CLASSIFICATION WITH ABSTENTION UNDER CLASS-CONDITIONAL ERROR CONSTRAINTS: MINIMAX RATE AND ALGORITHM

Anonymous authors
Paper under double-blind review

ABSTRACT

We study binary classification with abstention under separate class-conditional error constraints, with the objective of minimizing abstention while keeping both errors below prescribed thresholds. We characterize the distribution-free minimax rate of excess abstention risk, up to logarithmic factors, in terms of the complexity of the hypothesis class and the sample size. To make the framework amenable to computation with models such as neural networks, we introduce surrogate-loss formulations and derive finite-sample guarantees for excess surrogate ambiguity risk. We formulate the resulting learning task as a constrained optimization problem and characterize its computational complexity in the convex setting. Finally, we evaluate our approach on various datasets and compare its performance with a competing method for this problem.

1 INTRODUCTION

In many decision-making problems, misclassification can have serious consequences. In such applications, it is often preferable for a classifier to abstain when the available evidence is insufficient to make a prediction (El-Yaniv & Wiener, 2010; Mozannar & Sontag, 2020). A medical diagnostic system, for example, may confidently identify some patients as healthy or at high risk, while referring uncertain cases to a specialist for further examination (Vazhentsev et al., 2026; Das et al., 2019; Hamid et al., 2017; da Rocha Neto et al., 2011). Similarly, in fraud detection, a system may automatically approve clearly legitimate transactions, block highly suspicious ones, and send ambiguous cases for manual review (Singh et al., 2026; Alves et al., 2023). Abstention allows a classification system to avoid unreliable decisions, but abstaining too often limits its practical usefulness. The central challenge is therefore to minimize abstention while controlling classification errors.

In many such applications, the two types of classification errors have different consequences, making it important to control them separately. In medical diagnosis, for instance, missing a high-risk patient may have severe health consequences, while incorrectly flagging a healthy patient may lead to unnecessary tests or interventions. Separate control is also important when the classes are highly imbalanced or one class is scarce, since a small overall classification error does not necessarily imply a small error within each class. One common approach is to constrain the error associated with one class below a prescribed level while optimizing performance on the other, as in Neyman–Pearson classification (Scott & Nowak, 2005; Rigollet & Tong, 2011; Tong, 2013; Tong et al., 2020; Kalan et al., 2025b). In applications where both types of errors are critical, it is natural to instead require both class-conditional errors to remain below prescribed thresholds. However, because the two errors trade off against each other, these requirements may not be simultaneously achievable when a prediction is made for every observation. Abstention can then make it possible to satisfy both error constraints.

In this work, we study the problem of minimizing abstention while requiring the two class-conditional errors to remain below given levels α_0 and α_1. Given a finite labeled sample and a hypothesis class, the goal is to learn class-specific decision rules whose prediction regions capture sufficient probability mass under the corresponding class distributions to keep each class-conditional error below its prescribed threshold. These two regions may overlap, and their intersection forms the ambiguity region on which the classifier abstains. Thus, minimizing abstention amounts to min-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

imizing how often observations fall in the ambiguity region, subject to keeping the class-conditional errors below the thresholds.

We first study the *distribution-free statistical complexity* of this problem under the 0-1 loss formulation and characterize its minimax rate up to logarithmic factors. For a hypothesis class with finite VC dimension, we establish finite-sample upper bounds based on empirical risk minimization and prove matching minimax lower bounds. We then introduce a surrogate-loss formulation that is more amenable to computation. For function classes with bounded Rademacher complexity, we derive finite-sample statistical guarantees for both the class-conditional constraints and the resulting abstention performance. The surrogate formulation leads to a constrained optimization problem, which we solve using a stochastic primal-dual algorithm (Mahdavi et al., 2012) and characterize its computational complexity in the convex setting. Finally, we evaluate our approach on synthetic and real-world datasets and compare its performance with the plug-in method of Lei (2014).

2 Related Work

Lei (2014) introduced a closely related framework, which considers classification under two class-conditional error constraints and minimizes the probability of the ambiguous region, on which the classifier abstains from making a prediction. Under suitable regularity conditions, Lei (2014) characterizes the universally population-optimal classification regions (i.e., not restricted to a hypothesis class) through an extension of the Neyman–Pearson lemma (Lehmann, 1986, Theorem 3.2.1), and develops plug-in estimation procedures based on estimates of the conditional class probabilities, together with convergence guarantees under suitable regularity and margin conditions. In contrast, we study this constrained ambiguity-minimization problem in a distribution-free setting over a prescribed hypothesis class. In particular, we characterize the statistical minimax rate in terms of the complexity of the hypothesis class by establishing finite-sample upper bounds and matching lower bounds. We further develop a score-function formulation with surrogate losses and derive finite-sample statistical guarantees for function classes with bounded Rademacher complexity, making the framework directly applicable to modern learning models, including neural networks. We also provide a practical constrained optimization procedure for training such models. Finally, our empirical study directly compares the two approaches and demonstrates improved performance across a collection of real and synthetic datasets.

Other constrained formulations of classification with abstention, or rejection, have also been studied (Pietraszek, 2005; 2007; Li & Sethi, 2006; Geifman & El-Yaniv, 2017; 2019; Denis & Hebiri, 2020; Shekhar et al., 2019; Gangrade et al., 2021). Among the formulations most closely related to ours, Hanczar & Dougherty (2008), in a setting closely related to that of Lei (2014), minimize the rejection rate subject to class-specific error constraints conditional on non-rejection. Their analysis, however, assumes a fixed classifier score and optimizes only the two rejection thresholds, without providing statistical guarantees that the resulting procedure converges to a population-optimal classifier. El-Yaniv & Wiener (2010) study the risk–rejection trade-off in the noise-free realizable setting, deriving bounds on achievable trade-offs with particular emphasis on minimizing rejection under zero classification error among non-rejected predictions. Relatedly, the bounded-improvement formulation, which maximizes the probability of making a non-abstained prediction subject to a prescribed threshold on the total classification error among non-abstained predictions, and the bounded-abstention formulation, which minimizes this conditional classification error subject to a prescribed threshold on the probability of making a non-abstained prediction, were considered by Pietraszek (2005) and further studied by Franc et al. (2023). Franc et al. (2023) characterize the corresponding population-optimal strategies in terms of the Bayes classifier and an appropriate, potentially randomized, abstention rule, without studying the sample complexity of these formulations.

Another widely studied formulation of classification with abstention is the cost-based approach, in which abstention is assigned a fixed cost and the objective is to minimize the expected cost of misclassification and abstention (Chow, 1970; Herbei & Wegkamp, 2006; Bartlett & Wegkamp, 2008; Yuan & Wegkamp, 2010; Ramaswamy et al., 2015). The classical formulation goes back to Chow (1970), who characterized the Bayes-optimal rejection rule obtained by trading off the cost of misclassification against the cost of rejection. Herbei & Wegkamp (2006) studied statistical properties of plug-in and empirical risk minimization procedures under the classical cost-based reject-option formulation of Chow (1970). Bartlett & Wegkamp (2008) introduced a hinge-type convex surrogate

loss for this formulation and established consistency and excess-risk guarantees. Yuan & Wegkamp (2010) provided a more general analysis of convex surrogate losses, giving conditions for consistency and bounds relating excess surrogate risk to excess reject-option risk. In contrast to our formulation, these cost-based approaches control the trade-off between overall classification error and abstention through a single penalized objective, rather than imposing separate class-conditional error constraints; moreover, although some establish excess-risk convergence rates, they do not derive matching minimax upper and lower bounds.

Finally, our formulation is also related to Neyman–Pearson classification, which addresses asymmetric class-conditional errors by controlling one error while minimizing the other (Scott & Nowak, 2005; Rigollet & Tong, 2011; Tong, 2013; Zhao et al., 2016; Tong et al., 2018; 2020; Kalan & Kpotufe, 2024; Mousavi Kalan & Kpotufe, 2024; Kalan et al., 2025a; Mousavi Kalan et al., 2026). Without abstention, the two class-conditional errors trade off against each other and therefore cannot in general be simultaneously controlled at arbitrarily small levels. Our formulation allows both errors to be controlled, with abstention serving as the price for this simultaneous control.

3 Problem Setting

Let $(\mathcal{X}, \Sigma)$ be a measurable feature space, and let (X, Y) be a random pair taking values in $\mathcal{X} \times \{0, 1\}$ with joint probability distribution P. We denote by P_X the marginal distribution of X. For each $j \in \{0, 1\}$, let $\pi_j := \mathbb{P}(Y = j)$, and assume that $\pi_j > 0$. Let P_j denote the class-conditional distribution of X given $Y = j$.

Let $\mathcal{H}$ be a class of measurable functions $h : \mathcal{X} \to \{0, 1\}$. We choose a pair $(h_0, h_1) \in \mathcal{H} \times \mathcal{H}$ and define

$$C_j(h_j) := \{x \in \mathcal{X} : h_j(x) = 1\}, \qquad j \in \{0, 1\}.$$

The sets $C_0(h_0)$ and $C_1(h_1)$ are viewed as candidate regions for classifying points as class 0 and class 1, respectively. Points in $C_0(h_0) \setminus C_1(h_1)$ are classified as class 0, and points in $C_1(h_1) \setminus C_0(h_0)$ are classified as class 1. On the intersection $C_0(h_0) \cap C_1(h_1)$, the procedure abstains from classification; we call this intersection the ambiguous region. For $j \in \{0, 1\}$, define the class-conditional error

$$R_j(h_j) := P_j(h_j(X) = 0).$$

Thus, $R_j(h_j)$ is the probability that the class-j candidate rule fails to include a point drawn from class j. For user-specified thresholds $\alpha_0, \alpha_1 \in (0, 1)$, the goal is to choose $(h_0, h_1) \in \mathcal{H} \times \mathcal{H}$ such that the class-conditional errors for classes 0 and 1 are at most α_0 and α_1, while minimizing the probability of the ambiguous region, denoted by

$$R_{\mathrm{amb}}(h_0, h_1) := P_X\big(h_0(X) = 1,\ h_1(X) = 1\big).$$

This leads to the population problem

$$\underset{(h_0,h_1)\in\mathcal{H}\times\mathcal{H}}{\text{minimize}} \quad R_{\mathrm{amb}}(h_0, h_1) \quad \text{subject to} \qquad R_0(h_0) \le \alpha_0, \qquad R_1(h_1) \le \alpha_1. \tag{1}$$

We assume throughout that the feasible set of equation 1 is nonempty. We denote a solution to equation 1 by (h_0^*, h_1^*) and its optimal value by R_{amb}^*. The formulation in equation 1 does not require $C_0(h_0) \cup C_1(h_1) = \mathcal{X}$. This does not create an additional undecided region: by convention, points in $\mathcal{X} \setminus (C_0(h_0) \cup C_1(h_1))$ may be assigned to class 0. This gives the rule

$$h_0'(x) := \mathbf{1}\{h_0(x) = 1 \text{ or } (h_0(x) = 0,\ h_1(x) = 0)\},$$

with h_1 unchanged. This modification does not change the ambiguity risk, and the class-conditional constraints remain satisfied. In proper learning, the final decision rule is required to belong to the hypothesis class $\mathcal{H}$; here, h_0' need not belong to $\mathcal{H}$. We therefore allow improper learning.

Given n i.i.d. samples $S^n = \{(X_i, Y_i)\}_{i=1}^n$, where $(X_i, Y_i) \sim P$, the goal is to learn $(\widehat{h}_0, \widehat{h}_1) \in \mathcal{H} \times \mathcal{H}$ whose ambiguity risk is close to the optimal value R_{amb}^* while satisfying the class-conditional constraints. We measure its statistical performance by the **excess ambiguity risk**

$$\mathcal{E}_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) := R_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) - R_{\mathrm{amb}}^*. \tag{2}$$

4 MINIMAX RATES

In this section, we characterize the statistical complexity of equation 1. We first prove an upper bound for an empirical constrained learner and then establish a matching minimax lower bound.

4.1 EMPIRICAL CONSTRAINED LEARNER AND UPPER BOUND

Let $n_j := \sum_{i=1}^n \mathbf{1}\{Y_i = j\}$ denote the number of samples from class j, for $j \in \{0,1\}$. For $h \in \mathcal{H}$, define the empirical class-conditional errors

$$\widehat{R}_j(h) := \frac{1}{n_j} \sum_{i:Y_i=j} \mathbf{1}\{h(X_i) = 0\}, \qquad j \in \{0,1\},$$

whenever $n_j > 0$. Similarly, for a pair $(h_0, h_1) \in \mathcal{H} \times \mathcal{H}$, define the empirical ambiguity risk

$$\widehat{R}_{\mathrm{amb}}(h_0, h_1) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h_0(X_i) = 1,\ h_1(X_i) = 1\}.$$

Let $\epsilon_0, \epsilon_1 > 0$ be slack parameters to be determined. The empirical constrained learner is defined as any solution $(\widehat{h}_0, \widehat{h}_1) \in \mathcal{H} \times \mathcal{H}$ to

$$\underset{(h_0,h_1)\in\mathcal{H}\times\mathcal{H}}{\text{minimize}} \quad \widehat{R}_{\mathrm{amb}}(h_0, h_1) \quad \text{subject to} \qquad \widehat{R}_0(h_0) \le \alpha_0 + \epsilon_0, \qquad \widehat{R}_1(h_1) \le \alpha_1 + \epsilon_1. \tag{3}$$

Theorem 1 (Upper bound). *Assume that $\mathcal{H}$ has finite VC dimension $d_{\mathcal{H}}$. Let $\delta \in (0,1)$ and define the event $\mathcal{E}_n := \{n_0 > 0,\ n_1 > 0\}$. On the event $\mathcal{E}_n$, for $j \in \{0,1\}$, set $\epsilon_j(n_j, \delta) := C\sqrt{\{d_{\mathcal{H}} \log(en_j) + \log(1/\delta)\}/n_j}$, where $C > 0$ is a universal constant. Let $(\widehat{h}_0, \widehat{h}_1)$ be any solution to equation 3 with slack parameters $\epsilon_0/2$ and $\epsilon_1/2$. Then, conditional on $\mathcal{E}_n$, with probability at least $1 - 2\delta - \frac{\delta}{\mathbb{P}(\mathcal{E}_n)}$,*

$$R_j(\widehat{h}_j) \le \alpha_j + \epsilon_j(n_j, \delta), \text{ for } j \in \{0,1\}, \quad \text{and} \quad \mathcal{E}_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) \le C\sqrt{\frac{d_{\mathcal{H}} \log(en) + \log(1/\delta)}{n}}.$$

for a universal constant $C > 0$.

The preceding upper bound relies on positive slack in the class-conditional constraints. This relaxation is unavoidable for deterministic learners: even with complete knowledge of the underlying distribution, imposing the strict constraints $R_0(h_0) < \alpha_0$ and $R_1(h_1) < \alpha_1$ may force every feasible pair to have maximal ambiguity excess risk. The following proposition formalizes this impossibility.

Proposition 1. *Fix $\alpha_0, \alpha_1 \in (0, 1/2)$. There exist a domain $\mathcal{X}$, a hypothesis class $\mathcal{H}$ with $d_{\mathcal{H}} > 0$, and a distribution P with $\pi_0 = \pi_1 = 1/2$ such that every pair $(h_0, h_1) \in \mathcal{H} \times \mathcal{H}$ satisfying $R_0(h_0) < \alpha_0$ and $R_1(h_1) < \alpha_1$ has ambiguity excess risk $\mathcal{E}_{\mathrm{amb}}(h_0, h_1) = 1$.*

We next define a randomized learning rule that satisfies the class-conditional constraints without slack. Let $(\widehat{h}_0, \widehat{h}_1)$ be the empirical hypotheses obtained by the procedure in equation 3. Let $\mathbf{1}$ denote the classifier that is identically equal to one on $\mathcal{X}$. On the event $\mathcal{E}_n = \{n_0 > 0, n_1 > 0\}$, define $q_j := \frac{\alpha_j}{\alpha_j + \epsilon_j(n_j,\delta)}$, $j \in \{0,1\}$. Given the data S^n, independently draw $B_j \sim \mathrm{Bernoulli}(q_j)$, and define the randomized classifier

$$\widetilde{h}_j := B_j \widehat{h}_j + (1 - B_j)\mathbf{1}, \qquad j \in \{0,1\}. \tag{4}$$

Let $\mathbb{E}_\zeta$ denote denote expectation with respect to the auxiliary random variables $\zeta = (B_0, B_1)$. Then, define

$$\overline{R}_j(\widetilde{h}_j) := \mathbb{E}_\zeta\left[R_j(\widetilde{h}_j)\right], \ \text{ for } j \in \{0,1\}, \qquad \text{and} \qquad \overline{R}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) := \mathbb{E}_\zeta\left[R_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1)\right].$$

Theorem 2 (Upper bound for the randomized learner). *Assume that $\mathcal{H}$ has finite VC dimension $d_{\mathcal{H}}$, and let $\delta \in (0,1)$. On the event $\mathcal{E}_n$, let $\epsilon_j(n_j, \delta)$ be as in Theorem 1. Let $(\widehat{h}_0, \widehat{h}_1)$ be any solution of equation 3 with slack parameters $\epsilon_0/2$ and $\epsilon_1/2$, and let $(\widetilde{h}_0, \widetilde{h}_1)$ be the randomized learner defined*

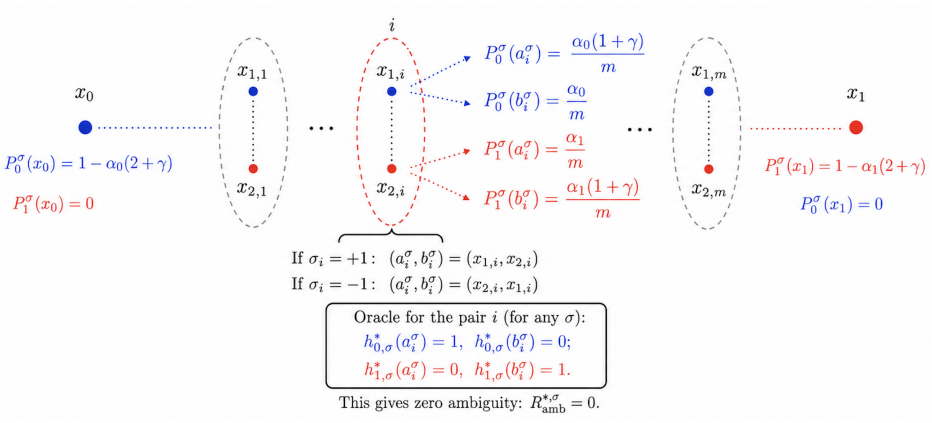

Figure 1: Illustration of the lower bound construction in Theorem 3. Here, $m = \left\lfloor \frac{d_{\mathcal{H}}-2}{2} \right\rfloor \sim d_{\mathcal{H}}$. The unknown sign vector $\sigma \in \{-1,+1\}^m$ determines ,in each pair $(x_{1,i}, x_{2,i})$, which point is a_i^σ and which is b_i^σ. The oracle accepts a_i^σ for class 0 and b_i^σ for class 1, satisfying $R_0 = \alpha_0$ and $R_1 = \alpha_1$ with zero ambiguity. Since the mass gap is only of order $\gamma \sim \sqrt{d_{\mathcal{H}}/n}$, the different values of σ are difficult to distinguish; errors in recovering σ, together with the feasibility constraints, orce overlap between the two regions, yielding an $\Omega\left(\sqrt{d_{\mathcal{H}}/n}\right)$ excess ambiguity lower bound.

in equation 4. Then, conditional on $\mathcal{E}_n$, with probability at least $1 - 2\delta - \frac{\delta}{\mathbb{P}(\mathcal{E}_n)} - \frac{2\exp(-cn\min\{\pi_0,\pi_1\})}{\mathbb{P}(\mathcal{E}_n)}$ with respect to the training sample S^n, we have

$$\overline{R}_j(\widetilde{h}_j) \le \alpha_j, \quad j \in \{0,1\}, \qquad \text{and} \qquad \overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \le C_{\alpha,\pi}\sqrt{\frac{d_{\mathcal{H}}\log(en) + \log(1/\delta)}{n}}.$$

Here, $\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) := \overline{R}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) - R^*_{\mathrm{amb}}$, and $C_{\alpha,\pi} > 0$ is a constant depending only on $\alpha_0, \alpha_1, \pi_0$, and π_1.

4.2 Minimax Lower Bound

To derive a matching minimax lower bound for the problem equation 1, we first specify the class of distributions over which the minimax problem is formulated.

Definition 1 (Class of distributions)**.** *Fix a hypothesis class $\mathcal{H}$ with finite VC dimension $d_{\mathcal{H}}$. For $\alpha_0, \alpha_1 \in (0,1)$, let $\mathcal{P}_{\mathcal{H}}(\alpha_0, \alpha_1, \pi_0, \pi_1)$ denote the class of distributions P on $\mathcal{X} \times \{0,1\}$ such that $P(Y=0) = \pi_0, P(Y=1) = \pi_1$, and for which the problem equation 1 is feasible; namely, there exist $h_0, h_1 \in \mathcal{H}$ such that $R_0(h_0) \le \alpha_0$ and $R_1(h_1) \le \alpha_1$.*

We next specify the class of learning rules considered in the minimax formulation. In line with the randomized learner used in the upper bound, we allow the learning rule to use internal randomness. The constraints are required to hold in expectation with respect to this internal randomness, with high probability over the training sample.

Definition 2 (Approximate randomized (α_0, α_1)-learner)**.** *Let $r_0, r_1 \ge 0$ and $0 < \delta < 1$. Let $\mathcal{H}^+ = \mathcal{H} \cup \{\mathbf{1}\}$, where $\mathbf{1}$ denotes the constant-one classifier. For every $n \ge 1$, we call a randomized learning rule $(S^n, \zeta) \longmapsto (\widetilde{h}_0, \widetilde{h}_1) \in \mathcal{H}^+ \times \mathcal{H}^+$ an (r_0, r_1, δ)-**approximate randomized** (α_0, α_1)-**learner** if the auxiliary randomness ζ is independent of S^n, and for every distribution $P \in \mathcal{P}_{\mathcal{H}}(\alpha_0, \alpha_1, \pi_0, \pi_1)$,*

$$\mathbb{P}_{S^n \sim P^n}\left(\overline{R}_0(\widetilde{h}_0) \le \alpha_0 + r_0,\ \overline{R}_1(\widetilde{h}_1) \le \alpha_1 + r_1\right) \ge 1 - \delta_0.$$

Theorem 3. *Fix a hypothesis class $\mathcal{H}$ with finite VC dimension $d_{\mathcal{H}} \ge 4$, and suppose that the constant-one classifier $\mathbf{1}$ belongs to $\mathcal{H}$. Let $\alpha_0, \alpha_1 \in (0, \frac{49}{100})$ and $0 < \delta < 1/6$. Suppose that n is large enough so that $n \ge c\, d_{\mathcal{H}}$ for a sufficiently large constant $c > 0$. Let $r_0, r_1 \ge 0$ satisfy $\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1} \le c_0\sqrt{\frac{d_{\mathcal{H}}}{n}}$ for a sufficiently small constant $c_0 > 0$. Then, for every randomized (r_0, r_1, δ)-approximate (α_0, α_1)-learner $(S^n, \zeta) \longmapsto (\widetilde{h}_0, \widetilde{h}_1) \in \mathcal{H} \times \mathcal{H}$, there exists a distribution $P \in$*

$\mathcal{P}_{\mathcal{H}}(\alpha_0, \alpha_1, \frac{1}{2}, \frac{1}{2})$ *such that*

$$\mathbb{P}_{S^n \sim P^n}\left(\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \geq c_1\sqrt{\frac{d_{\mathcal{H}}}{n}}\right) \geq c_2,$$

where $c_1, c_2 > 0$ are universal constants depending only on α_0 and α_1.

Remark 1. *Theorem 3 applies to the class of randomized (r_0, r_1, δ)-approximate learners whenever $\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1} \leq c_0\sqrt{\frac{d_{\mathcal{H}}}{n}}$ for a sufficiently small constant $c_0 > 0$. This class contains the exact-feasible case $r_0 = r_1 = 0$. It also contains deterministic learners as the special case in which the auxiliary randomness is degenerate. Thus the lower bound is robust: even after allowing both randomization and constraint slack of the same order as the target statistical rate, the ambiguity excess risk remains at least of order $\sqrt{\frac{d_{\mathcal{H}}}{n}}$.*

The intuition behind the proof of the lower bound is illustrated in Figure 1. The construction consists of a collection of paired points, with the number of pairs proportional to the VC dimension $d_{\mathcal{H}}$. For each pair, an unknown sign determines a small perturbation of the class-conditional probabilities, arranged in opposite directions for the two classes. If these signs were known, an oracle could select complementary points for the two class-specific regions, satisfy both error constraints, and achieve zero ambiguity. However, the perturbations are chosen small enough that, from n observations, the learner cannot reliably identify the correct choice across all pairs. Under the class-conditional feasibility requirements, such errors necessarily translate into overlap between the two prediction regions, and hence into nonzero ambiguity. Choosing the perturbation at the statistical indistinguishability scale then yields an excess ambiguity of order $\sqrt{d_{\mathcal{H}}/n}$.

5 Surrogate Formulation and Statistical Guarantees

The formulation developed so far is based on binary-valued classifiers and the 0–1 loss. Although this formulation provides a natural statistical description of the ambiguity minimization problem, directly optimizing the associated empirical objective is generally computationally difficult. In particular, the indicator losses appearing in both the class-conditional constraints and the ambiguity objective are discontinuous and typically lead to nonconvex combinatorial optimization problems. We therefore replace the binary hypotheses by real-valued score functions and the indicator losses by surrogate losses. In addition to yielding more tractable optimization problems, surrogate losses incorporate the distance of each observation from the corresponding decision boundary, thereby encouraging classifiers with larger margins and improved robustness.

Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$. Each $f \in \mathcal{F}$ induces the binary classifier $h_f(x) := \mathbf{1}\{f(x) \geq 0\}$. Accordingly, for a pair $(f_0, f_1) \in \mathcal{F} \times \mathcal{F}$, the point $x \in \mathcal{X}$ is declared ambiguous whenever $f_0(x) \geq 0$ and $f_1(x) \geq 0$. Thus, the induced 0–1 class-conditional and ambiguity losses are defined by $\ell_j^{0\text{-}1}(f_j; x) := \mathbf{1}\{f_j(x) < 0\}$, $j \in \{0, 1\}$, and $\ell_{\mathrm{amb}}^{0\text{-}1}(f_0, f_1; x) := \mathbf{1}\{f_0(x) \geq 0,\ f_1(x) \geq 0\}$.

Definition 3 (Surrogate loss)**.** *Assume that there exists $B > 0$ such that $\sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} |f(x)| \leq B$. A function $\phi : \mathbb{R} \to \mathbb{R}_+$ is called an L_ϕ-Lipschitz surrogate loss over $\mathcal{F}$ if it is nondecreasing, satisfies $\phi(0) = 1$, $|\phi(u) - \phi(v)| \leq L_\phi|u - v|$ for all $u, v \in [-B, B]$, and there exists a constant $C_\phi > 1$ such that $\sup_{|u|\leq B} \phi(u) \leq C_\phi$.*

If $\mathcal{F}$ is uniformly bounded on $\mathcal{X}$, then standard surrogate losses such as the hinge, logistic, and exponential losses satisfy all the conditions in Definition 3.

Definition 4 (Surrogate risks)**.** *Let $(u)_+ := \max\{u, 0\}$. For $(f_0, f_1) \in \mathcal{F} \times \mathcal{F}$, define $R_{\phi,j}(f_j) := \mathbb{E}_{P_j}[\phi(-f_j(X))]$, $j \in \{0, 1\}$, and the pointwise surrogate ambiguity losses by*

$$\ell_{\mathrm{amb},\phi}^{\mathrm{add}}(f_0, f_1; x) := [\phi(f_0(x)) + \phi(f_1(x)) - 1]_+, \qquad \ell_{\mathrm{amb},\phi}^{\mathrm{prod}}(f_0, f_1; x) := \phi(f_0(x))\phi(f_1(x)).$$

For $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$, define $R_{\mathrm{amb},\phi}^{\diamond}(f_0, f_1) := \mathbb{E}_{P_X}\left[\ell_{\mathrm{amb},\phi}^{\diamond}(f_0, f_1; X)\right]$.

Since ϕ is nondecreasing and $\phi(0) = 1$, these surrogate risks upper-bound their 0–1 counterparts: $R_j(h_{f_j}) \leq R_{\phi,j}(f_j)$ for $j \in \{0, 1\}$, and $R_{\mathrm{amb}}(h_{f_0}, h_{f_1}) \leq R_{\mathrm{amb},\phi}^{\diamond}(f_0, f_1)$ for $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$.

The additive formulation is convex whenever ϕ is convex and $\mathcal{F}$ admits a convex parameterization. However, it may assign a large loss even when only one classifier accepts the point, since a sufficiently large positive score can outweigh a negative score from the other classifier. In contrast, ambiguity occurs only when both classifiers accept the same point, and the multiplicative formulation assigns a smaller penalty whenever at least one score is negative. It is therefore better aligned with the ambiguity event, although it is generally nonconvex.

For $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$, consider the population optimization problem

$$\underset{(f_0,f_1)\in\mathcal{F}\times\mathcal{F}}{\text{minimize}} \quad R^{\diamond}_{\mathrm{amb},\phi}(f_0,f_1) \quad \text{subject to} \qquad R_{\phi,0}(f_0) \le \alpha_0, \qquad R_{\phi,1}(f_1) \le \alpha_1. \tag{5}$$

We denote a solution to equation 5 by $(f_0^{\diamond,*}, f_1^{\diamond,*})$ and its optimal value by $R^{\diamond,*}_{\mathrm{amb},\phi}$. The corresponding surrogate excess ambiguity risk is

$$\mathcal{E}^{\diamond}_{\mathrm{amb},\phi}(f_0,f_1) := R^{\diamond}_{\mathrm{amb},\phi}(f_0,f_1) - R^{\diamond,*}_{\mathrm{amb},\phi}.$$

We next introduce the empirical counterpart of equation 5, obtained by replacing the population surrogate risks with their empirical analogues and allowing small slacks in the class-conditional constraints. Recall that n_j denotes the number of samples from class j. For $n_j > 0$, define

$$\widehat{R}_{\phi,j}(f_j) := \frac{1}{n_j} \sum_{i:Y_i=j} \phi(-f_j(X_i)), \qquad j \in \{0,1\}.$$

For $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$, define the empirical surrogate ambiguity risk by

$$\widehat{R}^{\diamond}_{\mathrm{amb},\phi}(f_0,f_1) := \frac{1}{n} \sum_{i=1}^{n} \ell^{\diamond}_{\mathrm{amb},\phi}(f_0,f_1;X_i).$$

Let $\epsilon_0, \epsilon_1 \ge 0$ be slack parameters. For $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$, the empirical surrogate learner is defined as any solution $(\widehat{f}_0, \widehat{f}_1) \in \mathcal{F} \times \mathcal{F}$ to

$$\underset{(f_0,f_1)\in\mathcal{F}\times\mathcal{F}}{\text{minimize}} \quad \widehat{R}^{\diamond}_{\mathrm{amb},\phi}(f_0,f_1) \quad \text{subject to} \qquad \widehat{R}_{\phi,0}(f_0) \le \alpha_0 + \epsilon_0, \qquad \widehat{R}_{\phi,1}(f_1) \le \alpha_1 + \epsilon_1. \tag{6}$$

To analyze the statistical performance of the empirical solutions, we measure the complexity of the score class $\mathcal{F}$ through its Rademacher complexity.

Definition 5 (Rademacher complexity)**.** *Let $X_1, \dots, X_m$ be i.i.d. samples drawn from a distribution Q on $\mathcal{X}$. Define the empirical Rademacher complexity by*

$$\widehat{\mathfrak{R}}_m(\mathcal{F}) := \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(X_i) \right| \right],$$

where $\sigma_1, \dots, \sigma_m$ are independent uniform $\{-1,1\}$-valued random variables. The Rademacher complexity of $\mathcal{F}$ with respect to Q is defined by $\mathfrak{R}^Q_m(\mathcal{F}) := \mathbb{E}\left[\widehat{\mathfrak{R}}_m(\mathcal{F})\right]$, where the expectation is taken with respect to the i.i.d. sample.

Assumption 1. *There exists a constant $B_{\mathcal{F}} > 0$ such that $\mathfrak{R}^Q_m(\mathcal{F}) \le \frac{B_{\mathcal{F}}}{\sqrt{m}}$ for all distributions Q on $\mathcal{X}$.*

Remark 2. *Assumption 1 is satisfied by many commonly used function classes under standard boundedness conditions. For instance, linear predictors with bounded coefficients and neural networks with bounded weights have Rademacher complexity of order $m^{-1/2}$ when the input domain is bounded.*

For $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$, define $\kappa_{\mathrm{add}} := 1$, $\kappa_{\mathrm{prod}} := C_\phi$, $M_{\mathrm{add}} := 2C_\phi$, and $M_{\mathrm{prod}} := C_\phi^2$. For $m \ge 1$ and $\delta \in (0,1)$, let

$$\epsilon_\phi(m,\delta) := \frac{4L_\phi B_{\mathcal{F}}}{\sqrt{m}} + C_\phi \sqrt{\frac{\log(6/\delta)}{2m}}, \qquad \epsilon^{\diamond}_{\mathrm{amb},\phi}(n,\delta) := C\left(\frac{\kappa_\diamond L_\phi B_{\mathcal{F}}}{\sqrt{n}} + M_\diamond \sqrt{\frac{\log(6/\delta)}{2n}} \right),$$

where $C > 0$ is a universal constant.

Theorem 4. *Suppose that Definition 3 and Assumption 1 hold. Fix* $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$. *Let* $\mathcal{E}_n := \{n_0 > 0,\ n_1 > 0\}$, *and, on* $\mathcal{E}_n$, *let* $(\widehat{f}_0^{\diamond}, \widehat{f}_1^{\diamond})$ *be any solution of the empirical problem equation 6 with slack parameters* $\epsilon_j = \epsilon_\phi(n_j, \delta)$ *for* $j \in \{0,1\}$. *Then, conditional on* $\mathcal{E}_n$, *with probability at least* $1 - \frac{2\delta}{3} - \frac{\delta}{3\mathbb{P}(\mathcal{E}_n)}$, *we have*

$$R_{\phi,j}(\widehat{f}_j^{\diamond}) \le \alpha_j + 2\epsilon_\phi(n_j, \delta), \quad j \in \{0,1\}, \qquad \text{and} \qquad \mathcal{E}^{\diamond}_{\mathrm{amb},\phi}(\widehat{f}_0^{\diamond}, \widehat{f}_1^{\diamond}) \le 2\epsilon^{\diamond}_{\mathrm{amb},\phi}(n, \delta).$$

Consequently, $R_j(h_{\widehat{f}_j^{\diamond}}) \le \alpha_j + 2\epsilon_\phi(n_j, \delta)$ *for* $j \in \{0,1\}$.

6 Optimization and Computational Guarantees

In this section, we describe the optimization procedure used in our experiments to approximately solve the empirical problem equation 6. First we assume the two classifier f_0 and f_1 are parameterized by θ_0 and θ_1, respectively, where $\theta_0, \theta_1 \in \Theta \subseteq \mathbb{R}^d$. For $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$, The problem can then be rewritten as follows:

$$\begin{aligned} \underset{(\theta_0,\theta_1)\in\Theta\times\Theta}{\text{minimize}} \quad & \widehat{R}^{\diamond}_{\mathrm{amb},\phi}\big(f_{\theta_0}, f_{\theta_1}\big) \\ \text{subject to} \quad & \widehat{R}_{\phi,0}\big(f_{\theta_0}\big) \le \alpha_0 + \epsilon_0, \quad \widehat{R}_{\phi,1}\big(f_{\theta_1}\big) \le \alpha_1 + \epsilon_1. \end{aligned} \tag{7}$$

We further define $g(\theta_0, \theta_1) := \max\{\widehat{R}_{\phi,0}\big(f_{\theta_0}\big) - \alpha_0 - \epsilon_0, \widehat{R}_{\phi,1}\big(f_{\theta_1}\big) - \alpha_1 - \epsilon_1\}$, so that the two constraints can be written compactly as $g(\theta_0, \theta_1) \le 0$. To approximately solve this constrained problem, we employ the one-projection algorithm of Mahdavi et al. (2012). The algorithm first applies gradient descent–ascent to the associated Lagrangian problem

$$\min_{(\theta_0,\theta_1)\in\Theta\times\Theta} \max_{\lambda\ge 0} \left\{ \widehat{R}^{\diamond}_{\mathrm{amb},\phi}\big(f_{\theta_0}, f_{\theta_1}\big) + \lambda g(\theta_0, \theta_1) \right\}.$$

After T iterations, it performs a single projection of the parameters onto the feasible set. The complete procedure is presented in Algorithm 1. Although our experiments use multilayer perceptrons (MLPs) with both the additive and product ambiguity losses, the following guarantee applies to the additive formulation when the objective and constraints are convex in the parameters.

Theorem 5. *Consider* $\diamond = \mathrm{add}$ *and write* $\theta = (\theta_0, \theta_1)$. *Let* $B = \{\theta : \|\theta\|_2 \le 1\}$ *and* $K = \{\theta : g(\theta) \le 0\}$. *Assume that* $K \subseteq B$, *that there exists* $\theta \in \mathrm{int}(B)$ *with* $g(\theta) < 0$, *and that* $\widehat{\theta}^{\star} = (\widehat{\theta}_0^{\star}, \widehat{\theta}_1^{\star}) \in \arg\min_{\theta\in K} \widehat{R}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\theta_0}, f_{\theta_1})$ *satisfies* $\widehat{\theta}^{\star} \in \mathrm{int}(B)$. *Assume that* g *and each mapping* $\theta \mapsto \ell^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\theta_0}, f_{\theta_1}; X_i)$ *are convex on a neighborhood of* B. *For every* $\theta \in B$ *and* $i \in \{1, \dots, n\}$, *suppose*

$$\sup_{v\in\partial_\theta \ell^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\theta_0}, f_{\theta_1}; X_i)} \|v\|_2 \le G_1, \ \sup_{u\in\partial g(\theta)} \|u\|_2 \le G_2, \ |g(\theta)| \le C_g, \ \inf_{\theta:\, g(\theta)=0} \inf_{u\in\partial g(\theta)} \|u\|_2 \ge \rho > 0.$$

For $\delta \in (0,1)$, *define* $c_\delta := \sqrt{G_1^2 + C_g^2 + 4\left(1 + \ln\frac{2}{\delta}\right) G_1^2}, \mu := \frac{G_2^2}{c_\delta\sqrt{T}}, \eta := \frac{\mu}{2G_2^2}$. *Then Algorithm 1 outputs* $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1) \in K$ *such that, conditional on the training sample, with probability at least* $1 - \delta$

$$\widehat{R}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\widehat{\theta}_0}, f_{\widehat{\theta}_1}) - \widehat{R}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\widehat{\theta}_0^{\star}}, f_{\widehat{\theta}_1^{\star}}) \lesssim \frac{1}{\sqrt{T}} \left[\left(1 + \frac{G_1^2}{\rho^2}\right) c_\delta + \frac{G_1^2 G_2^2}{\rho^2 c_\delta} + G_1 \sqrt{\ln\frac{2}{\delta}} \right].$$

Consequently, taking $T \gtrsim \frac{1}{\left(\epsilon^{\mathrm{add}}_{\mathrm{amb},\phi}(n,\delta)\right)^2} \left[\left(1 + \frac{G_1^2}{\rho^2}\right) c_\delta + \frac{G_1^2 G_2^2}{\rho^2 c_\delta} + G_1 \sqrt{\ln\frac{2}{\delta}} \right]^2$, *under the assumptions of Theorem 4, preserves its feasibility guarantees and yields its ambiguity rate up to a constant factor, with an additional failure probability of at most* δ.

7 Experiments

We evaluate our approach on five benchmark datasets and synthetic Gaussian data, comparing with the method of Lei (2014). We use MLPs and, on BCI, linear models to parameterize the class-specific scores, with additive and product surrogates based on hinge or squared hinge loss. All experiments use stratified 70/30 training–test splits and target error levels $\alpha_0 = \alpha_1 = 0.05$. We report

Algorithm 1: SGD with One Projection

Input: $\theta_0^0 = \theta_1^0 = 0$, $\lambda_0 = 0$, stepsize η, $\mu > 0$, and $\diamond \in \{\mathrm{add}, \mathrm{prod}\}$.
for $t = 0, \ldots, T-1$ **do**
  Sample X_t uniformly from $\{X_i\}_{i=1}^n$ and Choose $G_t^\diamond \in \partial_{(\theta_0,\theta_1)} \ell_{\mathrm{amb},\phi}^\diamond(f_{\theta_0^t}, f_{\theta_1^t}; X_t)$ and $u_t \in \partial_{(\theta_0,\theta_1)} g(\theta_0^t, \theta_1^t)$
  $\theta' = (\theta_0^t, \theta_1^t) - \eta G_t^\diamond - \eta \lambda_t u_t$
  $(\theta_0^{t+1}, \theta_1^{t+1}) = \theta' / \max\{1, \|\theta'\|_2\}$
  $\lambda_{t+1} = [(1 - \mu\eta)\lambda_t + \eta g(\theta_0^t, \theta_1^t)]_+$
end
$\widehat{\theta} = \ell_2$ projection of $\frac{1}{T}\sum_{t=0}^{T-1}(\theta_0^t, \theta_1^t)$ onto the constraint set $\{(\theta_0, \theta_1) : g(\theta_0, \theta_1) \le 0\}$.
Output: $\widehat{\theta}$

test class-conditional errors and ambiguity risk over ten runs. Dataset descriptions, preprocessing, model architectures, and training details and further discussion of the results for each dataset are provided in Appendix H.

Table 1: Test class-conditional errors (R_0, R_1) and ambiguity risk (R_{amb}): mean $\pm$ standard deviation over ten runs. Target errors are $\alpha_0 = \alpha_1 = 0.05$. Add. and Prod. denote additive and product surrogates.

Model / loss / surrogate	R_0	R_1	R_{amb}
Vertebral Column (Barreto & Neto, 2005)			
Lei (Lei, 2014)	0.0333 ± 0.0351	0.0524 ± 0.0290	0.3505 ± 0.1152
MLP / hinge / Add.	0.0400 ± 0.0344	0.0524 ± 0.0375	0.3086 ± 0.1153
MLP / hinge / Prod.	0.0467 ± 0.0358	0.0540 ± 0.0383	0.2591 ± 0.1029
MAGIC Gamma Telescope (Bock, 2004)			
Lei	0.0501 ± 0.0046	0.0485 ± 0.0064	0.5966 ± 0.0197
MLP / hinge / Add.	0.0316 ± 0.0204	0.0520 ± 0.0232	0.5107 ± 0.0889
MLP / hinge / Prod.	0.0418 ± 0.0091	0.0493 ± 0.0092	0.4332 ± 0.0263
BCI Competition IV 2b (Leeb et al., 2008)			
Lei (Lei, 2014)	0.0351 ± 0.0361	0.0522 ± 0.0163	0.2293 ± 0.4065
Linear / hinge / Add.	0.0459 ± 0.0242	0.0450 ± 0.0147	0.0117 ± 0.0098
Linear / hinge / Prod.	0.0505 ± 0.0279	0.0450 ± 0.0147	0.0077 ± 0.0090
MLP / hinge / Add.	0.0523 ± 0.0268	0.0495 ± 0.0196	0.0050 ± 0.0113
MLP / hinge / Prod.	0.0658 ± 0.0321	0.0396 ± 0.0114	0.0072 ± 0.0085
MLP / squared hinge / Prod.	0.0459 ± 0.0270	0.0423 ± 0.0113	0.0158 ± 0.0133
Phoneme (pho)			
Lei (Lei, 2014)	0.0531 ± 0.0122	0.0540 ± 0.0069	0.5606 ± 0.0257
MLP / hinge / Add.	0.0387 ± 0.0308	0.0462 ± 0.0170	0.5292 ± 0.0607
MLP / hinge / Prod.	0.0551 ± 0.0204	0.0479 ± 0.0151	0.4482 ± 0.0341
Sensorless Drive Diagnosis (Bator, 2013)			
Lei (Lei, 2014)	0.0538 ± 0.0093	0.0505 ± 0.0059	0.2833 ± 0.1541
MLP / hinge / Add.	0.0441 ± 0.0203	0.0527 ± 0.0127	0.0644 ± 0.0527
MLP / hinge / Prod.	0.0437 ± 0.0154	0.0543 ± 0.0116	0.0166 ± 0.0123
Synthetic Gaussian (dim = 10)			
Lei (Lei, 2014)	0.0485 ± 0.0051	0.0510 ± 0.0104	0.6038 ± 0.0207
MLP / hinge / Add.	0.0473 ± 0.0173	0.0490 ± 0.0195	0.5056 ± 0.0921
MLP / hinge / Prod.	0.0611 ± 0.0092	0.0519 ± 0.0103	0.4089 ± 0.0392
MLP / squared hinge / Add.	0.0655 ± 0.0193	0.0496 ± 0.0174	0.4565 ± 0.0754
MLP / squared hinge / Prod.	0.0512 ± 0.0109	0.0479 ± 0.0098	0.4415 ± 0.0441

Table 1 shows that our approach achieves lower mean ambiguity risk than the method of Lei (2014) across all datasets and configurations. On Phoneme and synthetic Gaussian data, the additive hinge formulation improves all three metrics: both class-conditional errors and ambiguity risk. On other datasets, substantial reductions in ambiguity are achieved with errors close to the prescribed thresholds. For example, on Sensorless, the product formulation reduces ambiguity from 0.2833 to 0.0166, with mean errors $R_0 = 0.0437$ and $R_1 = 0.0543$, the latter only slightly above the target 0.05. The product surrogate generally yields lower ambiguity than the additive surrogate, although the relative performance depends on the model and dataset.

AI USE DISCLOSURE

We used generative AI tools to assist with editing and polishing the manuscript and checking proofs for potential errors and inconsistencies. The authors reviewed the AI suggestions and take full responsibility for the final content of the paper, including all mathematical claims and proofs.

REFERENCES

Phoneme. OpenML. URL https://www.openml.org/d/1489.

Jean V Alves, Diogo Leitão, Sérgio Jesus, Marco OP Sampaio, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. Fifar: A fraud detection dataset for learning to defer. *arXiv preprint arXiv:2312.13218*, 2023.

Guilherme Barreto and Ajalmar Neto. Vertebral column. UCI Machine Learning Repository, 2005. URL https://doi.org/10.24432/C5K89B.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Martyna Bator. Dataset for sensorless drive diagnosis. UCI Machine Learning Repository, 2013. URL https://doi.org/10.24432/C5VP5F.

R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. URL https://doi.org/10.24432/C52C8B.

Chao Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Ajalmar R da Rocha Neto, Ricardo Sousa, Guilherme de A. Barreto, and Jaime S Cardoso. Diagnostic of pathology on the vertebral column with embedded reject option. In *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 588–595. Springer, 2011.

Diptesh Das, Junichi Ito, Tadashi Kadowaki, and Koji Tsuda. An interpretable machine learning model for diagnosis of alzheimer's disease. *PeerJ*, 7:e6543, 2019.

Christophe Denis and Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 32(1):42–72, 2020.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In *International Conference on Artificial Intelligence and Statistics*, pp. 2179–2187. PMLR, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Kanza Hamid, Amina Asif, Wajid Abbasi, Durre Sabih, and Fayyaz-ul-Amir Afsar Minhas. Machine learning with abstention for automated liver disease diagnosis. In *2017 international conference on frontiers of information technology (FIT)*, pp. 356–361. IEEE, 2017.

Blaise Hanczar and Edward R Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–1895, 2008.

Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pp. 709–721, 2006.

Mohammadreza M. Kalan and Samory Kpotufe. Distribution-free rates in neyman–pearson classification. *arXiv preprint arXiv:2402.09560*, 2024.

Mohammadreza Mousavi Kalan, Eitan J. Neugut, and Samory Kpotufe. Transfer neyman-pearson algorithm for outlier detection. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pp. 4717–4725, 2025a.

Mohammadreza Mousavi Kalan, Eitan J. Neugut, and Samory Kpotufe. Transfer neyman-pearson algorithm for outlier detection. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4717–4725. PMLR, 03–05 May 2025b.

R. Leeb, C. Brunner, G. R. Müller-Putz, A. Schlögl, and G. Pfurtscheller. BCI Competition 2008 – Graz Data Set B. Graz University of Technology, 2008. URL `https://www.bbci.de/competition/iv/desc_2b.pdf`.

Erich L. Lehmann. *Testing Statistical Hypotheses*. John Wiley & Sons, New York, 2 edition, 1986.

Jing Lei. Classification with confidence. *Biometrika*, 2014.

Mingkun Li and Ishwar K Sethi. Confidence-based classifier design. *Pattern recognition*, 39(7): 1230–1240, 2006.

Mehrdad Mahdavi, Tianbao Yang, Rong Jin, Shenghuo Zhu, and Jinfeng Yi. Stochastic gradient descent with only one projection. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper_files/paper/2012/file/c52f1bd66cc19d05628bd8bf27af3ad6-Paper.pdf`.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2 edition, 2018.

Mohammadreza Mousavi Kalan and Samory Kpotufe. Tight rates in supervised outlier transfer learning. In *International Conference on Learning Representations*, 2024.

Mohammadreza Mousavi Kalan, Yuyang Deng, Eitan J Neugut, and Samory Kpotufe. Neyman-pearson classification under both null and alternative distributions shift. In *International Conference on Learning Representations*, volume 2026, pp. 30384–30408, 2026.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pp. 7076–7087, 2020.

Tadeusz Pietraszek. Optimizing abstaining classifiers using roc analysis. In *Proceedings of the 22nd international conference on Machine learning*, pp. 665–672, 2005.

Tadeusz Pietraszek. On the use of roc analysis for the optimization of abstaining classifiers. *Machine learning*, 68(2):137–169, 2007.

Harish G Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with a reject option. *arXiv preprint arXiv:1505.04137*, 2015.

Philippe Rigollet and Xin Tong. Neyman–pearson classification under a strict type i error constraint. *Journal of Machine Learning Research*, 12:2839–2873, 2011.

Clayton Scott and Robert Nowak. A neyman–pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

Shubhanshu Shekhar, Mohammad Ghavamzadeh, and Tara Javidi. Binary classification with bounded abstention rate. *arXiv preprint arXiv:1905.09561*, 2019.

Manpreet Singh, Akshatha Srikantha, and Shyamal Lakhanpal. Cost-sensitive conformal prediction and human-in-the-loop abstention for imbalanced high-stakes decision support: A multi-domain benchmark. *arXiv preprint arXiv:2607.27143*, 2026.

Xin Tong. A plug-in approach to neyman–pearson classification. *Journal of Machine Learning Research*, 14:3011–3040, 2013.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman–pearson classification algorithms and NP receiver operating characteristics. *Science Advances*, 4(2):eaao1659, 2018. doi: 10.1126/sciadv.aao1659.

Xin Tong, Lucy Xia, Jiacheng Wang, and Yang Feng. Neyman–pearson classification: Parametrics and sample size requirement. *Journal of Machine Learning Research*, 21(12):1–48, 2020.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009.

Artem Vazhentsev, Ivan Sviridov, Alvard Barseghyan, Gleb Kuzmin, Alexander Panchenko, Aleksandr Nesterov, Artem Shelmanov, and Maxim Panov. Uncertainty-aware abstention in medical diagnosis based on medical texts. *npj Artificial Intelligence*, 2026.

Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(1), 2010.

Anqi Zhao, Yang Feng, Lie Wang, and Xin Tong. Neyman–pearson classification under high-dimensional settings. *Journal of Machine Learning Research*, 17(212):1–39, 2016.

APPENDIX

A PROOF OF THEOREM 1 [UPPER BOUND]

Proof. Let $\mathcal{E}_n := \{n_0 > 0, n_1 > 0\}$. For $j \in \{0, 1\}$, define

$$Z_j = \sup_{h_j \in \mathcal{H}} \left| \widehat{R}_j(h_j) - R_j(h_j) \right|.$$

For every integer $m \geq 1$, conditional on the event $\{n_j = m\}$, the observations $\{X_i : Y_i = j\}$ are i.i.d. from P_j. Hence, there exists a universal constant $C > 0$ such that

$$\mathbb{P}\left(Z_j \leq C\sqrt{\frac{d_{\mathcal{H}} \log(em) + \log(1/\delta)}{m}} \;\middle|\; n_j = m \right) \geq 1 - \delta.$$

Increasing C if necessary, define the random radius

$$\epsilon_j(n_j, \delta) = C\sqrt{\frac{d_{\mathcal{H}} \log(en_j) + \log(1/\delta)}{n_j}}, \qquad n_j \geq 1.$$

Then, for each $j \in \{0, 1\}$,

$$\begin{aligned}\mathbb{P}\left(Z_j > \frac{\epsilon_j(n_j, \delta)}{2} \;\middle|\; \mathcal{E}_n \right) &= \sum_{m=1}^{n-1} \mathbb{P}\left(Z_j > \frac{\epsilon_j(m, \delta)}{2} \;\middle|\; n_j = m \right) \mathbb{P}\left(n_j = m \mid \mathcal{E}_n\right) \\ &\leq \sum_{m=1}^{n-1} \delta\, \mathbb{P}\left(n_j = m \mid \mathcal{E}_n\right) = \delta.\end{aligned}$$

Define the event

$$\mathcal{A} := \mathcal{E}_n \cap \left\{ Z_j \leq \frac{\epsilon_j(n_j, \delta)}{2} \text{ for } j = 0, 1 \right\}.$$

Therefore, by a union bound, we have $\mathbb{P}(\mathcal{A} \mid \mathcal{E}_n) \geq 1 - 2\delta$. On the event $\mathcal{A}$, since $(\widehat{h}_0, \widehat{h}_1)$ is feasible for the empirical problem, i.e.,

$$\widehat{R}_j(\widehat{h}_j) \leq \alpha_j + \frac{\epsilon_j(n_j, \delta)}{2}, \qquad j \in \{0, 1\},$$

we have

$$R_j(\widehat{h}_j) \leq \widehat{R}_j(\widehat{h}_j) + Z_j \leq \alpha_j + \epsilon_j(n_j, \delta), \qquad j \in \{0, 1\}. \tag{8}$$

It remains to control the ambiguity term. Let

$$\mathcal{G} = \{ g_{h_0, h_1} : x \mapsto h_0(x) h_1(x) : h_0, h_1 \in \mathcal{H} \}.$$

For a function $g \in \mathcal{G}$, we write $P_X g := \mathbb{E}_{X \sim P_X}[g(X)]$ and $P_n g := n^{-1} \sum_{i=1}^n g(X_i)$. Now observe that

$$R_{\mathrm{amb}}(h_0, h_1) = P_X\big(h_0(X) = 1,\ h_1(X) = 1\big) = P_X g_{h_0, h_1}.$$

Similarly, $\widehat{R}_{\mathrm{amb}}(h_0, h_1) = P_n g_{h_0, h_1}$. Therefore,

$$\sup_{h_0, h_1 \in \mathcal{H}} \left| R_{\mathrm{amb}}(h_0, h_1) - \widehat{R}_{\mathrm{amb}}(h_0, h_1) \right| = \sup_{g \in \mathcal{G}} |P_X g - P_n g|. \tag{9}$$

By Lemma 1, the class $\mathcal{G}$ satisfies $\mathrm{VC}(\mathcal{G}) \leq C d_{\mathcal{H}}$, for a universal constant $C > 0$. Let

$$\Delta_{\mathrm{amb}}(n, \delta) := C\sqrt{\frac{d_{\mathcal{H}} \log(en) + \log(1/\delta)}{n}},$$

and define the event

$$\mathcal{B} := \left\{ \sup_{h_0, h_1 \in \mathcal{H}} \left| R_{\mathrm{amb}}(h_0, h_1) - \widehat{R}_{\mathrm{amb}}(h_0, h_1) \right| \leq \Delta_{\mathrm{amb}}(n, \delta) \right\}.$$

Then, by the uniform convergence bound over the class $\mathcal{G}$ and equation 9, we obtain $\mathbb{P}(\mathcal{B}^c) \leq \delta$. Hence, conditional on $\mathcal{E}_n$,

$$\mathbb{P}(\mathcal{B}^c \mid \mathcal{E}_n) = \frac{\mathbb{P}(\mathcal{B}^c \cap \mathcal{E}_n)}{\mathbb{P}(\mathcal{E}_n)} \leq \frac{\delta}{\mathbb{P}(\mathcal{E}_n)}.$$

Now we consider the event $\mathcal{A} \cap \mathcal{B}$. Since (h_0^*, h_1^*) is feasible for the population problem,

$$R_j(h_j^*) \leq \alpha_j, \qquad j \in \{0, 1\}.$$

Using $\mathcal{A}$, we obtain

$$\widehat{R}_j(h_j^*) \leq R_j(h_j^*) + Z_j \leq \alpha_j + \frac{\epsilon_j(n_j, \delta)}{2}, \qquad j \in \{0, 1\}.$$

Thus (h_0^*, h_1^*) is feasible for the empirical problem. Therefore, by the optimality of $(\widehat{h}_0, \widehat{h}_1)$ in the empirical problem,

$$\widehat{R}_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) \leq \widehat{R}_{\mathrm{amb}}(h_0^*, h_1^*).$$

Using $\mathcal{B}$, we obtain

$$\begin{aligned} R_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) - R_{\mathrm{amb}}(h_0^*, h_1^*) &\leq \left[R_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) - \widehat{R}_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1)\right] \\ &\quad + \left[\widehat{R}_{\mathrm{amb}}(h_0^*, h_1^*) - R_{\mathrm{amb}}(h_0^*, h_1^*)\right] \\ &\leq 2\Delta_{\mathrm{amb}}(n, \delta). \end{aligned}$$

Since $R_{\mathrm{amb}}(h_0^*, h_1^*) = R_{\mathrm{amb}}^*$, this gives

$$R_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) - R_{\mathrm{amb}}^* \leq 2\Delta_{\mathrm{amb}}(n, \delta). \tag{10}$$

Finally,

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B} \mid \mathcal{E}_n) \geq 1 - \mathbb{P}(\mathcal{A}^c \mid \mathcal{E}_n) - \mathbb{P}(\mathcal{B}^c \mid \mathcal{E}_n) \geq 1 - 2\delta - \frac{\delta}{\mathbb{P}(\mathcal{E}_n)}.$$

Therefore, conditional on $\mathcal{E}_n$, with probability at least $1 - 2\delta - \frac{\delta}{\mathbb{P}(\mathcal{E}_n)}$, the two class-conditional bounds in equation 8 and the excess ambiguity bound in equation 10 hold simultaneously. □

B Proof of Proposition 1

Proof. Let $\mathcal{X} = \{a, b\}$, and define $g_0(a) = 0$, $g_0(b) = 1$, $g_1(a) = 1$, and $g_1(b) = 0$. Let $\mathcal{H} = \{\mathbf{1}, \mathbf{0}, g_0, g_1\}$. VC dimension of this class is 2. Consider the distribution with equal class priors $\pi_0 = \pi_1 = 1/2$ and class-conditional distributions $P_0(X = a) = \alpha_0$, $P_0(X = b) = 1 - \alpha_0$, $P_1(X = a) = 1 - \alpha_1$, and $P_1(X = b) = \alpha_1$. We have $R_0(g_0) = \alpha_0$ and $R_1(g_1) = \alpha_1$. Therefore, the pair (g_0, g_1) is feasible. Moreover, $R_{\mathrm{amb}}(g_0, g_1) = 0$. It follows that $R_{\mathrm{amb}}^* = 0$.

The class-0 risks of the classifiers in $\mathcal{H}$ are

$$R_0(\mathbf{1}) = 0, \qquad R_0(g_0) = \alpha_0, \qquad R_0(g_1) = 1 - \alpha_0, \qquad R_0(\mathbf{0}) = 1.$$

Hence, the only classifier satisfying $R_0(h_0) < \alpha_0$ is $h_0 = \mathbf{1}$. Similarly, the only classifier satisfying $R_1(h_1) < \alpha_1$ is $h_1 = \mathbf{1}$. Thus, the only pair satisfying both strict constraints is $(h_0, h_1) = (\mathbf{1}, \mathbf{1})$. Its ambiguity risk is $R_{\mathrm{amb}}(\mathbf{1}, \mathbf{1}) = 1$. Therefore, $\mathcal{E}_{\mathrm{amb}}(\mathbf{1}, \mathbf{1}) = 1 - R_{\mathrm{amb}}^* = 1$, which proves the result. □

C Proof of Theorem 2 [Upper bound for the randomized learner]

Proof. Conditioning on $\mathcal{E}_n$, let $\mathcal{B}$ denote the event on which the conclusions of Theorem 1 hold for the empirical solution $(\widehat{h}_0, \widehat{h}_1)$. Thus, we have $\mathbb{P}(\mathcal{B} \mid \mathcal{E}_n) \geq 1 - 2\delta - \frac{\delta}{\mathbb{P}(\mathcal{E}_n)}$. On $\mathcal{B}$, we have

$$R_j(\widehat{h}_j) \leq \alpha_j + \epsilon_j(n_j, \delta), \qquad j \in \{0, 1\},$$

and

$$\mathcal{E}_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) \le C\sqrt{\frac{d_{\mathcal{H}} \log(en) + \log(1/\delta)}{n}}. \tag{11}$$

$\mathbb{E}_\zeta$ denotes expectation only with respect to the internal randomness, with the data S^n fixed. Since the constant-one classifier has class-conditional error equal to zero, we have

$$\overline{R}_j(\widetilde{h}_j) = \mathbb{E}_\zeta \left[R_j\big(B_j \widehat{h}_j + (1 - B_j)\mathbf{1}\big) \right] = q_j R_j(\widehat{h}_j).$$

Therefore,

$$\overline{R}_j(\widetilde{h}_j) \le \frac{\alpha_j}{\alpha_j + \epsilon_j(n_j, \delta)} \big(\alpha_j + \epsilon_j(n_j, \delta)\big) = \alpha_j, \qquad j \in \{0, 1\}.$$

We next bound the ambiguity risk. For every realization of $\zeta = (B_0, B_1)$,

$$R_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \le R_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) + \mathbf{1}\{B_0 = 0\} + \mathbf{1}\{B_1 = 0\}.$$

Taking expectation with respect to ζ gives

$$\overline{R}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \le R_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) + (1 - q_0) + (1 - q_1).$$

Hence,

$$\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \le \mathcal{E}_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) + \sum_{j=0}^{1} (1 - q_j) \le \mathcal{E}_{\mathrm{amb}}(\widehat{h}_0, \widehat{h}_1) + \sum_{j=0}^{1} \frac{\epsilon_j(n_j, \delta)}{\alpha_j}. \tag{12}$$

Next, define the event $\mathcal{N}_n := \big\{ n_j \ge \frac{\pi_j n}{2},\ j \in \{0, 1\} \big\}$. Since $n_j \sim \mathrm{Binomial}(n, \pi_j)$ and $\pi_j > 0$, a standard binomial concentration bound gives

$$\mathbb{P}(\mathcal{N}_n^c) \le 2 \exp\left(-cn \min\{\pi_0, \pi_1\}\right),$$

for a universal constant $c > 0$. Therefore,

$$\mathbb{P}(\mathcal{N}_n^c \mid \mathcal{E}_n) \le \frac{2 \exp\left(-cn \min\{\pi_0, \pi_1\}\right)}{\mathbb{P}(\mathcal{E}_n)}.$$

On $\mathcal{N}_n$, for each $j \in \{0, 1\}$,

$$\epsilon_j(n_j, \delta) = C\sqrt{\frac{d_{\mathcal{H}} \log(en_j) + \log(1/\delta)}{n_j}} \le C_\pi \sqrt{\frac{d_{\mathcal{H}} \log(en) + \log(1/\delta)}{n}}.$$

Combining the bounds equation 11 and equation 12, on $\mathcal{B} \cap \mathcal{N}_n$,

$$\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \le C_{\alpha,\pi} \sqrt{\frac{d_{\mathcal{H}} \log(en) + \log(1/\delta)}{n}}.$$

Finally, conditional on $\mathcal{E}_n$,

$$\mathbb{P}(\mathcal{B} \cap \mathcal{N}_n \mid \mathcal{E}_n) \ge 1 - 2\delta - \frac{\delta}{\mathbb{P}(\mathcal{E}_n)} - \frac{2 \exp\left(-cn \min\{\pi_0, \pi_1\}\right)}{\mathbb{P}(\mathcal{E}_n)}.$$

This concludes the proof. □

D Proof of Theorem 3 [Minimax Lower bound]

We will use the following proposition to construct a well-separated packing of the parameter space.

Proposition 2 (Gilbert–Varshamov bound)**.** *Let $m \ge 8$. Then there exists a subset $\{\sigma_0, \ldots, \sigma_M\}$ of $\{-1, +1\}^m$ such that $\sigma_0 = (1, 1, \ldots, 1)$,*

$$d_H(\sigma_j, \sigma_k) \ge \frac{m}{8}, \qquad \forall\, 0 \le j < k \le M, \qquad \textit{and} \qquad M \ge 2^{m/8},$$

where $d_H(\sigma, \sigma') = \#\{i \in [m] : \sigma(i) \ne \sigma'(i)\}$ is the Hamming distance.

We will also use the following proposition to convert a packing into a lower bound on the probability of estimation error. This proposition is a direct consequence of the Fano/Tsybakov minimax lower bound (Tsybakov, 2009, Theorem 2.5), proved in Section G.

Proposition 3. *Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model, and let S^n denote a sample drawn from $P_\theta^{\otimes n}$. Assume that $M \geq 2$ and that the function $\mathrm{dist}(\cdot,\cdot)$ is a semi-metric. Furthermore, suppose that $\{\Pi_{\theta_j}\}_{\theta_j \in \Theta}$ is a family of distributions indexed over the parameter space Θ, and that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$ such that:*

(i) $\mathrm{dist}(\theta_i, \theta_j) \geq 2s > 0, \forall\, 0 \leq i < j \leq M$.

(ii) $\Pi_{\theta_j} \ll \Pi_{\theta_0}, \forall\, j = 1, \ldots, M$, *and* $\frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}\big(\Pi_{\theta_j} \,\big\|\, \Pi_{\theta_0}\big) \leq \eta \log M$ *with* $0 < \eta < \frac{1}{8}$, *where D_{KL} denotes the KL divergence.*

Define $L_{M,\eta} := \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\eta - \sqrt{\frac{2\eta}{\log M}}\right)$. *Then*

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} P_\theta^{\otimes n}\left(\mathbb{E}_\zeta\left[d\big(\widetilde{\theta}(S^n, \zeta), \theta\big)\right] \geq \frac{s L_{M,\eta}}{2}\right) \geq \frac{L_{M,\eta}}{2 - L_{M,\eta}},$$

where the infimum is over all randomized estimators of the form $\widetilde{\theta} = \widetilde{\theta}(S^n, \zeta)$, with ζ denoting auxiliary randomness of the estimator, independent of the sample S^n.

We divide the proof into two cases: (I) $d_{\mathcal{H}} \geq 18$ and (II) $4 \leq d_{\mathcal{H}} \leq 17$.

Case (I): $d_{\mathcal{H}} \geq 18$. Let $m = \lfloor \frac{d_{\mathcal{H}}-2}{2} \rfloor$. We choose $2m + 2$ points $x_0, x_1, x_{1,1}, \ldots, x_{1,m}, x_{2,1}, \ldots, x_{2,m} \in \mathcal{X}$ that are shattered by $\mathcal{H}$. Let $\sigma = (\sigma_1, ..., \sigma_m) \in \{-1, +1\}^m$. For each $i \in [m]$, define

$$(a_i^\sigma, b_i^\sigma) := \begin{cases} (x_{1,i}, x_{2,i}), & \text{if } \sigma_i = 1, \\ (x_{2,i}, x_{1,i}), & \text{if } \sigma_i = -1. \end{cases}$$

Construction of the distributions: Fix a small number $\gamma \in (0, \frac{4}{100})$, to be chosen later as $\gamma = c\sqrt{\frac{m}{n}}$ for a sufficiently small universal constant $c > 0$. For every $\sigma \in \{-1, +1\}^m$, define P^σ as follows. The label prior is

$$P^\sigma(Y = 0) = P^\sigma(Y = 1) = \frac{1}{2}.$$

The conditional distribution $P_0^\sigma := P^\sigma_{X|Y=0}$ is defined by $P_0^\sigma(x_0) = 1 - \alpha_0(2 + \gamma)$, $P_0^\sigma(x_1) = 0$, and, for each $i \in [m]$,

$$P_0^\sigma(a_i^\sigma) = \frac{\alpha_0(1+\gamma)}{m}, \qquad P_0^\sigma(b_i^\sigma) = \frac{\alpha_0}{m}.$$

Similarly, define $P_1^\sigma = X \mid Y = 1$ by $P_1^\sigma(x_1) = 1 - \alpha_1(2 + \gamma)$, $P_1^\sigma(x_0) = 0$, and, for each $i \in [m]$,

$$P_1^\sigma(b_i^\sigma) = \frac{\alpha_1(1+\gamma)}{m}, \qquad P_1^\sigma(a_i^\sigma) = \frac{\alpha_1}{m}.$$

Because $\alpha_0, \alpha_1 < \frac{49}{100}$ and $\gamma < \frac{4}{100}$, all masses are nonnegative.

Oracle pair has zero ambiguity: Because the selected points are shattered by $\mathcal{H}$, there exist classifiers $h^*_{0,\sigma}, h^*_{1,\sigma} \in \mathcal{H}$ such that

$$h^*_{0,\sigma}(x_0) = 1, \quad h^*_{0,\sigma}(x_1) = 0, \quad h^*_{0,\sigma}(a_i^\sigma) = 1, \quad h^*_{0,\sigma}(b_i^\sigma) = 0, \quad \forall i \in [m],$$

and

$$h^*_{1,\sigma}(x_0) = 0, \quad h^*_{1,\sigma}(x_1) = 1, \quad h^*_{1,\sigma}(a_i^\sigma) = 0, \quad h^*_{1,\sigma}(b_i^\sigma) = 1, \quad \forall i \in [m].$$

For $h^*_{0,\sigma}$,

$$R_0^\sigma(h^*_{0,\sigma}) = P_0^\sigma\big(h^*_{0,\sigma}(X) = 0\big) = \sum_{i=1}^{m} P_0^\sigma(b_i^\sigma) = \alpha_0.$$

Similarly, for $h^*_{1,\sigma}$,

$$R_1^\sigma(h^*_{1,\sigma}) = \sum_{i=1}^m P_1^\sigma(a_i^\sigma) = \alpha_1.$$

Thus the pair is feasible. Moreover, we have $h^*_{0,\sigma}(x)h^*_{1,\sigma}(x) = 0$, which implies that $R^{*,\sigma}_{\mathrm{amb}} = R^\sigma_{\mathrm{amb}}(h^*_{0,\sigma}, h^*_{1,\sigma}) = 0$.

Ambiguity–Hamming reduction: Let $(S^n, \zeta) \mapsto (\widetilde{h}_0, \widetilde{h}_1) \in \mathcal{H}^+ \times \mathcal{H}^+$ be a randomized learning rule, where ζ is independent of S^n. We next define a decoding map that maps each pair $(\widetilde{h}_0, \widetilde{h}_1) \in \mathcal{H}^+ \times \mathcal{H}^+$ to a sign vector $\widehat{\sigma}(\widetilde{h}_0, \widetilde{h}_1) \in \{-1, +1\}^m$. This allows us to relate the ambiguity of $(\widetilde{h}_0, \widetilde{h}_1)$ to the Hamming distance between $\widehat{\sigma}(\widetilde{h}_0, \widetilde{h}_1)$ and the underlying parameter σ.

For a fixed sample S^n, a fixed auxiliary random variable ζ, and a coordinate $i \in [m]$, define

$$A_i^\zeta := \widetilde{h}_0(a_i^\sigma), \qquad B_i^\zeta := \widetilde{h}_0(b_i^\sigma), \qquad C_i^\zeta := \widetilde{h}_1(a_i^\sigma), \qquad D_i^\zeta := \widetilde{h}_1(b_i^\sigma).$$

The oracle pattern at coordinate i is $(A_i, B_i, C_i, D_i) = (1, 0, 0, 1)$. Define the overlap count $O_i^\zeta = A_i^\zeta C_i^\zeta + B_i^\zeta D_i^\zeta$. Let

$$\begin{aligned} O^\zeta &= \sum_{i=1}^m O_i^\zeta, \\ M_0^\zeta &= \sum_{i=1}^m (1 - A_i^\zeta), \quad M_1^\zeta = \sum_{i=1}^m (1 - D_i^\zeta), \\ E_0^\zeta &= \sum_{i=1}^m B_i^\zeta, \qquad\qquad E_1^\zeta = \sum_{i=1}^m C_i^\zeta. \end{aligned}$$

We write

$$\overline{O} := \mathbb{E}_\zeta[O^\zeta], \qquad \overline{M}_0 := \mathbb{E}_\zeta[M_0^\zeta], \qquad \overline{M}_1 := \mathbb{E}_\zeta[M_1^\zeta], \qquad \overline{E}_0 := \mathbb{E}_\zeta[E_0^\zeta], \qquad \overline{E}_1 := \mathbb{E}_\zeta[E_1^\zeta].$$

Consider the feasibility event

$$\overline{\mathcal{E}}_\sigma := \left\{ \overline{R}_0^\sigma(\widetilde{h}_0) \le \alpha_0 + r_0, \quad \overline{R}_1^\sigma(\widetilde{h}_1) \le \alpha_1 + r_1 \right\}.$$

Let $\Pi^\sigma := (P^\sigma)^{\otimes n}$ denote the distribution of the sample S^n. By the definition of an (r_0, r_1, δ)-approximate randomized (α_0, α_1)-learner, we have $\Pi^\sigma(\overline{\mathcal{E}}_\sigma) \ge 1 - \delta$. On $\overline{\mathcal{E}}_\sigma$, the constraint on $\widetilde{h}_0$ gives $\mathbb{E}_\zeta P_0^\sigma(h_0(X) = 1) \ge 1 - \alpha_0 - r_0$. On the other hand, the oracle satisfies $P_0^\sigma(h^*_{0,\sigma}(X) = 1) = 1 - \alpha_0$. Therefore,

$$P_0^\sigma(h^*_{0,\sigma}(X) = 1) - \mathbb{E}_\zeta P_0^\sigma(\widetilde{h}_0(X) = 1) \le r_0. \tag{13}$$

Now, by the definitions of M_0^ζ and E_0^ζ, we have $M_0^\zeta = \#\{i : \widetilde{h}_0(a_i^\sigma) = 0\}$ and $E_0^\zeta = \#\{i : \widetilde{h}_0(b_i^\sigma) = 1\}$. Compared with the oracle $h^*_{0,\sigma}$, every missed point a_i^σ loses mass $\frac{\alpha_0(1+\gamma)}{m}$, and every extra accepted point b_i^σ gains mass $\frac{\alpha_0}{m}$. Hence, equation 13 implies that

$$\frac{\alpha_0}{m}\left((1+\gamma)\overline{M}_0 - \overline{E}_0\right) \le r_0,$$

or equivalently

$$(1+\gamma)\overline{M}_0 - \overline{E}_0 \le \frac{m r_0}{\alpha_0}. \tag{14}$$

Similarly, the class-1 constraint gives

$$(1+\gamma)\overline{M}_1 - \overline{E}_1 \le \frac{m r_1}{\alpha_1}. \tag{15}$$

For every ζ, we have the pointwise inequalities $B_i^\zeta \le B_i^\zeta D_i^\zeta + (1 - D_i^\zeta)$, and $C_i^\zeta \le A_i^\zeta C_i^\zeta + (1 - A_i^\zeta)$. Summing over i and taking expectation with respect to ζ, we obtain

$$\overline{E}_0 \le \overline{O} + \overline{M}_1, \tag{16}$$

and

$$\overline{E}_1 \le \overline{O} + \overline{M}_0. \tag{17}$$

Combining equation 14 with equation 16, we get

$$(1+\gamma)\overline{M}_0 \le \overline{O} + \overline{M}_1 + \frac{mr_0}{\alpha_0}.$$

Combining equation 15 with equation 17, we get

$$(1+\gamma)\overline{M}_1 \le \overline{O} + \overline{M}_0 + \frac{mr_1}{\alpha_1}.$$

Adding these two inequalities yields

$$\overline{M}_0 + \overline{M}_1 \le \frac{2\overline{O}}{\gamma} + \frac{m}{\gamma}\left(\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1}\right). \tag{18}$$

For each realization of ζ, define the decoded sign vector $\widetilde{\sigma} = \widehat{\sigma}(\widetilde{h}_0, \widetilde{h}_1) \in \{-1, +1\}^m$ as follows: at coordinate i, choose the sign whose oracle pattern is closer to the induced pattern of $(\widetilde{h}_0, \widetilde{h}_1)$, breaking ties arbitrarily. If $\widetilde{\sigma}_i \neq \sigma_i$, then the induced pattern cannot equal the true oracle pattern. Hence, at least one of the following must occur:

$$A_i^\zeta = 0, \qquad D_i^\zeta = 0, \qquad A_i^\zeta C_i^\zeta + B_i^\zeta D_i^\zeta \ge 1.$$

Therefore, for every ζ,

$$\mathbf{1}\{\widetilde{\sigma}_i \neq \sigma_i\} \le (1 - A_i^\zeta) + (1 - D_i^\zeta) + A_i^\zeta C_i^\zeta + B_i^\zeta D_i^\zeta.$$

Summing over i and taking expectation over ζ, we obtain

$$\mathbb{E}_\zeta\left[d_H(\widetilde{\sigma}, \sigma)\right] \le \overline{M}_0 + \overline{M}_1 + \overline{O}.$$

Using equation 18, this gives

$$\mathbb{E}_\zeta\left[d_H(\widetilde{\sigma}, \sigma)\right] \le \frac{3\overline{O}}{\gamma} + \frac{m}{\gamma}\left(\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1}\right).$$

Thus,

$$\boxed{\overline{O} \ge \frac{\gamma}{3}\mathbb{E}_\zeta\left[d_H(\widetilde{\sigma}, \sigma)\right] - \frac{m}{3}\left(\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1}\right).} \tag{19}$$

This is the key reduction: a large Hamming distance implies a large overlap, unless the learning rule constraints are violated.

Now we translate the averaged overlap count into averaged ambiguity risk. Each point among the paired points has marginal mass at least $\frac{c_\alpha}{m}$, where $c_\alpha := \frac{1}{2}\min\{\alpha_0, \alpha_1\}$. Therefore,

$$\overline{R}^\sigma_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) = \mathbb{E}_\zeta\left[R^\sigma_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1)\right] \ge \frac{c_\alpha}{m}\overline{O}.$$

Combining this with equation 19, we get

$$\boxed{\overline{R}^\sigma_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \ge c_\alpha\left[\frac{\gamma}{3m}\mathbb{E}_\zeta\left[d_H(\widetilde{\sigma}, \sigma)\right] - \frac{1}{3}\left(\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1}\right)\right].} \tag{20}$$

This is the ambiguity-to-Hamming reduction.

Packing by Gilbert–Varshamov: By Proposition 2, there exists $\Sigma \subset \{-1, +1\}^m$ such that $|\Sigma| - 1 \ge 2^{m/8}$, and for every distinct $\sigma, \sigma' \in \Sigma$, we have $d_H(\sigma, \sigma') \ge \frac{m}{8}$.

KL divergence bound: We now compute KL divergence between P^σ and $P^{\sigma'}$. Because the label prior, namely the marginal distribution of Y, is the same for all σ, we have

$$D_{\mathrm{KL}}(P^\sigma \| P^{\sigma'}) = \frac{1}{2}D_{\mathrm{KL}}(P_0^\sigma \| P_0^{\sigma'}) + \frac{1}{2}D_{\mathrm{KL}}(P_1^\sigma \| P_1^{\sigma'}).$$

If σ and σ' differ in k coordinates, then, for each coordinate in which they differ, the two conditional distributions interchange the two masses $\frac{\alpha_j(1+\gamma)}{m}$ and $\frac{\alpha_j}{m}$. Hence, each differing coordinate contributes

$$\frac{\alpha_0(1+\gamma)}{m}\log\frac{\alpha_0(1+\gamma)/m}{\alpha_0/m}+\frac{\alpha_0}{m}\log\frac{\alpha_0/m}{\alpha_0(1+\gamma)/m}$$

to $D_{\mathrm{KL}}\left(P_0^{\sigma}\,\middle\|\,P_0^{\sigma'}\right)$. Therefore,

$$D_{\mathrm{KL}}\left(P_0^{\sigma}\,\middle\|\,P_0^{\sigma'}\right)=\frac{k\alpha_0}{m}\left[(1+\gamma)\log(1+\gamma)+\log\frac{1}{1+\gamma}\right]\le\frac{k\alpha_0\gamma^2}{m}\le\alpha_0\gamma^2.$$

Consequently,

$$D_{\mathrm{KL}}\left(P^{\sigma}\,\middle\|\,P^{\sigma'}\right)\le\frac{1}{2}(\alpha_0+\alpha_1)\gamma^2.$$

For n i.i.d. samples, we have

$$D_{\mathrm{KL}}\left((P^{\sigma})^{\otimes n}\,\middle\|\,(P^{\sigma'})^{\otimes n}\right)\le\frac{1}{2}n(\alpha_0+\alpha_1)\gamma^2.$$

For $\gamma=\frac{1}{16}\sqrt{\frac{m}{n}}$, we have

$$D_{\mathrm{KL}}\left((P^{\sigma})^{\otimes n}\,\middle\|\,(P^{\sigma'})^{\otimes n}\right)\le\eta\log|\Sigma|. \tag{21}$$

where $\eta=\frac{1}{32}$.

Reduction to randomized Tsybakov's method (Proposition 3): The randomized learning rule induces the randomized estimator $\widetilde{\sigma}=\widehat{\sigma}(\widetilde{h}_0,\widetilde{h}_1)=\widehat{\sigma}(S^n,\zeta)$ of the unknown parameter σ. We apply Proposition 3 with parameter space $\Theta=\Sigma$, semi-metric $\mathrm{dist}(\sigma,\sigma')=d_H(\sigma,\sigma')$, and statistical model $\{\Pi^{\sigma}:\sigma\in\Sigma\}$, where $\Pi^{\sigma}=(P^{\sigma})^{\otimes n}$ denotes the distribution of the sample S^n. For any distinct $\sigma,\sigma'\in\Sigma$, we have $d_H(\sigma,\sigma')\ge\frac{m}{8}$. Thus condition (i) of Proposition 3 holds with $s=\frac{m}{16}$. Moreover, by equation 21, condition (ii) of Proposition 3 holds with $\eta=1/32$. Therefore, for $M+1=|\Sigma|$ with

$$L_{M,\eta}:=\frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\eta-\sqrt{\frac{2\eta}{\log M}}\right),$$

Proposition 3 yields

$$\inf_{\widetilde{\sigma}}\sup_{\sigma\in\Sigma}\Pi^{\sigma}\left(\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma},\sigma)\right]\ge\frac{sL_{M,\eta}}{2}\right)\ge\frac{L_{M,\eta}}{2-L_{M,\eta}}.$$

Since $s=m/16$, $\eta=1/32$, and $M\ge2$, we can write

$$\inf_{\widetilde{\sigma}}\sup_{\sigma\in\Sigma}\Pi^{\sigma}\left(\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma},\sigma)\right]\ge\frac{m}{128}\right)\ge\frac{1}{5}.$$

The bound applies in particular to the randomized estimator induced by the learning rule. Hence there exists $\sigma\in\Sigma$ such that

$$\Pi^{\sigma}\left(\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma},\sigma)\right]\ge\frac{m}{128}\right)\ge\frac{1}{5}. \tag{22}$$

We have $\Pi^{\sigma}(\bar{\mathcal{E}}_{\sigma})\ge1-\delta$. Then, we obtain that

$$\Pi^{\sigma}\left(\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma},\sigma)\right]\ge\frac{m}{128}\text{ and }\mathcal{E}_{\sigma}\right)\ge\frac{1}{5}-\delta\ge c.$$

On this event, using equation 20 we get

$$\begin{aligned}\overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0,\widetilde{h}_1)&\ge c_{\alpha}\left[\frac{\gamma}{3m}\cdot\frac{m}{128}-\frac{1}{3}\left(\frac{r_0}{\alpha_0}+\frac{r_1}{\alpha_1}\right)\right]\\&=c_{\alpha}\left[\frac{1}{6144}\sqrt{\frac{m}{n}}-\frac{1}{3}\left(\frac{r_0}{\alpha_0}+\frac{r_1}{\alpha_1}\right)\right]\ge c_1\sqrt{\frac{d_{\mathcal{H}}}{n}}\end{aligned}$$

for some constant $c_1 > 0$ depending only on α_0, α_1, provided that $\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1} \le c'\sqrt{\frac{d_{\mathcal{H}}}{n}}$ for a sufficiently small constant c'. Since the oracle pair has zero ambiguity, $R^{*,\sigma}_{\mathrm{amb}} = 0$, we have $\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) = \overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1)$. Therefore,

$$\Pi^{\sigma}\left(\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \ge c_1\sqrt{\frac{d_{\mathcal{H}}}{n}}\right) \ge c.$$

Consequently,

$$\sup_{\sigma\in\Sigma}\Pi^{\sigma}\left(\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \ge c_1\sqrt{\frac{d_{\mathcal{H}}}{n}}\right) \ge c.$$

This proves the minimax lower bound.

Case (II): $4 \le \mathbf{d}_{\mathcal{H}} \le 17$. In this case, we use a binary testing construction. Since $d_{\mathcal{H}} \ge 4$, we can choose four points $x_0, x_1, x_{1,1}, x_{2,1} \in \mathcal{X}$ that are shattered by $\mathcal{H}$. We use the same construction as in Case (I), specialized to $m = 1$. In particular, let $\sigma \in \{-1, +1\}$ and define

$$(a^{\sigma}, b^{\sigma}) := \begin{cases}(x_{1,1}, x_{2,1}), & \text{if } \sigma = 1,\\ (x_{2,1}, x_{1,1}), & \text{if } \sigma = -1.\end{cases}$$

For each $\sigma \in \{-1, +1\}$, define P^{σ} as in Case (I), with $\gamma = \frac{1}{8\sqrt{n}}$. The same argument as in Case (I) shows that there exists a feasible oracle pair $(h^*_{0,\sigma}, h^*_{1,\sigma})$ such that $R^{*,\sigma}_{\mathrm{amb}} = 0$. Moreover, the ambiguity-to-Hamming reduction equation 20, specialized to $m = 1$, gives, on the feasibility event $\overline{\mathcal{E}}_{\sigma}$,

$$\overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \ge c_{\alpha}\left[\frac{\gamma}{3}\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma}, \sigma)\right] - \frac{1}{3}\left(\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1}\right)\right].$$

It remains to lower bound the error in estimating the binary parameter σ. Let

$$\Pi^{\sigma} = (P^{\sigma})^{\otimes n}, \qquad \sigma \in \{-1, +1\}.$$

By the same KL-divergence calculation as in Case (I),

$$D_{\mathrm{KL}}\left(\Pi^{+1} \| \Pi^{-1}\right) \le \frac{1}{2}n(\alpha_0 + \alpha_1)\gamma^2 = \frac{\alpha_0 + \alpha_1}{128}.$$

Since $\alpha_0, \alpha_1 < 49/100$, Pinsker's inequality implies

$$\mathrm{TV}\left(\Pi^{+1}, \Pi^{-1}\right) \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\Pi^{+1} \| \Pi^{-1}\right)} \le \frac{1}{16}.$$

Therefore, by Le Cam's two-point method, for every randomized estimator $\widetilde{\sigma} = \widetilde{\sigma}(S^n, \zeta)$,

$$\max_{\sigma\in\{-1,+1\}} \mathbb{E}_{\Pi^{\sigma}}\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma}, \sigma)\right] \ge \frac{1}{2}\left(1 - \mathrm{TV}\left(\Pi^{+1}, \Pi^{-1}\right)\right) \ge \frac{15}{32}.$$

Hence, for every randomized learner, there exists $\sigma \in \{-1, +1\}$ such that $\mathbb{E}_{\Pi^{\sigma}}\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma}, \sigma)\right] \ge \frac{15}{32}$. Define $Q_{\sigma}(S^n) := \mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma}, \sigma)\right]$. Since $0 \le Q_{\sigma}(S^n) \le 1$, we have

$$\mathbb{E}_{\Pi^{\sigma}}[Q_{\sigma}] \le \frac{1}{8} + \frac{7}{8}\Pi^{\sigma}\left(Q_{\sigma} \ge \frac{1}{8}\right).$$

Therefore,

$$\Pi^{\sigma}\left(\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma}, \sigma)\right] \ge \frac{1}{8}\right) \ge \frac{11}{28}.$$

Since $\Pi^{\sigma}(\overline{\mathcal{E}}_{\sigma}) \ge 1 - \delta$, we obtain

$$\Pi^{\sigma}\left(\left\{\mathbb{E}_{\zeta}\left[d_H(\widetilde{\sigma}, \sigma)\right] \ge \frac{1}{8}\right\} \cap \overline{\mathcal{E}}_{\sigma}\right) \ge \frac{11}{28} - \delta \ge c.$$

On this event, the ambiguity-to-Hamming reduction gives

$$\overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \ge c_{\alpha}\left[\frac{\gamma}{24} - \frac{1}{3}\left(\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1}\right)\right].$$

Since $\gamma = 1/(8\sqrt{n})$, we have

$$\overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \geq c_\alpha \left[\frac{1}{192\sqrt{n}} - \frac{1}{3}\left(\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1} \right) \right].$$

By assumption,

$$\frac{r_0}{\alpha_0} + \frac{r_1}{\alpha_1} \leq c' \sqrt{\frac{d_{\mathcal{H}}}{n}} \leq c' \sqrt{\frac{17}{n}}.$$

Thus, for a sufficiently small constant $c' > 0$, we have $\overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \geq \frac{c_1}{\sqrt{n}}$, for some constant $c_1 > 0$ depending only on α_0 and α_1. Since $d_{\mathcal{H}} \leq 17$, we have $\frac{1}{\sqrt{n}} \geq \frac{1}{\sqrt{17}} \sqrt{\frac{d_{\mathcal{H}}}{n}}$. Consequently, for some constant $c_2 > 0$ depending only on α_0 and α_1, we have $\overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \geq c_2 \sqrt{\frac{d_{\mathcal{H}}}{n}}$. Since $R^{*,\sigma}_{\mathrm{amb}} = 0$, we have $\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) = \overline{R}^{\sigma}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1)$. Therefore,

$$\Pi^{\sigma}\left(\overline{\mathcal{E}}_{\mathrm{amb}}(\widetilde{h}_0, \widetilde{h}_1) \geq c_2 \sqrt{\frac{d_{\mathcal{H}}}{n}} \right) \geq c.$$

This proves the lower bound for the case $4 \leq d_{\mathcal{H}} \leq 17$.

E Proof of Theorem 4 [Surrogate upper bound]

Proof. For $j \in \{0, 1\}$, consider the composed class

$$\phi \circ (-\mathcal{F}) := \{ x \mapsto \phi(-f(x)) : f \in \mathcal{F} \}.$$

By the contraction inequality (Bartlett & Mendelson, 2002, Theorem 12) and Assumption 1, for every $m \geq 1$,

$$\mathfrak{R}_m\big(\phi \circ (-\mathcal{F})\big) \leq 2L_\phi \mathfrak{R}_m(\mathcal{F}) \leq \frac{2L_\phi B_{\mathcal{F}}}{\sqrt{m}}. \tag{23}$$

Moreover, every function in $\phi \circ (-\mathcal{F})$ takes values in $[0, C_\phi]$. For each $j \in \{0, 1\}$, define

$$Z_j := \sup_{f \in \mathcal{F}} \left| R_{\phi,j}(f) - \widehat{R}_{\phi,j}(f) \right|.$$

For every integer $m \geq 1$, conditional on the event $\{n_j = m\}$, the observations $\{X_i : Y_i = j\}$ are i.i.d. from P_j. Therefore, the Rademacher uniform deviation inequality (Mohri et al., 2018, Theorem 3.3) implies that

$$\mathbb{P}\left(Z_j > 2\mathfrak{R}_m\big(\phi \circ (-\mathcal{F})\big) + C_\phi \sqrt{\frac{\log(6/\delta)}{2m}} \,\middle|\, n_j = m \right) \leq \frac{\delta}{3}.$$

Using equation 23, we obtain

$$\mathbb{P}\left(Z_j > \frac{4L_\phi B_{\mathcal{F}}}{\sqrt{m}} + C_\phi \sqrt{\frac{\log(6/\delta)}{2m}} \,\middle|\, n_j = m \right) = \mathbb{P}\left(Z_j > \epsilon_\phi(m, \delta) \,|\, n_j = m \right) \leq \frac{\delta}{3}.$$

Because $\mathcal{E}_n$ is determined by the class counts and implies $n_j \geq 1$, the law of total probability yields Then,

$$\begin{aligned} \mathbb{P}\left(Z_j > \epsilon_\phi(n_j, \delta) \,|\, \mathcal{E}_n \right) &= \sum_{m=1}^{n-1} \mathbb{P}\left(Z_j > \epsilon_\phi(m, \delta) \,|\, n_j = m \right) \mathbb{P}\left(n_j = m \,|\, \mathcal{E}_n \right) \\ &\leq \frac{\delta}{3} \sum_{m=1}^{n-1} \mathbb{P}\left(n_j = m \,|\, \mathcal{E}_n \right) \leq \frac{\delta}{3}. \end{aligned}$$

Thus, Defining the event $\mathcal{A} := \mathcal{E}_n \cap \{ Z_j \leq \epsilon_\phi(n_j, \delta) \text{ for } j = 0, 1 \}$, we have $\mathbb{P}\left(\mathcal{A} \,|\, \mathcal{E}_n \right) \geq 1 - \frac{2\delta}{3}$.

We next control the surrogate ambiguity objective. Define

$$\mathcal{G}_\diamond := \left\{ x \mapsto \ell^\diamond_{\mathrm{amb},\phi}(f_0, f_1; x) : (f_0, f_1) \in \mathcal{F} \times \mathcal{F} \right\}.$$

For the additive surrogate ambiguity loss, the fact that the positive-part map is 1-Lipschitz gives

$$|[u + v - 1]_+ - [u' + v' - 1]_+| \le |u - u'| + |v - v'|.$$

Consequently,

$$\begin{aligned}\left|\ell^{\mathrm{add}}_{\mathrm{amb},\phi}(f_0, f_1; x) - \ell^{\mathrm{add}}_{\mathrm{amb},\phi}(g_0, g_1; x)\right| &\le |\phi(f_0(x)) - \phi(g_0(x))| + |\phi(f_1(x)) - \phi(g_1(x))| \\ &\le L_\phi \left(|f_0(x) - g_0(x)| + |f_1(x) - g_1(x)| \right).\end{aligned}$$

Moreover, $0 \le \ell^{\mathrm{add}}_{\mathrm{amb},\phi}(f_0, f_1; x) \le 2C_\phi$. For the multiplicative surrogate ambiguity loss, we have

$$\begin{aligned}&|\phi(f_0(x))\phi(f_1(x)) - \phi(g_0(x))\phi(g_1(x))| \\ &\qquad \le \phi(f_1(x)) \, |\phi(f_0(x)) - \phi(g_0(x))| + \phi(g_0(x)) \, |\phi(f_1(x)) - \phi(g_1(x))| \\ &\qquad \le C_\phi L_\phi \left(|f_0(x) - g_0(x)| + |f_1(x) - g_1(x)| \right),\end{aligned}$$

Also, $0 \le \ell^{\mathrm{prod}}_{\mathrm{amb},\phi}(f_0, f_1; x) \le C_\phi^2$. Recalling that $\kappa_{\mathrm{add}} = 1$, and $\kappa_{\mathrm{prod}} = C_\phi$, the vector contraction inequality (Maurer, 2016, Corollary 1 and Section 4.1) and Assumption 1 imply that there exists a universal constant $C > 0$ such that

$$\mathfrak{R}_n(\mathcal{G}_\diamond) \le C \kappa_\diamond L_\phi \left(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{F}) \right) \le C \frac{\kappa_\diamond L_\phi B_\mathcal{F}}{\sqrt{n}},$$

Define

$$Z^\diamond_{\mathrm{amb}} := \sup_{(f_0, f_1) \in \mathcal{F} \times \mathcal{F}} \left| R^\diamond_{\mathrm{amb},\phi}(f_0, f_1) - \widehat{R}^\diamond_{\mathrm{amb},\phi}(f_0, f_1) \right|.$$

Since every function in $\mathcal{G}_\diamond$ takes values in $[0, M_\diamond]$, the Rademacher uniform deviation inequality gives

$$\mathbb{P}\left(Z^\diamond_{\mathrm{amb}} > 2\mathfrak{R}_n(\mathcal{G}_\diamond) + M_\diamond \sqrt{\frac{\log(6/\delta)}{2n}} \right) = \mathbb{P}\left(Z^\diamond_{\mathrm{amb}} > \epsilon^\diamond_{\mathrm{amb},\phi}(n, \delta) \right) \le \frac{\delta}{3}.$$

Let $\mathcal{A}^\diamond_{\mathrm{amb}} := \left\{ Z^\diamond_{\mathrm{amb}} \le \epsilon^\diamond_{\mathrm{amb},\phi}(n, \delta) \right\}$. Then we obtain

$$\mathbb{P}\left(\mathcal{A} \cap \mathcal{A}^\diamond_{\mathrm{amb}} \,|\, \mathcal{E}_n \right) \ge 1 - \frac{2\delta}{3} - \frac{\delta}{3\mathbb{P}(\mathcal{E}_n)}$$

It remains to establish the claimed conclusions, conditional on $\mathcal{E}_n$, on the event $\mathcal{A} \cap \mathcal{A}^\diamond_{\mathrm{amb}}$. On $\mathcal{A}$, for $j \in \{0, 1\}$ we have

$$\widehat{R}_{\phi,j}(f_j^{\diamond,*}) \le R_{\phi,j}(f_j^{\diamond,*}) + \epsilon_\phi(n_j, \delta) \le \alpha_j + \epsilon_\phi(n_j, \delta).$$

Thus, $f^{\diamond,*} = (f_0^{\diamond,*}, f_1^{\diamond,*})$ is feasible for the empirical optimization problem. By the empirical optimality of $\widehat{f}^\diamond = (\widehat{f}^\diamond_0, \widehat{f}^\diamond_1)$, we have $\widehat{R}^\diamond_{\mathrm{amb},\phi}(\widehat{f}^\diamond_0, \widehat{f}^\diamond_1) \le \widehat{R}^\diamond_{\mathrm{amb},\phi}(f_0^{\diamond,*}, f_1^{\diamond,*})$. On $\mathcal{A}^\diamond_{\mathrm{amb}}$, it follows that

$$\begin{aligned}R^\diamond_{\mathrm{amb},\phi}(\widehat{f}^\diamond_0, \widehat{f}^\diamond_1) \le \widehat{R}^\diamond_{\mathrm{amb},\phi}(\widehat{f}^\diamond_0, \widehat{f}^\diamond_1) + \epsilon^\diamond_{\mathrm{amb},\phi}(n, \delta) &\le \widehat{R}^\diamond_{\mathrm{amb},\phi}(f_0^{\diamond,*}, f_1^{\diamond,*}) + \epsilon^\diamond_{\mathrm{amb},\phi}(n, \delta) \\ &\le R^\diamond_{\mathrm{amb},\phi}(f_0^{\diamond,*}, f_1^{\diamond,*}) + 2\epsilon^\diamond_{\mathrm{amb},\phi}(n, \delta).\end{aligned}$$

Therefore, $\mathcal{E}^\diamond_{\mathrm{amb},\phi}(\widehat{f}^\diamond_0, \widehat{f}^\diamond_1) \le 2\epsilon^\diamond_{\mathrm{amb},\phi}(n, \delta)$. Finally, empirical feasibility gives

$$\widehat{R}_{\phi,j}(\widehat{f}^\diamond_j) \le \alpha_j + \epsilon_\phi(n_j, \delta), \qquad j \in \{0, 1\}.$$

Hence, on $\mathcal{A}$,

$$R_{\phi,j}(\widehat{f}^\diamond_j) \le \widehat{R}_{\phi,j}(\widehat{f}^\diamond_j) + \epsilon_\phi(n_j, \delta) \le \alpha_j + 2\epsilon_\phi(n_j, \delta).$$

Since the surrogate loss upper bounds the corresponding indicator loss, $R_j(h_{\widehat{f}^\diamond_j}) \le R_{\phi,j}(\widehat{f}^\diamond_j)$, and therefore

$$R_j(h_{\widehat{f}^\diamond_j}) \le \alpha_j + 2\epsilon_\phi(n_j, \delta), \qquad j \in \{0, 1\}.$$

This completes the proof. □

F Proof of Theorem 5

Proof. According to Mahdavi et al. (2012), Algorithm 1 guarantees that, conditional on the training sample, with probability at least $1-\delta$, we have

$$\widehat{R}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\widehat{\theta}_0}, f_{\widehat{\theta}_1}) - \widehat{R}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\widehat{\theta}_0^\star}, f_{\widehat{\theta}_1^\star}) \lesssim \frac{1}{\sqrt{T}} \left[\left(1 + \frac{G_1^2}{\rho^2}\right) c_\delta + \frac{G_1^2 G_2^2}{\rho^2 c_\delta} + G_1 \sqrt{\ln \frac{2}{\delta}} \right].$$

Taking

$$T \gtrsim \frac{1}{\left(\epsilon^{\mathrm{add}}_{\mathrm{amb},\phi}(n,\delta)\right)^2} \left[\left(1 + \frac{G_1^2}{\rho^2}\right) c_\delta + \frac{G_1^2 G_2^2}{\rho^2 c_\delta} + G_1 \sqrt{\ln \frac{2}{\delta}} \right]^2$$

with a sufficiently large implicit constant ensures that

$$\widehat{R}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\widehat{\theta}_0}, f_{\widehat{\theta}_1}) - \widehat{R}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\widehat{\theta}_0^\star}, f_{\widehat{\theta}_1^\star}) \le \epsilon^{\mathrm{add}}_{\mathrm{amb},\phi}(n,\delta).$$

Following the same uniform-deviation argument as in the proof of Theorem 4, this additional optimization error yields

$$\mathcal{E}^{\mathrm{add}}_{\mathrm{amb},\phi}(f_{\widehat{\theta}_0}, f_{\widehat{\theta}_1}) \le 3\epsilon^{\mathrm{add}}_{\mathrm{amb},\phi}(n,\delta).$$

Since $\widehat{\theta} \in K$, the empirical constraints hold, so the feasibility guarantees follow exactly as in Theorem 4. A union bound accounts for the additional failure probability of at most δ. □

G Auxiliary Results

Proof of Proposition 3. Fix an arbitrary randomized estimator $\widetilde{\theta} = \widetilde{\theta}(S^n, \zeta)$. Let ν denote the distribution of ζ. Under parameter θ, the joint law of (S^n, ζ) is $P_\theta^{\otimes n} \otimes \nu$. For every $j = 1, ..., M$,

$$D_{\mathrm{KL}}\left(P_{\theta_j}^{\otimes n} \otimes \nu \,\middle\|\, P_{\theta_0}^{\otimes n} \otimes \nu\right) = D_{\mathrm{KL}}\left(P_{\theta_j}^{\otimes n} \,\middle\|\, P_{\theta_0}^{\otimes n}\right).$$

Therefore,

$$\frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}\left(P_{\theta_j}^{\otimes n} \otimes \nu \,\middle\|\, P_{\theta_0}^{\otimes n} \otimes \nu\right) \le \eta \log M.$$

Now view $\widetilde{\theta}(S^n, \zeta)$ as an estimator based on the enlarged observation (S^n, ζ). Applying the usual Fano/Tsybakov lower bound [Tsybakov (2009), Theorem 2.5] to the family of distributions $\left\{P_{\theta_j}^{\otimes n} \otimes \nu : j = 0, 1, \ldots, M\right\}$, we obtain

$$\sup_{0 \le j \le M} \left(P_\theta^{\otimes n} \otimes \nu\right)\left(d\big(\widetilde{\theta}(S^n, \zeta), \theta\big) \ge s\right) \ge L_{M,\eta}.$$

Equivalently,

$$\sup_{0 \le j \le M} \mathbb{E}_{S^n \sim P_{\theta_j}^{\otimes n}} \mathbb{E}_\zeta \left[\mathbf{1}\left\{d\big(\widetilde{\theta}(S^n, \zeta), \theta_j\big) \ge s\right\}\right] \ge L_{M,\eta}.$$

Hence there exists some $j^\star \in \{0, \ldots, M\}$ such that

$$\mathbb{E}_{\theta_{j^\star}} \mathbb{E}_\zeta \left[\mathbf{1}\left\{d\big(\widetilde{\theta}(S^n, \zeta), \theta_{j^\star}\big) \ge s\right\}\right] \ge L_{M,\eta}.$$

Define

$$Q(S^n) := \mathbb{E}_\zeta \left[\mathbf{1}\left\{d\big(\widetilde{\theta}(S^n, \zeta), \theta_{j^\star}\big) \ge s\right\}\right].$$

For any $t \in (0, L_{M,\eta})$, we have

$$L_{M,\eta} \le \mathbb{E}_{\theta_{j^\star}}[Q(S^n)] \le t + (1-t) P_{\theta_{j^\star}}^{\otimes n}\left(Q(S^n) \ge t\right).$$

Therefore, $P_{\theta_{j^*}}^{\otimes n}\left(Q(S^n) \ge t\right) \ge \frac{L_{M,\eta} - t}{1-t}$. For $t = \frac{L_{M,\eta}}{2}$, we get

$$P_{\theta_{j^*}}^{\otimes n}\left(Q(S^n) \ge \frac{L_{M,\eta}}{2}\right) \ge \frac{L_{M,\eta}}{2 - L_{M,\eta}}.$$

Now observe that, given S^n,

$$\mathbb{E}_\zeta\left[d\big(\widetilde{\theta}(S^n,\zeta),\theta_{j^*}\big)\right] \geq sQ(S^n).$$

Hence,

$$\left\{Q(S^n) \geq \frac{L_{M,\eta}}{2}\right\} \subseteq \left\{\mathbb{E}_\zeta\left[d\big(\widetilde{\theta}(S^n,\zeta),\theta_{j^*}\big)\right] \geq \frac{sL_{M,\eta}}{2}\right\}.$$

Therefore,

$$P_{\theta_{j^*}}^{\otimes n}\left(\mathbb{E}_\zeta\left[d\big(\widetilde{\theta}(S^n,\zeta),\theta_{j^*}\big)\right] \geq \frac{sL_{M,\eta}}{2}\right) \geq \frac{L_{M,\eta}}{2-L_{M,\eta}}.$$

This holds for every randomized estimator $\widetilde{\theta}$. Taking the infimum over all randomized estimators gives

$$\inf_{\widetilde{\theta}} \sup_{\theta\in\Theta} P_\theta^{\otimes n}\left(\mathbb{E}_\zeta\left[d\big(\widetilde{\theta}(S^n,\zeta),\theta\big)\right] \geq \frac{sL_{M,\eta}}{2}\right) \geq \frac{L_{M,\eta}}{2-L_{M,\eta}}.$$

This concludes the proof. □

Lemma 1. *Let $\mathcal{H}$ be a hypothesis class with $0 < d_\mathcal{H} < \infty$, and define*

$$\mathcal{G} := \{g_{h_0,h_1} : x \mapsto h_0(x)h_1(x) \ : \ h_0, h_1 \in \mathcal{H}\}.$$

Then $d_\mathcal{G} \leq Cd_\mathcal{H}$ for a universal constant $C > 0$.

Proof. Let $\Pi_\mathcal{H}(m)$ and $\Pi_\mathcal{G}(m)$ denote the growth functions of $\mathcal{H}$ and $\mathcal{G}$, respectively. For any fixed points $x_1, \ldots, x_m$, each labeling induced by a function $g_{h_0,h_1} \in \mathcal{G}$ is completely determined by the two labelings

$$\big(h_0(x_1), \ldots, h_0(x_m)\big) \qquad \text{and} \qquad \big(h_1(x_1), \ldots, h_1(x_m)\big).$$

Therefore, $\Pi_\mathcal{G}(m) \leq \Pi_\mathcal{H}(m)^2$. By Sauer's lemma, for $m \geq d_\mathcal{H}$, we have $\Pi_\mathcal{H}(m) \leq \left(\frac{em}{d_\mathcal{H}}\right)^{d_\mathcal{H}}$, and hence $\Pi_\mathcal{G}(m) \leq \left(\frac{em}{d_\mathcal{H}}\right)^{2d_\mathcal{H}}$. Now suppose that $\mathcal{G}$ shatters m points. If $m < d_\mathcal{H}$, then trivially $m = O(d_\mathcal{H})$. Otherwise, $m \geq d_\mathcal{H}$, and since $\Pi_\mathcal{G}(m) = 2^m$, we obtain $2^m \leq \left(\frac{em}{d_\mathcal{H}}\right)^{2d_\mathcal{H}}$. Taking $m = 10d_\mathcal{H}$, we obtain $2^{10d_\mathcal{H}} \leq (10e)^{2d_\mathcal{H}}$, which is impossible. Thus $\mathcal{G}$ cannot shatter $10d_\mathcal{H}$ points, and consequently $d_\mathcal{G} < 10d_\mathcal{H}$. This proves $d_\mathcal{G} \leq Cd_\mathcal{H}$ for a universal constant C. □

H Additional Experimental Details

This appendix complements the experiments in Section 7, providing dataset descriptions, preprocessing procedures, model architectures, training details, and a more detailed discussion of the results for each dataset. We compare the additive and product surrogate formulations with the method of Lei (2014), which constructs classification regions by thresholding nonparametric estimates of conditional class probabilities.

H.1 Vertebral Column Dataset

The Vertebral Column dataset (Barreto & Neto, 2005) contains 310 patients described by six biomechanical features, with 100 normal and 210 pathological cases. We use stratified splits of 217 training and 93 test observations, standardizing features using training-sample statistics. We compare a three-layer MLP with hinge loss under the additive and product formulations against Lei's method (Lei, 2014), with class-conditional error levels set to $\alpha_0 = \alpha_1 = 0.05$. We report the mean and standard deviation of test ambiguity risk and class-conditional errors over ten runs.

As shown in Table 2, both surrogate formulations achieve lower mean ambiguity risk than Lei's method (Lei, 2014), with class-conditional errors close to the target level of 0.05. The product formulation achieves the lowest mean ambiguity risk (0.2591, compared with 0.3086 for the additive formulation and 0.3505 for Lei's method (Lei, 2014)), with slightly higher mean class-conditional errors than the additive formulation.

Table 2: Test performance on the Vertebral Column dataset (Barreto & Neto, 2005) with target class-conditional error levels $\alpha_0 = \alpha_1 = 0.05$. We compare Lei's method (Lei, 2014) with three-layer MLPs trained using hinge loss under the additive and product surrogate formulations. Values are means $\pm$ standard deviations over ten runs. R_0 and R_1 denote class-conditional errors, and R_{amb} denotes ambiguity risk.

Method	R_0	R_1	R_{amb}
Lei (Lei, 2014)	0.0333 ± 0.0351	0.0524 ± 0.0290	0.3505 ± 0.1152
MLP, hinge, additive	0.0400 ± 0.0344	0.0524 ± 0.0375	0.3086 ± 0.1153
MLP, hinge, product	0.0467 ± 0.0358	0.0540 ± 0.0383	0.2591 ± 0.1029

Table 3: Test performance on the MAGIC Gamma Telescope dataset (Bock, 2004) with target class-conditional error levels $\alpha_0 = \alpha_1 = 0.05$. We compare Lei's method (Lei, 2014) with three-layer MLPs trained using hinge loss under the additive and product surrogate formulations. Values are means $\pm$ standard deviations over ten runs. R_0 and R_1 denote class-conditional errors, and R_{amb} denotes ambiguity risk.

Method	R_0	R_1	R_{amb}
Lei (Lei, 2014)	0.0501 ± 0.0046	0.0485 ± 0.0064	0.5966 ± 0.0197
MLP, hinge, additive	0.0316 ± 0.0204	0.0520 ± 0.0232	0.5107 ± 0.0889
MLP, hinge, product	0.0418 ± 0.0091	0.0493 ± 0.0092	0.4332 ± 0.0263

H.2 MAGIC Gamma Telescope Dataset

The MAGIC Gamma Telescope dataset (Bock, 2004) contains 19,020 simulated atmospheric-shower events, each described by ten continuous features. The task is to distinguish gamma showers (12,332 observations) from hadronic showers (6,688 observations). We use a stratified split of 13,314 training and 5,706 test observations and compare three-layer MLPs with hinge loss under the additive and product formulations against Lei's method (Lei, 2014), with $\alpha_0 = \alpha_1 = 0.05$.

As shown in Table 3, both formulations achieve lower mean ambiguity risk than Lei's method (Lei, 2014). The product formulation performs better than the additive formulation in terms of mean ambiguity risk (0.4332 versus 0.5107), while keeping both mean class-conditional errors below 0.05.

H.3 BCI Competition IV Dataset 2b

The BCI Competition IV Dataset 2b (Leeb et al., 2008) contains EEG recordings from nine subjects performing left- and right-hand motor-imagery tasks. We report results for Subject 4, whose original competition data comprise 400 training trials and 320 evaluation trials. Our experiments use a stratified 70%/30% training–test split, with preprocessing and feature parameters estimated from the training data. We compare Lei's method with linear and MLP models using hinge or squared hinge loss under the additive and product formulations, with $\alpha_0 = \alpha_1 = 0.05$.

As shown in Table 4, the linear model with hinge loss and the product formulation achieves a low mean ambiguity risk of 0.0077, with mean class-conditional errors of 0.0505 and 0.0450, approximately meeting the prescribed thresholds. This compares favorably with the mean ambiguity risk of 0.2293 achieved by the method of Lei (Lei, 2014).

H.4 Phoneme Dataset

Thew Phoneme dataset (pho) from OpenML contains 5,404 observations described by five acoustic features, with the task of distinguishing nasal from oral vowel sounds. We use a stratified 70%/30% training–test split and compare MLP models with hinge loss under the additive and product formulations against the method of Lei (Lei, 2014), with $\alpha_0 = \alpha_1 = 0.05$.

Table 4: Test performance for Subject 4 of the BCI Competition IV Dataset 2b (Leeb et al., 2008) with target class-conditional error levels $\alpha_0 = \alpha_1 = 0.05$. We compare Lei's method (Lei, 2014) with linear and MLP models using hinge or squared hinge loss under the additive and product formulations. Values are means $\pm$ standard deviations over ten runs. R_0 and R_1 denote class-conditional errors, and R_{amb} denotes ambiguity risk.

Method	R_0	R_1	R_{amb}
Lei (Lei, 2014)	0.0351 ± 0.0361	0.0252 ± 0.0163	0.2293 ± 0.4065
Linear, hinge, product	0.0505 ± 0.0279	0.0450 ± 0.0147	0.0077 ± 0.0090
Linear, hinge, additive	0.0459 ± 0.0242	0.0450 ± 0.0147	0.0117 ± 0.0098
MLP, hinge, product	0.0658 ± 0.0321	0.0396 ± 0.0114	0.0072 ± 0.0085
MLP, hinge, additive	0.0523 ± 0.0268	0.0495 ± 0.0196	0.0050 ± 0.0113
MLP, squared hinge, product	0.0459 ± 0.0270	0.0423 ± 0.0113	0.0158 ± 0.0133

Table 5: Test performance on the Phoneme dataset (pho) with target class-conditional error levels $\alpha_0 = \alpha_1 = 0.05$. We compare the method of Lei (Lei, 2014) with MLP models trained using hinge loss under the additive and product formulations. Values are means $\pm$ standard deviations over ten runs. R_0 and R_1 denote class-conditional errors, and R_{amb} denotes ambiguity risk.

Method	R_0	R_1	R_{amb}
Lei (Lei, 2014)	0.0531 ± 0.0122	0.0540 ± 0.0069	0.5606 ± 0.0257
MLP, hinge, product	0.0551 ± 0.0204	0.0479 ± 0.0151	0.4482 ± 0.0341
MLP, hinge, additive	0.0387 ± 0.0308	0.0462 ± 0.0170	0.5292 ± 0.0607

As shown in Table 5, the additive formulation achieves mean class-conditional errors below both the prescribed thresholds and those of Lei's method (Lei, 2014), while also yielding a lower mean ambiguity risk (0.5292 versus 0.5606). The product formulation further reduces mean ambiguity risk to 0.4482, although its mean $R_0 = 0.0551$ slightly exceeds the target level.

H.5 Sensorless Drive Diagnosis Dataset

The Sensorless Drive Diagnosis dataset (Bator, 2013) contains 48 features extracted from motor-current signals under different operating conditions. We consider classes 6 and 9, yielding 10,638 observations equally divided between the two classes, with 7,446 training and 3,192 test observations. We compare MLP models with hinge loss under the additive and product formulations against the method of Lei (Lei, 2014), with $\alpha_0 = \alpha_1 = 0.05$.

As shown in Table 6, the product formulation achieves the lowest mean ambiguity risk (0.0166, compared with 0.0644 for the additive formulation and 0.2833 for Lei (Lei, 2014)). Both formulations yield mean R_0 below the prescribed threshold, while their mean R_1 slightly exceeds it.

H.6 Synthetic Gaussian Data

We consider a binary classification problem in dimension $d = 10$, with class-conditional distributions

$$X \mid Y = 0 \sim \mathcal{N}(\mathbf{0}_{10}, I_{10}), \qquad X \mid Y = 1 \sim \mathcal{N}(0.5\mathbf{1}_{10}, 2I_{10}).$$

We generate 10,000 observations, equally divided between the two classes, and use a stratified split of 7,000 training and 3,000 test observations. We compare MLP models using hinge and squared hinge losses under the additive and product formulations against the method of Lei (Lei, 2014), with $\alpha_0 = \alpha_1 = 0.05$.

As shown in Table 7, the additive formulation with hinge loss achieves mean class-conditional errors below both prescribed thresholds and those of Lei's method (Lei, 2014), while also yielding a lower mean ambiguity risk (0.5056 versus 0.6038). The product formulation with squared hinge loss reduces mean ambiguity risk to 0.4415, with mean $R_0 = 0.0512$ slightly above the target and mean $R_1 = 0.0479$ below it.

Table 6: Test performance on the Sensorless Drive Diagnosis dataset (Bator, 2013) (classes 6 versus 9) with target class-conditional error levels $\alpha_0 = \alpha_1 = 0.05$. We compare the method of Lei (Lei, 2014) with MLP models trained using hinge loss under the additive and product formulations. Values are means $\pm$ standard deviations over ten runs. R_0 and R_1 denote class-conditional errors, and R_{amb} denotes ambiguity risk.

Method	R_0	R_1	R_{amb}
Lei (Lei, 2014)	0.0538 ± 0.0093	0.0505 ± 0.0059	0.2833 ± 0.1541
MLP, hinge, product	0.0437 ± 0.0154	0.0543 ± 0.0116	0.0166 ± 0.0123
MLP, hinge, additive	0.0441 ± 0.0203	0.0527 ± 0.0127	0.0644 ± 0.0527

Table 7: Test performance on the 10-dimensional Gaussian dataset with $X \mid Y = 0 \sim \mathcal{N}(\mathbf{0}_{10}, I_{10})$ and $X \mid Y = 1 \sim \mathcal{N}(0.5\mathbf{1}_{10}, 2I_{10})$, using 7,000 training observations and target error levels $\alpha_0 = \alpha_1 = 0.05$. Values are means $\pm$ standard deviations over ten runs. R_0 and R_1 denote class-conditional errors, and R_{amb} denotes ambiguity risk.

Method	R_0	R_1	R_{amb}
Lei (Lei, 2014)	0.0485 ± 0.0051	0.0510 ± 0.0104	0.6038 ± 0.0207
MLP, hinge, product	0.0611 ± 0.0092	0.0519 ± 0.0103	0.4089 ± 0.0392
MLP, hinge, additive	0.0473 ± 0.0173	0.0490 ± 0.0195	0.5056 ± 0.0921
MLP, squared hinge, product	0.0512 ± 0.0109	0.0479 ± 0.0098	0.4415 ± 0.0441
MLP, squared hinge, additive	0.0655 ± 0.0193	0.0496 ± 0.0174	0.4565 ± 0.0754