Sharp Error Bounds for Imbalanced Classification: How many Examples in the Minority Class?

Anass Aghbalou
LTCI, Télécom Paris,
Institut Polytechnique de Paris,
France

François Portier
CREST, Ensai
Université de Rennes, France

Anne Sabourin
Université Paris Cité, CNRS, MAP5,
F-75006 Paris, France

Abstract

When dealing with imbalanced classification data, reweighting the loss function is a standard procedure allowing to equilibrate between the true positive and true negative rates within the risk measure. Despite significant theoretical work in this area, existing results do not adequately address a main challenge within the imbalanced classification framework, which is the negligible size of one class in relation to the full sample size and the need to rescale the risk function by a probability tending to zero. To address this gap, we present two novel contributions in the setting where the rare class probability approaches zero: (1) a non asymptotic fast rate probability bound for constrained balanced empirical risk minimization, and (2) a consistent upper bound for balanced nearest neighbors estimates. Our findings provide a clearer understanding of the benefits of class-weighting in realistic settings, opening new avenues for further research in this field.

1 Introduction

Consider the problem of binary classification with covariate X and target $Y \in \{-1, 1\}$. The flagship approach to this problem in statistical learning is Empirical Risk Minimization (ERM), which produces approximate minimizers of $\mathcal{R}(g) = \mathbb{E}[\ell(g(X), Y)]$, given a loss function ℓ and a family of candidate classifiers $g \in \mathcal{G}$, with the help of observed data. with classifier g, $\ell_g(X, Y) = \ell(g(X), Y)$. However, when the underlying distribution is imbalanced, that is $p = \mathbb{P}(Y = +1)$ is relatively small, minimizing empirical version of $\mathcal{R}$ often leads to trivial classification rules for which the majority class is always predicted, because minimizing $\mathcal{R}(g)$ in that case is similar to minimizing $\mathbb{E}[\ell(g(X), Y) \mid Y = -1]$. Indeed by the law of total probabilities, $\mathcal{R}(g) = p\mathbb{E}[\ell(g(X), Y) \mid Y = +1] + (1 - p)\mathbb{E}[\ell(g(X), Y) \mid Y = -1]$ and the former term is negligible with respect to the latter when $p \ll 1$. For this reason, even though standard ERM approaches might enjoy satisfactory generalization properties over imbalanced distributions, with respect to the standard risk $\mathcal{R}$, they may lead to unpleasantly high false negative rates and in general the average error on the minority class has no reason to be small, as its contribution to the overall risk $\mathcal{R}$ is negligible. This is typically what should be avoided in many applications when false negatives are of particular concern, among which medical diagnosis or anomaly detection for aircraft engines, considering the tremendous cost of an error regarding a positive example.

Bypassing the shortcoming described above is the main goal of many works regarding imbalanced classification. The existing literature may be roughly divided into oversampling approaches such as SMOTE and GAN (Chawla et al., 2002; Mariani et al., 2018), undersampling procedures (Liu et al., 2009; Triguero et al., 2015) and risk balancing procedures also known as *cost-sensitive learning* (Scott, 2012; Xu et al., 2020). Here we focus on the latter approach which enjoys numerous benefits, including simplicity, improved decision-making (Elkan, 2001a; Viaene and Dedene, 2005), improved class probability estimation (Wang et al., 2019; Fu et al., 2022), better resource allocation (Xiong et al., 2015; Ryu et al., 2017) and increased fairness (Menon and Williamson, 2018; Agarwal et al., 2018). By incorporating the varying costs of misclassification into the learning process, it enables models to make more informed and accurate predictions for the minority class, leading to higher-quality predictions. Balancing the risk consists of minimizing risk measures that differ

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

significantly from the standard empirical risk, by means of an appropriate weighting of the negative and positive errors, in order to achieve a balance between the contributions of the positive and negative classes to the overall risk. In the present paper our main focus is the balanced-risk, $\mathcal{R}_p(g) = \mathbb{E}\left[\ell(g(X), Y) \mid Y = +1\right] + \mathbb{E}\left[\ell(g(X), Y) \mid Y = -1\right]$. Other metrics might be considered as detailed for instance in Table 1 in Menon et al. (2013) which we do not analyze here for the sake of conciseness, even though our techniques of proof may be straightforwardly extended to handle these variants.

Empirical risk minimization based on the balanced risk is a natural idea, which is widely exploited by practitioners and has demonstrated its practical relevance in several operational contexts (Elkan, 2001b; Sun et al., 2007; Wang et al., 2016; Khan et al., 2018; Pathak et al., 2022). From a theoretical perspective, class imbalance has been the subject of several works. For instance, the consistency of the resulting classifier is investigated in Koyejo et al. (2014). Several different risk measures and loss functions are considered in Menon et al. (2013) where results of asymptotic nature are established, for fixed $p > 0$, as $n \to \infty$. Also in the recent work by Xu et al. (2020), generalization bounds are established for the imbalanced multi-class problem for a robust variant of the balanced risk considered here. Their main results from the perspective of class imbalance, is their Theorem 1 where the upper bound on the (robust) risk includes a term scaling as $1/(p\sqrt{n})$. A related subject is weighted ERM where the purpose is to learn from biased data (see *e.g.* Vogel et al. (2020); Bertail et al. (2021) and the references therein), that is, the training distribution and the target distribution differ. The imbalanced classification problem may be seen as a particular instance of this transfer learning problem, where the training distribution is imbalanced and the target is a balanced version of it with equal class weights. A necessary assumption in Bertail et al. (2021) is that the density of the target with respect to the source is bounded, which in our context is equivalent to requiring that p is bounded away from 0, an explicit assumption in Vogel et al. (2020) where the main results impose that $p > \epsilon$ for some fixed $\epsilon > 0$.

The common working assumption in the cited references that p is bounded from below, renders their application disputable in concrete situations where the number of positive examples is negligible with respect to a wealth of negative instances. To our best knowledge the literature is silent regarding such a situation. More precisely, we have not found neither asymptotic results covering the case where p depends on n in such a way that $p \to 0$ as $n \to \infty$; nor finite sample bounds which would remain sharp even in situations where p is much smaller than $1/\sqrt{n}$. Such situations arise in many examples in machine learning (see *e.g.* the motivating examples in the next section). However, existing works assume that the sizes of both classes are of comparable magnitude, which leaves a gap between theory and practice. A possible explanation is that existing works do not exploit the full potential of the *low variance* of the loss functions on the minority class typically induced by boundedness assumptions combined with a low expected value associated with a small p.

It is the main purpose of this work to overcome this bottleneck and obtain generalization guarantees for the balanced risk which remain sharp even for very small p, that is, under sever class imbalance. Our purpose is to obtain upper bounds on the deviations of the empirical risk (and thus on the empirical risk minimizer) matching the state-of-the art, up to replacing the sample size n with np, the mean size of the rare class. To our best knowledge, the theoretical results which come closest to this goal are normalized Vapnik-type inequalities (Theorem 1.11 in Lugosi (2002)) and relative deviations (Section 5.1 in Boucheron et al. (2005)). However the latter results only apply to binary valued functions and as such do not extend immediately to general real valued loss functions which we consider in this paper, nor do they yield fast rates for *imbalanced* classification problems, although relative deviations play a key role in establishing fast rates in *standard* classification as reviewed in Section 5 from Boucheron et al. (2005). Also, as explained above, we have not found any theoretical result regarding imbalanced classification which would leverage these bounds in order to obtain guarantees with leading terms depending on np instead of n.

Our main tools are (i) Bernstein-type concentration inequalities (that is, upper bounds including a variance term) for empirical processes that are consequences of Talagrand inequalities such as in Giné and Guillou (2001), (ii) fine controls of the expected deviations of the supremum error in the vicinity of the Bayes classifier, by means of local Rademacher complexities Bartlett et al. (2005); Bartlett and Mendelson (2006). Our contributions are two-fold.

1. We establish an estimation error bound on the balanced risk which holds true for VC classes of functions, which scales as $1/\sqrt{np}$ instead of the typical rate $1/\sqrt{n}$ in well-balanced problem, or $1/(p\sqrt{n})$ in existing works regarding the imbalanced case (*e.g.* as in Xu et al. (2020)). Thus, in practice, our setting encompasses the case where $p \ll 1$ (severe class imbalanced) and our upper bound constitutes a crucial improvement by a factor $\sqrt{p}$ compared with existing works in imbalanced classification. Applying the previous bound to the k-nearest neighbor classification rule, we obtain the following new consistency result: as soon as kp goes to infinity, the nearest neighbors classification rule is consistent in case of relative rarity.

2. We obtain fast rates for empirical risk minimization procedures under an additional classical assumption called a Bernstein condition. Namely we prove upper bounds on the excess risk scaling as $1/(np)$, which matches fast rate results in the standard, balanced case, up to replacing the full sample size n with the expected minority class size np. To our best knowledge such fast rates are the first of their kind in the imbalanced classification literature.

Outline. Some mathematical background about imbalanced classification and some motivating examples are given in Section 2. In Section 3, we state our first non-asymptotic bound on the estimation error over VC class of functions and consider application to k-nearest neighbor classification rules. In Section 4, fast convergence rates are obtained and an application to ERM is given. Finally, some numerical experiments are provided in Section 5 to illustrate the theory developed in the paper. All proofs of the mathematical statements are in the supplementary material.

2 Definitions and Notations

Consider a standard binary classification problem where random covariates X, defined over a space $\mathcal{X}$, are employed to distinguish between two classes defined by their labels $Y = 1$ and $Y = -1$. The underlying probability measure is denoted by $\mathbb{P}$ and the associated expectancy, by $\mathbb{E}$. The law of (X, Y) on the sample space $\mathcal{X} \times \mathcal{Y} := \mathcal{X} \times \{-1, 1\}$, is denoted by P. We assume that the label $Y = 1$ corresponds to minority class, i.e., $p = \mathbb{P}(Y = 1) \ll 1$. In the sequel we assume that $p > 0$, even though p may be arbitrarily small.

We adopt notation from empirical process theory. Given a measure μ on $\mathcal{X} \times \mathcal{Y}$ and a real function f defined over $\mathcal{X} \times \mathcal{Y}$, we denote $\mu(f) = \int f d\mu$. When $f = \mathbb{1}_C$ for a measurable set C, we may write interchangeably $\mu(f) = \mu(\mathbb{1}_C) = \mu(C)$. We denote by P_+ the conditional law of (X, Y) given that $Y = +1$, thus

$$P_+(f) = \frac{\mathbb{E}(f(X, Y)\mathbb{1}\{Y = 1\})}{p} = \mathbb{E}(f(X, Y) \mid Y = 1).$$

In addition, we denote by $\mathrm{Var}_+(f)$ the conditional variance of $f(X, Y)$ given that $Y = +1$. The conditional distribution and variance P_- and Var_- are defined similarly, conditional to $Y = -1$. Consider more generally the weighted probability measures, for $q \in (0, 1)$,

$$P_q(f) = \frac{1}{2}\big(q^{-1}P(fI_+) + (1 - q)^{-1}P(fI_-)\big),$$

where $fI_s = f(x, y)\mathbb{1}\{y = s.1\}$, $s \in \{+, -\}$. Notice that $P_p f = (P_+ f + P_- f)/2$.

In this paper we consider general discrimination functions (also called *scores*) $g : \mathcal{X} \to \mathbb{R}$ and loss functions $\ell : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}$, and our results will hold under boundedness and Vapnik-type complexity assumptions detailed below in Sections 3, 4. Given a score function g and a loss ℓ, it is convenient to introduce the function $\ell_g : (x, y) \mapsto \ell(g(x), y)$. Thus the (unbalanced) risk of the score function g is $\mathcal{R}(g) = \mathbb{E}[\ell_g(X, Y)]$. Notice that the standard $0-1$ misclassification risk, $\mathcal{R}^{0-1}(g) = \mathbb{P}(g(X) \neq Y)$, is retrieved when g is real valued and $\ell(g(x), y) = \mathrm{sign}(-g(x)y)$. Allowing for more general scores and losses is a standard approach in statistical learning allowing to bypass the NP-hardness of the minimization problem associated with $\mathcal{R}^{0-1}$. Typically (although not required here) $\ell_g(x, y) = \phi(-g(x)y)$, where ϕ is convex and differentiable with $\phi'(0) < 0$ (Zhang, 2004; Bartlett et al., 2006). This ensures that the loss is classification calibrated and that $\mathcal{R}(g) = \mathbb{E}[\ell_g(X, Y)]$ is a convex upper bound of $\mathcal{R}^{0-1}(g)$. Various consistency results ensuring that $g^* = \arg\min_{g \in \mathbb{R}^{\mathcal{X}}} \mathcal{R}(g) = \arg\min_{g \in \mathbb{R}^{\mathcal{X}}} \mathcal{R}^{0-1}(g)$ can be found in Bartlett et al. (2006). Examples include the logistic ($\phi(u) = \log(1 + e^{-u})$), exponential ($\phi(u) = e^{-u}$), squared ($\phi(u) = (1 - u)^2$), and hinge loss ($\phi(u) = \max(0, 1 - u)$).

The balanced $0-1$ risk, defined as the arithmetic mean $\mathcal{R}_p^{0-1}(g) = (P_+(Y \neq g(X)) + P_-(Y \neq g(X)))/2$ is called the *AM risk* in Menon et al. (2013). The minimizer of the latter risk, g_p^*, is known as the balanced Bayes classifier. It returns 1 when $\eta(X) = \mathbb{P}(Y = +1 \mid X) \geq p$ and -1 otherwise (see *e.g.* Th. 2 or Prop. 2 in Koyejo et al. (2014)). Here we consider general weighted risks and real-valued loss function ℓ_g, defined for $g \in \mathcal{G}$, $q \in (0, 1)$ as

$$\mathcal{R}_q(g) = P_q(\ell_q).$$

Of particular interest is the case $q = p$, for which $\mathcal{R}_p$ is called the balanced risk.

Given an independent and identically distributed sample $(X_i, Y_i)_{1 \leq i \leq n}$ according to P, we denote by P_n the empirical measure, $P_n(f) = (1/n)\sum_{i=1}^n f(X_i, Y_i)$, for any measurable and real-valued function f on $\mathcal{X} \times \mathcal{Y}$. While the standard empirical probability is simply expressed as $P_n(f)$ for any measurable function f, the weighted empirical probability with weight $q \in (0, 1)$ is

$$P_{n,q}(f) = \frac{1}{2}\big(q^{-1}P_n(fI_+) + (1 - q)^{-1}P_n(fI_-)\big).$$

The balanced empirical probability $P_{n,\hat{p}}(f)$ must be defined in terms of $\hat{p}$ if p is unkown. We shall sometimes use that $P_{n,\hat{p}}(f) = (P_{n,+}f + P_{n,-}f)/2$, where $P_{n,+}(f) = \hat{p}^{-1}P_n(fI_+)$ (by convention we set $P_{n,+}(f) = 0$ when $\hat{p} = P_n(Y = 1) = 0$). The empirical measure of the negative class, $P_{n,-}$, is defined in a similar manner. For $q \in (0,1)$ the weighted q-empirical risk is

$$\mathcal{R}_{n,q}(g) = P_{n,q}(\ell_g).$$

Finally the balanced empirical risk considered in this paper is

$$\mathcal{R}_{n,\hat{p}}(g) = P_{n,\hat{p}}(\ell_g) = \frac{1}{2}\left(P_{n,+}(\ell_g) + P_{n,-}(\ell_g)\right).$$

For simplicity we make the standard assumption throughout this paper that for all $q \in (0,1)$, a minimizer g_q^* (resp. $\hat{g}_q$) of $\mathcal{R}_q$ (resp. $\mathcal{R}_{n,q}$) exists, in particular that g_p^* (resp. $\hat{g}_{\hat{p}}$), a minimizer of the balanced risk $\mathcal{R}_p$(resp. $\mathcal{R}_{n,\hat{p}}$), exists.

Motivating examples We now present two examples where the probability $p \to 0$ as $n \to \infty$:

1. The first example is the problem of contaminated data which is central in the robustness literature. A common theoretical assumption is that the number of anomalies n_0 grows sub-linearly with the sample size, as discussed in (Xu et al., 2012; Staerman et al., 2021). In this context, $n_0 = n^a$ for some $a < 1$ and consequently, $p = n^{a-1} \to 0$.

2. The second example pertains to Extreme Value Theory (EVT) (Resnick, 2013; Goix et al., 2015; Jalalzai et al., 2018; Aghbalou et al., 2023). Consider a continuous positive random variable T, predicting exceedances over arbitrarily high threshold t may be viewed as a binary classification problem. Indeed for fixed t, consider the binary target $Y = \mathbb{1}\{T > t\} - \mathbb{1}\{T \leq t\}$ with marginal class probability $p = P(T > t)$. The goal is thus to predict Y, by means of the covariate vector X. In practice, EVT based approaches set the threshold t as the $1-\alpha$ quantile of T with $\alpha = k/n \to 0$ and $k = o(n)$. This approach essentially assumes that the positive class consists of the $k = o(n)$ largest observations of T so that $P(T > t) = P(Y = 1) = k/n \to 0$.

The considered framework "p going to 0 with n" is also valuable from a theoretical perspective as it allows to picture a learning frontier defined in terms of the relative order of magnitude of p and n, above which learning is consistent while below this threshold, one would not be able to estimate the quantity of interest. In a similar spirit, the in high-dimensional statistics, it is customary to let the underlying dimension to grow with the sample size.

3 Standard Learning Rates Under Relative Rarity

3.1 A First Deviation Inequality for Balanced Risks

The primary goal of this paper is to assess the error associated with estimating the balanced risk $\mathcal{R}_p(g)$ using the empirical balanced risk $\mathcal{R}_{n,\hat{p}}(g)$. Given the definition of the balanced risk, the quantity of interest takes the form $(P_{n,+} - P_+)(f)$, and a similar analysis applies to $(P_{n,-} - P_-)(f)$. In this paper we control the complexity of the function class *via* the following notion of VC-complexity. Let $(S, \mathcal{S})$ be a measurable space. Let $\mathcal{F}$ be a class of real valued functions defined on S and Q be a probability measure on $(S, \mathcal{S})$. Given $\mathcal{F} \subset L_2(Q)$, i.e., $Q(f^2) < \infty$, for each $f \in \mathcal{F}$, the ϵ-covering number, denoted by $\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon)$, is defined as the smallest number of closed $L_2(Q)$-balls of radius $\epsilon > 0$ needed to cover $\mathcal{F}$. For a given class of functions $\mathcal{F}$, F is called an envelope if $|f(x)| \leq F(x)$, for all $x \in S$ and all $f \in \mathcal{F}$.

Definition 3.1. The family of functions $\mathcal{F}$ is said to be of VC-type with constant envelope $U > 0$ and parameters $v \geq 1$ and $A \geq 1$ if all functions in $\mathcal{F}$ are bounded by U and for any $0 < \epsilon < 1$ and any probability measure Q on $(S, \mathcal{S})$, we have

$$\mathcal{N}\left(\mathcal{F}, L_2(Q), \epsilon U\right) \leq (A/\epsilon)^v.$$

The connection between the usual VC definition (Vapnik and Chervonenkis, 1971) and Definition 3.1 can be directly established through Haussler's inequality (Haussler, 1995), which indicates that the covering number of a class of binary classifiers with VC dimension v (in the sense of Vapnik and Chervonenkis (1971)) is given by

$$\mathcal{N}\left(\mathcal{F}, L_2(Q), \epsilon\right) \leq e(v+1)(2e/\epsilon^2)^v \leq (e^2/\epsilon)^{2v}.$$

Thus a VC-class of functions in the sense of Vapnik and Chervonenkis (1971) is necessarily a VC-type class in the sense of Definition 3.1.

Notice that within a class $\mathcal{F}$ with envelope $U > 0$, the following variance bounds are automatically satisfied:

$$\sigma_+^2, \sigma_-^2 = \sup_{f \in \mathcal{F}} \mathrm{Var}_+(f), \sup_{f \in \mathcal{F}} \mathrm{Var}_-(f) \leq U^2.$$

The following theorem states a uniform generalization bound that incorporates the probability of each class in such a way that the deviations of the empirical measures are controlled by the expected number of examples in each class, np and $n(1-p)$. Interestingly the deviations may be small even for small p, as soon as the product np is large. The bound also incorporates the conditional variance of a class (σ_+^2, σ_-^2), which will play a key role in our application to nearest neighbors.

Theorem 3.2. *Let $\mathcal{F}$ be of VC-type with constant envelope U and parameters (v, A). For any n and δ*

such that

$$np \geq \max\left[\frac{U^2}{\sigma_+^2} v \log\left(\frac{KAU}{\delta\sigma_+\sqrt{p}}\right), 8\log(1/\delta)\right]$$

we have with probability $1 - 2\delta$,

$$\sup_{f\in\mathcal{F}} |P_{n,+}(f) - P_+(f)| \leq K\sigma_+ \sqrt{\frac{v}{np}\log\left(\frac{KAU}{\delta\sigma_+\sqrt{p}}\right)}$$

for some universal explicit constant $K > 0$.

Remark 3.1. *This upper bounds extends Theorem 1.11 in Lugosi (2002), which is limited to a binary class of functions characterized by finite shatter coefficients. The extension is possible by utilizing Berstein type inequalities for empirical processes defined on VC-type classes of functions as in Giné and Guillou (2001); Portier (2021). It is crucial to recognize that most existing non-asymptotic statistical rates in the imbalanced classification literature (Menon et al., 2013; Koyejo et al., 2014; Xu et al., 2020) follow the rate* $1/(p\sqrt{n})$, *leading to a trivial upper bound when* $p \leq 1/\sqrt{n}$. *In our analysis, the upper bound remains consistent provided that* $np \to \infty$, *thereby emphasizing the merits of using concentration inequalities incorporating the variance of the positive class* $\mathrm{Var}(f\mathbb{1}\{y = 1\}) \leq Up \ll 1$.

The next corollary, which derives from Theorem 3.2 together with standard arguments, provides generalization guarantees for ERM algorithms based upon the balanced risk. Namely it gives an upper bound on the excess risk of a minimizer of the balanced risk. The proof is provided in the supplementary material for completeness.

Corollary 3.3. Suppose that $\{\ell_g : g \in \mathcal{G}\}$ is VC-type with envelope U and parameters (v, A). Under the conditions of Theorem 3.2 and that $p \leq 1/2$, we have, with probability $1 - \delta$,

$$\mathcal{R}_p(\hat{g}_p) \leq \mathcal{R}_p(g_p^*) + K\sigma_{\max}\sqrt{\frac{v}{np}\log\left(\frac{KAU}{\delta\sigma_{\min}\sqrt{p}}\right)},$$

with $\sigma_{\max} = \sigma_- \vee \sigma_+$, $\sigma_{\min} = \sigma_- \wedge \sigma_+$ and $K > 0$ is an explicit universal constant (σ_- and σ_+ are defined as before but with $\{\ell_g : g \in \mathcal{G}\}$ instead of $\mathcal{F}$).

The previous result shows that whenever $np \to \infty$, learning from ERM based on a VC-type class of functions is consistent. Another application of our result pertains to k-nearest neighbor classification algorithms. In this case the sharpness of our bound is fully exploited by leveraging the variance term σ_+. This is the subject of the next section.

Remark 3.2. *The AM risk objective may be viewed as a Lagrangian formulation of the constrained optimization problem associated with the Neyman-Pearson framework (Scott and Nowak, 2005). For instance, Rigollet and Tong (2011) demonstrate an upper bound of order* $1/\sqrt{np}$ *under the condition that the number of observations in each class remains fixed. Their proof technique heavily relies on this fixed sample size assumption for each class. It is unclear how to leverage this result to our context where* n^+ *and* n^- *are random. Another key result in the given reference (Corollary 6) is also close to our framework, however the stated upper bound holds with probability* $1-\delta-\exp(-np^2)$. *Thus the probability of the adverse event becomes large as soon as* $p \leq 1/\sqrt{n}$, *rendering the guarantee vacuous. Incidentally, note that the analysis in the referenced work relies on the assumption that the family of classifiers is finite.*

3.2 Balanced k-Nearest Neighbor

In the context of imbalanced classification, we consider here a balanced version of the standard k-nearest neighbor (k-NN for short) rule, which is designed in relation with the balanced risk $R^*_{bal}(g)$. We establish the consistency of the balanced k-NN classifier with respect to the balanced risk.

Let $x \in \mathbb{R}^d$ and $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^d$. Denote by $B(x, \tau)$ the set of points $z \in \mathbb{R}^d$ such that $\|x - z\| \leq \tau$. For $n \geq 1$ and $k \in \{1, \ldots, n\}$, the k-NN radius at x is defined as

$$\hat{\tau}_x := \inf\left\{\tau \geq 0 : \sum_{i=1}^n 1_{B(x,\tau)}(X_i) \geq k\right\}.$$

Let $\hat{I}(x)$ be the set of index i such that $X_i \in B(x, \hat{\tau}_x)$ and define the estimate of the regression function $\eta(x)$ as

$$\hat{\eta}(x) = \frac{1}{k}\sum_{i\in\hat{I}(x)} \mathbb{1}_{Y_i=1}.$$

While standard NN classification rule is a majority vote following $\hat{\eta}(x)$, i.e., predict 1 whenever $\hat{\eta}(x) \geq 1/2$, it is natural, in view of well known results recalled in Section 2, to consider a balanced classifier $\hat{g}$ for imbalanced data which predicts 1 whenever $\hat{\eta}(x) \geq \hat{p}$, that is $\hat{g} = \mathrm{sign}(\hat{\eta}(x)/\hat{p} - 1)$.

The analysis of the k-NN classification rule is conducted for covariates X that admit a density with respect to the Lebesgue measure. We will need in addition that the support S_X is well shaped and that the density is lower bounded away from zero. These standard regularity conditions in the k-NN literature are recalled below.

(X1) The random variable X admits a density f_X with compact support $S_X \subset \mathbb{R}^d$.

(X2) There is $c > 0$ and $T > 0$ such that $\forall \tau \in (0, T]$ and $\forall x \in S_X$,

$$\lambda(S_X \cap B(x,\tau)) \geq c\lambda(B(x,\tau)),$$

where λ is the Lebesgue measure.

(X3) There is $0 < b_X \leq U_X < +\infty$ such that

$$b_X \leq f_X(x) \leq U_X, \qquad \forall x \in S_X.$$

In light of Proposition A.4 (stated in the supplement), we consider the estimation of $\eta(x)/p$ using the k-NN estimate $\hat{\eta}(x)/\hat{p}$. The proof, which is postponed to the supplementary file, crucially relies on arguments from the proof of our Theorem 3.2 combined with known results concerning the VC dimension of Euclidean balls (Wenocur and Dudley, 1981).

Theorem 3.4. *Suppose that (X1) (X2) and (X3) are fulfilled and that $x \mapsto \eta(x)/p$ is L-Lipschitz on S_X (L does not depend on n). Then whenever $p \to 0$, $pn/\log(n) \to \infty$, $k/\log(n) \to \infty$ and $k/n \to 0$, we have, with probability 1,*

$$\sup_{x\in\mathcal{X}} \left| \frac{\hat{\eta}(x)}{\hat{p}} - \frac{\eta(x)}{p} \right| = O\left(\sqrt{\frac{\log(n)}{kp}} + \left(\frac{k}{n}\right)^{1/d} \right).$$

The consistency of the balanced k-NN with respect to the AM risk, encapsulated in the next corollary, follows from Theorem 3.4 combined with an additional result (Lemma A.4) relating the deviations of the empirical regression function with the excess balanced risk.

Corollary 3.5. Suppose that (X1) (X2) and (X3) are fulfilled and that $x \mapsto \eta(x)/p$ is L-Lipschitz on S_X. Then whenever $p \to 0$, $kp/\log(n) \to \infty$ and $k/n \to 0$, we have, with probability 1,

$$\mathcal{R}_p(\hat{g}_{\hat{p}}) \to \mathcal{R}_p(g_p^*).$$

The main interest of Corollary 3.5 is that the condition for consistency involves the product of the number of neighbors k with the rare class probability p. The take-home message is that learning nonparametric decision rules is possible with imbalanced data, as soon as kp is large enough. In other words local averaging process should be done carefully to ensure a sufficiently large *expected* number of neighbors from the rare class.

4 Fast Rates Under Relative Rarity

We now state and prove a concentration inequality that is key to obtain fast convergence rates for the excess risk in the context of balanced ERM. The following condition regarding a class of functions $\mathcal{F}$ is a prevalent concept within the fast rates literature (Bartlett and Mendelson, 2006; Klochkov and Zhivotovskiy, 2021). A class of function $\mathcal{F}$ satisfies a Bernstein condition relative to a probability measure P on $\mathcal{X} \times \mathcal{Y}$ if there exists $B > 0$ such that

(B1) for all $f \in \mathcal{F}$, $Pf^2 \leq BPf$.

Prior to stating our main result, we introduce classes of functions that are constructed as convex combinations of differences between functions in the original loss class and minimizers of the weighted risk $\mathcal{R}_q$. For $q \in (0,1)$ recall that g_q^* minimizes the $\mathcal{R}_q$-risk over the class of score functions $\mathcal{G}$ and let $\mathcal{H}_q = \{\ell_g - \ell_{g_q^*}, g \in \mathcal{G}\}$. Let

$$\mathcal{H} = \{(1-q)h_q I_+ + q h_q I_-,\ q \in (0,1), h_q \in \mathcal{H}_q\}.$$

With these notations notice already that for $h = (1-q)h_q I_+ + q h_q I_- \in \mathcal{H}$, with $h_q = \ell_g - \ell_{g_q^*}$, the quantity Ph should be interpreted as an excess of weighted risk, since

$$\begin{aligned}\frac{1}{q(q-1)}Ph &= \frac{1}{q}P(h_q I_+) + \frac{1}{1-q}P(h_q I_-) \\ &= P_q h_q = P_q(\ell_g - \ell_{g_q^*}) \\ &= \mathcal{R}_q(g_q) - \mathcal{R}_q(g_q^*).\end{aligned}$$

It turns out that the class $\mathcal{H}$ indeed satisfies a Bernstein assumption under standard assumptions regarding the *original* loss class $\mathcal{L} = \{\ell_g, g \in \mathcal{G}\}$. Namely it is enough to assume that the latter satisfies a strong convexity property and a Lipschitz property, which are commonly satisfied in Machine Learning problems. The proof of the following statement is deferred to the Supplementary material

Lemma 4.1 (Sufficient conditions for $\mathcal{H}$ to satisfy a Bernstein-condition)**.** *Assume that $\mathcal{G}$ is a convex subset of a normed vector space, and that there exists $L, \lambda > 0$, such that for $s \in \{+,-\}$ the functions $g \mapsto P\ell_g I_s$ are respectively $(p\lambda)$-strongly convex and $((1-p)\lambda)$-strongly convex. Assume also that for $g_1, g_2 \in \mathcal{G}$, $\sqrt{P((\ell_{g_1} - \ell_{g_2})^2 I_+)} \leq (\sqrt{p}\,L)\|g_1 - g_2\|$ and $\sqrt{P((\ell_{g_1} - \ell_{g_2})^2 I_-)} \leq (\sqrt{1-p}\,L)\|g_1 - g_2\|$. Then $\mathcal{H}$ satisfies the Bernstein condition (B1), with $B = L^2/\lambda$.*

Example 4.1. Assume that the domain $\mathcal{X}$ is bounded in $\mathbb{R}^d$ *i.e.*, there exists some $\Delta_X > 0$ such as $\forall x \in \mathcal{X}, \|x\| \leq \Delta_X$ for some norm $\|\cdot\|$. Consider the family of classifier and the loss function $\mathcal{G}_u = \{g_\beta : x \mapsto \beta^T x\ \ \|\beta\| \leq u\}$ and $\ell_{g_\beta}(X,Y) = \phi(\beta^T XY)$, where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a twice continuously differentiable non-decreasing function which is μ-strongly convex for some $\mu > 0$. Then we have that the domain $I = \{\beta^\top xy, , \|\beta\| \leq u, x \in \Delta_x, y \in \{-1,+1\}\}$ is bounded and the derivative ϕ' satisfies $\sup_{t\in I} |\phi'|(t) = D < \infty$. Let $V_s = \mathbb{E}\left[XX^\top \,|\, Y = s.1\right], s = +,-$ be the second moment matrix of each class and $\sigma^2_{\max}, \sigma^2_{\min}$ their maximum and minimum eigenvalues. Direct computations show that the Lipschitz and convexity constraints in Lemma 4.1 are respectively $L = D\sigma_{\max}$, and $\lambda = \mu\sigma^2_{\min}$. Under the condition that each class distribution is non degenerate, *i.e.* does not concentrate on any lower dimensional subspace of $\mathbb{R}^d$, we have $\sigma_{\min} > 0$, thus $\lambda > 0$. The assumptions of Lemma 4.1 are thus satisfied, so that Bernstein condition (B1) holds true with

$$B = L^2/\lambda = \frac{D^2\sigma^2_{\max}}{\mu\sigma^2_{\min}}.$$

Another useful property of $\mathcal{H}$ is that it inherits the regularity property of $\mathcal{L}$, see the Supplementary material for details.

Lemma 4.2 (VC-property of $\mathcal{H}$). *If $\mathcal{L} = \{\ell_g, g \in \mathcal{G}\}$ is of VC-type with envelope U and parameters (v, A) then $\mathcal{H}$ also is VC with envelope $2U$ and parameters $(\tilde{v} = 4v + 1, \tilde{A} = 6A)$.*

We now state our main result (fast rates for the deviations of weighted risks) in the light of the good properties of $\mathcal{H}$ stated below. To wit, in our application to empirical risk minimization (Corollaries 4.4, 4.5), the classes $\mathcal{F}_q$ in the following statement will be chosen as $\mathcal{H}_q$, so that we shall have $\mathcal{F} = \mathcal{H}$.

Theorem 4.3. *[Fast rates for the deviations of weighted probabilities] Let $(\mathcal{F}_q, q \in (0,1))$ be a family of classes of functions with common envelope $2U > 0$. Assume that the class of convex combinations $\{(1-q)f_q I_+ + qf_q I_-,\ q \in (0,1), f_q \in \mathcal{F}_q\}$ satisfies Bernstein condition (B1) for some $B \geq 2U$, and that it is of VC-type with parameters $(\tilde{v}, \tilde{A})$.*

Then the deviations of the weighted probabilities P_q over the classes $\mathcal{F}_q$ are uniformly controlled as follows: for any $K > 1$ and every $\delta > 0$, with probability at least $1 - \delta$, for all $q \in (0,1)$ and for all $f_q \in \mathcal{F}_q$,

$$P_q(f_q) \leq \frac{K}{K-1} P_{n,q}(f_q) + \frac{c_1 BK\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{2nq(1-q)},$$

where $c_1 > 0$ is an explicit universal constant given in the proof.
Also, with probability at least $1 - \delta$, $\forall q, \forall f_q \in \mathcal{F}_q$,

$$P_{n,q}(f_q) \leq \frac{K+1}{K} P_q(f_q) + \frac{c_1 BK\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{2nq(1-q)}.$$

Sketch of proof. The main tool for the proof is a fast rate result (Theorem 3.3 in Bartlett et al. (2005), recalled for completeness in the supplementary material as Theorem C.4). The argument from the cited reference relies on a fixed point technique relative to a sub-root function upper bounding some local Rademacher complexity. Leveraging fine controls of the latters (Section C.1) we establish that the fixed point of the sub-root function is of order $O(\log(n)/n)$ and we obtain an explicit control of the deviations of the (standard) empirical measure, under a Bernstein condition (see Proposition C.5). Finally the main result is obtained by applying the latter proposition to the specific class of convex combinations defined in the statement, and rescaling the obtained bound by the quantity $2q(1-q)$, see Section C.1 for details.

□

Discussion. Similar proof techniques can be found in the standard classification literature, for example Corollary 3.7 in Bartlett et al. (2005). Nevertheless, this particular work primarily concentrates on loss functions with binary values, namely $\{0, 1\}$. The proof is based upon the fact that these functions are positive, and it employs the initial definition of the VC dimension (Vapnik and Chervonenkis, 1971). In contrast, other existing works (e.g. Theorem 2.12 in Bartlett and Mendelson (2006) or Example 7.2 in Giné and Koltchinskii (2006)) demonstrate accelerated convergence rates for the *typical* empirical risk minimizers, which do not extend to their balanced counterparts. The present result is more general, as it is uniformly applicable to a broader range of bounded functions and encompasses a more extensive definition of the VC class. This notable extension facilitates the establishment of fast convergence rates for the excess risk of (ML) algorithms employed in imbalanced classification scenarios, such as cost-sensitive logistic regression and balanced boosting (Menon et al., 2013; Koyejo et al., 2014; Tanha et al., 2020; Xu et al., 2020). In the remainder of this section we provide examples of algorithms that verify the assumptions of Theorem 4.3.

As an application of Theorem 4.3, we derive fast rates for the excess risk of empirical risk minimizers. Our result (Corollary 4.4 below) is stated in terms of excess of $\mathcal{R}_{\hat{p}}$ risk.

Corollary 4.4. Assume that $\mathcal{L} = \{\ell_g : g \in \mathcal{G}\}$ is of VC-type with envelope $U > 0$ and parameters (v, A) and assume that $\mathcal{H}$ defined at the beginning of this section satisfies the Bernstein condition (B1) for some $B \geq 2U$ (this is the case *e.g.* under the Assumptions of Lemma 4.1). Let $\hat{g}_{\hat{p}}$ be a minimizer of the empirical balanced risk $\mathcal{R}_{n,\hat{p}}$ considered in Section 3. Then for $\delta > 0$, with probability $1 - \delta$,

$$\mathcal{R}_{\hat{p}}(\hat{g}_{\hat{p}}) - \mathcal{R}_{\hat{p}}(g^*_{\hat{p}}) \leq \frac{c_1 B\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{2n\hat{p}(1-\hat{p})},$$

where $(\tilde{v} = 4v + 1, \tilde{A} = 6A)$, and where the constant $c_1$is the same as in Theorem 4.3.

Proof. We consider the classes of functions $\mathcal{F}_q = \mathcal{H}_q$. The class of convex combinations from the statement of Theorem 4.3 is precisely $\mathcal{H}$, which is B-Bernstein by assumption and it is also of VC-type with parameters $(\tilde{v}, \tilde{A})$ by virtue of Lemma 4.2. We may thus apply Theorem 4.3. Because the result of Theorem 4.3 holds uniformly over $q \in (0,1)$, $f \in \mathcal{F}_q$, one may choose $q = \hat{p}$. Also we choose $f_{\hat{p}} \in \mathcal{F}_{\hat{p}} = \mathcal{H}_{\hat{p}}$ as $f_{\hat{p}} = \ell_{\hat{g}_{\hat{p}}} - \ell_{g^*_{\hat{p}}}$. Then the first term on the right-hand side of the first upper bound in Theorem 4.3 is nonpositive, and the result follows upon choosing $K = 2$. □

Remark 4.1 (Weighting with $\hat{p}$ or p)**.** *Under mild conditions on np and δ, we have that $\hat{p}(1-\hat{p}) \geq p(1-p)/2$ (as indicated by Chernoff's multiplicative bound in Theorem A.1), so that Corollary 4.4 immediately yields a rate of convergence in terms of the true value p rather than its empirical counter part. However whether it is possible to replace $\mathcal{R}_{\hat{p}}$ with $\mathcal{R}_p$ in the statement remains an open question. The main bottleneck seems to be that replacing $\hat{p}$ with p in expressions of the kind '$P_n(fI_+)/\hat{p} + P_n(fI_-)/(1-\hat{p})$', where one term in the summand may be negative, induces additional slow rate terms of order $O(1/\sqrt{np})$.*

We conclude this section by illustrating the significance of our results, through the concrete setting of Example 4.1. The following corollary is a direct consequence of Corollary 4.4 and guarantees fast rates of convergence for constrained ERM, specifically for algorithms of the form $\hat{g}_{u,\hat{p}}(x) = \hat{\beta}_u^T x$ with

$$\hat{\beta}_u = \underset{\|\beta\| \leq u}{\arg\min}\, n^{-1} \sum_{i=1}^{n} \phi(\beta^\top X_i Y_i)\Big(p^{-1}\mathbb{1}\{Y_i = 1\} + \cdots \\ (1-p)^{-1}\mathbb{1}\{Y_i = -1\}\Big).$$

Then from standard arguments the class $\mathcal{L} = \{\ell(x,y) = \phi(\beta^\top\}$ is of VC-type with parameters $(v = 2(d+1), A)$ for some $A > 0$ (see e.g.van der Vaart and Wellner (1996), Chap.2.6) depending on ϕ. The following result derives immediately from the argument of Example 4.1, from Lemma 4.1 and from Corollary 4.4.

Corollary 4.5. In the setting of Example 4.1, let $(v = 2(d+1), A)$ be the parameters of the VC-type class $\mathcal{L}$ defined above the statement. The excess risk of $\hat{g}_u$ verifies, for any $\delta > 0$, with probability $1 - 4\delta$,

$$\mathcal{R}_{\hat{p}}\left(\hat{g}_{\hat{p}}\right) \leq \mathcal{R}_{\hat{p}}(g^*_{\hat{p}}) + \frac{c_1(d+1)D^2\sigma^2_{\max}}{\mu\sigma^2_{\min}} \frac{\log(30A\sqrt{n}/\delta)}{n\hat{p}(1-\hat{p})},$$

where c_1 is as in Theorem 4.3.

Discussion. In the context of constrained logistic regression, where $\phi(x) = \log(1+e^{-x})$, the latter corollary yields fast convergence rates with constants and $L' = 1$, along with $\lambda = e^{-u}$. Corollary 4.5 further establishes accelerated convergence rates for constrained empirical *balanced* risk minimization with respect to losses such as mean squared error, squared hinge, and exponential loss, among others. This outcome aligns with expectations, as constrained empirical risk minimization is equivalent to penalization (Lee et al., 2006; Homrighausen and McDonald, 2017). Numerous studies have demonstrated the effectiveness of penalization in achieving rapid convergence rates (Koren and Levy, 2015; van Erven et al., 2015). This aspect is particularly significant in the present context, as the standard convergence rate for imbalanced classification is $1/\sqrt{np}$, and accelerating the convergence rate leads to a more pronounced impact.

5 Numerical Illustration

In this section, we illustrate on synthetic data our theoretical results on k-NN classification (Corollary 3.5) and on logistic regression (Corollary 4.5). In both cases, particular attention is paid to highly imbalanced settings where $p = n^{-a}$ for some $0 < a < 1$. Due to space constraint, the real data experiments are postponed to the supplement. We use the following simple data generation process to obtain a binary classification i.i.d dataset $(X_i, Y_i)_{i=1,\ldots,n}$, with $X_i \in \mathbb{R}^2$ and $Y \in \{-1, 1\}$. The Y_i are Bernoulli variables with parameter $p = 1/n^a$, for some $a < 1$. Then, given $Y_i = y$, X_i is drawn according to a t-multivariate-student distribution, with parameters (μ_y, σ_y, ν_y), where $(\mu_{-1}, g\mu_1) = ((0,0),(1,1))$, $\sigma_1 = 3\sigma_{-1} = 3I$ and $(\nu_{-1}, \nu_1) = (2.5,\, 1.1)$.

5.1 Balanced k-Nearest Neighbors

Corollary 3.5 provides sufficient conditions on k, n and p for consistency of the k-NN classification rule, the key being that kp should go to ∞. This suggests the existence of a learning frontier on the set (k, p) above which consistent learning is ensured. Our experiments aim at illustrating this fact. Here the training size is $n = 1e4$. We set $p = p_n = 1/n^a$ and $k = n^b$, where a, b vary within the interval $[1/4, 3/4]$ and cover different cases ranging from $pn \to 0$ to $pn \to \infty$. The AM-risk for the classification error associated to the balanced k-NN classifier (estimated with 20 simulations) is displayed as a function of (k, p) in Figure 1. For small values of kp, the performance of the k-nearest neighbors classifier mirrors that of a random guess, maintaining an AM risk near 0.5, while kp large ensures good performance.This observation illustrates (and extends) the conclusion of Corollary 3.5, supporting that consistency is obtained if (and only if) $kp \to \infty$.

5.2 Balanced ERM

Our goal is to demonstrate empirically that the fast convergence rate of order $1/(np)$ obtained in Corollary

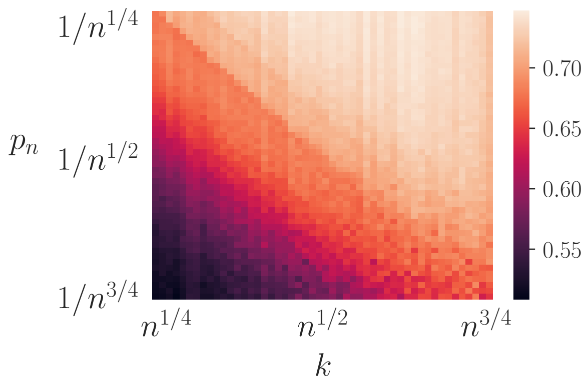
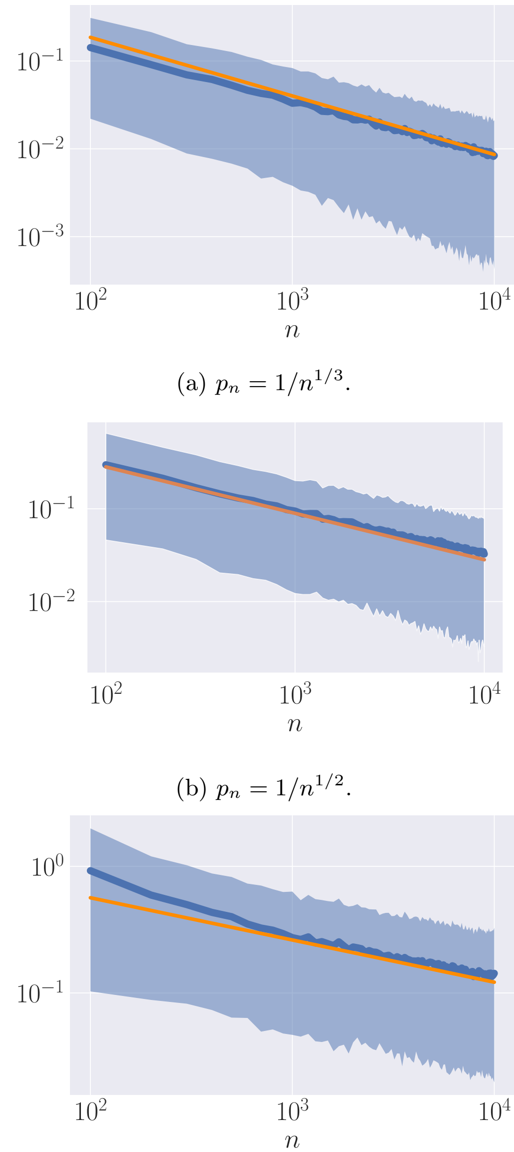


Figure 1: AM risk of the balanced k-NN (heatmap).



(a) $p_n = 1/n^{1/3}$.

(b) $p_n = 1/n^{1/2}$.

(c) $p_n = 1/n^{2/3}$.

Figure 2: Excess balanced risk (log-scale) of logistic regression as a function of n, when $p = p_n \to 0$. Orange line: curve $1/np$. Blue area: inter-quantile range $[0.1, 0.9]$.

.

4.5 is sharp and can be observed in practice for a wide range of values of p. We consider the linear setting of Example 4.1, with the logistic loss: $\ell_g(X,Y) = \log(1-e^{-g(X)Y})$, $g(X) = \beta^T X$ and $\|\beta\| \leq u = 10$. The sample size n ranges over the grid $[100, 1e4]$ and we let $p = n^{-a}$, $a \in \{1/3, 1/2, 2/3\}$. Some Monte-Carlo simulations ($N = 1e5$ simulations) are needed to estimate $g^*_{\hat{p}}$. For simplicity and to alleviate the computational burde we consider g^*_p for fixed p instead, and the balanced risk $\mathcal{R}_p$ instead of $\mathcal{R}_{\hat{p}}$. The risk function $\mathcal{R}_p$ is also estimated based on Monte-Carlo simulations ($N' = 1e4$). We thus obtain both $\mathcal{R}_p(g^*_p)$ and $\mathcal{R}_p(\hat{g}_{\hat{p}})$, and the value of the excess p-risk follows We perform report the average and the upper 0.10 and 0.90 quantile of the absolute error obtained over the $n_{simu} = 1e4$ experiments. Figure 2 displays the excess risk as a function of the sample size n on a logarithmic scale, for $a \in \{1/3, 1/2, 2/3\}$. Other figures exploring other values of a are reported in the supplementary material. We notice that the excess risk vanishes in the same way as the function $n \mapsto 1/np$ confirming the accuracy of the upper bound from Corollary 4.5.

6 Conclusion

In this paper, we have derived upper bounds for the balanced risk in highly imbalanced classification scenarios. Notably, our bounds remain consistent even under severe class imbalance ($p \to 0$), setting our work apart from existing studies in imbalanced classification (Menon et al., 2013; Koyejo et al., 2014; Xu et al., 2020). Furthermore, it is worth to highlight that this is the first study to achieve fast rates in imbalanced classification, marking a significant advancement in the field. Our findings confirm that both risk-balancing approaches and cost-sensitive learning are consistent across nearly all imbalanced classification scenarios. This aligns with experimental works previously documented in the literature (Elkan, 2001b; Wang et al., 2016; Khan et al., 2018; Wang et al., 2019; Pathak et al., 2022). Furthermore, the methodologies and proof techniques presented in this paper are adaptable to other imbalanced classification metrics beyond balanced classification. Potential extensions include demonstrating consistency for metrics such as the F_1-measure, Recall, and their respective variants.

Acknowledgements

We thank anonymous reviewers for their useful comments that helped improve the final version of this paper. Anne Sabourin acknowledges the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project), reference ANR-23-CE40-0009-01 (EXSTA project), and of the chair DSAIDIS from Télécom Paris.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Aghbalou, A., Portier, F., Sabourin, A., and Zhou, C. (2023). Tail inverse regression for dimension reduction with extreme response.

Ausset, G., Clémençon, S., and Portier, F. (2021). Nearest neighbour based estimates of gradients: Sharp nonasymptotic bounds and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 532–540. PMLR.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537.

Bartlett, P. L., Jordan, M. I., and Mcauliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability theory and related fields*, 135(3):311–334.

Bertail, P., Clémençon, S., Guyonvarch, Y., and Noiry, N. (2021). Learning from biased data: A semiparametric approach. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 803–812. PMLR.

Biau, G. and Devroye, L. (2015). *Lectures on the nearest neighbor method*, volume 246. Springer.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Elkan, C. (2001a). The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle*, 1.

Elkan, C. (2001b). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17-1, pages 973–978. Lawrence Erlbaum Associates Ltd.

Fu, S., Yu, X., and Tian, Y. (2022). Cost sensitive ν-support vector machine with linex loss. *Information Processing & Management*, 59(2):102809.

Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Annales de l'IHP Probabilités et statistiques*, 37(4):503–522.

Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143 – 1216.

Giné, E. and Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields*, 143(3-4):569–596.

Goix, N., Sabourin, A., and Clémençon, S. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on Learning Theory*, pages 843–860. PMLR.

Hagerup, T. and Rüb, C. (1990). A guided tour of chernoff bounds. *Information processing letters*, 33(6):305–308.

Haussler, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232.

Homrighausen, D. and McDonald, D. J. (2017). Risk consistency of cross-validation with lasso-type procedures. *Statistica Sinica*, pages 1017–1036.

Jalalzai, H., Clémençon, S., and Sabourin, A. (2018). On binary classification in extreme regions. In *NeurIPS*, pages 3096–3104.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., and Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587.

Klochkov, Y. and Zhivotovskiy, N. (2021). Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076.

Koren, T. and Levy, K. (2015). Fast rates for exp-concave empirical risk minimization. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. *Advances in neural information processing systems*, 27.

Lee, S.-I., Lee, H., Abbeel, P., and Ng, A. Y. (2006). Efficient l˜ 1 regularized logistic regression. In *Aaai*, volume 6, pages 401–408.

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550.

Lugosi, G. (2002). Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. (2018). Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. (2013). On the statistical consistency of algorithms for binary classification under class imbalance. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28-3 of *Proceedings of Machine Learning Research*, pages 603–611, Atlanta, Georgia, USA. PMLR.

Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR.

Pathak, Y., Shukla, P., Tiwari, A., Stalin, S., Singh, S., and Shukla, P. (2022). Deep transfer learning based classification model for covid-19 disease. *IRBM*, 43(2):87–92.

Plassier, V., Portier, F., and Segers, J. (2023). Risk bounds when learning infinitely many response functions by ordinary linear regression. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 59-1, pages 53–78. Institut Henri Poincaré.

Portier, F. (2021). Nearest neighbor process: weak convergence and non-asymptotic bound.

Resnick, S. I. (2013). *Extreme values, regular variation and point processes*. Springer.

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

Ryu, D., Jang, J.-I., and Baik, J. (2017). A transfer cost-sensitive boosting approach for cross-project defect prediction. *Software Quality Journal*, 25:235–272.

Scott, C. (2012). Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6(none):958 – 992.

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

Staerman, G., Laforgue, P., Mozharovskyi, P., and d'Alché Buc, F. (2021). When ot meets mom: Robust estimation of wasserstein distance. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 136–144. PMLR.

Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378.

Talagrand, M. (1994). Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28 – 76.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., and Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7(1).

Triguero, I., Galar, M., Vluymans, S., Cornelis, C., Bustince, H., Herrera, F., and Saeys, Y. (2015). Evolutionary undersampling for imbalanced big data classification. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, pages 715–722.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D., and Williamson, R. C. (2015). Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54):1793–1861.

Van Handel, R. (2014). Probability in high dimension. Technical report, PRINCETON UNIV NJ.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

Viaene, S. and Dedene, G. (2005). Cost-sensitive learning and decision making revisited. *European Journal of Operational Research*, 166(1):212–220. Metaheuristics and Worst-Case Guarantee Algorithms: Relations, Provable Properties and Applications.

Vogel, R., Achab, M., Clémencon, S., and Tillier, C. (2020). Weighted empirical risk minimization: Transfer learning based on importance sampling. In *28th*

European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pages 515–520. i6doc. com.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., and Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4368–4374.

Wang, X., Zhang, H. H., and Wu, Y. (2019). Multiclass probability estimation with support vector machines. *Journal of Computational and Graphical Statistics*, 28(3):586–595.

Wenocur, R. S. and Dudley, R. M. (1981). Some special vapnik-chervonenkis classes. *Discrete Mathematics*, 33(3):313–318.

Xiong, P., Chi, Y., Zhu, S., Moon, H. J., Pu, C., and Hacgümüş, H. (2015). Smartsla: Cost-sensitive management of virtualized resources for cpu-bound database services. *IEEE Transactions on Parallel and Distributed Systems*, 26(5):1441–1451.

Xu, H., Caramanis, C., and Mannor, S. (2012). Outlier-robust pca: The high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572.

Xu, Z., Dan, C., Khim, J., and Ravikumar, P. (2020). Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning*, pages 10544–10554. PMLR.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85.

Supplementary material for "Sharp Error Bounds for Imbalanced Classification: How many Examples in the Minority Class?"

A Auxiliary results

The following standard Chernoff inequality is stated and proven in Hagerup and Rüb (1990).

Theorem A.1. *Let $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables valued in $\{0,1\}$. Set $\mu = nP(Z_1)$ and $S = \sum_{i=1}^n Z_i$. For any $\delta \in (0,1)$ and all $n \geq 1$, we have with probability $1-\delta$:*

$$S \geq \left(1 - \sqrt{\frac{2\log(1/\delta)}{\mu}}\right)\mu.$$

In addition, for any $\delta \in (0,1)$ and $n \geq 1$, we have with probability $1-\delta$:

$$S \leq \left(1 + \sqrt{\frac{3\log(1/\delta)}{\mu}}\right)\mu.$$

The following is taken from Portier (2021). Other similar results are given in Giné and Guillou (2001) or Plassier et al. (2023) with non-explicit constants.

Theorem A.2. *Let $(Z_1, \dots, Z_n)$ be an independent and identically distributed collection of random variables with common distribution P on $(S, \mathcal{S})$. Let $\mathcal{G}$ be of VC-type with parameters $v \geq 1$, $A \geq 1$ and uniform envelope $U \geq \sup_{g\in\mathcal{G},\, x\in S} |g(x)|$. Let σ be such that $\sup_{g\in\mathcal{G}} \mathrm{Var}_P(g) \leq \sigma^2 \leq U^2$. For any $n \geq 1$ and $\delta \in (0,1)$, it holds, with probability at least $1-\delta$,*

$$\sup_{g\in\mathcal{G}} \left| \sum_{i=1}^n \{g(Z_i) - P(g)\} \right| \leq K_1 \sqrt{vn\sigma^2 \log\left(9AU/(\sigma\delta)\right)} + K_2 U v \log\left(9AU/(\sigma\delta)\right),$$

with $K_1 = 5C$, $K_2 = 64C^2$ and $C = 12$.

Lemma A.3. *Suppose that $\mathcal{F}$ (resp. $\mathcal{G}$) defined on $(S, \mathcal{S})$ is of VC-type with envelope U and parameter (v, A) and let $E \in \mathcal{S}$. The following holds:*

1. *$\{fI_E : f \in \mathcal{F}\}$ is of VC-type with envelope U and parameter (v, A),*
2. *$\mathcal{F} - \mathcal{G} = \{f - g : f \in \mathcal{F},\, g \in \mathcal{G}\}$ is of VC-type with envelope $2U$ and parameter $(2v, 2A)$,*
3. *$\{f - P(f|E) : f \in \mathcal{F}\}$ is of VC-type with envelope $2U$ and parameter $(2v, A)$,*
4. *$\{qf + (1-q)g : f \in \mathcal{F},\, g \in \mathcal{G},\, q \in [0,1]\}$ is of VC-type with envelope U and parameter $(2v+1, 3A)$.*

Proof. Let Q be a probability measure on $(S,\mathcal{S})$. Let $(f_k)_{k=1,\dots,K}$ be the center of an ϵU-covering of $(\mathcal{F}, Q)$. The first statement follow from the fact that $\|f1_E - f_k 1_E\|_{L_2(Q)} \leq \|f - f_k\|_{L_2(Q)}$. For the second statement consider $U\epsilon$-covers $(f_k, k \leq K)$ and $(g_j, j \leq J)$ respectively of $\mathcal{F}$ and $\mathcal{G}$. Then the triangle inequality shows that $(f_k - g_j), k \leq K, j \leq J$ forms a $2U\epsilon$-cover of $\mathcal{F} + \mathcal{G}$ and the result follows. Now let $(\tilde{f}_k)_{k=1,\dots,K}$ be the center of an ϵU-covering of $(\mathcal{F}, P_E)$ with $P_E(\cdot) = P(\cdot|E)$. Consider the covering induces by the centers $(f_k - P_E(\tilde{f}_j))_{1\leq k,j\leq K}$ made of K^2 elements. Suppose that $f \in \mathcal{F}$. Then there is k and j such that

$$\begin{aligned}
\|(f - P_E(f)) - (f_k - P_E(\tilde{f}_j)\|_{L_2(Q)} &\leq \|f - f_k\|_{L_2(Q)} + P_E(f - \tilde{f}_j) \\
&\leq \|f - f_k\|_{L_2(Q)} + \|f - \tilde{f}_j\|_{L_2(P)} \\
&\leq 2U\epsilon.
\end{aligned}$$

Hence we have found a $2U\epsilon$-covering of size K^2 which by assumption is smaller than $(A/\epsilon)^{2v}$. This implies the third statement of the lemma. For the last statement, let $\mathcal{H} = \{qf + (1-q)g \,:\, f \in \mathcal{F},\, g \in \mathcal{G},\, q \in [0,1]\}$. Let $(f_k)_{k=1,\dots,K}$ (resp. g_ℓ, $(g_\ell)_{\ell=1,\dots,L}$) be the center of an ϵU-covering of $(\mathcal{F}, Q)$ (resp. $(\mathcal{G}, Q)$). Let q_i, $i = 1, \dots, \lfloor 1/\epsilon \rfloor$ be an ϵ-covering of $[0,1]$. Let $h = qf + (1-q)g$ be such that $q \in [0,1]$, $f \in \mathcal{F}$ and $g \in \mathcal{G}$. There is f_k, g_ℓ, q_i such that

$$\begin{aligned}
&\|qf + (1-q)g - (q_i f_k + (1-q_i)g_\ell)\|_{L_2(Q)} \\
&\leq \|q(f - f_k) + (1-q)(g - g_\ell)\|_{L_2(Q)} + \|(q - q_i)f_k + (q_i - q)g_\ell\|_{L_2(Q)} \\
&\leq qU\epsilon + (1-q)U\epsilon + \epsilon U + \epsilon U = 3\epsilon U.
\end{aligned}$$

Hence the element $(q_i f_k + (1-q_i)g_\ell)$ form an $3\epsilon U$-covering in the space $L_2(Q)$. There are $KL\lceil 1/\epsilon \rceil \leq KL/\epsilon$ such elements. As a consequence, since $\mathcal{F}$ and $\mathcal{G}$ are are of VC-type, it follows that

$$\mathcal{N}(\mathcal{H}, L_2(Q), 3\epsilon U) \leq (A/\epsilon)^{2v}(1/\epsilon) \leq (A/\epsilon)^{2v+1}$$

which implies the stated result.

□

The next lemma generalizes Theorem 17.1 from Biau and Devroye (2015) to the balanced type classifiers.

Lemma A.4. *For any classifier g that writes $g(x) = \operatorname{sign}(\nu(x) - 1)$, $x \in \mathcal{X}$, we have*

$$\mathcal{R}_p(g) - \mathcal{R}_p(g_p^*) = \mathbb{E}\left[\mathbb{1}_{g(X) \neq g_p^*(X)} \frac{|\eta(X) - p|}{p(1-p)}\right],$$

where g_p^* is the balanced Bayes classifier (introduced in Section 2). Furthermore, whenever $p \leq 1/2$,

$$\mathcal{R}_p(g) - \mathcal{R}_p(g_p^*) \leq 2\mathbb{E}\left[|\nu(X) - \nu^*(X)|\right]$$

where $\nu^*(x) = \eta(x)/p$.

Proof. The balanced risk writes as

$$\begin{aligned}
\mathcal{R}_p(g) &= P_+\left(\nu(X) < 1\right) + P_-\left(\nu(X) \geq 1\right) \\
&= \mathbb{E}\left[\frac{\mathbb{I}_{(\nu(X)<1)}\mathbb{I}_{Y=1}}{p} + \frac{\mathbb{I}_{(\nu(X)\geq 1)}\mathbb{I}_{Y=-1}}{1-p}\right].
\end{aligned}$$

In addition, using a conditioning argument yields,

$$\mathcal{R}_p(g) = \mathbb{E}\left[\frac{\mathbb{I}_{(\nu(X)<1)}\eta(X)}{p} + \frac{\mathbb{I}_{(\nu(X)\geq 1)}(1 - \eta(X))}{1-p}\right].$$

Similarly we have

$$\mathcal{R}_p(g^*) = \mathbb{E}\left[\frac{\mathbb{I}_{(\nu^*(X)<1)}\eta(X)}{p} + \frac{\mathbb{1}_{(\nu^*(X)\geq 1)}(1 - \eta(X))}{1-p}\right].$$

It follows that

$$\begin{aligned}\mathcal{R}_p(g) - \mathcal{R}_p(g_p^*) &= \mathbb{E}\left[\mathbb{I}_{\text{sign}(\nu^*(X)-1)\neq\text{sign}(\nu(X)-1)}\frac{|\eta(X)-p|}{p(1-p)}\right] \\ &= \mathbb{E}\left[\mathbb{I}_{g^*(X)\neq g(X)}\frac{|\eta(X)-p|}{p(1-p)}\right],\end{aligned}$$

This concludes the first part. For the second part, it remains to note that for any real numbers (x, y)

$$\text{sign}(x-1) \neq \text{sign}(y-1) \implies |y-1| \leq |x-y|,$$

so that, using that $\nu^* = \eta^*/p$, we obtain

$$\begin{aligned}\mathcal{R}_p(g) - \mathcal{R}_p(g^*) &= \mathbb{E}\left[\mathbb{I}_{\text{sign}(\nu^*(X)-1)\neq\text{sign}(\nu(X)-1)}\frac{|\eta(X)-p|}{p(1-p)}\right] \\ &= \mathbb{E}\left[\mathbb{I}_{\text{sign}(\nu^*(X)-1)\neq\text{sign}(\nu(X)-1)}\frac{|\nu^*(X)-1|}{(1-p)}\right] \\ &\leq \frac{\mathbb{E}\left[|\nu^*(X)-\nu(X)|\right]}{1-p},\end{aligned}$$

but since $p \leq 1/2$ we obtain the desired result.

□

B Standard rates proofs

B.1 Proof of Theorem 3.2

We start with the following lemma which is a simple consequence of Theorem A.1.

Lemma B.1. *Let $z_n = \sqrt{2\log(1/\delta)/(np)}$ and suppose that $z_n \leq 1$. Then, with probability at least $1-\delta$, we have*

$$\frac{p}{\hat{p}} - 1 \leq \frac{z_n}{1-z_n} \tag{1}$$

and whenever $z_n \leq 1/2$ we obtain that $p/\hat{p} - 1 \leq 2z_n \leq 1$, with probability greater than $1-\delta$.

We have that

$$P_{n,+}(f) - P_+(f) = \frac{P_n\left((f - P_+(f))\,\mathbb{1}_{\{Y=1\}}\right)}{\hat{p}} \tag{2}$$

we focus on each term, denominator and numerator, separately. For the numerator, the term $(f - P_+(f))\,\mathbb{1}_{\{Y=1\}}$ has mean 0. In virtue of Lemma A.3, the class $(f - P_+(f))\mathbb{1}_{\{Y=1\}}$ is still bounded by $2U$ and is still of VC-type with VC parameter $(2v, A)$. As a consequence, we can use Theorem 2 in Portier (2021), stated as Theorem A.2 in the present supplementary file. The variance is bounded as follows

$$\text{Var}\left((f - P_+(f))\,\mathbb{1}_{\{Y=1\}}\right) = P\left((f - P_+(f))^2\mathbb{1}_{\{Y=1\}}\right) = \text{Var}_+(f)p = \sigma_+^2 p,$$

by definition of σ_+^2. As a consequence, Theorem A.2 gives that

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left|P_n\left((f - P_+(f))\,\mathbb{1}_{\{Y=1\}}\right)\right| &\leq K'\left(\sqrt{\frac{v\sigma_+^2 p}{n}\log\left(\frac{K'AU}{\delta\sigma_+\sqrt{p}}\right)} + \frac{Uv}{n}\log\left(\frac{K'AU}{\delta\sigma_+\sqrt{p}}\right)\right) \\ &\leq 2K'\sqrt{\frac{v\sigma_+^2 p}{n}\log\left(\frac{K'AU}{\delta\sigma_+\sqrt{p}}\right)},\end{aligned}$$

where the last inequality has been obtained using the stated condition on n and δ. For the denominator, using Lemma B.1, we have that, with probability $1-\delta$, $p \leq 2\hat{p}$, by using the condition on n and δ. Using the union bound, we get, with probability $1-2\delta$,

$$\sup_{f\in\mathcal{F}}\frac{P_n\left((f - P_+(f))\,\mathbb{1}_{\{Y=1\}}\right)}{\hat{p}} \leq 4K'\sqrt{\frac{v\sigma_+^2}{np}\log\left(\frac{K'AU}{\delta\sigma_+\sqrt{p}}\right)}$$

and the proof is complete.

B.2 Proof of Corollary 3.3

First, using the definition of $\hat{g}_{\hat{p}}$ yields

$$\mathcal{R}_{n,\hat{p}}\left(\hat{g}_{\hat{p}}\right) - \mathcal{R}_{n,\hat{p}}\left(g_p^*\right) \leq 0,$$

so that

$$\begin{aligned}
\mathcal{R}_p(\hat{g}_{\hat{p}}) - \mathcal{R}_p(g_p^*) &\leq \mathcal{R}_p(\hat{g}_{\hat{p}}) - \mathcal{R}_{n,\hat{p}}(\hat{g}_{\hat{p}}) - \left(\mathcal{R}_p(g_p^*) - \mathcal{R}_{n,\hat{p}}(g_p^*)\right) \\
&\leq 2 \sup_{g\in\mathcal{G}} \left|\mathcal{R}_p(g) - \mathcal{R}_{n,\hat{p}}(g)\right| \\
&\leq \sup_{g\in\mathcal{G}} \left|P_{n,-}(g) - P_-(g)\right| + \sup_{g\in\mathcal{G}} \left|P_{n,+}(g) - P_+(g)\right|.
\end{aligned}$$

It remains to use Theorem 3.2 twice, one time with $Y = 1$ (as stated) and one more time with $Y = -1$. The end of the proof consists in verifying that the stated bound is an upper bound for each of the two previously obtained upper bounds.

B.3 Proof of Theorem 3.4

Even though the results are different, the proof is inspired from the ones of Theorem 1 in Ausset et al. (2021) and Theorem 6 in Portier (2021). First we recall two results that will be useful in the proof. In each, we assume that (X1) (X2) and (X3) are fulfilled. The following Lemma (Portier, 2021, Lemma 4) controls the size of the k-NN balls uniformly over all $x \in S_X$.

Lemma B.2 (Portier (2021, Lemma 4))**.** *For all $n \geq 1$, $\delta \in (0,1)$ and $1 \leq k \leq n$ such that $8d\log(12n/\delta) \leq k \leq T^d n b_X c V_d/2$, it holds, with probability at least $1-\delta$:*

$$\sup_{x\in S_X} \hat{\tau}_x \leq \overline{\tau}_{n,k} := \left(\frac{2k}{nb_X cV_d}\right)^{1/d}, \tag{3}$$

where $V_d = \lambda(B(0,1))$.

The next lemma is a consequence of Theorem A.2. Let

$$\mathcal{G} = \{g(Y,X) = (\mathbb{1}_{Y=1} - \eta(X))\mathbb{I}_{\|X-x\|\leq\tau} : \tau \leq \overline{\tau}_{n,k},\, x \in \mathbb{R}^d\}$$

which is of VC-type as shown in Lemma 9 in Portier (2021) (see also Wenocur and Dudley (1981)). Because S_X is compact and η/p continuous, there exists C such that $\eta(x) \leq pC$ for all $x \in S_X$. The variance of each element in the class is bounded as

$$\mathrm{Var}(g(Y,X)) \leq E(\mathbb{1}_{Y=1}\mathbb{I}_{\|X-x\|\leq\tau}) \leq \int \eta(z)\mathbb{I}_{\|z-x\|\leq\tau} f_X(z)dz \leq CpU_X\overline{\tau}_{n,k}^d V_d.$$

Injecting the previous variance bound (which scales as pk/n) in the upper-bound given in Theorem A.2 we obtain the following statement.

Lemma B.3. *We have with probability at least $1-\delta$,*

$$\sup_{g\in\mathcal{G}} \left|\sum_{i=1}^n g(Y_i,X_i)\right| \leq C_1\left(\sqrt{kp\log\left(\frac{C_2 n}{p\delta}\right)} + \log\left(\frac{C_2 n}{p\delta}\right)\right) \tag{4}$$

where C_1 and C_2 are constants that does not depend on n, k and p but on the dimension d, the VC parameter of $\mathcal{G}$, and the probability measure P_X.

Define the event E_n as the union of (1), (3) and (4). By the previous two lemmas and Lemma B.1, using the union bound, we obtain that $P(E_n) \geq 1 - 3\delta$. In light of Borel-Cantelli Lemma we choose $\delta = 1/n^2$ so that $\sum_n(1 - P(E_n))$ is finite and the event $\liminf_n E_n$ has probability 1. It then suffices to show that E_n implies that $\hat{\eta}(x)/\hat{p} - \nu^*(x) = O(\sqrt{\log(n)/kp} + (k/n)^{1/d})$. Note that under E_n, when n is large enough, by Lemma B.1, $p/\hat{p} \leq 2$. Let $M_i = \mathbb{1}_{Y_i=1} - \eta(X_i)$ and $B_i(x) = \eta(X_i) - \eta(x)$. We have

$$\frac{\hat{\eta}(x)}{\hat{p}} - \nu^*(x) = \frac{\sum_{i\in\hat{I}(x)} M_i}{k\hat{p}} + \frac{\sum_{i\in\hat{I}(x)} B_i(x)}{k\hat{p}} + \eta(x)\left(\frac{1}{\hat{p}} - \frac{1}{p}\right). \tag{5}$$

On the event E_n, the function $(Y,X) \mapsto (\mathbb{1}_{Y=1} - \eta(X))\mathbb{1}_{\|X-x\| \le \hat{\tau}_x}$ belongs to the space $\mathcal{G}$. Consequently, the first term in (5) is smaller than

$$(k\hat{p})^{-1} \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} g(Y_i, X_i) \right|$$

which by Lemma B.1 and (4) is $O(\sqrt{\log(n)/kp})$. Using the assumption that $x \mapsto \eta(x)/p$ is L-Lipschitz we get that, on E_n, the second term in (5) is such that

$$\frac{\sum_{i \in \hat{I}(x)} B_i(x)}{k\hat{p}} \le \frac{p}{\hat{p}} L \overline{\tau}_{n,k},$$

which, using Lemma B.1, is $O((k/n)^{1/d})$. The third term in (5) is smaller than $(\eta(x)/p)\,(p/\hat{p} - 1)$ which is, using again the Lipschitz assumption and Lemma B.1 again, $O(\sqrt{\log(n)/(np)}$. The latter bound is smaller than $\sqrt{\log(n)/(kp)}$ so it does not appear in the stated bound.

C Fast rates proofs

C.1 Intermediate results

Before moving to the main proof we remind some necessary concepts and provide two technical lemmas inspired from several papers dealing with empirical processes on VC-type classes Giné and Guillou (2001); Giné and Nickl (2009). First, let us recall the definition of a *sub-root* functions.

Definition C.1. A function $\psi : [0,\infty) \to [0,\infty)$ is sub-root if it is nonnegative, nondecreasing and if $r \mapsto \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.

Let $\sigma_1, \dots, \sigma_n$ denote a collection of independent Rademacher variables, i.e., for each i, $\sigma_i \in \{-1, 1\}$ and $P(\sigma_i = 1) = 1/2$. The Rademacher variables are independent from the collection $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$. In the sequel, we will focus on the so called Rademacher complexity of functional classes $\mathcal{F}$, defined as

$$R_n(\mathcal{F}) = \frac{1}{n} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(Z_i).$$

Namely a central object in our proof will be the Rademacher complexity of classes $\mathcal{F}_q$ (differences between a given loss function and an optimal one) introduced at the beginning of Section 4. The conditional expectation given $(Z_i)_{1 \le i \le n}$ (taken with respect to the Rademacher variables $(\sigma_i)_{1 \le i \le n}$ only), will be denoted E_σ. Define

$$\mathcal{F}_{n,r} = \{f \in \mathcal{F} \,:\, P_n(f^2) \le r\}.$$

When $r > U^2$, since $P_n(f^2) \le U^2$, we have that $\mathcal{F}_{n,r} = \mathcal{F}_{n,U^2}$. Therefore we assume subsequently that $r \le U^2$. In the next lemma we derive an upper bound for $E_\sigma R_n(\mathcal{F}_r)$.

Lemma C.2. *Let $\mathcal{F}$ be a class of functions that is VC-type with envelope $U > 0$ and parameter $v, A \ge 1$. For any $r \le U^2$, it holds that, with probability 1,*

$$E_\sigma R_n(\mathcal{F}_{n,r}) \le C\sqrt{rn^{-1} v \log(eAU/\sqrt{r})}.$$

with $C = 12 \int_1^\infty s^{-2}\sqrt{1 + \log(s)}ds$.

Proof. Using Dudley's entropy integral bound, see for instance Corollary 5.25 in Van Handel (2014), one has,

$$E_\sigma \left[R_n(\mathcal{F}_{n,r})\right] \le \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}\left(\mathcal{F}_{n,r}, L_2(P_n), \epsilon\right)} d\epsilon$$

By definition of $\mathcal{F}_{n,r}$, it holds that $\mathcal{N}\left(\mathcal{F}_{n,r}, L_2(P_n), \epsilon\right) = 1$ as soon as $\epsilon \ge \sqrt{r}$. Hence, using some variable changes, we obtain

$$
\begin{aligned}
E_\sigma\left[R_n(\mathcal{F}_{n,r})\right] &\leq \frac{12}{\sqrt{n}}\int_0^{\sqrt{r}} \sqrt{\log \mathcal{N}\left(\mathcal{F}_{n,r}, L_2(P_n), \epsilon\right)}d\epsilon \\
&= \frac{12U}{\sqrt{n}}\int_0^{\sqrt{r}/U} \sqrt{\log \mathcal{N}\left(\mathcal{F}_{n,r}, L_2(P_n), U\epsilon\right)}d\epsilon \\
&\leq \frac{12U\sqrt{v}}{\sqrt{n}}\int_0^{\sqrt{r}/U} \sqrt{\log(A/\epsilon)}d\epsilon \\
&= \frac{12UA\sqrt{v}}{\sqrt{n}}\int_{AU/\sqrt{r}}^{\infty} \sqrt{\log(s)}s^{-2}ds \\
&= \frac{12\sqrt{vr}}{\sqrt{n}}\int_1^{\infty} \sqrt{\log(sAU/\sqrt{r})}s^{-2}ds
\end{aligned}
$$

Now since $A \geq 1$, we have $\log(eAU/\sqrt{r}) \geq 1$ and we can write

$$
\log(sAU/\sqrt{r}) \leq \log(eAU/\sqrt{r})(1+\log(s))
$$

which implies that

$$
\begin{aligned}
E_\sigma\left[R_n(\mathcal{F}_{n,r})\right] &\leq 12\sqrt{\frac{vr\log(eAU/\sqrt{r})}{n}}\int_1^{\infty} s^{-2}\sqrt{1+\log(s)}ds \\
&= C\sqrt{\frac{vr\log(eAU/\sqrt{r})}{n}}.
\end{aligned}
$$

□

A further similar result is now given about the class

$$
\mathcal{F}_r = \{f \in \mathcal{F} : P(f^2) \leq r\}.
$$

.

Lemma C.3. *Let $\mathcal{F}$ be a class of functions that is VC-type with envelope $U > 0$ and parameter $v, A \geq 1$. We have, for any $r \leq U^2$,*

$$
E[R_n\{\mathcal{F}_r\}] \leq C\sqrt{n^{-1}vr\log(5AU/\sqrt{r})} + 8UC^2n^{-1}v\log(5AU/\sqrt{r})
$$

where C is defined in Lemma C.2. Moreover if $r \geq n^{-1}U^2$, we obtain

$$
E[R_n\{\mathcal{F}_r\}] \leq C\sqrt{vn^{-1}r\log(5A\sqrt{n})} + 8C^2Uvn^{-1}\log(5A\sqrt{n}).
$$

Proof. First we apply Lemma C.2 with the largest possible r given by $\hat{\sigma}_n^2 = \sup_{f\in\mathcal{F}_r} P_n(f^2)$. Note that by definition $P_n(f^2) \leq \hat{\sigma}_n^2$ for all $f \in \mathcal{F}_r$, so that

$$
\mathcal{F}_r = \{f \in \mathcal{F}_r \,:\, P_n(f^2) \leq \hat{\sigma}_n^2\}
$$

and we obtain

$$
E_\sigma[R_n\{\mathcal{F}_r\}] \leq C\sqrt{\hat{\sigma}_n^2 n^{-1}v\log(eAU/\hat{\sigma}_n)}].
$$

Using twice Jensen inequality (functions $\sqrt{x}$ and $ax\log(b/x)$ are both concave), we get

$$
\begin{aligned}
E[R_n\{\mathcal{F}_r\}] &\leq C\sqrt{E[\hat{\sigma}_n^2 n^{-1}v\log(eAU/\hat{\sigma}_n)]} \\
&\leq C\sqrt{E[\hat{\sigma}_n^2]n^{-1}v\log((eAU)^2/E[\hat{\sigma}_n^2])/2} \\
&\leq C\sqrt{E[\hat{\sigma}_n^2]n^{-1}v\log(9e(AU)^2/E[\hat{\sigma}_n^2])/2}
\end{aligned}
$$

where the last point is just convenient for the increasing function property which will be used in the next few line. From Corollary 3.4 in Talagrand (1994), we obtain

$$E[\hat{\sigma}_n^2] \leq r + 8UE[R_n(\mathcal{F}_r)].$$

Now remark that $x\log(b/x)$ is increasing for $x \leq b/e$. This is always satisfied for $x = r + 8Un^{-1}E[R_n(\mathcal{F}_r)]$ because this quantity is smaller than $9U^2$ which itself is smaller or equal to b/e when $b = 9e(AU)^2$ because $A \geq 1$. We therefore obtain

$$\begin{aligned} E[R_n\{\mathcal{F}_r\}] &\leq C\sqrt{n^{-1}v(r + 8UE[R_n(\mathcal{F}_r)])\log(9e(AU)^2/(r + 8UE[R_n(\mathcal{F}_r)])]/2} \\ &\leq C\sqrt{n^{-1}v(r + 8UE[R_n(\mathcal{F}_r)])\log(9e(AU)^2/r)/2} \end{aligned}$$

Therefore

$$E[R_n(\mathcal{F}_r)]^2 \leq C^2n^{-1}v(r + 8UE[R_n(\mathcal{F}_r)])\log(3\sqrt{e}AU/\sqrt{r}) = bE[R_n(\mathcal{G})] + c$$

with $b = 8UC^2n^{-1}v\log(3\sqrt{e}AU/\sqrt{r})$ and $c = rC^2n^{-1}v\log(3\sqrt{e}AU/\sqrt{r})$. It implies that $E[R_n(\mathcal{F}_r)] \leq b + \sqrt{c}$ and the first statement follows by remarking that $3\sqrt{e} \leq 5$.

□

To proceed further, a central tool Theorem 3.3 in Bartlett et al. (2005) which we state below (Theorem C.4 with the functional $T(f) = P(f^2)$ for which $\mathrm{Var}(f) \leq T(f)$ and $T(\alpha f) \leq \alpha^2 T(f)$.

Theorem C.4. *Let $\mathcal{F}$ be a class of functions with envelope $U > 0$ and suppose that $P(f^2) \leq BPf$ for all $f \in \mathcal{F}$ for some $B \geq U$. Let ψ be a sub-root function and let $r^\star$ be the fixed point of ψ,* i.e. *$\psi(r^\star) = r^\star$. Assume that ψ satisfies, for any $r \geq r^\star$,*

$$\psi(r) \geq BE[R_n\{f \in \mathcal{F} : P(f^2) \leq r\}]$$

Then, for any $K > 1$ and every $\delta > 0$, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F} \quad Pf \leq \frac{K}{K-1}P_nf + \frac{6Kr^\star}{B} + \frac{\log(1/\delta)B\,(22 + 5K)}{n}.$$

Also, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F} \quad P_nf \leq \frac{K+1}{K}Pf + \frac{6Kr^\star}{B} + \frac{\log(1/\delta)B\,(22 + 5K)}{n}.$$

The next Proposition is key to obtain our main fast rates result Theorem 4.3.

Proposition C.5. *Let $\mathcal{F}$ be a VC-type class of functions with envelope $U > 0$ and parameters $(\tilde{A}, \tilde{v})$. Assume that $\mathcal{F}$ satisfies the Bernstein condition (B1) with constant $B > U$ relative to a probability P on $\mathcal{X} \times \mathcal{Y}$. Then with probability $1 - \delta$, for all $f \in \mathcal{F}$,*

$$P(f) \leq \frac{K}{K-1}P_n(f) + \frac{c_1BK\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{n},$$

where $c_1 > 0$ is an explicit universal constant given in the proof.

Also, with probability at least $1 - \delta$, $\forall f \in \mathcal{F}$,

$$P_n(f) \leq \frac{K+1}{K}P(f) + \frac{c_1BK\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{n}.$$

Proof. In light of the upper bound given in Lemma C.3, and in order to apply Theorem C.4, we introduce

$$\psi(r) = b\sqrt{r} + c$$

with $b = BC\sqrt{\tilde{v}n^{-1}\log(5\tilde{A}\sqrt{n})}$ and $c = 8B^2C^2n^{-1}\tilde{v}\log(5\tilde{A}\sqrt{n})$. The function ψ, defined on $\mathbb{R}_+$, is sub-root with unique fixed point r^* given by $\sqrt{r^*} = (b + \sqrt{b^2 + 4c})/2 \leq b + \sqrt{c}$. Therefore

$$r^* \leq 2(b^2 + c) \leq 18B^2C^2n^{-1}\tilde{v}\log(5\tilde{A}\sqrt{n}). \tag{6}$$

We have that $r \geq r^*$ implies that $r \geq \psi(r) \geq c \geq 8B^2C^2n^{-1}v$ which is larger than U^2n^{-1} by our assumption on the constant B. Hence the second inequality in Theorem C.3 holds true, which implies that whenever $r \geq r^\star$, we have

$$BE\left(R_n\{f \in \tilde{\mathcal{F}} : P(f^2) \leq r\}\right) \leq BC\sqrt{\tilde{v}n^{-1}r\log(5\tilde{A}\sqrt{n})} + 8BUC^2n^{-1}\tilde{v}\log(5\tilde{A}\sqrt{n}) \leq \psi(r).$$

Applying Theorem C.4 combined with the previous bound for r^*, we obtain, with probability $1-\delta$,

$$\forall f \in \mathcal{F}, \quad P(f) \leq \frac{K}{K-1}P_n(f) + \frac{BS}{n},$$

where

$$S = c_1K\tilde{v}\log(5\tilde{A}\sqrt{n}) + \log(1/\delta)(22+5K),$$

and $c_1 = 18 \times 6C^2 = 108C^2$. It remains to check that, since $K > 1$ and $v \geq 1$,

$$S \leq c_2Kv\log(5A\sqrt{n}/\delta),$$

with $c_2 = (c_1 \vee 27) = c_1$. We have established the first statement. For the second statement, the argument is similar, using this time the second statement of Theorem C.4. □

C.2 Proof of Theorem 4.3

We start by applying Proposition C.5 to the class $\mathcal{F} = \{(1-q)f_qI_+ + qf_qI_-,\ q \in (0,1), f_q \in \mathcal{F}_q\}$ which, by assumption, is of VC-type with parameters $(\tilde{A}, \tilde{v})$ and envelope $2U$. By construction, any $f \in \mathcal{F}$ writes as $f = (1-q)f_qI_+ + qf_qI_-$ for some $f_q \in \mathcal{F}_q$. The first statement of Proposition C.5 writes in this context as follows: With probability at least $1-\delta$, for all $q \in (0,1)$, for all $f_q \in \mathcal{F}_q$,

$$P\big((1-q)f_qI_+ + qf_qI_-\big) \leq \frac{K}{K-1}P_n\big((1-q)f_qI_+ + qf_qI_-\big) + \frac{c_1BK\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{n}.$$

Dividing both sides by $2q(1-q)$ yields

$$\frac{1}{2}\big(q^{-1}P(f_qI_+) + (1-q)^{-1}P(f_qI_-)\big) \leq \frac{K}{K-1}\frac{1}{2}\big(q^{-1}P_n(f_qI_+) + (1-q)^{-1}P_n(f_qI_-)\big) + \frac{c_1BK\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{2nq(q-1)},$$

which means precisely, by definition of P_q and $P_{n,q}$ (see Section 2),

$$P_qf_q \leq \frac{K}{K-1}P_{n,q} + \frac{c_1BK\tilde{v}\log(5\tilde{A}\sqrt{n}/\delta)}{2nq(q-1)}.$$

□

C.3 Proof of Lemma 4.1

Fix $q \in (0,1)$ and consider the function $\varphi(g) = P((1-q)\ell_gI_+ + q\ell_gI_-), g \in \mathcal{G}$. Notice that $\varphi(g) = 2q(1-q)\mathcal{R}_q(g)$. The function φ is a convex combination of the functions $\varphi_+(g) = P(\ell_gI_+)$ and $\varphi_-(g) = P(\ell_gI_-)$ with coefficients $(1-q, q)$. By assumption of the statement, φ_+ and φ_- are both strongly convex with respective parameters $p\lambda, (1-p)\lambda$, meaning that for all $\alpha \in (0,1)$ and g_1, g_2, it holds that for $s \in \{+,-\}$

$$\varphi_s(\alpha g_1 + (1-\alpha)g_2) \leq \alpha\varphi_s(g_1) + (1-\alpha)\varphi_s(g_2) - \frac{\alpha(\alpha-1)}{2}\pi_s\lambda\|g_1 - g_2\|^2,$$

with $\pi_+ = p$ and $\pi_- = 1-p$. By convex combination with coefficients $q, 1-q$ we obtain

$$\varphi(\alpha g_1 + (1-\alpha)g_2) \leq \alpha\varphi(g_1) + (1-\alpha)\varphi(g_2)\frac{\alpha(\alpha-1)}{2}\left((1-q)p + q(1-p)\right)\lambda\|g_1 - g_2\|^2.$$

Thus φ is λ'-strongly convex with $\lambda' = ((1-q)p + q(1-p))\,\lambda$. Now the score g_q^* is a minimizer of $\mathcal{R}_q$, whence it is also a minimizer of $\varphi = 2q(1-q)\mathcal{R}_q$. By strong convexity and because is a minimizer of φ we obtain that for any $g \in \mathcal{G}$,

$$P\big((1-q)(\ell_g - \ell_{g_q^*})I_+ + q(\ell_g - \ell_{g_q^*})I_-\big) = \varphi(g) - \varphi(g_q^*) \geq \lambda'\|g - g_q^*\|^2. \tag{7}$$

In other words for all $h \in \mathcal{H}$ of the form $h = h_{q,g} = (1-q)(\ell_g - \ell_{g_q^*})I_+ + q(\ell_g - \ell_{g_q^*})I_-$ we have

$$P(h_{q,g}) \leq \lambda'\|g - g_q^*\|^2. \tag{8}$$

On the other hand, the second assumption of the statement is that

$$\|(\ell_{g_1} - \ell_{g_2})I_s\|_{L_2(P)} \leq \sqrt{\pi_s}L\|g_1 - g_2\|,$$

Thus we may write for all $g_1, g_2 \in \mathcal{G}$,

$$\begin{aligned}\|(1-q)(\ell_{g_1} - \ell_{g_2})I_+ + q(\ell_{g_1} - \ell_{g_2})I_-\|_{L^2(P)} &\leq (1-q)\|(\ell_{g_1} - \ell_{g_2})I_+\|_{L_2(P)} + q\|(\ell_{g_1} - \ell_{g_2})I_-\|_{L_2(P)} \\ &\leq L\left((1-q)\sqrt{p} + q\sqrt{1-p}\right)\|g_1 - g_2\|.\end{aligned}$$

In other words,

$$P\Big[\big((1-q)(\ell_{g_1} - \ell_{g_2})I_+ + q(\ell_{g_1} - \ell_{g_2})I_-\big)^2\Big] \leq (L')^2\|g_1 - g_2\|^2,$$

with $L' = L\left((1-q)\sqrt{p} + q\sqrt{1-p}\right)$. Choosing $g_1 = g$ and $g = g_q^*$ yields that for any $h = h_{q,g} \in \mathcal{H}$ as above,

$$P(h_{q,g}^2) \leq L'^2\|g - g_q^*\|^2. \tag{9}$$

Combining (8) and (9), we obtain

$$P(\tilde{\ell}_{q,g} - \tilde{\ell}_{q,g_q^*})^2 \leq \frac{L'^2}{\lambda'}P(\tilde{\ell}_{q,g} - \tilde{\ell}_{q,g_q^*}).$$

Now by Jensen inequality applied to the function $t \mapsto t^2$ and the convex combination with coefficents $(1-q), q$, we have

$$(L')^2 \leq L^2\left((1-q)\sqrt{p}^2 + q\sqrt{1-p}^2\right),$$

so that $(L')^2/\lambda' \leq L^2/\lambda$, which concludes the proof.

C.4 Proof of Lemma 4.2

Write

$$\mathcal{H} = \{q(\ell_g - \ell_{g_q^*})I_+ + (1-q)(\ell_g - \ell_{g_q^*})I_-\} \subset \{qf + (1-q)g, q \in (0,1), f \in \mathcal{L}'_+, g \in \mathcal{L}'_-\} = \mathcal{F}'$$

where $\mathcal{L}'_s = \{(\ell_1 - \ell_2)I_s \;\; \ell_1, \ell_2 \in \mathcal{L}\}$. From Lemma A.3(1., 2.), each $\mathcal{L}'_s$ is of VC-type with envelope $2U$ and parameters $(2v, 2A)$. Using the 4^{th} statement of the same lemma, also $\mathcal{F}'$ is of VC type with envelope 2U and parameters $(2(2v+1), 3*2A)$.

D Numerical experiments: Real world dataset

Our aim, just as in the main paper, is to illustrate the decision boundary of the k-nn classifiers on real-world datasets. To do so, we follow the same procedure as the main paper, but instead of using synthetic data, we employ six real-world datasets (Pima, Breast, Cardio, Sattelite, Annthyroid, Ionosphere) from the ODDS repository[1]. Figures 3 to 8 display the balanced accuracy $(1-\mathcal{R}_p^{0-1})$ of the balanced k-nn as function of (k,p), we make the proportion of positive class p vary by randomly removing positive examples. Similar to the findings on synthetic data, these experiments suggest that a large number of neighbors k should be chosen relative to $p = p_n$ to ensure the consistency of the nearest neighbors method. It's important to note, however, that the learning boundary appears somewhat more noisy than in the synthetic data case. This is indeed not surprising since the number of examples available is significantly smaller in comparison to the previous simulation.

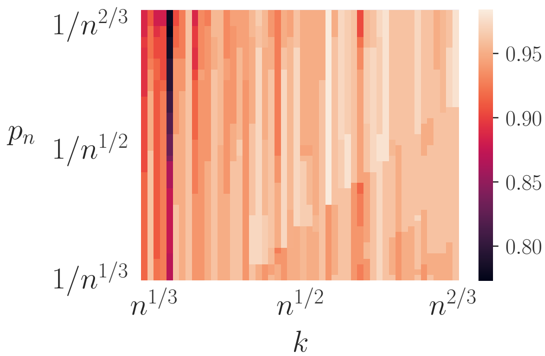


Figure 3: Balanced accuracy heat map for the Breast dataset.

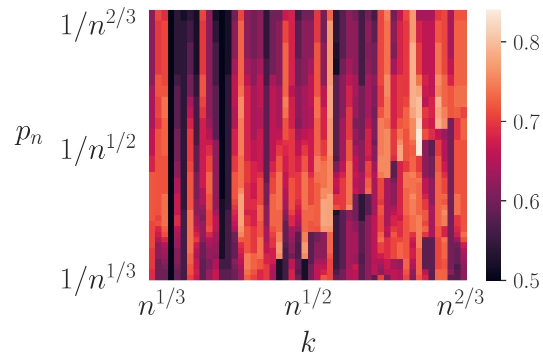


Figure 4: Balanced accuracy heat map for the Ionosphere dataset.

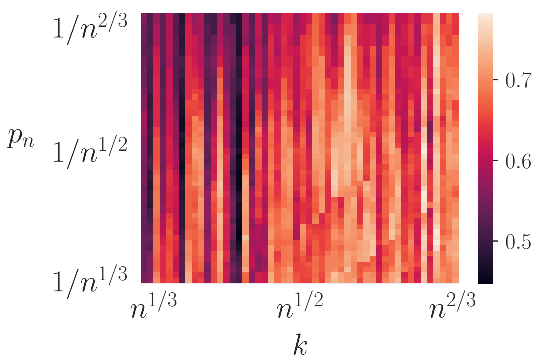


Figure 5: Balanced accuracy heat map for the Pima dataset.

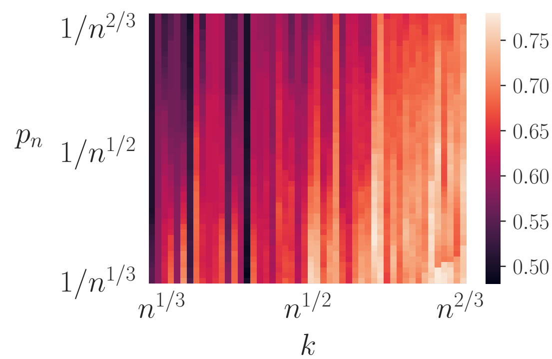


Figure 6: Balanced accuracy heat map for the Annthyroid dataset.

[1] http://odds.cs.stonybrook.edu

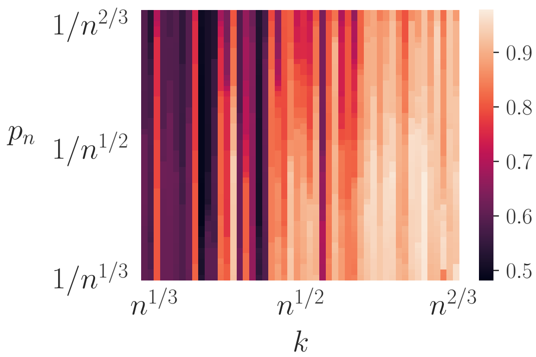


Figure 7: Balanced accuracy heat map for the Cardio dataset.

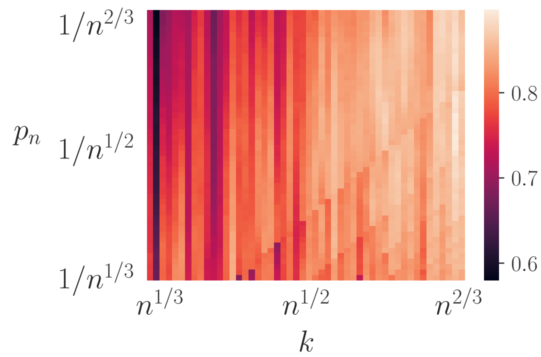


Figure 8: Balanced accuracy heat map for the Satellite dataset.