A Notion of Uniqueness for the Adversarial Bayes Classifier*

Natalie S. Frank†

Abstract. We propose a new notion of uniqueness for the adversarial Bayes classifier in the setting of binary classification. Analyzing this notion of uniqueness produces a simple procedure for computing all adversarial Bayes classifiers for a well-motivated family of one dimensional data distributions. This characterization is then leveraged to show that as the perturbation radius increases, certain notions of regularity improve for adversarial Bayes classifiers. We demonstrate with various examples that the boundary of the adversarial Bayes classifier frequently lies near the boundary of the Bayes classifier.

Key words. Robust Learning, Calculus of Variations, ∞-Wasserstein Metric

MSC codes. 62A995

1. Introduction.

A crucial reliability concern for machine learning models is their susceptibility to adversarial attacks. Neural nets are particularly susceptible to small perturbations to data. For instance, [5, 19] 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, [16] achieves 71% robust accuracy for ℓ_∞ perturbations size 8/255 while [7] 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 [12, 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, 4, 17] 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 classifi-

*Submitted to the editors DATE.

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

†Courant Institute, New York, NY (nf1066@nyu.edu, https://natalie-frank.github.io/).

cation risk. The Bayes classifier is unique of 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.

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 show 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. Prior work [23] discusses the tradeoff between robustness and standard accuracy, and such examples illustrate that this tradeoff could be mitigated by a careful selection of an adversarial Bayes classifier. Followup work [8] 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.

2. Background.

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 $\mathbb{R}^d$. [todo: is this clear enough] 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). Most of our results will assume that $\mathbb{P}_0$ and $\mathbb{P}_1$ are absolutely continuous with respect to the Lebesgue measure μ. 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:

$$R(A) = \int \mathbf{1}_{A^C} d\mathbb{P}_1 + \int \mathbf{1}_A d\mathbb{P}_0. \tag{2.1}$$

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 as

$$R(f) = \int C(\eta(\mathbf{x}), f) d\mathbb{P}(\mathbf{x}). \tag{2.2}$$

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

$$C(\eta, \alpha) = \eta \mathbf{1}_{\alpha \leq 0} + (1 - \eta)\mathbf{1}_{\alpha > 0}, \quad C^*(\eta) = \inf_\alpha C(\eta, \alpha). \tag{2.3}$$

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

$$p_1(\mathbf{x}) - p_0(\mathbf{x}) = 0 \tag{2.4}$$

A central goal of this paper is extending (2.4) 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]

$$R^\epsilon(A) = \int S_\epsilon(\mathbf{1}_{A^C}) d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0 \tag{2.5}$$

where the S_ϵ operation on a function g is defined as

$$S_\epsilon(g)(\mathbf{x}) = \sup_{\|\mathbf{h}\| \leq \epsilon} g(\mathbf{x} + \mathbf{h}). \tag{2.6}$$

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

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

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 (2.5), referred to as *adversarial Bayes classifiers* [2, 4, 9, 18], see [9, Theorem 1] for an existence theorem that matches the setup of this paper. Finding minimizers to (2.5) is difficult because unlike the standard classification problem, one cannot write the integrand of (2.5) so that it can be minimized in a pointwise manner. Furthermore, prior research [6] on the structure of minimizers to R^ϵ proves:

Lemma 2.1. *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 A 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 [21] . 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

[1] In order to define the adversarial classification risk, one must show that A^ϵ is measurable for measurable A. A full discussion of this issue is delayed to subsection 5.2.

$\{a_i\}$, $\{b_i\}$ 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:

$$R^\epsilon(A) = \cdots + \int_{b_{i-1}-\epsilon}^{a_i+\epsilon} p_1(x)dx + \int_{a_i-\epsilon}^{b_i+\epsilon} p_0(x)dx + \int_{b_i-\epsilon}^{a_{i+1}+\epsilon} p_1(x)dx + \cdots \tag{2.7}$$

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

$$p_1(a_i + \epsilon) - p_0(a_i - \epsilon) = 0 \tag{2.8a}$$

$$p_0(b_i + \epsilon) - p_1(b_i - \epsilon) = 0 \tag{2.8b}$$

When $\epsilon = 0$, these equations reduce to the condition describing the boundary of the Bayes classifer in (2.4). Prior work shows that when p_0, p_1 are well-behaved, this necessary condition holds for sufficiently small ϵ.

Theorem 2.2 ([21]). *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 (2.8).*

For a proof, see the discussion of Equation (4.1) and Theorem 5.4 in [21]. A central goal of this paper is producing necessary conditions analogous to (2.8) that hold for all ϵ.

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 (2.2). The Wasserstein-∞ metric from optimal transport and the minimax theorems in [9, 18] 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, see [22] for details. We denote the ϵ ball in the W_∞ metric around a measure $\mathbb{Q}$ by

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

Prior work [18, 20] 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

$$\bar{R}(\mathbb{P}_0', \mathbb{P}_1') = \int C^* \left(\frac{d\mathbb{P}_1'}{d(\mathbb{P}_0' + \mathbb{P}_1')} \right) d(\mathbb{P}_0' + \mathbb{P}_1') \tag{2.9}$$

where C^* is defined by (2.3). Theorem 1 of [9] relates this risk to R^ϵ.

Theorem 2.3. *Let $\bar{R}$ be defined by (2.9). Then*

$$\inf_{A \text{ Borel}} R^\epsilon(A) = \sup_{\substack{\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0) \\ \mathbb{P}_1' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\mathbb{P}_0', \mathbb{P}_1') \tag{2.10}$$

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

This minimax theorem then implies complimentary slackness conditions that characterize optimal A and $\mathbb{P}_0^*, \mathbb{P}_1^*$. See Appendix B for a proof.

Theorem 2.4. *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)

$$\int \mathbf{1}_{A^C} d\mathbb{P}_1^* = \int S_\epsilon(\mathbf{1}_{A^C}) d\mathbb{P}_1 \quad \text{and} \quad \int \mathbf{1}_A d\mathbb{P}_0^* = \int \mathbf{1}_A d\mathbb{P}_0 \tag{2.11}$$

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

$$\eta^*(\mathbf{y})\mathbf{1}_{A^C} + (1 - \eta^*(\mathbf{y}))\mathbf{1}_A = C^*(\eta^*(\mathbf{y})) \quad \mathbb{P}^*\text{-a.e.} \tag{2.12}$$

3. Main Results.

Definitions. As discussed in subsection 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 $p(x) = 3/5$, $\eta(x) = 1$ for $1 \geq |x| > 1/4$ and $p(x) = 1/5$, $\eta(x) = 0$ for $|x| \leq 1/4$ (see Figure 2c for an illustration of p_0 and p_1 for this distribution). 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 is an adversarial Bayes classifier as well. (These claims are rigorously justified in Example 4.6.) Thus 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 3.1. *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 2.1, 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 3.2. *A set S is* degenerate *for an adversarial Bayes classifier A if for all Borel E with $A - S \subset E \subset A \cup S$, the set E is also an adversarial Bayes classifier.*

Equivalently, a set S is degenerate for A if for all disjoint subsets $S_1, S_2 \subset S$, the set $A \cup S_1 - S_2$ 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 these properties 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 3.3) and furthermore, every adversarial Bayes classifier has a 'regular' representative when $d = 1$ (Theorem 3.5). The differentiation argument in subsection 2.1 then produces necessary conditions characterizing regular adversarial Bayes classifiers in one dimension (Theorem 3.7). These conditions provide a tool for understanding how the adversarial Bayes classifier depends on ϵ; see Theorem 3.9 and Propositions 4.7 to 4.9. Identifying all adversarial Bayes classifiers then requires characterizing degenerate sets, and we provide such a criterion under specific assumptions. Furthermore, Theorem 3.4 provides alternative criteria for equivalence up to degeneracy, which are helpful for understanding degenerate sets.

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

Theorem 3.3. *If $\mathbb{P} \ll \mu$, then equivalence up to degeneracy is an equivalence relation.*

Example 5.4 shows that the assumption $\mathbb{P} \not\ll \mu$ is necessary for this result. Additionally, uniqueness up to degeneracy generalizes multiple notions of uniqueness for the Bayes classifier.

Theorem 3.4. *Assume that $\mathbb{P} \ll \mu$. 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 A), Item B), and Item C) are equivalent notions of uniqueness of the Bayes classifier (see subsection 2.1). When $\mathbb{P} \not\ll \mu$, Theorem 3.3 is false although Item B) and Item C) are still equivalent (see Example 5.4 and Lemma C.5). This equivalence suggests a different notion of uniqueness for such distributions, see the subsection 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 3.5. *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 3.6. *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 (2.8) always hold for a regular adversarial Bayes classifier.

Theorem 3.7. *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 second order necessary conditions (2.8a) (resp. (2.8b)). 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 first order necessary conditions (3.1a) (resp. (3.1b)).

$$p_1'(a_i + \epsilon) - p_0'(a_i - \epsilon) \geq 0 \tag{3.1a}$$

$$p_0'(b_i + \epsilon) - p_1'(b_i - \epsilon) \geq 0 \tag{3.1b}$$

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, as noted in subsection 2.1. Section 4 applies this procedure above to several example distributions, see Example 4.1 for a crisp example. This analysis reveals interesting patterns across several example distributions. First, boundary points of the adversarial Bayes classifier are frequently within ϵ of boundary points of the Bayes classifier. Proposition 4.9 and Proposition 4.10 prove that this phenomenon occurs when either $\mathbb{P}$ is a uniform distribution on an interval or $\eta \in \{0, 1\}$, and Proposition 4.7 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 3.7 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 3.8. *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 $\overline{(\operatorname{supp}\mathbb{P}^\epsilon)^C} \cup \partial A$. 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 4.6 presents a counterexample where $\operatorname{supp}\mathbb{P}$ is an interval and $\mathbb{P}(\eta \in \{0,1\}) > 0$ while Example 6.5 presents a counterexample for which $\mathbb{P}(\eta \in \{0,1\}) = 0$ but $\operatorname{supp}\mathbb{P}$ is not an interval.

Prior work [2, 6] shows that a certain form of regularity for adversarial Bayes classifiers improves as ϵ increases. Theorem 3.5 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 in a set A.

Theorem 3.9. *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 $\operatorname{comp}(A_1^C \cap I^{\epsilon_1}) \geq \operatorname{comp}(A_2^C \cap I^{\epsilon_2})$.*

Subsection 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 3.8, the assumptions of Theorem 3.9 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 3.9 and the stronger version in subsection 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:

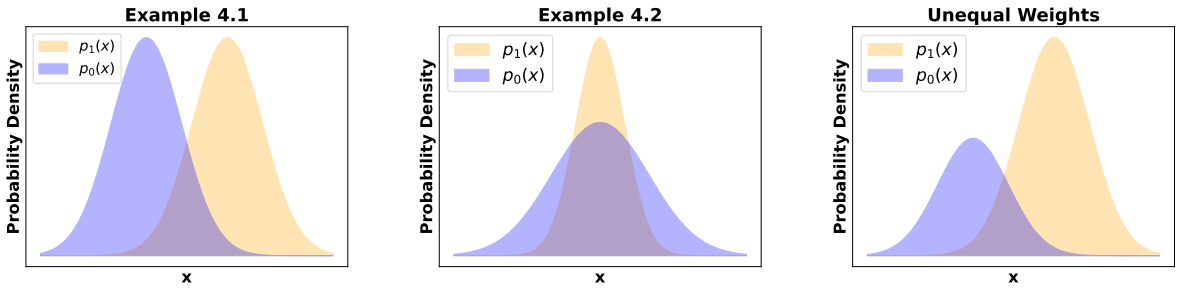
Theorem 3.10. *Let A be an adversarial Bayes classifier. Then A is equivalent up to degeneracy to a classifier A_1 for which $A_1 = C^\epsilon$ and a classifier A_2 for which $A_2^C = E^\epsilon$, for some sets C, E.*

Further understanding uniqueness up to degeneracy in higher dimension is an open question.

Paper Outline. Section 4 applies the tools presented above to compute adversarial Bayes classifiers for a variety of distributions. Subsequently, section 5 presents properties of equivalence up to degeneracy, including proofs of Theorems 3.3, 3.4, and 3.10. Subsections 5.2 and 5.3 further study degenerate sets, and these results are later applied in subsection 6.1 to prove Theorems 3.5 and 3.7. Subsection 6.2 further studies degenerate sets in one dimension to prove Theorem 3.8. Lastly, subsection 6.3 proves Theorem 3.9. Technical proofs and calculations appear in the appendix, which is organized so that it can be read sequentially.

4. Examples. The examples below find the equivalence classes under equivalence up to degeneracy for any $\epsilon > 0$. Examples 4.2 and 4.6 demonstrate a distribution for which the adversarial Bayes classifier is unique up to degeneracy for all ϵ while Example 4.5 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 4.1 and Example 4.4 describe an intermediate situations— uniqueness up to degeneracy fails fails only for a single value of ϵ in Example 4.4 and only for sufficiently large ϵ in Example 4.1. Lastly, Example 4.6 presents an example with a degenerate set.

Examples 4.5 and 4.6 exhibit situations where different adversarial Bayes classifiers have



(a) (b) (c)

Figure 1: (a) Gaussian Mixture with equal means and unequal variances as a in Example 4.2. (b) Gaussian Mixture with equal weights, unequal means, and equal variances as in Example 4.1. (c) Gaussian Mixture with unequal weights, unequal means, and equal variances.

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 4.2 exhibit a curious occurrence—the boundary of the adversarial Bayes classifier is within ϵ of the boundary of the Bayes classifier. Propositions 4.9 and 4.10 state conditions under which this phenomenon must occur. Next, Proposition 4.7 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 [17] 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 4.1 (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 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 (2.8) 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 (3.1b) (see Appendix K.1). Thus the candidate sets for the Bayes classifier are $\mathbb{R}$, $\emptyset$, and $(a(\epsilon), +\infty)$. The fourth bullet of Theorem 3.8 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 K.1 for details). Thus the adversarial Bayes classifier

is unique up to degeneracy only when $\epsilon < \frac{\mu_1 - \mu_0}{2}$.

When $\epsilon \le (\mu_1 - \mu_0)/2$, the set $(a(1/2), +\infty)$ 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 4.2 (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 1b for a depiction. Calculations similar to Example 4.1 show that the adversarial Bayes classifier is unique up to degeneracy for every ϵ, and is given by $(-b(\epsilon), b(\epsilon))$ where

$$b(\epsilon) = \frac{\epsilon\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) + \sqrt{\frac{4\epsilon^2}{\sigma_0^4\sigma_1^4} - 2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\ln\frac{(1-\lambda)\sigma_1}{\lambda\sigma_0}}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}}. \tag{4.1}$$

The computational details are similar to those of Example 4.1, and thus are delayed to Appendix K.2.

Unlike Example 4.1, when σ_0 and σ_1 are close, the Bayes and adversarial Bayes classifiers 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 4.3. *Consider a distribution for which* $\operatorname{supp}\mathbb{P}$ *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 K.3 for a proof.

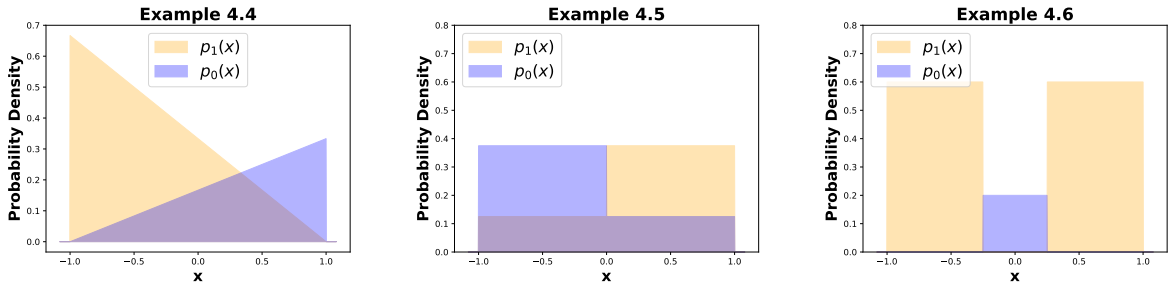
Example 4.4 (Uniqueness fails for a single value of ϵ). Consider a distribution for which

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

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

$$a(\epsilon) = \frac{2}{3}(1-\epsilon) \quad \text{and} \quad b(\epsilon) = \frac{2}{3} + \epsilon$$

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, $\epsilon < 1/5$. Then $p_0'(a(\epsilon) - \epsilon) = p_0'(b(\epsilon) + \epsilon) = 1/3$ and $p_1'(a(\epsilon) + \epsilon) = p_0'(b(\epsilon) - \epsilon) = 1/6$. Consequently, the point $a(\epsilon)$ fails to satisfy the second order necessary condition (3.1a). To identify all adversarial Bayes classifiers under uniqueness up to degeneracy for $\epsilon < 1/5$, it remains to compare the adversarial risks of $\emptyset$, $\mathbb{R}$, and $(-\infty, b(\epsilon))$, and Theorem 3.8 implies that none of these sets could be equivalent up to degeneracy. These values compute to $R^{\epsilon}(\emptyset) = 2/3$, $R^{\epsilon}(\mathbb{R}) = 1/3$, and $R^{\epsilon}((-\infty, b(\epsilon))) = \left(\frac{1+\epsilon}{2}\right)^2$. Therefore, for all $\epsilon < 1/5$, the set $(-\infty, b(\epsilon))$ is



(a) (b) (c)

Figure 2: (a) The distribution of Example 4.4. (b) The distribution of Example 4.5. (c) The distribution of Example 4.6.

an adversarial Bayes classifier iff $\epsilon \leq 2/\sqrt{3} - 1$ while $\mathbb{R}$ is an adversarial Bayes classifier iff $\epsilon \geq 2/\sqrt{3} - 1$. Theorem 3.9 then implies that this last statement holds without the restriction $\epsilon < 1/5$.

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 4.5 (Non-uniqueness for all $\epsilon > 0$). Let p be the uniform density on the interval $[-1, 1]$ 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 4.4, so we delay the details to Appendix K.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 3.8 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 [14] relates this condition to the sample complexity of learning from a function class. For the example above, Theorem 3.4 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 4.6 (Example of a degenerate set). Consider a distribution on $[-1, 1]$ with

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

Theorem 3.7 and Lemma 4.3 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 $\epsilon < 1/8$, 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 K.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 4.7. *Let $B = \bigcup_{i=1}^M (c_i, d_i)$, $A = \bigcup_{i=1}^M (a_i, b_i)$ be the 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 $R(A) - R(B) \leq 2\epsilon MK$.*

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

Proof.

$$R(A) - R(B) = \sum_{i=1}^{M} \int_{\min(a_i,c_i)}^{\max(a_i,c_i)} p_1 dx + \int_{\min(b_i,d_i)}^{\max(b_i,d_i)} p_0 dx \leq 2\epsilon MK$$

■

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 4.8. *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 (2.8a) and (2.8b) 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 (2.8a) within ϵ of z. Analogous reasoning shows that there is a solution to (2.8b) 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 1c) for an illustration. Just as in Example 4.1, the necessary conditions (2.8) reduce to linear equations and so there is at most one $a(\epsilon)$ solving (2.8a) and at most one $b(\epsilon)$ solving (2.8b). Thus Proposition 4.8 implies that the solutions to the first order necessary conditions (2.8) 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 4.8 implies that solutions to the first order necessary conditions (2.8) are within ϵ of solutions to $p_0(z) = p_1(z)$.

Proposition 4.9. *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 4.8, see Appendix K.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 4.10. *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 4.8, see Appendix K.7 for details.

5. Equivalence up to Degeneracy.

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 5.1. *Let $\mathbb{P} \ll \mu$. 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 5.1 is proved in Appendix C.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 complimentary slackness conditions of Theorem 2.4 implies that Item 2) and Item 3) equivalent, (see the proof of Lemma C.5 in Appendix C). 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. (Lemma C.3), and thus 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 complimentary slackness condition of (2.11) of Theorem 2.4 to argue that $D = A_1 \triangle A_2$ has $\mathbb{P}^*$-measure zero. First, we show

that if $D_1 = \operatorname{int} D \cap \mathbb{Q}^d$, $D_2 = \operatorname{int} D \cap (\mathbb{Q}^d)^C$ then $D_1^\epsilon = D_2^\epsilon = (\operatorname{int} D)^\epsilon$ (see Lemma C.7). 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 complimentary slackness condition (2.11) implies that $\mathbb{P}_0^*(\operatorname{int} D) = 0$. Similarly, one can argue that $\mathbb{P}_1^*(\operatorname{int} D) = 0$. Next, to prove $\mathbb{P}^*(D \cap \partial D) = 0$, we prove that when $\mathbb{P} \ll \mu$, the boundary ∂A is always a degenerate set for an adversarial Bayes classifier A when $\mathbb{P} \ll \mu$. Consequently:

Lemma 5.2. *Let A be an adversarial Bayes classifier. If $\mathbb{P} \ll \mu$, then A, $\overline{A}$, and $\operatorname{int} A$ are all equivalent up to degeneracy.*

See Appendix C.1 for a proof. Proposition 5.1 has several useful consequences for understanding degenerate sets, which we explore in subsection 5.3. Specifically, when $\mathbb{P} \ll \mu$, equivalence up to degeneracy is in fact an equivalence relation.

Proof of Theorem 3.3. Item 3) of Proposition 5.1 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. Thus equivalence up to degeneracy is an equivalence relation because equality of functions $\mathbb{P}^*$-a.e. is an equivalence relation. ∎

Furthermore, Proposition 5.1 implies Theorem 3.4. Item 2) of Proposition 5.1 is equivalent to Item B) of Theorem 3.4 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 5.1 is equivalent to Item C) of Theorem 3.4.

First, to show Item C) ⇒ Item 3), notice that the complimentary slackness condition in (2.12) implies that

$$\mathbf{1}_{\eta^* > 1/2} \leq \mathbf{1}_A \leq \mathbf{1}_{\eta^* \geq 1/2} \quad \mathbb{P}^*\text{-a.e.} \tag{5.1}$$

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 5.3. *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 $0 = \mathbb{P}^*(\hat{A}_2 \triangle \hat{A}_1) = \mathbb{P}^*(\eta^* = 1/2)$. See Appendix D for proofs of Theorem 3.4 and Lemma 5.3. 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 D.2 in Appendix D.1).

Theorem 3.3 is false when $\mathbb{P}$ is not absolutely continuous with respect to μ:

Example 5.4. Consider a distribution for which $\mathbb{P}_0 = \frac{1}{2}\delta_{-\epsilon}$ 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 $S_\epsilon(\mathbf{1}_{A^C})(-\epsilon) = 1$. If either case, $R^\epsilon(A) \geq 1/2$. The classifier $A = \mathbb{R}$ achieves adversarial classification $1/2$ and therefore $R_*^\epsilon = 1/2$. 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 5.1 are still equivalent when $\mathbb{P} \not\ll \mu$, as are Item B) and Item C) of Theorem 3.4 (see Lemma C.5 in Appendix C.2 and Proposition D.3 of Appendix D.2). As the proof of Theorem 3.3 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 $\mathbb{P} \not\ll \mu$.

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:

$$A^\epsilon = \{\mathbf{x} \colon \overline{B_\epsilon(\mathbf{x})} \text{ intersects } A\} \tag{5.2}$$

$$A^{-\epsilon} = \{\mathbf{x} : \overline{B_\epsilon(\mathbf{x})} \subset A\} \tag{5.3}$$

See Appendix E for a proof. Prior works [2, 6] note that applying the ϵ, $-\epsilon$ operations in succession can improve the regularity of an adversarial Bayes classifier. Additionally,

Lemma 5.5. *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 E 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, [18] 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 [10] states that

Theorem 5.6. *If A is universally measurable, then A^ϵ is as well.*

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

As $\mathbf{1}_{A^\epsilon} = S_\epsilon(\mathbf{1}_A)$, Theorem 5.6 implies that $S_\epsilon(\mathbf{1}_A)$ is a universally measurable function if A is universally measurable.

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 [9, 17, 18] study this minimization problem over the Borel σ-algebra. We show that these two approaches are equivalent:

Theorem 5.7. *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 F 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.

5.3. Describing Degenerate Sets and Proof of Theorem 3.10. Proposition 5.1 together with fundamental properties of the $^\epsilon$ and $^{-\epsilon}$ operations imply several results on degenerate sets.

First, Lemma 2.1 implies that countable unions and intersections of adversarial Bayes classifiers are adversarial Bayes classifiers. Item 3) of Proposition 5.1 then implies that countable unions and intersections preserve equivalence up to degeneracy. As a result:

Lemma 5.8. *Let $\mathbb{P} \ll \mu$. Then a countable union of degenerate sets is degenerate.*

See Appendix G.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 5.9. *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 5.10. *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 E for proofs of Lemma 5.9 and Lemma 5.10. Lemma 5.5 then implies:

Corollary 5.11. *Assume $\mathbb{P} \ll \mu$ and let A be an adversarial Bayes classifier. Then $A, (A^\epsilon)^{-\epsilon}, (A^{-\epsilon})^\epsilon$ are all equivalent up to degeneracy.*

Proof. Lemma 5.10 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 5.1 implies that A, $(A^{-\epsilon})^\epsilon)$, and $(A^\epsilon)^{-\epsilon}$ are all equivalent up to degeneracy. Lemma 5.9 then implies that $(A^\epsilon)^{-\epsilon} - (A^{-\epsilon})^\epsilon$ is a degenerate set. ∎

Theorem 3.10 is an immediate consequence of Corollary 5.11. Furthermore, Corollary 5.11 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 5.12. *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*

$$\bigcup \left\{ C : \text{connected components of } A \text{ or } A^C \text{ with } C^{-\epsilon} = \emptyset \right\} \tag{5.4}$$

is contained in a degenerate set of A.

See Appendix G.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 5.13. *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 G.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 5.14. *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 is an adversarial Bayes classifier, and furthermore A and A_2 are equivalent up to degeneracy by Proposition 5.1. Notice that the set A_2 satisfies

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

and then Lemma 5.10 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 5.1 implies that $\mathbf{1}_{A^\epsilon} = \mathbf{1}_{(A\cup D)^\epsilon}$-$\mathbb{P}_0$-a.e. which in turn implies $\mathbf{1}_{A^\epsilon} = \mathbf{1}_{(A\cup(D^\epsilon)^{-\epsilon})^\epsilon}$-$\mathbb{P}_0$-a.e. ∎

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.

6.1. Regular adversarial Bayes classifiers—Proof of Theorem 3.5 and Theorem 3.7. 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 and A^C of length strictly less than 2ϵ is contained in a degenerate set by Proposition 5.12.

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 6.1. *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 5.12 and Lemma 5.2.

Proof of Lemma 6.1. Lemma 5.2 implies that $\operatorname{int} A$ and $\overline{A}$ are both adversarial Bayes classifiers, and thus Corollary 5.11 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 for $\operatorname{int} A$ and $\overline{A}$ respectively. Thus Lemma 5.2 and Corollary 5.11 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 Corollary 5.11 and Lemma 5.9 imply 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 3.5, which is detailed in Appendix H.1, 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 $\mathbb{R}^d$ 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 A. 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 3.7 follows the fact that one can express the adversarial risk of $A = \bigcup_{i=m}^{M}(a_i, b_i)$ as (2.7) when A is regular.

Proof of Theorem 3.7. 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 (2.7). The first order necessary conditions for a minimizer then imply the first relation of (2.8) and the second order necessary conditions for a minimizer then imply the first relation of (3.1). The argument for b_i is analogous. ∎

6.2. Degenerate Sets in One Dimension—Proof of Theorem 3.8. 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 strong than Proposition 5.12.

Lemma 6.2. *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 6.2. 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 6.1. 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 of length less than or equal to 2ϵ are included in $\tilde{A}_2 - A$, which is a degenerate set. ∎

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

Corollary 6.3. *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 5.13 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 3.8. To start:

Lemma 6.4. *Let $\mathbb{P} \ll \mu$ and let A be an adversarial Bayes classifier. If $\operatorname{supp} \mathbb{P}$ is an interval and 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 I.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 $J = [d_3, d_4]$ contained in $\operatorname{supp} \mathbb{P}$, in the sense that if J' is a degenerate interval and $J \subset J'$ then $J' = J$. Corollary 6.3 implies that $|J| \leq 2\epsilon$ and Lemma 5.14 implies that J is a distance strictly more than 2ϵ from any other degenerate set. Thus the intervals $(d_3 - \epsilon, d_3)$, $(d_4, d_4 + \epsilon)$ do not intersect a degenerate subset of A, and these intervals must be entirely contained in A or A^C due to Lemma 6.2. 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', which imply $\eta = 1$ and $\eta = 0$, respectively on a set of positive measure.

Lemma 6.4 and Lemma 5.14 together imply the fourth bullet of Theorem 3.8. The argument is outlined below, with a formal proof in Appendix I.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 5.14 implies that $[x, y] \subset (D^\epsilon)^{-\epsilon}$ is degenerate, which would contradict Lemma 6.4. 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 3.8— The first bullet of Theorem 3.8 is Lemma 6.2, the second bullet is Corollary 6.3, the third bullet is Lemma 5.8, and the fourth bullet is shown in Appendix I.2.

Lemma 6.4 and the fourth bullet of Theorem 3.8 are false when $\operatorname{supp} \mathbb{P}$ is not an interval.

Example 6.5. Consider a probability distribution for which

$$p_1(x) = \begin{cases} \frac{1}{4\epsilon} & \text{if } 2\epsilon \leq |x| \leq 3\epsilon \\ \frac{1}{12\epsilon} & \text{if } |x| \leq \epsilon \\ 0 & \text{otherwise} \end{cases} \qquad p_0(x) = \begin{cases} \frac{1}{9\epsilon} & \text{if } 2\epsilon \leq |x| \leq 3\epsilon \\ \frac{1}{18\epsilon} & \text{if } |x| \leq \epsilon \\ 0 & \text{otherwise} \end{cases}$$

See Figure 3 for an illustration. Then there are no solutions x to the necessary conditions (2.8) 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 $\{-4\epsilon, -3\epsilon, -2\epsilon, -\epsilon, 0, \epsilon, 2\epsilon, 3\epsilon, 4\epsilon\}$.

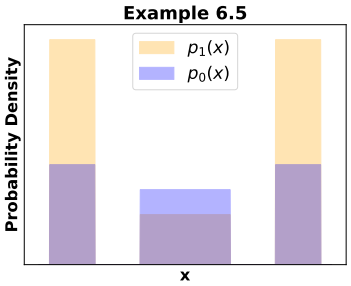


Figure 3: The distribution of Example 6.5.

By comparing the risks of all adversarial Bayes classifiers with endpoints in this set, one can show that $\mathbb{R}$ is an adversarial Bayes classifier. At the same time, $R^\epsilon(S^C) = R^\epsilon(\mathbb{R})$ for any subset S of $[-\epsilon, \epsilon]$. Thus S is a degenerate set, but $\eta(x) = 3/5$ on $[-\epsilon, \epsilon]$.

6.3. Regularity as ϵ Increases—Proof of Theorem 3.9. 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 (2.5) 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 and no connected component of A_2^C should contain a connected component of A_1.

We adopt an additional notation convention 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 6.6. *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 $(b_j^2, a_{j+1}^2) \cap I^{\epsilon_1}$.*

Example 4.1 demonstrates the the exception to the second bullet of this lemma— when $\epsilon \geq (\mu_1 - \mu_0)/2$, both $\mathbb{R}$ and $\emptyset$ are adversarial Bayes classifiers.

To show Lemma 6.6, notice that if $A_2 = \bigcup_{i=1}^M (a_i^2, b_i^2)$ is a regular adversarial Bayes classifier, 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 J 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.

7. Related Works. Prior work analyzes several variations of our setup, such as perturbations in open balls [6], alternative perturbation sets [4], attacks using general Wasserstein p-metrics [21, 20], minimizing R^ϵ over Lebesgue measurable sets [18], the multiclass setting [20], and randomized classifiers [11, 20]. 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 2.1 from [6]). 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 [20] but not binary classification [11].

Prior work [1, 4, 17] adopts a different method for identifying adversarial Bayes classifiers for various distributions. To prove a set is an adversarial Bayes classifier, [4] 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, 4, 17] 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 (2.8), and [1] relies on a result of [21] 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, 6] prove the existence of adversarial Bayes classifiers with one sided tangent balls. Theorem 3.10 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 $\text{comp}(A(\epsilon)) + \text{comp}(A(\epsilon)^C)$ 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 the stronger result of Lemma 6.6, and Lemma 6.6 does not imply this result of [1].

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.

REFERENCES

[1] Q. Aiken, B. Brewer, B. Fetsko, R. Murray, A. Pickarski, and H. Shugart, *A primal-dual method for tracking topological changes in optimal adversarial classification*, 2022, https://bruce2142.github.io/research/assets/pdfs/AVL_work_in_progress.pdf.

[2] P. Awasthi, N. S. Frank, and M. Mohri, *On the existence of the adversarial bayes classifier*, (2023), https://arxiv.org/abs/2112.01694.

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

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

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

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

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021, https://arxiv.org/abs/2010.11929.

[8] N. S. Frank, *The uniqueness of the adversarial bayes classifier and the adversarial consistency of surrogate risks for binary classification*, arxiv, (2024).

[9] N. S. Frank and J. Niles-Weed, *The adversarial consistency of surrogate risks for binary classification*, NeurIps, (2024).

[10] N. S. Frank and J. Niles-Weed, *Existence and minimax theorems for adversarial surrogate risks in binary classification*, JMLR, (2024).

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

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

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

[14] P. Massart and É. Nédélec, *Risk bounds for statistical learning*, The Annals of Statistics, 34 (2006).

[15] T. Nishiura, *Absolute Measurable Spaces*, Cambridge University Press, 2010.

[16] S. Peng, W. Xu, C. Cornelius, M. Hull, K. Li, R. Duggal, M. Phute, J. Martin, and D. H. Chau, *Robust principles: Architectural design principles for adversarially robust cnns*, 2023, https://arxiv.org/abs/2308.16258.

[17] M. S. Pydi and V. Jog, *Adversarial risk via optimal transport and optimal couplings*, ICML, (2020).

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

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

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

[21] N. G. Trillos and R. Murray, *Adversarial classification: Necessary conditions and geometric flows*, arxiv, (2022).

[22] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2nd ed., 2003.

[23] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, *Theoretically principled trade-off between robustness and accuracy*, 2019, https://arxiv.org/abs/1901.08573.

Appendix A. Proof of Lemma 2.1.

First, the S_ϵ operation satisfies a subadditivity property:

Lemma A.1. *Let S_1 and S_2 be two subsets of $\mathbb{R}^d$. Then*

$$S_\epsilon(\mathbf{1}_{S_1}) + S_\epsilon(\mathbf{1}_{S_2}) \geq S_\epsilon(\mathbf{1}_{S_1 \cap S_2}) + S_\epsilon(\mathbf{1}_{S_1 \cup S_2}) \tag{A.1}$$

Proof. First, notice that

$$S_\epsilon(\mathbf{1}_{S_1})(\mathbf{x}) + S_\epsilon(\mathbf{1}_{S_2})(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \notin S_1^\epsilon \text{ and } \mathbf{x} \notin S_2^\epsilon \\ 1 & \text{if } \mathbf{x} \in S_1^\epsilon \triangle S_2^\epsilon \\ 2 & \text{if } \mathbf{x} \in S_1^\epsilon \cap S_2^\epsilon \end{cases} \tag{A.2}$$
$$= \mathbf{1}_{S_1^\epsilon \cap S_2^\epsilon}(\mathbf{x}) + \mathbf{1}_{S_1^\epsilon \cup S_2^\epsilon}(\mathbf{x})$$

Next, one can always swap the order of two maximums but a min-max is always larger than a max-min. Therefore:

$$\begin{aligned} &S_\epsilon(\mathbf{1}_{S_1 \cap S_2}) + S_\epsilon(\mathbf{1}_{S_1 \cup S_2}) = S_\epsilon(\min(\mathbf{1}_{S_1}, \mathbf{1}_{S_2})) + S_\epsilon(\max(\mathbf{1}_{S_1}, \mathbf{1}_{S_2})) \\ &\leq \min(S_\epsilon(\mathbf{1}_{S_1}), S_\epsilon(\mathbf{1}_{S_2})) + \max(S_\epsilon(\mathbf{1}_{S_1}), S_\epsilon(\mathbf{1}_{S_2})) = \mathbf{1}_{S_1^\epsilon \cap S_2^\epsilon} + \mathbf{1}_{S_1^\epsilon \cup S_2^\epsilon} \end{aligned} \tag{A.3}$$

Comparing (A.2) and (A.3) results in (A.1). ∎

Therefore, the adversarial classification risk is sub-additive.

Corollary A.2. *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 2.1:

Proof of Lemma 2.1. Let A_1 and A_2 be two adversarial Bayes classifiers. Then Corollary A.2 implies that

$$R_*^\epsilon \geq R^\epsilon(A_1 \cup A_2) + R^\epsilon(A_1 \cap A_2)$$

and hence $A_1 \cap A_2$ and $A_1 \cup A_2$ must be adversarial Bayes classifiers as well. ∎

Appendix B. Complimentary Slackness– Proof of Theorem 2.4.

The complimentary slackness relations of Theorem 2.4 are a consequence of the minimax relation of Theorem 2.3 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 B.1. *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 [18] and Lemma 3 of [10] proved slightly different versions of this result, so we include a proof here for completeness.

Proof. Let $\mathbb{Q}'$ be any measure with $W_\infty(\mathbb{Q},\mathbb{Q}') \leq \epsilon$. 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 [13]. Then $S_\epsilon(\mathbf{1}_A)(\mathbf{x}) \geq \mathbf{1}_A(\mathbf{x}')$ γ-a.e. Thus

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

Now taking a supremum over all $\mathbb{Q}' \in \mathcal{B}_\epsilon^\infty(\mathbb{Q})$ concludes the proof. ∎

Subsequently, one can proof Theorem 2.4 with this result.

Proof of Theorem 2.4.

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:

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

$$= \int \eta^*\mathbf{1}_{A^C}d\mathbb{P}_1 + \int (1-\eta^*)\mathbf{1}_A d\mathbb{P}_0^* \geq C^*(\eta^*)d\mathbb{P}^* = \bar{R}(\mathbb{P}_0^*,\mathbb{P}_1^*) \tag{B.2}$$

The first inequality follows from Lemma B.1 while the second inequality follows from the definition of C^* in (2.3). By Theorem 2.3, the first expression of (B.1) and the last expression of (B.2) are equal. Thus all the inequalities above must in fact be equalities. Thus the fact that the inequality in (B.2) is an equality implies (2.12). Lemma B.1 and the fact that the inequality in (B.1) must be an equality implies (2.11).

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 (2.11) and (2.12). 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 2.3 implies that $R^\epsilon(A') \geq \bar{R}(\mathbb{P}_0',\mathbb{P}_1')$ for *any* Borel A' and $\mathbb{P}_0' \in \mathcal{B}_\epsilon^\infty(\mathbb{P}_0), \mathbb{P}_1'(\mathbb{P}_1)$. Thus if one can show

$$R^\epsilon(A) = \bar{R}(\mathbb{P}_0^*,\mathbb{P}_1^*), \tag{B.3}$$

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 (B.3). Applying (2.11) followed by (2.12), one can conclude that

$$R^\epsilon(A) = \int S_\epsilon(\mathbf{1}_A^C)d\mathbb{P}_1 + \int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 = \int \mathbf{1}_{A^C}d\mathbb{P}_1^* + \int \mathbf{1}_A d\mathbb{P}_0^* \qquad (2.11)$$

$$= \int C^*(\eta^*)d\mathbb{P}^* = \bar{R}(\mathbb{P}_0^*,\mathbb{P}_1^*) \qquad (2.12)$$

∎

Appendix C. Proof of Proposition 5.1 and Lemma 5.2.

The proof of Proposition 5.1 relies on Lemma 5.2.

C.1. Proof of Lemma 5.2.

The $^\epsilon$ operation on sets interacts particularly nicely with Lebesgue measure.

Lemma C.1. *For any set A, ∂A^ϵ has Lebesgue measure zero.*

This result is standard in geometric measure theory, see for instance the proof of Lemma 4 in [2] for a proof. Next, the closure and $^\epsilon$ operations commute:

Lemma C.2. *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 $\mathbf{x} = \mathbf{a} + \mathbf{h}$, where $\mathbf{a} \in \overline{A}$ 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, $\mathbf{a} + \mathbf{h} \in \overline{A^\epsilon}$. ∎

Proof of Lemma 5.2. We will show that if E is any set with $\operatorname{int} A \subset E \subset \overline{A}$, then E is an adversarial Bayes classifier.

First, Lemmas C.1 and C.2 imply that

$$\mathbb{P}_0(A^\epsilon) = \mathbb{P}_0(\overline{A^\epsilon} = \mathbb{P}_0(\overline{A}^\epsilon) \tag{C.1}$$

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

$$\mathbb{P}_1((A^C)^\epsilon) = \mathbb{P}_1((\overline{A}^C)^\epsilon) \tag{C.2}$$

A similar line of reasoning shows that

$$\mathbb{P}_1(\overline{(A^C)^\epsilon}) = \mathbb{P}_1(\overline{(A^C)}^\epsilon) = \mathbb{P}_1((\operatorname{int} A^C)^\epsilon) \tag{C.3}$$

and thus

$$\mathbb{P}_0(A^\epsilon) = \mathbb{P}_0((\operatorname{int} A)^\epsilon) \tag{C.4}$$

∎

Equations (C.1)–(C.4) imply 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.

C.2. Proof of Proposition 5.1.

The following lemma show that Item 2) implies Item 1).

Lemma C.3. *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*

$$S_\epsilon(\mathbf{1}_{A_1}) = S_\epsilon(\mathbf{1}_{A_2}) = S_\epsilon(\mathbf{1}_{A_1 \cap A_2}) = S_\epsilon(\mathbf{1}_{A_1 \cup A_2}) \quad \mathbb{P}_0\text{-a.e.} \tag{C.5}$$

and

$$S_\epsilon(\mathbf{1}_{A_1^C}) = S_\epsilon(\mathbf{1}_{A_2^C}) = S_\epsilon(\mathbf{1}_{(A_1 \cap A_2)^C}) = S_\epsilon(\mathbf{1}_{(A_1 \cup A_2)^C}) \quad \mathbb{P}_1\text{-a.e.} \tag{C.6}$$

See Appendix C.2.1 for a proof. As a result:

Corollary C.4. *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 (C.5) and (C.6) 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 C.5. *Let A_1 and A_2 be two adversarial Bayes classifiers. Then 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 of Lemma C.5. Assume that A_1 and A_2 are both adversarial Bayes classifiers. Lemma 2.1 then implies that $A_1 \cup A_2$, $A_1 \cap A_2$ are both adversarial Bayes classifiers. Equation (2.11) of Theorem 2.4 implies that

$$\int S_\epsilon(\mathbf{1}_{A_1 \cup A_2}) d\mathbb{P}_0 = \int \mathbf{1}_{A_1 \cup A_2} d\mathbb{P}_0^* = \int_{A_1 \cap A_2} \mathbf{1}_{A_1 \cap A_2} d\mathbb{P}_0^* + \mathbb{P}_0^*(A_1 \triangle A_2) = \int S_\epsilon(\mathbf{1}_{A_1 \cap A_2}) d\mathbb{P}_0 + \mathbb{P}_0^*(A_1 \triangle A_2)$$

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 C.3. Therefore, Corollary C.4 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 equivalence up to degeneracy with Item 3) is a consequence of Lemma 5.2 and a result on the $^\epsilon$ operation. See Appendices C.1 and C.2.2 proofs.

Lemma C.6. *If A is any adversarial Bayes classifier, then $\mathbb{P}_0(A^\epsilon) = \mathbb{P}_0(\overline{A}^\epsilon) = \mathbb{P}_0((\operatorname{int} A)^\epsilon)$ and $\mathbb{P}_1((A^C)^\epsilon) = \mathbb{P}_1(((\operatorname{int} A)^C)^\epsilon) = \mathbb{P}_1((\overline{A}^C)^\epsilon)$*

Lemma C.7. *Let U be an open set and let $\mathbb{Q}$ be the set of rational numbers. Then $U^\epsilon = (U \cap \mathbb{Q}^d)^\epsilon = (U \cap (\mathbb{Q}^d)^C)^\epsilon$.*

Proof of Proposition 5.1. Lemma C.5 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 C.4 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 C.3 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. Let $\mathbb{P}_0^*, \mathbb{P}_1^*$ be any maximizers of $\bar{R}$ and let $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$. Further let $A_3 = A_1 \cap A_2$, $A_4 = A_1 \cup A_2$, 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 $D := A_1 \triangle A_2$ into four 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 $D_4 = D \cap \partial D \cap A_4 - D_3$. Notice that these four sets satisfy $D = D_1 \sqcup D_2 \sqcup D_3 \sqcup D_4$.

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 C.7 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 (2.11) of Theorem 2.4 implies that $\mathbb{P}_0^*(A_3 \cup D_1) = \mathbb{P}_0^*(A_3 \cup \operatorname{int} D) = \mathbb{P}_0^*(A_3 \cup D_2)$. As D_1 and D_2 are disjoint sets whose union is $\operatorname{int} D$, it follows that $\mathbb{P}_0^*(\operatorname{int} D) = 0$. Analogously, comparing $S_\epsilon(\mathbf{1}_{(A_4 - D_1)^C})$, $S_\epsilon(\mathbf{1}_{(A_4 - D_2)^C})$, and $S_\epsilon(\mathbf{1}_{(A_4 - \operatorname{int} D)^C})$ results in $\mathbb{P}_1^*(\operatorname{int} D) = 0$.

Next we argue that $\mathbb{P}^*(D_3) = 0$. Lemma C.6 implies that $S_\epsilon(\mathbf{1}_{A_3 \cup D_3}) = S_\epsilon(\mathbf{1}_{A_3})$ $\mathbb{P}_0$-a.e., and (2.11) of Theorem 2.4 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 C.6 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 (2.11) of Theorem 2.4 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 $A_4 \cup D_4$.

C.2.1. Proof of Lemma C.3.

Proof of Lemma C.3. 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 to implies (C.5). ∎

C.2.2. Proof of Lemma C.7.

Before proving Lemma C.7, we reproduce another useful lemma from [2].

Lemma C.8. *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 C.7. 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 $B_\epsilon(\mathbf{u}) \subset (U \cap \mathbb{Q})^\epsilon$. 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 C.8 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 $U^\epsilon \subset (U \cap \mathbb{Q}^d)^\epsilon$.

Appendix D. Proof of Theorem 3.4.

D.1. Proof of Lemma 5.3. Lemma 24 of [10] 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 D.1. *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 $\mathbb{P}_i^*$ supported on Δ_ϵ as defined in (F.3). 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 [13] proves that when $W_\infty(\mathbb{Q}, \mathbb{Q}') \leq \epsilon$, there always exists a coupling γ between $\mathbb{Q}$ and $\mathbb{Q}'$ with $\operatorname{ess\,sup}_{(\mathbf{x},\mathbf{y})\sim\gamma} \leq \epsilon$.

Next, we argue that $\hat{A}_1 = \{\hat{\eta} > 1/2\}$ and $\hat{A}_2 = \{\hat{\eta} \geq 1/2\}$.

Proof of Lemma 5.3. Let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$, γ_0^*, and γ_1^* be the measures given by Proposition D.1 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 D.1. We will show that the classifier $\hat{A}_1 = \{\hat{\eta} > 1/2\}$ and $\hat{A}_2 = \{\hat{\eta} \geq 1/2\}$ satisfy the required properties by verifying the complimentary slackness conditions in Theorem 2.4.

Below, we verify these conditions for $\{\hat{\eta} > 1/2\}$, the argument for $\{\hat{\eta} \geq 1/2\}$ is analogous. First, Item 1 of Proposition D.1 implies that $\mathbf{1}_{\{\hat{\eta}>1/2\}} = \mathbf{1}_{\eta^*>1/2}$ $\mathbb{P}^*$-a.e. and $\mathbf{1}_{\{\hat{\eta}>1/2\}^C} = \mathbf{1}_{\hat{\eta}\leq 1/2\}} = \mathbf{1}_{\eta^*\geq 1/2}$ $\mathbb{P}^*$-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 D.1 implies that $\hat{\eta}$ assumes its maximum over closed ϵ-balls $\mathbb{P}_0$-a.e. and hence $S_\epsilon(\mathbf{1}_{\hat{\eta}>1/2}) = \mathbf{1}_{S_\epsilon(\hat{\eta})>1/2}$ $\mathbb{P}_0$-a.e. Additionally, Item 2 of Proposition D.1 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

$$\int S_\epsilon(\mathbf{1}_{\hat{\eta}>1/2})(\mathbf{x}) d\mathbb{P}_0(\mathbf{x}) = \int \mathbf{1}_{\hat{\eta}>1/2}(\mathbf{y}) d\gamma_0^*(\mathbf{x},\mathbf{y}) = \int \mathbf{1}_{\hat{\eta}>1/2} d\mathbb{P}_0^* \tag{D.1}$$

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 (D.1) verifies (2.11). ∎

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 D.2. *Let A be any adversarial Bays classifier and let $\hat{A}_1$, $\hat{A}_2$ be the two adversarial Bayes classifiers of Lemma 5.3. Then*

$$\int S_\epsilon(\mathbf{1}_{\hat{A}_1}) d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_A) d\mathbb{P}_0 \leq \int S_\epsilon(\mathbf{1}_{\hat{A}_2}) d\mathbb{P}_0 \tag{D.2}$$

and

$$\int S_\epsilon(\mathbf{1}_{\hat{A}_2^C})d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_A^C)d\mathbb{P}_1 \leq \int S_\epsilon(\mathbf{1}_{\hat{A}_1^C})d\mathbb{P}_1. \tag{D.3}$$

Proof. Let $\mathbb{P}_0^*$, $\mathbb{P}_1^*$, $\mathbb{P}^*$, and η^* be as described by Lemma 5.3. Then the complimentary slackness condition (2.11) implies that $\int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 = \int \mathbf{1}_A d\mathbb{P}_0^*$ and (2.12) implies (5.1), 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 5.3 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 complimentary slackness condition (2.11) applied to $\hat{A}_1$ and $\hat{A}_2$ then implies (D.2).

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 (D.3). ∎

D.2. Proving Theorem 3.4.

Proof of Theorem 3.4. **Item A)⇒Item B):** Assume that the the adversarial Bayes classifier is unique up to degeneracy. Then Item 2) of Proposition 5.1 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 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 2.1 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) of Theorem 3.4 implies Item 2) of Proposition 5.1. Consequently, Proposition 5.1 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 5.3 and $\hat{A}_1$ and $\hat{A}_2$ are the adversarial Bayes classifiers described by Lemma 5.3.

Item C) ⇒ Item A): Item 3) of Proposition 5.1 then implies that Item C) of Theorem 3.4 implies that the adversarial Bayes classifier is unique up to degeneracy.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 (5.1) 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 $\mathbf{1}_{A_1} = \mathbf{1}_{A_2}$ $\mathbb{P}_0^*$ for any two adversarial Bayes classifiers A_1, A_2, or in other words, $\mathbb{P}^*(A_1 \triangle A_2) = 0$. ∎

The same argument shows that Item B) and Item 3) are equivalent when $\mathbb{P} \not\ll \mu$:

Proposition D.3. *The following are equivalent:*

A) 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

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

Proof. First, to show that Item A) implies Item B), one can use the same argument as in the proof of Theorem 3.4 with Item 2) of Proposition 5.1 replaced with Item 2) of Lemma C.5.

Next, to show that Item B) implies Item A), assume that $\mathbb{P}^*(\eta^* = 1/2) = 0$ where $\mathbb{P}^* = \mathbb{P}_0^* + \mathbb{P}_1^*$ and $\eta^* = d\mathbb{P}_1^*/d\mathbb{P}^*$ for some $\mathbb{P}_0^*$, $\mathbb{P}_1^*$ that maximize the dual $\bar{R}$. Then the

same argument as Item C)⇒ Item A) of Theorem 3.4 with Item 3) of Lemma C.5 in place of Item 3) of Proposition 5.1 shows that $\mathbb{P}^*(\eta^* = 1/2) = 0$ implies that $\mathbb{P}^*(A_1 \triangle A_2) = 0$ for any two adversarial Bayes classifiers A_1 and A_2. Applying this statement to the adversarial Bayes classifiers $\hat{A}_1$, $\hat{A}_2$ of Lemma 5.3 implies that $\mathbf{1}_{\hat{A}_1} = \mathbf{1}_{\hat{A}_2}$ $\mathbb{P}_0^*$-a.e. and hence the complimentary slackness condition (2.11) implies that $\int S_\epsilon(\mathbf{1}_{\hat{A}_1})d\mathbb{P}_0 = \int S_\epsilon(\mathbf{1}_{\hat{A}_2})d\mathbb{P}_0$. Lemma D.2 then implies that $\int S_\epsilon(\mathbf{1}_A)d\mathbb{P}_0 = \int S_\epsilon(\mathbf{1}_{\hat{A}_1})d\mathbb{P}_0 = \int S_\epsilon(\mathbf{1}_{\hat{A}_2})d\mathbb{P}_0$ for any other adversarial Bayes classifier A. ■

Appendix E. 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 (5.2) and (5.3), Lemmas 5.5, 5.9, and 5.10. These results have all appeared elsewhere in the literature —[2, 6].

The characterization of the $^\epsilon$ and $^{-\epsilon}$ operations provided by (5.2) and (5.3) is an essential tool for understanding how $^\epsilon$ and $^{-\epsilon}$ interact.

Proof of Equation (5.2). To show (5.2), notice that $\mathbf{x} \in A^\epsilon$ iff $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$, where $\mathbf{a}$ is some element 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 (5.3) then follows directly from Equation (5.2):

Proof of Equation (5.3). By Equation (5.2),

$$\mathbf{x} \in (A^C)^\epsilon \Leftrightarrow \overline{B_\epsilon(\mathbf{x})} \text{ intersects } A^C$$

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 (5.2) and Equation (5.3) immediately imply Lemma 5.9:

Proof of Lemma 5.9. By Equation (5.2), Equation (5.3), $(A^\epsilon)^{-\epsilon}$ 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 5.9 then immediately implies Lemma 5.10:

Proof of Lemma 5.10. First, Lemma 5.9 implies that $A \subset (A^\epsilon)^{-\epsilon}$ and thus $A^\epsilon \subset ((A^\epsilon)^{-\epsilon})^\epsilon$. At the same time, Lemma 5.9 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 5.9 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 5.5 is then an immediate consequence of Lemma 5.10.

Appendix F. Measurability.

F.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 [3, Chapter 7] or [15] for more about this construction.

Due to Theorem 5.6, throughout this paper, we adopt the convention $\int S_\epsilon(\mathbf{1}_A) d\nu$ is the integral of $S_\epsilon(\mathbf{1}_A)$ with respect to the completion of ν.

F.2. Proof of Theorem 5.7. 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 2.3, 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 F.1. *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,

$$\int g d\mathbb{Q} = \int g d\widetilde{\mathbb{Q}} \quad \text{for any Borel function } g. \tag{F.1}$$

In addition to the W_∞ ball of Borel measures $\mathcal{B}^\infty_\epsilon(\mathbb{Q})$ around $\mathbb{Q}$, one can consider the W_∞ ball of universal completions of measures around $\mathbb{Q}$, which we will call $\widetilde{\mathcal{B}}^\infty_\epsilon(\mathbb{Q})$. The following result shows that if $\mathbb{Q}' \in \mathcal{B}^\infty_\epsilon(\mathbb{Q})$, then $W_\infty(\widetilde{\mathbb{Q}}, \widetilde{\mathbb{Q}}') \leq \epsilon$, and thus $\widetilde{\mathcal{B}}^\infty_\epsilon(\mathbb{Q})$ contains only measures that are within ϵ of $\widetilde{\mathbb{Q}}$ in the W_∞ metric.

Lemma F.2. *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* $W_\infty(\widetilde{\mathbb{Q}}, \widetilde{\mathbb{Q}}') \leq \epsilon$.

Explicitly, for a Borel measure $\mathbb{Q}$, let

$$\widetilde{\mathcal{B}}^\infty_\epsilon(\mathbb{Q}) = \{\widetilde{\mathbb{Q}}' : \mathbb{Q}' \in \mathcal{B}^\infty_\epsilon(\mathbb{Q})\}.$$

Next, to compare the values of $\bar{R}$ on $\mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \times \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)$ and $\widetilde{\mathcal{B}}^\infty_\epsilon(\mathbb{P}_0) \times \widetilde{\mathcal{B}}^\infty_\epsilon(\mathbb{P}_1)$, we show:

Corollary F.3. *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 F.2 and Corollary F.3 imply that

$$\sup_{\substack{\mathbb{P}'_0 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_0) \\ \mathbb{P}'_1 \in \mathcal{B}^\infty_\epsilon(\mathbb{P}_1)}} \bar{R}(\mathbb{P}'_0, \mathbb{P}'_1) = \sup_{\substack{\widetilde{\mathbb{P}}'_0 \in \widetilde{\mathcal{B}}^\infty_\epsilon(\mathbb{P}_0) \\ \widetilde{\mathbb{P}}'_1 \in \widetilde{\mathcal{B}}^\infty_\epsilon(\mathbb{P}_1)}} \bar{R}(\widetilde{\mathbb{P}}'_0, \widetilde{\mathbb{P}}'_1) \tag{F.2}$$

See Appendix F.3 for proofs of Lemma F.2 and Corollary F.3.

Then (F.1) and Lemma B.1 imply that

Corollary F.4. *Let $\mathbb{Q}$ be a finite positive measure on $\mathscr{U}(\mathbb{R}^d)$. Then for any universally measurable set A,*

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

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 F.5 (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 F.4 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

$$\inf_{A\in\mathscr{U}(\mathbb{R}^d)} R^\epsilon(A) \geq \sup_{\substack{\widetilde{\mathbb{P}}_0'\in\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0)\\ \widetilde{\mathbb{P}}_1'\in\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \inf_{A\in\mathscr{U}(\mathbb{R}^d)} \int \mathbf{1}_{A^C} d\widetilde{\mathbb{P}}_1' + \int \mathbf{1}_A d\widetilde{\mathbb{P}}_0' = \sup_{\substack{\widetilde{\mathbb{P}}_0'\in\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_0)\\ \widetilde{\mathbb{P}}_1'\in\widetilde{\mathcal{B}}_\epsilon^\infty(\mathbb{P}_1)}} \bar{R}(\widetilde{\mathbb{P}}_0',\widetilde{\mathbb{P}}_1') \qquad \blacksquare$$

This observation suffices to prove Theorem 5.7:

Proof of Theorem 5.7. 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 2.3 and (F.2) 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 in Lemma F.5 implies that the inequality above must actually be an equality. $\blacksquare$

F.3. Proofs of Lemma F.2 and Corollary F.3. 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

$$\Delta_\epsilon = \{(\mathbf{x},\mathbf{y})\in\mathbb{R}^d\times\mathbb{R}^d : \|\mathbf{x}-\mathbf{y}\|\leq\epsilon\}. \tag{F.3}$$

This notation is helpful in the proof of Lemma F.2.

Proof of Lemma F.2. 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 [13]. Let $\overline{\gamma}$ be the completion of γ restricted to $\sigma(\mathscr{U}(\mathbb{R}^d) \times \mathscr{U}(\mathbb{R}^d))$. 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. Then 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

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 $W_\infty(\widetilde{\mathbb{Q}}, \widetilde{\mathbb{Q}}') \leq \epsilon$. ∎

Next, we will show:

Lemma F.6. *Let ν, λ be two Borel measures with $\nu \ll \lambda$, and let $d\nu/d\lambda$ be the Raydon-Nikodym derivative. Then $d\widetilde{\nu}/d\widetilde{\lambda} = d\nu/d\lambda$ $\widetilde{\lambda}$-a.e.*

This result together with (F.1) with immediately implies Corollary F.3.

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 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. ∎

Appendix G. Deferred Proofs From subsection 5.3.

G.1. Proof of Lemma 5.8.

Proof of Lemma 5.8. Let $\{D_i\}_{i=1}^{\infty}$ be a countable sequence of degenerate sets for an adversarial Bayes classifier A. Then by Proposition 5.1, 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 5.1 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. ∎

G.2. Proof of Proposition 5.12.

Lemma G.1. *Let A be an adversarial Bayes classifier. If C is a connected component of A with $C^{-\epsilon} = \emptyset$, then*

$$C^\epsilon = \{\mathbf{y} \in A^C : \overline{B_\epsilon(\mathbf{y})} \text{ intersects } C\}^\epsilon \tag{G.1}$$

If C is a component of A^C with $C^{-\epsilon} = \emptyset$, then

$$C^\epsilon = \{\mathbf{y} \in A : \overline{B_\epsilon(\mathbf{y})} \text{ intersects } C\}^\epsilon \tag{G.2}$$

Proof. We will prove (G.1), the argument for (G.2) is analogous. Assume that C is a component of A, (5.2) implies the containment $\supset$ of (G.1).

Next, we prove the containment $\subset$ in (G.1). 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.

Assume first that $\mathbf{x} \in C$. Because $C^{-\epsilon} = \emptyset$, Equation (5.3) 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. This statement contradicts the fact that C is a connected component of A. Therefore, $\overline{B_\epsilon(\mathbf{x})}$ contains a point $\mathbf{y}$ in A^C, and thus $\mathbf{x} \in \overline{B_\epsilon(\mathbf{y})}$.

Next assume that $\mathbf{x} \in C^\epsilon$ but $\mathbf{x} \notin C$. Then Equation (5.2) states that the ball $\overline{B_\epsilon(\mathbf{x})}$ intersects C at some point $\mathbf{z}$. Consider the line defined by $\ell := \{t\mathbf{x} + (1-t)\mathbf{z} : 0 \le t \le 1\}$. 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 $\mathbf{x} \in \overline{B_\epsilon(\mathbf{y})}$.

Proof of Proposition 5.12. 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 5.11 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, (G.1) of Lemma G.1 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 5.10.

The argument for a connected component of A^C is analogous, with (G.2) in place of (G.1)

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 (5.4) is contained in the degenerate set $(A^\epsilon)^{-\epsilon} - (A^{-\epsilon})^\epsilon$. ∎

G.3. Proof of Lemma 5.13.

Proof of Lemma 5.13. 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 5.1 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, (5.2) and (5.3) imply that

$$\begin{aligned} D^\epsilon - A^\epsilon &= \{\mathbf{x} : \overline{B_\epsilon(\mathbf{x})} \text{ intersects } D \text{ but not } A - D\} \\ &\supset \{\mathbf{x} : \overline{B_\epsilon(\mathbf{x})} \subset D\} = D^{-\epsilon} \end{aligned}$$

∎

Thus $\mathbb{P}_0(D^{-\epsilon}) = 0$.

Appendix H. Deferred proofs from subsection 6.1.

Lemma 5.2 was proved in Appendix C.1.

H.1. Proof of Theorem 3.5.

Proof of Theorem 3.5. Let $\tilde{A}_1 \subset \tilde{A}_2$ be the adversarial Bayes classifiers defined in Lemma 6.1 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

$$\mathbb{R} = D \sqcup \bigcup_{i=m}^{M} (\tilde{a}_i, \tilde{b}_i) \sqcup \bigcup_{i=n}^{N} (\tilde{e}_i, \tilde{f}_i) \tag{H.1}$$

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 $\hat{A} \triangle A \subset D$, the set $\hat{A}$ is still an adversarial Bayes classifier.

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

$$\bigsqcup_{i=k}^{K}(a_i, b_i) = \bigcup_{i=m}^{M}(\hat{a}_i, \hat{b}_i)$$

■

Appendix I. Deferred Proofs from subsection 6.2.

I.1. Proof of Lemma 6.4. First, we show Lemma 6.4 for points near the boundary of $\operatorname{supp}\mathbb{P}$.

Lemma I.1. *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.*

- *Assume that $\operatorname{supp}\mathbb{P} = [\ell, \infty)$ or $\operatorname{supp}\mathbb{P} = [\ell, r]$.*
  If $y \in (\ell - \epsilon, \ell + \epsilon]$ then $[\ell - \epsilon, y]$ is a degenerate set. If furthermore $\operatorname{supp}\mathbb{P}$ is an interval, then for some $\delta > 0$, either $\eta \equiv 0$ or $\eta \equiv 1$ μ-a.e. on $[\ell, \ell + \delta]$.
- *Assume that $\operatorname{supp}\mathbb{P} = (-\infty, r]$ or $\operatorname{supp}\mathbb{P} = [\ell, r]$.*
  If $y \in (r - \epsilon, r + \epsilon]$ then $[y, r - \epsilon]$ is a degenerate set. If furthermore $\operatorname{supp}\mathbb{P}$ is an interval, 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 a_i is in $[\ell - \epsilon, \ell + \epsilon]$, the argument for b_i is analogous. Then because A is adversarial Bayes classifier:

$$0 \geq R^\epsilon(A) - R^\epsilon(A \cup [\ell - \epsilon, a_i]) = \int_\ell^{a_i+\epsilon} p dx - \int_\ell^{a_i+\epsilon} p_0 dx = \int_\ell^{a_i+\epsilon} p_1(x) dx. \tag{I.1}$$

By assumption $a_i > \ell - \epsilon$ and thus $\delta = a_i + \epsilon - \ell > 0$. Hence (I.1) implies that $\eta \equiv 0$ μ-a.e. on $[\ell, \ell + \delta]$.

Next we argue that the set $[\ell - \epsilon, a_i]$ is a degenerate set. Let D be an arbitrary measurable subset of $[\ell - \epsilon, a_i]$. Then

$$R^\epsilon(A \cup D) - R^\epsilon(A \cup [\ell - \epsilon, a_i]) \leq \int_\ell^{a_i+\epsilon} p dx - \int_\ell^{a_i+\epsilon} p_0 dx = \int_0^{a_i+\epsilon} p_1(x) dx$$

and this quantity must be zero by (I.1). ■

Proof of Lemma 6.4. Assume that the endpoints of I are d_1, d_2, so that $I = [d_1, d_2]$. 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 5.8. The interval J must be closed because the boundary of every adversarial Bayes classifier is a degenerate set when $\epsilon \ll \mu$. If $J \cap (\operatorname{supp}\mathbb{P}^\epsilon - \operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon})$ is nonempty, Lemma I.1 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 6.3 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 5.8 and Lemma 5.14 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. If this interval contained both points in A and A^C, this $[d_3 - \epsilon, d_3)$ would also be degenerate by Proposition 5.12. 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 6.3, 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 $[d_1 - \epsilon, d_2 + \epsilon]$, one can conclude that $p_1(x) = 0$ μ-a.e. As $[d_1, d_2] \subset \operatorname{supp}\mathbb{P}^{\epsilon}$ 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 6.3 implies that $|I| \leq 2\epsilon$. Then

$$0 = R^{\epsilon}(A \cup (J \cap \mathbb{Q})) - R^{\epsilon}(A \cup J) \geq \int_{d_3-\epsilon}^{d_4+\epsilon} p(x)dx - \left(\int_{d_3-\epsilon}^{d_4-\epsilon} p_0(x)dx + \int_{d_3-\epsilon}^{d_4+\epsilon} p(x)dx \right) \geq \int_{d_1-\epsilon}^{d_2-\epsilon} p_1(x)dx$$

Thus $p_1(x) = 0$ on $[d_1 - \epsilon, d_2 - \epsilon]$. Similarly, by considering $R^{\epsilon}(A \cup (J \cap \mathbb{Q})) - R^{\epsilon}(A - J)$, one can argue that $p_0(x) = 0$ on $[d_1 + \epsilon, d_2 + \epsilon]$.

Now if $[d_1, d_2] \subset \operatorname{supp}\mathbb{P}^{\epsilon}$, then at least one of $[d_1 - \epsilon, d_2 - \epsilon]$, $[d_1 + \epsilon, d_2 + \epsilon]$ intersects $\operatorname{supp}\mathbb{P}$ on an open interval. Thus either $\eta(x) =$ or $\eta(x) = 0$ on a set of positive measure.

I.2. Proof of the fourth bullet of Theorem 3.8. The following lemma implies $(\operatorname{supp}\mathbb{P}^{\epsilon})^C$ is a degenerate set.

Lemma I.2. *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 is a $\mathbf{a} \in A$, $\mathbf{b} \in B$ and $\mathbf{x} \in \overline{B_{\epsilon}(\mathbf{0})}$ 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 C.1 of Appendix C.1.

Lemma I.3. *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 I.2 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 C.1 implies that $\partial \operatorname{supp} \mathbb{P}^\epsilon)^C)^\epsilon$ has Lebesgue measure zero. Lemma 5.2 implies that ∂A is a degenerate set. Lastly, Lemma 5.8 implies that the union of these three sets 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 3.8 for regular adversarial Bayes classifiers.

Lemma I.4. *Assume that $\mathbb{P} \ll \mu$ 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 5.8 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 5.14 implies that $[x, y] \subset ((D \cup \partial A)^\epsilon)^{-\epsilon}$ is a degenerate set as well. This statement contradicts Lemma 6.4. 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 $\eta > 0$ 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 I.4 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 3.8. Any adversarial Bayes classifier A is equivalent up to degeneracy to a regular adversarial Bayes classifier A_r. Lemma I.4 implies that if D a degenerate set for A, then $D \subset \overline{(\operatorname{supp} \mathbb{P}^\epsilon)^C} \cup \partial A_r$. Let $\tilde{A}$ be any adversarial Bayes classifier equivalent up to degeneracy to A (and A_r).

We will show that $\partial A_r \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon = \partial \tilde{A} \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon$, and this statement together with Lemma I.3 will imply the desired result.

As $\partial \operatorname{int} A_r \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon = \partial \overline{A_r} \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon$ and

$$\operatorname{int} A_r \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon \subset \tilde{A} \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon \subset \overline{A_r} \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon,$$

it follows that $\partial \tilde{A} \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon = \partial A_r \cap \operatorname{int} \operatorname{supp} \mathbb{P}^\epsilon$. ∎

Appendix J. Deferred Proofs from subsection 6.3. In this appendix, we adopt the same notational convention as subsection 6.3 regarding the a_is and b_is: Namely, when $A = \bigcup_{i=m}^{M}$

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 6.6.

Lemma J.1. *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 values of

$$R^{\epsilon}(A) = \int_{\mathbb{R}} d\mathbb{P}_0 \quad R^{\epsilon}(\emptyset) = \int_{\mathbb{R}} d\mathbb{P}_1$$

are independent of the value of ϵ. Next, notice that $R^{\epsilon_2}(A) \geq R^{\epsilon_1}(A)$ for an set A. Therefore,

$$R^{\epsilon_2}_* \geq R^{\epsilon_1}_* = R^{\epsilon_1}(\emptyset) = R^{\epsilon_2}(\emptyset)$$

and 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

$$R^{\epsilon_1}(\mathbb{R}) = R^{\epsilon_2}(\emptyset) = R^{\epsilon_2}(\mathbb{R}) = R^{\epsilon_1}(\emptyset).$$

Consequently, $\mathbb{R}$ is also a minimizer of R^{ϵ_1}. ∎

Next, recall that Lemma I.1 implies that if the an endpoint of an adversarial Bayes classifier are 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 J.2. *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 $I^{\epsilon} - \operatorname{int} I^{-\epsilon}$.*

[todo: a_{i+1}, b_i conventions in the next lemma and the proof of lemma 5.6]

Next, proving Lemma 6.6 is simpler when $(a^2_j, b^2_j) \subset (b^1_i, a^1_{i+1})$ or $(b^2_j, a^2_{j+1}) \subset (a^1_i, b^1_i)$. The following lemma will allow us to always reduce to this scenario.

Proof of Lemma 6.6. We will show that $(b^1_i, a^1_{i+1}) \cap I^{\epsilon_1}$ does not include $(a^2_j, b^2_j) \cap I^{\epsilon_1}$, the argument for $(a^1_i, b^1_i) \cap I^{\epsilon_1}$ and $(a^2_j, b^2_{j+1}) \cap I^{\epsilon_1}$ is analogous. Fix an interval (a^2_j, b^2_j) and for contradiction, assume that $(a^2_j, b^2_j) \cap I^{\epsilon_1} \neq \emptyset$ and $(a^2_j, b^2_j) \cap I^{\epsilon_1} \subset (b^1_i, a^1_{i+1}) \cap I^{\epsilon_1}$.

First, notice that the assumption $\eta \neq 0, 1$ implies that none of the a^2_js, b^2_js are in $I^{\epsilon_2} - I^{-\epsilon_2}$ due to Corollary J.2. Thus if the intersection $(a^2_j, b^2_j) \cap I^{\epsilon_1}$ is non-empty, then either $I^{\epsilon_2} \subset (a^2_j, b^2_j)$ or at least one endpoint of (a^2_j, b^2_j) is in $I^{-\epsilon_2}$.

If in fact $(a^2_j, b^2_j) \supset I^{\epsilon_2}$, then (b^1_{i+1}, a_i) 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 J.1 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 6.6.

Thus, to show the second bullet, one can assume that $(a^2_j, b^2_j) \not\supset I^{\epsilon_2}$. As $b^2_j - a^2_j > 2\epsilon_2$ and the interval (a^2_j, b^2_j) is included in the adversarial Bayes classifier A_2, it follows that $R^{\epsilon}(A_2) \leq R^{\epsilon}(A_2 - (a^2_j, b^2_j))$ which implies

$$\int_{a^2_j - \epsilon_2}^{a^2_j + \epsilon_2} p dx + \int_{a^2_j + \epsilon_2}^{b^2_j - \epsilon_2} p_0 dx + \int_{b^2_j - \epsilon_2}^{b^2_j + \epsilon_2} p dx \leq \int_{a^2_j - \epsilon_2}^{b^2_j + \epsilon_2} p_1 dx$$

Which in turn implies

$$\int_{a_j^2-\epsilon_2}^{b_j^2+\epsilon_2} p_0 dx \leq \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_1 \tag{J.1}$$

Next, $b_j^2 - a_j^2 > 2\epsilon_2$ because A_2 is regular 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,

$$R^{\epsilon_1}(A_1) - R^{\epsilon_1}(A_1 \cup (a_j^2 + \epsilon_2 - \epsilon_1, b_j^2 - (\epsilon_2 - \epsilon_1))) = \int_{c_{i,j}}^{d_{i,j}} p_1 dx - \left(\int_{c_{i,j}}^{a_j^2+\epsilon_2} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx + \int_{b_j^2-\epsilon_2}^{d_{i,j}} p dx \right)$$

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 (J.1) implies that

$$\begin{aligned} \int_{c_{i,j}}^{d_{i,j}} p_1 dx &\geq \int_{c_{i,j}}^{a_j^2+\epsilon_2} p dx + \int_{b_j^2-\epsilon_2}^{d_{i,j}} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx + \int_{a_j^2-\epsilon_2}^{c_{i,j}} p_0 dx + \int_{d_{i,j}}^{b_j^2+\epsilon_2} p_0 dx \\ &> \int_{c_{i,j}}^{a_j^2+\epsilon_2} p dx + \int_{b_j^2-\epsilon_2}^{d_{i,j}} p dx + \int_{a_j^2+\epsilon_2}^{b_j^2-\epsilon_2} p_0 dx \end{aligned}$$

The last inequality follows because $c_{i,j} - (a_j^2 - \epsilon_2) \geq 2(\epsilon_2 - \epsilon_1) > 0$, $b_j^2 + \epsilon_2 - d_{i,j} \geq 2(\epsilon_2 - \epsilon_1) > 0$, $z \in \operatorname{int} I^{-\epsilon_2}$ implies that $z + \epsilon, z - \epsilon \operatorname{int} I$, and $(a_j^2, b_j^2 \not\supset I^{\epsilon_2}$ together with Corollary J.2 imply that $\operatorname{int} \operatorname{supp} \mathbb{P}^{\epsilon_2}$ must include at least one of a_j^2, b_j^2. This inequality would imply that $R^{\epsilon_1}(A_1) - R^{\epsilon_1}(A \cup (a_j^2 + \epsilon_2 - \epsilon_1, b_j^2 - (\epsilon_2 - \epsilon_1)) > 0$, which contradict the fact that A is an adversarial Bayes classifier.

Theorem 3.9 then directly follows from Lemma 6.6.

Proof of Theorem 3.9. The first bullet of Lemma 6.6 together with the fourth bullet of Theorem 3.8 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 J.2 implies that there are no $a_j^2, b_j^2 \in I^{\epsilon_2} - I^{-\epsilon_2}$. As $I^{-\epsilon_2} \subset I^{\epsilon_2} \subset I^{\epsilon_2}$ 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 $\text{comp}(A_2^C \cap I^{\epsilon_2} = \text{comp}(A_2^C \cap I^{\epsilon_1}$. Therefore, it remains to show that $\text{comp}(A_1 \cap I^{\epsilon_1}) \geq \text{comp}(A_2 \cap I^{\epsilon_1})$ 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 $(a_k^c, b_k^c) \cap I^{\epsilon_1}$, $(b_k^c, a_{k+1}^c) \cap I^{\epsilon_1}$ are intervals for $c = 1, k \in [m, M]$ and $c = 2, k \in [n, N]$. 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 6.6. 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 $\text{comp}(A_1 \cap I^{\epsilon_1}) \geq \text{comp}(A_2 \cap I^{\epsilon_1}) = \text{comp}(A_1 \cap I^{\epsilon_2})$.

Appendix K. Computational Details of Examples in section 4. The following lemma is helpful for verifying the second order necessary conditions for gaussian mixtures.

Lemma K.1. *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

$$g'(x) = -\frac{x-\mu}{\sigma^2} \cdot \frac{t}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = -\frac{x-\mu}{\sigma^2} g(x)$$

K.1. Further details from Example 4.1. It remains to verify two of the claims made in Example 4.1— namely, 1) that $b(\epsilon)$ does not satisfy the second order necessary condition (3.1b), 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 (3.1b). Due to Lemma K.1 the equation (3.1b) 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

$$\int_{-\infty}^{a(\epsilon)+\epsilon} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx \leq \int_{-\infty}^{a(\epsilon)-\epsilon} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}$$

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 $\epsilon \geq \frac{\mu_1-\mu_0}{2}$.

K.2. Further details of Example 4.2. 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 (2.8b) 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 (4.1) and

$$y(\epsilon) = \frac{\epsilon\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) - \sqrt{\frac{4\epsilon^2}{\sigma_0^4\sigma_1^4} - 2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)k}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}}.$$

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 (3.1b) (see Appendix K.2). Similarly, the only solution to the necessary conditions (2.8a) and (3.1a) 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 3.8 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 K.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 (3.1b), 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 (3.1b). First, notice that

$$y(\epsilon) \leq \frac{\epsilon\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_0^2}\right)-\sqrt{\frac{4\epsilon^2}{\sigma_0^4\sigma_1^4}}}{\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}} = \frac{\epsilon\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_0^2}\right)-\frac{2\epsilon}{\sigma_0\sigma_1}}{\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}} \tag{K.1}$$

This bound shows that $y(\epsilon)$ fails to satisfy the second order necessary condition (3.1b). One can compute the derivative p_i' in terms of p_i using Lemma K.1. 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 $p_0(y(\epsilon)+\epsilon) - p_1(y(\epsilon)-\epsilon) = 0$ together implies

$$p_0'(y(\epsilon)+\epsilon) - p_1'(y(\epsilon)-\epsilon) = \frac{p_0(y(\epsilon)+\epsilon)}{\sqrt{2\pi}}\left(y(\epsilon)\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)-\epsilon\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_0^2}\right)\right)$$

However, (K.1) implies that

$$\frac{p_0(y(\epsilon)+\epsilon)}{\sqrt{2\pi}}\left(y(\epsilon)\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)-\epsilon\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_0^2}\right)\right) \leq \frac{p_0(y(\epsilon)+\epsilon)}{\sqrt{2\pi}}\cdot\frac{-2\epsilon}{\sigma_0\sigma_1} < 0$$

Thus, the only solution to first (2.8b) and (3.1b) 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 $R^\epsilon((\infty, b(\epsilon)) > R^\epsilon((a(\epsilon), b(\epsilon))$:

$$R^\epsilon((a(\epsilon),\infty)) - R^\epsilon((a(\epsilon),b(\epsilon))) = \int_{b(\epsilon)-\epsilon}^{b(\epsilon)+\epsilon} p_1(x)+p_0(x)dx + \int_{b(\epsilon)+\epsilon}^{+\infty} p_0(x)-p_1(x)dx \tag{K.2}$$

First, notice that the first integral is always positive. Next, because $b(\epsilon)+\epsilon > b(0)$ and $p_0(x) > p_1(x)$ whenever $x > b(0)$, the second integral must always be positive as well. Therefore, $R^\epsilon((-\infty, b(\epsilon)) > R^\epsilon((a(\epsilon), b(\epsilon))$.

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 (K.2) one can show that $R^\epsilon(\mathbb{R}) - R^\epsilon((a(\epsilon), +\infty)) = R^\epsilon((-\infty, b(\epsilon))) - R^\epsilon((a(\epsilon), b(\epsilon)))$. 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 $R^\epsilon(\emptyset) > R^\epsilon((a(\epsilon), b(\epsilon)))$.

K.3. Proof of Lemma 4.3. Lemma I.1 of Appendix I.1 is helpful in proving Lemma 4.3.

Proof of Lemma 4.3. 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 two cases:

A) $|\ell - a_{i^*}| \leq |\ell - b_{j^*}|$, in which case we will show $A' = (-\infty, a_{i^*}) \cup A$ is the desired adversarial Bayes classifier

B) $|\ell - a_{i^*}| > |\ell - b_{j^*}|$, in which case we will show $A' = (-\infty, b_{j^*})^C \cap A$ is the desired adversarial Bayes classifier ∎

We will show Item A), the argument for Item B) is analogous.

First, Lemma I.1 implies that A and A' are equivalent up to degeneracy.

Next, we show that $A' := A \cup (-\infty, a_{i^*}]$ is a regular set. Because A is regular, the point a_{i^*} is more than 2ϵ from any other boundary point of ∂A. As $\partial(A \cup (-\infty, a_{i^*}]) \subset \partial A \cup \partial(-\infty, a_{i^*}) = \partial A$, the point a_i^* must be more than 2ϵ from any other boundary point of $(-\infty, a_{i^*}] \cup A$. Therefore, A' is regular.

Lastly, to show that $\partial A' \subset \operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon}$, we argue that A' has no boundary points in $(-\infty, \ell + \epsilon] = (\operatorname{int}\operatorname{supp}\mathbb{P}^{-\epsilon})^C$. First, 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.

K.4. Example 4.5 details. Theorem 3.7 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 4.3 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 $\epsilon \in [1/3, 1/2)$. Thus Theorem 3.9 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 $\epsilon \geq 1/3$.

K.5. Example 4.6 details. It remains to compare the adversarial risks of all sets whose boundary is included in $\{-1/4 \pm \epsilon, 1/4 \pm\}$ 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

$$R^\epsilon(A) \geq \frac{2\epsilon}{5} > \frac{1}{10} = R^\epsilon(\mathbb{R})$$

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

$$R^\epsilon(A) \geq \frac{3}{5}\cdot 2\epsilon \geq \frac{3}{20} > R^\epsilon(\mathbb{R}).$$

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

$$R^\epsilon(A) \geq \frac{3}{5}\cdot\frac{3}{4} > R^\epsilon(\mathbb{R}).$$

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 $\{1/4-\epsilon\}$, but 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 $\epsilon > 1/8$. 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.

K.6. Proof of Proposition 4.9.

Proof of Proposition 4.9. Due to Theorem 3.5 and Lemma 4.3, 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 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 (2.8) reduce to

(K.3a) $\eta(a+\epsilon) = 1-\eta(a-\epsilon)$ and (K.3b) $\eta(b-\epsilon) = 1-\eta(b+\epsilon)$

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 (K.3a). 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 (K.3b). ■

K.7. Proof of Proposition 4.10.

Proof of Proposition 4.10. Due to Theorem 3.5 and Lemma 4.3, 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 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 $(a(\epsilon) - \epsilon - r, a(\epsilon) + \epsilon + r)$ 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. Consequently, $a(\epsilon)$ cannot satisfy the necessary condition (2.8a), thus contradicting Theorem 3.7. ■

K.8. Example 6.5 details. It remains to compare the risks of all regular sets with endpoints in $\{-4\epsilon, -3\epsilon, -2\epsilon, -\epsilon, 0, \epsilon, 2\epsilon, 3\epsilon, 4\epsilon\}$, 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 I.1 (of Appendix I.1) implies that if A is a regular adversarial Bayes classifier and $y \in \{-4\epsilon, -3\epsilon, -2\epsilon\}$ is in ∂A, then $[-4\epsilon, y]$ is a degenerate set. Thus there is no need to consider classifiers with endpoints in $\{-4\epsilon, -3\epsilon, -2\epsilon\}$ when identifying all possible adversarial Bayes classifiers under equivalence up to degeneracy. Similarly, Lemma I.1 also implies that there is no need to consider $\{2\epsilon, 3\epsilon, 4\epsilon\}$ 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 $\{-\epsilon, 0, \epsilon\}$. As points in the boundary of a regular set are at most 2ϵ apart, one can rule out sets with more than one boundary point in $\{-\epsilon, 0, \epsilon\}$.

At the same time, if ∂A includes exactly one of $\{-\epsilon, 0, \epsilon\}$, then A fails to include either $(-\infty, -\epsilon)$ or (ϵ, ∞). Furthermore, the classifier pays a penalty of $p_0 + p_1$ on $(A^C)^{\epsilon} \cap A^{\epsilon}$, which must include either $(-\epsilon, 0)$ or $(0, \epsilon)$. Then

$$R^{\epsilon}(A) \geq \frac{1}{4} + \left(\frac{1}{12} + \frac{1}{18}\right) = \frac{1}{3} + \frac{1}{18}$$

while $R^{\epsilon}(\mathbb{R}) = 11/36 < 1/3$. Thus any set A for which ∂A includes exactly one of $\{-\epsilon, 0, \epsilon\}$ cannot be an adversarial Bayes classifier. It remains to compare $R^{\epsilon}(\mathbb{R}) = 11/36$ and $R^{\epsilon}(\emptyset) = 20/36$. Thus $\mathbb{R}$ is an adversarial Bayes classifier, and is unique up to degeneracy.