specifically, one can show:

Lemma 76. *For any function g and measures $\mathbb{Q}'$, $\mathbb{Q}$ with $W_\infty(\mathbb{Q}', \mathbb{Q}) \leq \epsilon$, the inequality $\int S_\epsilon(g) d\mathbb{Q} \geq \int g d\mathbb{Q}'$ holds.*

See Appendix C.4 for a proof.

Minimax theorems from prior work use this framework to introduce dual problems to the adversarial classification risks (4.6) and (4.7). Let $\mathbb{P}_0', \mathbb{P}_1'$ be finite Borel measures and define

$$\bar{R}(\mathbb{P}_0', \mathbb{P}_1') = \int C^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')} \right) d(\mathbb{P}_0' + \mathbb{P}_1') \tag{4.8}$$

where C^* is defined by (4.3). The next theorem states that maximizing $\bar{R}$ over W_∞ balls is in fact a dual problem to minimizing R^ϵ.

Theorem 77. *Let $\bar{R}$ be defined by (4.8).*

$$\inf_{\substack{f \text{ Borel} \\ \mathbb{R}\text{-valued}}} R^\epsilon(f) = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') \tag{4.9}$$

and furthermore equality is attained for some Borel measurable $\hat{f}$ *and* $\hat{\mathbb{P}}_1, \hat{\mathbb{P}}_0$ *with* $W_\infty(\hat{\mathbb{P}}_0, \mathbb{P}_0) \leq \epsilon$ *and* $W_\infty(\hat{\mathbb{P}}_1, \mathbb{P}_1) \leq \epsilon$.

The first to show such a theorem was Pydi and Jog [52]. In comparison to their Theorem 8, Theorem 77 removes the assumption that $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous with respect to Lebesgue measure and shows that the minimizer $\hat{f}$ is in fact Borel. We prove this theorem in Appendix C.5. Frank and Niles-Weed [25] prove a similar statement for the surrogate risk R^ϵ_ϕ. This time, the dual objective is

$$\bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) = \int C^*_\phi \left(\frac{d\mathbb{P}'_1}{d(\mathbb{P}'_0 + \mathbb{P}'_1)} \right) d(\mathbb{P}'_0 + \mathbb{P}'_1) \tag{4.10}$$

with C^*_ϕ defined by (4.4).

Theorem 78. *Assume that Assumption 2 holds, and define* $\bar{R}_\phi$ *by* (4.10). *Then*

$$\inf_{\substack{f \text{ Borel,} \\ f \ \mathbb{R}\text{-valued}}} R^\epsilon_\phi(f) = \sup_{\substack{\mathbb{P}'_0 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \\ \mathbb{P}'_1 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) \tag{4.11}$$

and furthermore equality in the dual problem is attained for some $\mathbb{P}^*_1, \mathbb{P}^*_0$ *with* $W_\infty(\mathbb{P}^*_0, \mathbb{P}_0) \leq \epsilon$ *and* $W_\infty(\mathbb{P}^*_1, \mathbb{P}_1) \leq \epsilon$.

Frank and Niles-Weed [25] proved this statement in Theorem 6 but with the infimum taken over $\overline{\mathbb{R}}$-valued functions. To extend the result to $\mathbb{R}$-valued functions as in Theorem 78, we show that $\inf_{f \text{ Borel}, f \ \overline{\mathbb{R}}\text{-valued}} R^\epsilon_\phi(f) = \inf_{f \text{ Borel}, f \ \mathbb{R}\text{-valued}} R^\epsilon_\phi(f)$ in Appendix C.2.

4.4 Adversarially Consistent Losses

This section contains our main results on adversarial consistency. In light of Proposition 73, our main task is to show that a loss satisfying $C^*_\phi(1/2) < \phi(0)$ is adversarially consistent.

At a high level, we will show that every minimizing sequence of R^ϵ_ϕ must also minimize R^ϵ. However, directly analyzing minimizing sequences $\{f_n\}$ of R^ϵ_ϕ and R^ϵ is challenging due to the supremums in the definitions of the adversarial risks. We therefore develop alternate characterizations of minimizing sequences to both functionals, based on complementary slackness conditions derived from the convex duality results of Section 3.2. However, unlike standard complementary slackness conditions well known from convex optimization, these theorems allow us to characterize minimizing sequences as well as minimizers.

4.4.1 Approximate Complementary Slackness

We first state this slackness result for the surrogate case, due to Frank and Niles-Weed [25, Lemmas 16 and 26] and Theorem 78.

Proposition 79. *Let $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ be any maximizers of $\bar{R}_\phi$ over $\mathcal{B}^\infty_\epsilon(\mathbb{P}_i)$. Define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$, $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. If f_n is a minimizing sequence for R^ϵ_ϕ, then the following hold:*

$$\lim_{n\to\infty} \int C_\phi(\eta^*, f_n) d\mathbb{P}^* = \int C^*_\phi(\eta^*) d\mathbb{P}^*. \tag{4.12}$$

$$\lim_{n\to\infty} \int S_\epsilon(\phi \circ f_n) d\mathbb{P}_1 - \int \phi \circ f_n d\mathbb{P}_1^* = 0, \quad \lim_{n\to\infty} \int S_\epsilon(\phi \circ -f_n) d\mathbb{P}_0 - \int \phi \circ -f_n d\mathbb{P}_0^* = 0 \tag{4.13}$$

Proof. Let $R^\epsilon_{\phi,*}$ be the minimal value of R^ϵ_ϕ and choose a $\delta > 0$. Then for sufficiently large N, $n \geq N$ implies that $R^\epsilon_\phi(f_n) \leq R^\epsilon_{\phi,*} + \delta$. Lemma 76 and the definition of C^*_ϕ in (4.4) further imply that

$$R^\epsilon_{\phi,*} + \delta \geq \int S_\epsilon(\phi \circ f_n) d\mathbb{P}_1 + \int S_\epsilon(\phi \circ -f_n) d\mathbb{P}_0 \geq \int \phi \circ f_n d\mathbb{P}_1^* + \int \phi \circ -f_n d\mathbb{P}_0^* \geq R^\epsilon_{\phi,*} \tag{4.14}$$

As $R^\epsilon_{\phi,*} = \int C^*_\phi(\eta^*) d\mathbb{P}^*$, this relation immediately implies (4.12).

Next, Lemma 76 again implies that

$$\int S_\epsilon(\phi \circ f_n) d\mathbb{P}_1 \geq \int \phi \circ f_n d\mathbb{P}_1^* \quad \text{and} \quad \int S_\epsilon(\phi \circ -f_n) d\mathbb{P}_0 \geq \int \phi \circ -f_n d\mathbb{P}_0^* \tag{4.15}$$

while (4.14) implies that

$$R_{\phi,*}^\epsilon - \int \phi \circ f_n d\mathbb{P}_1^* + \int \phi \circ -f_n d\mathbb{P}_0^* \leq 0.$$

Therefore, subtracting $\int \phi \circ f_n d\mathbb{P}_1^* + \int \phi \circ -f_n d\mathbb{P}_0^*$ from (4.14) results in

$$\delta \geq \left(\int S_\epsilon(\phi \circ f_n) d\mathbb{P}_1 - \int \phi \circ f_n d\mathbb{P}_1^* \right) + \left(\int S_\epsilon(\phi \circ -f_n) d\mathbb{P}_0 - \int \phi \circ -f_n d\mathbb{P}_0^* \right) \geq 0. \tag{4.16}$$

Again, (4.15) implies that the quantities on parentheses are both positive which implies (4.13).

□

Proposition 79 shows that minimizing sequences of R_ϕ^ϵ satisfy two properties: 1) The sequence $\{f_n\}$ must minimize the *standard* ϕ-risk R_ϕ with measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ in place of $\mathbb{P}_0, \mathbb{P}_1$, 2) At the limit, the measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ are best adversarial attacks on $\phi \circ f_n, \phi \circ -f_n$. In fact, one can show that $\{f_n\}$ is a minimizing sequence of R_ϕ^ϵ *if and only if* it satisfies these properties. Crucially, a very similar characterization holds for minimizers of the adversarial classification loss. We state and prove the 'only if' direction of this characterization in Proposition 80.

Proposition 80. *Let f_n be a sequence and let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ be measures in $\mathcal{B}_\epsilon^\infty(\mathbb{P}_i)$. Define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$, $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. If the following two conditions hold:*

$$\lim_{n\to\infty} \int C(\eta^*, f_n) d\mathbb{P}^* = \int C^*(\eta^*) d\mathbb{P}^* \tag{4.17}$$

$$\lim_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 - \int \mathbf{1}_{f_n \leq 0} d\mathbb{P}_1^* = 0, \quad \lim_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n > 0}) d\mathbb{P}_0 - \int \mathbf{1}_{f_n > 0} d\mathbb{P}_0^* = 0, \tag{4.18}$$

then f_n is a minimizing sequence of R^ϵ.

Proof. Equation 4.17 implies that the limit $\lim_{n\to\infty} C(\eta^*, f_n)d\mathbb{P}^*$ exists. Thus (4.17) and (4.18) imply that

$$\begin{aligned}\lim_{n\to\infty} R^\epsilon(f_n) &= \lim_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n\leq 0})d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_{f_n>0})d\mathbb{P}_0 = \lim_{n\to\infty} \int \mathbf{1}_{f_n\leq 0}d\mathbb{P}_1^* + \int \mathbf{1}_{f_n>0}d\mathbb{P}_0^* \\ &= \lim_{n\to\infty} \int C(\eta^*, f_n)d\mathbb{P}^* = \int C^*(\eta^*)d\mathbb{P}^* = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)\,.\end{aligned}$$

Therefore, Strong duality (Theorem 77) then implies that

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

and therefore, f_n is a minimizing sequence. □

We end this section by comparing the different criteria for consistency presented in Proposition 71 with Propositions 79 and 80. Together, Propositions 79 and 80 will allow us to compare minimizing sequences of R_ϕ^ϵ to those of R^ϵ by showing that any sequence satisfying (4.12)–(4.13) must also satisfy (4.17)–(4.18). This statement is the analog to 2) of Proposition 71. Indeed, because $C_\phi(\eta^*, f_n) \geq C_\phi^*(\eta^*)$, (4.12) is actually equivalent to to $C_\phi(\eta^*, f_n) \to C_\phi^*(\eta^*)$ in $L^1(\mathbb{P}^*)$. However, the extra criterion (4.18) implies an additional constraint on the structure of the minimizing sequence. This additional constraint is the reason 3) of Proposition 71 is false in the adversarial setting. In the restricted situation where $\bar{R}_\phi = \bar{R}$, Meunier et al. [42] show that (4.12) implies (4.17) (Proposition 4.2). However, this observation does not suffice to conclude consistency.

4.4.2 Adversarial Consistency

We are now in a position to prove consistency. Before presenting the full proof, we pause to discuss the overall strategy. Consistency will follow from three considerations. First, every minimizing sequence of R_ϕ^ϵ satisfies conditions (4.12) and (4.13). Second, conditions (4.12) and (4.13) imply the very similar conditions (4.17) and (4.18). Finally, any function sequence satisfying (4.17) and (4.18) must be a minimizing sequence to R^ϵ. The first and last steps are the content of Propositions 79 and 80, so it remains to justify the middle step.

Verifying that (4.12) implies (4.17) is straightforward. The relation (4.12) actually states that f_n minimizes the *standard* surrogate risk with respect to the distribution given by $\mathbb{P}_0^*$, $\mathbb{P}_1^*$. Therefore (4.12) implies (4.17) so long as ϕ is consistent.

The main difficulty is verifying (4.18), due to the discontinuity of $\mathbf{1}_{\alpha<0}$, $\mathbf{1}_{\alpha\geq 0}$ at 0. Due to this discontinuity, one cannot directly argue that (4.13) implies (4.18): to simplify the discussion, assume that ϕ is strictly decreasing on a neighborhood of the origin, in which case $\mathbf{1}_{\alpha<0} = \mathbf{1}_{\phi(\alpha)>\phi(0)}$ and $\mathbf{1}_{\alpha\geq 0} = \mathbf{1}_{\phi(-\alpha)\geq\phi(0)}$. Recall that according to (4.13), in the limit $n\to\infty$, $\mathbb{P}_0^*, \mathbb{P}_1^*$ are the strongest attack in $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\times\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$, or informally, $S_\epsilon(\phi\circ f_n)(\mathbf{x})$ approaches $\phi(f_n(\mathbf{x}'))$ for an optimal perturbation $\mathbf{x}'$ w.h.p., with a similar condition for $\phi\circ -f_n$. However, due to the discontinuity of $\mathbf{1}_{\phi(-\alpha)\geq\phi(0)}$ at $\phi(0)$, if $f_n(\mathbf{x}')\to 0$ as $n\to\infty$, this relation does not imply that $\mathbf{1}_{S_\epsilon(\phi\circ -f_n)(\mathbf{x})\geq\phi(0)}$ approaches $\mathbf{1}_{\phi\circ -f_n(\mathbf{x}')\geq 0}$.

Lemma 75 says that if $C_\phi^*(1/2) < \phi(0)$, minimizers of $C_\phi(\eta,\cdot)$ are uniformly bounded away from 0. This fact suggests that minimizing sequences will also be bounded away from the origin, which will allow us to avoid the discontinuity there. Concretely, we show:

Lemma 81. *Let $C_\phi^*(1/2) < \phi(0)$. Then there is a $\delta>0$ and a $c>0$ with $\phi(c)<\phi(0)$ for which $\alpha\in[-c,c]$ implies $C_\phi(\eta,\alpha)\geq C_\phi^*(\eta)+\delta$, uniformly in η. Furthermore, for this value of c, if $\alpha>c$ then $\phi(\alpha)<\phi(c)$.*

We prove this lemma in Appendix C.3. Because $C_\phi(\eta^*, f_n)\to C_\phi^*(\eta^*)$ in $L^1(\mathbb{P}^*)$, Lemma 81

implies that

$$\lim_{n\to\infty} \mathbb{P}^*(f_n \in [-c, c]) = 0. \tag{4.19}$$

This relation is the key fact that allows us to show that (4.13) implies (4.18). The condition $C^*_\phi(1/2) < \phi(0)$ is essential for this step of the argument.

Lastly, Lemma 76 implies that $\int S_\epsilon(\mathbf{1}_{f_n \geq 0}) d\mathbb{P}_1 \geq \int \mathbf{1}_{f_n \geq 0} d\mathbb{P}_1^*$ and thus to validate (4.18), it suffices to verify the opposite inequality in the limit $n \to \infty$.

Lemma 82. *Let f_n be a sequence of functions and let $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$. The equation*

$$\limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \liminf_{n\to\infty} \int \mathbf{1}_{f_n \leq 0} d\mathbb{P}_1^* \tag{4.20}$$

implies the first relation of (4.18) and

$$\limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n > 0}) d\mathbb{P}_0 \leq \liminf_{n\to\infty} \int \mathbf{1}_{f_n > 0} d\mathbb{P}_0^* \tag{4.21}$$

implies the second relation of (4.18).

See Appendix C.6 for a proof. These considerations suffice to prove the main result of this paper:

Theorem 83. *The loss ϕ is adversarially consistent if and only if $C^*_\phi(1/2) < \phi(0)$.*

Proof. The 'only if' portion of the statement is Proposition 73.

To show the 'if' statement, recall the standard analysis fact: $\lim_{n\to\infty} a_n = a$ iff for all subsequences $\{a_{n_j}\}$ of $\{a_n\}$, there is a further subsequence $a_{n_{j_k}}$ for which $\lim_{k\to\infty} a_{n_{j_k}} = a$. This result implies that to prove R^ϵ_ϕ is consistent, it suffices to show that every minimizing sequence f_n of R^ϵ_ϕ has a subsequence f_{n_j} that minimizes R^ϵ.

Let f_n be a minimizing sequence of R^ϵ_ϕ. For convenience, pick a subsequence f_{n_j} for which the limits $\lim_{j\to\infty} \int S_\epsilon(\mathbf{1}_{f_{n_j} < 0}) d\mathbb{P}_0$, $\lim_{j\to\infty} \int S_\epsilon(\mathbf{1}_{f_{n_j} \geq 0}) d\mathbb{P}_1$ both exist. For notational clarity, we drop the $_j$ subscript and denote this sequence as f_n.

By Proposition 79, the equations (4.12) and (4.13) hold. We will argue that f_n is in fact a minimizing sequence of R^ϵ by verifying the conditions of Proposition 80.

First, the relation (4.12) states that the sequence f_n minimizes the *standard* ϕ-risk for the distribution given by $\mathbb{P}_0^*$ and $\mathbb{P}_1^*$. As the loss ϕ is consistent by Proposition 74, the sequence f_n must minimize the standard classification risk for the distribution $\mathbb{P}_0^*, \mathbb{P}_1^*$. This statement implies (4.17). Next we will argue that (4.18) holds.

Let c, δ be as in Lemma 81. Because $C_\phi(\eta^*, f_n) \geq C_\phi^*(\eta^*)$, (4.12) implies that $C_\phi(\eta^*, f_n)$ converges to $C_\phi^*(\eta^*)$ in L^1. However, L^1 convergence implies convergence in measure (see for instance Proposition 2.29 of [22]), and therefore $\lim_{n\to\infty} \mathbb{P}^*\big(C_\phi(\eta^*, f_n) > C_\phi^*(\eta^*) + \delta\big) = 0$. Lemma 81 then implies that for $i = 0, 1$

$$\lim_{n\to\infty} \mathbb{P}_i^*(f_n \in [-c, c]) = 0. \tag{4.22}$$

Next, because ϕ is non-increasing, $f \leq 0$ implies $\phi(f) \geq \phi(0)$ and thus $\mathbf{1}_{f\leq 0} \leq \mathbf{1}_{\phi\circ f \geq \phi(0)}$. Furthermore, as the function $\alpha \mapsto \mathbf{1}_{\alpha\geq 0}$ is monotone and upper semi-continuous,

$$\int S_\epsilon(\mathbf{1}_{f_n\leq 0}) d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_{\phi\circ f_n \geq \phi(0)}) d\mathbb{P}_1 \leq \int \mathbf{1}_{S_\epsilon(\phi\circ f_n)\geq\phi(0)} d\mathbb{P}_1. \tag{4.23}$$

Let γ_i be a coupling between $\mathbb{P}_i$ and $\mathbb{P}_i^*$ for which $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma_i} \|\mathbf{x}-\mathbf{y}\| \leq \epsilon$. Then the measure γ_i is supported on $\Delta_\epsilon = \{(\mathbf{x}, \mathbf{y})\colon \|\mathbf{x} - \mathbf{y}\| \leq \epsilon\}$. Furthermore, as $S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi \circ f_n(\mathbf{x}')$ everywhere on Δ_ϵ, the relation $S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi \circ f_n(\mathbf{x}')$ actually holds γ_1-a.e. Therefore, (4.13) actually implies that $S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi \circ f_n(\mathbf{x}')$ converges in γ_1-measure to 0. In particular, since $\phi(c) < \phi(0)$, $\lim_{n\to\infty} \gamma_1\big(S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi(f_n(\mathbf{x}')) \geq \phi(0) - \phi(c)\big) = 0$ and

thus $\lim_{n\to\infty} \gamma_1(S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi(0) \cap \phi \circ f_n(\mathbf{x}') < \phi(c)) = 0$. Therefore,

$$\liminf_{n\to\infty} \mathbb{P}_1(S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi(0)) = \liminf_{n\to\infty} \gamma_1(S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi(0) \cap \phi \circ f_n(\mathbf{x}') \geq \phi(c))$$
$$\leq \liminf_{n\to\infty} \gamma_1(\phi \circ f_n(\mathbf{x}') \geq \phi(c)) = \liminf_{n\to\infty} \mathbb{P}_1^*(\phi \circ f_n(\mathbf{x}') \geq \phi(c))$$

This calculation implies

$$\liminf_{n\to\infty} \int \mathbf{1}_{S_\epsilon(\phi\circ f_n)(\mathbf{x})\geq\phi(0)} d\mathbb{P}_1 \leq \liminf_{n\to\infty} \int \mathbf{1}_{\phi\circ f_n(\mathbf{x}')\geq\phi(c)} d\mathbb{P}_1^* \leq \liminf_{n\to\infty} \int \mathbf{1}_{f_n\leq c} d\mathbb{P}_1^* \tag{4.24}$$

The last inequality follows because Lemma 81 states that $\alpha > c$ implies $\phi(\alpha) < \phi(c)$ and therefore $\mathbf{1}_{\phi\circ f_n\geq\phi(c)} \leq \mathbf{1}_{f_n\leq c}$. Equation 4.22 then implies

$$\liminf_{n\to\infty} \int \mathbf{1}_{\phi\circ f_n\geq\phi(0)} d\mathbb{P}_1^* \leq \liminf_{n\to\infty} \int \mathbf{1}_{f_n\leq c} d\mathbb{P}_1^* = \liminf_{n\to\infty} \int \mathbf{1}_{f_n\leq -c} d\mathbb{P}_1^*. \tag{4.25}$$

Recall that the sequence f_n was chosen so that the limit $\lim_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n\leq 0}) d\mathbb{P}_1$ exists. Combining this fact with (4.23), (4.24), and (4.25) results in

$$\limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n\leq 0}) d\mathbb{P}_1 \leq \liminf_{n\to\infty} \int \mathbf{1}_{f_n\leq -c} d\mathbb{P}_1^* \leq \liminf_{n\to\infty} \int \mathbf{1}_{f_n\leq 0} d\mathbb{P}_1^* \tag{4.26}$$

The first relation of (4.18) then follows from (4.26) together with Lemma 82.

A similar argument implies the second relation of (4.18). Because $\mathbf{1}_{f>0} = \mathbf{1}_{-f<0} \leq \mathbf{1}_{-f\leq 0}$, the same chain of inequalities as (4.23), (4.24), and (4.25) implies that

$$\limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n>0}) d\mathbb{P}_0 \leq \limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{-f_n\leq 0}) d\mathbb{P}_0 \leq \liminf_{n\to\infty} \int \mathbf{1}_{-f_n\leq -c} d\mathbb{P}_0^* = \liminf_{n\to\infty} \int \mathbf{1}_{f_n\geq c} d\mathbb{P}_0^*$$

As $c > 0$, it follows that $\limsup_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n>0}) d\mathbb{P}_0 \leq \liminf_{n\to\infty} \int \mathbf{1}_{f_n>0} d\mathbb{P}_0^*$. Once again, the second expression of (4.18) follows from this relation and Lemma 82. □

4.5 Quantitative Bounds for the ρ-Margin Loss

As discussed in the introduction, statistical consistency is not the only property one would want from a surrogate. Hopefully, minimizing a surrogate will also efficiently minimize the classification loss. Bartlett, Jordan, and McAuliffe [8], Reid and Williamson [53], and Steinwart [57] prove bounds of the form $R(f) - R_* \leq G_\phi(R^*_\phi(f) - R_{\phi,*})$ for a function G_ϕ and $R_* = \inf_f R(f)$, $R_{\phi,*} = \inf_f R_\phi(f)$. The function G_ϕ is an upper bound on the rate of convergence of the classification risk in terms of the rate of convergence of the surrogate risk. One would hope that G_ϕ is not logarithmic, as such a bound could imply that reducing $R(f) - R_*$ by a quantity Δ could require an exponential change of e^Δ in $R_\phi(f) - R_{\phi,*}$. Bartlett, Jordan, and McAuliffe [8] compute such G_ϕ for several popular losses in the standard classification setting. For example, they show the bounds $G_\phi(\theta) = \theta$ for the hinge loss $\phi(\alpha) = (1-\alpha)_+$ and $G_\phi(\theta) = \sqrt{\theta}$ for the squared hinge loss $\phi(\alpha) = (1-\alpha)_+^2$. On can prove an analogous bound for the ρ-margin loss in the adversarial setting:

Theorem 84. *Let $\phi_\rho = \min(1, \max(1 - \alpha/\rho, 0))$ be the ρ-margin loss, $R^\epsilon_* = \inf_f R^\epsilon(f)$, and $R^\epsilon_{\phi_\rho,*}(f) = \inf_f R^\epsilon_{\phi_\rho}(f)$. Then*

$$R^\epsilon(f) - R^\epsilon_* \leq R^\epsilon_{\phi_\rho}(f) - R^\epsilon_{\phi_\rho,*}.$$

Notice that this theorem immediately implies that the ρ-margin loss is in fact adversarially consistent. The proof below is completely independent of the argument in Section 4.4.

Proof. Notice that for the ρ-margin loss, $C^*_{\phi_\rho} = C^*$ and therefore, the optimal ϕ_ρ-risk $R^\epsilon_{\phi_\rho,*}$ equals the optimal adversarial classification risk R^ϵ_*. However, since $\phi_\rho(\alpha) \geq \mathbf{1}_{\alpha \leq 0}$ and $\phi_\rho(-\alpha) \geq \mathbf{1}_{\alpha > 0}$ for any α, one can conclude that $R^\epsilon(f) \leq R^\epsilon_{\phi_\rho}(f)$. Therefore,

$$R^\epsilon(f) - R^\epsilon_* = R^\epsilon(f) - R^\epsilon_{\phi_\rho,*} \leq R^\epsilon_{\phi_\rho}(f) - R^\epsilon_{\phi_\rho,*}$$

□

This bound implies that reducing the excess adversarial ρ-margin loss by Δ also reduces an upper bound on the excess adversarial classification loss by Δ. Thus, one would expect that minimizing the adversarial ρ-margin risk would be an effective procedure for minimizing the adversarial classification risk.

Extending Theorem 84 to other losses remains an open problem. In the non-adversarial scenario, many prior works develop techniques for computing such bounds. These include the Ψ-transform of [8], calibration analysis in [57], and special techniques for proper losses in [53].

Contemporary work [40] derives an $\mathcal{H}$-consistency surrogate risk bound for a variant of the adversarial ρ-margin loss.

4.6 Conclusion

In conclusion, we proved that the adversarial training procedure is consistent for perturbations in an ϵ-ball if an only if $C^*_\phi(1/2) < \phi(0)$. The technique that proved consistency extends to perturbation sets which satisfy existence and minimax theorems analogous to Theorems 77 and 78. Furthermore, we showed a quantitative excess risk bound for the adversarial ρ-margin loss. Finding such bounds for other losses remains an open problem. We hope that insights to consistency and the structure of adversarial learning will lead to the design of better adversarial learning algorithms.

Acknowledgements

Natalie Frank was supported in part by the Research Training Group in Modeling and Simulation funded by the National Science Foundation via grant RTG/DMS – 1646339.

Jonathan Niles-Weed was supported in part by a Sloan Research Fellowship.

5 — Adversarial Consistency and the Uniqueness of the Adversarial Bayes Classifier

5.1 Introduction

Robustness is a core concern in machine learning, as models are deployed in classification tasks such as facial recognition [72], medical imaging [47], and identifying traffic signs in self-driving cars [19]. Deep learning models exhibit a concerning security risk— small perturbations imperceptible to the human eye can cause a neural net to misclassify an image [14, 58]. 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 datapoints— also known as the *classification risk*. Minimizers to the classification risk are easy to compute analytically, and are known as *Bayes classifiers*. In the adversarial setting, each point is perturbed by a malicious adversary before the classifier makes a prediciton. The proportion of mislabeled data under such an attack is called the *adversarial classification risk*, and minimizers to this risk are called *adversarial Bayes classifiers*. Unlike the standard

classification setting, computing minimizers to the adversarial classification risk is a non-trivial task [11, 51]. Further studies [23, 29, 52, 62, 64] investigate additional properties of these minimizers, and Frank [23] 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 object and minimizing this quantity is computationally intractable. Instead, typical machine learning algorithms minimize a *surrogate risk* in place of the classification error. In the robust setting, the adversarial training algorithm uses a surrogate risk that computes the supremum of loss over the adversary's possible attacks, which we refer to as *adversarial surrogate risks*. However, one must verify that minimizing this adversarial surrogate will also minimize the classification risk. A loss function is *adversarially consistent* for a particular data distribution if every minimizing sequence of the associated adversarial surrogate risk also minimizes the adversarial classification risk. A loss is simply called *adversarially consistent* if it is adversarially consistent for all possible data distributions. Meunier et al. [42] show that no convex surrogate is adversarially consistent, in contrast to the standard classification setting where most convex losses are statistically consistent [8, 38, 43, 57, 75].

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

Frank [23] further demonstrates several distributions for which the adversarial Bayes classifier is unique, and thus a convex loss would be consistent. Understanding general conditions under which uniqueness occurs is an open question.

5.2 Related Works

Our results are inspired by prior work which showed that no convex loss is adversarially consistent [5, 42] yet a wide class of adversarial losses is adversarially consistent [26]. These consistency results rely on the theory of surrogate losses, studied by Bartlett, Jordan, and McAuliffe [8] and Lin [38] in the standard classification setting and by Frank and Niles-Weed [25] and Li and Telgarsky [36] in the adversarial setting. Furthermore, [3, 6, 57] study a property of related to consistency called *calibration*, which [42] relate to consistency. Complimenting this analysis, another line of research studies $\mathcal{H}$-consistency, which refines the concept of consistency to specific function classes [5, 49]. Our proof combines results on losses with minimax theorems for various adversarial risks, as studied by [25, 26, 52, 63]. Lastly, this work leverages recent results on the adversarial Bayes classifier, which are extensively studied by [11, 23, 51, 63].

5.3 Notation and Background

5.3.1 Surrogate Risks

This paper investigates binary classification on $\mathbb{R}^d$ with labels $\{-1, +1\}$. Class -1 is distributed according to a a measure $\mathbb{P}_0$ and while class $+1$ is distributed according to measure $\mathbb{P}_1$. A *classifier* is a Borel set A 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:

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

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

quantities is

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

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

$$C(\eta, b) = \eta b + (1-\eta)(1-b), \quad C^*(\eta) = \inf_{b \in \{0,1\}} C(\eta, b) = \min(\eta, 1-\eta). \tag{5.1}$$

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

$$\{\mathbf{x} : \eta(\mathbf{x}) > 1/2\} \quad \text{and} \quad \{\mathbf{x} : \eta(\mathbf{x}) \geq 1/2\} \tag{5.2}$$

are both Bayes classifiers.

While the Bayes classifier can be described mathematically, minimizing the empirical classification risk is a computationally intractable problem [9]. A common approach is to instead minimize a better-behaved alternative called a *surrogate risk*. As a surrogate to R, we consider:

$$R_\phi(f) = \int \phi(f) d\mathbb{P}_1 + \int \phi(-f) d\mathbb{P}_0 \,. \tag{5.3}$$

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

Assumption 3. *The loss ϕ is non-increasing, continuous, and* $\lim_{\alpha \to \infty} \phi(\alpha) = 0$.

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

$$R(f) = R(\{f > 0\}) \tag{5.4}$$

for a function f.

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

$$R_\phi(f) = \int C_\phi(\eta(\mathbf{x}), f(\mathbf{x}))d\mathbb{P} \tag{5.5}$$

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

$$C_\phi(\eta, \alpha) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha), \quad C_\phi^*(\eta) = \inf_\alpha C_\phi(\eta, \alpha) \tag{5.6}$$

for $\eta \in [0, 1]$. Thus a minimizer f of R_ϕ must minimize $C_\phi(\eta(\mathbf{x}), \cdot)$ almost everywhere according to the probability measure $\mathbb{P}$. The following lemma describes a method for mapping $\eta(\mathbf{x})$ to a minimizer of $C_\phi(\eta(\mathbf{x}), \cdot)$.

Lemma 85. *The function $\alpha_\phi : [0, 1] \to \overline{\mathbb{R}}$ that maps η to the smallest minimizer of $C_\phi(\eta, \cdot)$ is non-decreasing.*

See Appendix D.1 for a proof. Because α_ϕ is monotonic, the composition

$$\alpha_\phi(\eta(\mathbf{x})) \tag{5.7}$$

is always measurable, and thus this function is a minimizer of R_ϕ. Allowing for minimizers in extended real numbers $\overline{\mathbb{R}} = \{-\infty, +\infty\} \cup \mathbb{R}$ is necessary for certain losses— for instance when ϕ is the exponential loss, then $C_\phi(1, \alpha) = e^{-\alpha}$ does not assume its infimum on $\mathbb{R}$.

5.3.2 Adversarial Surrogate Risks

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

written in terms of a supremum. For any function $g : \mathbb{R}^d \to \mathbb{R}$, define

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

where $\overline{B_\epsilon(\mathbf{x})} = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\| \leq \epsilon\}$ 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 Equation (5.4), 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 as[2]

$$R^\epsilon_\phi(f) = \int S_\epsilon(\phi \circ f) d\mathbb{P}_1 + \int S_\epsilon(\phi \circ -f) d\mathbb{P}_0. \tag{5.8}$$

Theorem 9 of [25] then extends the construction of a minimizer in Equation (5.7) to the adversarial setting.

Theorem 86. *Let α_ϕ be the function in Lemma 85. Then for any distribution $\mathbb{P}_0$, $\mathbb{P}_1$, there is a function $\hat{\eta} : \mathbb{R}^d \to [0, 1]$ for which $\alpha_\phi(\hat{\eta}(\mathbf{x}))$ is a minimizer of R^ϵ_ϕ for any loss ϕ.*

The function $\hat{\eta}$ can be viewed as the conditional probability of label $+1$ under an 'optimal'

[1] The functions $S_\epsilon(\mathbf{1}_A)$, $S_\epsilon(\mathbf{1}_{A^C})$ must be measurable in order to define this integral. See [25, Section 3.3] for a treatment of this matter.

[2] Again, see See [25, Section 3.3] for a treatment of measurability.

adversarial attack [25]. Just as in the standard learning scenario, the function $\alpha(\hat{\eta}(\mathbf{x}))$ may be $\overline{\mathbb{R}}$-valued. Furthermore, recall that Bayes classifiers can be constructed by theshholding the conditional probability η at $1/2$, as in Equation (5.2). The function $\hat{\eta}$ plays an analogous role for adversarial learning.

Theorem 87. *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be finite measures and let $\hat{\eta}$ be the function described by Theorem 86. Then the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are adversarial Bayes classifiers. Furthermore, any adversarial Bayes classifier A satisfies*

$$\int S_\epsilon(\mathbf{1}_{\{\hat{\eta}\geq 1/2\}^C})d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2)^C})d\mathbb{P}_1 \tag{5.9}$$

and

$$\int S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2\}})d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_{\{\hat{\eta}\geq 1/2\}})d\mathbb{P}_0 \tag{5.10}$$

See Appendix D.3 for a proof and more about the function $\hat{\eta}$. Equations (5.9) and (5.10) imply that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ can be viewed as 'minimal' and 'maximal' adversarial Bayes classifiers.

5.3.3 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 88.

- *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.

[3]This concept is referred to as *calibration* in the non-adversarial machine learning context [8, 57]. We use the term 'consistent', as prior work on adversarial learning [4, 42] use 'calibration' to refer to a different but related concept.

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

A case of particular interest is convex ϕ, as these losses are ubiquitous in machine learning. In the non-adversarial context, Theorem 2 of [8] shows that a convex loss ϕ is consistent iff ϕ is differentiable at zero and $\phi'(0) < 0$. In contrast, Meunier et al. [42] show that no convex loss is adversarially consistent. Further results of [26] characterize the adversarially consistent losses in terms of the function C^*_ϕ:

Theorem 89. *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,

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

due to convexity. Notice that Theorem 89 does not preclude the adversarial consistency of a loss satisfying $C^*_\phi(1/2) = \phi(0)$ for any particular $\mathbb{P}_0, \mathbb{P}_1$. Prior work [26, 42] 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 $C^*_\phi(1/2) = \phi(0)$.

5.4 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 93 below). Frank [23] further develops a notion of uniqueness for adversarial Bayes classifiers.

Definition 90. *The adversarial Bayes classifiers A_1 and A_2 are* equivalent up to degeneracy *if any Borel set A with $A_1 \cap A_2 \subset A \subset A_1 \cup A_2$ is also an adversarial Bayes classifier. The adversarial Bayes classifier is* unique up to degeneracy *if any two adversarial Bayes classifiers are unique up to degeneracy.*

When $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, then equivalence up to degeneracy is an equivalence relation [23, Theorem 3.3]. The central result of this paper relates the consistency of convex losses to the uniqueness of the adversarial Bayes classifier.

Theorem 91. *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.*

Frank [23] provides the tools for verifying when the adversarial Bayes classifier is unique up to degeneracy for a wide class of one dimensional distributions. Below we highlight two interesting examples. Let p_1 be the density of $\mathbb{P}_1$ and p_0 be the density of $\mathbb{P}_0$.

- Consider mean zero gaussians with different variances: $p_0(x) = \frac{1}{2\sqrt{2\pi}\sigma_0} e^{-x^2/2\sigma_0^2}$ 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 for this distribution [23, Example 4.1].

- 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 $\epsilon < |\mu_1 - \mu_0|/2$ [23, Example 4.2].

Theorem 91 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 [23, Example 4.5] provide 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. Understanding when the adversarial Bayes classifier is unique up to degeneracy for reasonable distributions is an open problem.

5.5 Uniqueness up to Degeneracy implies Consistency

The proof of the forward direction in Theorem 91 relies on a dual formulation of the adversarial classification problem involving the Wasserstein-∞ metric. This tool is presented in the next section and is then used to prove the forward direction of Theorem 91 in Section 5.5.2.

5.5.1 Background— A Dual Problem for the Adversarial Classification Risk

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

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

Theorem 2.6 of [33] 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 92. *Let E be a Borel set. Then*

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

See Appendix D.2 for a proof. Consequently,

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

Does equality hold and can one swap the infimum and the supremum? [26, 52] answer this question in the affirmative:

Theorem 93. *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 Equation (5.1). Then

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

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

See Theorem 1 of [26] for a proof. Theorems 6, 8, and 9 of [25] show an analogous minimax theorem for surrogate risks.

Theorem 94. *Let $\mathbb{P}_0$, $\mathbb{P}_1$ be finite Borel measures. Define*

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

where the function C_ϕ^* is defined in Equation (5.6). Then

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

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

Just like R_ϕ, the risk R_ϕ^ϵ may not have an $\mathbb{R}$-valued minimizer. However, Lemma 8 of [26] states that

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

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

$$I_\epsilon(g) = \inf_{\mathbf{x}' \in \overline{B_\epsilon(\mathbf{x})}} g(\mathbf{x}') \tag{5.11}$$

Lemma 24 of [25] proves the following result:

Theorem 95. *There exists a function $\hat{\eta} : \mathbb{R}^d \to [0,1]$ and measures $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ for which*

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

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

This result implies Theorem 86: Item I) and Item II) imply that $R_\phi^\epsilon(\alpha_\phi(\hat{\eta})) = \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*)$ and Theorem 94 then implies that $\alpha_\phi(\hat{\eta})$ is a minimizer of R_ϕ^ϵ and $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ maximize $\bar{R}_\phi$.

(A similar argument is given in the proof of Lemma 180 of Appendix D.6.1 in this paper.) Furthermore, the relation $R^\epsilon_\phi(\alpha_\phi(\hat{\eta})) = \bar{R}_\phi(\mathbb{P}^*_0, \mathbb{P}^*_1)$ also implies

Lemma 96. *The $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ of Theorem 95 maximize $\bar{R}_\phi$ over $\mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \times \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$ for every ϕ.*

See [25, Lemma 26] for more details. Theorem 87 is proved analogously to Theorem 86 in Appendix D.3– Item I) and Item II) imply that $R^\epsilon(\{\hat{\eta} > 1/2\}) = \bar{R}(\mathbb{P}^*_0, \mathbb{P}^*_1) = R^\epsilon(\hat{\eta} \geq 1/2\})$ and consequently Theorem 93 implies that $\{\hat{\eta} > 1/2\}$, $\{\hat{\eta} \geq 1/2\}$ minimize R^ϵ and $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ maximize $\bar{R}$. Lastly, uniqueness up to degeneracy can be characterized in terms of these $\mathbb{P}^*_0$,$\mathbb{P}^*_1$.

Theorem 97. *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 95.

See Appendix D.4 for a proof of Theorem 97.

5.5.2 Proving that Uniqueness implies Consistency

Before presenting the full proof of consistency, we provide an overview the strategy of this argument. Approximate complementary slackness conditions derived in [26] describe minimizing sequences of R^ϵ_ϕ.

Proposition 98. *Assume that the measures $\mathbb{P}^*_0 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_0)$, $\mathbb{P}^*_1 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$ maximize $\bar{R}_\phi$. Then any minimizing sequence f_n of R^ϵ_ϕ must satisfy*

$$\lim_{n\to\infty} \int C_\phi(\eta^*, f_n) d\mathbb{P}^* = \int C^*_\phi(\eta^*) d\mathbb{P}^* \tag{5.12}$$

$$\lim_{n\to\infty}\int S_\epsilon(\phi\circ f_n)d\mathbb{P}_1-\lim_{n\to\infty}\int \phi\circ f_n d\mathbb{P}_1^*=0,\quad \lim_{n\to\infty}\int S_\epsilon(\phi\circ -f_n)d\mathbb{P}_0^*-\lim_{n\to\infty}\int \phi\circ -f_n d\mathbb{P}_0^*=0 \tag{5.13}$$

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

We will show that when $\mathbb{P}^*(\eta^*=1/2)=0$, every sequence of functions satisfying Equation (5.12) and Equation (5.13) must minimize R^ϵ. Specifically, we will prove that every minimizing sequence f_n of R_ϕ^ϵ must satisfy

$$\limsup_{n\to\infty}\int S_\epsilon(\mathbf{1}_{f_n\le 0})d\mathbb{P}_1\le\int \mathbf{1}_{\eta^*\le\frac{1}{2}}d\mathbb{P}_1^* \tag{5.14}$$

and

$$\limsup_{n\to\infty}\int S_\epsilon(\mathbf{1}_{f_n\ge 0})d\mathbb{P}_0\le\int \mathbf{1}_{\eta^*\ge\frac{1}{2}}d\mathbb{P}_0^* \tag{5.15}$$

for the measures $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ in Theorem 95. Consequently, $\mathbb{P}^*(\eta^*=1/2)=0$ implies that $\limsup_{n\to\infty}R^\epsilon(f_n)\le\bar{R}(\mathbb{P}_0^*,\mathbb{P}_1^*)$ and the strong duality relation in Theorem 93 implies that f_n must in fact be a minimizing sequence of R^ϵ.

We next describe the proof of Equation (5.14). 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 $C_\phi(\eta^*,f_n)\ge C_\phi^*(\eta^*)$, the condition Equation (5.12) implies that $C_\phi(\eta^*,f_n)$ converges to $C_\phi^*(\eta^*)$ in $L^1(\mathbb{P}^*)$, and the assumption that there is a single value of 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 92 states that $S_\epsilon(\phi\circ f_n)(\mathbf{x})\ge\phi\circ f_n(\mathbf{x}')$ γ_1^*-a.e., Equation (5.13) implies that $S_\epsilon(\phi\circ f_n)(\mathbf{x})-\phi\circ f_n(\mathbf{x}')$ converges to 0 in $L^1(\gamma_1^*)$. Consequently $S_\epsilon(\phi\circ f_n)(\mathbf{x})$ must converge to $\phi(\alpha_\phi(\eta^*(\mathbf{x}')))$ in $L^1(\gamma_1^*)$. As L^1 convergence

implies convergence in measure [22, Proposition 2.29], one can conclude that

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

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

$$\limsup_{n\to\infty} \int \mathbf{1}_{S_\epsilon(\phi(f_n))(\mathbf{x}) \ge \phi(0)} d\gamma_1^* \le \int \mathbf{1}_{\phi(\alpha_\phi(\eta^*(\mathbf{x}'))) < \phi(0) - c} d\gamma_1^* = \int \mathbf{1}_{\eta^* \ge \alpha_\phi^{-1} \circ \phi^{-1}(\phi(0) - c)} d\mathbb{P}_1^*. \tag{5.16}$$

Next, we will also assume that α_ϕ^{-1} is continuous and $\alpha_\phi(1/2) = 0$. (The exponential loss satisfies this assumption 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 smalaler 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 Equation (5.16) proves Equation (5.14).

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 99. *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 $|\eta - 1/2| \ge r$ and $\operatorname{sign}(\eta - 1/2)\alpha \le \alpha_r$ then $C_\phi(\eta, \alpha_r) - C_\phi^*(\eta) \ge k_r$, and this α_r satisfies $\phi(\alpha_r) < \phi(0)$.*

See Appendix D.5 for a proof. A minor modification of this argument proves our main result:

Proposition 100. *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 conditions of this propositon.

Proof. We will show that every minimizing sequence of R^ϵ_ϕ must satisfy Equation (5.14) and Equation (5.15). These equations together with the assumption $\mathbb{P}^*(\eta^* = 1/2) = 0$ imply that

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

The strong duality result of Theorem 93 then implies that f_n must be a minimizing sequence of R^ϵ.

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

$$\mathbb{P}^*(|\eta^* - 1/2| < r) < \delta. \tag{5.17}$$

Next, let α_r, k_r be as in Lemma 99.

Let γ_i^* be couplings between $\mathbb{P}_i$ and $\mathbb{P}_i^*$ supported on Δ_ϵ. Lemma 92 implies that $S_\epsilon(\phi \circ f_n)(\mathbf{x}) \geq \phi \circ f_n(\mathbf{x}')$ γ_1^*-a.e., and thus Equation (5.13) implies that $S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi \circ f_n(\mathbf{x}')$ converges to 0 in $L^1(\gamma_1^*)$. Because convergence in L^1 implies convergence in measure [22, Proposition 2.29], $S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi \circ f_n(\mathbf{x}')$ converges to 0 in γ_1^*-measure. Similarly, one can conclude that $S_\epsilon(\phi \circ -f_n)(\mathbf{x}) - \phi \circ -f_n(\mathbf{x}')$ converges to zero in γ_0^*-measure. Additionally, as $C^*_\phi(\eta^*, f_n) \geq C^*_\phi(\eta^*)$, Equation (5.12) implies that $C^*_\phi(\eta^*, f_n)$ converges in $\mathbb{P}^*$-measure to $C^*_\phi(\eta^*)$. Therefore, Proposition 98 implies that one can choose N large enough so that $n > N$ implies

$$\gamma_1^*\Big(S_\epsilon(\phi \circ f_n)(\mathbf{x}) - \phi \circ f_n(\mathbf{x}') \geq \phi(0) - \phi(\alpha_r)\Big) < \delta, \tag{5.18}$$

$$\gamma_0^*\Big(S_\epsilon(\phi \circ -f_n)(\mathbf{x}) - \phi \circ -f_n(\mathbf{x}') \geq \phi(0) - \phi(\alpha_r)\Big) < \delta, \tag{5.19}$$

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

that

$$\mathbb{P}^*\big(|\eta^* - 1/2| \geq r, f_n \operatorname{sign}(\eta^* - 1/2) \leq \alpha_r\big) < \delta \tag{5.20}$$

due to Lemma 99. 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 Equation (5.18) implies that for $n > N$, outside a set of γ_1^*-measure δ, $S_\epsilon(\phi \circ f_n)(\mathbf{x}) < (\phi \circ f_n)(\mathbf{x}') + \phi(0) - \phi(\alpha_r)$ and thus

$$\int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \gamma_1^*\big(\phi \circ f_n(\mathbf{x}') + \phi(0) - \phi(\alpha_r) > \phi(0)\big) + \delta \leq \mathbb{P}_1^*\big(\phi \circ f_n > \phi(\alpha_r)\big) + \delta \tag{5.21}$$

Next, the monotonicity of ϕ implies that $\mathbb{P}_1^*(\phi \circ f_n(\mathbf{x}') > \phi(\alpha_r)) \leq \mathbb{P}_1^*(f_n < \alpha_r)$ and thus Equation (5.17) implies

$$\int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 \leq \mathbb{P}_1^*(f_n < \alpha_r) + \delta \leq \mathbb{P}_1^*(f_n < \alpha_r, |\eta^* - 1/2| \geq r) + 2\delta. \tag{5.22}$$

Next, Equation (5.20) 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 Equation (5.14). Observe that $\mathbf{1}_{f \geq 0} = \mathbf{1}_{-f \leq 0}$, and thus the inequalities Equations (5.21) and (5.22) hold with $-f_n$ in place f_n, $\mathbb{P}_0$, $\mathbb{P}_0^*$, γ_0^* in place of $\mathbb{P}_1, \mathbb{P}_1^*$, γ_1^*, and Equation (5.19) in place of Equation (5.18) resulting in

$$\int S_\epsilon(\mathbf{1}_{f_n \geq 0}) d\mathbb{P}_0 \leq \mathbb{P}_0^*(f_n < -\alpha_r, |\eta^* - 1/2| \geq r) + 2\delta.$$

Next, Equation (5.20) implies $\mathbb{P}_1^*(\eta^* \leq 1/2 - r, f_n \geq -\alpha_r) < \delta$ and consequently

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

Because δ is arbitrary, this relation implies Equation (5.15).

□

5.6 Consistency Requires Uniqueness up to Degeneracy

We prove the reverse direction of Theorem 91 by constructing a sequence of of functions f_n that minimize R_ϕ^ϵ for which $R^\epsilon(f_n)$ is constant in n and not equal to the minimal adversarial Bayes risk.

Proposition 101. *Assume that $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous with respect to Lebesgue meaure the adversarial Bayes classifier is not unique up to degeneracy for the distribution $\mathbb{P}_0, \mathbb{P}_1$. Then any consistent loss ϕ satisfying $C_\phi^*(1/2) = \phi(0)$ is not adversarially consistent.*

First, Theorem 87 together with a result of [23] imply the adversarial Bayes classifier is unique iff $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are equivalent up to degeneracy, see Appendix D.6 for proof.

Lemma 102. *Assume $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure. Then adversarial Bayes classifier is unique up to degeneracy iff the adversarial Bayes classifiers $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are equivalent up to degeneracy.*

Therefore, if the adversarial Bayes classifier is not unique up to degeneracy, then there is a set $\tilde{A}$ that is not an adversarial Bayes classifier but $\{\hat{\eta} > 1/2\} \subset \tilde{A} \subset \{\hat{\eta} \geq 1/2\}$. Theorem 86 suggests that a minimizer of R_ϕ^ϵ can equal zero only when $\hat{\eta} = 1/2$. Thus we

select a sequence f_n that is strictly positive on $\tilde{A}$, strictly negative on $\tilde{A}^C$, and approaches 0 on $\{\hat{\eta} = 1/2\}$. Consider the sequence

$$f_n(\mathbf{x}) = \begin{cases} \alpha_\phi(\hat{\eta}(\mathbf{x})) & \hat{\eta}(\mathbf{x}) \neq 1/2 \\ \frac{1}{n} & \hat{\eta}(\mathbf{x}) = 1/2, \mathbf{x} \in \tilde{A} \\ -\frac{1}{n} & \hat{\eta}(\mathbf{x}) = 1/2, \mathbf{x} \notin \tilde{A} \end{cases} \tag{5.23}$$

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^ϵ_ϕ. However, f_n may assume the values $\pm\infty$ because the function α_ϕ is $\overline{\mathbb{R}}$-valued. A slight modification of these functions produces an $\mathbb{R}$-valued sequence that minimizes R^ϵ_ϕ but $R^\epsilon(f_n) = R^\epsilon(\tilde{A})$ for all n. See Appendix D.6 for a formal proof.

5.7 Conclusion

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

Acknowledgements

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.

6 — Conclusion

This thesis provides an array of tools for understanding adversarial risks. Insights from these tools include an explanation of the phenomenon of transfer attacks, formulas for minimizers of these risks, and a characterization of the consistency of these surrogate risks. Hopefully, the results from this research will assist in the development of algorithms for robust learning.

Appendices

A — Deferred Proofs from Chapter 2

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

A.1.1 Definition of the Universal σ-Algebra and Main Measurability Result

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

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

Let X be a separable metric space and let $\mathcal{B}(X)$ be the Borel σ-algebra on X. Denote the completion of $\mathcal{B}(X)$ with respect to a Borel measure ν by $\mathcal{L}_\nu(X)$. Let $\mathcal{M}_+(X)$ be the

set of all finite[1] positive Borel measures on X. Then the universal σ-algebra on X, $\mathscr{U}(X)$ is

$$\mathscr{U}(X) = \bigcap_{\nu \in \mathcal{M}_+(X)} \mathcal{L}_\nu(X). \tag{A.1}$$

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

Theorem 103. *If f is universally measurable and B is a compact set, then $S_B(f)$ is universally measurable.*

Thus, if $\mathbb{P}_0, \mathbb{P}_1$, and g are Borel, integrals of the form $\int S_\epsilon(g) d\mathbb{P}_i$ in (2.10) can be interpreted as the integral of $S_\epsilon(g)$ with respect to the completion of $\mathbb{P}_i$.

A.1.2 Proof Outline

To prove Theorem 103, we analyze the level sets of $S_B(g)$. One can compute the level set $[S_B(g)(\mathbf{x}) > a]$ using a direct sum.

Lemma 104. *Let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ be any function. For a set B, define $-B = \{-\mathbf{b}\colon \mathbf{b} \in B\}$. Then*

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

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

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

□

[1] Alternatively, one could compute the intersection in (A.1) over all σ-finite measures. These two approaches are equivalent because for every σ-finite measure λ and compact set K, the restriction $\lambda \llcorner K$ is a finite measure with $\mathcal{L}_{\lambda \llcorner K}(X) \supset \mathcal{L}_\lambda(X)$.

Therefore, to show that $S_B(g)$ is measurable for measurable g, it suffices to show that the direct sum of a measurable set and the compact set B is measurable. Thus, to prove Theorem 103, it suffices to demonstrate the following result:

Theorem 105. *Let $A \in \mathscr{U}(\mathbb{R}^d)$ and let B be a compact set. Then $A \oplus B \in \mathscr{U}(\mathbb{R}^d)$.*

The proof of Theorem 105 follows from fundamental concepts of measure theory. A classical measure theory result states that if $f : X \to Y$ is a continuous function, f^{-1} maps Borel sets in Y to Borel sets in X. Consider now the function $w\colon B \times \mathbb{R}^d \to B \times \mathbb{R}^d$ given by $w(\mathbf{h}, \mathbf{x}) = (\mathbf{h}, \mathbf{x}-\mathbf{h})$. Then w is invertible and the inverse of w is $w^{-1}(\mathbf{h}, \mathbf{x}+\mathbf{h})$. Furthermore, w^{-1} maps the set $B \times A$ to $B \times A \oplus B$. Therefore, if $A \in \mathcal{B}(\mathbb{R}^d)$, then $B \times A \oplus B$ is Borel in $\mathcal{B}(B \times \mathbb{R}^d)$. However, from this statement, *one cannot conclude that $A \oplus B$ is Borel in $\mathbb{R}^d$!* On the otherhand, one can use regularity of measures to conclude that $A \oplus B$ is in $\mathscr{U}(\mathbb{R}^d)$. Therefore, to prove Theorem 105, we prove the following two results:

Lemma 106. *Let $B \subset \mathbb{R}^d$ be a compact set. Then $B \times A \in \mathscr{U}(B \times \mathbb{R}^d)$ iff $A \in \mathscr{U}(\mathbb{R}^d)$.*

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

Lemma 107. *Let $f : X \to Y$ be a Borel measurable function. Then f is universally measurable as well.*

This result is stated on page 171 of [10], but we include a proof below for completeness.

Lemma 107 applied to w implies that the set $B \times A \oplus B$ is universally measurable while Lemma 106 implies that $A \oplus B$ is universally measurable.

A.1.3 Proof of Theorem 105

We begin by proving Lemma 107.

Proof of Lemma 107. Let A be a Borel set in Y. We will show that for any finite measure ν on X, $f^{-1}(A) \in \mathcal{L}_\nu(X)$. As ν is arbitrary, this statement will impy that $f^{-1}(A) \in \mathscr{U}(X)$.

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

On the other hand, the proof of Lemma 106 relies on the definition of a regular space X:

Definition 108. *A measure ν is* inner regular *if for every Borel set A,*

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

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

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

Theorem 109. *A σ-compact locally compact Hausdorff space is regular.*

This theorem is is a consequence of Theorem 7.8 of [22].

The notion of regularity extends to complete measures.

Lemma 110. *Let $\overline{\nu}$ be the completion of a measure ν on a regular space X. Then for any $A \in \mathcal{L}_\nu(X)$,*

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

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

Now using the concept of regularity, we prove Lemma 106.

Proof of Lemma 106. We first prove the forward direction. Consider the projection function $\Pi_2 \colon B \times \mathbb{R}^d \to \mathbb{R}^d$ given by $\Pi_2(\mathbf{x}, \mathbf{y}) = \mathbf{y}$. Then Π_2 is a continuous function and $\Pi_2^{-1}(A) =$

$B \times A$. Therefore Lemma 107 implies that if A is universally measurable in $\mathbb{R}^d$, then $B \times A$ is universally measurable in $B \times \mathbb{R}^d$.

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

Theorem 109 implies that $B \times \mathbb{R}^d$ is a regular space. Fix a Borel probability measure λ on B. Then $\lambda \times \nu$ is a finite Borel measure on $B \times \mathbb{R}^d$, so it is inner regular. Let $\overline{\lambda \times \nu}$ be the completion of $\lambda \times \nu$. Then by Lemma 110,

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

We will now argue that

$$\sup_{\substack{K \text{ compact} \\ K \subset B \times A}} \lambda \times \nu(K) = \sup_{\substack{K \text{ compact} \\ K \subset A}} \nu(K) \tag{A.2}$$

Let $K \subset B \times A$ and let Π_2 be projection onto the second coordinate. Because the continuous image of a compact set is compact, K'=$\Pi_2(K)$ is compact and contained in A. Thus $B \times A \supset B \times K' \supset K$, which implies (A.2). Now (A.2) applied to A^C implies that

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

Thus

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

Let K_n be a sequence of compact sets contained in A for which $\lim_{n\to\infty} \nu(K_n) = m$ and U_n a sequence of sets containing A for which U_n^C is compact and $\lim_{n\to\infty} \nu(U_n) = m$. Because a finite union of compact sets is compact, one can choose such sequences that satisfy $K_{n+1} \supset$

K_n and $U_{n+1} \subset U_n$. Then $B_1 = \bigcup K_n$, $B_2 = \bigcap U_n$ are Borel sets that satisfy $B_1 \subset A \subset B_2$ and $\nu(B_1) = \nu(B_2)$ so $A \in \mathcal{L}_\nu(\mathbb{R}^d)$.

□

Lastly, we formally prove Theorem 105.

Proof of Theorem 105. Consider the function $w\colon B \times \mathbb{R}^d \to B \times \mathbb{R}^d$ given by $w(\mathbf{h}, \mathbf{x}) = (\mathbf{h}, \mathbf{x} - \mathbf{h})$. Then w is continuous, invertible, and $w^{-1}(\mathbf{h}, \mathbf{x}) = (\mathbf{x}, \mathbf{x} + \mathbf{h})$.

Now let $A \in \mathscr{U}(\mathbb{R}^d)$. Then Lemma 106 implies that $B \times \mathbb{R}^d$ is universally measurable in $B \times A$. Lemma 107 then implies that $w^{-1}(B \times A) = B \times A \oplus B$ is universally measurable as well. Finally, Lemma 106 implies that $A \oplus B \in \mathscr{U}(\mathbb{R}^d)$ as well. □

A.2 Alternative Characterizations of the W_∞ Metric

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

Theorem 111. *Let X, Y be Borel sets and assume that $D \subset X \times Y$ is also Borel. Let D_x denote*

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

and

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

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

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

Assume that $f^*(x) > -\infty$ for all x. Then for any $\delta > 0$, there is a universally measurable

$\varphi\colon \operatorname{Proj}_X(D) \to Y$ *for which*

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

This statement is a consequence of Proposition 7.50 from [10].

We use the following results about universally measurable functions, see Lemma 7.27 of [10].

Lemma 112. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a universally measurable function and let $\mathbb{Q}$ be a Borel measure. Then there is a Borel measurable function φ for which $\varphi = g$ $\mathbb{Q}$-a.e.*

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

Lemma 113. *Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a universally measurable function and let $\mathbb{Q}$ be a Borel measure. Then there is a Borel measurable function φ for which $\varphi = g$ $\mathbb{Q}$-a.e.*

Proof. Let $\mathbf{e}_i$ denote the ith basis vector. Then $g_i := \mathbf{e}_i \cdot g$ is a universally measurable function from $\mathbb{R}^d$ to $\mathbb{R}$, so by Lemma 112, there is a Borel function φ_i for which $\varphi_i = g_i$ $\mathbb{Q}$-a.e. Then if we define $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_d)$, this function is equal to g $\mathbb{Q}$-a.e. □

Finally, we prove Lemma 3. Due to Lemmas 112 and 113, this lemma heavily relies on the fact that the domain of our functions is $\mathbb{R}^d$ rather than an arbitrary metric space.

Lemma 114. *Let $\mathbb{Q}$ be a finite positive Borel measure and let $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a Borel measurable function. Then*

$$\int S_\epsilon(f) d\mathbb{Q} = \sup_{\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})} \int f d\mathbb{Q}' \tag{2.13}$$

Recall that this paper defines the left-left hand side of (2.13) as the integral of $S_\epsilon(f)$ with respect to the completion of $\mathbb{Q}$.

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

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

Therefore, we can conclude that

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

We will show the opposite inequality by applying the measurable selection theorem. Theorem 111 implies for each $\delta > 0$, one can find a universally measurable function $\varphi \colon \mathbb{R}^d \to \overline{B_\epsilon(\mathbf{x})}$ for which $f(\varphi(\mathbf{x})) + \delta \geq S_\epsilon(f)(\mathbf{x})$. By Lemma 113, one can find a Borel measurable function T for which $T = \varphi$ $\mathbb{Q}$-a.e.

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

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

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

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

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

$\text{supp}(\gamma) \subseteq \Delta_\epsilon$. $\square$

Next we prove Lemma 4. We begin by presenting Strassen's theorem, see Corollary 1.28 of [65] for more details

Theorem 115 (Strassen's Theorem)**.** *Let $\mathbb{P}, \mathbb{Q}$ be positive finite measures with the same mass and let $\epsilon \geq 0$. Let $\Pi(\mathbb{P}, \mathbb{Q})$ denote the set couplings of $\mathbb{P}$ and $\mathbb{Q}$. Then*

$$\inf_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \pi(\{\|\mathbf{x} - \mathbf{y}\| > \epsilon) = \sup_{A \text{ closed}} \mathbb{Q}(A) - \mathbb{P}(A^\epsilon) \tag{A.3}$$

Strassen's theorem is usually written with A^ϵ in (A.3) replaced by $A^{\epsilon]} = \{\mathbf{x} \colon \text{dist}(\mathbf{x}, A) \leq \epsilon\}$—however, for closed sets $A^{\epsilon]} = A^\epsilon$. Strassen's theorem together with Urysohn's lemma then immediately proves Lemma 4.

Lemma 116 (Urysohn's Lemma)**.** *Let A and B be two closed and disjoint subsets of $\mathbb{R}^d$. Then there exists a function $f \colon \mathbb{R}^d \to [0, 1]$ for which $f = 0$ on A and $f = 1$ on B.*

See for instance result 4.15 of [22].

Lemma 117. *Let $\mathbb{P}, \mathbb{Q}$ be two finite positive Borel measures with $\mathbb{P}(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$. Then*

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

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

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

$$\sup_{A \text{ closed}} \mathbb{Q}(A) - \mathbb{P}(A^\epsilon) \leq \sup_{h \in C_b(\mathbb{R}^d)} \int h d\mathbb{Q} - \int S_\epsilon(h) d\mathbb{P} \tag{A.4}$$

Strassen's theorem will then imply that $W_\infty(\mathbb{P}, \mathbb{Q}) \leq \epsilon$. Let δ be arbitrary and let A be a closed set that satisfies $\sup_{A \text{ closed}} \mathbb{Q}(A) - \mathbb{P}(A^\epsilon) \leq \mathbb{Q}(A) - \mathbb{P}(A^\epsilon) + \delta$. Now because A is closed,

$A_n = A \oplus B_{1/n}(\mathbf{0})$ is a series of open sets decreasing to A and $A_n^\epsilon = A^\epsilon \oplus B_{1/n}(\mathbf{0})$ is a sequence of open sets decreasing to A^ϵ. Thus pick n sufficiently large so that $\mathbb{P}(A_n^\epsilon - \mathbb{P}(A^\epsilon) \leq \delta$. By Urysohn's lemma, one can choose a function h which is 1 on A, 0 on A_n^C, and between 0 and 1 on $A_n - A^C$. Then $S_\epsilon(h)$ is 1 on A^ϵ, 0 on $(A_n^\epsilon)^C$ and between 0 and 1 on $A_n^\epsilon - A^\epsilon$. Then $\int h d\mathbb{Q} - \mathbb{Q}(A) \geq 0$ and thus

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

Because δ was arbitrary, (A.4) follows. □

A.3 Minimizers of $C_\phi(\eta, \cdot)$: Proof of Lemma 25

Lemma 118. *Fix a loss function ϕ and let $\alpha_\phi(\eta)$ be as in (2.8). Then α_ϕ maps η to the smallest minimizer of $C_\phi(\eta, \cdot)$. Furthermore, the function $\alpha_\phi(\eta)$ non-decreasing in η.*

Proof. To start, we will show that $\alpha_\phi(\eta)$ as defined in (2.8) is a minimizer of $C_\phi(\eta, \cdot)$. Let S be the set of minimizers of $C_\phi^*(\eta, \cdot)$, which is non-empty due to the lower semi-continuity of ϕ. Let $a = \inf S = \alpha_\phi(\eta)$ and let $s_i \in S$ be a sequence converging to a. Then because ϕ is lower semi-continuous,

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

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

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

One can write

$$
\begin{aligned}
C_\phi(\eta_2, \alpha) &= \eta_2 \phi(\alpha) + (1 - \eta_2)\phi(-\alpha) \\
&= \eta_1 \phi(\alpha) + (1 - \eta_1)\phi(-\alpha) + (\eta_2 - \eta_1)(\phi(\alpha) - \phi(-\alpha)) \\
&= C_\phi(\eta_1, \alpha) + (\eta_2 - \eta_1)(\phi(\alpha) - \phi(-\alpha)) \qquad \text{(A.5)}
\end{aligned}
$$

Notice that the function $\alpha \mapsto \phi(\alpha) - \phi(-\alpha)$ is non-increasing. Then because $\alpha_\phi(\eta_1)$ is the smallest minimizer of $C_\phi(\eta_1, \alpha)$, if $\alpha < \alpha_\phi(\eta_1)$, then $C_\phi(\eta_1, \alpha) > C_\phi(\eta_1, \alpha_\phi(\eta_1))$. Furthermore, $\phi(\alpha) - \phi(-\alpha) \geq \phi(\alpha_\phi(\eta_1)) - \phi(-\alpha_\phi(\eta_1))$. Therefore, (A.5) implies that $C_\phi(\eta_2, \alpha) > C_\phi(\eta_2, \alpha_\phi(\eta_1))$, and thus α cannot be a minimizer of $C_\phi(\eta_2, \cdot)$. Therefore, $\alpha_\phi(\eta_2) \geq \alpha_\phi(\eta_1)$.

□

A.4 Continuity Properties of $\bar{R}_\phi$—Proof of Lemma 12

Recall the function $G(\eta, \alpha)$ defined by (2.47). With this notation, one can write the C_ϕ^* transform as $h_1^{C_\phi^*} = \sup_{\eta \in [0,1]} G(\eta, h_1)$.

Lemma 119. *Let $c > 0$ and consider $\alpha \geq c$. Let $a(\alpha) = \alpha^{C_\phi^*}$, where the C_ϕ^* transform is as in Lemma 22. Then there is a constant $k < 1$ for which*

$$
a(\alpha) = \sup_{\eta \in [0,k]} \frac{C_\phi^*(\eta) - \eta\alpha}{1 - \eta} \qquad \text{(A.6)}
$$

The constants k depends only on c.

Proof. Recall that the function $G(\eta, \alpha)$ is decreasing in α for fixed η and continuous on $[1, 0)$. Let $k = \sup\{\eta \colon G(\eta, c) > 0\}$. As c is strictly positive, one can conclude that $\lim_{\eta \to 1} G(\eta, c) =$

$-\infty$ and as a result $k < 1$. Because G is decreasing in α, one can conclude that $G(\eta, \alpha) \leq 0$ for all $\eta > k$ and $\alpha \geq c$. However, $\sup_{\eta \in [0,1]} G(\eta, \alpha) \geq 0$ because $G(0, \alpha) = 0$ for all α. Thus (A.6) holds.

□

Lemma 120. *Let $\{f_\alpha\}$ be a set of L-Lipschitz functions. Then $\sup_\alpha f_\alpha$ is also L-Lipschitz.*

This statement is proved in Box 1.8 of [55].

Lemma 121. *Let $\mathbb{Q}$ be any finite measure and assume that g is a non-negative function in $L^1(\mathbb{Q})$. Let $\delta > 0$. Then there is a lower semi-continuous function $\tilde{g}$ for which $\int |g - \tilde{g}| < \delta$ and $g \geq 0$.*

See Proposition 7.14 of Folland.

Lemma 122. *Let g be a lower semi-continuous function bounded from below. Then there is a sequence of Lipschitz functions that approaches g from below.*

This statement appears in Box 1.5 of [55].

Corollary 123. *Let h be an $L^1(\mathbb{Q})$ function with $h \geq 0$. Then for any δ, there exists a Lipschitz $\tilde{h}$ for which $\int |h - \tilde{h}| d\mathbb{Q} < \delta$.*

Proof. By Lemma 121, one can pick a lower semi-continuous $\tilde{g}$ for which $\tilde{g} \geq 0$ and $\int |h - \tilde{g}| d\mathbb{Q} < \delta/2$. Next, by Lemma 122, one can pick a Lipschitz $\tilde{h}$ for which $\int |\tilde{g} - \tilde{h}| d\mathbb{Q} \leq \delta/2$. Thus $\int |h - \tilde{h}| d\mathbb{Q} < \delta$. □

Lemma 124. *Let $K \subset \mathbb{R}^d$ be compact, $E = C_b(K^\epsilon) \times C_b(K^\epsilon)$, and $\mathbb{P}'_0, \mathbb{P}'_1 \in \mathcal{M}_+(K^\epsilon)$. Then*

$$\inf_{(h_0,h_1) \in S_\phi \cap E} \int h_1 d\mathbb{P}'_1 + \int h_0 d\mathbb{P}'_0 = \bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) \tag{2.27}$$

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

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

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

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

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

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

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

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

$$C^*_\phi(\eta) - \eta \tilde{h}_1 - (1-\eta)h_0 \leq 0 \tag{A.7}$$

Furthermore,

$$\int \tilde{h}_1 d\mathbb{P}'_1 + \int h_0 d\mathbb{P}'_0 < \bar{R}_\phi(\mathbb{P}'_0, \mathbb{P}'_1) + \delta. \tag{A.8}$$

Let k be the constant described by Lemma 119 corresponding to c. Now by Corollary 123, there is a Lipschitz function g for which $\int |h_1 - g| d\mathbb{P}' < \min((1-k)/k, 1)\delta$. Let $\hat{h}_1 = \max(g, c)$. Then Lemma 120 implies that $\hat{h}_1$ has the same Lipschitz constant as g, and the fact that

$\tilde{h}_1 \geq c$ implies that

$$\int |\tilde{h}_1 - \hat{h}_1| d\mathbb{P}' \leq \int |\tilde{h}_1 - g| d\mathbb{P}' < \min\left(\frac{1-k}{k}, 1\right)\delta \tag{A.9}$$

Now let $\hat{h}_0 = \hat{h}_1^{C_\phi^*}$. By Lemma 119, the supremum in the $^{C_\phi^*}$ transform for computing $\hat{h}_0$ can be taken over $[0, k]$. Therefore, if L is the Lipschitz constant of $\hat{h}_1$, Lemma 120 implies that the Lipschitz constant of $\hat{h}_0$ is at most $kL/(1-k)$. Furthermore, $\hat{h}_0, \hat{h}_1$ are bounded on K^ϵ because Lipschitz functions are bounded over compact sets. Thus $(\hat{h}_0, \hat{h}_1)$ is in $S_\phi \cap E$. Next, we will show that $\int \hat{h}_0$ is close to $\int h_0$.

$$\begin{aligned}
&\int \hat{h}_0 - h_0 d\mathbb{P}_0' = \int \sup_{[0,k]} \frac{C_\phi^*(\eta) - \eta\hat{h}_1}{1-\eta} - h_0 d\mathbb{P}_0' = \\
&\int \sup_{[0,k]} \frac{C_\phi^*(\eta) - \eta\hat{h}_1 - (1-\eta)h_0}{1-\eta} d\mathbb{P}_0' = \\
&\int \sup_{[0,k]} \left(\frac{C_\phi^*(\eta) - \eta\tilde{h}_1 - (1-\eta)h_0}{1-\eta} + \frac{\eta}{1-\eta}(\tilde{h}_1 - \hat{h}_1) \right) d\mathbb{P}_0' \leq \\
&\int \sup_{[0,k]} \frac{C_\phi^*(\eta) - \eta\tilde{h}_1 - (1-\eta)h_0}{1-\eta} + \sup_{[0,k]} \frac{\eta}{1-\eta}(\tilde{h}_1 - \hat{h}_1) d\mathbb{P}_0' \leq \\
&\int \sup_{[0,k]} \frac{\eta}{1-\eta}(\tilde{h}_1 - \hat{h}_1) d\mathbb{P}_0' = && \text{(Equation A.7)} \\
&\frac{k}{1-k} \int \tilde{h}_1 - \hat{h}_1 d\mathbb{P}_0' \leq \delta && \text{(Equation A.9)}
\end{aligned}$$

Therefore, by (A.8), (A.9), and the computation above,

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

As $\delta > 0$ is arbitrary, this inequality implies (2.27). Because K^ϵ is compact, the upper semi-continuity and concavity of $\bar{R}_\phi$ then follows from (2.27) together with the Reisz representation theorem. □

A.5 Duality for Distributions with Arbitrary Support—Proof of Lemma 14

We begin with the simple observation that weak duality holds for measures supported on $\mathbb{R}^d$. This argument is essentially swapping the order of an infimum and a supremum as presented in Section 2.4.1.

Lemma 125 (Weak Duality)**.** *Let ϕ be a non-increasing and lower semi-continuous loss function. Let S_ϕ be the set of pairs of functions defined in (2.25) for $K = \mathbb{R}^d$.*

Then

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

Proof. By Lemma 3,

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

Thus by swapping the inf and the sup,

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

□

The main strategy in this section is approximating measures with unbounded support by

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

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

Theorem 126 (Portmanteau Theorem)**.** *The following are equivalent:*

1) The sequence $\mathbb{Q}^n \in \mathcal{M}_+(\mathbb{R}^d)$ converges weakly to $\mathbb{Q}$

2) For all closed sets C, $\limsup_{n\to\infty} \mathbb{Q}^n(C) \leq \mathbb{Q}(C)$ and $\lim_{n\to\infty} \mathbb{Q}^n(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$

3) For all open sets U, $\liminf_{n\to\infty} \mathbb{Q}^n(U) \geq \mathbb{Q}(U)$ and $\lim_{n\to\infty} \mathbb{Q}^n(\mathbb{R}^d) = \mathbb{Q}(\mathbb{R}^d)$

See Theorem 8.2.3 of [15]. This result allows us to draw conclusions about restrictions of weakly convergent sequences.

Lemma 127. *Let $\mathbb{Q}^n, \mathbb{Q} \in \mathcal{M}_+(\mathbb{R}^d)$ and assume that $\mathbb{Q}^n$ converges weakly to $\mathbb{Q}$. Let K be a compact set with $\mathbb{Q}(\partial K) = 0$. Then $\mathbb{Q}^n|_K$ converges weakly to $\mathbb{Q}|_K$.*

Proof. We will verify 2) of Theorem 126 for the measures $\mathbb{Q}^n|_K$, $\mathbb{Q}$.

First, because $\mathbb{Q}(K) = \mathbb{Q}(\operatorname{int} K)$, Theorem 126 implies that

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

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

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

□

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

Theorem 128. *Let $\mathbb{Q}^n$ be a sequence of measures for which $\sup_n \mathbb{Q}^n(\mathbb{R}^d) < \infty$ and for all δ, there exists a compact K for which $\mathbb{Q}^n(K^C) < \delta$ for all n. Then $\mathbb{Q}^n$ has a weakly convergent subsequence.*

See Theorem 8.6.2 of [15]. These results imply that $\bar{R}_\phi$ is upper semi-continuous on $\mathcal{M}_+(\mathbb{R}^d) \times \mathcal{M}_+(\mathbb{R}^d)$.

Lemma 129. *The functional $\bar{R}_\phi$ is upper semi-continuous with respect to the weak topology on probability measures (in duality with $C_0(\mathbb{R}^d)$).*

Notice that Lemma 12 implies that $\bar{R}_\phi$ is upper semi-continuous on the space $\mathcal{M}_+(K^\epsilon) \times \mathcal{M}_+(K^\epsilon)$ for a *compact* set K. However, on $\mathbb{R}^d$, weak convergence of measures is defined with respect to the dual of $C_0(\mathbb{R}^d)$, the set of continuous functions vanishing at ∞. This set is strictly smaller than $C_b(\mathbb{R}^d)$, and thus the relation (2.27) would not immediately imply the the upper semi-continuity of R^ϵ_ϕ.

Proof. Let $\mathbb{Q}^n_0, \mathbb{Q}^n_1$ be sequences of measures converging to $\mathbb{Q}_0, \mathbb{Q}_1$ respectively. Set $\mathbb{Q} = \mathbb{Q}_0 + \mathbb{Q}_1$.

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

Let $\delta > 0$ be arbitrary and choose R large enough so that $F(R) < \delta / C^*_\phi(1/2)$ and F is continuous at R. Then notice that $\mathbb{P}(\partial B_R(\mathbf{0})) = 0$ and thus one can apply Lemma 127 with the set $\overline{B_R(\mathbf{0})}$.

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

$$|\bar{R}_\phi(\nu_0^R, \nu_1^R) - \bar{R}_\phi(\nu_0, \nu_1)| = \left| \int C^*_\phi(\eta) \mathbf{1}_{\overline{B_R(\mathbf{0})}} d\nu - \int C^*_\phi(\eta) d\nu \right| \leq C^*_\phi\left(\frac{1}{2}\right) \nu(\overline{B_R(\mathbf{0})^C}) \quad \text{(A.10)}$$

If we define $\mathbb{Q}_{i,R}, \mathbb{Q}^n_{i,R}$ via $\mathbb{Q}_{i,R} = \mathbb{Q}_i|_{\overline{B_R(\mathbf{0})}}$, $\mathbb{Q}^n_{i,R} = \mathbb{Q}^n_i = \mathbb{Q}^n_i|_{\overline{B_R(\mathbf{0})}}$, Lemma 127 implies that $\mathbb{Q}^n_{i,R}$ converges weakly to $\mathbb{Q}_{i,R}$ and $\lim_{n\to\infty} \mathbb{Q}^n(\overline{B_R(\mathbf{0})}^C) = \mathbb{Q}(B_R(\mathbf{0})^C) < \delta$. Therefore, for

sufficiently large n, $\mathbb{Q}^n(\overline{B_R(\mathbf{0})}^C) < 2\delta/C_\phi^*(1/2)$. By Lemma 12 and (A.10),

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

Because δ was arbitrary, the result follows.

□

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

Lemma 130. *Let $\mathbb{P}_0, \mathbb{P}_1$ be finite measures. Define $\mathbb{P}_i^n = \mathbb{P}_i|_{\overline{B_n(\mathbf{0})}}$ for $n \in \mathbb{N}$. Then $\mathbb{P}_0^n, \mathbb{P}_1^n$ converge weakly to $\mathbb{P}_0$, $\mathbb{P}_1$ respectively. Furthermore, there are measures $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0), \mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ for which*

$$\limsup_{n\to\infty} \sup_{\substack{\mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1^n) \\ \mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)^n}} \bar{R}_\phi(\mathbb{P}_0', \mathbb{P}_1') \le \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*) \tag{A.11}$$

Proof. Set $\mathbb{P} = \mathbb{P}_0 + \mathbb{P}_1$, $\mathbb{P}^n = \mathbb{P}_0^n + \mathbb{P}_1^n$. Notice that 2) of Theorem 126 implies that $\mathbb{P}_i^n$ converges weakly to $\mathbb{P}_i$. Let $\mathbb{P}_0^{*,n}, \mathbb{P}_1^{*,n}$ be maximizers of $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0^n) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1^n)$. Next, by Strassen's theorem (Theorem 115), $\mathbb{P}_i^n(\overline{B_r(\mathbf{0})}) \le \mathbb{P}_i^{n,*}(\overline{B_{r+\epsilon}(\mathbf{0})})$ and thus $\mathbb{P}_i(\overline{B_r(\mathbf{0})}^C) \ge \mathbb{P}_i^n(\overline{B_r(\mathbf{0})}^C) \ge \mathbb{P}_i^{n,*}(\overline{B_{r+\epsilon}(\mathbf{0})})$. Therefore, one can apply Prokhorov's theorem (Thereom 128) to conclude that $\mathbb{P}_0^{n,*}$, $\mathbb{P}_1^{n,*}$ have subsequences $\mathbb{P}_0^{n_k,*}$, $\mathbb{P}_1^{n_k,*}$ that converge to measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ respectively. The upper semi-continuity of R_ϕ (Lemma 129) then implies that $\mathbb{P}_0^*, \mathbb{P}_1^*$ satisfy (A.11).

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

□

Lemma 131. *Let ϕ be a non-increasing, lower semi-continuous loss function and let $\mathbb{P}_0, \mathbb{P}_1$*

be finite Borel measures supported on $\mathbb{R}^d$. *Let* S_ϕ *be as in* (2.25). *Then*

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

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

Proof. Let $\mathbb{P}_0^n$, $\mathbb{P}_1^n$, $\mathbb{P}_0^*, \mathbb{P}_1^*$ be the the measures described in Lemma 130. Notice that because $\mathbb{P}_0^n$, $\mathbb{P}_1^n$ are compactly supported, Lemma 13 applies. Define

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

Thus Lemmas 13 and Lemma 130 imply that

$$\limsup_{n\to\infty} \inf_{(h_0,h_1)\in S_\phi} \Theta^n(h_0,h_1) = \limsup_{n\to\infty} \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0^n)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1^n)}} \bar{R}_\phi(\mathbb{P}_0',\mathbb{P}_1') \le \bar{R}_\phi(\mathbb{P}_0^*,\mathbb{P}_1^*) \le \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0',\mathbb{P}_1'). \tag{A.12}$$

We will show

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0,h_1) \le \limsup_{n\to\infty} \inf_{(h_0,h_1)\in S_\phi} \Theta_n(h_0,h_1). \tag{A.13}$$

Equations A.12 and A.13 imply that

$$\inf_{(h_0,h_1)\in S_\phi} \Theta(h_0,h_1) \le \bar{R}_\phi(\mathbb{P}_0^*,\mathbb{P}_1^*) \le \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_\phi(\mathbb{P}_0',\mathbb{P}_1'). \tag{A.14}$$

This relation together with weak duality (Lemma 125) imply that the inequalities in (A.14) are actually equalities. Therefore strong duality holds and $\mathbb{P}_0^*, \mathbb{P}_1^*$ maximizes the dual.

Next, we prove the inequality in (A.13). Let $\delta > 0$ be arbitrary and choose an $n \in \mathbb{N}$ for which $n > 2\epsilon$ and

$$\mathbb{P}_1(\overline{B_{n-2\epsilon}(\mathbf{0})}^C) + \mathbb{P}_0(\overline{B_{n-2\epsilon}(\mathbf{0})}^C) \le \delta \tag{A.15}$$

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

$$\Theta^n(h_0^n, h_1^n) \leq \inf_{(h_0,h_1)\in S_\phi} \Theta^n(h_0, h_1) + \delta \tag{A.16}$$

Define

$$\tilde{h}_0^n = \begin{cases} h_0^n & \mathbf{x} \in \overline{B_{n-\epsilon}(\mathbf{0})} \\ C_\phi^*\left(\frac{1}{2}\right) & \mathbf{x} \notin \overline{B_{n-\epsilon}(\mathbf{0})} \end{cases} \quad \tilde{h}_1^n = \begin{cases} h_1^n & \mathbf{x} \in \overline{B_{n-\epsilon}(\mathbf{0})} \\ C_\phi^*\left(\frac{1}{2}\right) & \mathbf{x} \notin \overline{B_{n-\epsilon}(\mathbf{0})} \end{cases}$$

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

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

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

Now for each i, these bounds imply that

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

Then, applying this bound for each i,

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

The last inequality follows from Equations A.15 and A.16. Because δ arbitrary, (A.13) holds.

□

A.6 Complementary Slackness

Lemma 132. *Assume that $\mathbb{P}_0, \mathbb{P}_1$ are compactly supported. The functions h_0^*, h_1^* minimize Θ over S_ϕ and $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ maximize $\bar{R}_\phi$ over $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ iff the following hold:*

1)

$$\int h_1^* d\mathbb{P}_1^* = \int S_\epsilon(h_1^*) d\mathbb{P}_1 \quad \textit{and} \quad \int h_0^* d\mathbb{P}_0^* = \int S_\epsilon(h_0^*) d\mathbb{P}_0 \tag{2.30}$$

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

$$\eta^*(\mathbf{x}) h_1^*(\mathbf{x}) + (1 - \eta^*(\mathbf{x})) h_0^*(\mathbf{x}) = C_\phi^*(\eta^*(\mathbf{x})) \quad \mathbb{P}^*\textit{-a.e.} \tag{2.31}$$

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

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

$$\Theta(h_0^*, h_1^*) = \int S_\epsilon(h_1^*) d\mathbb{P}_1 + \int S_\epsilon(h_0^*) d\mathbb{P}_0 \geq \int h_1^* d\mathbb{P}_1^* + \int h_0^* d\mathbb{P}_0^* \tag{A.17}$$

$$= \int \eta^* h_1^* + (1 - \eta^*) h_0^* d\mathbb{P}^* \geq \int C_\phi^*(\eta^*) d\mathbb{P}^* = \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*) \tag{A.18}$$

By Lemma 14, both the first expression of (A.17) and the last expression of (A.18) are equal. Thus all the inequalities above must be equalities which implies (2.31). Next, because (A.18) implies that

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

and Lemma 3 implies that $\int S_\epsilon(h_0^*) d\mathbb{P}_0 \geq \int h_0^* d\mathbb{P}_0^*$ and $\int S_\epsilon(h_1^*) d\mathbb{P}_1 \geq \int h_1^* d\mathbb{P}_1^*$ we can conclude (2.30).

We will now show the opposite implication. Assume that $h_0^*, h_1^*, \mathbb{P}_0^*, \mathbb{P}_1^*$ satisfy (2.30) and (2.31). Then

$$\begin{aligned}
\Theta(h_0^*, h_1^*) &= \int S_\epsilon(h_1^*) d\mathbb{P}_1 + \int S_\epsilon(h_0^*) d\mathbb{P}_0 \\
&= \int h_1^* d\mathbb{P}_1^* + \int h_0^* d\mathbb{P}_0^* && \text{(Equation 2.30)} \\
&= \int \eta^* h_1^* + (1 - \eta^*) h_0^* d\mathbb{P}^* = \int C_\phi^*(\eta^*) d\mathbb{P}^* && \text{(Equation 2.31)} \\
&= \bar{R}_\phi(\mathbb{P}_0^*, \mathbb{P}_1^*)
\end{aligned}$$

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

□

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

A.7 Technical Lemmas from Section 2.6

A.7.1 Proof of Lemma 17

Lemma 133. *Let $\psi(\alpha) = e^{-\alpha}$. Then $C_\psi^*(\eta) = 2\sqrt{\eta(1-\eta)}$ and $\alpha_\psi(\eta) = 1/2\log(\eta/1-\eta)$ is the unique minimizer of $C_\psi(\eta,\cdot)$, with $\alpha_\psi(0), \alpha_\psi(1)$ interpreted as $-\infty$, $+\infty$ respectively. Furthermore, $\partial C_\psi^*(\eta)$ is the singleton $\partial C_\psi^*(\eta) = \{\psi(\alpha_\psi(\eta)) - \psi(-\alpha_\psi(\eta))\}$.*

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

Next, by definition

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

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

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

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

□

A.7.2 Proof of Lemma 18

Lemma 134. *Let (a_n, b_n) be a sequence for which $a_n, b_n \geq 0$ and*

$$\eta a_n + (1-\eta)b_n \geq C_\psi^*(\eta) \text{ for all } \eta \in [0,1] \tag{2.39}$$

and

$$\lim_{n\to\infty} \eta_0 a_n + (1-\eta_0) b_n = C_\psi^*(\eta_0) \tag{2.40}$$

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

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

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

Then (2.39) (2.40) imply that

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

$$\eta_0 a + (1-\eta_0) b = C_\psi^*(\eta_0) \tag{A.19}$$

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

$$a - b = \psi(\alpha_\psi(\eta_0)) - \psi(-\alpha_\psi(\eta_0)) \tag{A.20}$$

while (A.19) is equivalent to

$$\eta_0 a + (1-\eta_0) b = \eta_0 \psi(\alpha_\psi(\eta_0)) + (1-\eta_0)\psi(-\alpha_\psi(\eta_0)) \tag{A.21}$$

The equations (A.20) and (A.21) comprise a system of equations in two variables with a unique solution for a and b. □

A.7.3 Proof of Lemma 20

Lastly, we prove Lemma 20.

Lemma 135. *Let h_n be any sequence of functions. Then the sequence h_n satisfies*

$$\liminf_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\liminf_{n\to\infty} h_n) \tag{2.43}$$

and

$$\limsup_{n\to\infty} S_\epsilon(h_n) \geq S_\epsilon(\limsup_{n\to\infty} h_n) \tag{2.44}$$

Proof. We start by showing (2.43).

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

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

Equation 2.44 can then be proved by the same argument:

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

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

□

B — Deferred Proofs from Chapter 3

B.1 Proof of Lemma 27

First, the S_ϵ operation satisfies a subadditivity property:

Lemma 136. *Let S_1 and S_2 be two subsets of $\mathbb{R}^d$. Then*

$$S_\epsilon(\mathbf{1}_{S_1}) + S_\epsilon(\mathbf{1}_{S_2}) \geq S_\epsilon(\mathbf{1}_{S_1 \cap S_2}) + S_\epsilon(\mathbf{1}_{S_1 \cup S_2}) \tag{B.1}$$

Proof. First, notice that

$$\begin{aligned} S_\epsilon(\mathbf{1}_{S_1})(\mathbf{x}) + S_\epsilon(\mathbf{1}_{S_2})(\mathbf{x}) &= \begin{cases} 0 & \text{if } \mathbf{x} \notin S_1^\epsilon \text{ and } \mathbf{x} \notin S_2^\epsilon \\ 1 & \text{if } \mathbf{x} \in S_1^\epsilon \triangle S_2^\epsilon \\ 2 & \text{if } \mathbf{x} \in S_1^\epsilon \cap S_2^\epsilon \end{cases} \\ &= \mathbf{1}_{S_1^\epsilon \cap S_2^\epsilon}(\mathbf{x}) + \mathbf{1}_{S_1^\epsilon \cup S_2^\epsilon}(\mathbf{x}) \end{aligned} \tag{B.2}$$

Next, one can always swap the order of two maximums but a min-max is always larger

than a max-min. Therefore:

$$\begin{aligned} &S_\epsilon(\mathbf{1}_{S_1\cap S_2}) + S_\epsilon(\mathbf{1}_{S_1\cup S_2}) = S_\epsilon(\min(\mathbf{1}_{S_1}, \mathbf{1}_{S_2})) + S_\epsilon(\max(\mathbf{1}_{S_1}, \mathbf{1}_{S_2})) \\ &\leq \min(S_\epsilon(\mathbf{1}_{S_1}), S_\epsilon(\mathbf{1}_{S_2})) + \max(S_\epsilon(\mathbf{1}_{S_1}), S_\epsilon(\mathbf{1}_{S_2})) = \mathbf{1}_{S_1^\epsilon\cap S_2^\epsilon} + \mathbf{1}_{S_1^\epsilon\cup S_2^\epsilon} \end{aligned} \tag{B.3}$$

Comparing Equation (B.2) and Equation (B.3) results in Equation (B.1). □

Therefore, the adversarial classification risk is sub-additive.

Corollary 137. *Let S_1 and S_2 be any two sets. Then*

$$R^\epsilon(S_1 \cap S_2) + R^\epsilon(S_1 \cup S_2) \leq R^\epsilon(S_1) + R^\epsilon(S_2)$$

This result then directly implies Lemma 27:

Proof of Lemma 27. Let A_1 and A_2 be two adversarial Bayes classifiers, and let R_*^ϵ be the minimal adversarial Bayes risk. Then Corollary 137 implies that

$$2R_*^\epsilon \geq R^\epsilon(A_1 \cup A_2) + R^\epsilon(A_1 \cap A_2)$$

and hence $A_1 \cap A_2$ and $A_1 \cup A_2$ must be adversarial Bayes classifiers as well. □

B.2 Complementary Slackness– Proof of Theorem 30

The complementary slackness relations of Theorem 30 are a consequence of the minimax relation of Theorem 29 and properties of the W_∞ metric.

Integrating the maximum of an indicator function over an ϵ-ball is intimately linked to maximizing an integral over a W_∞ ball of measures:

Lemma 138. *Let $\mathbb{Q}$ be a positive measure. Then for any Borel set A*

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

Lemma 5.1 of [52] and Lemma 3 of [25] proved slightly different versions of this result, so we include a proof here for completeness.

Proof. Let $\mathbb{Q}'$ be any measure with $W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon$ and let γ be any coupling between for which

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

Such a coupling exists by Theorem 2.6 of [33]. Then $S_\epsilon(\mathbf{1}_A)(\mathbf{x}) \geq \mathbf{1}_A(\mathbf{y})$ γ-a.e. Thus

$$\int S_\epsilon(\mathbf{1}_A)(\mathbf{x})d\mathbb{Q}(\mathbf{x}) = \int S_\epsilon(\mathbf{1}_A)(\mathbf{x})d\gamma(\mathbf{x},\mathbf{y}) \geq \int \mathbf{1}_A d\gamma(\mathbf{x},\mathbf{y}) = \int \mathbf{1}_A(\mathbf{y})d\mathbb{Q}'(\mathbf{y})$$

Now taking a supremum over all $\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})$ concludes the proof. □

One can prove Theorem 30 with this result.

Proof of Theorem 30.

Forward Direction:

Let A be a minimizer of R^ϵ and assume that $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ maximize $\bar{R}$. Then:

$$R^\epsilon(A) = \int S_\epsilon(\mathbf{1}_{A^C})d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 \geq \int \mathbf{1}_{A^C} d\mathbb{P}_1^* + \int \mathbf{1}_A d\mathbb{P}_0^* \tag{B.4}$$

$$= \int \eta^* \mathbf{1}_{A^C} d\mathbb{P}_1 + \int (1-\eta^*)\mathbf{1}_A d\mathbb{P}_0^* \geq \int C^*(\eta^*)d\mathbb{P}^* = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*) \tag{B.5}$$

The first inequality follows from Lemma 138 while the second inequality follows from the definition of C^* in Equation (3.3). By Theorem 29, the first expression of Equation (B.4)

and the last expression of Equation (B.5) are equal. Thus all the inequalities above must in fact be equalities. Thus the fact that the inequality in Equation (B.5) is an equality implies Equation (3.12). Lemma 138 and the fact that the inequality in Equation (B.4) must be an equality implies Equation (3.11).

Backward Direction:

Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be measures satisfying $W_\infty(\mathbb{P}_0^*, \mathbb{P}_0) \leq \epsilon$, $W_\infty(\mathbb{P}_1^*, \mathbb{P}_1) \leq \epsilon$, and let A be a Borel set. Assume that A, $\mathbb{P}_0^*$, and $\mathbb{P}_1^*$ satisfy Equation (3.11) and Equation (3.12). We will argue that A is must be a minimizer of R^ϵ and $\mathbb{P}_0^*, \mathbb{P}_1^*$ must maximize $\bar{R}$.

First, notice that Theorem 29 implies that $R^\epsilon(A') \geq \bar{R}(\mathbb{P}_0', \mathbb{P}_1')$ for *any* Borel A' and $\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0), \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$. Thus if one can show

$$R^\epsilon(A) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*), \tag{B.6}$$

then A must minimize R^ϵ because for any other A',

$$R^\epsilon(A') \geq \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*) = R^\epsilon(A).$$

Similarly, one could conclude that $\mathbb{P}_0^*, \mathbb{P}_1^*$ maximize $\bar{R}$ because for any other $\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$ and $\mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$,

$$\bar{R}(\mathbb{P}_0', \mathbb{P}_1') \leq R^\epsilon(A) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*).$$

Hence it remains to show Equation (B.6). Applying Equation (3.11) followed by Equation (3.12), one can conclude that

$$\begin{aligned} R^\epsilon(A) &= \int S_\epsilon(\mathbf{1}_A^C) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0 = \int \mathbf{1}_{A^C} d\mathbb{P}_1^* + \int \mathbf{1}_A d\mathbb{P}_0^* && \text{Equation (3.11)} \\ &= \int \eta^* \mathbf{1}_{A^C} + (1-\eta^*)\mathbf{1}_A d\mathbb{P}^* = \int C^*(\eta^*) d\mathbb{P}^* = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*) && \text{Equation (3.12)} \end{aligned}$$

□

B.3 Proof of Proposition 51 and Lemma 52

The proof of Proposition 51 relies on Lemma 52.

B.3.1 Proof of Lemma 52

The $^\epsilon$ operation on sets interacts particularly nicely with Lebesgue measure.

Lemma 139. *For any set A and $\epsilon > 0$, ∂A^ϵ has Lebesgue measure zero.*

This result is standard in geometric measure theory, see for instance Lemma 4 in [2] for a proof. Next, the closure and $^\epsilon$ operations commute:

Lemma 140. *Let A be any set in $\mathbb{R}^d$. Then $\overline{A^\epsilon} = \overline{A}^\epsilon$.*

Proof. We show the two inclusions $\overline{A^\epsilon} \subset \overline{A}^\epsilon$ and $\overline{A^\epsilon} \supset \overline{A}^\epsilon$ separately.

Showing $\overline{A^\epsilon} \subset \overline{A}^\epsilon$: First, because the direct sum of a closed set and a compact set must be closed, $\overline{A}^\epsilon$ is a closed set that contains A^ϵ. Therefore, because $\overline{A^\epsilon}$ is the smallest closed set containing A^ϵ, the set $\overline{A^\epsilon}$ must be contained in $\overline{A}^\epsilon$.

Showing $\overline{A^\epsilon} \supset \overline{A}^\epsilon$: Let $\mathbf{x} \in \overline{A}^\epsilon$, we will show that $\mathbf{x} \in \overline{A^\epsilon}$. If $\mathbf{x} \in \overline{A^\epsilon}$, then $\mathbf{x} = \mathbf{a} + \mathbf{h}$ for some $\mathbf{a} \in \overline{A}$ and $\mathbf{h} \in \overline{B_\epsilon(\mathbf{0})}$. Let $\mathbf{a}_i$ be a sequence of points contained in A that converges to $\mathbf{a}$. Then $\mathbf{a}_i + \mathbf{h} \in A^\epsilon$, and $\mathbf{a}_i + \mathbf{h}$ converges to $\mathbf{a} + \mathbf{h}$. Therefore, $\mathbf{a} + \mathbf{h} \in \overline{A^\epsilon}$. □

Next, this result implies that the sets $(\operatorname{int} A)^\epsilon$, A^ϵ and $\overline{A}^\epsilon$ all have equal $\mathbb{P}_0$ measure while $((\operatorname{int} A)^C)^\epsilon$, $(A^C)^\epsilon$, and $(\overline{A}^C)^\epsilon$ have equal $\mathbb{P}_1$-measure.

Lemma 141. *If A is any adversarial Bayes classifier and $\epsilon > 0$, then $\mathbb{P}_0(A^\epsilon) = \mathbb{P}_0(\overline{A}^\epsilon) = \mathbb{P}_0((\operatorname{int} A)^\epsilon)$ and $\mathbb{P}_1((A^C)^\epsilon) = \mathbb{P}_1(((\operatorname{int} A)^C)^\epsilon) = \mathbb{P}_1((\overline{A}^C)^\epsilon)$.*

Proof. First, Lemmas 139 and 140 imply that

$$\mathbb{P}_0(A^\epsilon) = \mathbb{P}_0(\overline{A^\epsilon}) = \mathbb{P}_0(\overline{A}^\epsilon) \tag{B.7}$$

Furthermore, $\mathbb{P}_1((A^C)^\epsilon) \geq \mathbb{P}_1((\overline{A}^C)^\epsilon)$ and thus $R^\epsilon(A) \geq R^\epsilon(\overline{A})$. Consequently, $\overline{A}$ must be an adversarial Bayes classifier and

$$\mathbb{P}_1((A^C)^\epsilon) = \mathbb{P}_1((\overline{A}^C)^\epsilon) \tag{B.8}$$

A similar line of reasoning shows that

$$\mathbb{P}_1(\overline{(A^C)^\epsilon}) = \mathbb{P}_1(\overline{(A^C)}^\epsilon) = \mathbb{P}_1((\operatorname{int} A^C)^\epsilon) \tag{B.9}$$

and thus

$$\mathbb{P}_0(A^\epsilon) = \mathbb{P}_0((\operatorname{int} A)^\epsilon) \tag{B.10}$$

Equations (B.7) to (B.10) imply the desired result. □

Finally, Lemma 141 implies that $\operatorname{int} A$, A and $\overline{A}$ are all equivalent up to degeneracy.

Proof of Lemma 52. Lemma 141 implies that if E is any measurable set with $\operatorname{int} A \subset E \subset \overline{A}$, then $\mathbb{P}_0(E^\epsilon) = \mathbb{P}_0(A^\epsilon)$ and $\mathbb{P}_1((E^C)^\epsilon) = \mathbb{P}_1((A^C)^\epsilon)$. Therefore, E must be an adversarial Bayes classifier.

□

B.3.2 Proof of Proposition 51

The following lemma show that Item 2) implies Item 1).

Lemma 142. *Let A_1 and A_2 be adversarial Bayes classifiers for which 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. Then*

$$S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2}) = S_\epsilon(\mathbf{1}_{A_1 \cap A_2}) = S_\epsilon(\mathbf{1}_{A_1 \cup A_2}) \quad \mathbb{P}_0\text{-a.e.} \tag{B.11}$$

and

$$S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C}) = S_\epsilon(\mathbf{1}_{(A_1 \cap A_2)^C}) = S_\epsilon(\mathbf{1}_{(A_1 \cup A_2)^C}) \quad \mathbb{P}_1\text{-a.e.} \tag{B.12}$$

See Appendix B.3.2.1 for a proof. As a result:

Corollary 143. *Let A_1 and A_2 be two adversarial Bayes classifiers. Then $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$ $\mathbb{P}_0$-a.e. iff $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C})$ $\mathbb{P}_1$-a.e.*

Furthermore, the last equality in Equation (B.11) and Equation (B.12) implies that A_1 and A_2 are equivalent up to degeneracy.

This result suffices to prove the equivalence between Item 2) and Item 3), even when $\mathbb{P}$ is not absolutely continuous with respect to Lebesgue measure.

Lemma 144. *Let A_1 and A_2 be two adversarial Bayes classifiers for $\epsilon > 0$, and let $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ be a maximizer of $\bar{R}$. Define $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$.*

The following are equivalent:

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

3) $\mathbb{P}^*(A_2 \triangle A_1) = 0$

Proof. Assume that A_1 and A_2 are both adversarial Bayes classifiers. Lemma 27 then implies that $A_1 \cup A_2$, $A_1 \cap A_2$ are both adversarial Bayes classifiers. Equation (3.11) of Theorem 30 implies that

$$\begin{aligned}\int S_\epsilon(\mathbf{1}_{A_1 \cup A_2}) d\mathbb{P}_0 = \int \mathbf{1}_{A_1 \cup A_2} d\mathbb{P}_0^* &= \int \mathbf{1}_{A_1 \cap A_2} d\mathbb{P}_0^* + \mathbb{P}_0^*(A_1 \triangle A_2) \\ &= \int S_\epsilon(\mathbf{1}_{A_1 \cap A_2}) d\mathbb{P}_0 + \mathbb{P}_0^*(A_1 \triangle A_2)\end{aligned}$$

Because $S_\epsilon(\mathbf{1}_{A_1 \cap A_2}) \leq S_\epsilon(\mathbf{1}_{A_1 \cup A_2})$, $\mathbb{P}_0^*(A_1 \triangle A_2) = 0$ is equivalent to $S_\epsilon(\mathbf{1}_{A_1 \cap A_2}) = S_\epsilon(\mathbf{1}_{A_1 \cup A_2})$ $\mathbb{P}_0$-a.e. Next, $S_\epsilon(\mathbf{1}_{A_1 \cap A_2}) = S_\epsilon(\mathbf{1}_{A_1 \cup A_2})$ $\mathbb{P}_0$-a.e. is equivalent to $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$ $\mathbb{P}_0$-a.e. by Lemma 142. Therefore, Corollary 143 implies that $\mathbb{P}_0^*(A_1 \triangle A_2) = 0$ is equivalent to Item 2).

The same argument implies that $\mathbb{P}_1^*(A_1 \triangle A_2) = 0$ is equivalent to Item 2). Lastly, $\mathbb{P}^*(A_1 \triangle A_2) = 0$ is equivalent to $\mathbb{P}_0^*(A_1 \triangle A_2) = 0$ and $\mathbb{P}_1^*(A_1 \triangle A_2) = 0$. □

Next, the equivalence of Item 1) with Item 3) in Proposition 51 is a consequence of Lemma 141 and an additional result on the $^\epsilon$ operation.

Lemma 145. *Let U be an open set and let $\mathbb{Q}$ be the set of rational numbers. Further assume $\epsilon > 0$. Then $U^\epsilon = (U \cap \mathbb{Q}^d)^\epsilon = (U \cap (\mathbb{Q}^d)^C)^\epsilon$.*

See Appendix B.3.2.2 for a proof.

Proof of Proposition 51. Lemma 144 states that Item 3) implies Item 2). It remains to show Item 2) implies Item 1) and Item 1) implies Item 3).

Item 2) ⇒ Item 1): Assume that Item 2) holds; then Corollary 143 implies that both $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$ $\mathbb{P}_0$-a.e. and $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C})$ $\mathbb{P}_1$-a.e. Lemma 142 implies than any set A with $A_1 \cap A_2 \subset A \subset A_1 \cup A_2$ satisfies $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_A)$ $\mathbb{P}_0$-a.e. and $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A^C})$ $\mathbb{P}_1$-a.e. Therefore $R^\epsilon(A) = R^\epsilon(A_1)$ so A is also an adversarial Bayes classifier.

Item 1) ⇒ Item 3): Assume that for all A satisfying $A_1 \cap A_2 \subset A \subset A_1 \cup A_2$, the set A is an adversarial Bayes classifier. Define $A_3 = A_1 \cap A_2$, $A_4 = A_1 \cup A_2$, and $D = A_1 \triangle A_2$. As $A_3 \sqcup D \sqcup A_4^C = \mathbb{R}^d$, the boundary ∂D is included in $\partial A_3 \cup \partial A_4$.

We split D into four disjoint sets, $D_1 = \operatorname{int} D \cap \mathbb{Q}^d$, $D_2 = \operatorname{int} D \cap (\mathbb{Q}^d)^C$, $D_3 = D \cap \partial D \cap \partial A_3$, and $D_4 = D \cap \partial D \cap \partial A_4 - D_3$. Notice that these four sets satisfy $D = D_1 \sqcup D_2 \sqcup D_3 \sqcup D_4$. Next, we will prove that each for these four sets has $\mathbb{P}^*$-measure zero.

Because D is a degenerate set, the sets $A_3 \cup D_1$, $A_3 \cup D_2$, and $A_3 \cup \operatorname{int} D$ are all adversarial Bayes classifiers. However, Lemma 145 implies that $D_1^\epsilon = D_2^\epsilon = \operatorname{int} D^\epsilon$ and therefore $S_\epsilon(\mathbf{1}_{A_3 \cup D_1}) = S_\epsilon(\mathbf{1}_{A_3 \cup \operatorname{int} D}) = S_\epsilon(\mathbf{1}_{A_3 \cup D_2})$. Because each of these sets is an adversarial Bayes

classifier, Equation (3.11) of Theorem 30 implies that $\mathbb{P}_0^*(A_3 \cup D_1) = \mathbb{P}_0^*(A_3 \cup \text{int } D) = \mathbb{P}_0^*(A_3 \cup D_2)$. As D_1 and D_2 are disjoint sets whose union is int D, it follows that $\mathbb{P}_0^*(\text{int } D) = 0$. Analogously, comparing $S_\epsilon(\mathbf{1}_{(A_4 - D_1)^C})$, $S_\epsilon(\mathbf{1}_{(A_4 - D_2)^C})$, and $S_\epsilon(\mathbf{1}_{(A_4 - \text{int } D)^C})$ results in $\mathbb{P}_1^*(\text{int } D) = 0$.

Next we argue that $\mathbb{P}^*(D_3) = 0$. Lemma 141 implies that $S_\epsilon(\mathbf{1}_{A_3 \cup D_3}) = S_\epsilon(\mathbf{1}_{A_3})$ $\mathbb{P}_0$-a.e., and Equation (3.11) of Theorem 30 then implies that $\mathbb{P}_0^*(A_3 \cup D_3) = \mathbb{P}_0^*(A_3)$. Thus $\mathbb{P}_0^*(D_3) = 0$ because A_3 and D_3 are disjoint. Similarly, Lemma 141 implies that $S_\epsilon(\mathbf{1}_{(A_3 \cup D_3)^C}) = S_\epsilon(\mathbf{1}_{A_3^C - D_3}) = S_\epsilon(\mathbf{1}_{A_3^C})$ $\mathbb{P}_1$-a.e., and Equation (3.11) of Theorem 30 then implies that $\mathbb{P}_1^*(A_3^C - D_3) = \mathbb{P}_1^*(A_3^C)$, and thus $\mathbb{P}_1^*(D_3) = 0$.

Similarly, one can conclude that $\mathbb{P}^*(D_4) = 0$ by comparing A_4, $A_4 - D_4$, and $A_4 \cup D_4$.

□

B.3.2.1 Proof of Lemma 142

Proof of Lemma 142. We will assume that $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$ $\mathbb{P}_0$-a.e., the argument for $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C})$ $\mathbb{P}_1$-a.e. is analogous. If $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$ $\mathbb{P}_0$-a.e., then

$$S_\epsilon(\mathbf{1}_{A_1}) = \max(S_\epsilon(\mathbf{1}_{A_1}), S_\epsilon(\mathbf{1}_{A_2})) = S_\epsilon(\max(\mathbf{1}_{A_1}, \mathbf{1}_{A_2})) = S_\epsilon(\mathbf{1}_{A_1 \cup A_2}) \quad \mathbb{P}_0\text{-a.e.}$$

However, $S_\epsilon(\mathbf{1}_{A_1^C}) \geq S_\epsilon(\mathbf{1}_{(A_1 \cup A_2)^C})$. If this inequality were strict on a set of positive $\mathbb{P}_1$-measure, we would have $R^\epsilon(A_1 \cup A_2) < R^\epsilon(A_1)$ which would contradict the fact that A_1 is an adversarial Bayes classifier. Thus $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{(A_1 \cup A_2)^C})$ $\mathbb{P}_1$-a.e. The same argument applied to A_2 then shows that $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{(A_1 \cup A_2)^C}) = S_\epsilon(\mathbf{1}_{A_2^C})$ $\mathbb{P}_1$-a.e.

Now as $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C})$ $\mathbb{P}_1$-a.e., one can conclude that

$$S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C}) = \max(S_\epsilon(\mathbf{1}_{A_1^C}), S_\epsilon(\mathbf{1}_{A_2^C})) = S_\epsilon(\mathbf{1}_{(A_1 \cap A_2)^C}) \quad \mathbb{P}_1\text{-a.e.}$$

An analogous argument implies Equation (B.11). □

B.3.2.2 Proof of Lemma 145

Before proving Lemma 145, we reproduce another useful intermediate result from [2].

Lemma 146. *Let* $\mathbf{a}_n$ *be a sequence that approaches* $\mathbf{a}$*. Then* $B_\epsilon(\mathbf{a}) \subset \bigcup_{n=1}^{\infty} B_\epsilon(\mathbf{a}_n)$*.*

Proof. Let $\mathbf{y}$ be any point in $B_\epsilon(\mathbf{a})$ and let $\delta = \|\mathbf{y} - \mathbf{a}\|$. Pick n large enough so that $\|\mathbf{a} - \mathbf{a}_n\| < \epsilon - \delta$. Then

$$\|\mathbf{y} - \mathbf{a}_n\| \leq \|\mathbf{a} - \mathbf{a}_n\| + \|\mathbf{y} - \mathbf{a}\| < \epsilon - \delta + \delta = \epsilon$$

and thus $\mathbf{y} \in B_\epsilon(\mathbf{a}_n)$. □

Proof of Lemma 145. We will argue that $U^\epsilon = (U \cap \mathbb{Q}^d)^\epsilon$, the argument for $U \cap (\mathbb{Q}^d)^C$ is analogous.

First, $U \cap \mathbb{Q}^d \subset U$ implies that $(U \cap \mathbb{Q}^d)^\epsilon \subset U^\epsilon$.

For the opposite containment, let $\mathbf{u}$ be a point in U. We will argue that $\overline{B_\epsilon(\mathbf{u})} \subset (U \cap \mathbb{Q})^\epsilon$. Because U is open, there is a ball $B_r(\mathbf{u})$ contained in U. Because $\mathbb{Q}^d$ is dense in $\mathbb{R}^d$, for every $\mathbf{y} \in B_r(\mathbf{u})$, there is a sequence $\mathbf{y}_n \in \mathbb{Q}$ converging to $\mathbf{y}$. Thus Lemma 146 implies that

$$\overline{B_\epsilon(\mathbf{u})} \subset B_r(\mathbf{u})^\epsilon \subset (B_r(\mathbf{u}) \cap \mathbb{Q}^d)^\epsilon \subset (U \cap \mathbb{Q}^d)^\epsilon$$

Taking a union over all $\mathbf{u} \in U$ results in $U^\epsilon \subset (U \cap \mathbb{Q}^d)^\epsilon$.

□

B.4 Proof of Theorem 34

B.4.1 Proof of Lemma 53

Lemma 24 of [25] show that there exists a function $\hat{\eta}$ and maximizers $\mathbb{P}_0^*, \mathbb{P}_1^*$ of $\bar{R}$ for which optimal attacks on $\hat{\eta}$ are are given by $\mathbb{P}_0^*$, $\mathbb{P}_1^*$:

Proposition 147. *There exists a function $\hat{\eta} : \mathbb{R}^d \to [0,1]$ and measures $\mathbb{P}_0^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0)$, $\mathbb{P}_1^* \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ with the following properties:*

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

$$\hat{\eta}(\mathbf{y}) = \eta^*(\mathbf{y}) \quad \mathbb{P}^* - a.e.$$

2. *Let γ_i^* be a coupling between $\mathbb{P}_i$ and $\mathbb{P}_i^*$ for which $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma_i^*} \|\mathbf{x}-\mathbf{y}\| \le \epsilon$. Then for these $\mathbb{P}_0^*, \mathbb{P}_1^*$, $\hat{\eta}$ satisfies*

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

Recall that Theorem 2.6 of [33] proves that when $W_\infty(\mathbb{Q}, \mathbb{Q}') \le \epsilon$, there always exists a coupling γ between $\mathbb{Q}$ and $\mathbb{Q}'$ with $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x}-\mathbf{y}\| \le \epsilon$.

Next, we prove that one can take $\hat{A}_1 = \{\hat{\eta} > 1/2\}$ and $\hat{A}_2 = \{\hat{\eta} \ge 1/2\}$ in Lemma 53.

Proof of Lemma 53. Let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$, γ_0^*, and γ_1^* be the measures given by Proposition 147 and set $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Let $\hat{\eta}$ be the function described by Proposition 147. We will show that the classifiers $\hat{A}_1 = \{\hat{\eta} > 1/2\}$ and $\hat{A}_2 = \{\hat{\eta} \ge 1/2\}$ satisfy the required properties by verifying the complementary slackness conditions in Theorem 30.

Below, we verify these conditions for $\{\hat{\eta} > 1/2\}$, the argument for $\{\hat{\eta} \ge 1/2\}$ is analogous. First, Item 1 of Proposition 147 implies that $\mathbf{1}_{\{\hat{\eta}>1/2\}} = \mathbf{1}_{\eta^*>1/2}$ $\mathbb{P}^*$-a.e. and $\mathbf{1}_{\{\hat{\eta}>1/2\}^C} =$

$\mathbf{1}_{\{\hat{\eta}\leq 1/2\}} = \mathbf{1}_{\eta^*\leq 1/2}$ $\mathbb{P}^*$-a.e.

Therefore,

$$\eta^*\mathbf{1}_{\{\hat{\eta}>1/2\}^C} + (1-\eta^*)\mathbf{1}_{\{\hat{\eta}>1/2\}} = C^*(\eta^*) \quad \mathbb{P}^*\text{-a.e.}$$

Next, Item 2 of Proposition 147 implies that $\hat{\eta}$ assumes its maximum over closed ϵ-balls $\mathbb{P}_0$-a.e. and hence $S_\epsilon(\mathbf{1}_{\hat{\eta}>1/2})(\mathbf{x}) = \mathbf{1}_{S_\epsilon(\hat{\eta}(\mathbf{x}))>1/2}$ $\mathbb{P}_0$-a.e. Additionally, Item 2 of Proposition 147 implies that $\mathbf{1}_{S_\epsilon(\hat{\eta})(\mathbf{x})>1/2} = \mathbf{1}_{\hat{\eta}(\mathbf{y})>1/2}$ γ_0^*-a.e. Therefore, one can conclude that

$$\int S_\epsilon(\mathbf{1}_{\hat{\eta}>1/2})(\mathbf{x})d\mathbb{P}_0(\mathbf{x}) = \int \mathbf{1}_{\hat{\eta}(\mathbf{y})>1/2}d\gamma_0^*(\mathbf{x},\mathbf{y}) = \int \mathbf{1}_{\hat{\eta}>1/2}d\mathbb{P}_0^* \tag{B.13}$$

Similarly, using the fact that $I_\epsilon(\hat{\eta})(\mathbf{x}) = \hat{\eta}(\mathbf{y})$ γ_1^*-a.e., one can show that $\int S_\epsilon(\mathbf{1}_{\hat{\eta}\leq 1/2})d\mathbb{P}_1 = \int \mathbf{1}_{\hat{\eta}\geq 1/2}d\mathbb{P}_1^*$. This statement together with Equation (B.13) verifies Equation (3.11). □

The classifiers $\hat{A}_1$ and $\hat{A}_2$ are *minimal* and *maximal* classifiers in the sense that

$$\int S_\epsilon(\mathbf{1}_{\hat{A}_1})d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_{\hat{A}_2})d\mathbb{P}_0$$

for any other adversarial Bayes classifier A.

Lemma 148. *Let A be any adversarial Bayes classifier and let $\hat{A}_1$, $\hat{A}_2$ be the two adversarial Bayes classifiers of Lemma 53. Then*

$$\int S_\epsilon(\mathbf{1}_{\hat{A}_1})d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_{\hat{A}_2})d\mathbb{P}_0 \tag{B.14}$$

and

$$\int S_\epsilon(\mathbf{1}_{\hat{A}_2^C})d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_{A^C})d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_{\hat{A}_1^C})d\mathbb{P}_1. \tag{B.15}$$

Proof. Let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$, $\mathbb{P}^*$, and η^* be as described by Lemma 53. Then the complementary slackness condition Equation (3.11) implies that $\int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 = \int \mathbf{1}_A d\mathbb{P}_0^*$ and Equation (3.12) implies Equation (3.15), and hence $\int \mathbf{1}_{\eta^*>1/2}d\mathbb{P}_0^* \leq \int \mathbf{1}_A d\mathbb{P}_0^* \leq \int \mathbf{1}_{\eta^*\geq 1/2}d\mathbb{P}_0^*$. Lemma 53

implies that $\int \mathbf{1}_{\hat{A}_1} d\mathbb{P}_0^* \leq \int \mathbf{1}_A d\mathbb{P}_0^* \leq \int \mathbf{1}_{\hat{A}_2} d\mathbb{P}_0^*$. The complementary slackness condition (3.11) applied to $\hat{A}_1$ and $\hat{A}_2$ then implies Equation (B.14).

The fact that $\int S_\epsilon(\mathbf{1}_{A^C}) = R_*^\epsilon - \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0$ for any adversarial Bayes classifier A then implies Equation (B.15). □

B.4.2 Proving Theorem 34

To start, we prove that Item B) and Item C) are equivalent even when $\mathbb{P} \not\ll \mu$:

Proposition 149. *The following are equivalent:*

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

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

Proof. **Item B) ⇒ Item C):** Assume that $\mathbb{P}_0(A_1^\epsilon) = \mathbb{P}_0(A_2^\epsilon)$ for any two adversarial Bayes classifiers. Then Lemma 27 implies that $\mathbb{P}_0((A_1 \cup A_2)^\epsilon) = \mathbb{P}_0((A_1 \cap A_2)^\epsilon)$. Then $\mathbf{1}_{(A_1 \cup A_2)^\epsilon} = \mathbf{1}_{(A_1 \cap A_2)^\epsilon}$ $\mathbb{P}_0$-a.e. because $(A_1 \cap A_2)^\epsilon \subset (A_1 \cup A_2)^\epsilon$. As A_1^ϵ and A_2^ϵ are strictly between $(A_1 \cap A_2)^\epsilon$ and $(A_1 \cup A_2)^\epsilon$, one can conclude that

$$S_\epsilon(\mathbf{1}_{A_1}) = \mathbf{1}_{A_1^\epsilon} = \mathbf{1}_{A_2^\epsilon} = S_\epsilon(\mathbf{1}_{A_2}) \quad \mathbb{P}_0\text{-a.e.}$$

Similarly, if $\mathbb{P}_1((A_1^C)^\epsilon) = \mathbb{P}_1((A_2^C)^\epsilon)$ implies $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$. Therefore, Item B) implies Item 2) of Lemma 144. Consequently, Lemma 144 implies that $\mathbb{P}^*(\hat{A}_1 \triangle \hat{A}_2) = \mathbb{P}^*(\eta^* = 1/2) = 0$, where $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ are the measures described by Lemma 53 and $\hat{A}_1$ and $\hat{A}_2$ are the adversarial Bayes classifiers described by Lemma 53.

Item C) ⇒ Item B): Assume there is a maximizer $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ of $\bar{R}$ for which $\mathbb{P}^*(\eta^* = 1/2) = 0$, where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Then Equation (3.15) must hold, and

$\mathbb{P}^*(\eta^* = 1/2) = 0$ implies that $\mathbf{1}_A = \mathbf{1}_{\eta^*>1/2}$ $\mathbb{P}^*$-a.e. for any adversarial Bayes classifier A. Consequently, $\mathbf{1}_{A_1} = \mathbf{1}_{A_2}$ $\mathbb{P}^*$-a.e. for any two adversarial Bayes classifiers A_1, A_2 or equivalently, $\mathbb{P}^*(A_1 \triangle A_2) = 0$. Corollary 143 and Lemma 144 imply that $S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2})$ $\mathbb{P}_0$-a.e. and $S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C})$ $\mathbb{P}_1$-a.e., which implies Item B). □

Finally, this result together with Proposition 51 implies Theorem 34.

Proof of Theorem 34. Proposition 149 states that Item B) implies Item C). It remains to show Item A) implies Item B) and Item C) implies Item A).

Item A)⇒Item B): Assume that the the adversarial Bayes classifier is unique up to degeneracy. Then Item 2) of Proposition 51 implies that $\mathbb{P}_1(A_1^\epsilon) = \mathbb{P}_1(A_2^\epsilon)$ for any two adversarial Bayes classifiers A_1 and A_2.

Item C) ⇒ Item A): Assume that $\mathbb{P}^*(\eta^* = 1/2) = 0$ for some $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ that maximize $\bar{R}$, where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$. Then Equation (3.15) implies that $\mathbf{1}_{\eta^*>1/2} = \mathbf{1}_A$ $\mathbb{P}^*$-a.e. for any adversarial Bayes classifier A. Thus if $\mathbb{P}^*(\eta^* = 1/2) = 0$ then $\mathbf{1}_{A_1} = \mathbf{1}_{A_2}$ $\mathbb{P}_0^*$-a.e. for any two adversarial Bayes classifiers A_1, A_2, or in other words, $\mathbb{P}^*(A_1 \triangle A_2) = 0$. Item 3) of Proposition 51 then implies that A_1 and A_2 are equivalent up to degeneracy. As these adversarial Bayes classifiers were arbitrary, the adversarial Bayes classifier is unique up to degeneracy. □

B.5 More about the $^\epsilon$, $^{-\epsilon}$, and S_ϵ operations

This appendix provides a unified exposition of several results relating to the $^\epsilon$ and $^{-\epsilon}$ relations—namely Equations (3.16) and (3.17), Lemmas 55, 59 and 60. These results have all appeared elsewhere in the literature —[2, 16].

The characterizations of the $^\epsilon$ and $^{-\epsilon}$ operations provided by Equation (3.16) and Equation (3.17) are an essential tool for understanding how $^\epsilon$ and $^{-\epsilon}$ interact.

Proof of Equation (3.16). To show Equation (3.16), notice that $\mathbf{x} \in A^\epsilon$ iff $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$ for some element $\mathbf{a}$ of A. Thus:

$$\mathbf{x} \in A^\epsilon \Leftrightarrow \mathbf{x} \in \overline{B_\epsilon(\mathbf{a})} \text{ for some } \mathbf{a} \in A \Leftrightarrow \mathbf{a} \in \overline{B_\epsilon(\mathbf{x})} \text{ for some } \mathbf{a} \in A \Leftrightarrow \overline{B_\epsilon(\mathbf{x})} \text{ intersects } A$$

□

Equation (3.17) then follows directly from Equation (3.16):

Proof of Equation (3.17). By Equation (3.16),

$$\mathbf{x} \in (A^C)^\epsilon \Leftrightarrow \overline{B_\epsilon(\mathbf{x})} \text{ intersects } A^C$$

Now $A^{-\epsilon} = ((A^C)^\epsilon)^C$, and so taking compliments of the relation above implies

$$\mathbf{x} \in A^{-\epsilon} \Leftrightarrow \overline{B_\epsilon(\mathbf{x})} \text{ does not intersect } A^C \Leftrightarrow \overline{B_\epsilon(\mathbf{x})} \subset A$$

□

Next, Equation (3.16) and Equation (3.17) immediately imply Lemma 59:

Proof of Lemma 59. By Equation (3.16), Equation (3.17), $(A^\epsilon)^{-\epsilon}$ is the set of points $\mathbf{x}$ for which $\overline{B_\epsilon(\mathbf{x})} \subset A^\epsilon$. For any point $\mathbf{a} \in A$, $\overline{B_\epsilon(\mathbf{a})} \subset A^\epsilon$ and thus $A \subset (A^\epsilon)^{-\epsilon}$. Applying this statement to the set A^C and then taking compliments results in $(A^{-\epsilon})^\epsilon \subset A$. □

Lemma 59 then immediately implies Lemma 60:

Proof of Lemma 60. First, Lemma 59 implies that $A \subset (A^\epsilon)^{-\epsilon}$ and thus $A^\epsilon \subset ((A^\epsilon)^{-\epsilon})^\epsilon$. At the same time, Lemma 59 implies that $((A^\epsilon)^{-\epsilon})^\epsilon = \left(\left(A^\epsilon\right)^{-\epsilon}\right)^\epsilon \subset A^\epsilon$. Therefore, $((A^\epsilon)^{-\epsilon})^\epsilon = A^\epsilon$. Applying this result to A^C and then taking compliments then results in $((A^{-\epsilon})^\epsilon)^{-\epsilon} = A^{-\epsilon}$.

Next, Lemma 59 implies that $(A^{-\epsilon})^{\epsilon} \subset A$ and hence $((A^{-\epsilon})^{\epsilon})^{\epsilon} \subset A^{\epsilon}$. Applying this result to A^C and then taking compliments $((A^{\epsilon})^{-\epsilon})^{-\epsilon} \supset A^{-\epsilon}$.

□

Lemma 55 is then an immediate consequence of Lemma 60.

B.6 Measurability

B.6.1 Defining the Universal σ-algebra

Let $\mathcal{M}_+(\mathbb{R}^d)$ be the set of finite positive measures on the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$. For a Borel measure ν in $\mathcal{M}_+(\mathbb{R}^d)$, let $\mathcal{L}_\nu(\mathbb{R}^d)$ be the completion of $\mathcal{B}(\mathbb{R}^d)$ under ν. Then the *universal σ-algebra* $\mathscr{U}(\mathbb{R}^d)$ is defined as

$$\mathscr{U}(\mathbb{R}^d) = \bigcap_{\nu \in \mathcal{B}(\mathbb{R}^d)} \mathcal{L}_\nu(\mathbb{R}^d)$$

In other words, $\mathscr{U}(\mathbb{R}^d)$ is the σ-algebra of sets which are measurable with respect to the completion of *every* finite positive Borel measure ν. See [10, Chapter 7] or [45] for more about this construction.

Due to Theorem 56, throughout this paper, we adopt the convention that $\int S_\epsilon(\mathbf{1}_A) d\nu$ is the integral of $S_\epsilon(\mathbf{1}_A)$ with respect to the completion of ν.

B.6.2 Proof of Theorem 57

First, notice that because every Borel set is universally measurable, $\inf_{A \in \mathcal{B}(\mathbb{R}^d)} R^\epsilon(A) \geq \inf_{A \in \mathscr{U}(\mathbb{R}^d)} R^\epsilon(A)$. The opposite inequality relies on a duality statement similar to Theorem 29, but with the primal minimized over universally measurable sets and the dual maximized over measures on $\mathscr{U}(\mathbb{R}^d)$.

For a Borel measure $\mathbb{Q}$, there is a canonical extension to the universal σ-algebra called the *universal completion*.

Definition 150. *The* universal completion $\widetilde{\mathbb{Q}}$ *of a Borel* $\mathbb{Q}$ *is the completion of* $\mathbb{Q}$ *restricted to the universal* σ*-algebra.*

Notice that $\mathbb{Q}(E) = \widetilde{\mathbb{Q}}(E)$ for any Borel measure $\mathbb{Q}$ and Borel set E. As a consequence,

$$\int g d\mathbb{Q} = \int g d\widetilde{\mathbb{Q}} \quad \text{for any Borel function } g. \tag{B.16}$$

In addition to the W_∞-ball of Borel measures $\mathcal{B}_\epsilon^\infty(\mathbb{Q})$ around $\mathbb{Q}$, one can consider the W_∞ ball of universal completions of measures around $\mathbb{Q}$, which we will call $\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{Q})$.

Explicitly, for a Borel measure $\mathbb{Q}$, define

$$\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{Q}) = \{\widetilde{\mathbb{Q}}' : \mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})\}.$$

The following result shows that if $\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})$, then $W_\infty(\widetilde{\mathbb{Q}}, \widetilde{\mathbb{Q}}') \leq \epsilon$, and thus $\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{Q})$ contains only measures that are within ϵ of $\widetilde{\mathbb{Q}}$ in the W_∞ metric.

Lemma 151. *Let* $\mathbb{Q}$ *and* $\mathbb{Q}'$ *be Borel measures with* $W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon$ *and let* $\widetilde{\mathbb{Q}}, \widetilde{\mathbb{Q}}'$ *be their universal completions. Then* $W_\infty(\widetilde{\mathbb{Q}}, \widetilde{\mathbb{Q}}') \leq \epsilon$.

Next, to compare the values of $\bar{R}$ on $\mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \times \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)$ and $\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \times \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)$, we show:

Corollary 152. *Let* $\mathbb{P}_0, \mathbb{P}_1$ *be two Borel measures and let* $\widetilde{\mathbb{P}}_0, \widetilde{\mathbb{P}}_1$ *be their universal completions. Then* $\bar{R}(\mathbb{P}_0, \mathbb{P}_1) = \bar{R}(\widetilde{\mathbb{P}}_0, \widetilde{\mathbb{P}}_1)$.

Thus Lemma 151 and Corollary 152 imply that

$$\sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') = \sup_{\substack{\widetilde{\mathbb{P}}_0' \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \\ \widetilde{\mathbb{P}}_1' \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\widetilde{\mathbb{P}}_0', \widetilde{\mathbb{P}}_1') \tag{B.17}$$

See Appendix B.6.3 for proofs of Lemma 151 and Corollary 152.

Furthermore,Lemma 138 and Equation (B.16) imply:

Corollary 153. *Let $\mathbb{Q}$ be a finite positive measure on $\mathscr{U}(\mathbb{R}^d)$. Then for any universally measurable set A,*

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

See Appendix B.6.3.3 for a proof.

This result implies a weak duality relation between the primal R^ϵ minimized over $\mathscr{U}(\mathbb{R}^d)$ and the dual $\bar{R}$ maximized over $\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \times \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)$:

Lemma 154 (Weak Duality)**.** *Let $\mathbb{P}_0, \mathbb{P}_1$ be two Borel measures. Then*

$$\inf_{A \in \mathscr{U}(\mathbb{R}^d)} R^\epsilon(A) \geq \sup_{\substack{\widetilde{\mathbb{P}}'_0 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \\ \widetilde{\mathbb{P}}'_1 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\widetilde{\mathbb{P}}'_0, \widetilde{\mathbb{P}}'_1)$$

Proof. Let A be any universally measurable set and let $\widetilde{\mathbb{P}}'_0$, $\widetilde{\mathbb{P}}'_1$ be any measures in $\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0)$ and $\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)$ respectively.

Then Corollary 153 implies that

$$\inf_{A \in \mathscr{U}(\mathbb{R}^d)} R^\epsilon(A) \geq \inf_{A \in \mathscr{U}(\mathbb{R}^d)} \sup_{\substack{\widetilde{\mathbb{P}}'_0 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \\ \widetilde{\mathbb{P}}'_1 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \int \mathbf{1}_{A^C} d\widetilde{\mathbb{P}}'_1 + \int \mathbf{1}_A d\widetilde{\mathbb{P}}'_0$$

However, because inf-sup is always larger than a sup-inf, one can conclude that

$$\inf_{A \in \mathscr{U}(\mathbb{R}^d)} R^\epsilon(A) \geq \sup_{\substack{\widetilde{\mathbb{P}}'_0 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \\ \widetilde{\mathbb{P}}'_1 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{A \in \mathscr{U}(\mathbb{R}^d)} \int \mathbf{1}_{A^C} d\widetilde{\mathbb{P}}'_1 + \int \mathbf{1}_A d\widetilde{\mathbb{P}}'_0 = \sup_{\substack{\widetilde{\mathbb{P}}'_0 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0) \\ \widetilde{\mathbb{P}}'_1 \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\widetilde{\mathbb{P}}'_0, \widetilde{\mathbb{P}}'_1)$$

□

This observation suffices to prove Theorem 57:

Proof of Theorem 57. First, because every Borel set is universally measurable,

$$\inf_{A\in\mathcal{B}(\mathbb{R}^d)} R^\epsilon(A) \geq \inf_{A\in\mathscr{U}(\mathbb{R}^d)} R^\epsilon(A).$$

Thus the strong duality result of Theorem 29 and Equation (B.17) imply that

$$\inf_{A\in\mathscr{U}(\mathbb{R}^d)} R^\epsilon(A) \leq \inf_{A\in\mathcal{B}(\mathbb{R}^d)} R^\epsilon(A) = \sup_{\substack{\mathbb{P}_0'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_0)\\ \mathbb{P}_1'\in\mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0',\mathbb{P}_1') = \sup_{\substack{\widetilde{\mathbb{P}}_0'\in\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0)\\ \widetilde{\mathbb{P}}_1'\in\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\widetilde{\mathbb{P}}_0',\widetilde{\mathbb{P}}_1').$$

However, the weak duality statement of Lemma 154 implies that the inequality above must actually be an equality. □

B.6.3 Proofs of Lemma 151 and Corollaries 152 and 153

Lemma 7.26 of [10] provides a useful result for translating statements for $\mathcal{B}(\mathbb{R}^d)$ to $\mathcal{B}(\mathbb{R}^d)$.

Lemma 155. *The set E is universally measurable iff given any Borel measure $\mathbb{Q}$, there are Borel sets B_1, B_2 for which $B_1 \subset E \subset B_2$ and $\mathbb{Q}(B_1) = \mathbb{Q}(B_2)$.*

The proofs of Lemma 151 and Corollaries 152 and 153 all rely on this result.

B.6.3.1 Proof of Lemma 151

Notice that if γ is a coupling between two Borel measures, $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x}-\mathbf{y}\| \leq \epsilon$ iff $\gamma(\Delta_\epsilon^C) = 0$, where Δ_ϵ is the set defined by

$$\Delta_\epsilon = \{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d : \|\mathbf{x}-\mathbf{y}\| \leq \epsilon\}. \tag{B.18}$$

This notation is helpful in the proof of Lemma 151.

Proof of Lemma 151. Let γ be the Borel coupling between $\mathbb{Q}$ and $\mathbb{Q}'$ for which $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x}-\mathbf{y}\| \leq \epsilon$, which exists by Theorem 2.6 of [33]. Let $\overline{\gamma}$ be the completion of γ restricted to

$\sigma(\mathscr{U}(\mathbb{R}^d) \times \mathscr{U}(\mathbb{R}^d))$, the σ-algebra generated by $\mathscr{U}(\mathbb{R}^d) \times \mathscr{U}(\mathbb{R}^d)$. We will show $\overline{\gamma}$ is the desired coupling between $\widetilde{\mathbb{Q}}$ and $\widetilde{\mathbb{Q}}'$. Let S be an arbitrary universally measurable set in $\mathbb{R}^d$. Then Lemma 155 states that there are Borel sets E_1, E_2 for which $E_1 \subset S \subset E_2$ and $\widetilde{\mathbb{Q}}(E_1) = \widetilde{\mathbb{Q}}(S) = \widetilde{\mathbb{Q}}(E_2)$. Then because γ and $\overline{\gamma}$ are equal on Borel sets,

$$\widetilde{\mathbb{Q}}(E_1) = \mathbb{Q}(E_1) = \gamma(E_1 \times \mathbb{R}^d) = \overline{\gamma}(E_1 \times \mathbb{R}^d)$$

and similarly,

$$\widetilde{\mathbb{Q}}(E_2) = \mathbb{Q}(E_2) = \gamma(E_2 \times \mathbb{R}^d) = \overline{\gamma}(E_2 \times \mathbb{R}^d)$$

Therefore,

$$\widetilde{\mathbb{Q}}(S) = \overline{\gamma}(E_1 \times \mathbb{R}^d) = \overline{\gamma}(E_2 \times \mathbb{R}^d) = \overline{\gamma}(S \times \mathbb{R}^d).$$

Similarly, one can argue

$$\widetilde{\mathbb{Q}}'(S) = \overline{\gamma}(\mathbb{R}^d \times S)$$

Therefore, $\overline{\gamma}$ is a coupling between $\widetilde{\mathbb{Q}}$ and $\widetilde{\mathbb{Q}}'$. Next, recall that $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x}-\mathbf{y}\| \leq \epsilon$ iff $\gamma(\Delta_\epsilon^C) = 0$, where Δ_ϵ defined by Equation (B.18).

Therefore, because Δ_ϵ is closed (and thus Borel),

$$\overline{\gamma}(\Delta_\epsilon^C) = \gamma(\Delta_\epsilon^C) = 0$$

Consequently, $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\overline{\gamma}} \|\mathbf{x}-\mathbf{y}\| \leq \epsilon$ and thus $W_\infty(\widetilde{\mathbb{Q}}, \widetilde{\mathbb{Q}}') \leq \epsilon$. □

B.6.3.2 Proof of Corollary 152

Next, we will show:

Lemma 156. *Let ν, λ be two Borel measures with $\nu \ll \lambda$, and let $d\nu/d\lambda$ be the Radon-Nikodym derivative. Then $d\widetilde{\nu}/d\widetilde{\lambda} = d\nu/d\lambda$ $\widetilde{\lambda}$-a.e.*

This result together with Equation (B.16) immediately implies Corollary 152.

Proof. First, if a function g is Borel measurable, $(g^{-1} : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}^d))$, then it is necessarily universally measurable $(g^{-1} : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}^d, \mathscr{U}(\mathbb{R}^d)))$. Thus the Radon-Nikodym derivative $d\nu/d\lambda$ is both Borel measurable and universally measurable.

Next, if $S \in \mathscr{U}(\mathbb{R}^d)$ then Lemma 155 implies there is a Borel set E and λ-null sets N_1, N_2 for which $S = E \cup N_1 - N_2$. Because ν is absolutely continuous with respect to λ, the sets N_1 and N_2 are null under ν as well. Therefore, by the defintion of the Radon-Nikodym derivative $d\nu/d\lambda$ and the fact that $\int g d\lambda = \int g d\widetilde{\lambda}$ for all Borel functions g,

$$\widetilde{\nu}(S) = \nu(E) = \int_E \frac{d\nu}{d\lambda} d\lambda = \int_E \frac{d\nu}{d\lambda} d\widetilde{\lambda} = \int_S \frac{d\nu}{d\lambda} d\widetilde{\lambda}$$

Because the Radon-Nikodym derivative is unique $\widetilde{\lambda}$-a.e., it follows that $d\widetilde{\nu}/d\widetilde{\lambda} = d\nu/d\lambda$ $\widetilde{\lambda}$-a.e. □

B.6.3.3 Proof of Corollary 153

Proof of Corollary 153. Fix a $\mathbb{Q}' \in \widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{Q})$ and assume that $\mathbb{Q}' = \tilde{\lambda}$ for some $\lambda \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})$. Then Lemma 155 states that there is a Borel set $B_1 \subset A$ for which

$$\lambda(B_1) = \tilde{\lambda}(B_1) = \mathbb{Q}'(B_1) = \mathbb{Q}'(A).$$

Thus Lemma 138 and Equation (B.16) imply that $\int \mathbf{1}_{B_1} d\mathbb{Q}' \leq \int S_\epsilon(\mathbf{1}_{B_1}) d\mathbb{Q}$. Furthermore, $B_1 \subset A$ implies $S_\epsilon(\mathbf{1}_{B_1}) \leq S_\epsilon(\mathbf{1}_A)$ and consequently:

$$\int \mathbf{1}_A d\mathbb{Q}' = \int \mathbf{1}_{B_1} d\mathbb{Q}' \leq \int S_\epsilon(\mathbf{1}_{B_1}) d\mathbb{Q} \leq \int S_\epsilon(\mathbf{1}_A) d\mathbb{Q}$$

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

□

B.7 Deferred Proofs From Section 3.5.3

B.7.1 Proof of Lemma 58

Proof of Lemma 58. Let $\{D_i\}_{i=1}^{\infty}$ be a countable sequence of degenerate sets for an adversarial Bayes classifier A. Then by Proposition 51, one can conclude that $S_\epsilon(\mathbf{1}_A) = S_\epsilon(\mathbf{1}_{A\cup D_i}) = \mathbf{1}_{A^\epsilon \cup D_i^\epsilon}$ $\mathbb{P}_0$-a.e. and $S_\epsilon(\mathbf{1}_{A^C}) = S_\epsilon(\mathbf{1}_{A^C\cup D_i}) = \mathbf{1}_{(A^C)^\epsilon\cup D_i^\epsilon}$ $\mathbb{P}_1$-a.e. for every i. Countable additivity then implies that $S_\epsilon(\mathbf{1}_A) = \mathbf{1}_{A^\epsilon\cup\bigcup_{i=1}^\infty D_i^\epsilon} = S_\epsilon(\mathbf{1}_{A\cup\bigcup_{i=1}^\infty D_i})$ $\mathbb{P}_0$-a.e. and $S_\epsilon(\mathbf{1}_{A^C}) = \mathbf{1}_{(A^C)^\epsilon\cup\bigcup_{i=1}^\infty D_i^\epsilon} = S_\epsilon(\mathbf{1}_{A^C\cup\bigcup_{i=1}^\infty D_i})$. Therefore, Proposition 51 implies that A, $A\cup\bigcup_{i=1}^\infty D_i$, and $A-\bigcup_{i=1}^\infty D_i$ are all equivalent up to degeneracy. Consequently, $\bigcup_{i=1}^\infty D_i$ is a degenerate set. □

B.7.2 Proof of Proposition 62

Lemma 157. *Let A be an adversarial Bayes classifier. If C is a connected component of A with $C^{-\epsilon} = \emptyset$, then*

$$C^\epsilon = \{\mathbf{y}\in A^C : \overline{B_\epsilon(\mathbf{y})} \text{ intersects } C\}^\epsilon \tag{B.19}$$

If C is a component of A^C with $C^{-\epsilon} = \emptyset$, then

$$C^\epsilon = \{\mathbf{y}\in A : \overline{B_\epsilon(\mathbf{y})} \text{ intersects } C\}^\epsilon \tag{B.20}$$

Proof. We will prove Equation (B.19), the argument for Equation (B.20) is analogous. Assume that C is a component of A, Equation (3.16) implies the containment $\supset$ of Equation (B.19).

Next, we prove the containment $\subset$ in Equation (B.19). Specifically, we will show that for every $\mathbf{x}\in C^\epsilon$, there is a $\mathbf{y}\in A^C$ for which $\mathbf{x}\in\overline{B_\epsilon(\mathbf{y})}$ and $\overline{B_\epsilon(\mathbf{y})}$ intersects C.

To show the opposite containment, we show that for every $\mathbf{x}\in C^\epsilon$, there is a $\mathbf{y}\in A^C$ for

which $\mathbf{x} \in \overline{B_\epsilon(\mathbf{y})}$ and $\overline{B_\epsilon(\mathbf{y})}$ intersects C.

Let $\mathbf{x} \in C$. Because $C^{-\epsilon} = \emptyset$, Equation (3.17) implies that $\overline{B_\epsilon(\mathbf{x})}$ is not entirely contained in C. Thus the set $C \cup \overline{B_\epsilon(\mathbf{x})}$ is connected and strictly contains C. Recall that a connected component of a set A is a maximal connected subset. If $\overline{B_\epsilon(\mathbf{x})}$ were entirely contained in A, $C \cup \overline{B_\epsilon(\mathbf{x})}$ would be a connected subset of A that strictly contains C, and then C would not be a maximal connected subset of A. Therefore, $\overline{B_\epsilon(\mathbf{x})}$ contains a point $\mathbf{y}$ in A^C, and $\overline{B_\epsilon(\mathbf{y})}$ intersects C at the point $\mathbf{x}$.

Next assume that $\mathbf{x} \in C^\epsilon$ but $\mathbf{x} \notin C$. Then Equation (3.16) states that the ball $\overline{B_\epsilon(\mathbf{x})}$ intersects C at some point $\mathbf{z}$. Consider the line defined by $\ell := \{t\mathbf{x} + (1-t)\mathbf{z} : 0 \leq t \leq 1\}$. Again ℓ is a connected set that intersects C, so $\ell \cup C$ is connected as well. However, ℓ also contains a point not in C and thus if ℓ were entirely contained in A, then $C \cup \ell$ would be a connected subset of A that strictly contains C. As C is a maximal connected subset of A, the set ℓ is not entirely contained in A. Let $\mathbf{y}$ be any point in $A^C \cap \ell$, then $\overline{B_\epsilon(\mathbf{y})}$ intersects C at the point $\mathbf{z}$ and contains $\mathbf{x}$. □

Proof of Proposition 62. First assume that C is a connected component of A with $C^{-\epsilon} = \emptyset$. We will argue that $C \subset (A^\epsilon)^{-\epsilon} - (A^{-\epsilon})^\epsilon$, and then Corollary 61 will imply that C is a degenerate set for A.

If C is a component of A, then $C^\epsilon \subset A^\epsilon$ and thus $C \subset (C^\epsilon)^{-\epsilon} \subset (A^\epsilon)^{-\epsilon}$. Next, Equation (B.19) of Lemma 157 implies that $C^\epsilon \subset (A^C)^\epsilon$ and thus $C \subset (C^\epsilon)^{-\epsilon} \subset ((A^C)^\epsilon)^{-\epsilon} = ((A^{-\epsilon})^\epsilon)^C$. Therefore, C is disjoint from $(A^{-\epsilon})^\epsilon$. Consequently, C is contained in $(A^\epsilon)^{-\epsilon} - (A^{-\epsilon})^\epsilon$, which is degenerate by Lemma 60.

The argument for a connected component of A^C is analogous, with Equation (B.20) in place of Equation (B.19)

As each connected component of A or A^C is contained in the degenerate set $(A^\epsilon)^{-\epsilon} - (A^{-\epsilon})^\epsilon$, it follows that the set in (3.18) is contained in the degenerate set $(A^\epsilon)^{-\epsilon} - (A^{-\epsilon})^\epsilon$. □

B.7.3 Proof of Lemma 63

Proof of Lemma 63. We will show that $\mathbb{P}_0(D^{-\epsilon}) = 0$, the argument for $\mathbb{P}_1$ is analogous. As both $A - D$ and $A \cup D$ are adversarial Bayes classifiers, Proposition 51 implies that $\mathbb{P}_0((A-D)^\epsilon \cup D^\epsilon) = \mathbb{P}_0((A-D)^\epsilon)$ and thus $\mathbb{P}_0(D^\epsilon - (A-D)^\epsilon) = 0$. However, Equation (3.16) and Equation (3.17) imply that

$$\begin{aligned} D^\epsilon - (A-D)^\epsilon &= \{\mathbf{x} : \overline{B_\epsilon(\mathbf{x})} \text{ intersects } D \text{ but not } A - D\} \\ &\supset \{\mathbf{x} : \overline{B_\epsilon(\mathbf{x})} \subset D\} = D^{-\epsilon} \end{aligned}$$

Thus $\mathbb{P}_0(D^{-\epsilon}) = 0$.

□

B.8 Proof of Theorem 35

Proof of Theorem 35. Let $\tilde{A}_1 \subset \tilde{A}_2$ be the adversarial Bayes classifiers defined in Lemma 65 with

$$\tilde{A}_1 = \bigcup_{i=m}^{M} (\tilde{a}_i, \tilde{b}_i), \quad \tilde{A}_2^C = \bigcup_{j=n}^{N} (\tilde{e}_j, \tilde{f}_j).$$

for which $D = \tilde{A}_2 - \tilde{A}_1$ is a degenerate set. Then one can write

$$\mathbb{R} = D \sqcup \bigcup_{i=m}^{M} (\tilde{a}_i, \tilde{b}_i) \sqcup \bigcup_{i=n}^{N} (\tilde{e}_i, \tilde{f}_i) \tag{B.21}$$

For each i, define

$$\hat{a}_i = \inf\{x : (x, \tilde{b}_i) \text{ does not intersect } \tilde{A}_2^C\}$$

$$\hat{b}_i = \sup\{x : (\tilde{a}_i, x) \text{ does not intersect } \tilde{A}_2^C\}$$

and let

$$\hat{A} = \bigcup_{i=m}^{M} (\hat{a}_i, \hat{b}_i)$$

Notice that $(\hat{a}_i, \hat{b}_i) \supset (\tilde{a}_i, \tilde{b}_i)$ so that $\hat{b}_i - \hat{a}_i > 2\epsilon$. Similarly, by the definition of the $\hat{a}_i$ and $\hat{b}_i$, every interval $(\hat{b}_i, \hat{a}_{i+1})$ with $i, i+1 \in [m, M]$ must include some $(\tilde{e}_j, \tilde{f}_j)$ and thus $\hat{b}_i - \hat{a}_{i+1} > 2\epsilon$. As $\tilde{A} \triangle A \subset D$, the set $\tilde{A}$ is an adversarial Bayes classifier equivalent to A.

Next, we will show that any two intervals $(\hat{a}_k, \hat{b}_k)$, $(\hat{a}_p, \hat{b}_p)$ are either disjoint or equal. Assume that $(\hat{a}_k, \hat{b}_k)$ and $(\hat{a}_p, \hat{b}_p)$ intersect at a point x. By the definition of $\hat{b}_k$, $(x, \hat{b}_k)$ does not intersect $\tilde{A}_2^C$ and thus $\hat{b}_p \geq \hat{b}_k$. Reversing the roles of $\hat{b}_p$ and $\hat{b}_k$, one can then conclude that $\hat{b}_p = \hat{b}_k$. One can show that $\hat{a}_p = \hat{a}_k$ via a similar argument. Thus we can choose (a_i, b_i) be unique disjoint intervals for which

$$\bigsqcup_{i=k}^{K} (a_i, b_i) = \bigcup_{i=m}^{M} (\hat{a}_i, \hat{b}_i)$$

□

B.9 Deferred Proofs from Section 3.6.2

B.9.1 Proof of Lemma 68

First, we show Lemma 68 for intervals near the boundary of $\operatorname{supp} \mathbb{P}$.

Lemma 158. *Assume $\mathbb{P} \ll \mu$ and let $A = \bigcup_{i=m}^{M} (a_i, b_i)$ be a regular adversarial Bayes classifier for radius ϵ. Let y represent any of the a_is or b_is. Let I be an interval for which* $\operatorname{supp} \mathbb{P} \subset I$

- *Assume that $I = [\ell, \infty)$ or $I = [\ell, r]$.*

  If $y \in (\ell - \epsilon, \ell + \epsilon]$ then $[\ell - \epsilon, y]$ is a degenerate set. If furthermore $\operatorname{supp} \mathbb{P} = I$*, then for some $\delta > 0$, either $\eta \equiv 0$ or $\eta \equiv 1$ μ-a.e. on $[\ell, \ell + \delta]$.*

- *Assume that* $I = (-\infty, r]$ *or* $I = [\ell, r]$.

  If $y \in [r - \epsilon, r + \epsilon)$ *then* $[y, r - \epsilon]$ *is a degenerate set. If furthermore* $\operatorname{supp} \mathbb{P} = I$*, then for some* $\delta > 0$*, either* $\eta \equiv 0$ *or* $\eta \equiv 1$ μ*-a.e. on* $[r - \delta, r]$.

Proof. We will prove the first bullet; the second bullet follows from the first by considering distributions with densities $\tilde{p}_0(x) = p_0(-x)$ and $\tilde{p}_1(x) = p_1(-x)$.

Assume that some $y = a_i$ is in $(\ell - \epsilon, \ell + \epsilon]$, the argument for $y = b_i$ is analogous. Then because A is adversarial Bayes classifier:

$$0 \geq R^\epsilon(A) - R^\epsilon(A \cup [\ell - \epsilon, a_i]) = \int_\ell^{a_i + \epsilon} p dx - \int_\ell^{a_i + \epsilon} p_0 dx = \int_\ell^{a_i + \epsilon} p_1(x) dx. \tag{B.22}$$

Consequently, $\int_\ell^{a_i+\epsilon} p_1(x)dx = 0$ and thus the set $A \cup [\ell - \epsilon, a_i]$ must be an adversarial Bayes classifier as well.

First, we prove that the interval $[\ell - \epsilon, a_i]$ is a degenerate set. Let D_1, D_2 be arbitrary measurable subsets of $[\ell - \epsilon, a_i]$. Then

$$R^\epsilon(A \cup D_1 - D_2) - R^\epsilon(A \cup [\ell - \epsilon, a_i]) \leq \int_\ell^{a_i + \epsilon} p dx - \int_\ell^{a_i + \epsilon} p_0 dx = \int_\ell^{a_i + \epsilon} p_1(x) dx$$

and this quantity must be zero by Equation (B.22). Therefore, the set $A \cup D_1 - D_2$ is an adversarial Bayes classifier.

Next, we will show that if $\operatorname{supp} \mathbb{P} = I$, then $\eta = 0$ μ-a.e. on a set of positive measure. By assumption $a_i > \ell - \epsilon$ and thus $\delta = a_i + \epsilon - \ell > 0$. As $[\ell, \ell + \delta] \subset \operatorname{supp} \mathbb{P}$, Equation (B.22) implies that $\eta \equiv 0$ μ-a.e. on $[\ell, \ell + \delta]$. □

Proof of Lemma 68. Assume that the endpoints of I are d_1, d_2, so that $I = [d_1, d_2]$ (Corollary 67 implies that $|I| < \infty$). Define an interval J via

$$J = \bigcup_{\substack{I' \supset I: \\ I \text{ degenerate interval}}} I'$$

Because each interval I' includes I, the interval J can be expressed as a countable union of intervals of length at least $|I|$ and thus is a degenerate set as well by Lemma 58. The interval J must be closed because the boundary of every adversarial Bayes classifier is a degenerate set when $\mathbb{P} \ll \mu$. If $J \cap (\operatorname{supp} \mathbb{P}^{\epsilon} - \operatorname{int} \operatorname{supp} \mathbb{P}^{-\epsilon})$ is nonempty, Lemma 158 implies that $\eta \in \{0, 1\}$ on a set of positive measure under $\mathbb{P}$. It remains to consider the case $J \subset \operatorname{int} \operatorname{supp} \mathbb{P}^{-\epsilon}$. Corollary 67 implies that J has finite length and so one can express J as $J = [d_3, d_4]$. Now if any point $\{x\}$ in $[d_3 - \epsilon, d_3)$ were a degenerate set, then Lemma 58 and Lemma 64 would imply that $((J \cup \{x\})^{\epsilon})^{-\epsilon} = [x, d_4]$ would be a degenerate interval strictly containing J, which would contradict the definition of J. Thus $[d_3 - \epsilon, d_3)$ cannot contain any degenerate sets. Similarly, if this interval contains both points in A and A^C, Corollary 61 and Proposition 62 imply that there would be an interval I' that strictly contains J. Thus $[d_3 - \epsilon, d_3)$ must be contained entirely in A or A^C. Similarly, $(d_4, d_4 + \epsilon]$ must be contained entirely in A or A^C.

We will analyze the two cases $(d_3 - \epsilon, d_3]$, $[d_4, d_4 + \epsilon) \subset A$ and $(d_3 - \epsilon, d_3] \subset A$, $[d_4, d_4 + \epsilon) \subset A^C$. The cases $(d_3 - \epsilon, d_3]$, $[d_4, d_4 + \epsilon) \subset A^C$ and $(d_3 - \epsilon, d_3] \subset A^C$, $[d_4, d_4 + \epsilon) \subset A$ are analogous.

Assume first that $(d_3 - \epsilon, d_3]$, $[d_4, d_4 + \epsilon) \subset A$. Then because J is degenerate and $J^{\epsilon} \subset \operatorname{supp} \mathbb{P}$, Corollary 67 implies that $|J| \leq 2\epsilon$. Hence one can conclude

$$0 = R^{\epsilon}(A - J) - R^{\epsilon}(A \cup J) = \int_{d_3 - \epsilon}^{d_4 + \epsilon} p(x)dx - \int_{d_3 - \epsilon}^{d_4 + \epsilon} p_0(x)dx = \int_{d_3 - \epsilon}^{d_4 + \epsilon} p_1(x)dx \geq \int_{d_1 - \epsilon}^{d_2 + \epsilon} p_1(x)dx.$$

Thus on the interval $[d_1 - \epsilon, d_2 + \epsilon]$, one can conclude that $p_1(x) = 0$ μ-a.e. As $[d_1, d_2] \subset \operatorname{int} \operatorname{supp} \mathbb{P}^{-\epsilon}$ and $d_2 > d_1$, one can conclude that $[d_1 - \epsilon, d_2 + \epsilon]$ intersects $\operatorname{supp} \mathbb{P}$ on an open set. Thus $\eta(x) = 0$ μ-a.e. on a set of positive measure.

Next assume that $(d_3 - \epsilon, d_3] \subset A$, $[d_4, d_4 + \epsilon) \subset A^C$. Again, Corollary 67 implies that

$|I| \leq 2\epsilon$. Then:

$$
\begin{aligned}
0 &= R^\epsilon(A \cup (J \cap \mathbb{Q}) - (J \cap \mathbb{Q}^C)) - R^\epsilon(A \cup J) \\
&\geq \int_{d_3-\epsilon}^{d_4+\epsilon} p(x)dx - \left(\int_{d_3-\epsilon}^{d_4-\epsilon} p_0(x)dx + \int_{d_4-\epsilon}^{d_4+\epsilon} p(x)dx \right) \geq \int_{d_3-\epsilon}^{d_4+\epsilon} p_1(x)dx \geq \int_{d_1-\epsilon}^{d_2-\epsilon} p_1(x)dx
\end{aligned}
$$

Thus $p_1(x) = 0$ on $[d_1 - \epsilon, d_2 - \epsilon]$.

Now $[d_1, d_2] \subset \operatorname{int} \operatorname{supp} \mathbb{P}^{-\epsilon}$ implies that $[d_1 - \epsilon, d_2 - \epsilon]$, intersects $\operatorname{supp} \mathbb{P}$ on an open interval. Thus $\eta(x) = 0$ on a set of positive measure. □

B.9.2 Proof of the fourth bullet of Theorem 38

The following lemma implies $(\operatorname{supp} \mathbb{P}^\epsilon)^C$ is a degenerate set.

Lemma 159. *If A and B^ϵ are disjoint, then A^ϵ and B are disjoint.*

Proof. We will show the contrapositive of this statement: if A^ϵ and B intersect, then A and B^ϵ intersect.

If A^ϵ an B intersect, then there are $\mathbf{a} \in A$, $\mathbf{b} \in B$ and $\mathbf{h} \in \overline{B_\epsilon(\mathbf{0})}$ for which $\mathbf{a} + \mathbf{h} = \mathbf{b}$ and thus $\mathbf{a} = \mathbf{b} - \mathbf{h} \in B^\epsilon$. Thus A and B^ϵ intersect. □

Next, we argue that the set $\overline{(\operatorname{supp} \mathbb{P}^\epsilon)^C} \cup \partial A$ is indeed degenerate for any regular adversarial Bayes classifier A. The proof of this result relies on Lemma 139.

Lemma 160. *Assume that $\mathbb{P} \ll \mu$ and let A be a regular adversarial Bayes classifier. Then the set $\overline{(\operatorname{supp} \mathbb{P}^\epsilon)^C} \cup \partial A$ is degenerate for A.*

Proof. First, $\operatorname{supp} \mathbb{P}^\epsilon$ and $(\operatorname{supp} \mathbb{P}^\epsilon)^C$ are disjoint, so Lemma 159 implies that $\operatorname{supp} \mathbb{P}$ and $((\operatorname{supp} \mathbb{P}^\epsilon)^C)^\epsilon)$ are disjoint. Thus $\mathbb{P}((\operatorname{supp} \mathbb{P}^\epsilon)^C)^\epsilon) = 0$, and so $(\operatorname{supp} \mathbb{P}^\epsilon)^C$ is a degenerate set. Next, Lemma 52 implies that $\overline{(\operatorname{supp} \mathbb{P}^\epsilon)^C}$ is a degenerate set. Lemma 52 implies that ∂A is a degenerate set. Lastly, Lemma 58 implies that the union of the three sets ∂A, $(\operatorname{supp} \mathbb{P}^\epsilon)^C$, and $\partial(\operatorname{supp} \mathbb{P}^\epsilon)^C$ is a degenerate set. □

Next, using the fact that $\overline{(\operatorname{supp}\mathbb{P}^\epsilon)^C}$ is degenerate, one can prove the fourth bullet of Theorem 38 for regular adversarial Bayes classifiers.

Lemma 161. *Assume that* $\mathbb{P} \ll \mu$, $\mathbb{P}(\eta = 0 \text{ or } 1) = 0$, *and* $\operatorname{supp}\mathbb{P}$ *is an interval. Then if* D *is a degenerate set for a regular adversarial Bayes classifier* A, *then* $D \subset \overline{(\operatorname{supp}\mathbb{P}^\epsilon)^C} \cup \partial A$.

Proof. Let D be a degenerate set disjoint from $\overline{(\operatorname{supp}\mathbb{P}^\epsilon)^C}$. We will show that $D \subset \partial A$. First, we use a proof by contradiction to argue that the points in $D \cup \partial A$ are strictly greater than 2ϵ apart. If ∂A and D are both degenerate, Lemma 58 implies that $D \cup \partial A$ is degenerate as well. For contradiction, assume that $x \leq y$ are two points in $D \cup \partial A$ with $y - x \leq 2\epsilon$. Then Lemma 64 implies that $[x, y] \subset ((D \cup \partial A)^\epsilon)^{-\epsilon}$ is a degenerate set as well. This statement contradicts Lemma 68. Therefore, $D \cup \partial A$ is comprised of points that are at least 2ϵ apart.

Next, we will show that a degenerate set cannot include any points in $\operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon$ which are more than 2ϵ from ∂A. Let z be any point in $\operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon$ that is strictly more than 2ϵ from ∂A. Assume first that $z \in A$. Then

$$R^\epsilon(A - \{z\}) - R^\epsilon(A) = \int_{z-\epsilon}^{z+\epsilon} \eta(x) d\mathbb{P}$$

However, if $z \in \operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon$ then $(z-\epsilon, z+\epsilon) \not\subset \operatorname{supp}\mathbb{P}^C$ and thus has positive measure under $\mathbb{P}$. As $\eta(x) > 0$ on $\operatorname{supp}\mathbb{P}$, one can conclude that $R^\epsilon(A - \{z\}) - R^\epsilon(A) > 0$. Similarly, if $z \in A^C$, then one can show that $R^\epsilon(A \cup \{z\}) - R^\epsilon(A) > 0$. Therefore z cannot be in any degenerate set.

In summary: $D \cup \partial A$ is comprised of points that are at least 2ϵ apart, but no more than 2ϵ from ∂A. Therefore, one can conclude that $D \subset \partial A$. □

Finally, one can extend Lemma 161 to all adversarial Bayes classifiers by comparing the boundary of a given adversarial Bayes classifier A to the boundary of an equivalent regular adversarial Bayes classifier A_r.

Proof of the fourth bullet of Theorem 38. Any adversarial Bayes classifier A is equivalent up to degeneracy to a regular adversarial Bayes classifier A_r. Lemma 161 implies that $(A_r \triangle A) \cap \operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon \subset \partial A_r \cap \operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon$, where $E_1 \triangle E_2 = E_1 \cap E_2^C \cup E_2 \cap E_1^C$ is the symmetric difference between two sets. Thus there are disjoint sets $S_1, S_2 \subset \partial A_r$ for which $A \cap \operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon = (A_r \cup S_1 - S_2) \cap \operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon$. Because A_r, A_r^C are unions of intervals of length at least 2ϵ, then $\partial A_r = \partial(A_r \cup S_1 - S_2)$ and consequently, $\partial A_r \cap \operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon = \partial A \cap \operatorname{int}\operatorname{supp}\mathbb{P}^\epsilon$. This statement together with Lemma 160 implies the result. □

B.10 Deferred Proofs from Section 3.6.3

In this appendix, we adopt the same notational convention as Section 3.6.3 regarding the a_is and b_is: Namely, when $A = \bigcup_{i=m}^{M}(a_i, b_i)$ is a regular adversarial Bayes classifier, a_{M+1} is defined to be $+\infty$ if M is finite and b_{m-1} is defined to be $-\infty$ if m is finite.

The following observation will assist in proving the first bullet of Lemma 70.

Lemma 162. *Let $\epsilon_2 > \epsilon_1$. If $\mathbb{R}$ minimizes R^{ϵ_2} but $\emptyset$ minimizes R^{ϵ_1}, then both $\mathbb{R}$ and $\emptyset$ minimize both R^{ϵ_1} and R^{ϵ_2}.*

Similarly, if $\emptyset$ minimizes R^{ϵ_2} but $\mathbb{R}$ minimizes R^{ϵ_1}, then both $\mathbb{R}$ and $\emptyset$ minimize both R^{ϵ_1} and R^{ϵ_2}.

Proof. First, assume that $\mathbb{R}$ minimizes R^{ϵ_2} and $\emptyset$ minimizes R^{ϵ_1}. The quantities

$$R^\epsilon(\mathbb{R}) = \int_{\mathbb{R}} d\mathbb{P}_0 \quad R^\epsilon(\emptyset) = \int_{\mathbb{R}} d\mathbb{P}_1$$

are independent of the value of ϵ. Next, notice that $R^{\epsilon_2}(A) \geq R^{\epsilon_1}(A)$ for an set A. Therefore,

$$R_*^{\epsilon_2} \geq R_*^{\epsilon_1} = R^{\epsilon_1}(\emptyset) = R^{\epsilon_2}(\emptyset),$$

where $R_*^\epsilon = \inf_A R^\epsilon(A)$. Thus $\emptyset$ also minimizes R^{ϵ_2}. As a result, the sets $\mathbb{R}$ and $\emptyset$ achieve

the same R^{ϵ_2} risk, and so

$$R^{\epsilon_1}(\mathbb{R}) = R^{\epsilon_2}(\mathbb{R}) = R^{\epsilon_2}(\emptyset) = R^{\epsilon_1}(\emptyset).$$

Consequently, $\mathbb{R}$ is also a minimizer of R^{ϵ_1}.

Next, swapping the roles of $\mathbb{P}_0$ and $\mathbb{P}_1$ shows that if $\emptyset$ minimizes R^{ϵ_2} and $\mathbb{R}$ minimizes R^{ϵ_1} then $\mathbb{R}, \emptyset$ minimize both R^{ϵ_1} and R^{ϵ_2} □

Next, recall that Lemma 158 implies that if the an endpoint of an adversarial Bayes classifier is too close to the boundary of $\operatorname{supp} \mathbb{P}$, then that endpoint must be in the boundary of a degenerate interval. As a result:

Corollary 163. *Assume $\mathbb{P} \ll \mu$ is a measure for which $\operatorname{supp} \mathbb{P}$ is an interval I, and $\mathbb{P}(\eta = 0 \text{ or } 1) = 0$. Then if A is a regular adversarial Bayes classifier at radius ϵ, then A has no finite endpoints in $I^{\epsilon} - \operatorname{int} I^{-\epsilon}$.*

This result implies in the proof of Lemma 70, one only need consider $a_i^1, b_i^1, a_j^2, b_j^2$ contained in $I^{-\epsilon_2}$.

Proof of Lemma 70. We will show that $(b_i^1, a_{i+1}^1) \cap I^{\epsilon_1}$ does not include any non-empty $(a_j^2, b_j^2) \cap I^{\epsilon_1}$, the argument for $(a_i^1, b_i^1) \cap I^{\epsilon_1}$ and $(a_j^2, b_{j+1}^2) \cap I^{\epsilon_1}$ is analogous. Fix an interval (a_j^2, b_j^2) and for contradiction, assume that $(a_j^2, b_j^2) \cap I^{\epsilon_1} \neq \emptyset$ and $(a_j^2, b_j^2) \cap I^{\epsilon_1} \subset (b_i^1, a_{i+1}^1) \cap I^{\epsilon_1}$.

First, notice that the assumption $\eta \neq 0, 1$ implies that none of the a_j^2s, b_j^2s are in $I^{\epsilon_2} - \operatorname{int} I^{-\epsilon_2}$ due to Corollary 163. Thus because the intersection $(a_j^2, b_j^2) \cap I^{\epsilon_1}$ is non-empty, then either $I^{\epsilon_2} \subset (a_j^2, b_j^2)$ or at least one endpoint of (a_j^2, b_j^2) is in $I^{-\epsilon_2}$.

If in fact $(a_j^2, b_j^2) \supset I^{\epsilon_2}$, then $(b_{i+1}^1, a_i^1) \supset (a_j^2, b_j^2)$ must include I^{ϵ_1}. Thus $R^{\epsilon_1}(A_1) = R^{\epsilon_1}(\emptyset)$ while $R^{\epsilon_2}(A_2) = R^{\epsilon_2}(\mathbb{R})$. Lemma 162 then implies that $\mathbb{R}$, $\emptyset$ are both adversarial Bayes classifiers for both perturbation sizes ϵ_1 and ϵ_2, which implies the first bullet of Lemma 70.

Thus, to show the second bullet of Lemma 70, it remains to consider $(a_j^2, b_j^2) \not\supset I^{\epsilon_2}$. As $b_j^2 - a_j^2 > 2\epsilon_2$ and the interval (a_j^2, b_j^2) is included in the adversarial Bayes classifier A_2, it follows that $R^\epsilon(A_2) \leq R^\epsilon(A_2 - (a_j^2, b_j^2))$ which implies

$$\int_{a_j^2-\epsilon_2}^{a_j^2+\epsilon_2} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx + \int_{b_j^2-\epsilon_2}^{b_j^2+\epsilon_2} p dx \leq \int_{a_j^2-\epsilon_2}^{b_j^2+\epsilon_2} p_1 dx$$

and consequently

$$\int_{a_j^2-\epsilon_2}^{b_j^2+\epsilon_2} p_0 dx \leq \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_1 dx. \tag{B.23}$$

Next, $b_j^2 - a_j^2 > 2\epsilon_2$ and thus $(b_j^2 - (\epsilon_2 - \epsilon_1)) - (a_j^2 + (\epsilon_2 - \epsilon_1)) > 2\epsilon_1$. Notice that

$$(a_j^2 + \epsilon_2 - \epsilon_1, b_j^2 - (\epsilon_2 - \epsilon_1)) \cap I^{\epsilon_1} \subset (a_j^2, b_j^2) \cap I^{\epsilon_1} \subset (b_i^1, a_{i+1}^1) \cap I^{\epsilon_1}$$

is then a connected component of $(A_1 \cup (a_j^2 + (\epsilon_2 - \epsilon_1), b_j^2 - (\epsilon_2 - \epsilon_1))) \cap I^{\epsilon_1}$. Therefore,

$$\begin{aligned} &R^{\epsilon_1}(A_1) - R^{\epsilon_1}(A_1 \cup (a_j^2 + \epsilon_2 - \epsilon_1, b_j^2 - (\epsilon_2 - \epsilon_1))) \\ &\qquad = \int_{c_{i,j}}^{d_{i,j}} p_1 dx - \left(\int_{c_{i,j}}^{a_j^2+\epsilon_2} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx + \int_{b_j^2-\epsilon_2}^{d_{i,j}} p dx \right) \end{aligned}$$

where $c_{i,j} = \max(b_i^1 + \epsilon_1, a_j^2 + \epsilon_2 - 2\epsilon_1)$ and $d_{i,j} = \min(a_{i+1}^1 + \epsilon_1, b_j^2 - \epsilon_2 + 2\epsilon_1)$. We will now argue that this quantity is positive, which will contradict the fact that A_1 is an adversarial Bayes classifier.

Adding

$$\int_{c_{i,j}}^{a_j^2+\epsilon_2} p_1 dx + \int_{b_j^2-\epsilon_2}^{d_{i,j}} p_1 dx$$

to both sides of Equation (B.23) implies that

$$\int_{c_{i,j}}^{d_{i,j}} p_1 dx \geq \int_{c_{i,j}}^{a_j^2+\epsilon_2} p dx + \int_{b_j^2-\epsilon_2}^{d_{i,j}} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx + \int_{a_j^2-\epsilon_2}^{c_{i,j}} p_0 dx + \int_{d_{i,j}}^{b_j^2+\epsilon_2} p_0 dx$$
$$> \int_{c_{i,j}}^{a_j^2+\epsilon_2} p dx + \int_{b_j^2-\epsilon_2}^{d_{i,j}} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx \tag{B.24}$$

We will now prove that this last inequality is in fact strict. First, recall that the interval (a_j^2, b_j^2) does not contain I^{ϵ_2} and thus Corollary 163 implies that at least one of a_j^2, b_j^2 must be in $\operatorname{int} I^{-\epsilon_2}$. Consequently, $\operatorname{supp} \mathbb{P}$ must contain at least one of $a_j^2 - \epsilon$ and $b_j^2 + \epsilon$. Lastly, $c_{i,j} - (a_j^2 - \epsilon_2) \geq 2(\epsilon_2 - \epsilon_1) > 0$ and $b_j^2 + \epsilon - d_{i,j} \geq 2(\epsilon_2 - \epsilon_1) > 0$ and thus at least one of the intervals $[a_j^2 - \epsilon, c_{i,j}]$, $[d_{i,j}, b_j^2 + \epsilon_2]$ must have positive $\mathbb{P}$-measure. The assumption $\mathbb{P}(\eta = 0 \text{ or } 1) = 0$ implies $\operatorname{supp} \mathbb{P}_0 = \operatorname{supp} \mathbb{P}_1$ and consequently one of these intervals must have positive $\mathbb{P}_0$-measure.

The strict inequality in Equation (B.24) implies $R^{\epsilon_1}(A_1) - R^{\epsilon_1}(A \cup (a_j^2 + \epsilon_2 - \epsilon_1, b_j^2 - (\epsilon_2 - \epsilon_1))) > 0$, which contradicts the fact that A is an adversarial Bayes classifier.

□

Theorem 39 then directly follows from Lemma 70.

Proof of Theorem 39. The first bullet of Lemma 70 together with the fourth bullet of Theorem 38 imply that if both $\emptyset$, $\mathbb{R}$ are adversarial Bayes classifiers for perturbation size ϵ_i, then either $A \cap I^{\epsilon_i} = \mathbb{R} \cap I^{\epsilon_i}$ and $A^C \cap I^{\epsilon_i} = \emptyset \cap I^{\epsilon_i}$, or $A \cap I^{\epsilon_i} = \emptyset \cap I^{\epsilon_i}$ and $A^\epsilon \cap I^{\epsilon_i} = \mathbb{R} \cap I^{\epsilon_i}$. In either case, one can conclude that $\operatorname{comp}(A \cap I^{\epsilon_1}) + \operatorname{comp}(A^C \cap I^{\epsilon_1}) = 1$ and $\operatorname{comp}(A \cap I^{\epsilon_2}) + \operatorname{comp}(A^C \cap I^{\epsilon_2}) = 1$.

Next, assume that for perturbation size ϵ_1, the sets $\mathbb{R}$, $\emptyset$ are not both adversarial Bayes classifiers. Corollary 163 implies that there are no $a_j^2, b_j^2 \in I^{\epsilon_2} - I^{-\epsilon_2}$. As $I^{-\epsilon_2} \subset I^{\epsilon_1} \subset I^{\epsilon_2}$ are all intervals which are connected sets, one can conclude that $\operatorname{comp}(A_2 \cap I^{\epsilon_2}) = \operatorname{comp}(A_2 \cap I^{\epsilon_1})$ and $\operatorname{comp}(A_2^C \cap I^{\epsilon_2}) = \operatorname{comp}(A_2^C \cap I^{\epsilon_1})$. Therefore, it remains to show that $\operatorname{comp}(A_1 \cap I^{\epsilon_1}) \geq$

$\text{comp}(A_2 \cap I^{\epsilon_1})$ and $\text{comp}(A_1^C \cap I^{\epsilon_1}) \geq \text{comp}(A_2^C \cap I^{\epsilon_1})$. We will show the statement for $A_1 \cap I^{\epsilon_1}$ and $A_2 \cap I^{\epsilon_1}$, the argument for $A_1^C \cap I^{\epsilon_1}$ and $A_2^C \cap I^{\epsilon_1}$ is analogous.

Let

$$A_1 = \bigcup_{i=m}^{M} (a_i^1, b_i^1), \quad A_2 = \bigcup_{j=n}^{N} (a_j^2, b_j^2).$$

Because I^{ϵ_1} is an interval, the intersections $(a_i^1, b_i^1) \cap I^{\epsilon_1}$, $(b_j^2, a_{j+1}^2) \cap I^{\epsilon_1}$ are intervals for $i \in [m, M]$ and $j \in [n, N]$. If the interval $(a_i^1, b_i^1) \cap I^{\epsilon_1}$ intersects both the intervals $(a_j^2, b_j^2) \cap I^{\epsilon_1}$ and $(a_{j+1}^2, b_{j+1}^2) \cap I^{\epsilon_1}$ for some j, then $(a_i^1, b_i^1) \cap I^{\epsilon_1}$ must contain some $(b_j^2, a_{j+1}^2) \cap I^{\epsilon_1}$ for some j, which contradicts Lemma 70. Thus there is at most one interval $(a_j^2, b_j^2) \cap I^{\epsilon_1}$ for each interval $(a_i^1, b_i^1) \cap I^{\epsilon_1}$, which implies that $\text{comp}(A_1 \cap I^{\epsilon_1}) \geq \text{comp}(A_2 \cap I^{\epsilon_1}) = \text{comp}(A_2 \cap I^{\epsilon_2})$. □

B.11 Computational Details of Examples and proofs of Propositions 49 and 50

The following lemma is helpful for verifying the second order necessary conditions for gaussian mixtures.

Lemma 164. *Let* $g(x) = \frac{t}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. *Then* $g'(x) = -\frac{x-\mu}{\sigma^2} g(x)$.

Proof. The chain rule implies that

$$g'(x) = -\frac{x-\mu}{\sigma^2} \cdot \frac{t}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = -\frac{x-\mu}{\sigma^2} g(x)$$

□

B.11.1 Further details from Example 41

It remains to verify two of the claims made in Example 41— namely, 1) that $b(\epsilon)$ does not satisfy the second order necessary condition Equation (3.13b), and 2) Comparing the adversarial risks of $\mathbb{R}$, $\emptyset$, $(a(\epsilon), +\infty)$ to prove that $(a(\epsilon), +\infty)$ is an adversarial Bayes classifier iff $\epsilon \leq \frac{\mu_1 - \mu_0}{2}$ and $\mathbb{R}$, $\emptyset$ are adversarial Bayes classifiers iff $\epsilon \geq \frac{\mu_1 - \mu_0}{2}$.

1) Showing $b(\epsilon)$ doesn't satisfy the second order necessary condition Equation (3.13b)

Due to Lemma 164 the equation Equation (3.13b) reduces to

$$p_0'(b(\epsilon) + \epsilon) - p_1'(b(\epsilon) - \epsilon) = -\frac{b(\epsilon) + \epsilon - \mu_0}{\sigma^2} p_0(b(\epsilon) - \epsilon) + \frac{b(\epsilon) - \epsilon - \mu_1}{\sigma^2} p_1(b(\epsilon) + \epsilon)$$

Furthermore, the first order necessary condition $p_0(b(\epsilon) - \epsilon) - p_1(b(\epsilon) + \epsilon) = 0$ implies that

$$p_0'(b(\epsilon) + \epsilon) - p_1'(b(\epsilon) - \epsilon) = \frac{p_1(b + \epsilon)}{\sigma^2} \left(-(b(\epsilon) + \epsilon - \mu_0) + (b(\epsilon) - \epsilon - \mu_1)\right) = \frac{p_1(b + \epsilon)}{\sigma^2}(\mu_0 - \mu_1 - 2\epsilon)$$

This quantity is negative due to the assumption $\mu_1 > \mu_0$.

2) Comparing the adversarial risks of $\mathbb{R}$, $\emptyset$, and $(a(\epsilon), +\infty)$

First, notice that $R^\epsilon(\emptyset) = R^\epsilon(\mathbb{R}) = \frac{1}{2}$.

Thus it suffices to compare the risks of $(a(\epsilon), +\infty)$ and $\mathbb{R}$. Let

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

be the cdf of a standard gaussian. Then $R^\epsilon((a(\epsilon), +\infty)) \leq R^\epsilon(\mathbb{R})$ iff

$$\int_{-\infty}^{a(\epsilon)+\epsilon} p_1(x)dx + \int_{a(\epsilon)-\epsilon}^{+\infty} p_0(x)dx \leq \int_{-\infty}^{+\infty} p_0(x)dx.$$

Furthermore, because p_0 and p_1 are strictly positive the equation above is equivalent to

$$\int_{-\infty}^{a(\epsilon)+\epsilon} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx \leq \int_{-\infty}^{a(\epsilon)-\epsilon} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}$$

which is also equivalent to $\Phi\left(\frac{a(\epsilon)+\epsilon-\mu_1}{\sigma}\right) \leq \Phi\left(\frac{a(\epsilon)-\epsilon-\mu_0}{\sigma}\right)$. As the function Φ is strictly increasing, this relation is equivalent to the inequality

$$\frac{a(\epsilon)+\epsilon-\mu_1}{\sigma} \leq \frac{a(\epsilon)-\epsilon-\mu_0}{\sigma}$$

which simplifies as $\epsilon \leq \frac{\mu_1-\mu_0}{2}$. Therefore, $(-\infty, a(\epsilon))$ is an adversarial Bayes classifier iff $\epsilon \leq \frac{\mu_1-\mu_0}{2}$ and $\mathbb{R}, \emptyset$ are adversarial Bayes classifiers iff $\epsilon \geq \frac{\mu_1-\mu_0}{2}$.

B.11.2 Further details of Example 42

The constant $k = \ln \frac{(1-\lambda)\sigma_1}{\lambda\sigma_0}$ will feature prominently in subsequent calculations, notice that the assumption $\frac{\lambda}{\sigma_1} > \frac{1-\lambda}{\sigma_0}$ implies that $k < 0$. The equation Equation (3.8b) requires solving $\frac{1-\lambda}{\sigma_0} e^{-(b+\epsilon)^2/2\sigma_0^2} = \frac{\lambda}{\sigma_1} e^{-(b-\epsilon)^2/2\sigma_1^2}$, with solutions Equation (3.14) and

$$y(\epsilon) = \frac{\epsilon\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) - \sqrt{\frac{4\epsilon^2}{\sigma_0^2\sigma_1^2} - 2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)k}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}}. \tag{B.25}$$

The discriminant is positive as $k < 0$ and $\sigma_0 > \sigma_1$. However, one can show that $y(\epsilon)$ does not satisfy the second order necessary condition Equation (3.13b) (see below). Similarly, the only solution to the necessary conditions Equation (3.8a) and Equation (3.13a) is $a(\epsilon) = -b(\epsilon)$.

Thus there are five candidate sets for the adversarial Bayes classifier: $\emptyset$, $\mathbb{R}$, $(-\infty, b(\epsilon))$, $(a(\epsilon), +\infty)$ and $(a(\epsilon), b(\epsilon))$. Theorem 38 implies that none of these sets could be equivalent up to degeneracy. By comparing the adversarial classification risks, one can show that $(a(\epsilon), b(\epsilon))$ has the strictly smallest adversarial classification risk from these five options (see Appendix B.11.2). Therefore, $(a(\epsilon), b(\epsilon))$ is the adversarial Bayes classifier for all ϵ.

It remains to verify two of the claims above— namely, 1) that $y(\epsilon)$ does not satisfy the second order necessary condition Equation (3.13b), and 2) Proving that $(a(\epsilon), b(\epsilon))$ is always the adversarial Bayes classifier by comparing the risks of $(a(\epsilon), b(\epsilon)$, $\mathbb{R}$, $\emptyset$, $(a(\epsilon), \infty)$, and $(-\infty, b(\epsilon))$.

1) The point $y(\epsilon)$ does not satisfy the second order necessary condition Equation (3.13b)

First, notice that

$$y(\epsilon) \leq \frac{\epsilon\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) - \sqrt{\frac{4\epsilon^2}{\sigma_0^2\sigma_1^2}}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}} = \frac{\epsilon\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) - \frac{2\epsilon}{\sigma_0\sigma_1}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}} \tag{B.26}$$

This bound shows that $y(\epsilon)$ fails to satisfy the second order necessary condition Equation (3.13b). One can compute the derivative p_i' in terms of p_i using Lemma 164. Specifically, $p_i'(x) = \frac{-x}{\sigma_i^2} p_i(x)$ and therefore

$$p_0'(y(\epsilon)+\epsilon) - p_1'(y(\epsilon)-\epsilon) = -\frac{y(\epsilon)+\epsilon}{\sigma_0^2} p_0(y(\epsilon)+\epsilon) + \frac{y(\epsilon)-\epsilon}{\sigma_1^2} p_1(y(\epsilon)-\epsilon)$$

The first order condition $p_0(y(\epsilon)+\epsilon) - p_1(y(\epsilon)-\epsilon) = 0$ implies

$$p_0'(y(\epsilon)+\epsilon) - p_1'(y(\epsilon)-\epsilon) = p_0(y(\epsilon)+\epsilon)\left(y(\epsilon)\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) - \epsilon\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)\right)$$

However, Equation (B.26) implies that

$$p_0(y(\epsilon)+\epsilon)\left(y(\epsilon)\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)-\epsilon\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_0^2}\right)\right) \leq p_0(y(\epsilon)+\epsilon)\cdot\frac{-2\epsilon}{\sigma_0\sigma_1} < 0$$

Thus, the only solution to first Equation (3.8b) and Equation (3.13b) is $b(\epsilon)$.

2) Comparing the risks of $(a(\epsilon), b(\epsilon))$, $\mathbb{R}$, $\emptyset$, $(a(\epsilon), \infty)$, and $(-\infty, b(\epsilon))$

First, we argue that $R^\epsilon\left((a(\epsilon), \infty)\right) > R^\epsilon\left((a(\epsilon), b(\epsilon)\right)$:

$$\begin{aligned} R^\epsilon((a(\epsilon),\infty)) - R^\epsilon((a(\epsilon),b(\epsilon))) &= \int_{b(\epsilon)+\epsilon}^{+\infty} p_0(x) - p_1(x)dx - \int_{b(\epsilon)-\epsilon}^{b(\epsilon)+\epsilon} p_1(x)dx \\ &= \int_{b(\epsilon)}^{\infty} p_0(x+\epsilon) - p_1(x-\epsilon)dx \end{aligned} \tag{B.27}$$

The same calculation that solves for $b(\epsilon)$ in Equation (3.14) and $y(\epsilon)$ in Equation (B.25) then shows that $p_0(x+\epsilon) - p_1(x-\epsilon)$ is strictly positive when $x > b(\epsilon)$.

Additionally, $R^\epsilon((a(\epsilon), +\infty)) = R^\epsilon((-\infty, b(\epsilon)))$ because $a(\epsilon) = -b(\epsilon)$ and p_0, p_1 are symmetric around zero. Furthermore, by writing out the integrals as in the first line of Equation (B.27), one can show that $R^\epsilon(\mathbb{R}) - R^\epsilon((-\infty, b(\epsilon))) = R^\epsilon((a(\epsilon), \infty)) - R^\epsilon((a(\epsilon), b(\epsilon)))$. Thus

$$R^\epsilon(\mathbb{R}) - R^\epsilon((a(\epsilon), b(\epsilon))) = 2(R^\epsilon((a(\epsilon), \infty)) - R^\epsilon((a(\epsilon), b(\epsilon))) > 0$$

and hence one can conclude that $R^\epsilon((a(\epsilon), b(\epsilon))) < R^\epsilon(\mathbb{R})$ and $R^\epsilon((a(\epsilon), b(\epsilon))) < R^\epsilon((-\infty, b(\epsilon)))$. Similarly, one can show that

$$R^\epsilon(\emptyset) - R^\epsilon((a(\epsilon), b(\epsilon))) = 2(R^\epsilon((a(\epsilon), \infty)) - R^\epsilon((a(\epsilon), b(\epsilon)))) > 0$$

and thus $R^\epsilon(\emptyset) > R^\epsilon((a(\epsilon), b(\epsilon)))$.

B.11.3 Proof of Lemma 43

Lemmas 158 and 160 of Appendix B.9.1 are used in the proof of Lemma 43.

Proof of Lemma 43. Without loss of generality one can assume that A is a union of open intervals due to Lemma 52.

There is nothing to show if $\operatorname{supp} \mathbb{P} = \mathbb{R}$.

We now consider smaller support— for concreteness, we will assume that $\operatorname{supp} \mathbb{P} = [\ell, \infty)$, the cases $\operatorname{supp} \mathbb{P} = [\ell, r]$, $\operatorname{supp} \mathbb{P} = (-\infty, r]$ have analogous reasoning.

Let

$$i^* = \operatorname*{argmin}_{a_i \geq \ell} a_i - \ell$$

$$j^* = \operatorname*{argmin}_{b_i \geq \ell} b_i - \ell$$

We will now consider four cases:

I) $|\ell - a_{i^*}| \leq |\ell - b_{j^*}|$ and $a_{i^*} > \ell + \epsilon$; in which case $A' = (a_{i^*}, +\infty) \cap A$ is the desired adversarial Bayes classifier

II) $|\ell - a_{i^*}| \leq |\ell - b_{j^*}|$ and $a_{i^*} \leq \ell + \epsilon$; in which case $A' = (-\infty, a_{i^*}] \cup A$ is the desired adversarial Bayes classifier

III) $|\ell - a_{i^*}| > |\ell - b_{j^*}|$ and $b_{j^*} > \ell + \epsilon$; in which case $A' = (-\infty, b_{j^*}) \cup A$ is the desired adversarial Bayes classifier

IV) $|\ell - a_{i^*}| > |\ell - b_{j^*}|$ and $b_{j^*} \leq \ell + \epsilon$; we will show $A' = (b_{j^*}, \infty) \cap A$ is the desired adversarial Bayes classifier

We will show Items I) and II), the arguments for Items III) and IV) is analogous.

Item I): First, we argue that A and A' are equivalent. Lemma 160 implies that $(-\infty, \ell - \epsilon]$ is a degenerate set. Next, there can be at most one point of ∂A in $[\ell - \epsilon, \ell]$ because A is

regular. By the definition of i_*, if there is some point of ∂A in $[\ell-\epsilon, \ell]$, that point must be b_{i^*-1}.

- If $b_{i^*-1} \notin [\ell-\epsilon, \ell]$, then $A' = A - (-\infty, \ell-\epsilon]$ and thus A and A' are equivalent.
- If $b_{i^*-1} \in [\ell-\epsilon, \ell]$, then Lemma 158 implies that $[\ell-\epsilon, b_{i^*-1}]$ is a degenerate set, and thus $A' = A - ((-\infty, \ell-\epsilon] \cup [\ell-\epsilon, b_{i^*-1}])$ and consequently A and A' are equivalent.

Next, we show that $A' := (a_{i^*}, \infty) \cap A$ is a regular set. Because A is regular, the point a_{i^*} is more than 2ϵ from any other boundary point of ∂A. As $\partial((a_{i^*}, +\infty) \cap A) \subset \partial(-\infty, a_{i^*}) \cup \partial A = \partial A$, the point a_i^* must be more than 2ϵ from any other boundary point of $(a_{i^*}, +\infty) \cap A$. Therefore, A' is regular.

The assumption $a_{i^*} > \ell + \epsilon$ implies that $A' \subset (\ell + \epsilon + \delta, +\infty)$ for some $\delta > 0$ and consequently A' can only have boundary points in $(\ell+\epsilon, +\infty) = \operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$.

Finally, A' is open as it is the intersection of open sets.

Item II): First, we argue that A and A' are equivalent. Lemmas 158 and 160 imply that the sets $(-\infty, \ell-\epsilon]$ and $[\ell-\epsilon, a_{i^*}]$ are degenerate sets for A. Therefore, A and A' are equivalent up to degeneracy.

Next, the same argument as Item I) shows that $A' = A \cup (-\infty, a_{i^*}]$ is a regular set: $\partial(A \cup (-\infty, a_{i^*}]) \subset \partial A \cup \partial(-\infty, a_{i^*}] = \partial A$. Thus the boundary points of A' must be at least 2ϵ apart because A is regular.

Further, the set A' is open because $(-\infty, a_{i^*}] \cup (a_{i^*}, b_{i^*}) = (-\infty, b_{i^*})$ and consequently, $A' = (-\infty, b_{i^*}) \cup A$.

Finally, to show that $\partial A' \subset \operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$, we argue that A' has no boundary points in $(-\infty, \ell+\epsilon] = (\operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon})^C$. As $(-\infty, b_{i^*}) \subset A'$, the set A' has no boundary points in $(-\infty, b_{i^*}]$. However, the interval $(-\infty, b_{i^*}]$ contains $(-\infty, \ell+\epsilon]$ as $b_{i^*} - a_{i^*} > 2\epsilon$ because A is regular. □

B.11.4 Example 45 details

Theorem 37 implies that when $\epsilon < 1/2$ the candidate solutions for the a_i, b_i are $[-\epsilon, \epsilon] \cup \{-1-\epsilon, -1+\epsilon, 1-\epsilon, 1+\epsilon\}$. However, Lemma 43 implies that one only needs to consider points a_i, b_i in $[-\epsilon, \epsilon]$ when identifying adversarial Bayes classifiers under equivalence up to degeneracy. However, $R^\epsilon((y, \infty)) < R^\epsilon((-\infty, y))$ for any $y \in [-\epsilon, \epsilon]$ because $p_1(x) > p_0(x)$ for $x > \epsilon$ while $p_1(x) - p_0(x) < 0$ for any $x < -\epsilon$. Thus, the candidate sets for the adversarial Bayes classifier are $\mathbb{R}$, $\emptyset$, and (y, ∞) for any $y \in [-\epsilon, \epsilon]$. Next, any point $y \in [-\epsilon, \epsilon]$ achieves the same risk: $R^\epsilon((y, \infty)) = \epsilon + \frac{1}{4}(1 - \epsilon)$ while $R^\epsilon(\mathbb{R}) = R^\epsilon(\emptyset) = 1/2$. Thus $\emptyset, \mathbb{R}$ are adversarial Bayes classifiers when $\epsilon \in [1/3, 1/2)$ and (y, ∞) is an adversarial Bayes classifier only when $\epsilon \leq 1/3$. Thus Theorem 39 implies that (y, ∞) is an adversarial Bayes classifier for any $y \in [-\epsilon, \epsilon]$ iff $\epsilon \leq 1/3$ while $\mathbb{R}$, $\emptyset$ are adversarial Bayes classifiers iff $\epsilon \geq 1/3$.

B.11.5 Example 46 details

It remains to compare the adversarial risks of all sets whose boundary is included in $\{-1/4 \pm \epsilon, 1/4 \pm \epsilon\}$ for all $\epsilon > 0$. As points in the boundary of a regular adversarial Bayes classifier must be more than 2ϵ apart, the boundary of a regular adversarial Bayes classifier can include at most one of $\{-\frac{1}{4} - \epsilon, -\frac{1}{4} + \epsilon\}$ and at most one of $\{\frac{1}{4} - \epsilon, \frac{1}{4} + \epsilon\}$. Let $\mathcal{S}$ be the set of open sets with at most one boundary point in $\{-\frac{1}{4} - \epsilon, -\frac{1}{4} + \epsilon\}$, at most one boundary point in $\{\frac{1}{4} - \epsilon, \frac{1}{4} + \epsilon\}$, and no other boundary points.

Instead of explicitly computing the adversarial risk of each set in $\mathcal{S}$, we will rule out most combinations by understanding properties of such sets, and then comparing to the adversarial risk of $\mathbb{R}$, for which $R^\epsilon(\mathbb{R}) = 1/10$ for all possible ϵ. We consider three separate cases:

When $\epsilon > 1/4$: If a set A includes at least one endpoint in $\operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$, then

$$R^\epsilon(A) \geq 2\epsilon \inf_{x \in \operatorname{supp}\mathbb{P}} p(x) \geq \frac{2\epsilon}{5} > \frac{1}{10} = R^\epsilon(\mathbb{R})$$

The only two sets in $\mathcal{S}$ that have no endpoints in $\operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$ are $\mathbb{R}$ and $\emptyset$, but $R^\epsilon(\emptyset) = 9/10$. Thus if $\epsilon > 1/4$, then $\mathbb{R}$ is an adversarial Bayes classifier, and this classifier is unique up to degeneracy.

When $1/8 \leq \epsilon \leq 1/4$: If either $1/4 + \epsilon, -1/4 - \epsilon$ are in the boundary of a set A, then

$$R^\epsilon(A) \geq \int_{y-\epsilon}^{y+\epsilon} p(x)dx = \frac{3}{5} \cdot 2\epsilon \geq \frac{3}{20} > R^\epsilon(\mathbb{R}).$$

(The value y above is either $1/4 + \epsilon$ or $-1/4 - \epsilon$.) Consequently, for these values of ϵ, only sets in $\mathcal{S}$ with at most one endpoint in $\{-1/4 + \epsilon\}$ and at most one endpoint in $\{1/4 - \epsilon\}$ can be adversarial Bayes classifiers.

Next, if a set A in $\mathcal{S}$ excludes either $(-\infty, -1/4)$ or $(1/4, \infty)$, then

$$R^\epsilon(A) \geq \int_S p_1(x)dx \geq \frac{3}{5} \cdot \frac{3}{4} > R^\epsilon(\mathbb{R}).$$

(The set S above represents either $(-\infty, -1/4)$ or $(1/4, \infty)$.) As a result, such a set cannot be an adversarial Bayes classifier.

However, $\mathbb{R}$ and $(-\infty, -1/4 + \epsilon) \cup (1/4 - \epsilon, \infty)$ are the only two sets in $\mathcal{S}$ with at most one endpoint in $\{-1/4 + \epsilon\}$ and at most one endpoint in $\{1/4 - \epsilon\}$, that include $(-\infty, -1/4) \cup (1/4, \infty)$. The set $(-\infty, -1/4 + \epsilon) \cup (1/4 - \epsilon, \infty)$ is not a regular set when $\epsilon \geq 1/8$. Consequently, When $\epsilon \in (1/8, 1/4]$, the set $\mathbb{R}$ is an adversarial Bayes classifier, and this classifier is unique up to degeneracy.

When $\epsilon < 1/8$: First, if A excludes $[-1-\epsilon, -1/4-\epsilon)$ or $(1/4+\epsilon, 1+\epsilon]$, then

$$R^\epsilon(A) \geq \frac{3}{5} \cdot (\frac{3}{4} - \epsilon) \geq \frac{3}{5} \cdot (\frac{3}{4} - \frac{1}{8}) = \frac{3}{8} > R^\epsilon(\mathbb{R}).$$

There are only five sets in $\mathcal{S}$ that satisfy this requirement: $A_1 = (-\infty, -1/4+\epsilon) \cup (1/4-\epsilon, \infty)$, $A_2 = (-\infty, -1/4-\epsilon) \cup (1/4-\epsilon, \infty)$, $A_3 = (-\infty, -1/4+\epsilon) \cup (1/4+\epsilon, \infty)$, $A_4 = (-\infty, -1/4-\epsilon) \cup (1/4+\epsilon, \infty)$, and $A_5 = \mathbb{R}$. All of these sets are regular when $\epsilon < 1/8$. One can compute:

$$R^\epsilon(A_1) = \frac{4\epsilon}{5}, R^\epsilon(A_2) = R^\epsilon(A_3) = \frac{8\epsilon}{5}, \text{ and } R^\epsilon(A_4) = \frac{6}{5}\epsilon$$

Of these five alternatives, the set A_1 has the strictly smallest risk when $\epsilon \in (0, 1/8)$. Consequently, when $\epsilon \in (0, 1/8)$, the set A_1 is the adversarial Bayes classifier and is unique up to degeneracy.

B.11.6 Proof of Proposition 49

Proof of Proposition 49. Due to Theorem 35 and Lemma 43, any adversarial Bayes classifier is equivalent up to degeneracy to a regular adversarial Bayes classifier $A = \bigcup_{i=m}^{M} (a_i, b_i)$ for which all the finite a_i and b_i are contained in $\operatorname{int} \operatorname{supp} \mathbb{P}^{-\epsilon}$. Consequently, if there is some a_i or b_i in $\operatorname{int} \operatorname{supp} \mathbb{P}^{-\epsilon}$, then $\epsilon < |\operatorname{supp} \mathbb{P}|/2$.

For every point x in $\operatorname{int} \operatorname{supp} \mathbb{P}^{-\epsilon}$, the densities p_0 and p_1 are both continuous at $x - \epsilon$ and $x + \epsilon$. Consequently, the necessary conditions Equation (3.8) reduce to

$$\eta(a+\epsilon) = 1 - \eta(a-\epsilon) \qquad \text{(B.28a)} \qquad \eta(b-\epsilon) = 1 - \eta(b+\epsilon) \qquad \text{(B.28b)}$$

on this set. If a is more than ϵ away from a point z satisfying $\eta(z) = 1/2$, the continuity of η implies that $\eta(a+\epsilon), \eta(a-\epsilon)$ are either both strictly larger than $1/2$ or strictly smaller than $1/2$, and thus a would not satisfy Equation (B.28a). As a result, every a_i must be within ϵ of a solution to $\eta(z) = 1/2$. An analogous argument shows that the same holds for solutions to Equation (B.28b). □

B.11.7 Proof of Proposition 50

Proof of Proposition 50. Due to Theorem 35 and Lemma 43, any adversarial Bayes classifier is equivalent up to degeneracy to a regular adversarial Bayes classifier $A = \bigcup_{i=m}^{M}(a_i, b_i)$ for which all the finite a_i and b_i are contained in $\operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$. Consequently, if there is some a_i or b_i in $\operatorname{int}\operatorname{supp}\mathbb{P}^{\epsilon}$, then $\epsilon < |\operatorname{supp}\mathbb{P}|/2$.

For contradiction, assume that a_i is not within ϵ of any point in $\partial\{\eta = 1\}$. Then for some $r > 0$, η is either identically 1 or identically 0 on $(a_i(\epsilon)-\epsilon-r, a_i(\epsilon)+\epsilon+r)$ and thus $p_1 = p\eta$ is continuous on this set. Furthermore, because $a_i \in \operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$ but $\epsilon < |\operatorname{supp}\mathbb{P}|/2$, $p_1(a_i+\epsilon)$ is strictly positive while $p_0(a_i - \epsilon) = 0$. Consequently, $a(\epsilon)$ cannot satisfy the necessary condition Equation (3.8a), thus contradicting Theorem 37. □

B.11.8 Example 69 details

It remains to compare the risks of all regular sets with endpoints in $\{-\frac{7}{2}, -\frac{5}{2}\epsilon, -\frac{3}{2}\epsilon, -\frac{1}{2}\epsilon, +\frac{1}{2}\epsilon, +\frac{3}{2}\epsilon, +\frac{5}{2}\epsilon, +\frac{7}{2}\epsilon\}$, and show that $\mathbb{R}$ is indeed an adversarial Bayes classifier. Rather than explicitly writing out all such sets and computing their adversarial risks, we show that one need not consider certain sets in $\mathcal{S}$ because if they were adversarial Bayes classifiers, they would be equivalent up to degeneracy to other sets in $\mathcal{S}$.

First, Lemma 158 with $I = [-\frac{5}{2}\epsilon, +\frac{5}{2}\epsilon]$ implies that if A is a regular adversarial Bayes classifier and $y \in \{-\frac{7}{2}\epsilon, -\frac{5}{2}\epsilon, -\frac{3}{2}\epsilon\}$ is in ∂A, then $[-\frac{7}{2}\epsilon, y]$ is a degenerate set. Thus there is no need to consider classifiers with endpoints in $\{-\frac{7}{2}\epsilon, -\frac{5}{2}\epsilon, -\frac{3}{2}\epsilon\}$ when identifying all possible adversarial Bayes classifiers under equivalence up to degeneracy. Similarly, Lemma 158 also implies that there is no need to consider $\{+\frac{3}{2}\epsilon, +\frac{5}{2}\epsilon, +\frac{7}{2}\epsilon\}$ as possible values of the a_is or b_is. Thus it remains to compare the risks of regular sets whose boundary is contained in $\{-\frac{1}{2}\epsilon, -\frac{1}{2}\epsilon\}$. As points in the boundary of a regular set are at least 2ϵ apart, one can rule out sets with more than one boundary point in $\{-\frac{1}{2}\epsilon, +\frac{1}{2}\epsilon\}$.

Consequently, it remains to compare the adversarial risks of six sets: $\mathbb{R}$, $\emptyset$, $(-\frac{1}{2}\epsilon, +\infty)$, $(-\infty, -\frac{1}{2}\epsilon)$, $(+\frac{1}{2}\epsilon, +\infty)$, and $(-\infty, +\frac{1}{2}\epsilon)$. The adversarial risks of these sets are:

$$R^\epsilon(\mathbb{R}) = \frac{14}{25} \qquad R^\epsilon(\emptyset) = \frac{11}{25}$$

$$R^\epsilon\left((-\frac{1}{2}\epsilon, +\infty)\right) = R^\epsilon\left((+\frac{1}{2}\epsilon, +\infty)\right) = \frac{21}{25}$$

$$R^\epsilon\left((-\infty, -\frac{1}{2}\epsilon)\right) = R^\epsilon\left((-\infty, +\frac{1}{2}\epsilon)\right) = \frac{9}{25}$$

Therefore, the set $(-\infty, -\frac{1}{2}\epsilon)$ is an adversarial Bayes classifier.

C — Deferred Proofs from Chapter 4

C.1 An Alternative Characterization of Consistency– Proof of Proposition 71

First, prior work computes the minimum standard ϕ-risk.

Lemma 165. *Let ϕ be any monotonic loss function. Then*

$$\inf_{f \text{ measurable}} R_\phi(f) = \int C_\phi^*(\eta) d\mathbb{P}$$

This result appears on page 4 of [8]. Notice that Lemma 165 is Theorem 78 with $\epsilon = 0$. Next, one can use the following lemma to compare minimizing sequences of $C_\phi(\eta, \cdot)$ and $C(\eta, \cdot)$.

Lemma 166. *Assume that Assumption 2 holds, ϕ is consistent, and $0 \in \operatorname{argmin} C_\phi(\eta, \cdot)$. Then $\eta = 1/2$.*

Proof. Consider a distribution for which $\eta(\mathbf{x}) \equiv \eta$ is constant. Then $R_\phi(f) = C_\phi(\eta, f)$ and $R(f) = C(\eta, f)$. The consistency of ϕ implies that 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, $\eta = 1/2$. □

We use this result to prove Proposition 71 together with a standard argument from analysis:

Lemma 167. *Let $\{a_n\}$ be a sequence in $\mathbb{R} \cup \{\infty\}$. Then the following are equivalent:*

1) $\lim_{n\to\infty} a_n = a$

2) Every subsequence $\{a_{n_j}\}$ of $\{a_n\}$ has a subsequence $\{a_{j_k}\}$ for which $\lim_{k\to\infty} a_{j_k} = a$

As a result:

Corollary 168. *If every minimizing sequence f_n of R_ϕ has a subsequence f_{n_j} that minimizes R, then ϕ is consistent.*

Furthermore, this corollary can be applied to a distribution with constant $\eta(\mathbf{x})$ to conclude:

Corollary 169. *If every minimizing sequence α_n for $C_\phi(\eta, \cdot)$ has a subsequence α_{n_j} that minimizes $C(\eta, \cdot)$ then one can conclude that every minimizing sequence is of $C_\phi(\eta, \cdot)$ is also a minimizing sequence of $C(\eta, \cdot)$.*

We now prove a result slightly stronger than Proposition 71.

Theorem 170. *The following are equivalent:*

1) For all distributions, f_n is a minimizing sequence of R_ϕ implies that f_n is a minimizing sequence of R.

2) For all $\eta \in [0, 1]$, α_n is a minimizing sequence of $C_\phi(\eta, \cdot)$ implies that α_n is a minimizing sequence of $C(\eta, \cdot)$.

3) Every minimizer of $C_\phi(\eta, \cdot)$ is also a minimizer of $C(\eta, \cdot)$.

4) Every minimizer of R_ϕ is a minimizer of R

The proof is essentially the "pointwise" argument discussed in Section 4.3.

Proof. We show that 1) $\Leftrightarrow$ 2), 2) $\Leftrightarrow$ 3), and 3) $\Leftrightarrow$ 4).

Showing 1) is equivalent to 2):

To show that 1) implies 2), consider a distribution for which $\eta(\mathbf{x}) \equiv \eta$ is constant.

For the other direction, let f_n be any minimizing sequence of R_ϕ. Then $C_\phi(\eta, f_n) \geq C_\phi^*(\eta)$ and Lemma 165 implies that the sequence $C_\phi(\eta, f_n)$ actually converges to $C_\phi^*(\eta)$ in $L^1(\mathbb{P})$. Thus one can pick a subsequence f_{n_j} for which $C_\phi(\eta, f_{n_j})$ converges to $C_\phi^*(\eta)$ $\mathbb{P}$-a.e. (See for instance Corollary 2.32 of [22]). Then 2) implies that the function sequence f_{n_j} minimizes $C(\eta, \cdot)$ and therefore it also minimizes R by Corollary 168.

Showing 2) is equivalent to 3):

To show that 2) implies 3), notice that if α is a minimizer of $C_\phi(\eta, \cdot)$, 2) immediately implies that the sequence $\alpha_n \equiv \alpha$ also minimizes $C(\eta, \cdot)$.

For the other direction, assume that every minimizer of $C_\phi(\eta, \cdot)$ is also a minimizer of $C(\eta, \cdot)$. Let α_n be a minimizing sequence of $C_\phi(\eta, \cdot)$. Over the extended real numbers $\overline{\mathbb{R}}$, α_n has a subsequence α_{n_j} that converges to a limit point a, which must be a minimizer of $C_\phi(\eta, \cdot)$. Now if $a \neq 0$, both $\mathbf{1}_{\alpha \leq 0}, \mathbf{1}_{\alpha > 0}$ are continuous at a so that one can conclude that α_{n_j} also minimizes $C(\eta, \cdot)$. If in fact $a = 0$, Lemma 166 implies that $\eta = 1/2$ and thus *any* α minimizes $C(1/2, \cdot)$. Thus Corollary 169 implies that α_n minimizes $C(\eta, \cdot)$.

Showing 3) is equivalent to 4)

To show that 4) implies 3), consider a distribution for which $\eta(\mathbf{x}) \equiv \eta$ is constant.

For the other direction, let f^* be a minimizer of R_ϕ. Then $C_\phi(\eta(\mathbf{x}), f^*(\mathbf{x})) \geq C_\phi^*(\eta(\mathbf{x}))$ but $R_\phi(f^*) = \int C_\phi^*(\eta) d\mathbb{P}$ by Lemma 165. Therefore $C_\phi(\eta(\mathbf{x}), f^*(\mathbf{x})) = C_\phi^*(\eta(\mathbf{x}))$ $\mathbb{P}$-a.e. Item 3) then implies the result. □

C.2 Minimizing R_ϕ^ϵ over real valued functions

In this appendix, we will show

Lemma 171. *Let R_ϕ^ϵ be defined as in (4.7). Then*

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

Integrals of functions assuming values in $\mathbb{R} \cup \{\infty\}$ can still be defined using standard measure theory, see for instance [22].

Recall that [25] originally proved their minimax result for $\overline{\mathbb{R}}$-valued functions and thus this lemma is essential for the statement of Theorem 78.

Proof of Lemma 171. Let f be an $\overline{\mathbb{R}}$-valued function for with $R_\phi^\epsilon(f) < \infty$. We will show that the truncation $f_N = \min(\max(f, -N), N)$ satisfies $\lim_{N\to\infty} R_\phi^\epsilon(f_N) = R_\phi^\epsilon(f)$. Lemma 171 then follows from this statement.

Define a function $\sigma_{[a,b]} \colon \overline{\mathbb{R}} \to [a, b]$ by

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

Notice that $\sigma_{[a,b]}(-\alpha) = -\sigma_{[-b,-a]}(\alpha)$. Thus if $a = -b$, then $\sigma_{[a,b]}$ is anti-symmetric. Furthermore, because ϕ is continuous and non-increasing, for any function g,

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

and as $\sigma_{[a,b]}(\alpha)$ is continuous and non-decreasing,

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

Now let $f_N = \sigma_{[-N,N]}(f)$. Then $S_\epsilon(\phi \circ f_N)$, $S_\epsilon(\phi \circ -f_N)$ satisfy

$$S_\epsilon(\phi(f_N)) = \sigma_{[\phi(N),\phi(-N)]}(S_\epsilon(\phi \circ f)), \quad S_\epsilon(\phi(-f_N)) = \sigma_{[\phi(N),\phi(-N)]}(S_\epsilon(\phi \circ -f))$$

Therefore, $S_\epsilon(\phi \circ f_N)$,$S_\epsilon(\phi \circ -f_N)$ converge pointwise to $S_\epsilon(\phi \circ f)$, $S_\epsilon(\phi \circ -f)$. Furthermore, for $N \geq 1$, $\phi(f_N) \leq \phi(f) + \phi(1)$ which is integrable with respect to $\mathbb{P}_1$. Similarly, $\phi(-f_N) \leq \phi(-f) + \phi(1)$ which is integrable with respect to $\mathbb{P}_0$. Therefore, the dominated convergence theorem implies that

$$\lim_{N \to \infty} R^\epsilon_\phi(f_N) = R^\epsilon_\phi(f)$$

□

C.3 Further Properties of Adversarially Consistent Losses– Proofs of Lemma 75, Lemma 81, and Proposition 74

Recall the condition $C_\phi^*(1/2) < \phi(0)$ implies that minimizers of $C_\phi(1/2, \alpha)$ are bounded away from zero. Lemma 172 states that this property actually holds for *all* η. To prove this fact, we decompose $C_\phi(\eta, \alpha)$ into $C_\phi(1/2, \alpha)$ and a monotonic function:

$$C_\phi(\eta, \alpha) = \eta\phi(\alpha) + (1-\eta)\phi(-\alpha) = (\eta - 1/2)(\phi(\alpha) - \phi(-\alpha)) + \frac{1}{2}(\phi(\alpha) + \phi(-\alpha)). \quad \text{(C.1)}$$

Lemma 172. *Assume that $C_\phi^*(1/2) < \phi(0)$. Then there exists an $a > 0$ for which $|\alpha| < a$ implies $C_\phi(\eta, \alpha) \neq C_\phi^*(\eta)$ for all η. This a satisfies $\phi(a) < \phi(0)$.*

Proof. Let S be the set of non-negative minimizers of $C_\phi(1/2, \cdot)$ and define $a = \inf S$. Because ϕ is continuous, a is also a minimizer of $C_\phi(1/2, \cdot)$ and thus $C_\phi(1/2, a) = C_\phi^*(1/2) < \phi(0) = C_\phi(1/2, 0)$. Therefore, $\phi(a) < \phi(0)$ follows from the fact that $\phi(-a) \geq \phi(0)$.

We will now show that $C_\phi(\eta, \cdot)$ does not achieve its optimum on $(-a, a)$ for any η. First, this statement holds for $\eta = 1/2$ due to the definition of a. Next, we will assume that $\eta > 1/2$, the case $\eta < 1/2$ is analogous. To start, we can decompose the quantity $C_\phi(\eta, \alpha)$ as in (C.1). Subsequently, because a is the smallest positive minimizer of $C_\phi(1/2, \cdot)$, $1/2(\phi(\alpha) + \phi(-\alpha))$ assumes its infimum over $[-a, a]$ only at $-a$ and a. Next, notice that $\phi(\alpha) - \phi(-\alpha)$ is non-increasing on $[-a, a]$. Furthermore, because $\phi(a) < \phi(0)$, one can conclude that $\phi(-a) - \phi(a) > 0 > \phi(a) - \phi(-a)$, and thus the function $\alpha \mapsto \phi(\alpha) - \phi(-\alpha)$ is non-constant on $[-a, a]$. Therefore, (C.1) achieves its optimum over $[-a, a]$ only at $\alpha = a$. Thus, any $\alpha \in (-a, a)$ cannot be a minimizer of $C_\phi(\eta, \cdot)$ because $C_\phi(\eta, \alpha) > C_\phi(\eta, a) \geq C_\phi^*(\eta)$.

□

Proof of Lemma 75. Lemma 172 (above) immediately implies the forward direction.

For the backwards direction, note that if there is an a for which $|\alpha^*| \geq a$ for any minimizer $C_\phi(\eta, \cdot)$ for all η, then 0 does not minimize $C_\phi(1/2, \cdot)$. Therefore $C_\phi^*(1/2) < C_\phi(1/2, 0) = \phi(0)$.

□

Proof of Proposition 74. We will argue that for each η, every minimizer of $C_\phi(\eta, \cdot)$ over $\overline{\mathbb{R}}$ is also a minimizer of $C(\eta, \cdot)$. Proposition 71 will then imply that ϕ is consistent. To start, notice that *every* α is a minimizer of $C(1/2, \cdot)$. Next, we will show that for $\eta > 1/2$, every minimizer of $C_\phi(\eta, \cdot)$ is also a minimizer of $C(\eta, \cdot)$. The argument for $\eta < 1/2$ is analogous.

Consider the decomposition of $C_\phi(\eta, \alpha)$ in (C.1). Let a be as in Lemma 172 and notice that if $\alpha > a$ then $\phi(\alpha) < \phi(-\alpha)$. Hence as $\eta > 1/2$, then $C_\phi(\eta, \alpha) < C_\phi(\eta, -\alpha)$. Furthermore, Lemma 172 implies that there is no minimizer to $C_\phi(\eta, \cdot)$ in $(-a, a)$ and thus every minimizer to $C_\phi(\eta, \cdot)$ must be strictly positive. Therefore, every minimizer of $C_\phi(\eta, \cdot)$ also minimizes $C(\eta, \cdot)$.

□

Next, Lemma 81 is a quantitative version of Lemma 172.

Proof of Lemma 81. Let a be as in Lemma 172 and define ϕ^- by

$$\phi^-(y) = \sup\{\alpha : \phi(\alpha) \geq y\}.$$

The function ϕ^- is the right inverse of ϕ— this function satisfies $\phi(\phi^-(y)) = y$ while $\phi^-(\phi(\alpha)) \geq \alpha$.

Set $k = 1/2(\phi(0) + \phi(a))$, $c = \phi^-(k) = \sup\{\alpha \colon \phi(\alpha) \geq k\}$. From the definition of c, one can conclude that $\alpha > c$ implies that $\phi(\alpha) < \phi(c)$.

Because $\phi(a) < k = \phi(c) < \phi(0)$ and ϕ is non-increasing, $0 < c < a$. Thus $[-c, c] \subset (-a, a)$ and Lemma 172 implies that for all $\alpha \in [-c, c]$ and $\eta \in [0, 1]$, $C_\phi(\eta, \alpha) - C_\phi^*(\eta) > 0$. As this expression is jointly continuous in the variables η, α and $[-c, c] \times [0, 1]$ is compact, one can define

$$\delta = \inf_{\substack{\alpha \in [-c,c] \\ \eta \in [0,1]}} C_\phi(\eta, \alpha) - C_\phi^*(\eta)$$

and then it holds that $\delta > 0$ and $C_\phi(\eta, \alpha) \geq C_\phi^*(\eta) + \delta$ for all $\alpha \in [-c, c]$.

□

C.4 Optimal Transport Facts— Proof of Lemma 76

Proof of Lemma 76. Let $\mathbb{Q}'$ be any measure with $W_\infty(\mathbb{Q}', \mathbb{Q}) \leq \epsilon$. Let γ be a coupling with marginals $\mathbb{Q}$ and $\mathbb{Q}'$ for which $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x} - \mathbf{y}\| \leq \epsilon$. Such a coupling exists by Theorem 2.6 of [33]. This measure γ is supported on $\Delta_\epsilon = \{(\mathbf{x}, \mathbf{y}) \colon \|\mathbf{x} - \mathbf{y}\| \leq \epsilon\}$. Then

$$\begin{aligned}
&\int g d\mathbb{Q}' = \int g(\mathbf{x}') d\gamma(\mathbf{x}, \mathbf{x}') = \int g(\mathbf{x}') \mathbf{1}_{\|\mathbf{x}' - \mathbf{x}\| \leq \epsilon} d\gamma(\mathbf{x}, \mathbf{x}') \\
&\leq \int S_\epsilon(g)(\mathbf{x}) \mathbf{1}_{\|\mathbf{x}' - \mathbf{x}\| \leq \epsilon} d\gamma(\mathbf{x}, \mathbf{x}') = \int S_\epsilon(g)(\mathbf{x}) d\gamma(\mathbf{x}, \mathbf{x}') = \int S_\epsilon(g) d\mathbb{Q}
\end{aligned}$$

□

C.5 Proof of Theorem 77

As observed in Section 4.5, the ρ-margin loss satisfies $R^\epsilon_{\phi_\rho}(f) \geq R^\epsilon(f)$ while $C^*_{\phi_\rho}(\eta) = C^*(\eta)$. Theorem 78 then implies that

$$\sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}_{\phi_\rho}(\mathbb{P}_0', \mathbb{P}_1') = \inf_f R^\epsilon_{\phi_\rho}(f) \geq \inf_f R^\epsilon(f)$$

The opposite inequality follows from swapping an inf and a sup— a form of weak duality. We prove this weak duality for $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$-valued functions in order to later apply a result from [25] which is also stated for $\overline{\mathbb{R}}$-valued functions.

Lemma 173 (Weak Duality)**.** *Let R^ϵ be the adversarial classification loss. Then*

$$\inf_{\substack{f \text{ Borel,} \\ f \ \overline{\mathbb{R}}\text{-valued}}} R^\epsilon(f) \geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') \tag{C.2}$$

Proof. Notice that Lemma 76 implies that for any function g,

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

Applying this inequality to the functions $\mathbf{1}_{f \leq 0}, \mathbf{1}_{f>0}$ in the expression for $R^\epsilon(f)$ results in

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

Thus by swapping the inf and the sup and defining $\mathbb{P}' = \mathbb{P}_0' + \mathbb{P}_1'$, $\eta' = d\mathbb{P}_1'/d\mathbb{P}'$,

$$\begin{aligned}
&\inf_{\substack{f \text{ Borel} \\ f \ \overline{\mathbb{R}}\text{-valued}}} \int S_\epsilon(\mathbf{1}_{f\leq 0}) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_{f>0}) d\mathbb{P}_0 \geq \inf_{\substack{f \text{ Borel} \\ f \ \overline{\mathbb{R}}\text{-valued}}} \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \int \mathbf{1}_{f\leq 0} d\mathbb{P}_1' + \int \mathbf{1}_{f>0} d\mathbb{P}_0' \\
&\geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{\substack{f \text{ Borel} \\ f \ \overline{\mathbb{R}}\text{-valued}}} \int \mathbf{1}_{f\leq 0} d\mathbb{P}_1' + \int \mathbf{1}_{f>0} d\mathbb{P}_0' \\
&= \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{\substack{f \text{ Borel} \\ f \ \overline{\mathbb{R}}\text{-valued}}} \int C(\eta', f) d\mathbb{P}' \geq \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \int C^*(\eta') d\mathbb{P}' = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1')
\end{aligned}$$

□

Strong duality and existence of maximizers/minimizers then follows from weak duality.

Proof of Theorem 77. Let $\phi_\rho(\alpha)$ be the ϕ-margin loss $\phi_\rho = \min(1, \max(1 - \alpha/\rho, 0))$. Then as

discussed in Section 4.5, one can bound the adversarial classification risk $R^\epsilon(f)$ by $R^\epsilon(f) \leq R^\epsilon_{\phi_\rho}(f)$ but $C^*_{\phi_\rho}(\eta) = C^*(\eta)$ and thus $\bar{R}_{\phi_\rho} = \bar{R}$.

The minimax theorem for surrogate losses in [25] (Theorem 6) states that there is an $\overline{\mathbb{R}}$-valued function f^*, and measures $\mathbb{P}_0^*, \mathbb{P}_1^*$ for which $R^\epsilon_{\phi_\rho}(f^*) = \bar{R}_{\phi_\rho}(\mathbb{P}_0^*, \mathbb{P}_1^*)$. Thus weak duality (Lemma 173) implies

$$\bar{R}_{\phi_\rho}(\mathbb{P}_0^*, \mathbb{P}_1^*) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*) \leq R^\epsilon(f^*) \leq R^\epsilon_{\phi_\rho}(f^*).$$

However, the fact that $R^\epsilon_{\phi_\rho}(f^*) = \bar{R}_{\phi_\rho}(\mathbb{P}_0^*, \mathbb{P}_1^*)$ implies that the inequalities above must actually be equalities. This relation proves strong duality for the adversarial classification risk (Equation 4.9) and that f^* minimizes R^ϵ and $(\mathbb{P}_0^*, \mathbb{P}_1^*)$ maximizes $\bar{R}$ over $\mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \times \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$.

Next, let $\hat{f} = \min(1, \max(\hat{f}, -1))$. Then $\hat{f}$ is $\mathbb{R}$-valued and $R^\epsilon(\hat{f}) = R^\epsilon(f^*)$. Thus $\hat{f}$ is an $\mathbb{R}$-valued minimizer of R^ϵ. □

C.6 Proof of Lemma 82

Proof of Lemma 82. Lemma 76 implies that for each n,

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

Therefore, writing $\ell_n = \int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1$ and $r_n = \int \mathbf{1}_{f_n \leq 0} d\mathbb{P}_1^*$, we have

$$\liminf_{n\to\infty} r_n \leq \liminf_{n\to\infty} \ell_n \leq \limsup_{n\to\infty} \ell_n . \tag{C.3}$$

Therefore, (4.20) implies both that that the limit $\lim_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1$ exists and that

$$\lim_{n\to\infty} \int S_\epsilon(\mathbf{1}_{f_n \leq 0}) d\mathbb{P}_1 = \liminf_{n\to\infty} \int \mathbf{1}_{f_n \leq 0} d\mathbb{P}_1^* \tag{C.4}$$

Similarly, because $\limsup_{n\to\infty} \ell_n \geq \limsup_{n\to\infty} r_n \geq \liminf_{n\to\infty} r_n$, the relation (4.20) implies that the limit $\lim_{n\to\infty} \int \mathbf{1}_{f_n\leq 0} d\mathbb{P}_1^*$ exists. The first relation of (4.18) then follows from (C.4) and the existence of the limit of $\int \mathbf{1}_{f_n\leq 0} d\mathbb{P}_1^*$.

An analogous argument shows that (4.21) implies the second relation of (4.18). □

D — Deferred Proofs from Chapter 5

D.1 Proof of Lemma 85

Lemma 174. *The smallest minimizer of $C_\phi(\eta, \cdot)$ is well-defined.*

Proof. First, define

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

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

The next result implies that α_ϕ is non-decreasing.

Lemma 175. *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 $C_\phi(\eta_2, \alpha) >$

$C_\phi(\eta_2, \alpha_\phi(\eta_1))$. Thus every minimizer of $C_\phi(\eta_2, \cdot)$ must be greater than or equal to $\alpha_\phi(\eta_1)$. □

Proof of Lemma 85. Lemma 174 proves that α_ϕ is well-defined. For $\eta_2 > \eta_1$, Lemma 175 with the choice $\alpha_2^* = \alpha_\phi(\eta_2)$ proves that the function α_ϕ is non-decreasing.

□

D.2 Proof of Lemma 92

Proof of Lemma 92. 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.

□

D.3 Proof of Theorem 87

We prove that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ minimize R_ϕ^ϵ by showing that $R_\phi^\epsilon(\{\hat{\eta} > 1/2) = \bar{R}(\mathbb{P}_0^*, \mathbb{P}_1^*)$ for the measures $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ in Theorem 95.

Proposition 176. *Let $\hat{\eta}$ be the function in Theorem 86. Then the sets $\{\hat{\eta} > 1/2\}$, $\hat{\eta} \geq 1/2\}$ are both Bayes classifiers.*

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

Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be the measures of Theorem 95 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 II) implies that the function $\hat{\eta}(\mathbf{x})$ assumes its infimum on an ball $\overline{B_\epsilon(\mathbf{x})}$ γ_1^*-a.e. and therefore $S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2\}^C})(\mathbf{x}) = \mathbf{1}_{\{I_\epsilon(\hat{\eta})(\mathbf{x})>1/2\}^C}$ γ_1^*-a.e. (Recall the notation I_ϵ was defined in Equation (5.11).) Item II) further implies that $\mathbf{1}_{\{I_\epsilon(\hat{\eta})(\mathbf{x})>1/2\}^C} = \mathbf{1}_{\{\hat{\eta}(\mathbf{x}')>1/2\}^C}$ γ_1^*-a.e. and consequently,

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

An analogous argument shows

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

Equations (D.1) and (D.2) then imply that

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

Next Item I) of Theorem 95 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 93 implies that $\{\hat{\eta} > 1/2\}$ must minimize R^ϵ. □

Finally, the complementary slackness conditions from [23, Theorem 2.4] characterize minimizers of R^ϵ and maximizers of $\bar{R}$, and this characterization proves Equations (5.9) and (5.10). Verifying these conditions would be another method of proving Proposition 176.

Theorem 177. *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)

$$\int S_\epsilon(\mathbf{1}_{A^C})d\mathbb{P}_1 = \int \mathbf{1}_{A^C} d\mathbb{P}_1^* \quad \textit{and} \quad \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 = \int \mathbf{1}_A d\mathbb{P}_0^* \tag{D.3}$$

2)

$$C(\eta^*, \mathbf{1}_A(\mathbf{x}')) = C^*(\eta^*(\mathbf{x}')) \quad \mathbb{P}^*\textit{-a.e.} \tag{D.4}$$

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 92 implies that $\mathbf{1}_{A^C}(\mathbf{x}') \leq S_\epsilon(\mathbf{1}_{A^C})(\mathbf{x})$ γ_1^*-a.e. and $\mathbf{1}_A(\mathbf{x}') \leq S_\epsilon(\mathbf{1}_A)(\mathbf{x})$, the complementary slackness condition in Equation (D.3) is equivalent to

$$S_\epsilon(\mathbf{1}_{A^C})(\mathbf{x}) = \mathbf{1}_{A^C}(\mathbf{x}') \quad \gamma_1^*\text{-a.e.} \quad \text{and} \quad S_\epsilon(\mathbf{1}_A)(\mathbf{x}) = \mathbf{1}_A(\mathbf{x}') \quad \gamma_0^*\text{-a.e.} \tag{D.5}$$

This observation completes the proof of Theorem 87.

Proof of Theorem 87. First, Proposition 176 proves that the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are in fact adversarial Bayes classifiers.

Next, let $\hat{\eta}$, $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ be the function and measures of Theorem 95. Let $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$, $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$, and 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. If A is any adversarial Bayes classifier, the complementary slackness condition Equation (D.4) implies that $\mathbf{1}_{\eta^*>1/2} \leq \mathbf{1}_A \leq \mathbf{1}_{\eta^*\geq 1/2}$ $\mathbb{P}^*$-a.e. Thus Item I) 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 Equation (D.5) then implies Equations (5.9) and (5.10).

□

D.4 Proof of Theorem 97

Theorem 3.4 of [23] proves the following result:

Theorem 178. *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 178 is equivalent to Item B) of Theorem 97. We will apply the complementary slackness conditions of Theorem 177.

Proof of Theorem 97. Let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ be the measures of Theorem 95.

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

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

where R_*^ϵ 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.

Consequently, Item 2) of Theorem 178 and the fact that $\{\hat{\eta} > 1/2\} \subset \{\hat{\eta} \geq 1/2\}$ imply that

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

The complementary slackness condition Equation (D.3) implies that

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

and subsequently, Item I) of Theorem 95 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 87. The complimentary slackness conditions in Theorem 177 and the first inequality in Theorem 87 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. □

D.5 Proof of Lemma 99

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

Lemma 179. *Let ϕ be a consistent loss. Then if $0 \in \operatorname{argmin} C_\phi(\eta, \cdot)$, then $\eta = 1/2$.*

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

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

□

The proof of Lemma 99 also uses Lemma 175 from Appendix D.1.

Proof of Lemma 99. Notice that $C_\phi(\eta, \alpha) = C_\phi(1-\eta, -\alpha)$ and thus it suffices to consider $\eta \geq 1/2 + r$.

Lemma 179 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 175, 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 $\alpha_r > 0$. □

D.6 Proof of Proposition 101

First, we show that replacing the value of $\alpha_\phi(1/2)$ with 0 in Theorem 86 results in a minimizer of R_ϕ^ϵ.

Lemma 180. *Let $\alpha_\phi : [0, 1] \to \mathbb{R}$ be as in Lemma 85 and define a function $\tilde{\alpha}_\phi : [0, 1] \to \overline{\mathbb{R}}$ by*

$$\tilde{\alpha}_\phi(\eta) = \begin{cases} \alpha_\phi(\eta) & \text{if } \eta \neq 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{D.6}$$

Let $\hat{\eta} : \mathbb{R}^d \to [0,1]$ be the function described in Theorem 95. If ϕ is consistent and $C^*_\phi(1/2) = \phi(0)$, then $\tilde{\alpha}(\hat{\eta}(\mathbf{x}))$ is a minimizer of R^ϵ_ϕ.

See Appendix D.6.1 for a proof of this result. Next, we formally prove that if the adversarial Bayes classifier is not unique up to degeneracy, then the sets $\{\hat{\eta} > 1/2\}$ and $\{\hat{\eta} \geq 1/2\}$ are not equivalent up to degeneracy.

This result in Lemma 102 relies on a characterization of equivalence up to degeneracy from [23].

Theorem 181. *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 $S_\epsilon(\mathbf{1}_{A_2^C}) = S_\epsilon(\mathbf{1}_{A_1^C})$-$\mathbb{P}_1$-a.e.

Notice that when there is a single equivalence class, the equivalence between Item 1) and Item 2) is simply the equivalence between Item 1) and Item 2) in Theorem 178. This result together with Theorem 87 proves Lemma 102:

Proof of Lemma 102. 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 87 and Item 2) of Theorem 181 imply that $S_\epsilon(\mathbf{1}_A) = S_\epsilon(\mathbf{1}_{\{\hat{\eta}>1/2\}})$ $\mathbb{P}_0$-a.e. Item 2) of Theorem 181 again implies that A and $\{\hat{\eta} > 1/2\}$ must be equivalent up to degeneracy. □

Thus, if the adversarial Bayes classifier is not unique up to degeneracy, then there is a set $\tilde{A}$ with $\{\hat{\eta} > 1/2\} \subset \tilde{A} \subset \{\hat{\eta} \geq 1/2\}$ that is not an adversarial Bayes classifier, and this set is used to construct the sequence f_n in Equation (5.23). Next, we show that f_n minimizes R^ϵ_ϕ but not R^ϵ.

Proposition 182. *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 $\overline{\mathbb{R}}$-valued functions that minimize R_ϕ^ϵ but $R^\epsilon(f_n)$ is constant in n and not equal to the adversarial Bayes risk.

Proof. By Lemma 102, 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 Equation (5.23) and let $\tilde{\alpha}_\phi$ be the function in Lemma 180. Lemma 99 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 $n \to \infty$, the function sequence $S_\epsilon(\phi \circ f_n)$ is bounded above by $S_\epsilon(\phi \circ \tilde{\alpha}_\phi(\hat{\eta}))$ while $S_\epsilon(\phi \circ -f_n)$ is bounded above by $S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta}))$. This result will imply that f_n is a minimizing sequence of R_ϕ^ϵ.

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

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

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

$$S_\epsilon(\phi \circ \tilde{\alpha}_\phi(\hat{\eta})) = \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})), \phi(0)) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \tag{D.7}$$

$$S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta})) = \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})), \phi(0)) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \tag{D.8}$$

and similarly

$$S_\epsilon(\phi \circ f_n) \leq \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})), \phi(-\frac{1}{n})) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \tag{D.9}$$

$$S_\epsilon(\phi \circ -f_n) \leq \begin{cases} \max(\tilde{S}_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})), \phi(-\frac{1}{n})) & \mathbf{x} \in \{\hat{\eta} = 1/2\}^\epsilon \\ S_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})) & \mathbf{x} \notin \{\hat{\eta} = 1/2\}^\epsilon \end{cases} \tag{D.10}$$

Therefore, by comparing Equation (D.9) with Equation (D.7) and Equation (D.10) with Equation (D.8), one can conclude that

$$\limsup_{n\to\infty} S_\epsilon(\phi \circ f_n) \leq S_\epsilon(\phi \circ \tilde{\alpha}_\phi(\hat{\eta})) \quad \text{and} \quad \limsup_{n\to\infty} S_\epsilon(\phi \circ -f_n) \leq S_\epsilon(\phi \circ -\tilde{\alpha}_\phi(\hat{\eta})). \tag{D.11}$$

Furthermore, Equation (D.9) implies that $S_\epsilon(\phi \circ f_n) \leq S_\epsilon(\phi \circ \alpha_\phi(\hat{\eta})) + \phi(-1)$ and Equation (D.10) implies that $S_\epsilon(\phi \circ -f_n) \leq S_\epsilon(\phi \circ -\alpha_\phi(\hat{\eta})) + \phi(-1)$. Thus the dominated convergence theorem and Equation (D.11) implies that

$$\limsup_{n\to\infty} R^\epsilon_\phi(f_n) \leq R^\epsilon_\phi(\tilde{\alpha}_\phi(\hat{\eta}))$$

and thus f_n minimizes R^ϵ_ϕ.

□

Lastly, it remains to construct an $\mathbb{R}$-valued sequence that minimizes R^ϵ_ϕ but not R^ϵ. To construct this sequence, we threshhold 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 threshholded at N, then $\lim_{N\to\infty} R^\epsilon_\phi(g^{(N)}) = R^\epsilon_\phi(g)$.

Lemma 183. *Let g be an $\overline{\mathbb{R}}$-valued function and let $g^{(N)} = \min(\max(g, -N), N)$. Then $\lim_{N\to\infty} R^\epsilon_\phi(g^{(N)}) = R^\epsilon_\phi(g)$.*

See Appendix D.6.2 for a proof. Proposition 101 then follows from this lemma and Proposition 182:

Proof of Proposition 101. Let f_n be the $\overline{\mathbb{R}}$-valued sequence of functions in Proposition 182, and let f_{n_j} be a subsequence for which $R^\epsilon_\phi(f_{n_j}) - \inf_f R^\epsilon_\phi(f) < 1/j$. Next, Lemma 183

implies that for each j one can pick a threshhold N_j for which $|R^\epsilon_\phi(f_{n_j}) - R^\epsilon_\phi(f^{(N_j)}_{n_j})| \leq 1/j$. Consequently, $f^{(N_j)}_{n_j}$ is an $\mathbb{R}$-valued sequence of functions that minimizes R^ϵ_ϕ. However, notice that $\{f \leq 0\} = \{f^{(T)} \leq 0\}$ and $\{f > 0\} = \{f^{(T)} > 0\}$ for any strictly positive threshhold T. Thus $R^\epsilon(f^{(N_j)}_{n_j}) = R^\epsilon(f_{n_j})$ and consequently $f^{(N_j)}_{n_j}$ does not minimize R^ϵ. □

D.6.1 Proof of Lemma 180

The proof of Lemma 180 follows the same outline as the argument for Proposition 176: we show that $R^\epsilon_\phi(\tilde{\alpha}_\phi(\hat{\eta})) = \bar{R}_\phi(\mathbb{P}^*_0, \mathbb{P}^*_1)$ for the measures $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ in Theorem 95, and then Theorem 94 implies that $\tilde{\alpha}_\phi(\hat{\eta})$ must minimize R^ϵ_ϕ. Similar to the proof of Proposition 176, 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 184. *If $C^*_\phi(1/2) = \phi(0)$, then the function $\tilde{\alpha}_\phi : [0,1] \to \overline{\mathbb{R}}$ defined in Equation (D.6) is non-decreasing and maps each η to a minimizer of $C_\phi(\eta, \cdot)$.*

Proof. Lemma 85 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 85 implies that $\tilde{\alpha}_\phi$ is non-decreasing on $[0, 1/2)$ and $(1/2, 1]$. However, Lemma 99 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 $\mathbb{P}^*_0$, $\mathbb{P}^*_1$ suffice to prove Lemma 180.

Proof of Lemma 180. Let $\mathbb{P}^*_0, \mathbb{P}^*_1$ be the measures of Theorem 95 and set $\mathbb{P}^* = \mathbb{P}^*_0 + \mathbb{P}^*_1$, $\eta^* = d\mathbb{P}^*_1/d\mathbb{P}^*$. We will prove that $R^\epsilon_\phi(\tilde{\alpha}_\phi(\hat{\eta})) = \bar{R}_\phi(\mathbb{P}^*_0, \mathbb{P}^*_1)$ and thus Theorem 94 will imply that $\tilde{\alpha}_\phi(\hat{\eta})$ minimizes R^ϵ_ϕ. Let γ^*_0 and γ^*_1 be the couplings supported on Δ_ϵ between $\mathbb{P}_0$, $\mathbb{P}^*_0$ and $\mathbb{P}_1$, $\mathbb{P}^*_1$ respectively. Item II) of Theorem 95 and Lemma 184 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 the notation I_ϵ was introduced in Equation (5.11).) Therefore,

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

Next, Item I) of Theorem 95 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 94 implies that $\tilde{\alpha}_\phi(\hat{\eta})$ must minimize R_ϕ^ϵ.

□

D.6.2 Proof of Lemma 183

This argument is taken from the proof of Lemma 8 in [26].

Proof of Lemma 183. Define

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

Notice that

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

and

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

for any functions g and h. Therefore,

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

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

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

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

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

□

Bibliography

[1] Quinn Aiken et al. *A primal-dual method for tracking topological changes in optimal adversarial classification*. 2022.

[2] Pranjal Awasthi, Natalie S. Frank, and Mehryar Mohri. "On the Existence of the Adversarial Bayes Classifier (Extended Version)". In: *arxiv* (2023).

[3] Pranjal Awasthi et al. "A Finer Calibration Analysis for Adversarial Robustness". In: *arxiv* (2021).

[4] Pranjal Awasthi et al. "Calibration and Consistency of Adversarial Surrogate Losses". In: *NeurIps* (2021).

[5] Pranjal Awasthi et al. "H-Consistency Bounds for Surrogate Loss Minimizers". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Proceedings of Machine Learning Research. PMLR, 2022.

[6] Han Bao, Clayton Scott, and Masashi Sugiyama. "Calibrated Surrogate Losses for Adversarially Robust Classification". In: *arxiv* (2021).

[7] Viorel Barbu and Teodor Precupanu. *Convexity and Optimization in Banach Spaces*. 4th. Springer Monographs in Mathematics, 2012.

[8] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. "Convexity, Classification, and Risk Bounds". In: *Journal of the American Statistical Association* 101.473 (2006).

[9] Shai Ben-David, Nadav Eiron, and Philip M. Long. "On the Difficulty of Approximately Maximizing Agreements". In: *Journal of Computer System Sciences* (2003).

[10] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.

[11] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. "Lower bounds on adversarial robustness from optimal transport". In: *Advances in Neural Information Processing Systems*. 2019, pp. 7498–7510.

[12] Robi Bhattacharjee and Kamalika Chaudhuri. "Consistent Non-Parametric Methods for Maximizing Robustness". In: *NeurIps* (2021).

[13] Robi Bhattacharjee and Kamalika Chaudhuri. "When are Non-Parametric Methods Robust?" In: *PMLR* (2020).

[14] Battista Biggio et al. “Evasion attacks against machine learning at test time”. In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2013, pp. 387–402.

[15] Vladimir I. Bogachev. *Measure Theory*. Vol. II. Springer, 2007.

[16] Leon Bungert, Nicolás García Trillos, and Ryan Murray. “The Geometry of Adversarial Training in Binary Classification”. In: *arxiv* (2021).

[17] Thierry Champion, Luigi De Pascale, and Petri Juutinen. “The ∞-Wasserstein Distance: Local Solutions and Existence of Optimal Transport Maps”. In: *SIAM Journal on Mathematical Analysis* 40.1 (2008), pp. 1–20. DOI: 10.1137/07069938X.

[18] Ambra Demontis et al. “Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks”. In: *CoRR* (2018).

[19] Yao Deng et al. *An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models*. 2020.

[20] Carles Domingo-Enrich et al. *A mean-field analysis of two-player zero-sum games*. 2021.

[21] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021.

[22] Gerald B Folland. *Real analysis: modern techniques and their applications*. Vol. 40. John Wiley & Sons, 1999.

[23] Natalie S. Frank. “A Notion of Uniqueness for the Adversarial Bayes Classifier in Binary Classification”. In: *arxiv* (2024). The results in this reference are largely the same as those presented in Chapter 3 and Appendix B of this thesis.

[24] Natalie S. Frank. “The Uniqueness of The Adversarial Bayes classifier and the Adversarial Consistency of Surrogate Risks for Binary Classification”. In: *arxiv* (2024). The results in this reference are largely the same as those presented in Chapter 5 and Appendix D of this thesis.

[25] Natalie S. Frank and Jonathan Niles-Weed. “Existence and Minimax Theorems for Adversarial Surrogate Risks in Binary Classification”. In: *JMLR* (2023). The results in this reference are largely the same as those presented in Chapter 2 and Appendix A of this thesis.

[26] Natalie S. Frank and Jonathan Niles-Weed. “The Adversarial Consistency of Surrogate Risks for Binary Classification”. In: *NeurIps* (2023). The results in this reference are largely the same as those presented in Chapter 4 and Appendix C of this thesis.

[27] Rui Gao, Xi Chen, and Anton J. Kleywegt. *Wasserstein Distributionally Robust Optimization and Variation Regularization*. 2022.

[28] Lucas Gnecco-Heredia et al. *On the Role of Randomization in Adversarially Robust Classification*. 2023.

[29] Lucas Gnecco-Heredia et al. *On the Role of Randomization in Adversarially Robust Classification*. 2023.

[30] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *ICLR* (2014).

[31] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain". In: *CoRR* (2017).

[32] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[33] Heikki Jylhä. "The L^∞ Optimal Transport: Infinite Cyclical Monotonicity and the Existence of Optimal Transport Maps". In: *Calculus of Variations and Partial Differential Equations* (2014).

[34] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. "Adversarial Logit Pairing". In: *CoRR* (2018).

[35] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. "Adversarial Machine Learning at Scale". In: *ICLR* (2017).

[36] Justin D. Li and Matus Telgarsky. *On Achieving Optimal Adversarial Test Error*. 2023.

[37] Yujie Li et al. "Adaptive Square Attack: Fooling Autonomous Cars With Adversarial Traffic Signs". In: *IEEE Internet of Things Journal* 8.8 (2021).

[38] Yi Lin. "A note on margin-based loss functions in classification". In: *Statistics & Probability Letters* 68.1 (2004), pp. 73–82.

[39] Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *ICLR* (2019).

[40] Anqi Mao, Mehryar Mohri, and Yutao Zhong. *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*. 2023.

[41] Pascal Massart and Élodie Nédélec. "Risk bounds for statistical learning". In: *The Annals of Statistics* 34 (2006).

[42] Laurent Meunier et al. "Towards Consistency in Adversarial Classification". In: *arXiv* (2022).

[43] Shivani Agarwal Mingyuan Zhang. "Consistency vs. H-consistency: The Interplay between Surrogate Loss functions and the Scoring Function Class". In: *NeurIps* (2020).

[44] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. *A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach*. 2019.

[45] Togo Nishiura. *Absolute Measurable Spaces*. Cambridge University Press, 2010.

[46] Nicolas Papernot et al. "Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples". In: *CoRR* abs/1602.02697 (2016).

[47] Magdalini Paschali et al. "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging". In: *Springer* (2018).

[48] ShengYun Peng et al. *Robust Principles: Architectural Design Principles for Adversarially Robust CNNs*. 2023.

[49] Rocco A. Servedio Philip M. Long. "Consistency versus Realizable H-Consistency for MultiClass classification". In: *ICML* (2013).

[50] Muni Sreenivas Pydi and Varun Jog. "Adversarial risk via optimal transport and optimal couplings". In: *ICML* (2020).

[51] Muni Sreenivas Pydi and Varun Jog. "Adversarial risk via optimal transport and optimal couplings". In: *ICML* (2020).

[52] Muni Sreenivas Pydi and Varun Jog. "The Many Faces of Adversarial Risk". In: *Neural Information Processing Systems* (2021).

[53] 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: Association for Computing Machinery, 2009.

[54] Andras Rozsa, Manuel Günther, and Terrance E. Boult. "Are Accuracy and Robustness Correlated?" In: *CoRR* (2016).

[55] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. 1st. Birkhäuser, 2015.

[56] Ali Shafahi et al. "Adversarial Training for Free!" In: *CoRR* (2019).

[57] Ingo Steinwart. "How to Compare Different Loss Functions and Their Risks". In: *Constructive Approximation* (2007).

[58] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[59] Ambuj Tewari and Peter L. Bartlett. "On the Consistency of Multiclass Classification Methods". In: *Journal of Machine Learning Research* 8.36 (2007).

[60] Florian Tramèr et al. "The Space of Transferable Adversarial Examples". In: *arXiv* (2017).

[61] Camilo Garcia Trillos and Nicolas Garcia Trillos. *On adversarial robustness and the use of Wasserstein ascent-descent dynamics to enforce it*. 2023.

[62] Nicolas Garcia Trillos, Matt Jacobs, and Jakwang Kim. *On the existence of solutions to adversarial training in multiclass classification*. 2023.

[63] Nicolas Garcia Trillos, Matt Jacobs, and Jakwang Kim. "The Multimarginal Optimal Transport Formulation of Adversarial Multiclass Classification". In: *arXiv* (2022).

[64] Nicolas Garcia Trillos and Ryan Murray. "Adversarial Classification: Necessary conditions and geometric flows". In: *arxiv* (2022).

[65] Cédric Villani. *Topics in Optimal Transportation*. 2nd. American Mathematical Society, 2003.

[66] Guillaume Wang and Lénaïc Chizat. *An Exponentially Converging Particle Method for the Mixed Nash Equilibrium of Continuous Games*. 2023.

[67] Yisen Wang et al. "On the Convergence and Robustness of Adversarial Training". In: *ICML* (2021).

[68] Eric Wong, Leslie Rice, and J. Zico Kolter. "Fast is better than free: Revisiting adversarial training". In: *CoRR* abs/2001.03994 (2020).

[69] Eric Wong, Frank Schmidt, and Zico Kolter. "Wasserstein Adversarial Examples via Projected Sinkhorn Iterations". In: *Proceedings of the 36th International Conference on Machine Learning*. Proceedings of Machine Learning Research. PMLR, 2019.

[70] Kaiwen Wu, Allen Houze Wang, and Yaoliang Yu. *Stronger and Faster Wasserstein Adversarial Attacks*. 2020.

[71] Cihang Xie et al. "Feature Denoising for Improving Adversarial Robustness". In: *CoRR* (2018).

[72] Ying Xu et al. "Adversarial Attacks on Face Recognition Systems". In: *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Ed. by Christian Rathgeb et al. Cham: Springer International Publishing, 2022, pp. 139–161.

[73] Yao-Yuan Yang et al. "Robustness for Non-Parametric Classification: A Generic Attack and Defense". In: *Proceedings of Machine Learning Research* (2020).

[74] Hongyang Zhang et al. "Theoretically Principled Trade-off between Robustness and Accuracy". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. 2019, pp. 7472–7482.

[75] Tong Zhang. "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization". In: *The Annals of Statistics* (2004).