Ask for More Than Bayes Optimal: A Theory of Indecisions for Classification

Mohamed Ndaoud[1], Peter Radchenko[2], and Bradley Rava[2]

Abstract

Selective classification is a powerful tool for automated decision-making in high-risk scenarios, allowing classifiers to act only when confident and abstain when uncertainty is high. Given a target accuracy, our goal is to minimize indecisions, observations we do not automate. For difficult problems, the target accuracy may be unattainable without abstention. By using indecisions, we can control the misclassification rate to any user-specified level, even below the Bayes optimal error rate, while minimizing overall indecision mass.

We provide a complete characterization of the minimax risk in selective classification, establishing continuity and monotonicity properties that enable optimal indecision selection. We revisit selective inference via the Neyman–Pearson testing framework, where indecision enables control of type II error given fixed type I error probability. For both classification and testing, we propose a finite-sample calibration method with non-asymptotic guarantees, proving plug-in classifiers remain consistent and that accuracy-based calibration effectively controls indecision mass. In the binary Gaussian mixture model, we uncover the first sharp phase transition in selective inference, showing minimal indecision can yield near-optimal accuracy even under poor class separation. Experiments on Gaussian mixtures and real datasets confirm that small indecision proportions yield substantial accuracy gains, making indecision a principled tool for risk control.

Keywords: Selective Inference, Finite-sample Calibration, Indecision, Phase Transition.

1 Introduction

We address the problem of controlling a classifier's accuracy at any user-specified level through selective classification, regardless of the problem's inherent difficulty. Traditional classification frameworks are designed to approximate the Bayes optimal error rate as closely as possible. However, with the growing deployment of artificial intelligence (AI) systems in automated, high-stakes decision-making, it has become critical to ensure reliable control over a classifier's accuracy and to guarantee accurate predictions for all individuals.

When the underlying problem is truly difficult, achieving control over the error rate of an automated decision-making system may be impossible. This is particularly true when the number of potential classes is large or when the distributions of these classes are close enough, significantly increasing the difficulty of the problem. This phenomenon is illustrated in Figure 1, where the task

[1]Department of Decisions Sciences, ESSEC Business School, ndaoud@essec.edu

[2]Discipline of Business Analytics, University of Sydney Business School, peter.radchenko@sydney.edu.au, bradley.rava@sydney.edu.au

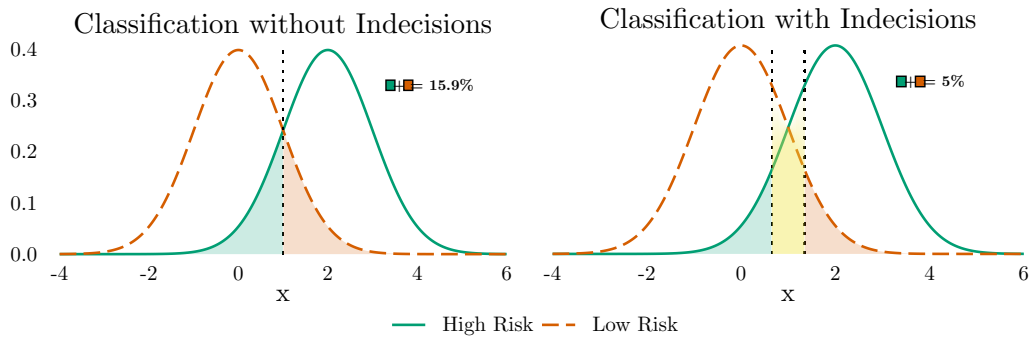

Figure 1: An example of a classification scenario where the data comes from two different normal distributions. Low Risk $\sim N(0,1)$ and High Risk $\sim N(2,1)$. Left plot: Classification with no indecisions. Right Plot: Classification with indecisions (highlighted in yellow). The indecisions do not contribute to the risk of our classifier. By including the indecisions, we are able to obtain a much lower specified level of control over the risk.

is to classify various observations as High-Risk or Low-Risk, while maintaining an error rate of 5%. In this example, the High-Risk and Low-Risk classes are modeled as mixtures of two normal distributions with means of 2 and 0, respectively, and a common variance of 1. The Bayes optimal decision boundary is represented by the dotted line in the leftmost plot of Figure 1.

In this scenario, the Bayes optimal error rate is 15.9%, significantly exceeding our target classification error of 5%. To achieve the desired level of accuracy, it becomes necessary to identify the most challenging observations to classify and abstain from making decisions on them, opting instead for an indecision. The traditional classification approach is depicted in the leftmost plot of Figure 1, while our proposed solution is illustrated in the rightmost plot. In both cases, the misclassification rate is represented by the shaded regions under the High-Risk (green / solid line) and Low-Risk (orange / dashed line) density curves.

This selective classification framework enables us to achieve *any* desired level of accuracy from an automated decision-making system. In the example shown in Figure 1, the misclassification rate within the selected region in the rightmost plot can be precisely 5%, whereas the leftmost plot is limited by the minimum achievable classification error of the Bayes classifier, which in this case is approximately 15.9%.

Indecisions are observations that are intentionally excluded from automated classification because their inherent difficulty prevents the algorithm from achieving the desired level of accuracy. This approach is particularly valuable in high risk decision-making scenarios, as these observations, lacking sufficient confidence for automated classification, can instead be referred for human review. This process facilitates effective Human-AI interaction by ensuring that only confident decisions are automated, while challenging cases are escalated for manual evaluation. Importantly, indecisions do not contribute to the classifier's error rate, allowing practitioners to reliably control the accuracy of the system while efficiently allocating human oversight to the most critical cases.

Critically, the use of indecisions through selective classification is most valuable when the desired accuracy cannot be achieved by the Bayes classifier, and consequently, by any standard classifier. We illustrate this tradeoff in Figure 2, using the same simulation setup as in Figure 1, but now varying the distance between class means, denoted by $\Delta = |\mu_{\text{High Risk}} - \mu_{\text{Low Risk}}|/2$. The figure

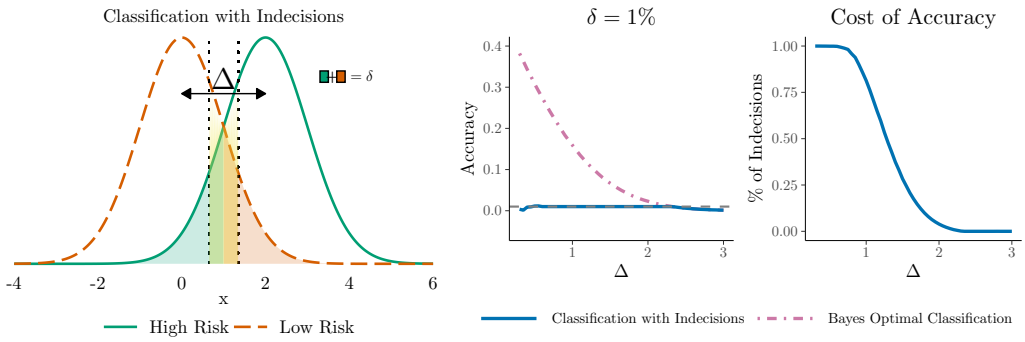

Figure 2: The best level of accuracy that can be obtained by the bayes classifier and classification with indecisions as the distance between the underlying data distributions gets further apart. Δ represents the amount of separation between the High and Low risk classes. A larger Δ means that the classification problem is easier.

highlights an ideal use case for indecisions: our target accuracy level of 1% is unattainable until the level of separation is greater than 4.

Figure 2 is based on one million simulated observations, with the threshold for the indecision region determined on an independent i.i.d. dataset to exactly enforce a 1% error rate among automatically classified cases.[1] When the two class distributions overlap substantially, the Bayes error exceeds our target misclassification rate of 1%. Selective classification, however, can effectively always maintain this exact error rate, and consequently the level of accuracy, by rejecting the most uncertain cases. Once the class separation is sufficiently large, the Bayes error rate falls below 1%, at which point even standard classifiers could succeed. The tradeoff is made explicit in the rightmost panel of Figure 2, which shows the proportion of indecisions required to achieve the desired accuracy.

Optimal selective classifiers should control the error rate of their automatic decisions exactly, while using as few indecisions as possible. Achieving an error rate of 0 is trivial since one could always assign all observations as indecisions, but this is largely unhelpful if the same error rate could be achieved with a smaller amount of indecisions. Instead, our goal is to match the user's target error rate with the minimal necessary amount of indecisions. This design reflects a practitioner's tolerance for mistakes and parallels advances in selective classification and multiple hypothesis testing (Benjamini and Hochberg, 1995; Sun and Wei, 2015; Gang et al., 2022; Wang et al., 2024; Rava et al., 2025). In practice, this framework enables users to specify how much risk they are willing to take on, while still automatically classifying as many subjects as possible.

We notice several motivating features that emerge from this example. When class separation is small, the Bayes error rate can be extremely high, sometimes exceeding 40%. By contrast, our selective approach holds the error rate exactly at the specified level of 1%. As separation increases, the Bayes error rate eventually drops below 1%, in which case practitioners could either relax the error constraint or dispense with indecisions altogether. Yet, making such adjustments requires understanding how class separation interacts with the number of required indecisions, an aspect often overlooked in practice.

The rest of this paper is structured as follows. In Sections 1.1 and 1.2 we formulate our problem

[1] Although illustrated with Gaussian mixtures, the phenomenon extends to arbitrary data distributions.

and give an extensive literature review on related fields. In Sections 2 and 3 we derive minimax optimal classifiers for the general binary classification setting and for the hypothesis testing setting in order to control the type I and type II error of our classifier. This is done for both the case of a fixed proportion of indecisions as well as a fixed level of accuracy, while also providing a general algorithm for calibration in finite sample. Section 4 further provides a theoretical analysis about the sharp phase transition under the Gaussian mixture model, in order to give insight into when practitioners can expect dramatic gains in accuracy for virtually no indecisions. This is followed by an analysis of plug-in rules where the conditional density function η is learned. Section 6 presents two simulation studies, the first showing that we can recover the phase transition presented earlier in theory and the second demonstrates our finite sample algorithm for the hypothesis testing setting. The second simulation study demonstrates that even if a practitioner cannot calibrate the type II error threshold well, any amount of indecisions will nearly always improve upon classifiers that do not use indecisions. Building upon the simulation analysis, Section 7 replicates our hypothesis testing setting with real data from the COMPAS algorithm used to predict recidivists (Angwin et al., 2016). Section 5 is dedicated to extensions of our method under the maximum likelihood ratio property as well as presenting a minimax framework for the multi-classification setting. Proofs of all theorems are in the appendix.

1.1 Problem Formulation

We observe a random variable X on a measurable space $(\mathcal{X},\mathcal{U})$ such that X is distributed according to a mixture model, where with probability p_1 its probability measure is given by P_1 and with probability p_2 its probability measure P_2. We assume that $P_2 \neq P_1$. Let f_1 and f_2 be densities of P_1 and P_2 with respect to some dominating measure that we will further denote by μ. Denote by Y the labeling quantity such that $Y = 1$ if the distribution of X is P_1 and $Y = 2$ if it is P_2. We are interested in the problem of predicting the true label Y with an estimator $\widehat{Y}$, with the quality of our estimate measured either conditional on making a decision through supervised classification, or conditional on the true label Y through controlling type I and type II errors, otherwise known as hypothesis testing.

As estimators of Y, we consider any measurable functions $\widehat{Y} = \widehat{Y}(X)$ taking values in $\{1, 2\}$. Such estimators will be called *classifiers*. We define the loss of a classifier $\widehat{Y}$ as the indicator of whether a mistake is made, that is $\mathbf{1}\{\widehat{Y} \neq Y\}$, where $\mathbf{1}(\cdot)$ is the indicator function. The performance of $\widehat{Y}$ is measured by its expected risk $\mathbf{P}_Y(\widehat{Y} \neq Y)$, also known as classification error or misclassification rate, or by $\mathbf{P}_Y(\widehat{Y} = Y)$, referred to as accuracy.

We denote by $\mathbf{E}_Y$ the expectation with respect to probability measure $\mathbf{P}_Y$ of X with labeling Y. Observe that $\mathbf{P}_Y(\widehat{Y} \neq Y) = p_1\mathbf{P}_1(\widehat{Y} = 2) + p_2\mathbf{P}_2(\widehat{Y} = 1)$, which is a weighted sum of the type I and type II errors. The classical theory of classification gives a precise characterization of the minimax risk, $\inf_{\tilde{Y}} \mathbf{P}_Y(\tilde{Y} \neq Y)$, where $\inf_{\tilde{Y}}$ denotes the infimum over all measurable classifiers. In particular, it is well known that the optimal classifier is given by the Bayes classification rule Y^* defined as $Y^*(X) = (3 - \operatorname{sign}(p_1 f_1(X) - p_2 f_2(X)))/2$. Moreover, the corresponding risk is given by

$$\inf_{\tilde{Y}} \mathbf{P}_Y(\tilde{Y} \neq Y) = \mathbf{P}_Y(Y^* \neq Y) = \int (p_1 f_1 \wedge p_2 f_2) d\mu = \frac{1}{2} - \frac{1}{2}\int |p_1 f_1 - p_2 f_2| d\mu.$$

In particular, the minimax risk is bounded from below by a quantity that represents the separation

between the two distributions. When f_1 and f_2 are close, any classifier preforms poorly, which serves as motivation for the present work. Our goal is to introduce and study a framework where arbitrarily large accuracy can be achieved with the help of indecisions.

In order to break the statistical barrier given by the Bayes risk, we allow our estimator a degree of freedom where it only makes a decision when it is sufficiently confident. Depending on the targeted accuracy level, the classifier may have to discard some of the observations. More precisely, given an *indecision level* γ, we will consider the new risk:

$$\mathcal{R}(\gamma) := \inf_{\tilde{Y}_\gamma} \mathbf{P}_Y \left(\tilde{Y}_\gamma \neq Y | \tilde{Y}_\gamma \neq 0 \right), \tag{1}$$

where $\inf_{\tilde{Y}_\gamma}$ denotes the infimum over all classifiers taking values in $\{0, 1, 2\}$ such that $\mathbf{P}(\tilde{Y}_\gamma = 0) = \gamma$. In other words, we are interested in the best accuracy given that we only make decisions for a pre-specified proportion of observations.

1.2 Related Literature

The concept of binary classification with indecisions has been well studied by different communities. It is known by several names, such as "Classification with a Reject Option", "Selective Classification", "No-decision classification", "Classification with abstention" and "Human-AI collaboration". The corresponding approaches involve classifiers that are allowed to not make a decision when the class probabilities used for making a decision are too close to each other. For clarity, throughout this paper we refer to all of these literature areas under the umbrella term of selective classification (El-Yaniv and Wiener, 2010).

Selective classification has traditionally encompassed two primary forms of observation rejection, referred to in this paper as indecision: ambiguity rejection and novelty rejection (Hendrickx et al., 2024). Ambiguity rejection arises when a model cannot confidently differentiate between two or more classes for a given observation (Chow, 1957; Hellman, 1970; Fukunaga and Kessell, 1972). In contrast, novelty rejection applies to observations that cannot be reliably assigned to any predefined class (Cordella et al., 1995; Seo et al., 2000; Vailaya and Jain, 2000). In this work, we introduce a distinct rejection paradigm that shares characteristics with both ambiguity and novelty rejection but is fundamentally driven by classification accuracy. Specifically, we propose accuracy rejection, a paradigm that rejects observations that cannot be classified without exceeding a predefined accuracy threshold, irrespective of whether they exhibit ambiguity or novelty. This type of rejection has been explored in prior work under a variety of assumptions and guarantees (Shekhar et al., 2019; Rava et al., 2025). While accuracy rejection bears similarities to ambiguity rejection, it explicitly prioritizes the maintenance of a specified classification accuracy level.

1.2.1 Classification with Reject

In the binary classification setting, the classification with reject paradigm seeks to incorporate indecisions into a classifier by optimally picking the cost of indecisions d and the threshold δ, then minimizing the modified cost function $\mathbf{P}\{H \cdot f(x) < -\delta\} + d \cdot \mathbf{P}\{|H \cdot f(x)| \leq \delta\}$, where $H := 2Y - 3$. Although closely related to our proposed selective classification framework, classification with reject differs in several important aspects. In particular, existing reject based approaches do not provide

clear guidance about how to obtain rigorous accuracy guarantees while simultaneously minimizing the number of identified indecisions.

There have also been recent investigations of how to best incorporate selective classification into modern machine learning algorithms, through the lens of convex optimization (Yuan and Wegkamp, 2010). The works of Grandvalet et al. (2008) and Wegkamp and Yuan (2011) studied incorporating selective classification along side support vector machines while (Cortes et al., 2016) investigated learning the simultaneously learning a classifier for a given selection rule. Recently, there has been work investigating the use of selective classification in order to manage limited resources (Valade et al., 2024). The impact of plug-in classifiers on oracle selection rules has been also studied by Denis and Hebiri (2020) and Lei (2014), while assuming continuity around the decision threshold. This work was further explored by Shekhar et al. (2019), who was able to control the abstention constraint with high probability while, also dealing with discontinuities in the empirical cdf.

1.2.2 Multiple Testing, Outlier Detection, and Conformal Inference

Another stream of work looks at identifying a calibrated selected set of observations that focuses on controlling coverage over a smaller set of observations, up to a user specified level Lei (2014). There are overlaps with conformal inference where the goal is to create prediction sets that contain the true classification label, up to a user specified level of coverage Vovk et al. (1999, 2005). Other works have aimed to bridge the gap between overall coverage and calibrated decision making. For binary classification, Lei (2014) constructed confidence sets that can be calibrated for each class at a user specified level. Another stream of literature offers calibrated decision making through control over the False Selection Rate (FSR), which is defined as the expected number erroneous decisions over the number of selected observations (Gang et al., 2022; Huo et al., 2024; Marandon, 2024; Jin and Candes, 2023; Rava et al., 2025; Zhao and Su, 2023). In a similar spirit, Sun and Wei (2015) developed a decision theoretic framework that utilized indecisions to control the FSR and Wang et al. (2024) recently used indecisions in the sequential setting.

On the modern application side, selective classification has also been used to address societal issues, such as fairness in decision making. The works of Schreuder and Chzhen (2021) and Rava et al. (2025) have independently investigated how to transform off the shelf classifiers into fair selective classifiers through empirical risk minimization and calibrated selection rules, respectively.

1.3 Our contributions

We start with a full characterization of the minimax risk (1) in the case of binary classification (Section 2), which we later generalize to multi-class classification (Section 5.2). Our theory is general and covers both continuous and discrete distributions. Along the way, we show that the map $\gamma \to \mathcal{R}(\gamma)$ is continuous and non-increasing. In other words, for any given (reachable) level of accuracy, we can find the optimal matching indecision level of and the corresponding classifier. These findings are extended to the problem of hypothesis testing, where given a type I error we wish to control the type II error. To the best of our knowledge, this setup was not previously explored in the context of selective inference.

Sections 2.3 and 3.3 are dedicated to our fully adaptive methodology, where both the training and calibration sets are finite. We offer a novel finite sample analysis of both, classification and hypothesis testing settings, with corresponding simulation and real data analyses in sections 6.2

and 7. We explain how to calibrate the indecision region given a plug-in rule $\hat{\eta}$, in order to either achieve a level of accuracy or match a level of indecisions for both problems of classification and testing.

In Section 4.1, we focus on the binary Gaussian mixture model given a fixed separation between the centers. We fully characterize the "sharp" phase transition of classification in terms of indecisions. It is well established that, in order to achieve a level of accuracy of order $1-\delta$, the separation between centers Δ has to be of the order $\sqrt{2\log(1/\delta)}$ where the constant 2 is sharp. When the separation is of order $c\sqrt{2\log(1/\delta)}$ for some $c<1$, we need indecisions to reach the level of accuracy $1-\delta$. We give a sharp characterization of indecisions in this case. Interestingly, as long as $1/2<c<1$, we show that the optimal amount of indecisions is of order $o(1)$, meaning that by allowing only a negligible proportion of indecisions we can reach the level of misclassification δ even in the case where the class distributions are not well-separated. These findings are illustrated by numerical experiments in Section 6.1. More generally, the optimal procedure is based on thresholding the likelihood ratio between distributions f_1 and f_2, which can be encoded through the regression function η. In practice, we can use a training sample to learn η. In Section 4.2, we quantify the loss induced by the estimation of η. First, under reasonable assumptions similar to the usual margin condition, we show that for a fixed level of indecisions, the accuracy of the plug-in procedure is comparable to that of the oracle and, in general, we can expect consistency of the plug-in approach. Second, we also show that if calibration is done with respect to the accuracy, i.e., if we tune the plug-in classifier to reach a given accuracy level, then the amount of indecisions is also controlled as the sample size grows, although not necessarily consistently.

Finally, we suggest two extensions of our theory in section 5. First, we emphasize, the special case where the likelihood ratio f_1/f_0 satisfies the "Monotone Likelihood Ratio" property. In this setting, we do not need a training sample, as we can simply threshold the observations themselves instead of the scores $\eta(\cdot)$. This is typically the case for location models under log-concave distributions. We also show how to calibrate our procedure in this setting. Second, we extend our classification theory to the multi-class setting.

1.4 Notation

Throughout the paper we use the following notation. For given quantities a_n and b_n, we write $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) when $a_n \leq cb_n$ ($a_n \geq cb_n$) for some absolute constant $c>0$. In the case $a_n/b_n \to 0$, we use the notation $a_n = o(b_n)$. We also write $a_n \approx b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For any $a, b \in \mathbf{R}$, we denote by $a \vee b$ ($a \wedge b$) the maximum (the minimum) of a and b. Finally c_0, c_1, c are used for positive constants whose values may vary from theorem to theorem. For error rate control, we interchangeably use α_1 and α_2 to denote a user specified level of control for both the case of binary classification (class 1 and class 2) and hypothesis testing (type I and type II errors). A fixed proportion of indecisions is denoted as γ and the smallest proportion of indecisions that controls the specified error rate is denoted as γ^*.

2 General Binary Classification

In this setting, we want to find the *smallest* possible indecision region that is able to control the accuracy (or, similarly, misclassification risk) of our classifier. By using the minimum necessary

amount of indecisions, we are able to automate as many decisions as possible, and delegate only the smallest necessary amount of indecisions over for potentially costly human review.

For a given classifier $\tilde{Y}$, our objective is to control the conditional misclassification risk, given that a decision has been made. We define the conditional minimax risk as:

$$\mathcal{R}(\gamma) := \inf_{\tilde{Y}_\gamma} \mathbf{P}_Y \left(\tilde{Y}_\gamma \neq Y | \tilde{Y}_\gamma \neq 0 \right),$$

where the infimum is over all classifiers taking values in $\{0, 1, 2\}$ such that $\mathbf{P}(\tilde{Y}_\gamma = 0) = \gamma$. Our goal at the end of this section is to ensure that this conditional risk does not exceed a pre-specified threshold α, while minimizing the number of indecisions. To accomplish this, we first study the comparatively easier setting where for a fixed proportion of indecisions γ, we aim to find the best achievable level of accuracy. Understanding the case with a fixed amount of indecisions provides a framework for developing a procedure that guarantees classifier accuracy.

2.1 Fixed proportion of indecisions

Here, we focus on the binary case where we only have two classes. For a given level of indecisions $\gamma \in [0, 1)$, we define the optimal indecision region Θ_γ, satisfying $\mathbf{P}_Y(\Theta_\gamma) = \gamma$. We will show that there exists a value $\tau_\gamma \in [1/2, 1]$ such that

$$\Theta_\gamma := \left\{ 1 - \tau_\gamma < \frac{p_1 f_1(X)}{p_1 f_1(X) + p_2 f_2(X)} < \tau_\gamma \right\} \cup \mathcal{M}_\gamma, \tag{2}$$

where $\mathcal{M}_\gamma$ is any subset of $\left\{ \frac{p_1 f_1(X) \vee p_2 f_2(X)}{p_1 f_1(X) + p_2 f_2(X)} = \tau_\gamma \right\}$ such that $\mathbf{P}_Y(\Theta_\gamma) = \gamma$.

Define $\eta(\cdot)$ as the conditional density function $\eta(x) = \mathbf{P}(X = x | Y = 1)$, which is defined as

$$\eta(X) = \frac{p_1 f_1(X)}{p_1 f_1(X) + p_2 f_2(X)}. \tag{3}$$

It is natural to observe that the optimal indecision region concentrates around where $\eta(X)$ is close to $1/2$. We note that our threshold τ_γ plays a similar role to the constant d in Herbei and Wegkamp (2006). We also note that when $\eta(X) \vee (1 - \eta(X)) = \tau_\gamma$, i.e., we are at the frontier of making an indecision, then we might randomly choose to reject or not. The Bayes oracle classifier with a γ proportion of indecisions is given by

$$Y^*_\gamma = \arg\max_{i \in \{1,2\}} (p_i f_i(X)) \mathbf{1}(\Theta^c_\gamma), \tag{4}$$

as shown in the following result.

Theorem 1. *Given γ, the classifier Y^*_γ is minimax optimal for the risk $\mathcal{R}(\gamma)$. Moreover, we have that*

$$\mathcal{R}(\gamma) = \mathbf{P}_Y(Y^*_\gamma \neq Y | Y^*_\gamma \neq 0) = \frac{\int_{\Theta^c_\gamma} (p_1 f_1 \wedge p_2 f_2) d\mu}{1 - \gamma},$$

where Θ^c_γ denotes the compliment of the set Θ_γ.

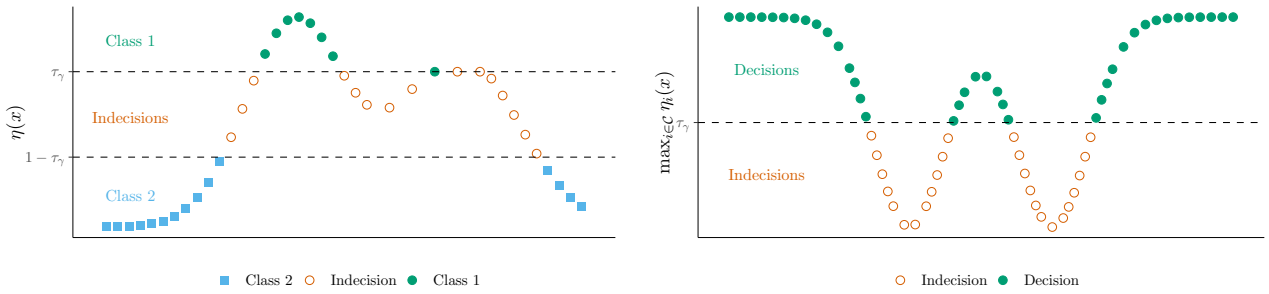

Figure 3: An example of a binary classification problem that includes indecisions (orange / open circle). For the left most figure, the indecisions lie in a region between the two classes: class 1 (green / solid circle) and class 2 (blue / square). A plateau at the threshold τ_γ indicates that some observations may be randomly classified as either class 1 or an indecision. In the right most figure, the indecisions lie below the threshold τ_γ in comparison to the largest conditional density across potentially many classes. This demonstrates that the indecision region may not be a simple interval.

It follows from our proof that τ_γ is an increasing function of γ, and as $\gamma \to 0$, we recover the classical classification result without indecisions.

On the one hand, when random variable $\eta(X)$ has atoms and, in particular, $\mathbf{P}(\eta(X) \vee (1 - \eta(X)) = \tau_\gamma) \neq 0$, the set $\mathcal{M}$ is non empty and we shall call the region $\mathcal{M}$ a "plateau" where the indecisions are picked randomly as shown in the left panel of Figure 3. On the other hand, if $\eta(X)$ has no atoms, then the indecision region is unique up to Lebesgue negligible sets.

We would like to emphasize that the indecision region is not necessarily an interval, as illustrated in right most plot in Figure 3. Consequently, constructing the indecision region requires prior knowledge of the conditional density η.

2.2 Fixed Error Rate

We will now show that understanding the case with a fixed amount of indecisions will allow us to control the accuracy or, similarly, the misclassification rate, of the optimal classifier at any user-specified level. We start with the following result on the properties of the risk function.

Proposition 1. *For any $0 \leq \gamma < 1$, we have*

$$\mathcal{R}(\gamma) = \mathbf{E}_Y \left(Z | Z < 1 - \tau_\gamma \text{ or } Z \in \mathcal{M}_\gamma \right),$$

where $Z := \frac{p_1 f_1 \wedge p_2 f_2}{p_1 f_1 + p_2 f_2}(X) = (\eta \wedge (1-\eta))(X)$. Moreover, $\gamma \mapsto \mathcal{R}(\gamma)$ is continuous and non-increasing.

Because function $\mathcal{R}(\gamma)$ is non-increasing and lower-bounded by 0, it has a limit as $\gamma \to 1$ that we shall denote $\mathcal{R}^* := \lim_{\gamma \to 1^-} \mathcal{R}(\gamma)$. We note that $\mathcal{R}(\gamma)$ interpolates between the misclassification rate we would get without indecisions and $\mathcal{R}^*$. Thanks to the continuity of $\mathcal{R}$, our result also shows that for any given misclassification level α above $\mathcal{R}^*$, we can find a γ^* such that $\mathcal{R}(\tilde{Y}_{\gamma^*}) = \alpha$ and this γ^* is the smallest possible. In other words, for any reachable level of accuracy, we are able to characterize the corresponding minimum number of indecisions.

Algorithm 1 Binary Classification with Indecisions

Input: Observed $\{(X_i, Y_i) : i \in \mathcal{D}\}$, accuracy level $1 - \alpha$.

Output: a selective classification rule $\{\hat{Y} \in \{0, 1, 2\}\}$ and the corresponding $\hat{\tau}$.

1: Randomly split $\mathcal{D}$ into $\mathcal{D}^{train}$ and $\mathcal{D}^{cal}$.
2: Train a machine learning model on $\{(X_i, Y_i) : i \in \mathcal{D}^{train}\}$.
3: Predict the conditional density $\hat{\eta}_i$ overall for $i \in \mathcal{D}^{cal}$.
4: Order $\hat{\tau}_i := 1 - (\hat{\eta}_i \wedge (1 - \hat{\eta}_i))$ from smallest to largest, $\hat{\tau}_{(1)} \leq \cdots \leq \hat{\tau}_{(m)}$.
5: Compute the empirical conditional misclassification error $\hat{R}_i$ using $\mathcal{D}^{cal}$, for all the candidate thresholds $\hat{\tau}_{(i)}$ starting with $\hat{\tau}_{(1)}$.
6: Ensure that the estimated conditional error is monotonic by keeping track of the minimum of all estimated errors $\tilde{R}_i = \min_{j \leq i} \hat{R}_j$.
7: Stop once $\tilde{R}_i$ is below α and return the corresponding $\hat{\tau}_{(i)}$.

Lemma 1. *Suppose that* $\mathbf{P}((\eta \wedge (1 - \eta))(X) \leq \varepsilon) > 0$ *for every* $\varepsilon > 0$. *Then,* $\lim_{\gamma \to 1^-} \tau_\gamma = 1$ *and* $\mathcal{R}^* = 0$.

Proof. Given any $\varepsilon > 0$, there exists a γ such that

$$\mathbf{P}((\eta \wedge (1 - \eta))(X) \leq \varepsilon) > 1 - \gamma \geq \mathbf{P}((\eta \wedge (1 - \eta))(X) \leq 1 - \tau_\gamma).$$

Consequently, $\tau_\gamma \geq 1 - \varepsilon$. Hence, $\mathcal{R}^* \leq \lim_{\gamma \to 1^-} 1 - \tau_\gamma = 0$ by Proposition 1. □

We conclude that any level of accuracy can be reached under the assumption of Lemma 1. This assumption is natural and can be interpreted as follows. In order to get the misclassification error as small as possible, we need the existence of regions where the likelihood of f_1 dominates that of f_2, and regions where the likelihood of f_2 dominates that of f_1, and hence we are more confident whenever we predict Y to be 1 or 2 in these regions.

2.3 Finite Sample Calibration

Our goal in this section is to present a calibration procedure in the practical setting where the data-generating process is unknown and the test set is finite. To do this, we will follow the theoretical framework presented in Sections 2.1 and 2.2. By doing so, we will demonstrate that effective selective classification rules can be calibrated according to our proposed theory.

We start with the misclassification risk. Given a misclassification error level $\alpha \in [\mathcal{R}^*, \mathcal{R}(0)]$, our goal is to construct a classifier that achieves misclassification level α using the minimal number of indecisions. From the results in the previous section, we know that there exists an indecision level γ^* such that $\mathcal{R}(\gamma^*) = \alpha$. We aim to to construct a classifier $\hat{Y}$ such that the accuracy of $\hat{Y}$ is at least $1 - \alpha$ and the proportion of indecisions is γ^*.

Let us define $\gamma_\alpha := \gamma(\alpha)$ to re-emphasize the fact that the optimal amount of indecisions depends on our desired level of accuracy. We recall that the optimal indecision region is such that $\mathbf{P}\left(\eta(X) \wedge (1 - \eta(X)) > 1 - \tau_\alpha\right) + \mathbf{P}(\mathcal{M}_{\gamma_\alpha}) = \gamma_\alpha$. Observe that τ_α corresponds to a quantile of the random variable $\eta(X) \wedge (1 - \eta(X))$ and can be easily computed. The (conditional) misclassification error of $Y^*_{\gamma_\alpha}$ is $\mathbf{P}_Y(Y^*_{\gamma_\alpha} \neq Y | Y^*_{\gamma_\alpha} \neq 0) = \alpha$. Since we do not have access to γ_α explicitly, we need to

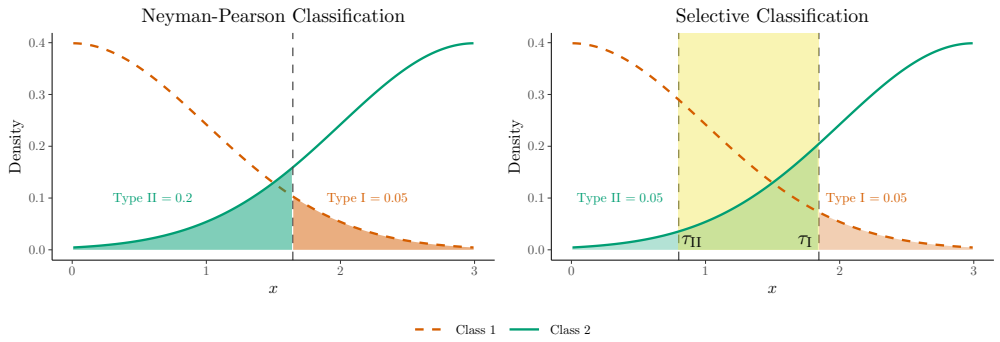

Figure 4: A comparison of Neyman-Pearson (NP) Classification (left) and Selective Classification (right), which can use indecisions. The NP-classifier is able to control the type I error at the correct level, at the compromise of the type II error. In contrast, selective classification is able to control both the type I and type II errors at the correct level, through the introduction of indecisions (yellow shaded region).

invert the function $\mathcal{R}(\cdot)$. In order to mimic the optimal classifier $Y^*_{\gamma_\alpha}$, and given an estimator $\hat{\eta}$, we wish to calibrate classifier $\hat{Y}$ of the form $\hat{Y} = \arg\max_{i\in\{1,2\}} \hat{\eta}_i \cdot \mathbf{1}\left(\hat{\eta}_i \wedge (1-\hat{\eta}_i) \leq 1-\hat{\tau}\right)$, where $\hat{\eta}_1 = \hat{\eta}$ and $\hat{\eta}_2 = 1-\hat{\eta}$ or, equivalently, $\hat{Y} = \mathbf{1}(\hat{\eta} \geq \hat{\tau}) + 2 \times \mathbf{1}(\hat{\eta} \leq 1-\hat{\tau})$. We estimate the value of $\hat{\tau}$ by using a calibration dataset, as we describe in Algorithm 1.

3 Controlling Type I and Type II Errors: Connection with the Neyman Pearson Paradigm

We now consider the hypothesis testing problem in which, given a fixed type I error probability, the goal is to achieve a desired level of type II error probability, using indecisions if necessary.

In the absence of indecisions, this framework can be viewed as an alternative to the classical Neyman-Pearson classification paradigm, which prioritizes controlling the Type I error, often at the expense of the Type II error (Cannon et al., 2002; Scott and Nowak, 2005; Rigollet and Tong, 2011; Tong, 2013). If the practitioner is willing to incorporate indecision into the decision making process, it becomes possible to forgo direct optimization of the Type II error rate and instead simultaneously control both the Type I and Type II error rates. We demonstrate the difference between the classic set-up, and our approach with indecisions in Figure 4.

In the context of indecisions, for a given value of γ, we define the (conditional) type I and type II error probabilities, respectively, as

$$\mathcal{P}_{\mathrm{I}}(\tilde{Y}_\gamma) = \mathbf{P}\left(\tilde{Y}_\gamma \neq Y \middle| Y = 1, \tilde{Y}_\gamma \neq 0\right) \quad \text{and} \quad \mathcal{P}_{\mathrm{II}}(\tilde{Y}_\gamma) = \mathbf{P}\left(\tilde{Y}_\gamma \neq Y \middle| Y = 2, \tilde{Y}_\gamma \neq 0\right),$$

where $\tilde{Y}_\gamma$ represents a classifier with a γ proportion of indecisions. We note that when the level of indecisions is $\gamma = 0$, we recover the classic definition of type I and type II error probabilities, because $\tilde{Y}$ will not be able to take the value 0. Given a type I error probability α_1 and an indecision level γ, the corresponding minimax risk is given by

$$\mathcal{P}(\alpha_1, \gamma) = \inf_{\tilde{Y}_\gamma} \mathcal{P}_{\mathrm{II}}(\tilde{Y}_\gamma),$$

where the infimum is taken over all classifiers taking values in $\{0,1,2\}$ such that $\mathbf{P}(\tilde{Y}_\gamma = 0) = \gamma$ and $\mathcal{P}_{\mathrm{I}}(\tilde{Y}_\gamma) = \alpha_1$. Our goal is to control both the type I and type II error probabilities at user-specified levels α_1 and α_2, respectively, using the smallest amount of indecisions

In some situations, such error control can be achieved without indecisions. However, when the problem is sufficiently difficult, it becomes necessary to identify the hardest to classify observations as indecisions in order to meet our objective.

3.1 Fixed Level of Indecisions

We first consider the case where the level of indecisions is fixed at γ. We start by considering all classifiers with indecision level γ and conditional type I error probability α_1 or, similarly, unconditional type I error probability $\alpha_1 \cdot (1-\gamma)$.

Among these classifiers, we then aim to minimize the type II error probability.

The corresponding minimax risk is given by:

$$\begin{aligned}\mathcal{P}(\alpha_1,\gamma) &= \inf_{\tilde{Y}_\gamma} \mathcal{P}_{\mathrm{II}}(\tilde{Y}_\gamma) \\ &= \inf_{\tilde{Y}_\gamma} \ \mathbf{P}_2(\tilde{Y}_\gamma = 1 \mid \tilde{Y}_\gamma \neq 0) \\ &= \inf_{\tilde{Y}_\gamma} \ \frac{\mathbf{P}_2(\tilde{Y}_\gamma = 1)}{1-\gamma}, \end{aligned} \tag{5}$$

where the infimum is taken over all classifiers $\tilde{Y}_\gamma$ taking values in $\{0,1,2\}$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$, and the unconditional type I error probability of $\tilde{Y}_\gamma$ is $\alpha_1 \cdot (1-\gamma)$.

We denote by Y^*_γ the optimal classifier achieving the infimum in (5). By construction, Y^*_γ has conditional type I error probability α_1, indecision level γ, and the smallest corresponding achievable type II error probability.

In the result below, we show that the optimal classifier is of the form

$$Y^*_\gamma \coloneqq Y^*_\gamma(\alpha_1) = \mathbf{1}\left(\{\eta > \tau_2\} \cup \mathcal{M}^2_{\alpha_1,\gamma}\right) + 2 \times \mathbf{1}\left(\{\eta \leq \tau_1\} \setminus \mathcal{M}^1_{\alpha_1,\gamma}\right),$$

where the set $\mathcal{M}^1_{\alpha_1,\gamma}$ is any subset of $\{\eta = \tau_1\}$ such that $\mathbf{P}_1\left(\eta < \tau_1\right) + \mathbf{P}(\mathcal{M}^1_{\alpha_1,\gamma}) = \alpha_1(1-\gamma)$ and the set $\mathcal{M}^2_{\alpha_1,\gamma}$ is any subset of $\{\eta = \tau_2\}$ such that $\mathbf{P}\left(\{\tau_1 < \eta \leq \tau_2\} \setminus \mathcal{M}^2_{\alpha_1,\gamma}\right) + \mathbf{P}(\mathcal{M}^1_{\alpha_1,\gamma}) = \gamma$.

Again, we drop the dependence of Y^*_γ on α_1 for simplicity of the presentation.

Theorem 2. *Classifier Y^*_γ is minimax optimal for the risk of type II $\mathcal{P}(\alpha_1,\gamma)$. Moreover, the optimal risk with a γ proportion of indecisions and a conditional type I error rate control of α_1 is given by*

$$\mathcal{P}(\alpha_1,\gamma) = \frac{\mathbf{P}_2(Y^*_\gamma = 1)}{1-\gamma} = \frac{\int_{\{Y^*_\gamma = 1\}} f_2 d\mu}{1-\gamma}.$$

3.2 Fixed Error Rate

We will now focus on the scenario where a practitioner wishes to simultaneously control the type I and type II error probabilities, at any (potentially different) user specified levels. We start with a result that is in the same spirit as Proposition 1, illustrating the connection between the case considered in Section 3.1 and the desired control of the type I and type II error probabilities.

Algorithm 2 Neyman Pearson Classification with Indecisions

Input: $\{(X_i, Y_i) : i \in \mathcal{D}^{\text{train}}\}$, $\{(X_i, Y_i) : i \in \mathcal{D}^{\text{test}}\}$, type I, type II levels $\{\alpha_c : c = 1, 2\}$, and a grid of candidate indecision levels $\Gamma = \left\{ \frac{k}{|\mathcal{D}^{\text{cal}}|}, \quad \text{for } k = 0, 1, \ldots, |\mathcal{D}^{\text{cal}}| \right\}$.

Output: a selective classification rule $\left\{ \hat{Y} \in \{0, 1, 2\} \right\}$ and the corresponding $\hat{\tau}_1$, $\hat{\tau}_2$.

1: Train a machine learning model on $\{(X_i, Y_i) : i \in \mathcal{D}^{train}\}$.

2: Predict the conditional density $\hat{\eta}_i$ for $i \in \mathcal{D}^{cal}$ and order them from smallest to largest as $\hat{\eta}_{(1)} \leq \cdots \leq \hat{\eta}_{(n)}$.

3: For each candidate indecision level γ_k, set $\hat{\tau}_1(k)$ to control the type I error at level $(1 - \gamma_k) \cdot \alpha_1$, starting from $k = 0$. Notice that $\hat{\tau}_1(k) := \hat{\eta}_{\tilde{k}}$ for some $\tilde{k}$.

4: Assign the following $\gamma_k \cdot |\mathcal{D}^{\text{cal}}| = k$ observations as an indecision, and set the upper threshold $\hat{\tau}_2(k)$ as $\hat{\tau}_2 := \hat{\eta}_{(\tilde{k}+k)}$.

5: Compute the corresponding candidate estimator $\hat{Y}_{\gamma_k}$:

$$\hat{Y}_{\gamma_k} = \mathbf{1}(\hat{\eta} \geq \hat{\tau}_2(k)) + 2 \cdot \mathbf{1}(\hat{\eta} \leq \hat{\tau}_1(k)).$$

6: Estimate the type II error of your candidate prediction $\hat{Y}_{\gamma_k}$.

7: Pick the smallest level of indecisions such that the type II error is controlled

$$\gamma^* = \arg\min_{\gamma_k \in \Gamma} \{\gamma_k \mid \text{Type II error}(\gamma_k) \leq \alpha_2\}$$

8: Return $\hat{Y}$ as the final prediction rule and the corresponding thresholds $\hat{\tau}_1, \hat{\tau}_2$.

Proposition 2. *For any $\alpha_1 \in [0, 1]$, function $\gamma \mapsto \mathcal{P}(\alpha_1, \gamma)$ is continuous and non-increasing. Moreover, if for each $\varepsilon > 0$ we have $\mathbf{P}(f_2(X) \leq \varepsilon \cdot f_1(X)) > 0$, then $\lim_{\gamma \to 1^-} \mathcal{P}(\alpha_1, \gamma) = 0$.*

Because $\mathcal{P}(\alpha_1, \cdot)$ is non-increasing and lower-bounded by 0, $\lim_{\gamma \to 1^-} \mathcal{P}(\alpha_1, \gamma)$ exists and shall be denoted $\mathcal{P}^*(\alpha_1)$. Hence, any value of the type II error within the range $[\mathcal{P}^*(\alpha_1), \mathcal{P}(\alpha_1, 1)]$ can be reached using only the necessary amount of indecisions.

Condition $\mathbf{P}\left(p_2 f_2(X) \leq \varepsilon \cdot p_1 f_1(X)\right) > 0$ for small ε can be interpreted as follows. In order to get a Type II error as small as possible, we need the existence of regions where the likelihood of f_1 dominates that of f_2, and hence we are more confident whenever we predict Y to be 1 in these regions.

3.3 Finite Sample Calibration

We are now equipped to use the results in Sections 3.1 and 3.2 to calibrate the thresholds for type I and type II error control. In contrast to the binary classification setting in Section 2, developing valid calibration procedures is more challenging due to the error rates' dependence on the true label Y instead of the prediction $\hat{Y}$. However, our analysis suggests that as long as the type I error probability can be controlled, adding indecisions can only lower the type II error probability relative to the methods that do not use indecisions.

For given levels $\alpha_1 \in [0, 1]$ and $\alpha_2 \in [\mathcal{P}^*(\alpha_1, \mathcal{P}(\alpha_1, 1)]$ of type I and type II error probabilities, respectively, there exists a γ^* such that $\mathcal{P}(\alpha_1, \gamma^*) = \alpha_2$. Our goal now is to find a classifier $\hat{Y}$ such

that the type I error probability of $\hat{Y}$ is at most α_1, the type II error probability is lat most α_2, and the proportion of indecisions is of the same order as the minimum amount of indecisions γ^*. Our suggested estimator $\hat{Y}$ is of the form $\hat{Y} = \mathbf{1}(\{\hat{\eta} > \hat{\tau}_2\}) + 2 \times \mathbf{1}(\{\hat{\eta} \leq \hat{\tau}_1\})$.

Ideally, we would like the type I error to be controlled by $\mathbf{P}_1(\hat{\eta} \leq \hat{\tau}_1) \leq \alpha_1 \cdot (1 - \gamma^*)$, with the optimal indecision region to be given by $\mathbf{P}(\hat{\tau}_1 < \hat{\eta} < \hat{\tau}_2) = \gamma^*$, and the type II error of $\hat{Y}$ to satisfy $\mathbf{P}_2(\hat{\eta} \geq \hat{\tau}_2) \leq \alpha_2(1 - \gamma^*)$. Because we do not have access to γ^* explicitly, we need to invert the function $\mathcal{P}(\alpha_1, \cdot)$. We estimate the values $\hat{\tau}_1, \hat{\tau}_2$ using a calibration dataset, as we describe in Algorithm 2.

4 Theory

In this section we further investigate the theoretical properties of our selective classification framework. The first section demonstrates a novel perspective in selective inference and demonstrates the sharp phase transition of the risk of our classifier and gives light to a new all-or-nothing phenomenon. The second section looks at plug-in rules where the true density function η is replaced with a learned function $\hat{\eta}$.

4.1 Explicit indecisions for the Gaussian Mixture Model: A sharp phase transition

Selective classification is the most attractive when through the use of virtually no indecisions, practitioners can expect dramatic gains in accuracy, as demonstrated in the introduction through Figure 2. To the best of our knowledge, there has been no work around understanding when practitioners are working in this critical regime. Providing deeper insight into this regime is essential for the practical adoption of selective classifiers. When a task is too difficult, selective classifiers may identify too many indecisions, overburdening human reviewers. Conversely, when the task is too easy, indecisions are unnecessary since standard classifiers can already meet the desired accuracy. The greatest benefit of selective classification thus arises in the intermediate regime, where standard models fall short of the accuracy target, yet only a modest rate of indecisions is required to achieve it. In this section we will analyze the sharp phase transition in the risk of our classifier when the Bayes classifier is unable to match our desired level of accuracy. We demonstrate that in many scenarios we can obtain a dramatic increase in accuracy (or, similarly, a dramatic decrease in risk) with virtually no indecisions.

This section is devoted to the asymptotic behavior of the optimal amount of indecisions as the risk gets smaller. We focus on the symmetric Gaussian mixture model. In particular, we assume that $p_1 = p_2 = 1/2$ and that $f_1(x) = \frac{1}{\sqrt{2\pi}} \exp(-(x - \Delta)^2/2) = f_2(-x)$ for some separation $\Delta > 0$. In this case, the Monotone Likelihood Ratio (MLR) property for symmetric likelihoods holds. We describe further the benefits of the MLR property in the appendix section 5.1. For a given level of misclassification rate $\delta \to 0$, we are interested in the asymptotic behavior of γ_δ as a function of separation Δ. Naturally, we would expect γ_δ to be non-increasing in Δ. We assume that $\delta \to 0$ and let parameter Δ depend on δ, omitting subscript δ whenever no ambiguity arises.

The asymptotic property we study here is $\delta-$*consistency*, which is inspired by consistency in classification Minsker et al. (2025) or, similarly, exact recovery in Gaussian mixtures as defined

in Ndaoud (2022). We establish a complete characterization of the sharp phase transition for $\delta-$consistency.

Definition 1. *Let $(\gamma_\delta)_{0\leq\delta\leq 1}$ be a class of indecision masses.*

- *We say that* $\delta-$consistency is impossible *for $(\gamma_\delta)_{0\leq\delta\leq 1}$ if*

$$\liminf_{\delta\to 0} \mathcal{R}(\gamma_\delta)/\delta > 1.$$

- *We say that* $\delta-$consistency is possible *for $(\gamma_\delta)_{0\leq\delta\leq 1}$ if there exists a classifier $\hat{Y}_{\gamma_\delta} := \hat{Y}(\gamma_\delta, \cdot)$, such that $\mathbf{P}(\hat{Y}_{\gamma_\delta} = 0) = \gamma_\delta$ for all δ, and*

$$\limsup_{\delta\to 0} \mathbf{P}_Y(\hat{Y}_{\gamma_\delta} \neq Y | \hat{Y}_{\gamma_\delta} \neq 0)/\delta \leq 1.$$

In this case, we say that $\hat{Y}$ achieves $\delta-$consistency.

In order to derive the phase transition of interest, let us first recall the equations that relate γ_δ to Δ_δ and δ. For a misclassification level δ, we have, under the MLR property (Section 5.1), that $\mathbf{P}(\xi \geq \Delta_\delta + t_\delta) = (1-\gamma_\delta)\delta$, where ξ is a standard normal random variable, and t_δ is a threshold that can be related to τ_δ. Moreover, the indecision level is given by $\mathbf{P}(\xi \geq \Delta_\delta - t_\delta) - \mathbf{P}(\xi \geq \Delta_\delta + t_\delta) = \gamma_\delta$. Since there is a one to one correspondence between δ and γ_δ, it is easy to see that the same holds for t_δ as well. Our proof strategy works as follows. For a given $t \geq 0$:

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq \Delta + t) + \mathbf{P}(\xi \geq t - \Delta)} \leq \delta \quad \text{if and only if} \quad \mathbf{P}(\xi \geq \Delta - t) - \mathbf{P}(\xi \geq \Delta + t) \geq \gamma_\delta.$$

We use the following parameterizations for Δ_δ and γ_δ: $\Delta_\delta = c\sqrt{2\log(1/\delta)}$ for $0 < c < 1$, and

$$\gamma_\delta = \begin{cases} 1 - \delta^m & \text{if } 0 < c < 1/2, \\ \delta^m & \text{if } 1/2 < c < 1. \end{cases}$$

We also define $m^*(c)$ such that

$$m^*(c) = \begin{cases} (c - 1/(4c))^2 & \text{if } 0 < c < 1/2, \\ (2c-1)^2 & \text{if } 1/2 < c < 1. \end{cases} \tag{6}$$

The next result describes a "phase transition" for γ_δ for the problem of $\delta-$consistency.

Theorem 3. *For any $\varepsilon > 0$ and $c > 1/2$.*

(i) Let $m \leq m^*(c)$. Then, the classifier $Y^*_{\gamma_\delta}$ defined in (4), achieves $\delta-$consistency.

(ii) Moreover, if $m \geq (1+\varepsilon)m^*(c)$ then $\delta-$consistency is impossible.

For any $\varepsilon > 0$ and $c < 1/2$.

(i) Let $m \geq m^*(c)$. Then, the classifier $Y^*_{\gamma_\delta}$ defined in (4), achieves $\delta-$consistency.

(ii) Moreover, if $m \leq (1-\varepsilon)m^*(c)$ then $\delta-$consistency is impossible.

Theorem 3 shows that δ-consistency holds if and only if

$$m(\gamma) \leq m^*(c), \quad \text{for } 1/2 < c < 1, \tag{7}$$

$$m(\gamma) \geq m^*(c), \quad \text{for } 0 < c < 1/2. \tag{8}$$

It is worth noting here that while in the classical setup (without indecisions) we need $c \geq 1$ to achieve $\delta-$consistency, we require almost no indecisions provided that $c > 1/2$, as $\delta^{(2c-1)^2} = o(1)$. We also note the following interesting all-or-nothing phenomenon. By observing the asymptotic behavior of γ_δ, it seems that γ_δ either goes to 0 or 1, depending on whether c is greater or smaller than $1/2$. Asymptotically, the optimal behavior corresponds to either full indecisions or almost no indecisions.

4.2 Plug-in rules

In this section, we provide theoretical guarantees for Algorithm 1. For simplicity, we assume an infinite calibration set, allowing us to focus on the error introduced by estimating η from a finite training sample. Analogous results for hypothesis testing (Algorithm 2) can be derived in the same manner and are therefore omitted to avoid redundancy.

In what follows, we replace classification probability function η with a learned function $\widehat{\eta}$. Given an indecision level, we quantify the loss in the accuracy due to the estimation of η. In addition, we also investigate what happens to the indecision level if we calibrate the $\widehat{\eta}$-based method to achieve a pre-specified level of accuracy.

Similar results have been established in the literature. In particular, Denis and Hebiri (2020) derive results on the asymptotic performance of the plug-in classifier in the general setting of our Theorem 4. However, they assume that $\eta(X)$ has a continuous distribution near the decision threshold, while we allow for a point mass on the decision boundary. Furthermore, Lei (2014) establishes results that are similar to our Theorem 5. However, while we fix the conditional misclassification error given that a decision has been made, Lei (2014) focuses on the unconditional accuracy. Moreover, like Denis and Hebiri (2020), he also assumes that $\eta(X)$ has a continuous distribution near the decision threshold.

The accuracy and the level of indecisions are linked through the choice of the threshold τ. Our analysis is more challenging than the one in Herbei and Wegkamp (2006), because it relies on ensuring that $\hat{\tau}$ and τ relatively close, which requires stronger assumptions than the usual margin condition. For the rest of this section, and for the sake of simplicity, we will assume that all levels of misclassification $\alpha > 0$ can be reached using our framework. In what follows we shall only provide theoretical results for the risk of the plug-in classifier in the classification setting. Similar results can be derived for testing as well.

4.2.1 Fixing the probability of an indecision

Let $\widehat{Y}_\gamma$ be the plug-in classifier for the indecision level γ, i.e.,

$$\widehat{Y}_\gamma(X) = 1 \times \mathbf{1}\{\widehat{\eta}(X) > \widehat{\tau}_\gamma\} + 2 \times \mathbf{1}\{\widehat{\eta}(X) < 1 - \widehat{\tau}_\gamma\},$$

where $\widehat{\tau}_\gamma$ is chosen so that $\mathbf{P}\Big(\widehat{\tau}_\gamma \geq \widehat{\eta}(X) \geq 1 - \widehat{\tau}_\gamma \,|\, \widehat{\eta}\Big) = \gamma$. Note that we may need to use only a subset of $\{\widehat{\eta}(X) = \widehat{\tau}_\gamma\} \cup \{\widehat{\eta}(X) = 1 - \widehat{\tau}_\gamma\}$ rather than the full set to get the exact equality above, same as we did for Y^*_γ.

We will write τ_γ^* for the corresponding threshold for Y_γ^*. Given a classifier $\tilde{Y}$, we let $R(\tilde{Y}) = \mathbf{P}(\tilde{Y} \neq Y | \tilde{Y} \neq 0)$. To simplify expressions, we write η and $\widehat{\eta}$ instead of $\eta(X)$ and $\widehat{\eta}(X)$, respectively.

Lemma 2. *For all $\gamma \in [0, 1)$,*

$$R(\widehat{Y}_\gamma) - R(Y_\gamma^*) = \tfrac{1}{1-\gamma}\mathbf{E}|\tau_\gamma^* - \eta|\big(\mathbf{1}\{Y_\gamma^* = 1, \widehat{Y}_\gamma \neq Y_\gamma^*\} + \mathbf{1}\{\widehat{Y}_\gamma = 1, \widehat{Y}_\gamma \neq Y_\gamma^*\}\big) + \tfrac{1}{1-\gamma}\mathbf{E}|1 - \tau_\gamma^* - \eta|\big(\mathbf{1}\{Y_\gamma^* = 2, \widehat{Y}_\gamma \neq Y_\gamma^*\} + \mathbf{1}\{\widehat{Y}_\gamma = 2, \widehat{Y}_\gamma \neq Y_\gamma^*\}\big).$$

Remark 1. *We can bound the above expression as follows:*

$$R(\widehat{Y}_\gamma) - R(Y_\gamma^*) \leq 2\mathbf{E}\big((|\tau_\gamma^* - \eta| \vee |1 - \tau_\gamma^* - \eta|)\mathbf{1}\{\widehat{Y}_\gamma \neq Y_\gamma^*\} \big| \widehat{Y}_\gamma \neq 0 \text{ or } Y_\gamma^* \neq 0\big).$$

Remark 1 implies that if η does not have too much mass around the optimal thresholds τ_γ^* and $1 - \tau_\gamma^*$, and $\hat{\eta}$ is close to η, then we can expect consistency of the plug-in approach.

For the next result, we let $\eta_{\max} := \eta \vee (1 - \eta)$ and focus on the standard setting where we can bound the probability that $\eta_{\max}$ lies within ϕ of τ_γ^* by some nonnegative power of ϕ. More specifically, we assume

$$\mathbf{P}\big(|\eta_{\max} - \tau_\gamma^*| \leq \phi\big) \lesssim \phi^\beta \quad \text{and} \quad \mathbf{P}\big(|\eta_{\max} - \tau_\gamma^*| \leq \phi,\, \eta_{\max} \neq \tau_\gamma^*\big) \gtrsim \phi^{\beta'}, \tag{9}$$

for some $\beta' \geq \beta \geq 0$ and all sufficiently small positive ϕ.

Theorem 4. *Suppose that (9) holds for $0 < \phi \leq 3\phi_\gamma^*$ with $\beta' \geq \beta \geq 0$ and $\phi_\gamma^* > 0$. Then,*

$$R(\widehat{Y}_\gamma) - R(Y_\gamma^*) \lesssim \inf_{0 < \phi \leq \phi_\gamma^*} \tfrac{1}{1-\gamma}\Big\{\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) + \phi^{1+\beta}\Big\} + (\phi^{1-\beta'}\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) \wedge \phi). \tag{10}$$

In particular, if $\beta' \leq 1$, then

$$R(\widehat{Y}_\gamma) - R(Y_\gamma^*) \lesssim \tfrac{1}{1-\gamma} \inf_{0 < \phi \leq \phi_\gamma^*} \Big\{\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) + \phi^{1+\beta}\Big\}.$$

Remark 2. *In the statement of Theorem 4, we can replace $\frac{1}{1-\gamma}\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big)$ with $\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \eta_{max} > \tau_\gamma^*\big) + \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}_{max} > \widehat{\tau}_\gamma\big)$, where $\widehat{\eta}_{max} := \widehat{\eta} \vee (1 - \widehat{\eta})$. That is, we only need $\hat{\eta}$ to be close to η within the region of decisions. For example, this can be easily achieved if we have good control over the uniform bound $\|\widehat{\eta} - \eta\|_\infty$. We can also replace the term $\frac{\phi^{1+\beta}}{1-\gamma}$ by $\phi^{1+\beta}$ if we assume that the margin condition (9) holds conditionally on being in the region of decisions.*

Note that we will have a good estimator $\hat{\eta}$ of η as long as η is sufficiently smooth. When $\beta' \leq 1$, our result is similar to the corresponding one in Herbei and Wegkamp (2006), which covers the setting without indecisions. It is worth noting that, unlike in the setting without indecisions, we have an additional challenge in controlling the distance between thresholds $\hat{\tau}$ and τ. The lower bound in condition (9) helps get that control. As a consequence, when picking $\phi \approx 1/\sqrt{n}$, where n the training sample size, we can recover fast rates when $\beta = \beta' = 1$, which is typically the case for atom-less distributions. Going back to the bound (10) and taking $\phi \approx 1/\sqrt{n}$, we recover the slow rates without making any assumptions on the margin.

4.2.2 Fixing the misclassification level

We will use $R_{\widehat{\eta}}(\widehat{Y})$ to denote the conditional risk of the classifier $\widehat{Y}_\gamma$ given $\widehat{\eta}$. Let γ be a fixed indecision level and let $R^* := R(Y^*_\gamma)$. Here, we analyze the plug-in classifier corresponding to the misclassification level R^*. The classifier we consider is of the form $1 \times \mathbf{1}\{\widehat{\eta} > \widehat{\tau}\} + 2 \times \mathbf{1}\{\widehat{\eta} < 1 - \widehat{\tau}\}$. The indecision level of this classifier is $\widehat{\gamma} := \min\{\gamma : \ R_{\widehat{\eta}}(\widehat{Y}_\gamma) \leq R^*\}$, and its threshold is $\widehat{\tau} := \min\{\tau, \text{ s.t. } \mathbf{P}(\tau \geq \widehat{\eta} \geq 1 - \tau \,|\, \widehat{\eta}) \geq \widehat{\gamma}\}$.

Theorem 5. *Suppose that $\gamma < 1$ is a fixed indecision level and condition (9) holds for $0 < \phi \leq 2\phi^*_\gamma$ with $\beta' \geq \beta \geq 0$ and $\phi^*_\gamma > 0$. Then, there exists a positive universal constant c_1 such that*

$$\mathbf{P}\big(\widehat{\tau} - \tau^*_\gamma > \phi\big) \lesssim \frac{\mathbf{P}\big(|\widehat{\eta} - \eta| > c_1 \phi^{1+2\beta'-2\beta}\big)}{\phi^{1+2\beta'-\beta}},$$

for $0 < \phi \leq \phi^*_\gamma$. Moreover, we also have that

$$\mathbf{E}(\hat{\gamma}) - \gamma \lesssim \inf_{0<\phi\leq\phi^*_\gamma} \left\{ \phi^\beta + \frac{\mathbf{P}\big(|\widehat{\eta} - \eta| > c_1 \phi^{1+2\beta'-2\beta}\big)}{\phi^{1+2\beta'-\beta}} \right\}.$$

In the case $\beta = \beta' = 1$, we can expect to recover slow rates of classification, while in general consistency is not guaranteed, especially if $\eta_{\max}$ has some mass around τ^*_γ.

5 Extensions

First, we demonstrate that selective classifiers described in Sections 2 and 3 can completely avoid estimation of the regression function, through the monotone likelihood ratio (MLR) property. Second, we consider the multi-classification case and derive the minimax rules for this setting.

5.1 Adaptation under the MLR property

In both the binary and neyman pearson classification settings, we can avoid estimating the conditional density function $\eta(\cdot)$ in (3).

For simplicity, we assume for the remainder of this section that the distribution of X does not have any atoms, i.e., $\mathbf{P}(X = t) = 0 \ \forall t \in \mathbb{R}$. While the approach discussed above allows us to calibrate the optimal procedure with indecisions, it relies heavily on the prior knowledge of likelihood ratio $p_1 f_1/(p_2 f_2)$. In this section, we demonstrate how to achieve adaptation under the Monotone Likelihood Ratio (MLR) property, which is defined as follows:

The random variable X takes values in a subset of $\mathbb{R}$, the densities f_1 and f_2 have the same support, and $\frac{f_2}{f_1}(\cdot)$ is an increasing function on the support of the densities.

This property covers a large class of exponential family distributions. For example, it is satisfied for location models with a log-concave density. It is also satisfied for the chi-square location model where f_2 and f_1 are, respectively, a standard chi-square and a non-central chi-square densities. We refer the reader to Butucea et al. (2023) for more details about the MLR property.

More precisely, under the MLR property we can calibrate the oracle procedure based on observations of X and without prior knowledge of f_1 or f_2. For the Neyman Person testing problem, the optimal procedure under MLR is given by

$$Y^*_\gamma = \mathbf{1}(X \leq \tau_2) + 2 \times \mathbf{1}(X \geq \tau_1),$$

where τ_1, τ_2 are such that

$$\mathbf{P}_1(X \geq \tau_1) = \alpha_1(1-\gamma) \quad \text{and} \quad \mathbf{P}\left(\tau_2 \leq X \leq \tau_1\right) = \gamma.$$

The Type II error of Y_γ^* is given by

$$\mathcal{P}(\alpha_1, \gamma) = \frac{\mathbf{P}_2(X \leq \tau_2)}{1-\gamma}.$$

Recall that our goal is to calibrate γ so that $\mathcal{P}(\alpha_1, \gamma) = \alpha_2$.

Remark 3. *Observe that $\gamma > 0$ (i.e $\tau_2 < \tau_1$) if and only if $F_2^{-1}(\alpha_2) < 1 - F_1^{-1}(\alpha_1)$. In other words, we only need indecisions if the power of the NP test is below the target $1 - \alpha_2$.*

Given a calibration set of i.i.d. X_i and the corresponding labels, we can compute the above quantiles empirically and repeat the steps described above. It is interesting to note that under the MLR property the indecision set is an interval.

The case with accuracy is slightly more challenging, as the constraint

$$\mathbf{P}\left(1 - \tau_\gamma \leq \eta(X) \leq \tau_\gamma\right) = \gamma,$$

does not necessarily translate into a symmetric interval for X. This can be dealt with if we further assume that $\log\left(\frac{p_1 f_1}{p_2 f_2}\right)(\cdot)$ is an odd function. In particular, this is the case under mixtures of symmetric distribution such that $p_2 f_2(x) = p_1 f_1(-x)$. In that case, the optimal procedure becomes

$$Y_\gamma^* = 2 \times \mathbf{1}(X \geq \tau_\gamma) + \mathbf{1}(X \leq -\tau_\gamma),$$

where $\tau_\gamma \in [0, \infty)$ is such that

$$1 - 2\mathbf{P}(X \geq \tau_\gamma) = \mathbf{P}\left(-\tau_\gamma \leq X \leq \tau_\gamma\right) = \gamma.$$

Our goal here is to calibrate γ so that the misclassification rate of Y_γ^* satisfies

$$\mathbf{P}_Y(Y_\gamma^* \neq Y | Y_\gamma^* \neq 0) := \frac{p_1 \mathbf{P}_1\left(X \geq \tau_\gamma\right) + p_2 \mathbf{P}_2\left(X \leq -\tau_\gamma\right)}{1-\gamma} = \alpha.$$

Again, using a calibration set, we can estimate the above quantiles and recover the optimal classifier under indecisions.

5.2 Multi-classification

We now focus on the multi-class case. Assume that we observe a random variable X on a measurable space $(\mathcal{X}, \mathcal{U})$ such that X is distributed according to a mixture model, where with probability p_i its probability measure is given by P_i for $i = 1, \ldots, K$, and K is the number of classes. We assume that $P_i \neq P_j$ for any $i \neq j$. Let f_i be density of P_i with respect to some dominating measure that we will further denote by μ. Denote by Y the labeling quantity such that $Y = i$ if the distribution of X is P_i. We are interested in the problem of recovering the label Y.

As estimators of Y, we consider all measurable functions $\widehat{Y} = \widehat{Y}(X)$ of X taking values in $\{0, 1, \ldots, K\}$, where we allow for indecisions. Such estimators will be called *classifiers*. The performance of a classifier $\widehat{Y}$ is measured by its expected risk $\mathbf{P}_Y(\widehat{Y} \neq Y | \widehat{Y} \neq 0)$. We denote by $\mathbf{E}_Y$ the expectation with respect to probability measure $\mathbf{P}_Y$ of X for with labeling Y. Observe that

$$\mathbf{P}_Y(\widehat{Y} \neq Y | \widehat{Y} \neq 0) = \frac{\sum_{i=1}^K p_i \mathbf{P}_i(\widehat{Y} \notin \{i, 0\})}{\mathbf{P}(\widehat{Y} \neq 0)} = 1 - \frac{\sum_{i=1}^K p_i \mathbf{P}_i(\widehat{Y} = i)}{\mathbf{P}(\widehat{Y} \neq 0)}.$$

Given a level of indecisions γ, the minimax risk $\mathcal{R}$ is given by

$$\mathcal{R}(\gamma) := \inf_{\tilde{Y}} \mathbf{P}_Y(\tilde{Y} \neq Y | \tilde{Y} \neq 0),$$

where $\inf_{\tilde{Y}}$ denotes the infimum over all classifiers taking values in $\{1, \ldots, K\}$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$. Let us define the oracle classifier such that

$$Y_\gamma^* = \arg\max_i (p_i f_i) \mathbf{1}\left(\max_i (p_i f_i) \geq \tau_\gamma \sum_i p_i f_i \right), \tag{11}$$

where $\tau_\gamma \in [1, \infty)$ is such that

$$\mathbf{P}\left(\max_i (p_i f_i) \leq \tau_\gamma \sum_i p_i f_i) \right) = \gamma.$$

Theorem 6. *The classifier Y_γ^*, defined in (11), is minimax optimal for the risk $\mathcal{R}(\gamma)$. Moreover, we have*

$$\mathcal{R}(\gamma) = \mathbf{P}_Y(Y_\gamma^* \neq Y | Y_\gamma^* \neq 0) = 1 - \frac{\int_{Y^* \neq 0} \max_i (p_i f_i)}{1 - \gamma}.$$

Remark 4. *We can use plug-in scores to calibrate the procedure as we did in sections 2.3, 3.3, and 4.2. Note that for calibration, given a level of indecisions γ, the procedure does not require knowledge of the labels, which means we can calibrate it even in an unsupervised fashion.*

6 Simulations

We illustrate the theory established in this paper through three simulation studies. The first demonstrates that we can empirically recover the sharp phase transition in Section 4.1 under the two component gaussian mixture model. The second demonstrates the finite sample procedure for hypothesis testing in Section 3.2. A surprising empirical observation of this section is that as long as the type I error probability of the initial classifier can be controlled, any amount of indecisions will typically lower the type II error probability of our classifier, even if this amount cannot be picked optimally.

6.1 Phase Transition: Gaussian Mixture Model

In the symmetric Gaussian mixture model, and given a vanishing misclassification level δ, we wish to compare the theoretical versus the empirical values of γ_δ. We recall the parameters c and m

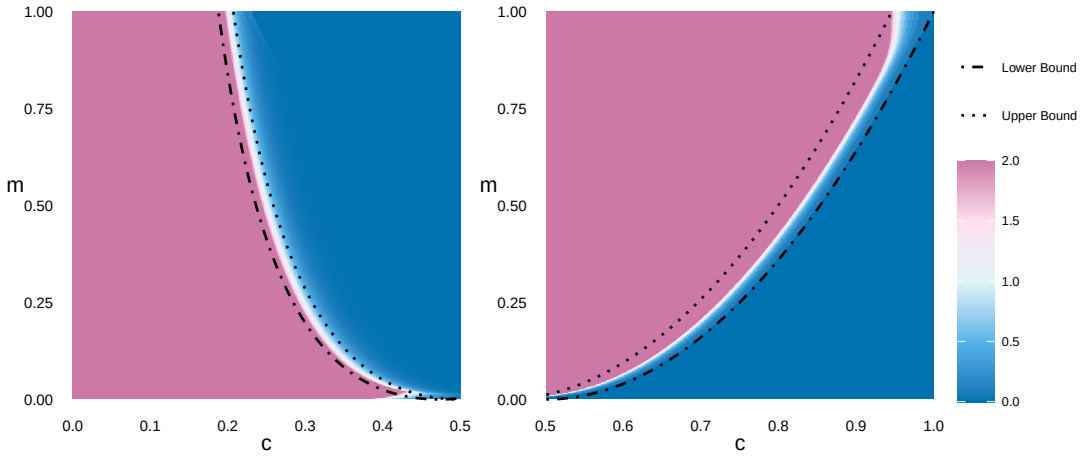

Figure 5: Computation of $\mathcal{R}(\gamma)/\delta$ with $\delta = 10^{-7}$ (left) and $\delta = 10^{-15}$ (right). The lower bound corresponds to the curve $m_*(c)$ while the upper bound corresponds to $m^*(c)$.

such that $\Delta(c) = c\sqrt{2\log(1/\delta)}$ and $\gamma_\delta(m) = (1-\delta^m)\mathbf{1}\{0 < c < 1/2\} + \delta^m \mathbf{1}\{1/2 < c < 1\}$, where we have made the dependence of γ_δ on m explicit.

According to Theorem 3, and using the above parameterization, δ-consistency is possible whenever

$$m \le m^*(c) \quad \text{if} \quad 0 < c < 1/2 \qquad \text{and} \qquad m \ge m^*(c) \quad \text{if} \quad 1/2 < c < 1,$$

Based on the proof of Theorem 3, it also holds that δ-consistency is impossible whenever

$$m \le m_*(c) \quad \text{if} \quad 0 < c < 1/2 \qquad \text{and} \qquad m \ge m_*(c) \quad \text{if} \quad 1/2 < c < 1,$$

where the lower bound $m_*(c)$ is given by

$$m_*(c) = \begin{cases} (c - (1-\varepsilon)/(4c))^2 & \text{if } 0 < c < 1/2, \\ (2c - 1 + \varepsilon)^2 & \text{if } 1/2 < c < 1, \end{cases} \tag{12}$$

for $\varepsilon = 1/2\log(4\pi\log(1/\gamma_\delta))/\log(1/\gamma_\delta)$. Observe that $m_*(c) \to m^*(c)$ as $\delta \to 0$.

In what follows, we fix $\delta = 10^{-15}$ for $c > 1/2$, and $\delta = 10^{-7}$ for $c < 1/2$. Our simulation setup is defined as follows. We set c on a uniform grid of 1000 points delimited by 0 and 1. Similarly, we set m on a uniform grid of 1000 points delimited by 0 and 1. For each combination of values of c and m, we first find t_γ using a grid search, such that $\mathbf{P}(\xi \ge \Delta(c) - t_\gamma) - \mathbf{P}(\xi \ge \Delta(c) + t_\gamma) = \gamma_\delta$, where ξ is a standard normal. Next, we compute $\mathcal{R}(\gamma_\delta) = \mathbf{P}(\xi \ge \Delta(c) + t_\gamma)/(1-\gamma_\delta)$. To ensure better interpretability, values of $\mathcal{R}(\gamma_\delta)/\delta$ outside the range $(0.5, 2)$ were capped at this range in Figure 5. As specified by our theory, the optimal amount of indecision $\log(1/\gamma_\delta)$ (or $\log(1/(1-\gamma_\delta))$) falls in the range delimited by m_* and m^*.

6.2 Binary Classification and the NP-testing Paradigm

We present our finite-sample calibration algorithms for binary classification and hypothesis testing, as detailed in Sections 2.3 and 3.3. These algorithms calibrate classifiers that estimate the regression function η, enabling control over error rates or type I / type II errors, as implemented in Algorithms

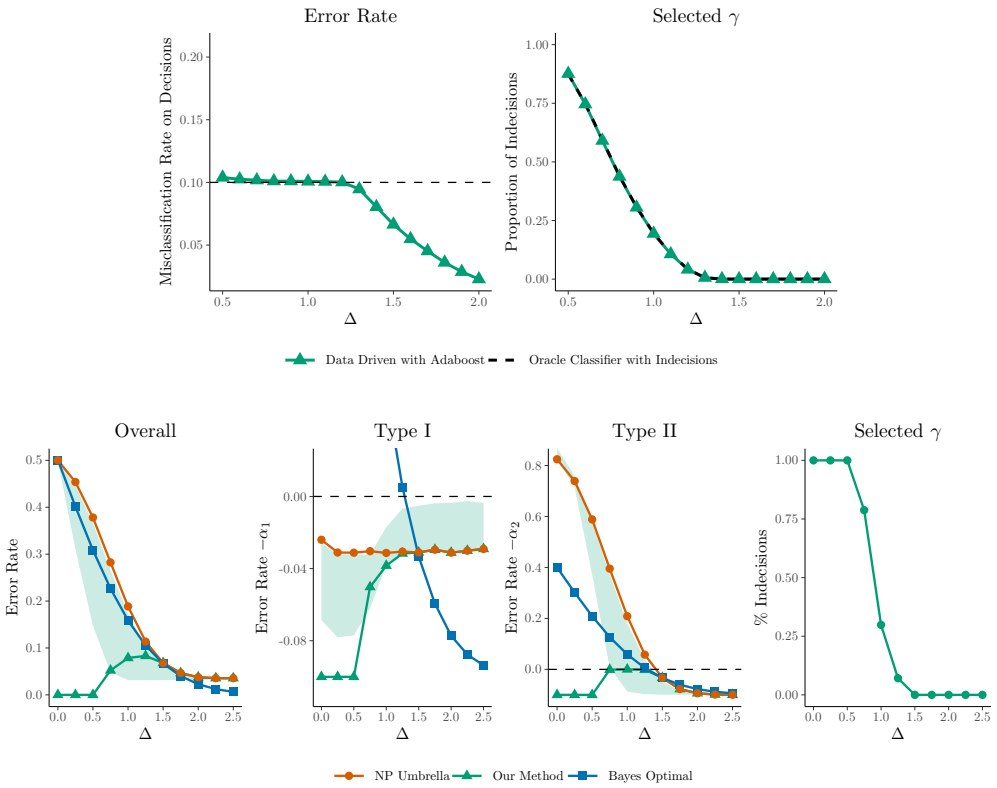

Figure 6: **Top:** Binary classification comparison between our GAM-based method (green triangles) and the Bayes optimal classifier (black dashed), which knows the true regression function η. The goal is to control the error rate on definitive decisions at 10%. Left: achieved error rate; right: proportion of indecisions required.**Bottom:** Controlling type I and type II errors at 10%, compared to the NP umbrella algorithm and Bayes classifier, which do not use indecisions. Panels show overall error, type I error, type II error, and selected indecision mass γ. Green shading indicates the 90% range of achievable error rates under varying indecision levels.

1 and 2. Each simulation uses $1,000$ training, calibration, and test samples drawn from a balanced Gaussian mixture with unit variance. The separation $\Delta = |\mu_1 - \mu_2|/2$ reflects task difficulty: smaller Δ increases classification challenge. For each Δ, results are averaged over $1,000$ simulations.

The top row of Figure 6 evaluates Algorithm 1 at the target error level $\alpha = 10\%$. The left plot compares our method (green / triangle), using GAM (Hastie et al., 2017), to the oracle classifier (black / dashed) that knows the true η. Both control error rates across all Δ, with the oracle achieving exact control. As Δ increases, our method becomes conservative, and the right plot shows its indecision rate closely matches the oracle's.

For hypothesis testing, Algorithm 2 requires estimating both error types. We first use the Neyman-Pearson (NP) Classification Umbrella Algorithm to control type I error, implemented via the `nproc` package in R (Tong et al., 2018), and then apply indecisions to control type II error. The second row of Figure 6 compares three methods: (1) NP-classifier (orange / circle), which controls only type I error; (2) our method (green / triangle), which controls both error types using selective inference; and (3) the Bayes optimal classifier (blue / square), which minimizes overall error without indecisions. All methods use LDA (Hastie et al., 2017) to estimate η . We target

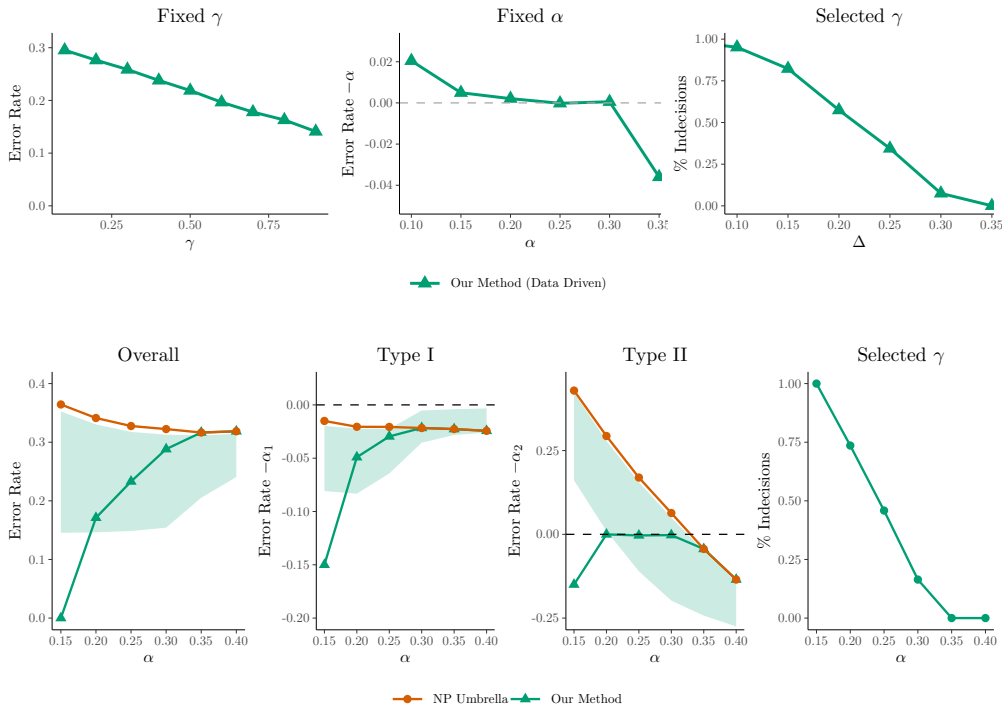

Figure 7: Predicting criminal recidivism on the COMPAS dataset using the NP Umbrella Algorithm and Our Method with indecisions. **Top:** Binary classification results for fixed indecision mass (left) and fixed error rate (middle), with the required indecision proportion shown on the right. **Bottom:** Overall error, type I error, type II error, and selected γ for controlling both error types. Green shading indicates the 90% range of achievable error rates under varying indecision levels.

$\alpha_1 = \alpha_2 = 0.1$.

We plot the overall, type I, and type II error rates across Δ. Our method consistently outperforms the NP-classifier in overall and type II error. In some cases, it even surpasses the Bayes classifier, as shown by the green shaded region falling below the oracle. The rightmost plot shows the optimal indecision mass γ needed to control both error types. As Δ increases, γ rapidly declines, aligning with the regime where the Bayes classifier also meets error constraints. Finally, we note that in this simulation setting, the use of indecisions almost always leads to a reduction in the classifier's type II error rate compared to the baseline NP-classifier. Therefore, even if a practitioner is uncertain about selecting the optimal value of γ for their specific application, they can expect a decrease in type II error and consequently, a reduction in the risk associated with definitive decisions.

7 Real Data

Predicting criminal recidivism is a well-studied application of automated decision systems. A prominent example is the COMPAS algorithm by NorthPoint Inc., used in the U.S. to assess a defendant's likelihood of reoffending. Given the high stakes of this task, ensuring prediction accuracy is critical, regardless of this task's inherent difficulty.

We analyze a dataset originally collected by ProPublica to investigate fairness in machine learn-

ing (Angwin et al., 2016). While fairness is not our focus here, our methodology could be extended to known protected groups. We approach the task using both binary classification and hypothesis testing, as outlined in Section 3, applying our data-driven Algorithms 1 and 2.

The dataset has $6,172$ defendants, with $2,990$ recidivating within a two-year window, which we use as the ground truth. We perform 100 random splits into observed and test sets, with the observed set further divided into training and calibration subsets. An ensemble model, averaging Naive Bayes, logistic regression, and AdaBoost classifiers is trained, and error rates and indecision masses γ are averaged across test sets (Hastie et al., 2017).

Figure 7 summarizes our results. The top row shows binary classification outcomes: the left plot varies indecision mass γ to minimize error, revealing that error decreases with more indecisions. The middle plot controls misclassification rate α, showing deviations from target accuracy, while the right plot tracks the required indecision mass. As α increases, the task becomes easier and fewer indecisions are needed.

The bottom row presents hypothesis testing results: overall error, type I and type II errors, and the minimal γ needed to control both. Green shading indicates the 5th–95th percentile range of achievable error rates. Middle plots show deviations from α, values above zero indicate uncontrolled error. Our method (green / triangle) consistently outperforms the NP-classifier (orange / circle), especially in reducing type II error. For small α, our method conservatively estimates type I error, but for $\alpha > 0.3$, control improves significantly. As the error tolerance increases, the required indecision mass drops sharply, consistent with our simulation findings.

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica 23*, 77–91.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Butucea, C., E. Mammen, M. Ndaoud, and A. B. Tsybakov (2023). Variable selection, monotone likelihood ratio and group sparsity. *The Annals of Statistics 51*(1), 312–333.

Cannon, A., J. Howse, D. Hush, and C. Scovel (2002). Learning with the neyman-pearson and min-max criteria. Technical Report LA-UR-02-2951, Los Alamos National Laboratory. Los Alamos National Laboratory Technical Report.

Chow, C. K. (1957, Dec). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers EC-6*(4), 247–254.

Cordella, L. P., C. De Stefano, C. Sansone, and M. Vento (1995). An adaptive reject option for lvq classifiers. In C. Braccini, L. DeFloriani, and G. Vernazza (Eds.), *Image Analysis and Processing*, Berlin, Heidelberg, pp. 68–73. Springer Berlin Heidelberg.

Cortes, C., G. DeSalvo, and M. Mohri (2016). Learning with rejection. In R. Ortner, H. U. Simon, and S. Zilles (Eds.), *Algorithmic Learning Theory*, Cham, pp. 67–82. Springer International Publishing.

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

El-Yaniv, R. and Y. Wiener (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research 11*(53), 1605–1641.

Fukunaga, K. and D. Kessell (1972). Application of optimum error-reject functions (corresp.). *IEEE Transactions on Information Theory 18*(6), 814–817.

Gang, B., Y. Shi, and W. Sun (2022). A locally adaptive shrinkage approach to false selection rate control in high-dimensional classification.

Grandvalet, Y., A. Rakotomamonjy, J. Keshet, and S. Canu (2008). Support vector machines with a reject option. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems*, Volume 21. Curran Associates, Inc.

Hastie, T., R. Tibshirani, and J. Friedman (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). Springer Series in Statistics. New York: Springer.

Hellman, M. E. (1970). The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics 6*(3), 179–185.

Hendrickx, K., L. Perini, D. V. der Plas, W. Meert, and J. Davis (2024). Machine learning with a reject option: A survey.

Herbei, R. and M. H. Wegkamp (2006). Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 34*(4), 709–721.

Huo, Y., L. Lu, H. Ren, and C. Zou (2024). Real-time selection under general constraints via predictive inference. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), *Advances in Neural Information Processing Systems*, Volume 37, pp. 61267–61305. Curran Associates, Inc.

Jin, Y. and E. J. Candes (2023). Selection by prediction with conformal p-values. *Journal of Machine Learning Research 24*(244), 1–41.

Lei, J. (2014). Classification with confidence. *Biometrika 101*(4), 755–769.

Marandon, A. (2024). Conformal link prediction for false discovery rate control. *TEST 33*, 1062–1083.

Minsker, S., M. Ndaoud, and Y. Shen (2025). Classification in the high dimensional anisotropic mixture framework: A new take on robust interpolation. *Journal of Machine Learning Research 26*(153), 1–39.

Ndaoud, M. (2022). Sharp optimal recovery in the two component gaussian mixture model. *The Annals of Statistics 50*(4), 2096–2126.

Rava, B., W. Sun, G. M. James, and X. Tong (2025). A burden shared is a burden halved: A fairness-adjusted approach to classification. *In: arXiv preprint arXiv:2110.05720 (2025)*.

Rigollet, P. and X. Tong (2011). Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research 12*(86), 2831–2855.

Schreuder, N. and E. Chzhen (2021, 27–30 Jul). Classification with abstention but without disparities. In C. de Campos and M. H. Maathuis (Eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Volume 161 of *Proceedings of Machine Learning Research*, pp. 1227–1236. PMLR.

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

Seo, S., M. Wallat, T. Graepel, and K. Obermayer (2000). Gaussian process regression: active data selection and test point rejection. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Volume 3, pp. 241–246 vol.3.

Shekhar, S., M. Ghavamzadeh, and T. Javidi (2019). Binary classification with bounded abstention rate.

Sun, W. and Z. Wei (2015, 05). Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika 102*(2), 267–280.

Tong, X. (2013). A plug-in approach to neyman-pearson classification. *Journal of Machine Learning Research 14*(92), 3011–3040.

Tong, X., Y. Feng, and J. J. Li (2018). Neyman-pearson classification algorithms and np receiver operating characteristics. *Science Advances 4*(2), eaao1659.

Vailaya, A. and A. Jain (2000). Reject option for vq-based bayesian classification. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Volume 2, pp. 48–51 vol.2.

Valade, F., M. Hebiri, and P. Gay (2024). Eero: Early exit with reject option for efficient classification with limited budget.

Vovk, V., A. Gammerman, and C. Saunders (1999). Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, San Francisco, CA, USA, pp. 444–453. Morgan Kaufmann Publishers Inc.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. New York, NY: Springer.

Wang, W., B. Gang, and W. Sun (2024). Sparse recovery with multiple data streams: An adaptive sequential testing approach. *Journal of Machine Learning Research 25*(304), 1–59.

Wegkamp, M. and M. Yuan (2011). Support vector machines with a reject option. *Bernoulli 17*(4), 1368–1385.

Yuan, M. and M. Wegkamp (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research 11*(5), 111–130.

Zhao, G. and Z. Su (2023). Controlling fsr in selective classification. *arXiv preprint arXiv:2311.03811*.

Ask for More Than Bayes Optimal: A Theory of Indecisions for Classification
Supplementary Material

By: Mohamed Ndaoud[1], Peter Radchenko[2], and Bradley Rava[2]

A Main Proofs

Lemma 3. *Let f and g be two positive functions and $c > 0$ and let $\mathcal{H}_c = \{A \ / \ \int_A f = c\}$. Assuming that $\mathcal{H}_c$ is not empty, we have that any*

$$A_c^* \in \arg\min_{A \in \mathcal{H}_c} \int_A g$$

is of the form $A_c^* := \{x \ / \ g(x) < t_c \cdot f(x)\} \cup \mathcal{M}_c$ for some $t_c \geq 0$ where $\mathcal{M}_c \subset \{x \ / \ g(x) = t_c \cdot f(x)\}$ such that $\int_{A_c^*} f = c$.

In particular if for all t, $|\{x \ / \ g(x) = t \cdot f(x)\}| = 0$, then A_c^* is unique up to Lebesgue negligible sets and $A_c^* := \{x \ / \ g(x) \leq t_c \cdot f(x)\}$ almost surely.

Proof. Observe that we may assume that $f > 0$, since for any $A \in \mathcal{H}_c$ we also have that $B = A \cap \{x \ / \ f(x) > 0\} \in \mathcal{H}_c$ and $\int_A g \geq \int_B g$. For the sake of generality we consider f and g to be simply positive.

Assuming that A_c^* exists, then for any $A \in \mathcal{H}_c$ we have

$$\begin{aligned}\int_A g - \int_{A_c^*} g &= \int_{A/A_c^*} g - \int_{A_c^*/A} g \\ &\geq t_c \int_{A/A_c^*} f - t_c \int_{A_c^*/A} f \\ &\geq t_c \left(\int_A f - \int_{A_c^*} f \right) \geq 0.\end{aligned}$$

It follows that

$$A_c^* \in \arg\min_{A \in \mathcal{H}_c} \int_A g.$$

We next show that, for any c, there exists $A_c^* := \{x \ / \ g(x) < t_c \cdot f(x)\} \cup \mathcal{M}_c$ for some $t_c \geq 0$ where $\mathcal{M}_c \subset \{x \ / \ g(x) = t_c \cdot f(x)\}$ and such that $\int_{A_c^*} f = c$. Let h be an application such that

$$h : t \to \int_{\mathcal{H}_t} f,$$

[1]Department of Decisions Sciences, ESSEC Business School, ndaoud@essec.edu

[2]Discipline of Business Analytics, University of Sydney Business School, peter.radchenko@sydney.edu.au, bradley.rava@sydney.edu.au

where $\mathcal{H}_t = \{x \ / \ g(x) \leq t \cdot f(x)\}$. It is clear that h is an increasing function and hence we can define for any $m \geq 0$

$$h^{-1}(m) = \inf\{t \ / \ h(t) \geq m\}.$$

Let us set $t_c = h^{-1}(c)$ and $\mathcal{F}_c := \{x \ / \ g(x) = t_c \cdot f(x)\}$.

If $h(t_c) = c$, then we are done with $\mathcal{M}_c = \mathcal{F}_c$. Otherwise $h(t_c) > c$ and for any $t < t_c$, $h(t) < c$. In particular $c^- := \lim_{t \to t_c^-} h(t) < h(t_c)$ and $h(t_c) - c^- = \int_{\mathcal{F}_c} f$.

We conclude that there exists $\mathcal{M}_c$ a subset of $\mathcal{F}_c$ such that

$$c - c^- = \int_{\mathcal{M}_c} f.$$

By setting $A_c^* = \{x \ / \ g(x) < t_c \cdot f(x)\} \cup \mathcal{M}_c$, it comes out that

$$\int_{A_c^*} f = c^- + \int_{\mathcal{M}_c} f = c.$$

It remains to show that any minimiser A^* satisfies almost surely

$$\{x \ / \ g(x) < t_c \cdot f(x)\} \subset A^* \subset \{x \ / \ g(x) \leq t_c \cdot f(x)\}.$$

Let us use the following notation $B_1 = \{x \ / \ g(x) > t_c \cdot f(x)\}, B_2 = \{x \ / \ g(x) < t_c \cdot f(x)\}$ and $B_3 = \{x \ / \ g(x) = t_c \cdot f(x)\}$. In that case we have

$$\int_{A^*} g = \int_{A^* \cap B_1} g + \int_{A^* \cap B_2} g + \int_{A^* \cap B_3} g.$$

It comes out that

$$0 = \int_{A^*} g - \int_{A_c^*} g = \int_{A^* \cap B_1} g + \int_{(A^*/A_c^*) \cap B_3} g - \int_{B_2/A^*} g - \int_{(A_c^*/A^*) \cap B_3} g.$$

Similarly we also have that

$$0 = \int_{A^*} f - \int_{A_c^*} f = t_c \int_{A^* \cap B_1} f + t_c \int_{(A^*/A_c^*) \cap B_3} f - t_c \int_{B_2/A^*} f - t_c \int_{(A_c^*/A^*) \cap B_3} f.$$

Combining both equations and the fact that $g(x) = t_c \cdot f(x)$ on B_3 leads to

$$\int_{A^* \cap B_1} (g - t_c \cdot f) = \int_{B_2/A^*} (g - t_c \cdot f) = 0.$$

So either we have that $A^* \cap B_1 = \emptyset$ or $B_2/A^* = \emptyset$. This concludes the proof. □

A.1 Proof of Theorem 1

Let us consider a classifier $\tilde{Y}(X)$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$ and let A be the set where $\tilde{Y} = 0$. We have that

$$\begin{aligned}\mathbf{P}_Y(\tilde{Y} \neq Y | \tilde{Y} \neq 0) &= \frac{p_1 \mathbf{P}_1(\tilde{Y} = 2) + p_2 \mathbf{P}_2(\tilde{Y} = 1)}{1 - \mathbf{P}(\tilde{Y} = 0)} \\ &= \frac{\int_{A^c} \mathbf{1}(\tilde{Y}(x) = 2) p_1 f_1(x) + \mathbf{1}(\tilde{Y}(x) = 1) p_2 f_2(x)}{1 - \gamma}.\end{aligned}$$

For each x, the integrand is minimized for $\tilde{Y} = (3 - \text{sign}(p_1 f_1 \geq p_2 f_2))/2$. Hence

$$\mathbf{P}_Y(\tilde{Y} \neq Y | \tilde{Y} \neq 0) \geq \frac{\int_{A^c}(p_1 f_1 \wedge p_2 f_2))}{1-\gamma}.$$

Invoking Lemma 3 we get further that the above quantity is minimized for

$$A^* := \left\{ x \, ; \, \frac{p_1 f_1 \wedge p_2 f_2}{p_1 f_1 + p_2 f_2}(x) > 1 - \tau_\gamma \right\} \cup \mathcal{M}_\gamma,$$

such that $\mathbf{P}(A^*) = \gamma$ and $\tau_\gamma \in [1/2, 1]$. The result follows and the expression of Y^* as well.

A.2 Proof of Proposition 1

The first part is straightforward from Theorem 1. Next observe that τ_γ is increasing by definition. Moreover $\lim_{\gamma \to 1^-} \tau_\gamma \leq 1$ and $\lim_{\gamma \to 0^+} \tau_\gamma \geq 1/2$.

On the one hand, let $\beta_1 \geq \beta_2$ and hence $\tau_{\beta_1} \geq \tau_{\beta_2}$. We have

$$\begin{aligned}
\mathcal{R}(\beta_2) &= \frac{\int_{A^{*c}_{\beta_2}} (p_1 f_1 \wedge p_2 f_2)}{1-\beta_2} \\
&= \frac{\int_{A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2)}{1-\beta_2} + \frac{\int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2)}{1-\beta_2} \\
&\geq \frac{1}{1-\beta_2} \left(\int_{A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2) + \frac{\int_{A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2)}{1-\beta_1} \int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}} (p_1 f_1 + p_2 f_2) \right),
\end{aligned}$$

where we have used, in the last inequality, the fact that on $A^{*c}_{\beta_2}/A^{*c}_{\beta_1}$ we have that

$$\frac{(p_1 f_1 \wedge p_2 f_2)}{(p_1 f_1 + p_2 f_2)} \geq (1 - \tau_{\beta_1}),$$

while on $A^{*c}_{\beta_1}$ we have

$$\frac{(p_1 f_1 \wedge p_2 f_2)}{(p_1 f_1 + p_2 f_2)} \leq (1 - \tau_{\beta_1}).$$

As a consequence we have that

$$\frac{(p_1 f_1 \wedge p_2 f_2)}{(p_1 f_1 + p_2 f_2)} \geq \frac{\int_{A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2)}{1-\beta_1}$$

on $A^{*c}_{\beta_2}/A^{*c}_{\beta_1}$. It comes out that

$$\mathcal{R}(\beta_2) \geq \frac{\int_{A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_{-1} f_{-1})}{1-\beta_2} \frac{1-\beta_2}{1-\beta_1} \geq \mathcal{R}(\beta_1).$$

It comes out that $\mathcal{R}(\gamma)$ is non-increasing.

On the other hand, we have that

$$\mathcal{R}(\beta_2) - \mathcal{R}(\beta_1) = \frac{(\beta_2 - \beta_1) \int_{A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2)}{(1-\beta_2)(1-\beta_1)} + \frac{\int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2)}{1-\beta_2}.$$

On the event $A^{*c}_{\beta_2}/A^{*c}_{\beta_1}$ we have that

$$p_1 f_1 \wedge p_2 f_2 \leq (p_1 f_1 + p_2 f_2)(1 - \tau_{\beta_2}).$$

Hence

$$\int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}} (p_1 f_1 \wedge p_2 f_2) \leq \frac{(1 - \tau_{\beta_2})(\beta_2 - \beta_1)}{1 - \beta_2}.$$

As a consequence we get that

$$0 \leq \mathcal{R}(\beta_2) - \mathcal{R}(\beta_1) \leq \frac{2(\beta_2 - \beta_1)}{1 - \beta_2}.$$

We conclude that $\gamma \to \mathcal{R}(\gamma)$ is continuous. This proof is complete.

A.3 Proof of Theorem 2

Let us consider a classifier $\tilde{Y}(X)$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$ and $\mathbf{P}_1(\tilde{Y} = 2) = \alpha(1 - \gamma)$. Let A be the set where $\tilde{Y} = 0$ and let B be the set where $\tilde{Y} = 2$. We have that

$$\begin{aligned}\frac{\mathbf{P}_2(\tilde{Y} = 1)}{1 - \gamma} &= \frac{\int_{B^c \cap A^c} f_2}{1 - \gamma} \\ &= \frac{\int_{A^c} f_2}{1 - \gamma} - \frac{\int_B f_2}{1 - \gamma}.\end{aligned}$$

Using Lemma 3 we get further that the above quantity is minimized for B^* such that

$$B^* = \left\{x \ / \ f_2 > \tau^u_{\alpha,\gamma} f_1\right\} \cup \mathcal{M}_{\alpha,\gamma}.$$

It comes out that

$$\frac{\mathbf{P}_2(\tilde{Y} = 1)}{1 - \gamma} \geq \frac{\int_{A^c} f_2 \mathbf{1}(B^{*c})}{1 - \gamma}.$$

Hence using Lemma 3 again we get

$$A^* := \left\{x \ ; \ f_1/\tau^l_{\alpha,\gamma} < f_2 \leq \tau^u_{\alpha,\gamma} f_1\right\} \setminus \mathcal{M}^c_{\alpha,\gamma} \cup \mathcal{A}_{\alpha,\gamma},$$

such that $\mathbf{P}(A^*) = \gamma$. The result follows and the expression of Y^* as well.

A.4 Proofs for Section 4.2

A.4.1 Proof of Lemma 2

Note that

$$\begin{aligned}(1 - \gamma)\big[R(\widehat{Y}_\gamma) - R(Y^*_\gamma)\big] = E\eta\big(&\mathbf{1}\{\widehat{Y}_\gamma = 2, Y^*_\gamma \neq 2\} - \mathbf{1}\{Y^*_\gamma = 2, \widehat{Y}_\gamma \neq 2\}\big) \\ &+ E(1 - \eta)\big(\mathbf{1}\{\widehat{Y}_\gamma = 1, Y^*_\gamma \neq 1\} - \mathbf{1}\{Y^*_\gamma = 1, \widehat{Y}_\gamma \neq 1\}\big). \qquad \text{(A.1)}\end{aligned}$$

Also note that

$$\begin{aligned}\mathbf{1}\{\widehat{Y}_\gamma = 2, Y^*_\gamma \neq 2\} - \mathbf{1}\{Y^*_\gamma = 2, \widehat{Y}_\gamma \neq 2\} + \mathbf{1}\{\widehat{Y}_\gamma = 1, Y^*_\gamma \neq 1\} - \mathbf{1}\{Y^*_\gamma = 1, \widehat{Y}_\gamma \neq 1\} \\ = \mathbf{1}\{Y^*_\gamma = 0\} - \mathbf{1}\{\widehat{Y}_\gamma = 0\}.\end{aligned}$$

Consequently, equality $\mathbf{P}(Y_\gamma^* = 0) = \mathbf{P}(\widehat{Y}_\gamma = 0)$ yields

$$E\big(\mathbf{1}\{\widehat{Y}_\gamma = 2, Y_\gamma^* \neq 2\} - \mathbf{1}\{Y_\gamma^* = 2, \widehat{Y}_\gamma \neq 2\}\big) + E\big(\mathbf{1}\{\widehat{Y}_\gamma = 1, Y_\gamma^* \neq 1\} - \mathbf{1}\{Y_\gamma^* = 1, \widehat{Y}_\gamma \neq 1\}\big) = 0. \tag{A.2}$$

Combining equations (A.1) and (A.2), we derive

$$\begin{aligned}(1-\gamma)\big[R(\widehat{Y}_\gamma) - R(Y_\gamma^*)\big] &= E\big(\eta - [1-\tau_\gamma^*]\big)\big(\mathbf{1}\{\widehat{Y}_\gamma = 2, Y_\gamma^* \neq 2\} - \mathbf{1}\{Y_\gamma^* = 2, \widehat{Y}_\gamma \neq 2\}\big) \\ &+ E\big(1-\eta - [1-\tau_\gamma^*]\big)\big(\mathbf{1}\{\widehat{Y}_\gamma = 1, Y_\gamma^* \neq 1\} - \mathbf{1}\{Y_\gamma^* = 1, \widehat{Y}_\gamma \neq 1\}\big).\end{aligned}$$

Note that $\eta < 1 - \tau_\gamma^*$ if and only if $Y_\gamma^* = 2$. Also note that $\eta > \tau_\gamma^*$ if and only if $Y_\gamma^* = 1$. Hence, we can rewrite the above display as follows:

$$\begin{aligned}(1-\gamma)\big[R(\widehat{Y}_\gamma) - R(Y_\gamma^*)\big] &= E|1-\tau_\gamma^* - \eta|\big(\mathbf{1}\{\widehat{Y}_\gamma = 2, Y_\gamma^* \neq 2\} + \mathbf{1}\{Y_\gamma^* = 2, \widehat{Y}_\gamma \neq 2\}\big) \\ &+ E|\tau_\gamma^* - \eta|\big(\mathbf{1}\{\widehat{Y}_\gamma = 1, Y_\gamma^* \neq 1\} + \mathbf{1}\{Y_\gamma^* = 1, \widehat{Y}_\gamma \neq 1\}\big),\end{aligned}$$

which completes the proof. □

A.4.2 Proof of Theorem 4

Define event A_ϕ as follows:

$$\begin{aligned}A_\phi &= \{\tau_\gamma^* < \eta \leq \widehat{\tau}_\gamma - \phi\} \cup \{\tau_\gamma^* < 1 - \eta \leq \widehat{\tau}_\gamma - \phi\} \\ &\quad \cup \{\widehat{\tau}_\gamma + \phi < \eta \leq \tau_\gamma^*\} \cup \{\widehat{\tau}_\gamma + \phi < 1 - \eta \leq \tau_\gamma^*\}.\end{aligned}$$

We will use the following result, which is proved in Section A.4.4.

Lemma 4. *For $0 < \phi \leq \phi_\gamma^*$, we have*

$$\mathbf{P}\big(A_\phi\big) \leq \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) \qquad \textit{and} \qquad \mathbf{P}\big(|\widehat{\tau}_\gamma - \tau_\gamma^*| > 2\phi\big) \lesssim \frac{\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big)}{\phi^{\beta'}}. \tag{A.3}$$

It follows from the proof of Lemma 2, that the equality in the statement of Lemma 2 continues to hold when τ_γ^* is replaced by an arbitrary constant c. Moreover, another small modification to the proof allows us to replace c with $\widehat{\tau}_\gamma$. We will focus on the first of the four terms in the resulting expression for $(1-\gamma)[R(\widehat{Y}_\gamma) - R(Y_\gamma^*)]$ – the other three terms can be handled by analogous arguments. The term of interest can be bounded as follows:

$$E|\widehat{\tau}_\gamma - \eta|\mathbf{1}\{Y_\gamma^* = 1, \widehat{Y}_\gamma \neq Y_\gamma^*\} \leq E_1 + E_2 + E_3,$$

where

$$\begin{aligned}E_1 &= \mathbf{P}\big(\eta > \widehat{\tau}_\gamma + \phi, Y_\gamma^* = 1, \widehat{Y}_\gamma \neq Y_\gamma^*\big), \\ E_2 &= \mathbf{P}\big(A_\phi\big), \quad \text{and} \\ E_3 &= \phi\mathbf{P}\big(|\eta - \widehat{\tau}_\gamma| \leq \phi, Y_\gamma^* = 1, \widehat{Y}_\gamma \neq Y_\gamma^*\big).\end{aligned}$$

Note that $\{Y_\gamma^* = 1, \widehat{Y}_\gamma \neq Y_\gamma^*\} = \{\eta > \tau_\gamma^*, \widehat{\eta} \leq \widehat{\tau}_\gamma\}$. Consequently, taking into account Lemma 4, we derive

$$E_1 + E_2 \leq 2\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big). \tag{A.4}$$

We also have

$$\begin{aligned} E_3 &= \phi\,\mathbf{P}\big(|\widehat{\tau}_\gamma - \tau^*_\gamma| > 2\phi,\ |\eta - \widehat{\tau}_\gamma| \le \phi,\ \eta > \tau^*_\gamma,\ \widehat{\eta} \le \widehat{\tau}_\gamma\big) \\ &\quad + \phi\,\mathbf{P}\big(|\widehat{\tau}_\gamma - \tau^*_\gamma| \le 2\phi,\ |\eta - \widehat{\tau}_\gamma| \le \phi,\ \eta > \tau^*_\gamma,\ \widehat{\eta} \le \widehat{\tau}_\gamma\big) \\ &\le \phi\,\mathbf{P}\big(|\widehat{\tau}_\gamma - \tau^*_\gamma| > 2\phi\big)\,\mathbf{P}\big(\eta > \tau^*_\gamma\big) + \phi\,\mathbf{P}\big(|\eta - \tau^*_\gamma| \le 3\phi\big) \\ &\lesssim \Big(\phi^{1-\beta'}\,\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) \wedge \phi\Big)\,(1-\gamma) + \phi^{1+\beta}. \end{aligned}$$

where we used Lemma 4 and condition (9) to derive the final bound. Thus, we get the desired bound for the first term in our resulting expression for $(1-\gamma)[R(\widehat{Y}_\gamma) - R(Y^*_\gamma)]$:

$$E|\widehat{\tau}_\gamma - \eta|\mathbf{1}\{Y^*_\gamma = 1, \widehat{Y}_\gamma \ne Y^*_\gamma\} \lesssim \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) + (\phi^{1-\beta'}\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) \wedge \phi)(1-\gamma) + \phi^{1+\beta}. \quad \text{(A.5)}$$

The other three terms can be similarly bounded using analogous arguments. This completes the proof of claim (10) in Theorem 4. □

A.4.3 Proof of Theorem 5

To simplify the notation, we will write $Y^{*\tau}$ and $\widehat{Y}^\tau$ for the classifiers $\widehat{Y}$ and Y^* that use τ as the threshold. For example,

$$\widehat{Y}^\tau(X) = 1 \times \mathbf{1}\{\widehat{\eta}(X) > \tau\} + 2 \times \mathbf{1}\{\widehat{\eta}(X) < 1 - \tau\}.$$

Recall that γ is the indecision level of the classifier Y^* that uses threshold τ^*_γ; also recall that $R^* = R(Y^{*\tau^*_\gamma})$. We will use the following result, which is proved in Section A.4.5.

Lemma 5. *For ϵ, ϕ such that $0 < \epsilon \le \phi \le \phi^*_\gamma \wedge (1/2 - \tau^*_\gamma/2)$, we have*

$$\begin{aligned} R^* - R(Y^{*\tau^*_\gamma+\phi}) &\gtrsim \phi^{1+2\beta'-\beta} \qquad \textit{and} \\ R_{\widehat{\eta}}(\widehat{Y}^{\tau^*_\gamma+\phi}) - R(Y^{*\tau^*_\gamma+\phi}) &\lesssim \mathbf{P}\{|\widehat{\eta} - \eta| > \epsilon \,|\, \widehat{\eta}\} + \epsilon\phi^\beta. \end{aligned}$$

By the definitions of $\widehat{\tau}$ and τ^*_γ, the event $\{\widehat{\tau} > \tau^*_\gamma + \phi\}$ implies $\{R_{\widehat{\eta}}(\widehat{Y}^{\tau^*_\gamma+\phi}) > R^*\}$. Hence,

$$\mathbf{P}\big(\widehat{\tau} > \tau^*_\gamma + \phi\big) \le \mathbf{P}\Big(R_{\widehat{\eta}}(\widehat{Y}^{\tau^*_\gamma+\phi}) - R(Y^{*\tau^*_\gamma+\phi}) > R^* - R(Y^{*\tau^*_\gamma+\phi})\Big).$$

Using Lemma 5 to bound the components of the event on the right-hand side in the above display, we derive

$$\begin{aligned} \mathbf{P}\big(\widehat{\tau} > \tau^*_\gamma + \phi\big) &\le \mathbf{P}\Big(\mathbf{P}\{|\widehat{\eta} - \eta| > \epsilon \,|\, \widehat{\eta}\} + \epsilon\phi^\beta \gtrsim \phi^{1+2\beta'-\beta}\Big) \\ &\le \mathbf{P}\Big(\mathbf{P}\{|\widehat{\eta} - \eta| > \epsilon \,|\, \widehat{\eta}\} \gtrsim \phi^{1+2\beta'-\beta}\Big) + \mathbf{P}\big(\epsilon\phi^\beta \gtrsim \phi^{1+2\beta'-\beta}\big) \\ &\lesssim \frac{\mathbf{P}\{|\widehat{\eta} - \eta| > \epsilon\}}{\phi^{1+2\beta'-\beta}} + \mathbf{P}\big(\epsilon \gtrsim \phi^{1+2\beta'-2\beta}\big). \end{aligned}$$

We take $\epsilon = c_1\phi^{1+2\beta'-2\beta}$ and note that we can choose c_1 sufficiently small to ensure that the second term in the line above is zero (recall that $\beta' \ge \beta$). This completes the proof of the first bound in Theorem 5. The second bound in Theorem 5 follows from the first bound together with condition (9). □

A.4.4 Proof of Lemma 4

Note that $\widehat{\tau}_\gamma$ is fully determined by $\widehat{\eta}$. When $\widehat{\tau}_\gamma \geq \tau^*_\gamma + \phi$, we have

$$\mathbf{P}\big(1 - \widehat{\tau}_\gamma + \phi \leq \eta \leq \widehat{\tau}_\gamma - \phi \,|\, \widehat{\eta}\big) = \gamma + \mathbf{P}\big(\eta \in A_\phi \,|\, \widehat{\eta}\big). \tag{A.6}$$

We also have

$$\begin{aligned} \mathbf{P}\big(1 - \widehat{\tau}_\gamma + \phi \leq \eta \leq \widehat{\tau}_\gamma - \phi \,|\, \widehat{\eta}\big) &\leq \mathbf{P}\big(1 - \widehat{\tau}_\gamma \leq \widehat{\eta} \leq \widehat{\tau}_\gamma \,|\, \widehat{\eta}\big) + \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big) \\ &= \gamma + \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big). \end{aligned} \tag{A.7}$$

Combining (A.6) and (A.7), we derive

$$\mathbf{P}\big(\eta \in A_\phi \,|\, \widehat{\eta}\big) \leq \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big). \tag{A.8}$$

When $\tau^*_\gamma - \phi < \widehat{\tau}_\gamma < \tau^*_\gamma + \phi$, we have $\mathbf{P}\big(\eta \in A_\phi \,|\, \widehat{\eta}\big) = 0$, and hence inequality (A.8) still holds. Now consider the last remaining case: $\widehat{\tau}_\gamma \leq \tau^*_\gamma - \phi$. Note that

$$\mathbf{P}\big(1 - \widehat{\tau}_\gamma - \phi \leq \eta \leq \widehat{\tau}_\gamma + \phi \,|\, \widehat{\eta}\big) = \gamma - \mathbf{P}\big(\eta \in A_\phi \,|\, \widehat{\eta}\big). \tag{A.9}$$

We also have

$$\begin{aligned} \gamma &= \mathbf{P}\big(1 - \widehat{\tau}_\gamma \leq \widehat{\eta} \leq \widehat{\tau}_\gamma \,|\, \widehat{\eta}\big) \\ &\leq \mathbf{P}\big(1 - \widehat{\tau}_\gamma - \phi \leq \eta \leq \widehat{\tau}_\gamma + \phi \,|\, \widehat{\eta}\big) + \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big). \end{aligned} \tag{A.10}$$

Combining (A.9) and (A.10), we again derive inequality (A.8), concluding that (A.8) holds for all possible $\widehat{\eta}$. Integrating (A.8) over $\widehat{\eta}$, we derive the first claim of Lemma 4.

For the second claim of Lemma 4, we will focus on bounding $\mathbf{P}\big(\widehat{\tau}_\gamma > \tau^*_\gamma + 2\phi\big)$; the complementary bound on $\mathbf{P}\big(\widehat{\tau}_\gamma < \tau^*_\gamma - 2\phi\big)$ follows analogously. Note that $\widehat{\tau}_\gamma > \tau^*_\gamma + 2\phi$ implies

$$\begin{aligned} \mathbf{P}\big(1 - \tau^*_\gamma - \phi \leq \eta \leq \tau^*_\gamma + \phi\big) &\leq \mathbf{P}\big(1 - \tau^*_\gamma - 2\phi \leq \widehat{\eta} \leq \tau^*_\gamma + 2\phi \,|\, \widehat{\eta}\big) + \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big) \\ &\leq \gamma + \mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big). \end{aligned} \tag{A.11}$$

By condition (9), we also have

$$\mathbf{P}\big(1 - \tau^*_\gamma - \phi \leq \eta \leq \tau^*_\gamma + \phi\big) \geq \gamma + c\phi^{\beta'}, \tag{A.12}$$

for some fixed positive constant c. Combining (A.11) and (A.12), we deduce that $\widehat{\tau}_\gamma > \tau^*_\gamma + 2\phi$ implies $\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big) \geq c\phi^{\beta'}$. Applying Markov inequality, we then conclude that

$$\mathbf{P}\big(\widehat{\tau}_\gamma > \tau^*_\gamma + 2\phi\big) \leq \mathbf{P}\Big(\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi \,|\, \widehat{\eta}\big) \geq c\phi^{\beta}\Big) \leq \frac{\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big)}{c\phi^{\beta'}}. \quad \square$$

A.4.5 Proof of Lemma 5

Let γ_ϕ be the indecision level corresponding to classifier Y^* with threshold $\tau^*_\gamma + \phi$, and let $\widehat{\gamma}_\phi$ be the indecision level corresponding to $\widehat{Y}$ with threshold $\tau^*_\gamma + \phi$. Define $\eta_{\min} = \eta \wedge (1 - \eta)$ and

$\eta_{\max} = \eta \vee (1-\eta)$, and note that

$$\begin{aligned}
R(Y^{*\tau_\gamma^*}) - R(Y^{*\tau_\gamma^*+\phi}) &= \tfrac{1}{(1-\gamma)}\mathbf{E}\eta_{\min}\mathbf{1}\{Y_\gamma^* \neq 0\} - \tfrac{1}{1-\gamma_\phi}\mathbf{E}\eta_{\min}\mathbf{1}\{Y_{\gamma_\phi}^* \neq 0\} \\
&= \tfrac{1}{(1-\gamma)}\mathbf{E}\eta_{\min}\big(\mathbf{1}\{Y_\gamma^* \neq 0\} - \mathbf{1}\{Y_{\gamma_\phi}^* \neq 0\}\big) + \big(\tfrac{1}{(1-\gamma)} - \tfrac{1}{1-\gamma_\phi}\big)\mathbf{E}\eta_{\min}\mathbf{1}\{Y_{\gamma_\phi}^* \neq 0\} \\
&= \tfrac{1}{(1-\gamma)}\mathbf{E}\eta_{\min}\mathbf{1}\{\tau_\gamma^* < \eta_{\max} \leq \tau_\gamma^* + \phi\} + \tfrac{(\gamma-\gamma_\phi)}{(1-\gamma)(1-\gamma_\phi)}\mathbf{E}\eta_{\min}\mathbf{1}\{Y_{\gamma_\phi}^* \neq 0\} \\
&\geq \tfrac{1}{(1-\gamma)}\Big[\mathbf{E}\eta_{\min}\mathbf{1}\{\tau_\gamma^* < \eta_{\max} \leq \tau_\gamma^* + \phi\} - (\gamma_\phi - \gamma)(1-\tau-\phi)\Big] \\
&= \tfrac{(\gamma_\phi-\gamma)}{(1-\gamma)}\Big[\phi - \mathbf{E}\Big(\eta_{\max} - \tau_\gamma^* \,\Big|\, \tau_\gamma^* < \eta_{\max} \leq \tau_\gamma^* + \phi\Big)\Big].
\end{aligned}$$

Condition (9) implies that $\gamma_\phi - \gamma \gtrsim \phi^{\beta'}$ and $\mathbf{E}\big(\eta_{\max} \,\big|\, \tau_\gamma^* < \eta_{\max} \leq \tau_\gamma^* + \phi\big) \leq \tau_\gamma^* + \phi^{1+\beta'-\beta}$. Indeed, using condition (9) once again, we have

$$\begin{aligned}
\mathbf{E}\Big(\eta_{\max} - \tau_\gamma^* \,\Big|\, \tau_\gamma^* < \eta_{\max} \leq \tau_\gamma^* + \phi\Big) &\leq \frac{\frac{\phi}{2}\mathbf{P}(0 < \eta_{\max} - \tau_\gamma^* \leq \phi/2) + \phi\left(\gamma_\phi - \gamma - \mathbf{P}(0 < \eta_{\max} - \tau_\gamma^* \leq \phi/2)\right)}{\gamma_\phi - \gamma} \\
&\leq \phi - \frac{\phi\,\mathbf{P}(0 < \eta_{\max} - \tau_\gamma^* \leq \phi/2)}{2\,\mathbf{P}(0 < \eta_{\max} - \tau_\gamma^* \leq \phi)} \\
&\leq \phi^{1+\beta'-\beta}.
\end{aligned}$$

Consequently, $R(Y^{*\tau_\gamma^*}) - R(Y^{*\tau_\gamma^*+\phi}) \gtrsim \phi^{1+2\beta'-\beta}$, and we have derived the first bound of Lemma 5.

Taking advantage of the fact that the threshold used by $\widehat{Y}^{\tau_\gamma^*+\phi}$ and $Y^{*\tau_\gamma^*+\phi}$ is the same, and repeating the standard argument in Herbei and Wegkamp (2006) while conditioning on $\widehat{\eta}$, we derive that

$$R_{\widehat{\eta}}(\widehat{Y}^{\tau_\gamma^*+\phi}) - R(Y^{*\tau_\gamma^*+\phi}) \;\lesssim\; \mathbf{P}\{|\widehat{\eta} - \eta| > \epsilon \,|\, \widehat{\eta}\} + \epsilon\big[\mathbf{P}\big(|\tau_\gamma^* + \phi - \eta| \leq \epsilon\big) + \mathbf{P}\big(|1 - \tau_\gamma^* - \phi - \eta| \leq \epsilon\big)\big].$$

Thus, using $\epsilon \leq \phi$ together with condition (9) we arrive at the second bound of Lemma 5. □

A.5 Proof of Theorem 6

Let us consider a classifier $\tilde{Y}(X)$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$ and let A be the set where $\tilde{Y} = 0$. We have that

$$\begin{aligned}
\mathbf{P}_Y(\tilde{Y} \neq Y | \tilde{Y} \neq 0) &= 1 - \frac{\sum_i p_i \mathbf{P}_i(\tilde{Y} = i)}{1 - \mathbf{P}(\tilde{Y} = 0)} \\
&= 1 - \frac{\int_{A^c} \sum_i \mathbf{1}(\tilde{Y}(x) = i) p_i f_i(x)}{1-\gamma}.
\end{aligned}$$

For each x, the integrand is minimized for $\tilde{Y} = \arg\max_i (p_i f_i)$. Hence

$$\mathbf{P}_Y(\tilde{Y} \neq Y | \tilde{Y} \neq 0) \geq 1 - \frac{\int_{A^c} \max_i (p_i f_i))}{1-\gamma}.$$

Invoking Lemma 3 we get further that the above quantity is minimized for

$$A^* := \left\{ x \;;\; \max_i p_i f_i \leq \tau_\gamma \sum_i p_i f_i \right\},$$

such that $\mathbf{P}(A^*) = \gamma$ and $\tau_\gamma \in [1, \infty)$. The result follows and the expression of Y^* as well.

A.6 Proof of Theorem 3

The following bound for the tail of the Gaussian distribution will be useful for this proof. For all $t \geq 0$, we have

$$\frac{\exp^{-t^2/2}}{\sqrt{2\pi}(t+1)} \leq \mathbf{P}(\xi \geq t) \leq \frac{\exp^{-t^2/2}}{\sqrt{2\pi}t}.$$

We start with the case $1/2 < c < 1$:

Remember that $\Delta = c\sqrt{2\log(1/\delta)}$. Let us choose $t = (1-c)\sqrt{2\log(1/\delta)}$. In that case

$$\frac{\delta}{\sqrt{2\pi}(\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq \Delta + t) \leq \frac{\delta}{\sqrt{4\pi\log(1/\delta)}},$$

and

$$\frac{\delta^{(2c-1)^2}}{\sqrt{2\pi}((2c-1)\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq \Delta - t) \leq \frac{\delta^{(2c-1)^2}}{(2c-1)\sqrt{4\pi\log(1/\delta)}}.$$

It comes out that

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq t - \Delta)} \leq \frac{\delta}{\sqrt{4\pi\log(1/\delta)}\left(1 - \frac{\delta^{(2c-1)^2}}{\sqrt{2\pi}((2c-1)\sqrt{2\log(1/\delta)})}\right)}.$$

It is now clear that for small values of δ we have

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq \Delta + t) + \mathbf{P}(\xi \geq t - \Delta)} \leq \delta.$$

As a consequence

$$\gamma_\delta \leq \frac{\delta^{(2c-1)^2}}{(2c-1)\sqrt{4\pi\log(1/\delta)}} - \frac{\delta}{\sqrt{2\pi}(\sqrt{2\log(1/\delta)}+1)}. \tag{A.13}$$

For $\varepsilon > 0$, let us now choose $t = (1-c-\varepsilon)\sqrt{2\log(1/\delta)}$. In that case

$$\frac{\delta^{1-\varepsilon}}{\sqrt{2\pi}((1-\varepsilon)\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq \Delta + t) \leq \frac{\delta^{1-\varepsilon}}{(1-\varepsilon)\sqrt{4\pi\log(1/\delta)}},$$

and

$$\frac{\delta^{(2c-1+\varepsilon)^2}}{\sqrt{2\pi}((2c-1+\varepsilon)\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq \Delta - t) \leq \frac{\delta^{(2c-1+\varepsilon)^2}}{(2c-1+\varepsilon)\sqrt{4\pi\log(1/\delta)}}.$$

It comes out that

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq t - \Delta)} \geq \frac{\delta^{1-\varepsilon}}{\sqrt{2\pi}((1-\varepsilon)\sqrt{2\log(1/\delta)+1})\left(1 - \frac{\delta^{(2c-1+\varepsilon)^2}}{\sqrt{2\pi}((2c-1+\varepsilon)\sqrt{2\log(1/\delta)}+1)}\right)}.$$

It is now clear that for small values of δ we have

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq \Delta + t) + \mathbf{P}(\xi \geq t - \Delta)} \geq \delta.$$

As a consequence for any $\varepsilon > 0$, we get

$$\gamma_\delta \geq \frac{\delta^{(2c-1+\varepsilon)^2}}{\sqrt{2\pi}((2c-1+\varepsilon)\sqrt{2\log(1/\delta)}+1)} - \frac{\delta^{1-\varepsilon}}{(1-\varepsilon)\sqrt{4\pi\log(1/\delta)}}. \tag{A.14}$$

Combining (A.13) and (A.14), we conclude that if $\log(1/\gamma) \leq (2c-1)^2 \log(1/\delta)$ then $\delta-$consistency is possible. On the other hand, if $\log(1/\gamma) \geq \sqrt{1+\varepsilon}(2c-1)^2 \log(1/\delta)$ then $\delta-$consistency is impossible.

We will next deal with the case $0 < c < 1/2$:

Remember that $\Delta = c\sqrt{2\log(1/\delta)}$. Let us choose $t = 1/(4c)\sqrt{2\log(1/\delta)}$. In that case

$$\frac{\delta^{(c+1/(4c))^2}}{\sqrt{2\pi}((c+1/(4c))\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq \Delta + t) \leq \frac{\delta^{(c+1/(4c))^2}}{(c+1/(4c))\sqrt{4\pi\log(1/\delta)}},$$

and

$$\frac{\delta^{(c-1/(4c))^2}}{\sqrt{2\pi}((1/(4c)-c)\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq t - \Delta) \leq \frac{\delta^{(c-1/(4c)))^2}}{(1/(4c)-c)\sqrt{4\pi\log(1/\delta)}}.$$

It comes out that

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq t - \Delta)} \leq \delta \frac{\sqrt{2\pi}((1/(4c)-c)\sqrt{2\log(1/\delta)}+1)}{(c+1/(4c))\sqrt{4\pi\log(1/\delta)}}.$$

It is now clear that for small values of δ and any $c > 0$ we have

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq \Delta + t) + \mathbf{P}(\xi \geq t - \Delta)} \leq \delta.$$

As a consequence

$$\gamma_\delta \leq 1 - \frac{\delta^{(c-1/(4c))^2}}{\sqrt{2\pi}((1/(4c)-c)\sqrt{2\log(1/\delta)}+1)} - \frac{\delta^{(c+1/(4c))^2}}{\sqrt{2\pi}((c+1/(4c))\sqrt{2\log(1/\delta)}+1)}. \tag{A.15}$$

For a choice of $\varepsilon > 0$ close to 0, let us now choose $t = (1-\varepsilon)/(4c)\sqrt{2\log(1/\delta)}$ such that $t > \Delta$. In that case

$$\frac{\delta^{(c+(1-\varepsilon)/(4c))^2}}{\sqrt{2\pi}((c+(1-\varepsilon)/(4c))\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq \Delta + t) \leq \frac{\delta^{(c+(1-\varepsilon)/(4c))^2}}{(c+(1-\varepsilon)/(4c))\sqrt{4\pi\log(1/\delta)}},$$

and

$$\frac{\delta^{(c-(1-\varepsilon)/(4c))^2}}{\sqrt{2\pi}(((1-\varepsilon)/(4c)-c)\sqrt{2\log(1/\delta)}+1)} \leq \mathbf{P}(\xi \geq t - \Delta) \leq \frac{\delta^{(c-(1-\varepsilon)/(4c)))^2}}{((1-\varepsilon)/(4c)-c)\sqrt{4\pi\log(1/\delta)}}.$$

It comes out that

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq t - \Delta)} \geq \delta^{1-\varepsilon} \frac{\sqrt{2\pi}(((1-\varepsilon)/(4c)-c)\sqrt{2\log(1/\delta)})}{\sqrt{2\pi}((c+(1-\varepsilon)/(4c))\sqrt{2\log(1/\delta)}+1)}.$$

It is now clear that for small values of δ we have

$$\frac{\mathbf{P}(\xi \geq \Delta + t)}{\mathbf{P}(\xi \geq \Delta + t) + \mathbf{P}(\xi \geq t - \Delta)} \geq \delta.$$

As a consequence, for small values of $\varepsilon > 0$, we get

$$\gamma_\delta \geq 1 - \frac{\delta^{(c-(1-\varepsilon)/(4c))^2}}{\sqrt{2\pi}(((1-\varepsilon)/(4c)-c)\sqrt{2\log(1/\delta)})} - \frac{\delta^{(c+(1-\varepsilon)/(4c))^2}}{\sqrt{2\pi}((c+(1-\varepsilon)/(4c))\sqrt{2\log(1/\delta)})}. \tag{A.16}$$

Combining (A.15) and (A.16), we conclude that if $\log(1/(1-\gamma)) \geq (c-1/(4c))^2 \log(1/\delta)$ then $\delta-$consistency is possible. On the other hand, if $\log(1/(1-\gamma)) \leq \sqrt{1-\varepsilon}(c-1/(4c))^2 \log(1/\delta)$ then $\delta-$consistency is impossible.