Optimal Binary Classification Beyond Accuracy

Shashank Singh
Max Planck Institute for Intelligent Systems
Tübingen, Germany
shashankssingh44@gmail.com

Justin Khim*
Amazon
New York, NY
jkhim@amazon.com

Abstract

The vast majority of statistical theory on binary classification characterizes performance in terms of accuracy. However, accuracy is known in many cases to poorly reflect the practical consequences of classification error, most famously in imbalanced binary classification, where data are dominated by samples from one of two classes. The first part of this paper derives a novel generalization of the Bayes-optimal classifier from accuracy to any performance metric computed from the confusion matrix. Specifically, this result (a) demonstrates that stochastic classifiers sometimes outperform the best possible deterministic classifier and (b) removes an empirically unverifiable absolute continuity assumption that is poorly understood but pervades existing results. We then demonstrate how to use this generalized Bayes classifier to obtain regret bounds in terms of the error of estimating regression functions under uniform loss. Finally, we use these results to develop some of the first finite-sample statistical guarantees specific to imbalanced binary classification. Specifically, we demonstrate that optimal classification performance depends on properties of class imbalance, such as a novel notion called Uniform Class Imbalance, that have not previously been formalized. We further illustrate these contributions numerically in the case of k-nearest neighbor classification.

1 Introduction

Many binary classification problems exhibit class imbalance, in which one of the two classes vastly outnumbers the other. Classifiers that perform well with balanced classes routinely fail for imbalanced classes, and developing reliable techniques for classification in the presence of severe class imbalance remains a challenging area of research [He and Ma, 2013, Krawczyk, 2016, Fernández et al., 2018]. Many practical approaches have been proposed to improve performance under class imbalance, including reweighting plug-in estimates of class probabilities [Lewis, 1995], resampling data to improve class imbalance [Chawla et al., 2002], or reformulating classification algorithms to optimize different performance metrics [Dembczynski et al., 2013, Fathony and Kolter, 2019, Joachims, 2005]. Extensive discussion of practical methods for handling class imbalance are surveyed in the books of He and Ma [2013] and Fernández et al. [2018].

Despite the pervasive challenge of class imbalance, our theoretical understanding of class imbalance is limited. The vast majority of theoretical performance guarantees for classification characterize classification accuracy or, equivalently, misclassification risk [Mohri et al., 2018], which is typically an uninformative measure of performance for imbalanced classes. Under measures that are used with imbalanced classes in practice, such as precision, recall, F_β scores, and class-weighted scores [Van Rijsbergen, 1974, 1979], existing theoretical guarantees are limited to statistical consistency, in that the algorithm under consideration asymptotically optimizes the metric of choice [Koyejo et al., 2014, Menon et al., 2013, Narasimhan et al., 2014]; specifically, there is no finite-sample theory that would

*The contributions in this paper were made prior to joining Amazon.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

https://doi.org/10.52202/068431-1325

allow comparison of an algorithm's performance to that of other algorithms or to theoretically optimal performance levels. Additionally, existing theory for classification does not explicitly model the effects of class imbalance, especially severe imbalance (i.e., as the proportion of samples from the rare class vanishes), and hence sheds little light on how severe imbalance influences optimal classification.

This paper provides two main contributions. First, in Section 4, we provide a novel characterization of classifiers optimizing general performance metrics that are functions of a classifier's confusion matrix. This characterization generalizes a classical result, that the Bayes classifier optimizes classification accuracy, to a much larger class of performance measures, including those commonly used in imbalanced classification, while relaxing certain empirically unverifiable distributional assumptions that pervade existing such results. Interestingly, we show that, in general, a Bayes classifier always exists if one considers stochastic classifiers, but not if one considers only deterministic classifiers. We then use this result to provide relative performance guarantees under these more general performance measures, in terms of the error of estimating the class probability (regression) function under uniform ($\mathcal{L}_\infty$) loss.

This motivates our second main contribution: an analysis of k-nearest neighbor (kNN) classification under uniform loss. In doing so, we also propose an explicit model of a sub-type of class imbalance, which we call Uniform Class Imbalance, and we show that the kNN classifier behaves quite differently under Uniform Class Imbalance than under other sub-types of class imbalance. To the best of our knowledge, such sub-types of class imbalance have not previously been distinguished in either the theoretical or practical literature, and we hope that identifying such relevant features of imbalanced datasets may facilitate development of classifiers that perform well on specific imbalance problems of practical importance. Collectively, these contributions provide some of the first finite-sample performance guarantees for nonparametric binary classification under performance metrics that are appropriate for imbalanced data and show how optimal performance depends on the nature of imbalance in the data.

2 Related Work

Here, we discuss how our results relate to existing theoretical guarantees for imbalanced binary classification and prior analyses of kNN methods.

2.1 Theoretical Guarantees for Imbalanced Binary Classification

Statistical learning theory has studied classification extensively in terms of accuracy [Mohri et al., 2018]. However, when classes are severely imbalanced, accuracy ceases to be an informative measure of performance [Cortes and Mohri, 2004], necessitating guarantees in terms of other performance metrics. Several papers have sought to address this [Narasimhan et al., 2014, 2015, Koyejo et al., 2014, Yan et al., 2018, Wang et al., 2019a] by generalizing the Bayes optimal classifier, a well-known classifier that provably optimizes accuracy, to more general performance measures better reflecting the desiderata of imbalanced classification. Relatedly, several works have investigated relationships between these different performance measures and demonstrated that they differ essentially in how they determine the optimal threshold between the two classes [Flach, 2003, Hernández-Orallo et al., 2013, Flach, 2016]. However, existing results make empirically unverifiable assumptions about the distribution of the data, leaving questions about their relevance to real data. We discuss these assumptions in detail in Section 4, where our main result, Theorem 3, leverages the idea of stochastic thresholding to relax these assumptions.

Another body of closely related theoretical work studies Neyman-Pearson classification, which attempts to minimize misclassification error on one class subject to constraints on misclassification error on other classes, analogous to the approach of statistical hypothesis testing. While substantial theoretical guarantees do exist for Neyman-Pearson classification [Rigollet and Tong, 2011, Tong, 2013, Tong et al., 2016], these focus on performance within the Neyman-Pearson framework, rather than under general performance measures as in our work, and we know of no work considering stochastic classification under the Neyman-Pearson framework. Interestingly, our use of stochastic classifiers in Theorem 3 parallels classical results in hypothesis testing [Lehmann and Romano, 2006], and our proof of Theorem 3 involves a reduction (Lemma 22 in the Appendix) of optimization of general classification performance measures to Neyman-Pearson classification.

https://doi.org/10.52202/068431-1325

Meanwhile, many practical approaches to handling class imbalance, such as class-weighting and resampling have been proposed, but the theoretical understanding of these methods is limited. Class-weighting is a natural choice in applications where costs, or cost ratios [Flach, 2003], can be explicitly assigned and, in the case of binary classification, is statistically equivalent to threshold selection, which we discuss later in this paper [Scott, 2012]. In practice, resampling appears to be the most popular approach to handling class imbalance [He and Ma, 2013]. Undersampling the dominant class is straightforward and can provide computational benefits with little loss in statistical performance [Fithian and Hastie, 2014], while interest in oversampling rare classes, sometimes referred to as data augmentation, has grown with the advent of sophisticated generative models to produce additional data [Mariani et al., 2018]. However, the theoretical ramifications of oversampling techniques used for imbalanced classification, most commonly variants of SMOTE [Chawla et al., 2002], are poorly understood.

2.2 kNN Classification and Regression

The kNN classifier is one of the oldest and most well-studied nonparametric classifiers Fix and Hodges [1951]. Early theoretical results include, Cover and Hart [1967], who showed that the misclassification risk of the kNN classifier with $k = 1$ is at most twice that of the Bayes-optimal classifier, and Stone [1977], who showed that the kNN classifier is Bayes-consistent if $k \to \infty$ and $k/n \to 0$. Extensive literature on the accuracy of kNN classification has since developed [Devroye et al., 1996, Györfi et al., 2002, Samworth, 2012, Chaudhuri and Dasgupta, 2014, Gottlieb et al., 2014, Biau and Devroye, 2015, Gadat et al., 2016, Döring et al., 2018, Kontorovich and Weiss, 2015, Gottlieb et al., 2018, Cannings et al., 2019, Hanneke et al., 2020].

Rather than accuracy bounds for kNN classification, the bounds on uniform error we present in Section 5 are most closely related to risk bounds for kNN regression, of which the results of Biau et al. [2010] are representative. Biau et al. [2010] gives convergence rates for kNN regression in $\mathcal{L}_2$ risk, weighted by the covariate distribution, in terms of noise variance and covering numbers of the covariate space. While closely related to our bounds on uniform ($\mathcal{L}_\infty$) risk, their results differ in at least three main ways. First, minimax rates under $\mathcal{L}_\infty$ risk are necessarily worse than under $\mathcal{L}_2$ risk by a logarithmic factor, as implied by our lower bounds. Second, the fact that Biau et al. [2010] use a risk that is weighted by the covariate distribution allows them to avoid our assumption that the covariate density is lower bounded away from 0, whereas, the lower boundedness assumption is unavoidable under $\mathcal{L}_\infty$ risk. Finally, Biau et al. [2010] assume additive noise with finite variance; Bernoulli noise is crucial for us to model severe class imbalance.

Extensive research on kNN for imbalanced classification has focused on algorithmic modifications, which are surveyed by Fernández et al. [2018]. Examples include prototype selection [Liu and Chawla, 2011, López et al., 2014, Vluymans et al., 2016], and gravitational methods [Cano et al., 2013, Zhu et al., 2015]. We are aware of no statistical guarantees exist for such methods.

3 Setup and Notation

Let $(\mathcal{X}, \rho)$ be a separable metric space, and let $\mathcal{Y} = \{0, 1\}$ denote the set of classes. For any $x \in \mathcal{X}$ and $\epsilon > 0$, $B(x, \epsilon) := \{z \in \mathcal{X} : \rho(x, z) < \epsilon\}$ denotes the open radius-ϵ ball around x. Consider n independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from a distribution $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X and P_Y. For positive sequences $\{a_n\}_{i=1}^\infty$ and $\{b_n\}_{i=1}^\infty$, $a_n \asymp b_n$ means $\liminf_{n\to\infty} a_n/b_n > 0$ and $\limsup_{n\to\infty} a_n/b_n < \infty$.

To optimize general performance metrics, we must consider stochastic classifiers. Formally, letting $\mathcal{B} := \{Y \sim \text{Bernoulli}(p) : p \in [0, 1]\}$ denote the set of binary random variables, a stochastic classifier can be modeled as a mapping $\widehat{Y} : \mathcal{X} \to \mathcal{B}$, where, for any $x \in \mathcal{X}$, $\mathbb{E}[\widehat{Y}(x)]$ is the probability that the classifier assigns x to class 1. We use $\mathcal{SC}$ to denote the class of stochastic classifiers.

The true *regression function* $\eta^* : \mathcal{X} \to [0, 1]$ is defined as $\eta^*(x) := \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$; that is, given an instance X_i, the label Y_i is Bernoulli-distributed with mean $\eta(X_i)$. As we show in the next section, an optimal classifier can always be written in terms of the true regression function η, motivating estimates $\widehat{\eta} : \mathcal{X} \to [0, 1]$ of η. Such estimates $\widehat{\eta}$ are referred to as "regressors".

https://doi.org/10.52202/068431-1325

4 Optimal Classification Beyond Accuracy

A famous result states that classification accuracy is maximized by the "Bayes" classifier

$$\widehat{Y}(x) \sim \text{Bernoulli}\left(1\{\eta^*(x) > 0.5\}\right). \tag{1}$$

Here, $\widehat{Y}(x)$ is simply a constant (deterministic) random variable that takes either the value 0 or the value 1 with probability 1 (depending on $\eta^*(x)$). Our reason for writing Eq. (1) in this seemingly redundant way will become clear with Definition 2 below.

Although η^* is unknown in practice, this result is a cornerstone of the statistical theory of binary classification because it provides an optimal performance benchmark against which a classifier can be evaluated in terms of accuracy [Devroye et al., 1996, Mitchell, 1997, James et al., 2013]. As discussed previously, accuracy can be a poor measure of performance in the imbalanced case. Therefore, the main contribution of this section, provided in Theorem 3 below, is to generalize this result to a broad class of classification performance measures, including those commonly used in imbalanced classification. First, we specify performance measures for which our results apply.

4.1 Confusion Matrix Measures (CMMs)

Nearly all measures of classification performance, including accuracy, precision, recall, F_β scores, and others, can be computed from the confusion matrix, which counts the number of test samples in each (true class, estimated class) pair. Formally, let $\mathcal{C} := \{C \in [0,1]^{2\times 2} : C_{1,1} + C_{1,2} + C_{2,1} + C_{2,2} = 1\}$ denote the set of all possible binary confusion matrices. Given a classifier $\widehat{Y}$, the *confusion matrix* $C_{\widehat{Y}} \in \mathcal{C}$ and *empirical confusion matrix* $\widehat{C}_{\widehat{Y}} \in \mathcal{C}$ are given by

$$C_{\widehat{Y}} = \begin{bmatrix} \text{TN}_{\widehat{Y}} & \text{FP}_{\widehat{Y}} \\ \text{FN}_{\widehat{Y}} & \text{TP}_{\widehat{Y}} \end{bmatrix}, \widehat{C}_{\widehat{Y}} = \begin{bmatrix} \widehat{\text{TN}}_{\widehat{Y}} & \widehat{\text{FP}}_{\widehat{Y}} \\ \widehat{\text{FN}}_{\widehat{Y}} & \widehat{\text{TP}}_{\widehat{Y}} \end{bmatrix}, \tag{2}$$

wherein the true positive probability $\text{TP}_{\widehat{Y}}$ and empirical true positive probability $\widehat{\text{TP}}_{\widehat{Y}}$ are given by

$$\text{TP}_{\widehat{Y}} = \mathbb{E}\left[\eta^*(X)\widehat{Y}(X)\right], \widehat{\text{TP}}_{\widehat{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i \widehat{Y}(X_i), \tag{3}$$

and the true and empirical false positive ($\text{FP}_{\widehat{Y}}$ and $\widehat{\text{FP}}_{\widehat{Y}}$), false negative ($\text{FN}_{\widehat{Y}}$ and $\widehat{\text{FN}}_{\widehat{Y}}$), and true positive ($\text{TP}_{\widehat{Y}}$ and $\widehat{\text{TP}}_{\widehat{Y}}$) probabilities are defined similarly. Note that the expectation in Eq. (3) is over randomness both in the data and in the classifier.

Intuitively, measures of a classifier's performance should improve as TN and TP increase and FN and FP decrease. We therefore define the class of Confusion Matrix Measures (CMMs) as follows:

Definition 1 (Confusion Matrix Measure (CMM))**.** *A function $M : \mathcal{C} \to \mathbb{R}$ is called a* confusion matrix measure (CMM) *if, for any confusion matrix*

$$C = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \in \mathcal{C}, \epsilon_1 \in [0, FP], \epsilon_2 \in [0, FN], \quad \textit{we have} \quad M(C) \leqslant M\left(\begin{bmatrix} TN + \epsilon_1 & FP - \epsilon_1 \\ FN - \epsilon_2 & TP + \epsilon_2 \end{bmatrix}\right).$$

Essentially, correcting an incorrect classification should not reduce a CMM. This is true of any reasonable measure of classification performance, and hence analyzing CMMs allows us to obtain theoretical guarantees for all performance measures used in practice. Specifically, by evaluating their gradients in the directions $\left[\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 0 & 0 \\ -1 & 1 \end{smallmatrix}\right]$, one can verify that most performance measures, such as weighted accuracy, precision, recall, F_β scores, and Matthew's Correlation Coefficient are CMMs. We note that the area under receiver operating characteristic (AUROC) and area under precision-recall curve (AUPRC) are not CMMs because they evaluate ($\mathbb{R}$-valued) scoring functions rather than ($\{0,1\}$-valued) classification functions. However, both AUROC and AUPR are averages of CMMs computed at various classification thresholds, and, as we discuss in Appendix A.1, our results for CMMs thus imply similar results for these measures. We next present our main result of Section 4, which generalizes the Bayes classifier (1) to arbitrary CMMs.

https://doi.org/10.52202/068431-1325

4.2 Generalizing the Bayes Classifier

The Bayes classifier thresholds the regression function deterministically at the value 0.5. The following generalizes this to a stochastic threshold:

Definition 2 (Regression-Thresholding Classifier (RTC))**.** *A classifier $\widehat{Y} : \mathcal{X} \to \mathcal{B}$ is called a* regression-thresholding classifier *(RTC) if, for some $p, t \in [0, 1]$ and $\eta : \mathcal{X} \to [0, 1]$,*

$$\widehat{Y}(x) \sim \textit{Bernoulli}(p \cdot 1\{\eta(x) = t\} + 1\{\eta(x) > t\}), \quad \textit{for all } x \in \mathcal{X}.$$

In the sequel, we will denote such classifiers $\widehat{Y}_{p,t,\eta}$, and refer to the pair (t, p) as the threshold*.*

Now we can state the main result of this paper:

Theorem 3. *For any CMM M and stochastic classifier $\widehat{Y}$, there is an RTC $\widehat{Y}_{p,t,\eta}$ with $M(\widehat{Y}_{p,t,\eta}) \geqslant M(\widehat{Y})$. In particular, if M is maximized by any stochastic classifier, then M is maximized by a RTC.*

As a special case of Theorem 3, the classical Bayes classifier corresponds to $M(C) = \text{TN} + \text{TP}$, $p = 0$, and $t = 0.5$. However, as discussed in the next paragraph, without stronger assumptions, Theorem 3 does *not* hold for deterministic classifiers. Since RTCs generalize both the RTC structure and optimality properties of the Bayes classifier, we also refer to them as *generalized Bayes classifiers*. We note that *existence* of any maximizer $\widehat{Y}$ of $M(C_{\widehat{Y}})$ may depend on specific properties, such as (semi)continuity or convexity of M, which we do not investigate here.

We emphasize that Theorem 3 makes absolutely no assumptions on the distribution of the data. In particular, all prior characterizations of optimal classifiers under general performance metrics assume that the distribution of the class probability $\eta(X)$ is absolutely continuous [Narasimhan et al., 2014, 2015, Koyejo et al., 2014, Yan et al., 2018, Wang et al., 2019a][2], and Wang et al. [2019a] claim that regularity assumptions on $\eta(X)$ such as absolute continuity "seem to be unavoidable". Our Theorem 3 is the first result to omit such assumptions, and we specifically show that this comes at the cost of the optimal classifier possibly being non-deterministic for a single atom of $\eta(X)$. Figure 1 visually compares the stochastic thresholding classifier in Theorem 3 to prior approaches.

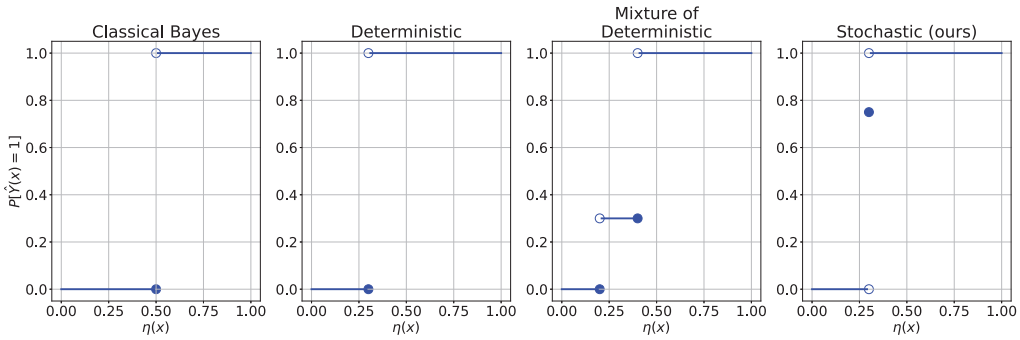


Figure 1: Examples of four different approaches to thresholding the regression function. Classical Bayes thresholding (Eq. (1)) always thresholds deterministically at $\eta(x) = 0.5$ to optimize accuracy. Koyejo et al. [2014], Narasimhan et al. [2014] and others have suggested using other Deterministic thresholds (e.g., $\eta(x) = 0.3$, shown here) to optimize other CMMs, assuming $\eta(X)$ is absolutely continuous. Wang et al. [2019a] showed that the optimal classifier can always be written as a Mixture of Deterministic (MD) classifiers (e.g., a $(0.3, 0.7)$-mixture of thresholds at $\eta(x) = 0.2$ and $\eta(x) = 0.4$, shown here). Finally, we propose using a single Stochastic threshold (e.g., $(t, p) = (0.3, 0.75)$, shown here). Only MD and Stochastic approaches are optimal in general (for arbitrary CMMs, without $\eta(X)$ absolutely continuous), while Stochastic thresholding is strictly simpler than MD.

The generality of Theorem 3 necessitates a significantly more complex proof than prior work. In particular, we prove Theorem 3 in Appendix A using a series of variational arguments. Roughly speaking, given a classifier $\widehat{Y}$, we construct a perturbation $\widehat{Y}'$ of $\widehat{Y}$ such that either $M(C_{\widehat{Y}}) <$

[2]Exceptions for the case of F_1 score are Zhao et al. [2013, Lemma 12] and Lipton et al. [2014, Theorem 1].

https://doi.org/10.52202/068431-1325

$M(C_{\widehat{Y}'})$ or $\widehat{Y}'$ is an RTC and $M(C_{\widehat{Y}}) \leqslant M(C_{\widehat{Y}'})$. Since, the classifier $\widehat{Y}$ might be quite poorly behaved (e.g., its behavior on sets of P_X-measure 0 could be arbitrary), the technical complexity lies in constructing admissible perturbations (i.e., those that are well-defined classifiers). For this reason, the proof of Theorem 3 involves a series of constructions of increasingly well-behaved classifiers.

Theorem 3 tells us that a generalized Bayes classifier can always be written in terms of the regression function η and two scalar parameters (t, p) depending on the distribution of $\eta(X)$ and the CMM M. The next example shows that this characterization cannot be simplified without stronger assumptions:

Example 4. Suppose $\mathcal{X} = \{0\}$ is a singleton, $\eta(0) \in (0, 1)$, and, for some $\theta > 0$, $M(C) = (\text{TP})^\theta\text{TN}$. One can check that M is a valid CMM. Suppose $\widehat{Y}$ is an RTC. It is straightforward to compute that $M(C_{\widehat{Y}}) = (p\eta(0))^\theta(1-p)(1-\eta(0))1\{t = \eta(0)\}$, and that $M(C_{\widehat{Y}})$ is uniquely maximized by $p = \frac{\theta}{\theta+1} \in (0, 1)$ and $t = \eta(0) \in (0, 1)$. This shows that both threshold parameters p and t in an RTC are necessary, in the absence of further assumptions on M or η. This example also illustrates the need for stochasticity to optimize general CMMs. Specifically, for any deterministic classifier $\widehat{Y}$, either $\widehat{Y}(0) = 0$ (so TP $= 0$) or $\widehat{Y}(0) = 1$ (so TN $= 0$); in either case, $M(C_{\widehat{Y}}) = 0$.

This performance gap between stochastic and deterministic classifiers is closely related to Theorem 1 of Cotter et al. [2019b], which provides a closely related lower bound on how well a stochastic classifier can be approximated by a deterministic one, in terms of the probability assigned to atoms of $\eta(X)$. However, Cotter et al. [2019b] only study how well stochastic classifiers can be approximated by deterministic ones (with the motivation of derandomizing classifiers), not whether stochastic classifiers can systematically outperform deterministic ones, as we show here.

4.3 Relative Performance Guarantees in terms of the Generalized Bayes Classifier

Theorem 3 motivates a two-step approach to imbalanced classification in which one first estimates the regression function η and then selects a stochastic threshold (t, p) that optimizes empirical performance $M(\widehat{C}_{\widehat{Y}})$. Such an approach has many practical advantages. For example, as we show in Appendix D, a simple algorithm can exactly optimize the threshold (t, p) over large datasets in $O(n \log n)$ time. Additionally, one can address covariate shift or retune a classifier trained under one CMM to perform well under another CMM, simply by re-optimizing (t, p), which is statistically and computationally much easier than retraining a classifier from scratch. In this section, we focus on an advantage for theoretical analysis, namely that the error of such a classifier decomposes into errors in selecting (t, p) and errors in estimating η, allowing the derivation of performance guarantees relative a generalized Bayes classifier. All results in this section are proven in Appendix B.

We first bound the performance difference of thresholding two regressors in terms of their $\mathcal{L}_\infty$ distance. This will allow us to bound error due to using a regressor $\widehat{\eta}$ instead of the true η.

Lemma 5. *For $p, t \in [0, 1]$, $\eta, \eta' : \mathcal{X} \to [0, 1]$, $\left\|C_{\widehat{Y}_{p,t,\eta}} - C_{\widehat{Y}_{p,t,\eta'}}\right\|_\infty \leqslant \mathbb{P}\left[|\eta(X) - t| \leqslant \|\eta - \eta'\|_\infty\right]$.*

Intuitively, Lemma 5 bounds the largest difference in the confusion matrices of $\widehat{Y}_{p,t,\eta}$ and $\widehat{Y}_{p,t,\eta'}$ by the probability that the threshold t lies between η and η'. As we will show later, under a margin assumption, this can be bounded by the $\mathcal{L}_\infty$ distance between η and η'.

Our next lemma bounds the worst-case error over thresholds $(t, p) \in [0, 1]$ of the empirical confusion matrix. This allows us to bound error due to using an empirical threshold $(\widehat{t}, \widehat{p})$ instead of the threshold (t^*, p^*) that is optimal for the true regression function.

Lemma 6. *Let $\eta : \mathcal{X} \to [0, 1]$ be any regression function. Then, with probability at least $1 - \delta$,*

$$\sup_{p,t\in[0,1]} \left\|\widehat{C}_{\widehat{Y}_{p,t,\eta}} - C_{\widehat{Y}_{p,t,\eta}}\right\|_\infty \leqslant \sqrt{\frac{8}{n} \log \frac{32(2n+1)}{\delta}}.$$

Lemma 6 follows from Vapnik-Chervonenkis (VC) bounds on the complexity of the set $\{\widehat{Y}_{p,t,\eta} : p, t \in [0, 1]\}$ of possible RTCs with fixed regression function η. In fact, Appendix B proves a more general bound on the error between empirical and true confusion matrices uniformly over any family $\mathcal{F}$ of stochastic classifiers in terms of the growth function of $\mathcal{F}$. Consequently, when $\mathcal{F}$ has finite VC dimension, we obtain uniform convergence at the fast rate $\sqrt{\log(n/\delta)/n}$. As we formalize later, this

https://doi.org/10.52202/068431-1325

suggests that the difficulty in tuning an imbalanced classifier to optimize a CMM M comes not from difficulty in estimating the confusion matrix but rather from the sensitivity of commonly used CMMs to the selected threshold. Because Theorem 3 shows that any CMM can be optimized by a RTC, we state here only the specific result for RTCs.

Before combining Lemmas 5 and 6 to give the main result of this section, we note a margin assumption, which characterizes separation between the two classes:

Definition 7 (Tsybakov Margin Condition). *Let $C, \beta \geqslant 0$, $t \in (0,1)$. A classification problem with covariate distribution P_X and regression function η satisfies a* (C, β)-margin condition around t *if, for any $\epsilon > 0$, $\mathbb{P}\left[|\eta(X) - t| \leqslant \epsilon\right] \leqslant C\epsilon^{\beta}$.*

The Tsybakov margin condition, introduced by Mammen and Tsybakov [1999] for $t = 0.5$, is widely used to establish convergence rates for classification in terms of accuracy [Audibert and Tsybakov, 2007, Arlot and Bartlett, 2011, Chaudhuri and Dasgupta, 2014]. Together with the margin condition and a Lipschitz condition on the M, Lemmas 5 and 6 give the following bound on sub-optimality of an RTC if the threshold is selected by maximizing M over the empirical confusion matrix:

Corollary 8. *Let $\eta : \mathcal{X} \to [0,1]$ be the true regression function and $\widehat{\eta} : \mathcal{X} \to [0,1]$ be any regressor.*

$$\textit{Let} \quad (\widehat{p}, \widehat{t}) := \underset{(t,p)\in[0,1]^2}{\operatorname{argmax}} M\left(\widehat{C}_{\widehat{Y}_{p,t,\widehat{\eta}}}\right) \quad \textit{and} \quad (p^*, t^*) := \underset{(t,p)\in[0,1]^2}{\operatorname{argmax}} M\left(C_{\widehat{Y}_{p,t,\eta}}\right)$$

denote the empirical and true optimal thresholds, respectively. Suppose M is Lipschitz continuous with constant L_M with respect to the uniform ($\mathcal{L}_\infty$) metric on $\mathcal{C}$. Finally, suppose P_X and η satisfy a (C, β)-margin condition around t^*. Then, with probability $\geqslant 1 - \delta$,

$$M\left(C_{\widehat{Y}_{p^*,t^*,\eta}}\right) - M\left(C_{\widehat{Y}_{\widehat{p},\widehat{t},\widehat{\eta}}}\right) \leqslant L_M\left(C\left\|\eta - \widehat{\eta}\right\|_\infty^\beta + 2\sqrt{\frac{8}{n}\log\frac{32(2n+1)}{\delta}}\right).$$

5 Uniform Error of the kNN Regressor

In the previous section, we bounded relative performance of an RTC in terms of uniform ($\mathcal{L}_\infty$) loss of the regression function estimate. Here, we bound uniform loss of one such regressor, the widely used k-nearest neighbor (kNN) regressor. Our analyses include a parameter r, introduced in Section 5, that characterizes a novel sub-type of class imbalance, which we call Uniform Class Imbalance. This leads to insights about how the behavior of the kNN classifier depends not only on the degree, but also on the structure, of class imbalance in a given dataset. We begin with some notation:

Definition 9 (k-Nearest Neighbor Regressor). *Given a point $x \in \mathcal{X}$, let $\sigma(x)$ denote a permutation of $\{1, ..., n\}$ such that $X_{\sigma_i(x)}$ is the i^{th}-nearest neighbor of x among $X_1, ..., X_n$. For integers $k \in [1, n]$, the kNN regressor $\widehat{\eta}_k : \mathcal{X} \to [0,1]$ is defined as*

$$\widehat{\eta}_k(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{\sigma_i(x)}, \quad \textit{for all } x \in \mathcal{X}. \tag{4}$$

We now formalize a novel sub-type of class imbalance:

Definition 10 (Uniform Class Imbalance (UCI)). *Write the regression function as $\eta = r\zeta$, where $r \in (0,1]$ and $\zeta : \mathcal{X} \to [0,1]$ is a regression function with $\sup_{x\in\mathcal{X}} \zeta(x) = 1$. A classification problem has Uniform Class Imbalance (UCI) in the number of samples n if $r \to 0$ as $n \to \infty$.*

Intuitively, in UCI, the class $Y = 1$ is rare regardless of X. This includes "difficult" classification problems where the covariate X provides only partial information about the class Y and examples from the rare class lie deep within the distribution of the common class. Examples include rare disease diagnosis [Schaefer et al., 2020] or fraud detection [Awoyemi et al., 2017]. In practice, the classifier's role is often to flag "high-risk" samples X, those with $\eta(X)$ relatively high, for follow-up investigation. UCI can be distinguished from "easier" problems in which, for some $x \in \mathcal{X}$, $\eta(X) \approx 1$ and so, given enough training data, a classifier can confidently assign the label $Y = 1$. These include well-separated classes or deterministic problems (e.g., protein structure prediction; Noé et al. [2020]).

To our knowledge, such notions of class imbalance have not previously been distinguished. In the particular case of data drawn from a logistic model, UCI reduces to the notion of class imbalance described in Wang [2020]; however, UCI applies in more general contexts.

https://doi.org/10.52202/068431-1325

5.1 Uniform Risk Bounds

We now present bounds (proven in Appendix C.1) on uniform error $\|\eta - \widehat{\eta}\|_\infty = \sup_{x \in \mathcal{X}} |\eta(x) - \widehat{\eta}(x)|$ of the kNN regressor $\widehat{\eta}_k$. First, recall two standard quantities, covering numbers and shattering coefficients, by which we measure complexity of the feature space:

Definition 11 (Covering Number). *Suppose $(\mathcal{X}, \rho)$ is a totally bounded metric space. Then, for any $\epsilon > 0$, the* ϵ-covering number $N(\epsilon)$ *of $(\mathcal{X}, \rho)$ is the smallest integer such that there exist $N(\epsilon)$ points $x_1, ..., x_{N(\epsilon)} \in X$ satisfying $\mathcal{X} \subseteq \bigcup_{i=1}^{N(\epsilon)} B(x_i, \epsilon)$.*

Definition 12 (Shattering Coefficient). *For integers $n > 0$, the* shattering coefficient of balls *in $(\mathcal{X}, \rho)$ is $S(n) = \sup_{x_1, ..., x_n \in \mathcal{X}} |\{\{x_1, ..., x_n\} \cap B(x, \epsilon) : x \in \mathcal{X}, \epsilon \geqslant 0\}|$.*

We now state two assumptions data distribution $P_{X,Y}$:

Assumption 13 (Dense Covariates Assumption). *For some $p_*, \epsilon^*, d > 0$, the marginal distribution P_X of covariates is lower bounded, for any $x \in \mathcal{X}$ and $\epsilon \in (0, \epsilon^*]$, by $P_X(B_\epsilon(x)) \geqslant p_* \epsilon^d$.*

Assumption 13 ensures that each query point's nearest neighbors are sufficiently near to be informative. We also assume that the regression function ζ is smooth:

Assumption 14 (Hölder Continuity). *For some $\alpha \in (0, 1]$, $L := \sup_{x \neq x' \in \mathcal{X}} \frac{|\zeta(x) - \zeta(x')|}{\rho^\alpha(x, x')} < \infty$.*

We now state our upper bound on uniform error:

Theorem 15. *Under Assumptions 13 and 14, whenever $k/n \leqslant p_*(\epsilon^*)^d/2$, for any $\delta > 0$, with probability at least $1 - N\left((2k/(p_* n))^{1/d}\right) e^{-k/4} - \delta$, we have*

$$\|\eta - \widehat{\eta}\|_\infty \leqslant 2^\alpha L r \left(\frac{2k}{p_* n}\right)^{\alpha/d} + \frac{2}{3k} \log \frac{2S(n)}{\delta} + \sqrt{\frac{2r}{k} \log \frac{2S(n)}{\delta}}. \tag{5}$$

If $r \in O\left((\log S(n))/n\right)$, this bound is minimized by $k \asymp n$, giving $\|\eta - \widehat{\eta}\|_\infty \in O_P\left((\log S(n))/n\right)$. Otherwise, this bound is minimized by $k \asymp n^{\frac{2\alpha}{2\alpha+d}} (\log S(n))^{\frac{d}{2\alpha+d}} r^{-\frac{d}{2\alpha+d}}$, giving

$$\|\eta - \widehat{\eta}\|_\infty \in O_P\left((\log S(n))/n)^{\frac{\alpha}{2\alpha+d}} r^{\frac{\alpha+d}{2\alpha+d}}\right).$$

Of the three terms in (5), the first term, of order $r(k/n)^{\alpha/d}$, comes from smoothing bias of the kNN classifier. The second and third terms are due to label noise, with the second term dominating under extreme class imbalance $r \in O\left(\log S(n)/n\right)$ and the third term dominating otherwise. Theorem 5 shows that, under UCI, one should use a much larger choice of the tuning parameter k than in the case of balanced classes; indeed, setting $k \asymp n^{\frac{2\alpha}{2\alpha+d}} (\log S(n))^{\frac{d}{2\alpha+d}}$, which is optimal in the balanced case, gives a rate that is suboptimal by a factor of $r^{-d/(4\alpha+d)}$.

The following example demonstrates how to apply Theorem 15 in a concrete setting of interest:

Corollary 16 (Euclidean, Absolutely Continuous Case). *Suppose $(\mathcal{X}, \rho) = ([0, 1]^d, \|\cdot\|_2)$ is the unit cube in $\mathbb{R}^d$, equipped with the Euclidean metric, and P_X has a density that is lower bounded away from 0 on $\mathcal{X}$. Then, for $k \asymp n^{\frac{2\alpha}{2\alpha+d}} (\log n)^{\frac{d}{2\alpha+d}} r^{-\frac{d}{2\alpha+d}}$, $\|\eta - \widehat{\eta}\|_\infty \in O_P\left((\log n)/n)^{\frac{\alpha}{2\alpha+d}} r^{\frac{\alpha+d}{2\alpha+d}}\right)$.*

The most problematic term in this bound is the exponential dependence on the dimension d of the covariates. Fortunately, since Theorem 15 utilizes covering numbers, it improves if the covariates exhibit structure, such as that of a low-dimensional manifold. We illustrate this in detail in Appendix C.1.

We close with a minimax lower bound, proven in Appendix C.2, on the uniform error of any estimator, over (α, L)-Hölder regression functions. Up to a polylogarithmic factor in r, the rate of this lower bound matches that in Theorem 15, suggesting that both bounds are quite tight.

Theorem 17. *Suppose $\mathcal{X} = [0, 1]^d$ is the d-dimensional unit cube and $X \sim \mathrm{Uniform}(\mathcal{X})$. Let $\Sigma^\alpha(L)$ denote the family of (α, L)-Hölder continuous regression function. Then, there exist constants*

https://doi.org/10.52202/068431-1325

n_0 *and* $c > 0$ *(depending only on* α*,* L*, and* d*) such that, for all* $n \geqslant n_0$ *and any estimator* $\widehat{\eta}$*,*

$$\sup_{\zeta \in \Sigma^\alpha(L)} \mathbb{P}\left[\|\eta - \widehat{\eta}\|_\infty \geqslant c\left(\frac{\log(nr)}{n}\right)^{\frac{\alpha}{2\alpha+d}} r^{\frac{\alpha+d}{2\alpha+d}}\right] \geqslant \frac{1}{8}.$$

Discussion Plugging the above upper bounds on $\|\eta - \widehat{\eta}\|_\infty$ into Corollary 8 provides error bound under arbitrary CMMs, in terms of the sample size n, hyperparameter k, UCI degree r, and complexity parameters (margin β, smoothness α, intrinsic dimension d, etc.) of $\mathcal{X}$ and $P_{X,Y}$. Thus, these results collectively give some of the first complete finite-sample guarantees under general performance metrics used for imbalanced classification. Our analysis shows that, under severe UCI, the optimal k is much larger than in balanced classification, whereas this same k leads to sub-optimal, or even inconsistent, estimates of the regression function under other (nonuniform) forms of class imbalance.

6 Numerical Experiments

We provide two numerical experiments to illustrate our results from Sections 4 and 5. We repeat each experiment 100 times and present average results with 95% confidence intervals computed using the central limit theorem. Python implementations and instructions for reproducing each experiment can be found at `https://gitlab.tuebingen.mpg.de/shashank/imbalanced-binary-classification-experiments`. Further technical details regarding the experiments can be found in Appendix E, while Appendix F explores some predictions of our theoretical results on real data from a credit card fraud detection problem.

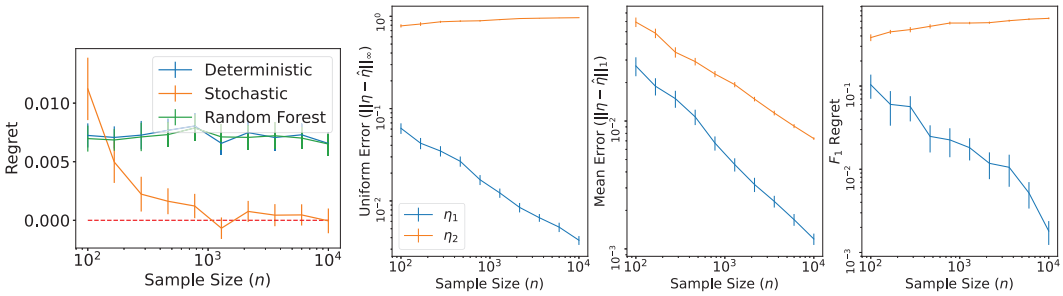
Experiment 1 Example 4 showed that, under general CMMs, deterministic RTCs are sometimes unable to approach optimal classification performance, necessitating stochastic RTCs. This experiment demonstrates this gap numerically. Suppose $\mathcal{X} = [0, 1]$, over which X is uniformly distributed, and for all $x \in \mathcal{X}$, $\eta(x) = 0.5 \cdot 1\{1/3 \leqslant x < 2/3\} + 1\{2/3 \leqslant x\}$.

Consider the CMM $M(C) = \text{TP} \cdot \text{TN}$. Similar to the analysis in Example 4, the optimal value of $M(C) = (5/12)^2 = \frac{25}{144}$ is achievable only by a stochastic classifier, whereas as deterministic classifiers achieve at most $M(C) = 1/6 < 25/144$.

For 10 logarithmically spaced values of n between 10^2 and 10^4, we drew n independent samples of (X, Y) according the above distribution. Using this training data, we selected optimal deterministic and stochastic thresholds $t \in [0, 1]$ and $(t, p) \in [0, 1]^2$ for the kNN classifier by maximizing $M(\widehat{C})$ over 10^4 uniformly spaced values in $[0, 1]$ and $[0, 1]^2$, respectively. Since, in this example, $\alpha = d = 1$, we set $k = \lfloor n^{2/3} \rfloor$ as suggested by Theorem 16. As another point of comparison, we also include a very different deterministic classifier, a random forest, trained with default parameters of Python's `scikit-learn` package. We estimated $M(C)$ using 1000 more independently generated test samples of (X, Y). Figure 2a shows regret, i.e., sub-optimality of each classifier relative to the optimal classifier, in terms of $M(C)$. Consistent with our analysis, regrets of the deterministic classifiers are bounded away from 0, while regret of the stochastic classifier vanishes as n increases.

Experiment 2 This experiment demonstrates that making classifiers robust to severe class imbalance requires distinguishing different sub-types of class imbalance, such as UCI. Suppose $\mathcal{X} = [0, 1]$, $X \sim \text{Uniform}([0, 1])$, and $r \in (0, 1)$. Consider two regression functions $\eta_1(x) = r(1 - x)$ and $\eta_2(x) = \max\{0, 1 - x/r\}$. η_1 and η_2 exhibit the same overall class imbalance, with $r/2$ proportion of samples from class 1. The regression function η_1 satisfies UCI of degree r, whereas η_2 does not satisfy a nontrivial degree of UCI. For sufficiently small $r \in (0, 1)$, specifically $r \in o\left(n^{-d/(2\alpha+2d)}\right)$, Theorem 15 gives that the optimal choice of k under η_1 satisfies $k \in \omega(rn)$. On the other hand, if $k \in \omega(rn)$, then, under η_2, $\mathbb{E}\left[\widehat{\eta}_k(0)\right] \to 0$, so that $\widehat{\eta}_k(0)$ is an inconsistent estimate of $\eta_2(0) = 1$.

For 10 logarithmically spaced values of n between 10^2 and 10^4, we drew n independent samples of (X, Y) according the joint distributions corresponding to each of η_1 and η_2. Since, in this example, $\alpha = d = 1$, to ensure $r \in o\left(n^{-d/(2\alpha+2d)}\right)$, we set $r = n^{-1/2}$. As indicated by Corollary 16, we set $k = \lfloor n^{2/3} r^{-1/3} \rfloor$. We then computed $\mathcal{L}_\infty$ and $\mathcal{L}_1$ distances between the kNN regressor (Eq. (4)) and true regression function. We also drew 1000 independent test samples of (X, Y) and used these to estimate the F_1 score of thresholding the kNN regressor at a threshold t determined by optimizing the



(a) Experiment 1 (b) Experiment 2

Figure 2: (a) Regret of deterministically and stochastically thresholding the kNN classifier, under the CMM $M(C) = \text{TP} \cdot \text{TN}$. (b) Uniform ($\mathcal{L}_\infty$) and average ($\mathcal{L}_1$) errors of kNN regressor, as well as F_1 regret of estimating regression functions η_1 and η_2. Error bars indicate 95% confidence intervals.

empirical F_1 score (over the training data) over 100 uniformly-spaced values of $t \in [0, 1]$. Figure 2b shows the uniform ($\mathcal{L}_\infty$) error, the average ($\mathcal{L}_1$) error, and the F_1 regret, which we bounded in Corollary 8. Consistent with our analysis above, the uniform ($\mathcal{L}_\infty$) error decays to 0 under η_1 but not under η_2. Meanwhile, the average error ($\mathcal{L}_1$) decays to 0 under both η_1 and η_2. Consistent with Corollary 8, the F_1 regret of the thresholded classifier, which decays to 0 under η_1 but not under η_2, mirrors performance of the regressor in uniform ($\mathcal{L}_\infty$) error rather than average ($\mathcal{L}_1$) error.

7 Conclusions

Our main conclusions are as follows. First, without any assumptions on the data-generating distribution, the Bayes-optimal classifier generalizes from accuracy to other performance metrics using a stochastic thresholding procedure, while, in general, deterministic classifiers may not achieve Bayes-optimality. This generalized Bayes classifier provides an optimal performance benchmark relative to which one can analyze classifiers that threshold estimates of the regression function. This includes the kNN classifier, for which we provided new guarantees, including minimax-optimally under uniform loss in the presence of Uniform Class Imbalance. Our results imply that the parameter k needs to be tuned differently for different sub-types of imbalanced classification, suggesting that developing reliable classifiers for severely imbalanced classes may require a more nuanced understanding of the data at hand. Further work is needed to (a) understand how sub-types of class imbalance can be distinguished in practice, (b) develop adaptive classifiers that perform well under multiple imbalance sub-types, and (c) extend our results to the multiclass case.

While this paper focused on statistical properties of stochastic classification, we should point out that using stochastic classifiers in real applications may require careful consideration of possible downstream consequences. On one hand, Theorem 3 provides justification for using (a limited degree of) stochasticity to break certain ties between classes: sometimes, this is provably necessary to optimize performance according to certain metrics. Stochastic classifiers can also be easier to train [Cotter et al., 2019a, Lu et al., 2020] or more robust to adversarial examples [Pinot et al., 2022]. However, stochastic classifiers have risks, including being harder to interpret, explain, or debug, and being vulnerable to manipulation by downstream users (e.g., a user might query the classifier multiple times to produce a desired prediction). Stochastic classifiers may also violate certain notions of fairness, as individuals with identical features might be assigned to different classes. Techniques for derandomizing classifiers [Cotter et al., 2019b, Wu et al., 2022] may help address these issues.

Acknowledgments and Disclosure of Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039B).

https://doi.org/10.52202/068431-1325

References

Sylvain Arlot and Peter L Bartlett. Margin-adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 2011.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

John O Awoyemi, Adebayo O Adetunmbi, and Samuel A Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9. IEEE, 2017.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer, 2015.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory*, 56(4):2034–2040, 2010.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003.

Timothy I Cannings, Thomas B Berrett, and Richard J Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *arXiv preprint arXiv:1704.00642 v3*, 2019.

Alberto Cano, Amelia Zafra, and Sebastián Ventura. Weighted data gravitation classification for standard and imbalanced data. *IEEE Transactions on Cybernetics*, 43(6):1672–1687, 2013.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. *Advances in Neural Information Processing Systems*, 16(16):313–320, 2004.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, pages 300–332. PMLR, 2019a.

Andrew Cotter, Harikrishna Narasimhan, and Maya R Gupta. On making stochastic classifiers deterministic. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Huellermeier. Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *Proceedings of the 30th International Conference on Machine Learning*. PMLR, 2013.

Krzysztof Dembczyński, Wojciech Kotłowski, Oluwasanmi Koyejo, and Nagarajan Natarajan. Consistency analysis for binary classification revisited. In *International Conference on Machine Learning*, pages 961–969. PMLR, 2017.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 1996.

Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *The Journal of Machine Learning Research*, 18(227):1–16, 2018.

Rick Durrett. *Probability: Theory and Examples*. Cambridge university press, fourth edition, 2010.

Rizal Fathony and J Zico Kolter. AP-perf: Incorporating generic performance metrics in differentiable learning. *arXiv preprint arXiv:1912.00965*, 2019.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from Imbalanced Data Sets*. Springer, 2018.

William Fithian and Trevor Hastie. Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of Statistics*, 42(5):1693, 2014.

Evelyn Fix and Joseph L Hodges. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Peter A Flach. The geometry of ROC space: understanding machine learning metrics through roc isometrics. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 194–201, 2003.

Peter A Flach. Classifier calibration. In *Encyclopedia of Machine Learning and Data Mining*. Springer US, 2016.

Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3):982–1009, 2016.

L. Gottlieb, A. Kontorovich, and P. Nisnevitch. Near-optimal sample compression for nearest neighbors. *IEEE Transactions on Information Theory*, 64(6):4120–4128, 2018.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2002.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–33. IEEE, 2020.

Haibo He and Yunqian Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons, 2013.

José Hernández-Orallo, Peter Flach, and César Ferri. Roc curves in cost space. *Machine learning*, 93 (1):71–91, 2013.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 377–384. ACM, 2005.

Aryeh Kontorovich and Roi Weiss. A Bayes consistent 1-NN classifier. In *Artificial Intelligence and Statistics*, pages 480–488. PMLR, 2015.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent Binary Classification with Generalized Performance Metrics. In *Advances in Neural Information Processing Systems 27*, pages 2744–2752. Curran Associates, Inc., 2014.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

David D Lewis. Evaluating and optimizing autonomous text classification systems. In *SIGIR*, volume 95, pages 246–254. Citeseer, 1995.

 https://doi.org/10.52202/068431-1325

Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer, 2014.

Wei Liu and Sanjay Chawla. Class confidence weighted knn algorithms for imbalanced data sets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 345–356. Springer, 2011.

Victoria López, Isaac Triguero, Cristóbal J Carmona, Salvador García, and Francisco Herrera. Addressing imbalanced classification with instance generation techniques: IPADE-ID. *Neurocomputing*, 126:15–28, 2014.

Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing GAN. *arXiv preprint arXiv:1803.09655*, 2018.

Colin McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998.

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611, 2013.

Tom M Mitchell. *Machine Learning*. McGraw-hill New York, 1997.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Advances in Neural Information Processing Systems*, pages 1493–1501, 2014.

Harikrishna Narasimhan, Harish Ramaswamy, Aadirupa Saha, and Shivani Agarwal. Consistent multiclass algorithms for complex performance measures. In *International Conference on Machine Learning*, pages 2398–2407. PMLR, 2015.

Frank Noé, Gianni De Fabritiis, and Cecilia Clementi. Machine learning for protein folding and dynamics. *Current Opinion in Structural Biology*, 60:77–84, 2020.

Rafael Pinot, Laurent Meunier, Florian Yger, Cédric Gouy-Pailler, Yann Chevaleyre, and Jamal Atif. On the robustness of randomized classifiers to adversarial examples. *Machine Learning*, pages 1–33, 2022.

Philippe Rigollet and Xin Tong. Neyman-Pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12(Oct):2831–2855, 2011.

Richard J Samworth. Optimal weighted nearest neighbour classifiers. *Annals of Statistics*, 40(5): 2733–2763, 2012.

Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, and Sylvia Thun. The use of machine learning in rare diseases: a scoping review. *Orphanet Journal of Rare Diseases*, 15(1):1–10, 2020.

Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

Xin Tong. A plug-in approach to Neyman-Pearson classification. *The Journal of Machine Learning Research*, 14(1):3011–3040, 2013.

Xin Tong, Yang Feng, and Anqi Zhao. A survey on Neyman-Pearson classification and suggestions for future research. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(2):64–81, 2016.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats*. Springer Series in Statistics. Springer, New York, 2009.

Cornelis Joost Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, 30(4):365–373, 1974.

Cornelis Joost Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, London, 2nd edition, 1979.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity*, pages 11–30. Springer, 2015.

Sarah Vluymans, Isaac Triguero, Chris Cornelis, and Yvan Saeys. EPRENNID: An evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data. *Neurocomputing*, 216:596–610, 2016.

HaiYing Wang. Logistic regression for massive data with rare events. In *International Conference on Machine Learning*, pages 9829–9836. PMLR, 2020.

Xiaoyan Wang, Ran Li, Bowei Yan, and Oluwasanmi Koyejo. Consistent classification with generalized metrics. *arXiv preprint arXiv:1908.09057*, 2019a.

Xin Wang, Hao Helen Zhang, and Yichao Wu. Multiclass probability estimation with support vector machines. *Journal of Computational and Graphical Statistics*, pages 1–18, 2019b.

Jimmy Wu, Yatong Chen, and Yang Liu. Metric-fair classifier derandomization. In *International Conference on Machine Learning*, pages 23999–24016. PMLR, 2022.

Bowei Yan, Sanmi Koyejo, Kai Zhong, and Pradeep Ravikumar. Binary classification with karmic, threshold-quasi-concave metrics. In *International Conference on Machine Learning*, pages 5531–5540. PMLR, 2018.

Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond Fano's inequality: Bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *The Journal of Machine Learning Research*, 14(1):1033–1090, 2013.

Yujin Zhu, Zhe Wang, and Daqi Gao. Gravitational fixed radius nearest neighbor for imbalanced problem. *Knowledge-Based Systems*, 90:224–238, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices A, B, and C.
3. If you ran experiments...

https://doi.org/10.52202/068431-1325

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code and instructions for reproducing the experimental results are included in the supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

https://doi.org/10.52202/068431-1325