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 frameworks are useful tools for automated decision making in highly risky scenarios, since they allow for a classifier to only make highly confident decisions, while abstaining from making a decision when it is not confident enough to do so, which is otherwise known as an indecision. For a given level of classification accuracy, we aim to make as many decisions as possible. For many problems, this can be achieved without abstaining from making decisions. But when the problem is hard enough, we show that we can still control the misclassification rate of a classifier up to any user specified level, while only abstaining from the minimum necessary amount of decisions, even if this level of misclassification is smaller than the Bayes optimal error rate. In many problem settings, the user could obtain a dramatic decrease in misclassification while only paying a comparatively small price in terms of indecisions.

Keywords: Selective Inference, Classification, Indecisions, Phase Transition.

1 Introduction

In this work, 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, as indicated by the distance between the true distributions for each decision class, 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 is to classify various observations as High-Risk or Low-Risk, while maintaining an error rate below 5%. In this example, the High-Risk and Low-Risk classes are modeled as mixtures of two normal distributions with means of 2 and 1, respectively, and a shared variance of 1. The Bayes classifier 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

[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

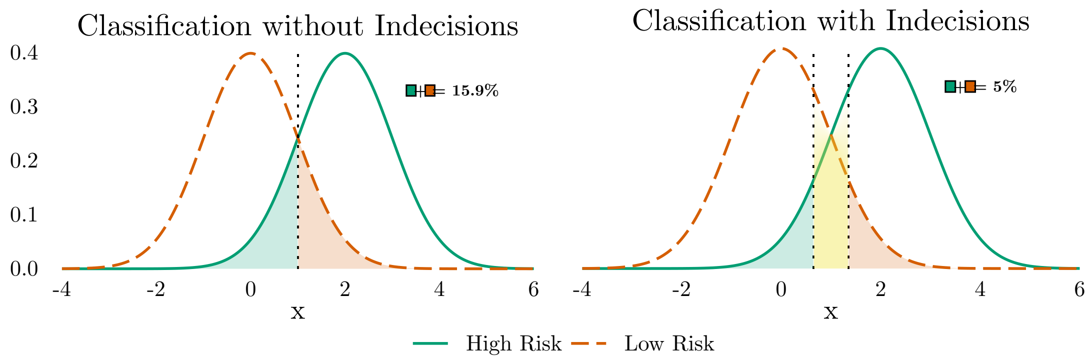

Figure 1: An example of a classification scenario where the data comes from two different normal distributions. Low Risk $\sim N(0,1)$ and High Risk $\sim N(2,1)$. Left plot: Classification with no indecisions. Right Plot: Classification with indecisions (highlighted in yellow). The indecisions do not 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.

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-stakes 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.

A key contribution of our work is demonstrating that, in many scenarios, achieving a lower misclassification rate than the Bayes error rate requires only a small number of indecisions. In other words, practitioners can realize significant improvements in accuracy while sacrificing only a small fraction of observations as indecisions.

Beyond focusing solely on accuracy, practical implementation necessitates minimizing the number of indecisions required to maintain the desired level of accuracy. This paper addresses two critical problems in selective classification. First, if a user is willing to allocate a fixed percentage of observations as indecisions, what improvement in accuracy can they expect? Second, if the user mandates a specified level of accuracy, how many indecisions will be necessary to achieve it? We present several contributions to these questions, which are discussed in detail in Section 1.3.

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 recovering the label Y, which can be viewed as hypothesis testing or as supervised classification when more labeled observations are given.

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 $\mathbf{1}(\cdot)$ denotes the indicator function.

The performance of a classifier $\widehat{Y}$ is measured by its expected risk $\mathbf{P}_Y(\widehat{Y} \neq Y)$ also known as the 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

$$Y^*(X) = \operatorname{sign}\left(p_1 f_1(X) - p_2 f_2(X)\right).$$

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 minimax risk, we might allow our estimator a degree of freedom where the classifier 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 $\mathcal{R}$ given by

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

where $\inf_{\tilde{Y}}$ denotes the infimum over all classifiers taking values in $\{0, 1, 2\}$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$. 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 different names, such as "Classification with a Reject Option", "Selective Classification", "No-decision classification", and "Human-AI collaboration". Regardless of the name used, the approaches involve classifiers that allowed to not make a decision when the class probabilities used for making a decision are too close to each other.

Some of the earliest work in this area dates back to 1957, when Chow (1957) explored the use of indecisions to improve the accuracy of character recognition systems. This was later extended by Chow (1970), who examined the trade-off between accuracy and the total number of indecisions within a Bayesian framework, assuming known data-generating distributions. Building on these foundational ideas, Dubuisson and Masson (1993) introduced the concept of leveraging indecisions to identify observations that lie far from the decision boundary. On the practical side, Fumera et al. (2000); Fumera and Roli (2004) advanced this work by addressing scenarios where density estimates are noisy due to estimation errors, where classifiers operate as ensembles, and by proposing the use of distinct thresholds for different classes. In the multi-class setting, Ni et al. (2019) investigated calibration challenges, highlighting its greater complexity compared to traditional binary classification. Additionally, Pillai et al. (2013) focused on minimizing the cost of indecisions and considered scenarios in which each instance could potentially belong to multiple classes.

In line with the focus of this paper, the seminal work of Herbei and Wegkamp (2006) developed a theoretical framework for binary classification with indecisions, employing empirical risk minimization to analyze the impact of plug-in classifiers and establish their convergence rates. Building on this foundation, Bartlett and Wegkamp (2008) proposed a convex surrogate loss function to enable efficient optimization when incorporating indecisions into learning algorithms, while also providing consistency and generalization bounds for their framework. Subsequent studies by El-Yaniv and Wiener (2010) and Wiener and El-Yaniv (2011) explored the trade-off between coverage and accuracy, along with achievable performance guarantees. Recently, Rava et al. (2024) also adopted this framework for the purposes of fairness.

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). Our theory is very 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.

The optimal procedure is based on thresholding the likelihood ratio between distributions f_0 and f_1 that can be encoded through the regression function η. In practice, we can use a training sample to learn η. In Section 3, 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.

Section 4 is dedicated to full adaptation, where both the training and calibration sets are finite. 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. We also 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 still show how to calibrate the procedure in this setting.

Finally, in Section 6 we study the phase transition of classification under a binary Gaussian mixture model for a fixed separation between the mixtures. 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 level 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. Our findings are illustrated by numerical experiments in Section 7.

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 $a_n \le cb_n$ ($a_n \ge cb_n$) 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.

2 General Binary Classification

Here, we focus on the binary case where we only have two classes. For a given value of $\gamma \in [0,1)$, let us define the optimal indecision region Θ_γ. We will show that there exists $\tau_\gamma \in [1/2,1]$ such that

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

where $\mathcal{M}_\gamma$ is any subset of $\left\{\frac{p_1 f_1(X) \wedge p_2 f_2(X)}{p_1 f_1(X)+p_2 f_2(X)} = 1-\tau_\gamma\right\}$ such that $\mathbf{P}_Y(\Theta_\gamma)=\gamma$.

If we define $\eta(\cdot)$ as the conditional density function such that $\eta(x) = \mathbf{P}(X=x|Y=1)$, we get that

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

It is natural to observe that the optimal indecision region concentrates around where $\eta(X)$ is close to $1/2$. It is interesting to compare our threshold τ_γ to the constant d in Herbei and Wegkamp (2006) as they play similar roles here. Observe also that when $\eta(X) \vee (1-\eta(X)) = \tau_\gamma$, i.e., we are at the frontier of indecision we might randomly choose to reject or not. The Bayes oracle classifier

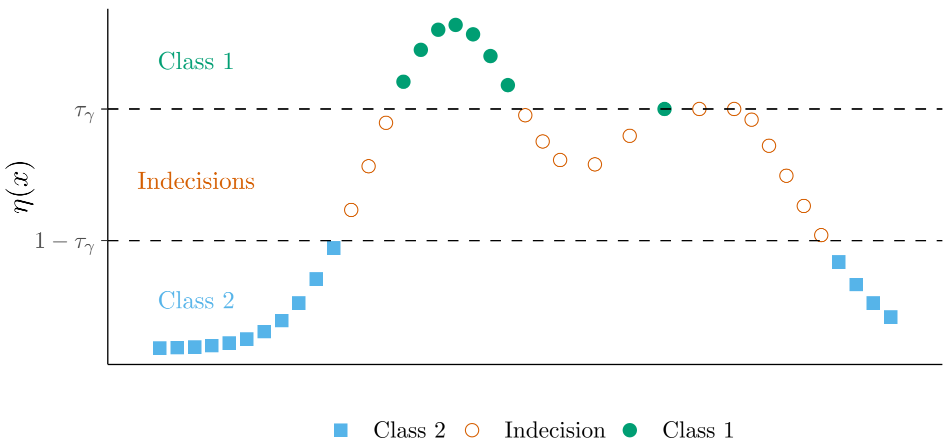

Figure 2: An example of binary classification with indecisions (orange / open circle). The threshold for class 1 (green / solid circle) and class 2 (blue / square) is τ_γ and $1-\tau_\gamma$ respectively. There is a plateau exactly at the threshold τ_γ, demonstrating that some observations may need to be classified into class 1 or as an indecision randomly.

is given by that

$$Y_\gamma^* = \arg\max_{i\in\{1,2\}} (p_i f_i(X))\mathbf{1}(\Theta_\gamma^c).$$

Theorem 1. *Given $\gamma \geq 0$, the classifier Y_γ^* is minimax optimal for the risk $\mathcal{R}(\gamma)$. Moreover, we have that*

$$\mathcal{R}(\gamma) = \mathbf{P}_Y(Y_\gamma^* \neq Y | Y_\gamma^* \neq 0) = \frac{\int_{\Theta_\gamma^c} (p_1 f_1 \wedge p_2 f_2) d\mu}{1-\gamma}.$$

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 the random variable $\eta(X)$ has atoms and $\mathbf{P}(\eta(X)\wedge(1-\eta(X)) = \tau_\gamma) \neq 0$, the set $\mathcal{M}$ is not trivial and we shall call the region $\mathcal{M}$ a "plateau" where the indecisions are picked randomly as shown in Figure 2. 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 Figure 3. Consequently, constructing the indecision region requires prior knowledge of η.

To get a better understanding of $\mathcal{R}(\gamma)$, we present the next proposition.

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

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

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 $\gamma \to \mathcal{R}(\gamma)$ is continuous and non-increasing .

Because $\mathcal{R}(\gamma)$ 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 shows also that

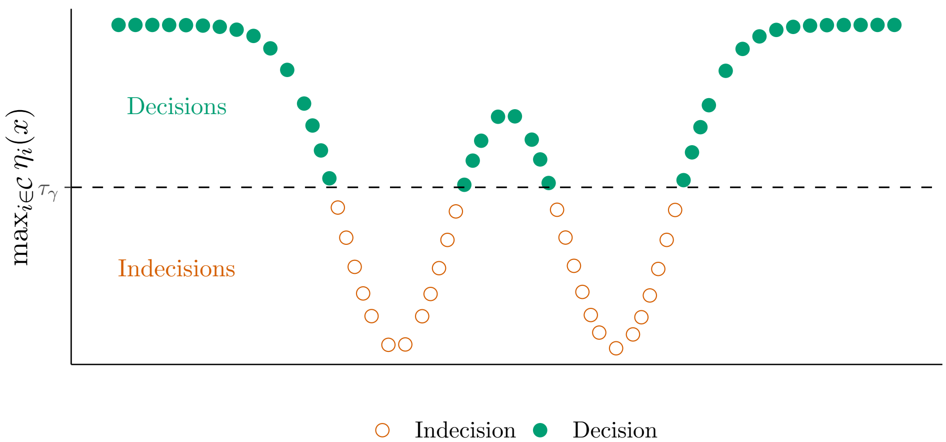

Figure 3: An example of classification with indecisions (orange / open-circle), with the threshold τ_γ separating the decisions (green / solid circle) and the indecisions. Here the indecision region is not an interval.

for any given misclassification level α above $\mathcal{R}^*$, we can find a γ such that $\mathcal{R}(\gamma) = \alpha$ and this γ is the smallest possible. In other words, for any reachable level for accuracy we are able characterize the minimum number of indecisions needed to reach that level.

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, as it supposes that the log-likelihood ratio between f_1 and f_2, given by $|\log(f_1/f_2)|$ shall go to ∞ if we wish to have high separation between the two distributions and hence a zero misclassification error.

2.1 A hypothesis testing perspective: Connection with the Neyman Pearson paradigm

We now consider the hypothesis testing problem, where given a type I error α we want to reach a certain level of the type II error, using indecisions if necessary. In this case, the risk to consider is

$$\mathcal{M}(\alpha, \gamma) := \inf_{\tilde{Y}} p_2 \mathbf{P}_2(\tilde{Y} = 1 | \tilde{Y} \neq 0) = \inf_{\tilde{Y}} \frac{p_2 \mathbf{P}_2(\tilde{Y} = 1)}{1 - \gamma},$$

where $\inf_{\tilde{Y}}$ denotes the infimum over all classifiers taking values in $\{0, 1, 2\}$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$, and the type I error is given by $p_1 \mathbf{P}_1(\tilde{Y} = 2) = \alpha(1 - \gamma)$. This means that we are interested in the best type II error given that we only make decisions for a given proportion of observations. Note that we normalized both errors by $(1-\gamma)$ as we have to think of these errors conditionally on the fact that we make a decision.

For given values of α and γ, let us define the following oracle classifier,

$$Y^*_{\alpha,\gamma} = \mathbf{1}\left(\left\{p_1 f_1 > \tau^l_{\alpha,\gamma} p_2 f_2\right\} \cup \mathcal{M}_{\alpha,\gamma}\right) + 2 \times \mathbf{1}\left(\left\{p_2 f_2 \geq \tau^u_{\alpha,\gamma} p_1 f_1\right\} \setminus \mathcal{A}_{\alpha,\gamma}\right),$$

for some $\tau^l_{\alpha,\gamma}, \tau^u_{\alpha,\gamma} \geq 0$ where:

- the set $\mathcal{M}_{\alpha,\gamma}$ is any subset of $\left\{p_2 f_2(X) = p_1 f_1(X) \tau^u_{\alpha,\gamma}\right\}$ such that

$$p_1 \mathbf{P}_1(p_2 f_2 > \tau^u_{\alpha,\gamma} p_1 f_1) + \mathbf{P}(\mathcal{M}_{\alpha,\gamma}) = \alpha(1-\gamma);$$

- and the set $\mathcal{A}_{\alpha,\gamma}$ is any subset of $\left\{p_1 f_1(X) = p_2 f_2(X) \tau^l_{\alpha,\gamma}\right\}$ such that

$$\mathbf{P}\left(\left\{1/\tau^l_{\alpha,\gamma} < \frac{p_2 f_2}{p_1 f_1}(X) \leq \tau^u_{\alpha,\gamma}\right\} \setminus \mathcal{M}_{\alpha,\gamma}\right) + \mathbf{P}(\mathcal{A}_{\alpha,\gamma}) = \gamma.$$

Theorem 2. *Classifier $Y^*_{\alpha,\gamma}$ is minimax optimal for the risk $\mathcal{M}(\alpha,\gamma)$. Moreover,*

$$\mathcal{M}(\alpha,\gamma) = \frac{p_2 \mathbf{P}_2(Y^*_{\alpha,\gamma} = 1)}{1-\gamma} = \frac{\int_{\{Y^*_{\alpha,\gamma}=1\}} p_2 f_2 d\mu}{1-\gamma}.$$

The next result is in the same spirit as Proposition 1, which focused on the accuracy.

Proposition 2. *For any α, function $\gamma \to \mathcal{M}(\alpha,\gamma)$ is continuous and non-increasing. Moreover, if for each $\varepsilon > 0$ we have $\mathbf{P}(p_2 f_2(X) \leq \varepsilon \cdot p_1 f_1(X)) > 0$, then $\lim_{\gamma \to 1^-} \mathcal{M}(\alpha,\gamma) = 0$.*

As claimed before, $\lim_{\gamma \to 1^-} \mathcal{M}(\alpha,\gamma)$ exists and shall be denoted $\mathcal{M}^*(\alpha)$. Hence, any value of the type II error within the range $[\mathcal{M}^*(\alpha), \mathcal{M}(\alpha,0)]$ can be reached using only the necessary amount of indecisions. The condition $\mathbf{P}(p_2 f_2(X) \leq \varepsilon \cdot p_1 f_1(X)) > 0$ 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 Plug-in rules

In this section, we replace classification probability function η by a learnt 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.

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 $\delta > 0$ can be reached using our framework.

3.1 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

$$\mathbf{P}\big(|\eta_{\max} - \tau_\gamma^*| \leq \phi\big) \lesssim \phi^\alpha \quad \text{and} \quad \mathbf{P}\big(|\eta_{\max} - \tau_\gamma^*| \leq \phi,\, \eta_{\max} \neq \tau_\gamma^*\big) \gtrsim \phi^{\alpha'}, \tag{2}$$

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

Theorem 3. *Suppose that (2) holds for $0 < \phi \leq 3\phi_\gamma^*$ with $\alpha' \geq \alpha \geq 0$ and $\phi_\gamma^* > 0$. Then,*

$$R(\widehat{Y}_\gamma) - R(Y_\gamma^*) \lesssim \inf_{0<\phi\leq\phi_\gamma^*} \tfrac{1}{1-\gamma}\left\{\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) + \phi^{1+\alpha}\right\} + (\phi^{1-\alpha'}\mathbf{P}\big(|\widehat{\eta} - \eta| > \phi\big) \wedge \phi). \tag{3}$$

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

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

Remark 2. *In the statement of Theorem 3, 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 $\frac{\phi^{1+\alpha}}{1-\gamma}$ by $\phi^{1+\alpha}$ if we assume that the margin condition (2) 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 $\alpha' \leq 1$, 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 controling the distance between thresholds $\hat{\tau}$ and τ. The lower bound in condition (2) 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 $\alpha = \alpha' = 1$, which is typically the case for atom-less distributions. Going back to the bound (3) and taking $\phi \approx 1/\sqrt{n}$, we recover the slow rates without making any assumptions on the margin.

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

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

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

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

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

4 Calibration/Methodology

We will start with the misclassification risk. Given a misclassification level δ in $[\mathcal{R}^*, \mathcal{R}(0)]$, our goal is to construct a classifier that achieves this misclassification level using the minimal number of indecisions.

4.1 The non adaptive case

Here, we assume prior knowledge of p_1, f_1, p_2, f_2. For a given misclassification level δ, there exists $\gamma(\delta)$ such that $\mathcal{R}(\gamma(\delta)) = \delta$. Next, we show next how to construct a classifier $\hat{Y}$ such that the accuracy of $\hat{Y}$ is at least $1 - \delta$ and the proportion of indecisions is of order $\gamma(\delta)$. In other words, $\hat{Y}$ is nearly minimax optimal. Let us define $\gamma_\delta := \gamma(\delta)$. We recall that the optimal indecision region is such that

$$\mathbf{P}\left(\eta(X) \wedge (1 - \eta(X)) > 1 - \tau_\delta\right) + \mathbf{P}(\mathcal{M}_{\gamma_\delta}) = \gamma_\delta.$$

Observe that τ_δ corresponds to a quantile of the random variable $\eta(X) \wedge (1-\eta(X))$ and can be easily computed. The misclassification error of $Y^*_{\gamma_\delta}$ is such that

$$\mathbf{P}_Y(Y^*_{\gamma_\delta} \neq Y | Y^*_{\gamma_\delta} \neq 0) = \delta.$$

Since we do not have access to γ_δ explicitly, we need to invert the function $\mathcal{R}(\cdot)$. In order to calibrate the value of τ_δ, we follow the steps below.

1. For each value of γ on a grid, compute τ_γ and then Y^*_γ. This should allow us to compute the misclassification of Y^*_γ given by $\mathcal{R}(\gamma)$.
2. Starting from $\gamma = 0$, stop at the first value of γ_δ such that $\mathcal{R}(\gamma_\delta) \leq \delta$.
3. Compute τ_δ for that value of γ_δ and return the corresponding classifier $Y^*_{\gamma_\delta}$.

Next, we move to the Neyman Pearson setup. For given levels α and $\beta \in [\mathcal{M}^*(\alpha), \mathcal{M}(\alpha, 0)]$ of Type I and Type II errors, respectively, there exists a $\gamma(\alpha,\beta)$ such that $\mathcal{M}(\gamma(\alpha,\beta)) = \beta$. Our goal now is to find a classifier $\hat{Y}$ such that the Type I error of $\hat{Y}$ is less than α, the Type II error is less than β, and the proportion of indecisions is of the same order as $\gamma(\alpha,\beta)$. We again define $\gamma_\beta := \gamma(\alpha,\beta)$. Recall that the Type I error is given by

$$p_1\mathbf{P}_1(p_2 f_2 > \tau^u_{\alpha,\gamma_\beta} p_1 f_1) + \mathbf{P}(\mathcal{M}_{\alpha,\gamma_\beta}) = \alpha(1-\gamma_\beta),$$

and the optimal indecision region is given by

$$\mathbf{P}\left(\left\{1/\tau^l_{\alpha,\gamma_\beta} < \frac{p_2 f_2}{p_1 f_1}(X) \leq \tau^u_{\alpha,\gamma_\beta}\right\} \setminus \mathcal{M}_{\alpha,\gamma_\beta}\right) + \mathbf{P}(\mathcal{A}_{\alpha,\gamma_\beta}) = \gamma_\beta.$$

The Type II error of $Y^*_{\gamma_\beta}$ is given by

$$\mathcal{M}(\alpha, \gamma_\beta) = \beta.$$

Since we do not have access to γ_β explicitly, we need to invert the function $\mathcal{M}(\alpha, \cdot)$. In order to calibrate the values of $\tau^u_\beta, \tau^l_\beta$, we follow the steps below.

1. For each value of γ on a grid, compute τ^u_γ and then τ^l_γ, which lead to Y^*_γ. This allows us to compute the Type II of Y^*_γ given by $\mathcal{M}(\alpha,\gamma)$.
2. Starting from $\gamma = 0$, stop at the first value of γ_β such that $\mathcal{M}(\alpha, \gamma_\beta) \leq \beta$.
3. Compute $\tau^u_\beta, \tau^l_\beta$ for that value of γ_β and return the corresponding classifier $Y^*_{\gamma_\beta}$.

4.2 Adaptation under the MLR property

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 take values in a subset of $\mathbf{R}$, 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

$$Y^*_\beta = \mathbf{1}(X \le \tau^l_{\alpha,\beta}) + 2 \times \mathbf{1}(X \ge \tau^u_{\alpha,\beta}),$$

where $\tau^u_{\alpha,\beta}, \tau^l_{\alpha,\beta}$ are such that

$$p_1 \mathbf{P}_1(X \ge \tau^u_{\alpha,\beta}) = \alpha(1-\gamma_\beta) \quad \text{and} \quad \mathbf{P}\left(\tau^l_{\alpha,\beta} \le X \le \tau^u_{\alpha,\beta}\right) = \gamma_\beta.$$

The Type II error of Y^*_β is given by

$$\mathcal{M}(\alpha, \gamma_\beta) = \frac{p_{-1}\mathbf{P}_2(X \le \tau^l_{\alpha,\beta})}{1-\gamma_\beta} = \beta.$$

Remark 3. *Observe that $\gamma_\beta > 0$ if and only if $F_2^{-1}(\beta/p_2) < 1 - F_1^{-1}(\alpha/p_1)$. In other words, we only allow indecisions if the power of the NP test is above the target β.*

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

$$\mathbf{P}\left(1 - \tau_{\gamma_\delta} \le \eta(X) \le \tau_{\gamma_\delta}\right) = \gamma_\delta,$$

does not necessarily translate into a symmetric interval on 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

$$Y^*_\delta = 2 \times \mathbf{1}(X \ge \tau_\delta) + \mathbf{1}(X \le -\tau_\delta),$$

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

$$1 - 2\mathbf{P}(X \ge \tau_\delta) = \mathbf{P}\left(-\tau_\delta \le X \le \tau_\delta\right) = \gamma_\delta.$$

The misclassification rate of Y^*_δ is given by

$$\mathbf{P}_Y(Y^*_\delta \ne Y | Y^*_\delta \ne 0) = \frac{p_1 \mathbf{P}_1\left(X \ge \tau_\delta\right) + p_2 \mathbf{P}_2\left(X \le -\tau_\delta\right)}{1-\gamma_\delta} = \delta.$$

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

5 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 $i = 1, \dots, K$, 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 $\{0, 1, \dots, K\}$, 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 $\{1, \dots, K\}$ such that $\mathbf{P}(\tilde{Y} = 0) = \gamma$. Let us define the oracle classifier such that

$$Y^*_\gamma = \arg\max_i (p_i f_i) \mathbf{1}\left(\max_i (p_i f_i) \geq t_\gamma \sum_i p_i f_i\right),$$

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

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

Theorem 5. *The classifier Y^*_γ 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 3 and 4. Note that the calibration, given a level of indecisions γ, does not require knowledge of the labels, which means we can calibrate the procedure even in an unsupervised fashion.*

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

This section is devoted to the asymptotic behavior of the risk as δ gets smaller. We will now focus on the symmetric Gaussian mixture model. In particular, we assume that $p_1 = p_{-1} = 1/2$ and that $f_1(x) = \frac{1}{\sqrt{2\pi}} \exp(-(x - \Delta)^2/2) = f_{-1}(-x)$ for some separation $\Delta > 0$. In this case, the MLR property for symmetric likelihoods holds. 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 that parameters Δ and γ depend on δ. For the sake of readability, we do not equip these parameters with the subscript δ.

The asymptotic property we study here is $\delta-$*consistency*, which is inspired by consistency in classification Minsker et al. (2021) 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 $\delta > 0$.*

- *We say that $\delta-$*consistency is impossible *for γ_δ if*

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

- *We say that $\delta-$*consistency is possible *for γ_δ if there exists an estimator $\hat{Y}$, with indecision level γ_δ, such that*

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

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 and δ. For a misclassification level δ, we have that

$$\mathbf{P}(\xi \geq \Delta + t_\delta) = (1 - \gamma_\delta)\delta,$$

where ξ is a standard normal random variable. Moreover, the indecision level is given by

$$\mathbf{P}(\xi \geq \Delta - t_\delta) - \mathbf{P}(\xi \geq \Delta + t_\delta) = \gamma_\delta.$$

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 parametrization: $\Delta = c\sqrt{2\log(1/\delta)}$ for $0 < c < 1$. We also use the following parametrization for γ:

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

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

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

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

Theorem 6. *For any $\varepsilon > 0$ and $c > 1/2$.*

(i) Let $m \leq m^*(c)$. Then, the oracle classifier defined before, 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 oracle classifier defined before achieves $\delta-$consistency.

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

Theorem 6 shows that $\delta-$consistency occurs if and only if

$$m(\gamma) \leq m^*(c), \tag{5}$$

for $1/2 < c < 1$, and

$$m(\gamma) \geq m^*(c), \tag{6}$$

for $0 < c < 1/2$.

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 $\delta^{(2c-1)^2} = o(1)$. Another phenomenon of interest here is an all-or-nothing phenomenon. By observing the asymptotic behavior of γ_δ, it seems that γ_δ either goes to 0 or 1, according to whether c is greater or smaller than 1/2. Asymptotically, the optimal behavior corresponds to either full indecisions or almost no indecisions.

7 Simulations

In the symmetric Gaussian mixture model, and given a vanishing misclassification level δ, we wish to compare the theoretical versus the empirical values of $\gamma(\delta)$. We recall the parameters c and m such that

$$\Delta(c) = c\sqrt{2\log(1/\delta)},$$

and

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

According to Theorem 6, and using the above parameterization, δ-consistency is possible whenever

$$\begin{array}{ll} m \geq m^*(c) & \text{if } 0 < c < 1/2, \\ m \leq m^*(c) & \text{if } 1/2 < c < 1. \end{array}$$

Based on the proof of Theorem 6, it also holds that δ-consistency is impossible whenever

$$\begin{array}{ll} m \leq m_*(c) & \text{if } 0 < c < 1/2, \\ m \geq m_*(c) & \text{if } 1/2 < c < 1, \end{array}$$

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

$$m_*(c) = \begin{cases} (c-(1-\varepsilon)/(4c))^2 & \text{if } 0 < c < 1/2, \\ (2c-1+\varepsilon)^2 & \text{if } 1/2 < c < 1, \end{cases} \tag{7}$$

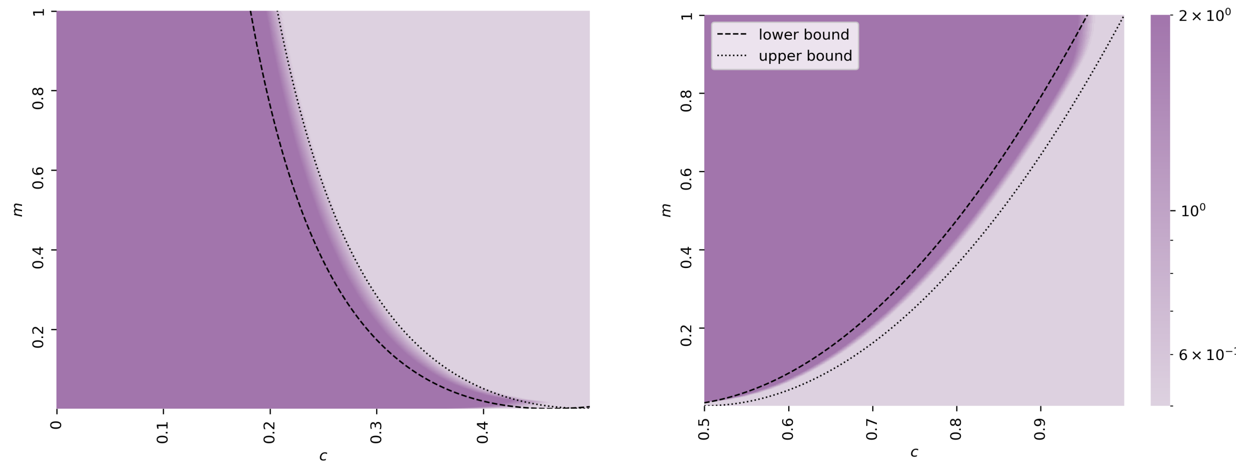

Figure 4: Computation of $\mathcal{R}(\gamma)/\delta$ with $\delta = 10^{-7}$ (left) and $\delta = 10^{-15}$ (right). The lower bound corresponds to the curve $m_*(c)$ while the upper bound corresponds to $m^*(c)$.

for $\varepsilon = 1/2 \log(4\pi \log(1/\gamma))/\log(1/\gamma)$. Observe that we get 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

$$\mathbf{P}(\xi \geq \Delta(c) - t_\gamma) - \mathbf{P}(\xi \geq \Delta(c) + t_\gamma) = \gamma(m),$$

where ξ is a standard normal. Next, we compute

$$\mathcal{R}(\gamma) = \frac{\mathbf{P}(\xi \geq \Delta(c) + t_\gamma)}{1 - \gamma(m)}.$$

To ensure better interpretability, values of $\mathcal{R}(\gamma)/\delta$ outside the range $(0.5, 2)$ were capped at this range in Figure 4. 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^*.

Funding

The work of Mohamed Ndaoud was supported by a Chair of Excellence in Data Science granted by the CY Initiative.

References

Bartlett, P. L. and M. H. Wegkamp (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research 9*(59), 1823–1840.

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.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory 16*(1), 41–46.

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

Dubuisson, B. and M. Masson (1993). A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition 26*(1), 155–165.

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

Fumera, G. and F. Roli (2004). Analysis of error-reject trade-off in linearly combined multiple classifiers. *Pattern Recognition 37*(6), 1245–1265.

Fumera, G., F. Roli, and G. Giacinto (2000). Reject option with multiple thresholds. *Pattern Recognition 33*(12), 2099–2101.

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.

Minsker, S., M. Ndaoud, and Y. Shen (2021). Minimax supervised clustering in the anisotropic gaussian mixture model: a new take on robust interpolation. *arXiv preprint arXiv:2111.07041*.

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

Ni, C., N. Charoenphakdee, J. Honda, and M. Sugiyama (2019). On the calibration of multiclass classification with rejection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Pillai, I., G. Fumera, and F. Roli (2013). Multi-label classification with a reject option. *Pattern Recognition 46*(8), 2256–2266.

Rava, B., W. Sun, G. M. James, and X. Tong (2024). A burden shared is a burden halved: A fairness-adjusted approach to classification. *In: arXiv preprint arXiv:2110.05720 (2024).*

Wiener, Y. and R. El-Yaniv (2011). Agnostic selective classification. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 24. Curran Associates, Inc.

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

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

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

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. $\square$

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 $\tilde{Y} = \operatorname{sign}(p_1 f_1 \geq p_2 f_2)$. 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

$$\begin{aligned}\mathcal{R}(\beta_2) &= \frac{\int_{A^{*c}_{\beta_2}}(p_1f_1 \wedge p_2f_2)}{1-\beta_2}\\ &= \frac{\int_{A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2)}{1-\beta_2} + \frac{\int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2)}{1-\beta_2}\\ &\geq \frac{1}{1-\beta_2}\left(\int_{A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2) + \frac{\int_{A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2)}{1-\beta_1}\int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}}(p_1f_1 + p_2f_2)\right),\end{aligned}$$

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

$$\frac{(p_1f_1 \wedge p_2f_2)}{(p_1f_1 + p_2f_2)} \geq (1-\tau_{\beta_1}),$$

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

$$\frac{(p_1f_1 \wedge p_2f_2)}{(p_1f_1 + p_2f_2)} \leq (1-\tau_{\beta_1}).$$

As a consequence we have that

$$\frac{(p_1f_1 \wedge p_2f_2)}{(p_1f_1 + p_2f_2)} \geq \frac{\int_{A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2)}{1-\beta_1}$$

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

$$\mathcal{R}(\beta_2) \geq \frac{\int_{A^{*c}_{\beta_1}}(p_1f_1 \wedge p_{-1}f_{-1})}{1-\beta_2}\frac{1-\beta_2}{1-\beta_1} \geq \mathcal{R}(\beta_1).$$

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

On the other hand, we have that

$$\mathcal{R}(\beta_2) - \mathcal{R}(\beta_1) = \frac{(\beta_2-\beta_1)\int_{A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2)}{(1-\beta_2)(1-\beta_1)} + \frac{\int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2)}{1-\beta_2}.$$

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

$$p_1f_1 \wedge p_2f_2 \leq (p_1f_1 + p_2f_2)(1-\tau_{\beta_2}).$$

Hence

$$\int_{A^{*c}_{\beta_2}/A^{*c}_{\beta_1}}(p_1f_1 \wedge p_2f_2) \leq \frac{(1-\tau_{\beta_2})(\beta_2-\beta_1)}{1-\beta_2}.$$

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 $p_1\mathbf{P}_1(\tilde{Y}=2)=\alpha(1-\gamma)$. Let A be the set where $\tilde{Y}=0$ and let B be the set where $\tilde{Y}=2$. We have that

$$\begin{aligned}\frac{p_2\mathbf{P}_2(\tilde{Y}=1)}{1-\gamma} &= \frac{\int_{B^c\cap A^c} p_2 f_2}{1-\gamma}\\ &= \frac{\int_{A^c} p_2 f_2}{1-\gamma} - \frac{\int_{B} p_2 f_2}{1-\gamma}.\end{aligned}$$

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

$$B^* = \left\{x \;/\; p_2 f_2 > \tau^u_{\alpha,\gamma} p_1 f_1\right\} \cup \mathcal{M}_{\alpha,\gamma}.$$

It comes out that

$$\frac{p_2\mathbf{P}_2(\tilde{Y}=1)}{1-\gamma} \geq \frac{\int_{A^c} p_2 f_2 \mathbf{1}(B^{*c})}{1-\gamma}.$$

Hence using Lemma 3 again we get

$$A^* := \left\{x \;;\; p_1 f_1/\tau^l_{\alpha,\gamma} < p_2 f_2 \leq \tau^u_{\alpha,\gamma} p_1 f_1\right\} \backslash \mathcal{M}^c_{\alpha,\gamma} \cup \mathcal{A}_{\alpha,\gamma},$$

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

A.4 Proofs for Section 3

A.4.1 Proof of Lemma 2

Note that

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

Also note that

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

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

$$\begin{aligned}&E\big(\mathbf{1}\{\widehat{Y}_\gamma = 2, Y^*_\gamma \neq 2\} - \mathbf{1}\{Y^*_\gamma = 2, \widehat{Y}_\gamma \neq 2\}\big)\\ &\qquad + E\big(\mathbf{1}\{\widehat{Y}_\gamma = 1, Y^*_\gamma \neq 1\} - \mathbf{1}\{Y^*_\gamma = 1, \widehat{Y}_\gamma \neq 1\}\big) = 0. \qquad (A.2)\end{aligned}$$

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

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

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:

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

which completes the proof. □

A.4.2 Proof of Theorem 3

Define event A_ϕ as follows:

$$A_\phi = \{\tau_\gamma^* < \eta \le \widehat{\tau}_\gamma - \phi\} \cup \{\tau_\gamma^* < 1-\eta \le \widehat{\tau}_\gamma - \phi\} \cup \{\widehat{\tau}_\gamma + \phi < \eta \le \tau_\gamma^*\} \cup \{\widehat{\tau}_\gamma + \phi < 1-\eta \le \tau_\gamma^*\}.$$

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

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

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

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:

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

where

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

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

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

We also have

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

where we used Lemma 4 and condition (2) to derive the final bound. Thus, we get the desired bound for the first term in our resulting expression for $(1-\gamma)[R(\widehat{Y}_\gamma) - R(Y_\gamma^*)]$:

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

The other three terms can be similarly bounded using analogous arguments. This completes the proof of claim (3) in Theorem 3. □

A.4.3 Proof of Theorem 4

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*

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

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

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

We take $\epsilon = c_1\phi^{1+2\alpha'-2\alpha}$ and note that we can choose c_1 sufficiently small to ensure that the second term in the line above is zero (recall that $\alpha' \ge \alpha$). This completes the proof of the first bound in Theorem 4. The second bound in Theorem 4 follows from the first bound together with condition (2). □

A.4.4 Proof of Lemma 4

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

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

We also have

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

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

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

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: $\widehat{\tau}_\gamma \le \tau_\gamma^* - \phi$. Note that

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

We also have

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

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

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

By condition (2), we also have

$$\mathbf{P}\big(1 - \tau_\gamma^* - \phi \le \eta \le \tau_\gamma^* + \phi\big) \ge \gamma + c\phi^{\alpha'}, \tag{A.12}$$

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

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

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

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

Condition (2) implies that $\gamma_\phi - \gamma \gtrsim \phi^{\alpha'}$ and $\mathbf{E}\big(\eta_{\max} \,\big|\, \tau_\gamma^* < \eta_{\max} \le \tau_\gamma^* + \phi\big) \le \tau_\gamma^* + \phi^{1+\alpha'-\alpha}$. Indeed, using condition (2) once again, we have

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

Consequently, $R(Y^{*\tau^*_\gamma}) - R(Y^{*\tau^*_\gamma+\phi}) \gtrsim \phi^{1+2\alpha'-\alpha}$, 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

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

Thus, using $\epsilon \leq \phi$ together with condition (2) we arrive at the second bound of Lemma 5. □

A.5 Proof of Theorem 5

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

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

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

A.6 Proof of Theorem 6

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.