Journal of Machine Learning Research 23 (2022) 1-63 Submitted 8/21; Revised 2/22; Published 4/22

Under-bagging Nearest Neighbors for Imbalanced Classification

Hanyuan Hang H.HANG@UTWENTE.NL
Department of Applied Mathematics, University of Twente
7522 NB Enschede, The Netherlands

Yuchao Cai YCCAI@RUC.EDU.CN
School of Statistics, Renmin University of China
100872 Beijing, China

Hanfang Yang HYANG@RUC.EDU.CN
Center for Applied Statistics, School of Statistics, Renmin University of China
100872 Beijing, China

Zhouchen Lin ZLIN@PKU.EDU.CN
Key Lab. of Machine Perception (MoE), School of Artificial Intelligence, Peking University
Institute for Artificial Intelligence, Peking University, 100872 Beijing, China
Peng Cheng Laboratory, 518055 Shenzhen, Guangdong, China

Editor: Nicolas Vayatis

Abstract

In this paper, we propose an ensemble learning algorithm called *under-bagging k-nearest neighbors* (*under-bagging k-NN*) for imbalanced classification problems. On the theoretical side, by developing a new learning theory analysis, we show that with properly chosen parameters, i.e., the number of nearest neighbors k, the expected sub-sample size s, and the bagging rounds B, optimal convergence rates for under-bagging k-NN can be achieved under mild assumptions w.r.t. the arithmetic mean (AM) of recalls. Moreover, we show that with a relatively small B, the expected sub-sample size s can be much smaller than the number of training data n at each bagging round, and the number of nearest neighbors k can be reduced simultaneously, especially when the data are highly imbalanced, which leads to substantially lower time complexity and roughly the same space complexity. On the practical side, we conduct numerical experiments to verify the theoretical results on the benefits of the under-bagging technique by the promising AM performance and efficiency of our proposed algorithm.

Keywords: Imbalanced classification, under-sampling, bagging, k-nearest neighbors, ensemble learning, arithmetic mean measure, optimal convergence rates, learning theory

1. Introduction

Imbalanced classification has been encountered in multiple areas such as telecommunication managements (Babu and Ananthanarayanan, 2018; Al_Janabi and Razaq, 2019), bioinformatics (Bugnon et al., 2019; Shanab and Khoshgoftaar, 2020), fraud detection (Li et al., 2021; Somasundaram and Reddy, 2019), and medical diagnosis (Zhao et al., 2020), and has been considered one of the top ten problems in data mining research (Yang and Wu, 2006; Rastgoo et al., 2016). In fact, the ratio of the size of the majority class to the minority class can be as huge as 10^6 (Wu et al., 2008a). It is noteworthy that the imbalanced classification is emerging as an important issue in designing classifiers (Weiss, 2004; Johnson and Khoshgoftaar, 2020). Two observations account for this point: on the one hand, the imbalanced classification is pervasive in a large number of domains of great importance in the machine learning community, on the other hand, most popular classification learning algorithms are reported to be inadequate when encountering the imbalanced classification problem. These classification algorithms involve k-nearest neighbors (Wang et al., 2021), support vector machines (Raskutti and Kowalczyk, 2004; Pisner and Schnyer, 2020), random forest (Chawla et al., 2004; Sohony et al., 2018), and neural networks (Johnson and Khoshgoftaar, 2019). Typically, imbalanced learning can be categorized into two conventional approaches, namely, data level approaches and algorithm level approaches (He and Ma, 2013). The typical data level approaches are based on resampling strategies which aims to develop the under-sampling or over-sampling techniques to compensate for imbalanced distributions of the original datasets, see e.g., Wilson (1972); Kubat et al. (1997); Liu et al. (2008); Smith et al. (2014); Arefeen et al. (2020) for under-sampling and Chawla et al. (2002); Batista et al. (2003); Han et al. (2005); He et al. (2008); Yan et al. (2019); Koziarski et al. (2019) for over-sampling with artificially synthetic data generation. However, under-sampling approaches may throw away potentially useful information of the data and over-sampling strategies add the computational burden. To overcome this problem, a variety of ensemble methods combined with resampling strategies have been described in the related researches, such as Chawla et al. (2003); Seiffert et al. (2009); Galar et al. (2011); Yang et al. (2018). Different from data-level approaches that change the distribution of the original data set, algorithm level approaches try to adjust the structure of standard classifiers to diminish the effect caused by class imbalance. A typical and popular method is cost-sensitive learning which generally assigns mis-classification costs for each class appropriately, see, e.g, Puthiya Parambath et al. (2014); Zhang et al. (2018); Zhang (2020). Besides, different algorithm modifications for imbalanced classification have been recently studied in Liu et al. (2018); Montahaei et al. (2018); Di et al. (2019); Wang et al. (2020). We refer the reader to Guo et al. (2017) for a general review on imbalanced classification.

k-nearest neighbors (k-NN) is included in the top ten most significant data mining algorithms (Wu et al., 2008b), as it has always been preferred for its non-parametric working principle, and ease of implementation (Duda et al., 2001). To be specific, k-NN is a lazy method without model training since it simply tags the new data entry based learning from historical data. On the other hand, k-NN is known to be a simple yet powerful non-parametric machine learning algorithm widely used in fields such as genetics (Ayyad et al., 2019; Arowolo et al., 2020), data compression (Salvador-Meneses et al., 2019; Deng et al., 2016), economic forecasting (Malini and Pushpa, 2017; Moldagulova and Sulaiman, 2017), recommendation and rating prediction (Patro et al., 2020; Al-Ghobari et al., 2021). In fact, companies like Amazon and Netflix use k-NN when recommending books to buy and movies to watch (Ahuja et al., 2019; Belacel et al., 2020). Constant research is going on to cope with the shortcomings resulting in various improvements of the k-NN algorithm. On the one hand, the data level approaches combine resampling techniques with the standard k-NN classifier to rebalance the training data (Wilson, 1972; Zhang and Mani, 2003;

Kurniawati et al., 2018; Nwe and Lynn, 2019). Recently, by adapting the ensemble learning methods to imbalance problem, Guo et al. (2016) implemented the powerful boosting framework Adaboost as the learning model, in which k-NN is chosen as the base classifier and boosting-by-resample method is used to generate the training set. On the other hand, there has been a flurry of algorithm approaches to design specific learning algorithms with k-NN for imbalanced classification, see e.g., Tan (2005); Wang et al. (2008); Li and Zhang (2011); Dubey and Pudi (2013); Zhang et al. (2017); Liu et al. (2018); Mullick et al. (2018); Yuan et al. (2021). Recently, Zhang (2020) for the first time designed two efficient cost-sensitive k-NN classification models by changing the distance function of the standard k-NN classification. However, these methods suffer from problems including the involvement of exhaustive search, the introduction of new parameters, and significant computational overhead, which hinder the scalability and easy implementation of k-NN. In addition, the effectiveness of these methods has not been investigated from a theoretical perspective.

In this study, we propose an ensemble learning algorithm named *under-bagging k-nearest neighbors*, where the drawback of the standard k-NN classifier for imbalanced classification is eliminated with the help of the under-bagging technique. More precisely, the under-bagging k-NN classifier creates an ensemble of base predictors over bootstrap training samples independently drawn following the *under-sampling* rule. In other words, at each *bagging* round, we sample several subsets independently from the majority class such that the expected size of each subset equals to that of the minority class. After that, we build a k-NN classifier on the newly created *balanced* training set. The under-bagging k-NN classifier is thus derived based on the averaged posterior probability function. It is worth pointing out that the under-bagging k-NN classifier enjoys three advantages. Firstly, it preserves the simplicity of standard k-NN as a lazy learner without building any discriminative function from the training data, compared with other model-based classification algorithms. Secondly, the under-bagging k-NN classifier is an effective method targeting at imbalanced classification by drawing relatively balanced data subsets used at each bagging round. Last but not least, the bagging technique helps to improve the computational efficiency greatly by reducing the number of training samples at bagging rounds, while the standard k-NN is problematic to keep all the training examples in memory, so as to search for all the k nearest neighbors for a test sample (Zhang, 2020).

The main purpose of this paper is to conduct a theoretical analysis on the under-bagging k-NN for imbalanced classification from a learning theory perspective (Cucker and Zhou, 2007; Steinwart and Christmann, 2008). In the analysis, we adopt the AM measure which is shown to be more general and powerful for imbalanced classification, see e.g., Chan and Stolfo (1998); Menon et al. (2013); Guo et al. (2017); Flach (2019); Grandini et al. (2020). In fact, as Menon et al. (2013) pointed out, since the Bayes classifier w.r.t. the AM measure usually differs from that w.r.t. the classification loss, standard algorithms expecting balanced class distribution are not consistent in terms of the AM measure. Therefore, we investigate the under-bagging k-NN classifier w.r.t. the AM measure and demonstrate its effectiveness and efficiency. On the other hand, since bagging is known to reduce the variance of the base learners, in our analysis, we use Bernstein's concentration inequality (Massart, 2007; van der Vaart and Wellner, 1996; Kosorok, 2008) to establish the convergence rates, which allow for localization due to their specific dependence on the variance (Hang and Steinwart, 2017; Hang et al., 2016). In this way, our results hold in sense of "with high probability", which is more closely related to practical needs than "in expectation" and "in probability" addressed commonly in existing statistical analysis, see e.g., Hall and Samworth (2005); Biau et al. (2010); Biau and Devroye (2015).

The contributions of this paper can be stated as follows.

(i) We present the learning theory analysis on the under-sampling and under-bagging k-NN for imbalanced multi-class classification w.r.t. the AM measure. We mention that the learning theory approach distinguishes our work from previous studies. Under the Hölder smoothness assumption and the margin condition, optimal convergence rates of both under-sampling k-NN and under-bagging k-NN are established w.r.t. the AM measure with high probability. It is worth pointing out that our finite sample results demonstrate the explicit relationship among bagging rounds B, the number of nearest neighbors k, and the expected under-sampling size s.

(ii) We conduct analysis on both time and space complexity of the under-bagging k-NN with parameter selection as shown in Theorem 3. We show that under roughly the same space complexity, with the help of under-bagging, the time complexity of construction can be reduced from $\mathcal{O}(n \log n)$ (for the standard k-NN) to $\mathcal{O}((\rho n \log(\rho n))^{d/(2\alpha+d)})$, and the time complexity in the testing stage can be reduced from $\mathcal{O}((n \log n)^{2\alpha/(2\alpha+d)})$ to $\mathcal{O}(\log^2(\rho n))$, where ρ represents the imbalance ratio. These results indicate that under-bagging helps to enhance the computational efficiency, especially when the data are highly imbalanced.

(iii) We conduct numerical experiments to verify the theoretical results. We first verify the relationship among parameters k, s, and B in Theorem 3 based on synthetic datasets. Then on real datasets, we compare our under-bagging k-NN with the standard k-NN and the under-sampling k-NN. The results show that the under-bagging k-NN enjoys higher AM performance on imbalanced datasets. Moreover, the under-bagging technique can significantly reduce the running time of k-NN classifiers, which verifies our analysis on the time complexity of under-bagging k-NN. Furthermore, we can reduce the running time by using fewer sub-samples with still competitive AM performance. In addition, we also compare our under-bagging k-NN with other sub-sampling algorithms to demonstrate the effectiveness and efficiency of our method.

The rest of this paper is organized as follows. Section 2 is a warm-up section for the introduction of some notations and definitions that are related to multi-class imbalanced classification. Then we propose the under-sampling and under-bagging k-NN for imbalanced classification in Section 3. We provide basic assumptions and our main results on the convergence rates of both the under-sampling and under-bagging k-NN classifiers in Section 4. Some comments and discussions concerning the main results will also be provided in this section. The analysis on bounding error terms is presented in Section 5. We conduct numerical experiments to verify our theoretical findings for the under-bagging k-NN classifier and show the improvements of the under-bagging technique in Section 6. Besides, we conduct numerical experiments to compare our under-bagging k-NN with other sub-sampling algorithms for imbalanced classification on real-world datasets in this Section. All the proofs of Sections 4 and 5 can be found in Section 7.

2. Preliminaries

2.1 Notations

For $1 \leq p < \infty$, the L_p-norm of $x = (x_1, \ldots, x_d)$ is defined as $\|x\|_p := (|x_1|^p + \ldots + |x_d|^p)^{1/p}$, and the L_∞-norm is defined as $\|x\|_\infty := \max_{i=1,\ldots,d} |x_i|$. For any $x \in \mathbb{R}^d$ and $r > 0$, we denote $B_r(x) := B(x, r) := \{x' \in \mathbb{R}^d : \|x' - x\|_2 \leq r\}$ as the closed ball centered at x with radius r. We use the notation $a_n \lesssim b_n$ and $a_n = \mathcal{O}(b_n)$ to denote that there exists a positive constant c independent of n such that $a_n \leq c b_n$, for all $n \in \mathbb{N}$. Similarly, $a_n \gtrsim b_n$ denotes that there exists some positive constant $c \in (0, 1)$ such that $a_n \geq c^{-1} b_n$. Finally, for a set $A \subset \mathbb{R}^d$, the cardinality of A is denoted by $\#(A)$ and the indicator function on A is denoted by $\mathbf{1}_A$ or $\mathbf{1}\{A\}$.

2.2 Imbalanced Classification

For a classification problem with M classes, we observe data points $(x, y) \in \mathcal{X} \times \mathcal{Y}$ from an unknown distribution P, where x denotes the feature vector, y is the corresponding label. Assume that $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset [M] := \{1, \ldots, M\}$. Given n independently observations $D_n := \{(X_i, Y_i) : i = 1, \ldots, n\}$ drawn from P, the conditional distribution $\mathrm{P}_{Y|X}$, i.e., *posterior probability*, is defined as $\eta : \mathcal{X} \to [0, 1]^M$, where

$$\eta_m(x) = \mathrm{P}(Y = m | X = x), \qquad m = 1, \ldots, M. \tag{1}$$

To analyze the theoretical properties of the classifier, there is a need to introduce some more notations to evaluate the performance. To this end, for any measurable decision function $\psi : \mathcal{X} \to [M]$, a loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ := [0, \infty)$ defines a penalty incurred on predicting $\psi(x) \in \mathcal{Y}$ when the true label is y. The *risk* is defined by $\mathcal{R}_{L,\mathrm{P}}(\psi) := \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, \psi(x)) \, d\mathrm{P}(x, y)$. The Bayes risk, which is the smallest possible risk w.r.t. P and L, is given by $\mathcal{R}^*_{L,\mathrm{P}} := \inf\{\mathcal{R}_{L,\mathrm{P}}(\psi) \,|\, \psi : \mathcal{X} \to [M] \text{ measurable}\}$, where the infimum is taken over all measurable functions $\psi : \mathcal{X} \to [M]$. In addition, a measurable function $\psi^*_{L,\mathrm{P}}$ satisfying $\mathcal{R}_{L,\mathrm{P}}(\psi^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ is called a Bayes decision function. For example, $\psi^*_{L_{\mathrm{cl}},\mathrm{P}}(x) = \arg\max_{m \in [M]} \eta_m(x)$ is the Bayes decision function w.r.t. the *classification loss* $L_{\mathrm{cl}}(x, y, \psi(x)) := \mathbf{1}\{\psi(x) \neq y\}$. We denote $\mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}} = \mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}}(\psi^*_{L_{\mathrm{cl}},\mathrm{P}})$ as the corresponding Bayes risk and $\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}}(\psi) - \mathcal{R}^*_{L,\mathrm{P}}$ as the *classification error* for a candidate classifier. For the sake of brevity, we write $\eta^*_{L_{\mathrm{cl}},\mathrm{P}}(x) := \eta_{\psi^*_{L_{\mathrm{cl}},\mathrm{P}}(x)}(x)$ in the following.

For imbalanced classification, we need to introduce some additional notations. To this end, for $m \in [M]$, let $D_{(m)} := \{(x, y) \in D_n \,|\, y = m\}$ and $n_{(m)} := \#(D_{(m)})$. Throughout this paper, without loss of generality, we assume that $n_{(1)} \leq \ldots \leq n_{(M)}$. In this case, $D_{(1)}$ is called the *minority class*. In addition, let $\pi_m = \mathrm{P}(Y = m)$ denote the proportion of each category. Moreover, we denote $\overline{\pi} := \max_{1 \leq m \leq M} \pi_m$ and $\underline{\pi} := \min_{1 \leq m \leq M} \pi_m$. Furthermore, for $1 \leq m \leq M$, we define the *weighted* posterior probability function

$$\eta^w_m(x) = \frac{\eta_m(x)/\pi_m}{\sum_{m=1}^M \eta_m(x)/\pi_m}, \tag{2}$$

where we assign lower weights to the highly populated classes whereas assign the largest weight to the minority class. By this means, we attach the same importance of the minority class and other classes in the evaluation of the model. In particular, when the class distribution is balanced, i.e. $\pi_1 = \ldots = \pi_m$, the weighted posterior probability function (2) is the same as (1). Additionally, we define the *imbalance ratio* $\rho \in [0, 1]$ as the ratio of the minority sample size $n_{(1)}$ to the averaged sample size in each class n/M, namely,

$$\rho := Mn_{(1)}/n. \tag{3}$$

Note that when $\rho = 1$, we have $n_{(1)} = \cdots = n_{(M)} = n/M$. In this case the problem is reduced to the balanced classification. It is easy to see that the smaller ρ, the higher level of class imbalance.

However, the usual misclassification loss is ill-suited as a performance measure for imbalanced classification, since it expects an equal misclassification cost on all classes (He and Garcia, 2009). In fact, in the presence of imbalanced training data, samples of the minority class occur sparsely in the data space. As a result, given a test sample, the calculated k-nearest neighbors bear higher probabilities of samples from the other classes. Hence, test samples from the minority class are prone to be incorrectly classified (Sun et al., 2009; Leevy et al., 2018). In particular, in the highly imbalanced binary classification, a classifier can achieve good performance w.r.t. classification error by predicting all test samples to the majority class. However, this results in undesirable performance on the minority class. To tackle this problem, a variety of performance measures have been proposed for evaluating multi-class classifiers in class-imbalance settings, see e.g., Tallón-Ballesteros and Riquelme (2014); Flach (2019); Opitz and Burst (2019).

In this paper, we study the statistical convergence of the algorithms w.r.t. one such performance measure, namely the arithmetic mean of the recall (AM), which was proposed in Chan and Stolfo (1998) and recently investigated in Menon et al. (2013), which studied the consistency of algorithms proposed for imbalanced binary classification. We refer the reader to Guo et al. (2017); Flach (2019); Grandini et al. (2020) for more details. The AM measure attempts to balance the errors on classes and is shown to be an effective performance measure for evaluating classifiers in imbalanced classification. We are confined to this measure since it can be reformulated as the sum of losses on individual samples as is illustrated in Section 5 and thus is available for theoretical analysis.

For any candidate classifier $\psi : \mathcal{X} \to [M]$ and $m \in [M]$, we first consider the *recall* of the class m defined by $r_m(\psi) = \mathrm{P}(\psi(x) = m \,|\, y = m)$. Larger recall indicates better prediction of samples in the class m. In particular, $r_m(\psi) = 1$ means that every sample from the class m is predicted correctly through the classifier ψ. Then we define the AM measure as the arithmetic mean of these values, that is,

$$r_{\mathrm{AM}}(\psi) = \frac{1}{M} \sum_{m=1}^{M} r_m(\psi). \tag{4}$$

In particular, we define the optimal AM performance by $r^*_{\mathrm{AM}} := \sup\{r_{\mathrm{AM}}(\psi) \,|\, \psi : \mathcal{X} \to [M] \text{ measurable}\}$. Moreover, we define the *AM-regret* of ψ by

$$\mathfrak{R}_{\mathrm{AM}}(\psi) = r^*_{\mathrm{AM}} - r_{\mathrm{AM}}(\psi). \tag{5}$$

It is generally considered that the usual classification error and the AM regret measure the goodness-of-fit of the classifier under different settings. More specifically, when the probability distribution is approximately balanced, the usual classification error is more suited when we focus on the prediction of each sample individually regardless of which class the sample is drawn from, let alone the class distribution. By contrast, in the imbalanced classification, it is argued by Grandini et al. (2020) that the AM regret could be a more reasonable choice since it is insensitive to imbalanced class distribution and it attaches equal importance to the recall of each class. In practice, the choice of the performance measure is usually decided by the type of data encountered. In the statistics and machine learning literature, the usual classification error has been studied extensively and understood well. In this study, we focus on the analysis of AM regret for the under-bagging algorithm in the imbalanced multi-class classification, which has not yet been well studied in the literature.

3. Main Algorithm

The usual Bayes optimal classifier that minimizes the classification error is not optimal w.r.t. AM regret (Menon et al., 2013; Narasimhan et al., 2015). In fact, according to Chaudhuri and Dasgupta (2014); Xue and Kpotufe (2018), the standard k-NN classifier is showed to converge to the Bayes error, and therefore it is not consistent w.r.t. AM regret. Thus it is emerging as an important issue to design k-NN based classifiers for imbalanced classification with both solid theoretical guarantees w.r.t. the AM measure and desirable practical performance. In this paper, we first consider the k-NN classifier built on the under-sampling data, namely *under-sampling k-NN*. Moreover, to make full use of the information that might be overlooked by the under-sampling, we introduce the bagging technique and propose the *under-bagging k-NN* classifier.

Before we proceed, we introduce the under-sampling strategy and the related probability measure. Specifically, suppose that an acceptance probability function $a(x,y) \in [0,1]$ is given for every data point $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Then each observation (x_i, y_i), $i = 1, \ldots, n$, is *independently* drawn from D_n with probability $a(x_i, y_i)$. Mathematically speaking, the under-sampling strategy can be stated as follows:

(i) Sample $(X,Y) \sim \mathrm{P}$, where P denotes the distribution of the input data.

(ii) Generate $Z(X,Y)$ from the Bernoulli distribution with parameter $a(X,Y) \in (0,1]$ which will be specified in the following sections.

(iii) If $Z(X,Y) = 1$, then *accept* the candidate (X,Y). Otherwise, *reject* (X,Y) and go to the beginning.

After the strategy is repeated n times, we obtain an under-sampling dataset $D_n^u = \{(X_i,Y_i) : Z(X_i,Y_i) = 1\}$ containing those *accepted* samples. The joint probability of $\{Z(X_i,Y_i)\}_{i=1}^n$ conditional on $\{(X_i,Y_i)\}_{i=1}^n$ is denoted by P_Z.

3.1 Under-sampling k-NN Classifier

The aim of under-sampling is to create more balanced data subsets from the class-imbalanced input data set, so that the multi-class classifiers expecting balanced class distribution can be adapted to imbalanced classification. To be specific, given the minority class $D_{(1)}$ and the other classes $D_{(m)}$, $2 \leq m \leq M$, the under-sampling method randomly subsamples

$D_n^u := \{(X_1^u, Y_1^u), \ldots, (X_{s_u}^u, Y_{s_u}^u)\}$, $s_u = \#(D_n^u)$, from D_n with *acceptance probability*

$$a(x, y) = n_{(1)}/n_{(m)}, \quad \text{if } y = m. \tag{6}$$

In this case, all the samples from the minority class are in the set D_n^u.

Let $X_{(i)}^u(x)$ denote the i-th nearest neighbor of x in the sub-sampling data D_n^u w.r.t. the Euclidean distance and $Y_{(i)}^u(x)$ denote its label. We define the *posterior probability estimate* $\widehat{\eta}^{k,u} : \mathcal{X} \to [0,1]^M$ with the m-th entry by

$$\widehat{\eta}_m^{k,u}(x) := \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}\{Y_{(i)}^u(x) = m\}. \tag{7}$$

Then the *under-sampling k-NN* is given by

$$\widehat{\psi}^{k,u}(x) = \operatorname*{arg\,max}_{m \in [M]} \widehat{\eta}_m^{k,u}(x). \tag{8}$$

3.2 Under-bagging k-NN Classifier

The main drawback of under-sampling is that potentially useful information contained in the samples not appearing in D_n^u is overlooked. Thus we use the bagging technique to further exploit the samples ignored by under-sampling, that is, samples in $D_n \setminus D_n^u$. More precisely, let $1 \leq s \leq M n_{(1)}$ be the expected number of bootstrap samples, then on the b-th round of bagging, we subsample $D_b^u := \{(X_1^{b,u}, Y_1^{b,u}), \ldots, (X_{s_b}^{b,u}, Y_{s_b}^{b,u})\}$ with acceptance probability

$$a(x, y) = s/(M n_{(m)}), \quad \text{if } y = m. \tag{9}$$

Our classifier is built upon D_b^u, that is, we compute the b-th posterior probability estimate $\widehat{\eta}^{b,u} : \mathcal{X} \to [0,1]^M$ on the set D_b^u for $1 \leq b \leq B$, respectively. To be specific, the m-th entry of $\widehat{\eta}^{b,u}$ is defined by

$$\widehat{\eta}_m^{b,u}(x) := \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}\big\{Y_{(i)}^{b,u}(x) = m\big\}. \tag{10}$$

Then the average posterior probability estimate for the class m is

$$\widehat{\eta}_m^{B,u}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\eta}_m^{b,u}(x). \tag{11}$$

Finally, the *under-bagging k-NN classifier* is defined by

$$\widehat{\psi}^{B,u}(x) = \operatorname*{arg\,max}_{m \in [M]} \widehat{\eta}_m^{B,u}(x). \tag{12}$$

We summarize the under-bagging k-NN classifier in Algorithm 1.

Algorithm 1: Under-bagging k-NN Classifier for Imbalanced Classification

Input: Minority class $D_{(1)}$ and the other classes $D_{(m)}$, $2 \leq m \leq M$.
Bagging rounds B and the expected subsample size s;
Parameter $k \in \mathbb{N}_+$.

for $b = 1 \to B$ **do**
  Randomly sample D_b^u from D_n with acceptance probability chosen by (9).
  Compute the b-th posterior probability estimate $\widehat{\eta}^{b,u}$ on the set D_b^u by (10).
end
Compute the bagged posterior probability estimate $\widehat{\eta}^{B,u}$ by (11).
Output: The under-bagging k-NN classifier (12).

4. Theoretical Results and Statements

In this section, we present main results on the convergence rates of the AM regret of $\widehat{\psi}^{k,u}$ and $\widehat{\psi}^{B,u}$ under mild conditions in Section 4.1 and 4.2. Then in Section 4.3 we also present some comments and discussions on the obtained main results.

Before we proceed, we need to introduce the following restrictions on the probability distribution to characterize which properties of a distribution most influence the performance of the classifier for imbalanced classification.

Assumption 1 *We make the following assumptions on the probability measure* P.

(i) ***[Smoothness]*** *The posterior probability function* η *defined by* (1) *is assumed to be* α*-Hölder continuous with a constant* $c_L \in (0, \infty)$*, that is, for any* $x, x' \in \mathcal{X}$*, we have*

$$|\eta(x') - \eta(x)| \leq c_L \|x' - x\|^{\alpha}. \tag{13}$$

(ii) ***[Margin]*** *For any* $x \in \mathcal{X}$*, let* $\eta^w_{(m)}(x)$ *denote the* m*-th largest element in* $\{\eta^w_m(x)\}_{m=1}^M$*, where* $\eta^w_m(x)$ *is defined by* (2)*. Assume that there exists a constant* $\beta > 0$ *and* $c_\beta > 0$ *such that for all* $t > 0$*, there holds*

$$\mathrm{P}\big(|\eta^w_{(1)}(X) - \eta^w_{(2)}(X)| \leq t\big) \leq c_\beta t^{\beta}. \tag{14}$$

The α-Hölder smoothness assumption *(i)* is commonly adopted for k-nearest neighbors classification, see, e.g., Chaudhuri and Dasgupta (2014); Döring et al. (2017); Xue and Kpotufe (2018); Khim et al. (2020). In fact, since $\eta^w(x)$ is the weighted posterior probability function, η^w is α-Hölder continuous as long as η is α-Hölder continuous. Note that when α is small, the posterior probability function fluctuates more sharply, which results in the difficulty of estimating η^w accurately and thus leads to a slower convergence rates for imbalanced classification. It is worth pointing out that the smoothness assumption endows our model with a global constraints, whereas the margin assumption only reflects the behavior of η^w near the decision boundary.

The margin assumption *(ii)* quantifies how well classes are separated on the decision boundary $\partial := \{x : \eta^w_{(1)}(x) = \eta^w_{(2)}(x)\}$, which was adopted for the weighted nearest neighbors for multi-class classification (Khim et al., 2020). In particular, in the usual binary classification problems, when the probability distribution is balanced in the sense of $\pi_1 = \pi_2$, this condition coincides with Tsybakov's margin condition (Audibert and Tsybakov, 2007). We mention that the adjusted posterior probability function $\eta^w(x)$ defined by (2) attaches larger weights to the small classes. The restriction on the margin of $\eta^w(x)$ is more reasonable than $\eta(x)$ for imbalanced classification. To give a clear explanation, let us consider the binary classification problem, where the decision boundary can be expressed by $\partial G := \{x : \eta_1(x)/\pi_1 = \eta_2(x)/\pi_2\}$. In this case, compared with the ordinary boundary $\partial G_0 = \{x : \eta_1(x) = \eta_2(x)\}$, the decision boundary ∂G move towards the majority class and more points would be categorized as the minority class. From (14), we clearly see that when β is smaller, $\eta^w_{(1)}(x)$ approaches $\eta^w_{(2)}(x)$ from above more steeply, which reflects a more complex behavior around the critical threshold ∂G. In this case, the imbalanced multi-class classification becomes more difficult. In particular, for $\beta = \infty$, $\eta^w_{(1)}(x)$ is far from $\eta^w_{(2)}(x)$ with a large probability, which makes the multi-class classification significantly easier. In general, the margin assumption (ii) does not affect the the smoothness of the posterior probability functions in condition (i) and vice versa.

In the following two sections, we present main results on the convergence rates for the under-sampling and under-bagging k-NN classifier w.r.t. the AM measure of type "with high probability". It is worth pointing out that our result is built upon the techniques from the approximation theory (Cucker and Zhou, 2007) and arguments from the empirical process theory (van der Vaart and Wellner, 1996; Kosorok, 2008), which is essentially different from the previous work on the consistency of algorithms w.r.t. the AM measure (Menon et al., 2013; Narasimhan et al., 2015), where several tools such as classification-calibrated losses (Bartlett et al., 2006) and regret bounds for cost-sensitive classification (Scott, 2012) have been developed for the study.

4.1 Results on Convergence Rates for the Under-sampling k-NN Classifier

Now we present the convergence rates for the under-sampling k-NN classifier w.r.t. the AM measure under the above assumptions.

Theorem 1 *Let $\widehat{\psi}^{k,u}$ be the under-sampling k-NN classifier defined as in* (8)*, where the acceptance probability is chosen as in* (6)*. Assume that* P *satisfies Assumptions 1 and* P_X *is the uniform distribution on* $[0,1]^d$*. Then there exists an* $N_1^* \in \mathbb{N}$*, which will be specified in the proof, such that for all* $n \geq N_1^*$*, by choosing*

$$k = s_u^{2\alpha/(2\alpha+d)} (\log s_u)^{d/(2\alpha+d)} \tag{15}$$

where $s_u = \#(D_n^u)$*, there holds*

$$\mathfrak{R}_{\mathrm{AM}}(\widehat{\psi}^{k,u}) \lesssim (\log n/n)^{\alpha(\beta+1)/(2\alpha+d)} \tag{16}$$

with probability $\mathrm{P}_Z \otimes \mathrm{P}^n$ *at least* $1 - 4/n^2$*.*

Compared with the standard k-NN where k is of order $n^{2\alpha/(2\alpha+d)}$ up to a logarithm factor, in Theorem 1 we prove that k is of order $(\rho n)^{2\alpha/(2\alpha+d)}$ up to a logarithm factor when under-sampling is introduced. Especially when the data is highly imbalanced, i.e., ρ is very small, the value of k can be significantly reduced by the under-sampling technique.

The following Theorem shows that up to a logarithm factor, the convergence rate (16) of the under-sampling k-NN classifier $\widehat{\psi}^{k,u}$ is minimax optimal w.r.t. the AM regret in the case $\alpha\beta < d$.

Theorem 2 *Let $\mathcal{F}$ be the set of all measurable functions $\psi_n : (\mathbb{R}^d \times \mathbb{R})^n \times \mathbb{R}^d \to \mathbb{R}$ and $\mathcal{P}$ be the set of all probability distributions satisfying Assumption 1 with $\alpha\beta < d$. Then we have*

$$\inf_{\psi_n \in \mathcal{F}} \sup_{\mathrm{P} \in \mathcal{P}} \mathfrak{R}_{\mathrm{AM}}(\psi_n) \gtrsim n^{-\alpha(\beta+1)/(2\alpha+d)}.$$

The lower bound in Theorem 2 coincides with that for standard classification w.r.t. the classification error (Audibert and Tsybakov, 2007), although the class of probability distribution considered in Theorem 2 is different from the one considered in Audibert and Tsybakov (2007) as stated in the beginning of Section 4.

4.2 Results on Convergence Rates for the Under-bagging k-NN Classifier

We now state our main results on the convergence of the under-bagging k-NN classifier w.r.t. the AM measure.

Theorem 3 *Let $\widehat{\psi}^{B,u}(x)$ be the under-bagging k-NN classifier defined as in (12). Assume that P satisfies Assumption 1 and P_X is the uniform distribution on $[0,1]^d$. Furthermore, let ρ be the imbalance ratio defined by (3). Then there exists an $N_2^* \in \mathbb{N}$, which will be specified in the proof, such that for all $n \geq N_2^*$, by choosing*

$$s \gtrsim \begin{cases} (\rho n)^{d/(2\alpha+d)} (\log(\rho n))^{2\alpha/(2\alpha+d)}, & \text{if } d > 2\alpha, \\ (\rho n \log(\rho n))^{1/2}, & \text{if } d \leq 2\alpha, \end{cases} \tag{17}$$

$$k = s(\log(\rho n)/\rho n)^{d/(2\alpha+d)}, \tag{18}$$

$$B = k\rho n/s = (\rho n)^{2\alpha/(2\alpha+d)} \log(\rho n)^{d/(2\alpha+d)}, \tag{19}$$

there holds

$$\mathfrak{R}_{\mathrm{AM}}(\widehat{\psi}^{B,u}) \lesssim (\log n/n)^{\alpha(\beta+1)/(2\alpha+d)} \tag{20}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 5/n^2$.

Theorem 3 together with Theorem 2 implies that up to a logarithm factor, the convergence rate (20) of the under-bagging k-NN classifier $\widehat{\psi}^{B,u}$ turns out to be minimax optimal w.r.t. the AM measure, if we choose the expected sub-sample size s, the number of nearest neighbors k, and the bagging rounds B according to (17), (18) and (19), respectively. In other words, when the bagging technique is combined with the under-sampling k-NN classifier, the convergence rates of $\widehat{\psi}^{B,u}$ is not only obtainable, but also the same with that of $\widehat{\psi}^{k,u}$.

Notice that for a given dataset, (18) and (19) yield that k and B is proportional to s and k/s, respectively. Therefore, only a few independent bootstrap samples are required to obtain the estimate $\widehat{\eta}_m^{b,u}$ in (10) for the posterior probability function at each bagging round. As a result, k is reduced to $\mathcal{O}(\log(\rho n))$ in (18), instead of $\mathcal{O}((\rho n)^{2\alpha/(2\alpha+d)}(\log(\rho n))^{d/(2\alpha+d)})$ in (15) for the under-sampling k-NN.

In particular, we show in Corollary 4 that k can be further reduced to a constant order and present the convergence rates of under-bagging 1-NN classifier w.r.t. the AM measure.

Corollary 4 *Let $\widehat{\psi}^{B,u}(x)$ be the bagged 1-NN classifier defined by Algorithm 1 with $k=1$. Furthermore, assume P satisfies Assumption 1 and P_X is the uniform distribution on $[0,1]^d$. Moreover, let ρ be the imbalance ratio defined by (3). Then there exists an $N_3^* \in \mathbb{N}$, which will be specified in the proof, such that for all $n \geq N_3^*$, with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1-5/n^2$, the following two statements hold:*

(i) If $d > 2\alpha$, by choosing $s = (\rho n)^{\frac{d}{2\alpha+d}}(\log(\rho n))^{\frac{2\alpha-d}{2\alpha+d}}$ and $B = (\rho n)^{\frac{2\alpha}{2\alpha+d}}(\log(\rho n))^{\frac{d-2\alpha}{2\alpha+d}}$, we have

$$\mathfrak{R}_{\mathrm{AM}}(\widehat{\psi}^{B,u}) \lesssim (\log^2 n/n)^{\alpha(\beta+1)/(2\alpha+d)}. \tag{21}$$

(ii) If $d \leq 2\alpha$, by choosing $s = (\rho n \log(\rho n))^{1/2}$ and $B = (\rho n/\log(\rho n))^{1/2}$, we have

$$\mathfrak{R}_{\mathrm{AM}}(\widehat{\psi}^{B,u}) \lesssim \max\bigl\{(\log n/n)^{\alpha/(2d)}, (\log^3 n/n)^{1/4}\bigr\}^{\beta+1}.$$

Again, Theorem 2 yields that up to a logarithm factor, the rate (21) of under-bagging 1-NN classifier is minimax optimal when $d > 2\alpha$, $\alpha \in [0,1]$, which is usually the case.

4.3 Comments and Discussions

The section presents some comments on the obtained theoretical results on the convergence rates of the under-sampling classifier $\widehat{\psi}^{k,u}$ and $\widehat{\psi}^{B,u}$, and compares them with related findings in the literature.

4.3.1 Comments on Convergence for Imbalanced Classification

In this paper, we focus on the imbalanced classification problem. As pointed out in Sections 1 and 2.2, in the context of imbalanced classification, the AM regret is a more reasonable performance measure instead of the usual classification error. Xue and Kpotufe (2018); Khim et al. (2020) show that the standard k-NN classifier converges to the Bayes risk for multi-class classification. Therefore, in general, it can not be consistent w.r.t. the AM regret, which explains the undesirable performance of standard k-NN classification on imbalanced data from the theoretical perspective. To tackle this problem, in this study, we consider the under-sampling and under-bagging k-NN classifiers. Both of them not only retain the scalability and easy implementation of k-NN method but also have the optimal convergence rates w.r.t. the AM measure (Theorems 1 and 3). In our analysis, we only make the α-Hölder continuity and the margin assumption for the posterior probability function η. As is pointed out in Section 4, for imbalanced classification, the decision boundary is determined by the weighted posterior probability function η^w by endowing smaller classes with larger weights. Moreover, it is worth pointing out that the results are of type "with high probability" by using Bernstein's concentration inequality that takes into account the variance information of the random variables within a learning theory framework (Cucker and Zhou, 2007; Steinwart and Christmann, 2008).

As mentioned in Section 1, in the literature, despite many classifiers designed to address imbalance have been proposed, theoretical studies on these methods w.r.t. the AM measure are relatively limited. Menon et al. (2013) proved the statistical consistency of two families of algorithms for imbalanced classification, where the first family of algorithms applies a suitable threshold to a class probability estimate obtained by minimizing an appropriate strongly proper loss, and the second one minimizes a suitably weighted form of an appropriate classification-calibrated loss, i.e., the cost-sensitive learning algorithms. However, it is well-known that consistency only measures the infinite-sample property of a classifier, while its finite-sample bounds of convergence rates can hardly be guaranteed. Besides, consistency did not directly reflect any degree of regularity or smoothness of the underlying posterior probability function.

4.3.2 Comparison with Previous Bagged k-NN Algorithms and Analysis

We also compare our results with previous theoretical analysis of the k-NN algorithm combined with bagging techniques. Hall and Samworth (2005) demonstrated the consistency properties of bagged nearest neighbor classifiers to the Bayes classifier. Biau et al. (2010) studied the rate of convergence of the bagged nearest neighbor estimate w.r.t. the mean squared error. They derived the optimal rate $\mathcal{O}(n^{-2/(2+d)})$ under the assumption that the regression function is Lipschitz. Samworth (2012) regarded the bagged nearest neighbor classifier as a weighted nearest neighbor classifier, and showed that the "infinite simulation" case of bagged nearest neighbors (with infinite bagging rounds) can attain the optimal convergence rate. It is worth pointing out that our analysis of the under-bagging k-NN presents in this study is essentially different from that in the previous works.

First of all, we highlight that different from previous statistical analysis, our theoretical analysis is conducted from a learning theory perspective (Cucker and Zhou, 2007; Steinwart and Christmann, 2008) using techniques such as approximation theory and empirical process theory (van der Vaart and Wellner, 1996; Kosorok, 2008).

Secondly, previous works only take into account the uniform resampling method based on the 1-NN classifier, where the weights of the bagged estimate have an explicit probability distribution, whereas our work aims at providing a theoretical analysis of the under-sampling k-NN algorithm (Theorem 1). To this end, we have to explore the more complex Generalized Pascal distribution (Section 7.2.2).

Thirdly, previous works consider the "infinite simulation" case of bagged k-NN when the number of bagging round $B \to \infty$, where the results fail to explain the success of bagging with finite resampling times in practice. By contrast, we provide results of convergence rate with finite B by exploiting arguments such as Bernstein's concentration inequality from the empirical process theory, which enable us to derive the relationships among the number of bagging rounds B, the number of nearest neighbors k and the expected sub-sample size s (Theorem 3). Moreover, (19) implies that $B = O((\rho n)^{2\alpha/(2\alpha+d)}(\log(\rho n)^{d/(2\alpha+d)}))$, which is relatively small especially when ρ is small and when d is large.

Last but not least, results in Biau et al. (2010) hold "in expectation" w.r.t. both the resampling distribution and input data, and results in Samworth (2012) hold "in probability", while results in our study hold "with high probability", which is a stronger claim since it gives us a confidence about how well the method has learned for a given data set D of fixed size n (Steinwart and Christmann, 2008). In other words, Theorem 1 and 3 imply that for most datasets sampled from P^n the classifiers $\widehat{\psi}^{k,u}$ and $\widehat{\psi}^{B,u}$ have an almost optimal performance whenever n is large.

4.3.3 Comments on Complexity

As a commonly-used algorithm, k-d tree (Bentley, 1975) is used to search the nearest neighbors in NN-based methods. In what follows, we show that under-bagging helps reducing the time complexity of both the construction and search stages, whereas maintaining roughly the same space complexity.

Friedman et al. (1977) shows that k-d tree has a time complexity $\mathcal{O}(nd\log n)$ and a space complexity $\mathcal{O}(nd)$ for the tree construction. By Theorem 3, it suffices to choose $B = O((\rho n)^{2\alpha/(2\alpha+d)}(\log(\rho n)^{d/(2\alpha+d)}))$ when $s = (\rho n)^{d/(2\alpha+d)}(\log(\rho n))^{2\alpha/(2\alpha+d)}$. Therefore, compared with the standard k-NN whose complexity is $\mathcal{O}(nd\log n)$, the time complexity of construction the k-d tree in our algorithm can be reduced to $\mathcal{O}((\rho n)^{d/(2\alpha+d)}d\log(\rho n))$ with parallel computing fully employed. Considering the bagging rounds B, the space complexity of our algorithm turns out to be $\mathcal{O}(Bsd) = \mathcal{O}(k\rho nd) = \mathcal{O}(\rho n\log(\rho n)d)$, whereas the space complexity of the standard k-NN is $\mathcal{O}(nd)$.

In the search of the k-th nearest neighbor for a test sample, the time complexity is $\mathcal{O}(k\log n)$ (Friedman et al., 1977). For the standard k-NN, since the number of nearest neighbors is $\mathcal{O}(n^{2\alpha/(2\alpha+d)})$ (Chaudhuri and Dasgupta, 2014; Zhao and Lai, 2021), the time complexity of the search stage turns out to be $\mathcal{O}(n^{2\alpha/(2\alpha+d)}\log n))$. According to Theorem 3, thanks to the under-sampling technique, for each base learner, we merely need to search $\mathcal{O}(\log(\rho n))$ neighbors among $s = (\rho n)^{d/(2\alpha+d)}(\log(\rho n))^{2\alpha/(2\alpha+d)}$ samples. Thus, the time complexity of the search stage can be reduced to $\mathcal{O}(\log^2(\rho n))$.

In summary, the bagging technique can enhance the computational efficiency to a considerable amount when parallel computation is fully employed. When the dimension gets higher, we typically require more samples in the input space, i.e., larger n, and thus an algorithm requires more time. This phenomenon is often referred to as the curse of dimensionality. We mention that the under-bagging technique can actually alleviate this problem by enjoying smaller time complexity. Furthermore, by adopting the under-sampling rule, the expected number of samples in each class is equal to the sample size of the minority class, and thus the size of training samples at each bagging round can be greatly reduced when the data distribution is highly imbalanced, reflected in a very small value of ρ.

5. Error Analysis

In this section, we conduct error analysis for the under-sampling and under-bagging k-NN classifier respectively by establishing its convergence rates, which are stated in the above section in terms of the AM measure. The downside of using the AM measure is that it does not admit an exact bias-variance decomposition and the usual techniques for classification error estimation may not apply directly. Nonetheless, if we introduce the *balanced* version of the classification loss,

$$L_{\mathrm{bal}}(x, y, \psi(x)) = L_{\mathrm{cl}}(x, m, \psi(x))/(M\pi_m) = \mathbf{1}\{\psi(x) \neq m\}/(M\pi_m), \qquad \text{if } y = m. \tag{22}$$

where a wrong classification of an instance from the minority class is punished stronger than a wrong classification of an instance from the majority class, we are able to reduce the problem of analyzing the AM regret to the problem of analyzing the expectation or sum of a loss on individual samples. In fact, the balanced loss is a useful tool to study the statistical consistency of algorithms for ranking and imbalanced classification (Kotlowski et al., 2011; Menon et al., 2013). According to Proposition 6 in Narasimhan et al. (2015), the Bayes classifier w.r.t. the balanced loss can be expressed as

$$\begin{aligned}\psi^*_{L_{\mathrm{bal}},\mathrm{P}}(x) &= \operatorname*{arg\,min}_{m\in[M]} \sum_{i=1}^M \eta_i(x) L_{\mathrm{bal}}(x,m,i) = \operatorname*{arg\,min}_{m\in[M]} \sum_{i=1}^M \eta_i(x) \boldsymbol{1}\{i \neq m\}/\pi_i \\ &= \operatorname*{arg\,min}_{m\in[M]} \biggl(\sum_{i=1}^M \eta_i(x)/\pi_i - \eta_m(x)/\pi_m \biggr) = \operatorname*{arg\,max}_{m\in[M]} \eta^w_m(x). \end{aligned} \tag{23}$$

For brevity, we write $\eta^*_{L_{\mathrm{bal}},\mathrm{P}}(x) := \eta_{\psi^*_{L_{\mathrm{bal}},\mathrm{P}}(x)}(x)$ in the following. From (23) we see that the Bayes classifier in terms of the balanced loss depends on the weighted posterior probability function η^w instead of η. To explain, let us consider the binary classification problem $\psi : \mathcal{X} \to \{+1, -1\}$ with $\pi_{+1} \geq \pi_{-1}$. Then (23) takes the following form

$$\begin{aligned}\psi^*_{L_{\mathrm{bal}},\mathrm{P}}(x) &= \operatorname{sign}\bigl(\eta_{+1}(x)/\pi_{+1} - \eta_{-1}(x)/\pi_{-1}\bigr) \\ &= \operatorname{sign}\bigl(\eta_{+1}(x)/\pi_{+1} - (1-\eta_{+1}(x))/\pi_{-1}\bigr) = \operatorname{sign}(\eta_{+1}(x) - \pi_{+1}),\end{aligned}$$

where $\operatorname{sign}(x) = 1$ if $x > 0$ and $\operatorname{sign}(x) = -1$ otherwise. It is easy to see that the decision boundary changes from $1/2$ for usual classification to π_{+1}, which expands the region where the prediction is the minority class. In particular, if $\pi_m = 1/M$ for $1 \leq m \leq M$, then we have $L_{\mathrm{bal}}(x, y, \psi(x)) = L_{\mathrm{cl}}(x, y, \psi(x))$, then the balanced loss is equal to the classification loss, which leads to the same Bayes classifier.

In what follows, Proposition 5, Theorem 6 and Proposition 7 tell us how to reduce the problem of bounding the AM regret to the problem of bounding the estimation error of the posterior probability function, which supplies the key to the proof of both under-sampling and under-bagging k-NN classifier. The proofs of these propositions are given in Section 7.

We start with Proposition 5 which indicates that to analyze the AM-regret of a classifier, it suffices to analyze its balance risk.

Proposition 5 *For any classifier* $\psi : \mathcal{X} \to [M]$*, we have* $r_{\mathrm{AM}}(\psi) = 1 - \mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\psi)$.

The above proposition directly yields that $r^*_{\mathrm{AM}} = 1 - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}}$, which implies that the AM-regret is equal to the excess balanced error. As a result, the AM-regret in (5) can be re-expressed as

$$\mathfrak{R}_{\mathrm{AM}}(\psi) = \mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\psi) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}}. \tag{24}$$

To bound the right-hand side of (24), the main idea here is to build a new probability distribution to convert the excess balanced error into the excess classification error so that the approximation theory and the Bernstein's concentration inequality for multi-class classification can be applied. To this end, note that (23) implies that the Bayes classifier in terms of the balanced loss depends on η^w, which inspires us to consider a new probability distribution P^w with the posterior probability function η^w. To be specific, let $\mathrm{P}(X, Y)$ be the probability distribution of the samples, then we define the *balanced* probability distribution $\mathrm{P}^w(X, Y)$ whose marginal distribution satisfies

$$\pi^w_m := \mathrm{P}^w(Y = m) = 1/M \qquad \text{for } 1 \leq m \leq M, \tag{25}$$

and the conditional density function satisfies

$$f^w(x|Y=m) = \eta_m(x) f_X(x)/\pi_m. \tag{26}$$

Consequently, combining (25) and (26), we obtain the marginal density $f_X^w(x)$ given by

$$f_X^w(x) = \sum_{m=1}^{M} \pi_m^w f^w(x|y=m) = \sum_{m=1}^{M} \eta_m(x) f_X(x)/(M\pi_m). \tag{27}$$

Thus, for the probability measure $\mathrm{P}^w(x,y)$, the Bayes classifier w.r.t. the classification loss is

$$\psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x) = \operatorname*{arg\,max}_{m\in[M]} \eta_m^w(x) = \psi^*_{L_{\mathrm{bal}},\mathrm{P}}(x). \tag{28}$$

With these preparations, we present the next theorem showing the equivalence between the excess balanced error w.r.t. P and the excess classification error w.r.t. P^w defined as above, which supplies the key to the proof of the convergence rates of the candidate classifier w.r.t. the AM measure.

Theorem 6 *Let P^w be the probability measure defined by (25) and (26). Then for any classifier $\psi : \mathcal{X} \to \mathcal{Y}$, we have $\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\psi) - \mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} = \mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\psi) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w}$.*

Combining Theorem 6 with (24), we see that the standard techniques can also be applied to analyzing the classification error w.r.t. the balanced probability distribution P^w to derive the convergence rates of AM regret.

The following Lemma enables us to reduce the problem of bounding the excess multi-class classification error to the problem of bounding the estimation error of the posterior probability function.

Proposition 7 *Let $\widehat{\eta} : \mathcal{X} \to [0,1]^M$ be an estimate of η^w and $\widehat{\psi}(x) = \operatorname{arg\,max}_{m\in[M]} \widehat{\eta}_m(x)$. If $(\mathrm{P}^w)^n(\|\widehat{\eta}(X) - \eta^w(X)\|_\infty \le \phi_n) \ge 1-\delta$, where (ϕ_n) is a positive sequence, then with probability $(\mathrm{P}^w)^n$ at least $1-\delta$, there holds $\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\widehat{\psi}) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} \le c_\beta (2\phi_n)^{\beta+1}$.*

Therefore, to further our analysis, we first need to bound the L_∞-distance $\|\widehat{\eta}^{k,u} - \eta^w\|_\infty$ and $\|\widehat{\eta}^{B,u} - \eta^w\|_\infty$, where $\widehat{\eta}^{k,u}$ and $\widehat{\eta}^{B,u}$ are the posterior probability function estimator for the under-sampling and under-bagging k-NN defined by (7) and (11), respectively. We present the error analysis for the under-sampling and under-bagging k-NN classifiers in the following two sections.

5.1 Analysis for the Under-Sampling k-NN Classifier

In this Section, we conduct error analysis for the under-sampling k-NN classifier by establishing error decomposition for the term $\|\widehat{\eta}^{k,u}(x) - \eta^w(x)\|_\infty$ in (32). Then in Sections 5.1.1-5.1.3, we bound the sample error, the approximation error and the under-sampling error respectively. These three terms play an essential role in establishing the convergence rates of the under-sampling k-NN classifier and the minimax lower bound as stated in Theorems 1 and 2 in Section 4.1. All the proofs related to Sections 5.1.1-5.1.3 are given in Section 7.1.

The downside of under-sampling strategy is that it changes the probability distribution of the training data, that is, the under-sampling subset D_n^u in Section 3.1 dose not have the distribution P. In the sequel, let P^u denote the probability distribution of the accepted samples by the under-sampling strategy discussed in Section 3.1 and $\eta^u(x)$ be the corresponding posterior probability function. By Lemma 19 in Section 7.1.1, for $1 \leq m \leq M$, $\eta_m^u(x)$ can be expressed as

$$\eta_m^u(x) = \frac{\eta_m(x)/n_{(m)}}{\sum_{m=1}^M \eta_m(x)/n_{(m)}}. \tag{29}$$

It thus follows that

$$\|\widehat{\eta}^{k,u}(x) - \eta^w(x)\|_\infty \leq \|\widehat{\eta}^{k,u}(x) - \eta^u(x)\|_\infty + \|\eta^u(x) - \eta^w(x)\|_\infty. \tag{30}$$

It is easy to see that the first term of the right-hand side of (30) represents the error for applying k-NN on the subset D_n^u and thus it admits the usual decomposition for error estimation whereas the second term, namely *under-sampling error* is brought about by the under-sampling strategy from the training data. To bound the first term $\|\widehat{\eta}^{k,u} - \eta^u\|_\infty$, we need to define $\overline{\eta}^{k,u} : \mathcal{X} \to [0,1]^M$, where its m-th entry

$$\overline{\eta}_m^{k,u}(x) = \mathbb{E}[\widehat{\eta}_m^{k,u}(x)|D_n^u] = \frac{1}{k}\sum_{i=1}^k \eta_m^u(X_{(i)}^u(x)). \tag{31}$$

In other words, $\overline{\eta}^{k,u}$ denotes the conditionally expectation of $\widehat{\eta}^{k,u}$ on the under-sampling data D_n^u. Thus we obtain the error decomposition for the posterior probability function w.r.t. the under-sampling k-NN classifier as follows:

$$\begin{aligned}&\|\widehat{\eta}^{k,u}(x) - \eta^w(x)\|_\infty \\ &\leq \|\eta^u(x) - \eta^w(x)\|_\infty + \|\widehat{\eta}^{k,u}(x) - \overline{\eta}^{k,u}(x)\|_\infty + \|\overline{\eta}^{k,u}(x) - \eta^u(x)\|_\infty.\end{aligned} \tag{32}$$

Apart from the under-sampling error mentioned above, the second term on the right-hand side of (32) is called the *sample error* since it is associated with the empirical measure D_n^u and the last term of (32) is called *approximation error* since it indicates how the error is propagated by the under-sampling k-NN algorithm.

5.1.1 Bounding the Sample Error Term

We now establish the oracle inequality for the under-sampling posterior probability function $\widehat{\eta}^{k,u}$ under L_∞-norm. This oracle inequality will be crucial in establishing the convergence results of the estimator.

Proposition 8 *Let $\widehat{\eta}^{k,u}$ and $\overline{\eta}^{k,u}$ be defined by* (7) *and* (31)*, respectively. Then there exists an $N_1 \in \mathbb{N}$, which will be specified in the proof, such that for all $n > N_1$, with probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - 1/n^2$, there holds*

$$\|\widehat{\eta}^{k,u}(x) - \overline{\eta}^{k,u}(x)\|_\infty \lesssim \sqrt{\log s_u / k}. \tag{33}$$

5.1.2 Bounding the Approximation Error Term

The result on bounding the deterministic error term shows that the L_∞-distance between $\overline{\eta}^{k,u}$ and η^u can be small by choosing k appropriately.

Proposition 9 *Let $\widehat{\psi}^{k,u}$ be the under-sampling k-NN classifier defined by* (8)*. Assume that P_X is the uniform distribution on $[0,1]^d$ and Assumption 1 is satisfied. Then there exists an $N_2 \in \mathbb{N}$, which will be specified in the proof, such that for all $n \geq N_2$, there holds*

$$\|\overline{\eta}^{k,u}(x) - \eta^u(x)\|_\infty \lesssim (k/s_u)^{\alpha/d}$$

with probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - 1/n^2$.

5.1.3 Bounding the Under-sampling Error Term

The next proposition shows that the L_∞-norm distance between η^u and η^w, which possess a polynomial decay w.r.t. the number of the training data. The following result is crucial in our subsequent analysis on the converge rates of both under-sampling and under-bagging k-NN classifier.

Proposition 10 *Let $\eta^w_m(x)$ and $\eta^u_m(x)$ be defined by* (2) *and* (29) *respectively. Then there exists an $N_3 \in \mathbb{N}$, which will be specified in the proof, such that for all $n \geq N_3$, there holds*

$$\|\eta^u(x) - \eta^w(x)\|_\infty \lesssim \sqrt{\log n/n} \tag{34}$$

with probability P^n at least $1 - 1/n^2$.

5.2 Analysis for the Under-bagging k-NN Classifier

In this Section, we proceed with the estimation of the posterior probability error term $\|\widehat{\eta}^{B,u}(x) - \eta^w(x)\|_\infty$ for the under-bagging k-NN classifier by establishing error decomposition for $\|\widehat{\eta}^{B,u}(x) - \eta^w(x)\|_\infty$ in (39). Then in Sections 5.2.1-5.2.3, we bound the bagging error, the bagged approximation error, and the bagged sample error, respectively. Together with the bound of the under-sampling error in Section 5.1.3, we are able to derive the convergence rates of the under-bagging k-NN classifier as stated in Theorem 3 and Corollary 4 in Section 4.2. All the proofs related to Sections 5.2.1-5.2.3 are given in Section 7.2.

We first show that the under-bagging k-NN classifier can be re-expressed as a weighted k-NN, which is amenable to statistical analysis. To be specific, let $X_{(i)}(x)$ be the i-th nearest neighbor of x in D_n w.r.t. the Euclidean distance and $Y_{(i)}(x)$ denote its label. Then for $1 \leq b \leq B$, we re-express the posterior probability estimate $\widehat{\eta}^{b,u} : \mathcal{X} \to [0,1]^M$ on the under-sampling set D^u_b with its m-th entry defined by $\widehat{\eta}^{b,u}_m(x) = \sum_{i=1}^n V_i^{b,u}(x) \boldsymbol{1}\{Y_{(i)}(x) = m\}$. Here, $V_i^{b,u}(x)$ equals $1/k$ if $\sum_{j=1}^i Z^b(X_{(j)}(x), Y_{(j)}(x)) \leq k$ and 0 otherwise, where $Z^b(x,y)$, $1 \leq b \leq B$, are i.i.d. Bernoulli random variables with parameter $a(x,y)$. Then the posterior probability estimate (11) can be re-expressed as

$$\widehat{\eta}^{B,u}_m(x) = \frac{1}{B}\sum_{b=1}^B \widehat{\eta}^{b,u}_m(x) = \frac{1}{B}\sum_{b=1}^B \sum_{i=1}^n V_i^{b,u}(x) \boldsymbol{1}\{Y_{(i)}(x) = m\}. \tag{35}$$

To bound $\|\widehat{\eta}^{B,u}(x)-\eta^w(x)\|_\infty$, we need to consider the bagged posterior probability function estimator, that is, we repeat under-sampling an infinite number of times, and take the average of the individual outcomes. To be specific, we define $\widetilde{\eta}^{B,u}:\mathcal{X}\to[0,1]^M$, with the m-th entry

$$\widetilde{\eta}_m^{B,u}(x)=\mathbb{E}_{\mathrm{P}_Z}^B[\widehat{\eta}_m^{B,u}(x)|\{(X_i,Y_i)\}_{i=1}^n]:=\sum_{i=1}^n \overline{V}_i^u(x)\mathbf{1}\{Y_{(i)}(x)=m\}, \tag{36}$$

where

$$\overline{V}_i^u(x)=\mathbb{E}_{\mathrm{P}_Z}\big[V_i^{b,u}(x)\big|\{(X_i,Y_i)\}_{i=1}^n\big]. \tag{37}$$

In fact, for any $x\in\mathcal{X}$, by the law of large numbers, we have $\widehat{\eta}_m^{B,u}(x)\to\widetilde{\eta}_m^{B,u}(x)$ almost surely as $B\to\infty$. Then we define the population version of the bagged estimator $\widetilde{\eta}_m^{B,u}(x)$ as follows:

$$\overline{\eta}_m^{B,u}(x)=\mathbb{E}[\widetilde{\eta}_m^{B,u}|X_1,\ldots,X_n]=\sum_{i=1}^n \overline{V}_i^u(x)\eta_m^u(X_{(i)}(x)), \tag{38}$$

where the conditional expectation is taken w.r.t. $(\mathrm{P}^u)_{Y|X}^n$. With these preparations, we are able to make the following error decomposition:

$$\begin{aligned}\|\widehat{\eta}^{B,u}(x)-\eta^w(x)\|_\infty &\leq \|\widehat{\eta}^{B,u}(x)-\widetilde{\eta}^{B,u}(x)\|_\infty+\|\overline{\eta}^{B,u}(x)-\widetilde{\eta}^{B,u}(x)\|_\infty\\ &\quad+\|\overline{\eta}^{B,u}(x)-\eta^u(x)\|_\infty+\|\eta^u(x)-\eta^w(x)\|_\infty.\end{aligned} \tag{39}$$

Compared with the analysis for the under-sampling k-NN classifier in (32), there are four terms on the right hand side of (39). Since we are not able to repeat the sampling strategy an infinite number of times, the bagging procedure brings about the error term $\|\widehat{\eta}^{B,u}-\widetilde{\eta}^{B,u}\|_\infty$, which is called *bagging error* in what follows. In addition, the second and the third term on the right hand-side of (39) can be viewed as the *bagged sample error* and the *bagged approximation error* for the bagged posterior probability function estimator $\widetilde{\eta}_m^{B,u}$ by similar arguments in the analysis for under-sampling k-NN classifier. Finally, the last term of (39) is the *under-sampling error* as mentioned in (32).

5.2.1 Bounding the Bagging Error Term

The next proposition shows that the bagging error term can be bounded in term of the number of bagging rounds B.

Proposition 11 *Let $\widehat{\eta}^{B,u}$ and $\widetilde{\eta}^{B,u}$ be defined by* (35) *and* (36)*, respectively. Suppose that $9B\geq 2(2d+3)\log n$. Then there exists an $N_4\in\mathbb{N}$, which will be specified in the proof, such that for all $n\geq N_4$, there holds*

$$\|\widetilde{\eta}^{B,u}(x)-\widehat{\eta}^{B,u}(x)\|_\infty\lesssim\sqrt{\log n/B}$$

with probability $\mathrm{P}_Z^B\otimes\mathrm{P}^n$ at least $1-1/n^2$.

5.2.2 Bounding the Bagged Approximation Error Term

We now show that the L_∞ distance between $\overline{\eta}^{B,u}$ and η^u can be bounded by two terms. The first term is determined by the ratio k/s and the smoothness of the posterior probability function whereas the second term results from the under-sampling strategy, which possess an exponential decay w.r.t. $(s/n)^2$.

Proposition 12 *Let $\overline{\eta}^{B,u}$ be defined by (38) with $k \geq \lceil 48(2d+9)\log n \rceil$ and η^u be defined by (29). Assume that P satisfies Assumption 1 and P_X is the uniform distribution on $[0,1]^d$. Moreover, suppose that $s\exp(-(s/M-k)^2/(2n)) \leq M\underline{\pi}/2$. Then there exists an $N_5 \in \mathbb{N}$, which will be specified in the proof, such that for all $n \geq N_5$, there holds*

$$\|\overline{\eta}^{B,u}(x) - \eta^u(x)\|_\infty \lesssim (k/s)^{\alpha/d} + \exp\bigl(-(s-k)^2/(2n)\bigr)$$

with probability P^n not less than $1 - 1/n^2$.

5.2.3 Bounding the Bagged Sample Error Term

We now establish the oracle inequality for the bagged posterior probability function $\widetilde{\eta}^{B,u}$ in terms of L_∞ norm. The oracle inequality will be crucial in establishing the convergence results for the under-bagging k-NN classifier.

Proposition 13 *Let $\widetilde{\eta}^{B,u}$ and $\overline{\eta}^{B,u}$ be defined by (36) and (38), respectively. Then there exists an $N_6 \in \mathbb{N}$, which will be specified in the proof, such that for all $n \geq N_6$, there holds*

$$\|\widetilde{\eta}^{B,u}(x) - \overline{\eta}^{B,u}(x)\|_\infty \lesssim \sqrt{s\log n/(kMn_{(1)})}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 1/n^2$.

6. Experiments

6.1 Performance Evaluation Metrics

Compared with typical supervised learning, imbalanced learning pays more attention to the classification performance of the minority classes. Therefore, instead of the overall accuracy, we use the AM measure defined by (4) to evaluate the performance of different classifiers.

6.2 Hyper-parameter Analysis

There are three hyper-parameters in the under-bagging k-NN for imbalanced classification: the bagging rounds B, the number of nearest neighbors k, and the expectation of subsample size s. Before we conduct parameter through synthetic experiments, we first introduce the data generation procedure. We use a simple toy dataset containing two interleaving half circles. In detail, we use two half moon functions with Gaussian noises added to synthesize the samples, where each half represents an unique class. The standard deviation of the Gaussian noises is 0.2. We generate $20,000$ positive samples and 200 negative samples in each run for training, and $200,000$ positive samples and $2,000$ negative samples for testing. We repeat the synthetic experiments for 100 times and record the averaged recall score. One

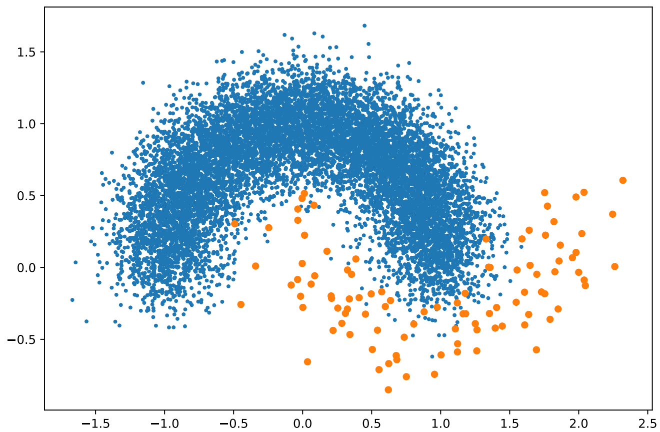

Figure 1: One Visualization of the Generated Synthetic Dataset

visualization of synthetic dataset is shown in Figure 1, with the ratio of the majority class (marked in blue) and the minority class (marked in orange) 100 : 1.

Parameter Analysis on B and k. To study these two hyper-parameters B and k, we fix the expected sub-sample size $s = Mn_{(1)}$, and vary the bagging rounds $B \in \{1, 2, 5, 10, 20, 50\}$ and the number of neighbors $k \in \{1, 2, \ldots, 30\}$. The averaged AM measure among different B and k are shown in Figure 2. As expected, we can choose a sufficient large B and the optimal number of neighbors k for a good performance. Moreover, as the bagging rounds B increases, the performance of a bagging classifier is robust under a wide range of hyper-parameter k, which reduces the difficulty of selecting the optimal hyper-parameter k. In addition, from the practical perspective, we can use a relatively large B to achieve a high AM measure performance with a low running time, since we can easily save running time under the parallelism of bagging rounds B.

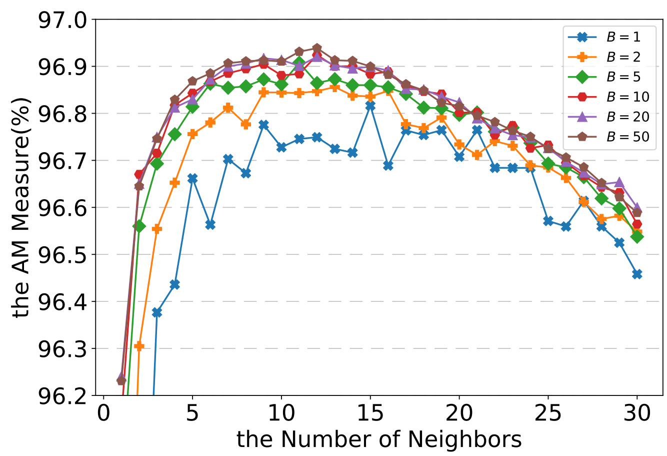

Figure 2: Parameter Analysis between B and k.

Parameter Analysis on the Expected Sub-Sample Size s. To study the empirical performance of under-bagging k-NN with different expected sub-sample size s, we fix the bagging rounds $B = 20$, and explore the performance under different sub-sample size $s = aMn_{(1)}$ with $a \in \{0.2, 0.4, \ldots, 1.0\}$. In fact, according to (6), a represents the acceptance probability of the samples in the minority class. As is shown in Figure 3, as the expected sub-sample size becomes larger, the results of the AM measure become better, since more samples are taken into consideration in each round of bagging. However, the difference in the best AM measure gets smaller while the expected sub-sample size is close to $Mn_{(1)}$, which means that when coping with massive imbalanced data, under-bagging k-NN achieves competitive empirical performance with a relatively small sub-sample size s w.r.t. n. Moreover, the optimal number of nearest neighbors k required reduces with the expected sub-sample size s, which coincides with the theoretical results that $k = s(\log(\rho n)/\rho n)^{d/(2\alpha+d)}$ in Theorem 3. The reduction of a and k results in a higher computational efficiency.

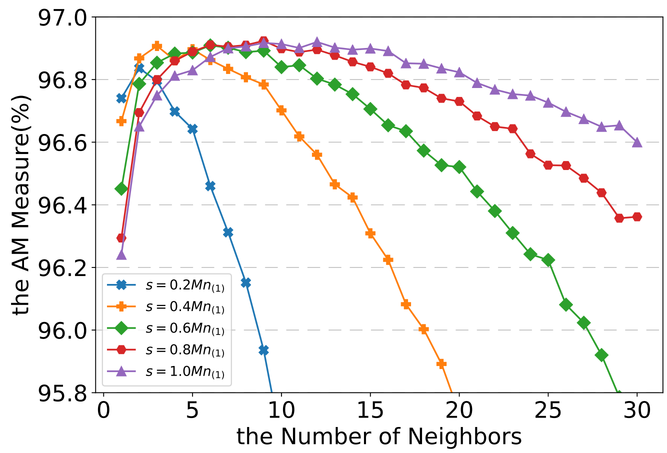

Figure 3: Parameter analysis of the expected sub-sample size s.

6.3 Numerical Comparison on Real-world Imbalanced Datasets

6.3.1 Experimental Settings

To verify the effectiveness of our proposed under-bagging k-NN for imbalanced classification, we conduct extensive experiments on nine real-world imbalanced datasets, including binary-class and multi-class data sets. These imbalanced datasets comes from the UCI Machine Learning Repository (Dua and Graff, 2017). As some data sets contain missing values, we impute the missing values of numerical features and categorical features with the mean value and the most frequent value of the non-missing values for those features respectively. The details of datasets, including size and dimension are listed in Table 1. Besides, we show the proportion of the majorities and minorities in each data set, and then calculate the imbalance ratio ρ. We mention that the number of features in Table 1 are provided after the preprocessing of the one-hot encoding for categorical features. We use the one-hot encoding to transform a categorical feature with k categories into k binary features. We apply standardization of datasets by scaling features to the range of $[0, 1]$. We apply 2 times

10-fold cross-validation with a total of 20 runs for repeated experiments, and then report the average results of AM measure and running time. All experiments are conducted on a 64-bit machine with 24-cores Intel Xeon 2.0GHz CPU (E5-4620) and 64GB main memory.

Table 1: Description of Real-World Data Sets

Datasets	#Instances	#Features	#Classes	%Majority	%Minority	ρ
`APSFailure`	76,000	170	2	98.19	1.81	3.62%
`ActivityRecognition`	22,646	9	4	90.69	1.48	5.92%
`Adult`	48,842	105	2	76.07	23.93	47.86%
`Bitcoin`	2,916,697	6	2	98.58	1.42	2.84%
`BuzzInSocialMedia`	140,707	77	2	99.16	0.84	1.68%
`CencusIncomeKDD`	299,285	503	2	93.8	6.2	12.4%
`CreditCardClients`	30,000	32	2	77.88	22.12	44.24%
`OccupancyDetection`	20,560	5	2	76.9	23.1	46.2%
`p53Mutants`	31,420	5408	2	99.52	0.48	0.96%

We verify the effectiveness of our under-bagging technique for the k-NN classifier, we compare the following cases:

(i) The standard k-NN classifier;

(ii) The k-NN classifier with under-sampling of the majority classes, corresponding to our under-bagging k-NN classifier with $B = 1$ and $a = 1$;

(iii) Our under-bagging k-NN classifier with $B = 5$ and $s = Mn_{(1)}$, which means that we use all samples of the minority class;

(iv) Our under-bagging k-NN classifier with $B = 5$ and $s = 0.5Mn_{(1)}$, which means that the expected sub-sample size turns out to be a half of that in *(iii)* for efficient computing.

We use the `scikit-learn` and `imbalanced-learn` implementations in `Python` and tune the number of neighbors k by cross-validation.

Moreover, we compare our under-bagging k-NN classifier to the following under-sampling and over-sampling methods, including

(i) `Edited Nearest Neighbours (ENN)` (Wilson, 1972) is an under-sampling technique which uses a selection criterion to remove samples from the class to be under-sampled. The selection creterion is to remove samples when the majority of their nearest-neighbbors do not belong to the the same class as the samples. The number of neighbors k to use in this technique is the hyper-parameter.

(ii) `NearMiss` (Mani and Zhang, 2003) is an under-sampling technique which uses some heuristic rules to select samples. It selects samples from majority classes for which the average distance to the N closest samples of the minority class is the smallest. The parameter N, the size of the neighborhood, is the hyper-parameter of this under-sampling technique.

(iii) `One Sided Selection (OSS)` (Kubat et al., 1997) is an under-sampling technique which uses a 1 nearest neighbor rule to iteratively decide if a sample from majority classes should be removed or not. TomekLinks is further used after the nearest neighbor rule to remove noisy samples.

(iv) `TomekLinks` (Tomek, 1976) is an under-sampling technique which uses some heuristic rules to clean the dataset. It uses the Tomek's links to detect paired samples and remove the point from the majority class.

(v) `Random Over-sampling (RO)` (Menardi and Torelli, 2014) aims to over-sample the minority class(es) by picking samples at random with replacement. Then the augmented data set is used instead of the original data set to train a classifier.

(vi) `SMOTE` (Chawla et al., 2002) is an over-sampling approach. It takes the difference between the samples under consideration and its nearest neighbor. Then a random point is selected along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general.

(vii) `ADASYN` (He et al., 2008) uses Adaptive Synthetic algorithm for oversampling. This method is similar to SMOTE but it generates different number of samples depending on an estimate of the local distribution of the class to be over-sampled.

We use the `imbalanced-learn` implementations in `Python`. To mention, we use the standard k-NN classifier on the under-sampled and over-sampled data points generated from these under-sampling and over-sampling methods.

6.3.2 Descriptions of Datasets

The data sets are from the UCI machine learning repository (Dua and Graff, 2017).

- `APSFailure`: The *APS Failure at Scania Trucks Data Set* (Biteus and Lindgren, 2017) contains 76,000 samples and 170 attributes. The dataset is collected for the prediction of failures in the Air Pressure System of Scania heavy trucks.

- `ActivityRecognition`: The goal of the dataset *Activity Recognition with Healthy Older People Using a Batteryless Wearable Sensor Data Set* (Shinmoto Torres et al., 2013) is to predict the activity type using a batteryless, wearable sensor on top of elder people's clothing. There are two room settings (S1 and S2), and we choose the S2 setting, where the sample size is 22,646 and the feature size is 9.

- `Adult`: The *Adult Data Set* is to predict whether the income is more than 50 thousands per year based on census data. It contains 48,842 data points with six numerical attributes and eight categorical attributes. We preprocess the categorical attributes by one-hot encodings, and the feature size of the preprocessed dataset is 105.

- `Bitcoin`: The *Bitcoin Heist Ransomware Address Data Set* (Akcora et al., 2020) has 2,916,697 data points and 6 predictive attributes. This dataset aims at identifying ransomware payments. The original labels are multiclass, containing the 'white' label (i.e., not known to be ransomware) and a series of ransomware types. As the sample size of each ransomware type is small, we combine all ransomware types into a 'ransomware' label to build a binary-class imbalanced dataset.

- `BuzzInSocialMedia`: The *Buzz in social media Data Set* (Kawala et al., 2013) contains two different social networks: Twitter and Tom's Hardware, and we choose the Twitter part. There are 140,707 samples with 77 features representing time-windows. The classification task is to predict whether these time-windows are followed by buzz events or not.

- `CensusIncomeKDD`: The *Census-Income (KDD) Data Set* contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau. There are 299,285 instances with 40 categorical and numerical attributes. We use one-hot encodings to preprocess the categorical attributes, leading to 503 features in the end.

- `CreditCardClients`: The response variable of the *Default of Credit Card Clients Data Set* (Yeh and Lien, 2009) is the default payment, with 23 numerical and categorical explanatory variables. We one-hot encode the categorical variables, and there are finally 32 attributes in total. The number of instances is 30,000.

- `OccupancyDetection`: The *Occupancy Detection Data Set* (Candanedo and Feldheim, 2016) uses temperature, humidity, light, and CO_2 to predict room occupancy. The sample size is 20,560 and the number of predictive variables is 5.

- `p53Mutants`: The classification task of the *p53 Mutants Data Set* (Danziger et al., 2006) is to predict the transcriptional activity (active vs inactive) of p53 proteins. There are 31,420 samples with 5,408 attributes.

6.3.3 Experimental Results

Tables 2(a)-2(c) summarize the averaged performance of the AM measure and Tables 3(a)-3(c) show the computational performance w.r.t. the averaged running time. The best results are marked in bold, the second best marked in <u>underline</u>, and the standard deviations are shown in parentheses.

The numerical experiments verify the theoretical results in the following ways:

(i) For imbalanced classification, under-sampling k-NN and under-bagging k-NN classifier significantly outperform the standard k-NN classifier by a large margin, which empirically verifies the results in Theorems 1 and 3 that under-sampling and under-bagging k-NN classifiers converge to the optimal classifier w.r.t. the AM measure, whereas the standard k-NN turns out to be inconsistent w.r.t. this measure.

(ii) The theoretical relationship between the expected sub-sample size s w.r.t. n shown in Theorem 3 is also numerically verified. When we compare our under-bagging k-NN with $s = Mn_{(1)}$ to our under-bagging k-NN with $s = 0.5Mn_{(1)}$, we find in Tables 2(a) and 3(a) that the AM performance trained with smaller sub-sample size does not degrade too much, and sometimes can be even slightly better, whereas the running time of our under-bagging k-NN with $s = 0.5Mn_{(1)}$ is much smaller than that with $s = Mn_{(1)}$. This indicates that we

Table 2: Average AM measure among Different Methods on Real-World Data Sets

(a) Part 1/3

Datasets	k-NN	Under-bagging with $s = Mn_{(1)}$ $B = 1$	$B = 5$	$s = 0.5Mn_{(1)}$ $B = 5$
APSFailure	0.7917 (0.0226)	0.9401 (0.0098)	0.9464 (0.0100)	0.9444 (0.0100)
ActivityRecognition	0.8439 (0.0228)	0.8499 (0.0181)	0.8571 (0.0203)	0.8485 (0.0210)
Adult	0.7491 (0.0054)	0.8020 (0.0067)	**0.8047 (0.0070)**	0.8046 (0.0057)
Bitcoin	0.5638 (0.0062)	0.6630 (0.0031)	**0.6740 (0.0024)**	0.6694 (0.0032)
BuzzInSocialMedia	0.7959 (0.0215)	0.9411 (0.0125)	**0.9433 (0.0122)**	0.9425 (0.0124)
CensusIncomeKDD	0.6799 (0.0058)	0.8412 (0.0042)	**0.8459 (0.0038)**	0.8435 (0.0033)
CreditCardClients	0.6373 (0.0084)	0.6829 (0.0095)	**0.6892 (0.0102)**	0.6868 (0.0094)
OccupancyDetection	0.9930 (0.0024)	0.9941 (0.0018)	0.9942 (0.0019)	0.9938 (0.0020)
p53Mutants	0.7733 (0.0601)	0.8824 (0.0449)	0.9006 (0.0408)	0.8911 (0.0460)

(b) Part 2/3

Datasets	ENN	NearMiss	OSS	TomekLinks
APSFailure	0.8938(0.0097)	0.8593(0.0121)	0.7978(0.0223)	0.7976(0.0223)
ActivityRecognition	0.8453(0.0228)	0.8181(0.0232)	0.8440(0.0233)	0.8437(0.0229)
Adult	0.8041(0.0061)	0.7480(0.0063)	0.7712(0.0053)	0.7712(0.0053)
Bitcoin	0.6511(0.0035)	0.5361(0.0016)	0.6206(0.0032)	0.6206(0.0031)
BuzzInSocialMedia	0.8893(0.0190)	0.8111(0.0073)	0.8078(0.0221)	0.8078(0.0221)
CensusIncomeKDD	0.8164(0.0128)	0.6958(0.0048)	0.7023(0.0063)	0.7016(0.0060)
CreditCardClients	0.6831(0.0089)	0.6539(0.0104)	0.6548(0.0088)	0.6548(0.0087)
OccupancyDetection	0.9945(0.0020)	0.9924(0.0034)	0.9933(0.0022)	0.9933(0.0022)
p53Mutants	0.8149(0.0635)	0.7608(0.0476)	0.7832(0.0595)	0.7832(0.0595)

(c) Part 3/3

Datasets	RO	SMOTE	ADASYN
APSFailure	0.9293(0.0143)	0.9481(0.0124)	**0.9507(0.0105)**
ActivityRecognition	0.8372(0.0374)	0.8692(0.0223)	**0.8705(0.0246)**
Adult	0.7948(0.0065)	0.7936(0.0056)	0.7821(0.0071)
Bitcoin	0.6360(0.0026)	0.6525(0.0027)	0.6512(0.0033)
BuzzInSocialMedia	0.9112(0.0105)	0.9294(0.0086)	0.9321(0.0081)
CensusIncomeKDD	0.8283(0.0071)	0.8277(0.0050)	0.8255(0.0057)
CreditCardClients	0.6716(0.0115)	0.6755(0.0085)	0.6667(0.0066)
OccupancyDetection	0.9954(0.0019)	**0.9960(0.0018)**	0.9955(0.0019)
p53Mutants	0.8629(0.0698)	**0.9129(0.0514)**	0.9129(0.0514)

can use a relative small sub-sample size to speed up the under-bagging k-NN and still keep a good performance in terms of the AM measure.

(iii) The running time of our under-bagging k-NN classifier demonstrates its computational efficiency compared with the standard k-NN classifier, especially when the sample size

Table 3: Average Running Time among Different Methods on Real-World Data Sets

(a) Part 1/3

Datasets	k-NN	Under-bagging with $s = Mn_{(1)}$ $B=1$	$B=5$	$s = 0.5Mn_{(1)}$ $B=5$
APSFailure	5.33 (0.07)	**0.48 (0.07)**	0.84 (0.07)	0.52 (0.05)
ActivityRecognition	0.56 (0.10)	**0.06 (0.02)**	0.17 (0.04)	0.13 (0.02)
Adult	4.37 (0.15)	2.05 (0.16)	2.92 (0.14)	**1.59 (0.10)**
Bitcoin	16885.49 (834.14)	10.21 (1.42)	11.88 (1.16)	**8.11 (0.86)**
BuzzInSocialMedia	21.86 (5.36)	0.69 (0.10)	1.08 (0.13)	**0.65 (0.10)**
CensusIncomeKDD	107.43 (0.41)	21.11 (1.01)	37.22 (0.60)	**19.00 (0.81)**
CreditCardClients	1.51 (0.20)	0.71 (0.05)	1.02 (0.06)	**0.65 (0.04)**
OccupancyDetection	0.49 (0.03)	0.24 (0.05)	0.40 (0.06)	0.25 (0.03)
p53Mutants	2.72 (0.00)	**1.51 (0.07)**	3.20 (0.08)	2.22 (0.07)

(b) Part 2/3

Datasets	ENN	NearMiss	OSS	TomekLinks
APSFailure	113.76(3.53)	2.06(0.10)	51.33(8.33)	0.98(1.24)
ActivityRecognition	1.01(0.17)	0.48(0.05)	1.57(0.06)	0.93(0.25)
Adult	36.98(0.73)	6.79(0.18)	30.35(0.37)	30.94(0.59)
Bitcoin	492.15(98.29)	83.43(12.23)	523.43(19.73)	449.59(101.17)
BuzzInSocialMedia	341.22(8.04)	4.02(0.13)	162.30(35.68)	244.52(6.45)
CensusIncomeKDD	1728.47(56.86)	82.72(0.88)	1008.89(206.21)	1367.94(8.67)
CreditCardClients	13.12(0.84)	2.40(0.28)	10.99(0.15)	11.38(0.24)
OccupancyDetection	0.24(0.05)	0.21(0.04)	0.61(0.01)	0.23(0.04)
p53Mutants	35.25(1.99)	3.56(0.20)	29.40(3.53)	32.70(0.53)

(c) Part 3/3

Datasets	RO	SMOTE	ADASYN
APSFailure	19.02(0.74)	20.24(1.12)	21.60(1.40)
ActivityRecognition	0.26(0.02)	0.29(0.02)	0.96(0.02)
Adult	6.35(0.36)	8.79(0.40)	17.24(2.71)
Bitcoin	84.70 (1.12)	84.90(1.29)	82.58(1.20)
BuzzInSocialMedia	55.81(1.92)	52.50(3.08)	53.36(1.39)
CensusIncomeKDD	348.61(23.67)	351.45(13.42)	464.12(13.89)
CreditCardClients	2.39(0.08)	3.13(0.04)	6.93(0.12)
OccupancyDetection	**0.09(0.00)**	0.11(0.00)	0.31(0.00)
p53Mutants	8.54(0.19)	10.78(0.26)	13.10(1.12)

is large. As our complexity analysis in Section 4.3.3 show, the time complexity of construction can be reduced from $\mathcal{O}(n \log n)$ (for the standard k-NN) to $\mathcal{O}((\rho n \log(\rho n))^{d/(2\alpha+d)})$, and the time complexity in the testing stage can be reduced from $\mathcal{O}((n \log n)^{2\alpha/(2\alpha+d)})$ to

$\mathcal{O}(\log^2(\rho n))$. To verify the complexity results, we compare the running time of the under-bagging k-NN ($B = 1$) with that of the standard k-NN, and we observe that the under-bagging technique significantly reduces the running time. In Table 3(a), the under-bagging technique can reduce at least half the running time of the standard k-NN. In particular, on `Bitcoin`, where the sample size is up to 3 million and the imbalance ratio is about 3%, the running time can even be reduced by 99.4%. Besides, when we adopt more bagging rounds $B = 5$, the performance of the under-bagging k-NN w.r.t. the AM measure further enhances with a mild increase of the running time.

(iv) From Tables 2(a), 2(b), 3(a), and 3(b), our under-bagging k-NN classifier not only outperforms the comparing under-sampling algorithms, but also it's competitive in computational efficiency. For example, the averaged AM measure score of our under-bagging k-NN with $s = Mn_{(1)}$ and $B = 5$ is 0.9464 on the dataset `APSFailure`, which significantly outperforms other four under-sampling algorithms with AM measure 0.8938, 0.8593, 0.7978, and 0.7976, respectively. Moreover, the averaged running time of our under-bagging k-NN is also much smaller than other competitive under-sampling methods, at least saving 60% of the running time (NearMiss 2.1s v.s. ours under-bagging 0.84s) for the dataset `APSFailure`. The advantages of our under-bagging k-NN in practice can be explained from two aspects. On the one hand, our under-bagging k-NN utilize more information from samples in the majority classes as the bagging size B increases. By contrast, under-sampling methods may throws away potentially useful information of the data as discussed in Section 1, which leads to the poor AM measure. On the other hand, we mention that our under-bagging k-NN is based on the random sampling procedure with smaller computation complexities than other under-sampling methods which either selects the best prototype or cleans noisy samples from the majority classes.

(v) In Tables 2(a), 2(c), 3(a) and 3(c), we compare our under-bagging k-NN with over-sampling algorithms. From the perspective of the AM measure, our under-bagging k-NN shows the best performance on 5 datasets, and shows comparable performance to SMOTE and ADADYN on the remaining 4 datasets. Note that our under-bagging k-NN outperforms RO on most datasets since random over-sampling doesn't significantly improve minority class recognition as discussed in Japkowicz (2000); Chawla et al. (2002). Moreover, from the computational perspective, our running time is significantly smaller than those of the over-sampling algorithms on most of the datasets, since our under-bagging k-NN doesn't involve additional synthetic data generation to rebalance the data and the size of training data at each bagging round is reduced by the under-sampling strategy.

7. Proofs

In this section, we first prove the fundamental results related to the AM measure in Section 5, which play an essential role in establishing the convergence rates for both under-sampling and under-bagging k-NN classifiers. Then in Section 7.1 and Section 7.2, we present the proofs of the theoretical results related to the under-sampling k-NN in Section 5.1 and the under-bagging k-NN in Section 5.2, respectively.

Proof [of Proposition 5] By the definition of the balanced loss in (22), we have

$$\begin{aligned}1-\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\psi) &= 1-\sum_{i=1}^{M}\mathrm{P}(y=i,\psi(x)\neq y)/(M\pi_i)=\frac{1}{M}\sum_{i=1}^{M}(\pi_i-\mathrm{P}(y=i,\psi(x)\neq y))/\pi_i\\ &=\frac{1}{M}\sum_{i=1}^{M}\mathrm{P}(y=i,\psi(x)=i)/\pi_i=\frac{1}{M}\sum_{i=1}^{M}\mathrm{P}(\psi(x)=i|y=i)=r_{\mathrm{AM}}(\psi).\end{aligned}$$

This completes the proof of Proposition 5. ■

The following lemma is needed in the proof of Theorem 6, which provides a new formulation of the excess risk w.r.t. the balanced loss.

Lemma 14 *Let the balanced loss be defined in* (23). *Then we have*

$$\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\psi)-\mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}}=\sum_{m=1}^{M}(\eta_m(x)/(M\pi_m))\mathbb{E}_{\mathrm{P}_X}\big[\eta^w_{L_{\mathrm{bal}},\mathrm{P}}(x)-\eta^w_{\psi(x)}(x)\big],$$

where we write $\eta^w_{L_{\mathrm{bal}},\mathrm{P}}(x):=\eta^w_{\psi^*_{L_{\mathrm{bal}},\mathrm{P}}(x)}(x)$.

Proof [of Lemma 14] By the definition of the balanced loss (23), we have

$$\begin{aligned}\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\psi) &= \mathbb{E}_{\mathrm{P}_X}\Big[\mathbb{E}_{\mathrm{P}_{Y|X}}\Big[\sum_{m=1}^{M}\mathbf{1}\{Y=m\}\mathbf{1}\{\psi(x)\neq Y\}/(M\pi_m)\Big]\Big|X=x\Big]\\ &= \mathbb{E}_{\mathrm{P}_X}\Big[\mathbb{E}_{\mathrm{P}_{Y|X}}\Big[\sum_{j=1}^{M}\Big(\mathbf{1}\{\psi(x)=j\}\sum_{m=1}^{M}\mathbf{1}\{Y=m\}\mathbf{1}\{Y\neq j\}/(M\pi_m)\Big)\Big]\Big|X=x\Big]\\ &= \mathbb{E}_{\mathrm{P}_X}\Big[\sum_{j=1}^{M}\Big(\mathbf{1}\{\psi(x)=j\}\mathbb{E}_{\mathrm{P}_{Y|X}}\Big[\sum_{m=1}^{M}\mathbf{1}\{Y=m\}\mathbf{1}\{Y\neq j\}/(M\pi_m)\Big]\Big)\Big|X=x\Big]\\ &= \mathbb{E}_{\mathrm{P}_X}\Big[\sum_{j=1}^{M}\Big(\mathbf{1}\{\psi(x)=j\}\sum_{m\neq j}^{M}\eta_m(x)/(M\pi_m)\Big)\Big]\\ &= \mathbb{E}_{\mathrm{P}_X}\Big[\sum_{m\neq\psi(x)}^{M}\eta_m(x)/(M\pi_m)\Big].\end{aligned}$$

Consequently we find

$$\begin{aligned}\mathcal{R}_{L_{\mathrm{bal}},\mathrm{P}}(\psi)-\mathcal{R}^*_{L_{\mathrm{bal}},\mathrm{P}} &= \mathbb{E}_{\mathrm{P}_X}\Big[\sum_{m\neq\psi(x)}^{M}\eta_m(x)/(M\pi_m)\Big]-\mathbb{E}_{\mathrm{P}_X}\Big[\sum_{m\neq\psi^*_{L_{\mathrm{bal}},\mathrm{P}}(x)}^{M}\eta_m(x)/(M\pi_m)\Big]\\ &= \sum_{m=1}^{M}(\eta_m(x)/(M\pi_m))\mathbb{E}_{\mathrm{P}_X}\big[\eta^w_{L_{\mathrm{bal}},\mathrm{P}}(x)-\eta^w_{\psi(x)}(x)\big],\end{aligned}$$

where $\eta^w_{L_{\mathrm{bal}},\mathrm{P}}(x) = \eta^w_{\psi^*_{L_{\mathrm{bal}},\mathrm{P}}(x)}(x)$. Thus we obtain the assertion. ■

Proof [of Theorem 6] An elementary calculation yields

$$\begin{aligned}\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\psi) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} &= \mathbb{E}_{\mathrm{P}^w_X}\big[\mathbb{E}_{\mathrm{P}^w_{Y|X}}[\mathbf{1}\{\psi(x) \neq Y\} - \mathbf{1}\{\psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x) \neq Y\}]\big|X = x\big] \\ &= \mathbb{E}_{\mathrm{P}^w_X}\big[\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) - \eta^w_{\psi(x)}(x)\big],\end{aligned}$$

where we write $\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) := \eta^w_{\psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x)}(x)$. By (28), we find

$$\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\psi) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} = \mathbb{E}_{\mathrm{P}^w_X}\big[\eta^{w,*}_{L_{\mathrm{bal}},\mathrm{P}}(x) - \eta^w_{\psi(x)}(x)\big],$$

where we write $\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w} := \eta^w_{f^*_{L_{\mathrm{bal}},\mathrm{P}}(x)}(x)$. This together with (27) implies

$$\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\psi) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} = \sum_{m=1}^M (\eta_m(x)/(M\pi_m)) \cdot \mathbb{E}_{\mathrm{P}_X}\big[\eta^w_{L_{\mathrm{bal}},\mathrm{P}}(x) - \eta^w_{\psi(x)}(x)\big].$$

Combining this with Lemma 14, we obtain the assertion. ■

To prove Proposition 7, we need the following Lemmas 15 and 16, which reduce the problem of analyzing the excess classification risk to the problem of analyzing the error estimation of posterior probability function.

Lemma 15 *Let $\widehat{\eta} : \mathcal{X} \to [0,1]^M$ be an estimate of η^w and $\widehat{\psi}(x) = \arg\max_{m\in[M]} \widehat{\eta}_m(x)$. If $(\mathrm{P}^w)^n(\|\widehat{\eta}(x) - \eta^w(x)\|_\infty \leq \phi) \geq 1-\delta$, then with probability $(\mathrm{P}^w)^n$ at least $1-\delta$, there holds $\|\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) - \eta^w_{\widehat{\psi}(x)}(x)\|_\infty \leq 2\phi$, where $\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) = \eta_{\psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x)}(x)$.*

Proof [of Lemma 15] Fix an $x \in \mathcal{X}$ with $\|\widehat{\eta}(x) - \eta^w(x)\|_\infty \leq \phi$. Let $m^* = \arg\max_{m\in[M]} \eta^w_m(x)$ and $m = \widehat{\psi}(x)$. Then we have $\eta^w_{m^*}(x) \leq \widehat{\eta}_{m^*}(x) + \phi$ and $\eta^w_m(x) \geq \widehat{\eta}_m(x) - \phi$. Thus, we find

$$\eta^w_{m^*}(x) - \eta^w_m(x) \leq (\widehat{\eta}_{m^*}(x) + \phi) - (\widehat{\eta}_m(x) - \phi) = (\widehat{\eta}_{m^*}(x) - \widehat{\eta}_m(x)) + 2\phi.$$

Since $m = \widehat{\psi}(x)$ is the maximum entry of $\widehat{\eta}_m(x)$, we have $\widehat{\eta}_{m^*}(x) \leq \widehat{\eta}_m(x)$ and consequently $\eta^w_{m^*}(x) - \eta^w_m(x) \leq 2\phi$. In other words, we show that the event $\{x \in \mathcal{X} : \|\widehat{\eta}(x) - \eta^w(x)\|_\infty \leq \phi\}$ is contained in $\{\eta^w_{m^*}(x) - \eta^w_m(x) \leq 2\phi\}$. This implies that for all $x \in \mathcal{X}$, with probability $(\mathrm{P}^w)^n$ at least $1-\delta$, there holds $\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) - \eta^w_{\widehat{\psi}(x)}(x) \leq 2\phi$, which finishes the proof. ■

Lemma 16 *Let $\widehat{\eta} : \mathcal{X} \to [0,1]^M$ be an estimate of η^w and $\widehat{\psi}(x) = \arg\max_{m\in[M]} \widehat{\eta}_m(x)$. Moreover, let Assumption 1 hold. Suppose that $\|\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w} - \eta^w_{\widehat{\psi}(x)}\|_\infty \leq 2\phi$ holds for some $\phi > 0$ with probability $(\mathrm{P}^w)^n$ at least $1-\delta$, where $\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) = \eta^w_{\psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x)}(x)$. Then with probability $(\mathrm{P}^w)^n$ at least $1-\delta$, there holds*

$$\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\widehat{\psi}) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} \leq (c_\beta/(M\underline{\pi})(2\phi)^{\beta+1}$$

Proof [of Lemma 16] Let $m = \psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x)$. By the definition of the excess risk, we have

$$\mathcal{R}_{L,\mathrm{P}^w}(\widehat{\psi}(x)) - \mathcal{R}^*_{L,\mathrm{P}^w} = \mathbb{E}_{\mathrm{P}^w_X}\big[\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x)\big].$$

Let $\eta^w_{(m)}(x)$ denote the m-th largest entry of the vector $\eta^w(x) = (\eta^w_1(x), \ldots, \eta^w_M(x))^\top$ and define $\delta(x) := \eta^w_{(1)}(x) - \eta^w_{(2)}(x)$. Moreover, let $\Delta_n := \{x : \mathcal{X} : \delta(x) \geq 2\phi\}$. Then we have

$$\mathbb{E}_{\mathrm{P}^w_X}[\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x)] = \int_{\Delta_n} (\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x))\, d\mathrm{P}^w_X(x) + \int_{\Delta_n^c} (\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x))\, d\mathrm{P}^w_X(x).$$

Now we consider the two regions Δ_n and Δ_n^c separately to bound the error. If $\delta(x) \geq 2\phi$, since $\eta^w_{(1)}(x) = \eta^w_m(x)$, we have $\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x) < 2\phi \leq \eta^w_{(1)}(x) - \eta^w_{(2)}(x)$. In other words, we find $\eta^w_{\widehat{\psi}(x)}(x)$ is larger than $\eta^w_{(2)}(x)$, which yields $\eta^w_{\widehat{\psi}(x)}(x) = \eta^w_{(1)}(x) = \eta^w_m(x)$. Consequently, we obtain

$$\int_{\Delta_n} (\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x))\, d\mathrm{P}^w_X(x) = 0. \tag{40}$$

Otherwise if $\delta(x) \leq 2\phi$, then by (27), we have

$$\begin{aligned}
\int_{\Delta_n^c} (\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x))\, d\mathrm{P}^w_X(x) &= \int_{\Delta_n^c} \sum_{m=1}^M (\eta_m(x)/(M\pi_m))(\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x))\, d\mathrm{P}_X(x) \\
&\leq \frac{1}{M\underline{\pi}} \int_{\Delta_n^c} (\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x))\, d\mathrm{P}_X(x) \\
&\leq \mathrm{P}_X(\delta(x) \leq 2\phi)/(M\underline{\pi}).
\end{aligned}$$

Using Condition (i) in Assumption 1, we get

$$\int_{\Delta_n^c} (\eta^w_m(x) - \eta^w_{\widehat{\psi}(x)}(x))\, d\mathrm{P}^w_X(x) \leq (c_\beta/(M\underline{\pi}))(2\phi)^{\beta+1}. \tag{41}$$

Combining (40) and (41), we obtain the assertion. ∎

Proof [of Proposition 7] Proposition 7 is a straightforward consequence of Lemma 15 and Lemma 16. ∎

Before we proceed, we list two lemmas that will be used frequently in the proofs. Lemma 17 is Hoeffding's inequality, which was established in Hoeffding (1963) and Lemma 18 is Bernstein's inequality, which was introduced in Bernstein (1946). Both concentration inequalities can be found in many statistical learning textbooks, see e.g., Massart (2007); Cucker and Zhou (2007); Steinwart and Christmann (2008).

Lemma 17 (Hoeffding's inequality) *Let $a < b$ be two real numbers, $n \geq 1$ be an integer, and $\xi_1, \ldots, \xi_n$ be independent random variables satisfying $\xi_i \in [a, b]$, for $1 \leq i \leq n$. Then, for all $\tau > 0$, we have*

$$\mathrm{P}\biggl(\frac{1}{n}\sum_{i=1}^n (\xi_i - \mathbb{E}_{\mathrm{P}}\xi_i) \geq (b-a)\sqrt{\frac{\tau}{2n}}\biggr) \leq e^{-\tau}.$$

Lemma 18 (Bernstein's inequality) *Let $B > 0$ and $\sigma > 0$ be real numbers, and $n \geq 1$ be an integer. Furthermore, let $\xi_1, \ldots, \xi_n$ be independent random variables satisfying $\mathbb{E}_{\mathrm{P}}\xi_i = 0$, $\|\xi_i\|_\infty \leq B$, and $\mathbb{E}_{\mathrm{P}}\xi_i^2 \leq \sigma^2$ for all $i = 1, \ldots, n$. Then for all $\tau > 0$, we have*

$$\mathrm{P}\biggl(\frac{1}{n}\sum_{i=1}^n \xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\biggr) \leq e^{-\tau}.$$

7.1 Proofs Related to the Under-sampling k-NN Classifier

In this section, we first present in Sections 7.1.1-7.1.3 the proof of the theoretical results on bounding the sample error in Section 5.1.1, the approximation error in Section 5.1.2, and the under-sampling error in Section 5.1.3, respectively. Then, in Section 7.1.4, we prove the main result on the convergence rates of the under-sampling k-NN classifier and the minimax lower bound, i.e., Theorems 1 and 2 in Section 4.1.

7.1.1 Proofs Related to Section 5.1.1

The following lemma providing an explicit expression for the under-sampling distribution discussed in Section 5.1, which supplies the key to the proof of Lemma 21 and Proposition 10.

Lemma 19 *Let P^u be the probability distribution of the accepted samples by the under-sampling strategy in Section 3.1. Then we have*

$$\mathrm{P}^u(X \in A, Y = m) = \frac{\int_A \eta_m(x) f_X(x)\, dx / n_{(m)}}{\sum_{m=1}^M \pi_m / n_{(m)}}.$$

Moreover, the marginal distribution can be expressed as

$$\pi_m^u := \mathrm{P}^u(Y = m) = \frac{\int_{\mathcal{X}} \eta_m(x) f_X(x)\, dx / n_{(m)}}{\sum_{m=1}^M \pi_m / n_{(m)}} = \frac{\pi_m / n_{(m)}}{\sum_{m=1}^M \pi_m / n_{(m)}}. \tag{42}$$

In addition, the conditional density function is $f^u(x|y = m) = \eta_m(x) f_X(x)/\pi_m$ and the posterior probability function is given by

$$\eta_m^u(x) := \mathrm{P}^u(Y = m | X = x) = \frac{\pi_m^u f^u(x|y = m)}{\sum_{m=1}^M \pi_m^u f^u(x|y = m)} = \frac{\eta_m(x)/n_{(m)}}{\sum_{m=1}^M \eta_m(x)/n_{(m)}} \tag{43}$$

Furthermore, the marginal distribution $f_X^u(x)$ can be expressed as

$$f_X^u(x) = \sum_{m=1}^M \pi_m^u(x) f(x|y = m). \tag{44}$$

Proof [of Lemma 19] Let $\mathrm{P}_{X,Y,Z} = \mathrm{P}_{X,Y} \times \mathrm{P}_{Z|(X,Y)}$ denote the joint probability measure. Then we can calculate the probability of $Z(X, Y) = 1$, that is, (X, Y) is accepted in the under-sampling strategy, as follows:

$$\mathrm{P}_{X,Y,Z}(Z(X, Y) = 1) = \int_{\mathcal{X}\times\mathcal{Y}} \mathrm{P}(Z(X, Y) = 1 | (X, Y) = (x, y))\, d\mathrm{P}(x, y)$$

$$= \sum_{m=1}^{M} \pi_m \int_{\mathcal{X}} \mathrm{P}(Z(X,Y)=1|(X,Y)=(x,m))\, d\mathrm{P}(x|y=m)$$
$$= \sum_{m=1}^{M} \pi_m \int_{\mathcal{X}} a(x,m) f(x|y=m)\, dx.$$

Thus, for any measurable set A of $\mathcal{X}$, an elementary calculation yields

$$\begin{aligned}
&\mathrm{P}_{X,Y,Z}(X \in A, Y=m, Z(X,Y)=1) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{1}\{x \in A\}\mathbf{1}\{Y=m\}\mathrm{P}(Z(x,y)=1|(X,Y)=(x,y))\, d\mathrm{P}(x,y) \\
&= \pi_m \int_{\mathcal{X}} \mathbf{1}\{x \in A\}\mathrm{P}(Z(x,m)=1|(X,Y)=(x,m))\, d\mathrm{P}(x|y=m) \\
&= \pi_m \int_{\mathcal{X}} \mathbf{1}\{x \in A\} a(x,m) f(x|y=m)\, dx \\
&= \pi_m \int_{A} a(x,m) f(x|y=m)\, dx.
\end{aligned}$$

Consequently, the distribution function of the accepted samples is given by

$$\begin{aligned}
\mathrm{P}^u(X \in A, Y=m) &= \mathrm{P}_{X,Y,Z}(X \in A, Y=m|Z(X,Y)=1) \\
&= \frac{\mathrm{P}^u(X \in A, Y=m, Z(X,Y)=1)}{\mathrm{P}^u(Z(X,Y)=1)} \\
&= \frac{\pi_m \int_A a(x,m) f(x|y=m)\, dx}{\sum_{m=1}^{M} \pi_m \int_{\mathcal{X}} a(x,m) f(x|y=m)\, dx}.
\end{aligned}$$

Combining this with (9), we find

$$\mathrm{P}^u(X \in A, Y=m) = \mathrm{P}(X \in A, Y=m|Z(X,Y)=1) = \frac{\int_A \eta_m(x) f_X(x)\, dx / n_{(m)}}{\sum_{m=1}^{M} \pi_m / n_{(m)}}. \quad (45)$$

Thus we finish the proof with straightforward application of the joint probability measure in (45) for calculating π_m^u, $f^u(x|y=m)$, η_m^u and $f_X^u(x)$. ■

To prove Proposition 8, we need to bound the number of reorderings of the data. To be specific, for fixed $x \in \mathbb{R}^d$, we reorder samples, $X_1, \ldots, X_n$, according to increasing values of $\|X_i - x\|$ with breaking ties by considering indices, i.e., $\|X_{\sigma_1} - x\| \leq \cdots \leq \|X_{\sigma_n} - x\|$, where $(\sigma_1, \ldots, \sigma_n)$ is a permutation of $(1, \ldots, n)$. Then we define the inverse of the permutation, namely the rank Σ_i by $\Sigma_i := \{1 \leq \ell \leq n : X_{\sigma_\ell} = X_i\}$. Since we break ties by considering indices, the rank Σ_i is unique for all $1 \leq i \leq n$. Therefore, the rank vector $(\Sigma_1, \ldots, \Sigma_n)$ for $x \in \mathbb{R}^d$ is well-defined. Let $\mathcal{S} = \{(\Sigma_1, \ldots, \Sigma_n), x \in \mathbb{R}^d\}$ be the set of all rank vectors one can observe by moving x around in space and we use the notation $|\mathcal{S}|$ to represent the cardinality of $\mathcal{S}$.

The next lemma provides the upper bound for the number of reorderings, which plays a crucial rule to derive the uniform bound for the proof of Propositions 8, 11 and 13.

Lemma 20 *For any $d \geq 1$ and all $n \geq 2d$, there holds $|\mathcal{S}| \leq (25/d)^d n^{2d}$.*

Proof [of Lemma 20] The hyperplane $\|x - X_i\|^2 = \|x - X_j\|^2$ generates a sign function

$$p_{ij}(x) = \begin{cases} 1 & \text{if } \|x - X_i\|^2 > \|x - X_j\|^2, \\ 0 & \text{if } \|x - X_i\|^2 = \|x - X_j\|^2, \\ -1 & \text{if } \|x - X_i\|^2 < \|x - X_j\|^2. \end{cases}$$

The collection of the sign functions $\{p_{ij}(x), 1 \leq i \leq j \leq n\}$, called the sign pattern, determines the ordering of $\|x - x_i\|^2$ and identifies all ties. Theorem 6.2.1 in Matousek (2013) shows that the maximal number of sign pattern is not larger than $(50DN/d)^d$ for any $d \geq 1$ and $N \geq 2d$, where D denote the maximum degree of the $\{p_{ij}(x), 1 \leq i \leq j \leq n\}$ and N denote the number of sign functions. Clearly, we have $d = 1$ and $N = \binom{n}{2}$. Therefore, for any $d \geq 1$ and all $n \geq 2d$, the number of sign pattern is not more than $(25/d)^d n^{2d}$. This completes the proof. ■

The next lemma states that the under-sampling distribution P^u is relatively close to the balanced distribution P^w with high probability. To be specific, (46) bounds the sub-sample size of each class $n_{(m)}$ and the marginal density function $f_X^u(x)$. In addition, (47) implies that if $\eta_m(x)$ is assumed to be α-Hölder continuous, $\eta_m^u(x)$ remains to be α-Hölder continuous. This lemma will be used several times in the sequel, which is crucial in the proof of Propositions 8 and 9.

Lemma 21 *Let $\eta_m^w(x)$, $\eta_m^u(x)$ and $f_X^u(x)$ be defined as in (2), (29), and (44), respectively. Assume that P satisfies Assumption 1 and P_X is the uniform distribution on $[0,1]^d$. Then there exists an $n_1 \in \mathbb{N}$ such that for all $1 \leq m \leq M$ and all $n \geq n_1$, there hold*

$$n\pi_m/2 \leq n_{(m)} \qquad \text{and} \qquad 1/(2M\overline{\pi}) \leq f_X^u(x) \leq 2/(M\underline{\pi}) \tag{46}$$

with probability P^n at least $1 - 2M/n^3$. Moreover, for $x, x' \in \mathcal{X}$, we have

$$|\eta_m^u(x') - \eta_m^u(x)| \leq 4c_L \|x - x'\|^\alpha. \tag{47}$$

Proof [of Lemma 21] For any $1 \leq m \leq M$, let $\zeta_i := \boldsymbol{1}\{Y_i = m\} - \pi_m$. Then ζ_i's are independent random variables such that $\mathbb{E}_{\mathrm{P}}[\zeta_i] = 0$ and $\mathbb{E}_{\mathrm{P}}\zeta_i^2 \leq 1/4$ for $1 \leq i \leq n$. Using Bernstein's inequality in Lemma 18, we obtain that for any $\tau > 0$, there holds

$$\mathrm{P}^n\biggl(\frac{1}{n}\sum_{i=1}^n \boldsymbol{1}\{Y_i = m\} \geq \pi_m + \sqrt{\frac{\tau}{2n}} + \frac{2\tau}{3n}\biggr) \leq e^{-\tau}.$$

Setting $\tau := 3\log n$, we get

$$\mathrm{P}^n\bigl(n_{(m)}/n \geq \pi_m + 2\sqrt{\log n/n}\bigr) \leq 1/n^3. \tag{48}$$

On the other hand, let $\zeta_i' := \boldsymbol{1}\{Y_i \neq m\} - (1 - \pi_m)$, then ζ_i''s are independent random variables such that $\mathbb{E}_{\mathrm{P}}[\zeta_i'] = 0$ and $\mathbb{E}_{\mathrm{P}}\zeta_i'^2 \leq 1/4$ for $1 \leq i \leq n$. Again, by using Bernstein's inequality in Lemma 18, we obtain that for any $\tau > 0$, there holds

$$\mathrm{P}^n\biggl(1 - \frac{1}{n}\sum_{i=1}^n \boldsymbol{1}\{Y_i = m\} \geq 1 - \pi_m + \sqrt{\frac{\tau}{2n}} + \frac{2\tau}{3n}\biggr) \leq 1/n^3.$$

Setting $\tau := 3 \log n$, we obtain

$$\mathrm{P}^n \big(\pi_m \geq n_{(m)}/n + 2\sqrt{\log n/n} \big) \leq 1/n^3. \tag{49}$$

Now, (48) together with (49) and a union bound argument yields

$$\mathrm{P}^n \Big(1 - (2/\underline{\pi})\sqrt{\log n/n} \leq n_{(m)}/(n\pi_m) \leq 1 + (2/\overline{\pi})\sqrt{\log n/n}, \ \forall 1 \leq m \leq M \Big) \geq 1 - 2M/n^3, \tag{50}$$

where $\overline{\pi} = \max_{1 \leq m \leq M} \pi_i$ and $\underline{\pi} = \min_{1 \leq m \leq M} \pi_i$. Let the event E be defined by

$$E := \Big\{ 1 - (2/\underline{\pi})\sqrt{\log n/n} \leq n_{(m)}/(n\pi_m) \leq 1 + (2/\overline{\pi})\sqrt{\log n/n}, \ \forall 1 \leq m \leq M \Big\}$$

and the integer $n_1 \in \mathbb{N}_+$ satisfy $\log n_1/n_1 \leq \min\{\overline{\pi}^2/4, \underline{\pi}^2/16\}$. The following arguments will be made on the event E if $n > n_1$.

(i) Since $n > n_1$, the definition of the event E implies that $1/2 < n_{(m)}/(n\pi_m)$ for $1 \leq m \leq M$. Consequently we have $n\pi_m/2 \leq n_{(m)} \leq n$.

(ii) By (42), we have for $1 \leq m \leq M$,

$$|\pi_m^u - 1/M| = \bigg| \frac{n\pi_m/n_{(m)}}{\sum_{m=1}^M n\pi_m/n_{(m)}} - 1/M \bigg| = \frac{|Mn\pi_m/n_{(m)} - \sum_{m=1}^M n\pi_m/n_{(m)}|}{M\sum_{m=1}^M n\pi_m/n_{(m)}}.$$

On the event E, there holds

$$|\pi_m^u - 1/M| \leq (4/(M\underline{\pi}))\sqrt{\log n/n} \cdot \big(1 + (2/\overline{\pi})\sqrt{\log n/n}\big) \cdot \big(1 - (2/\underline{\pi})\sqrt{\log n/n}\big)^{-1}.$$

Since $n > n_1$, we have $|\pi_m^u - 1/M| \leq (16/(M\underline{\pi})\sqrt{\log n/n}$ and consequently

$$1/(2M\overline{\pi}) \leq \sum_{m=1}^M \eta_m(x)/(2M\pi_m) \leq \sum_{m=1}^M \pi_m^u \eta_m(x)/\pi_m \leq \sum_{m=1}^M 2\eta_m(x)/(M\pi_m) \leq 2/(M\underline{\pi}).$$

This together with $f_X^u(x) = \sum_{m=1}^M \pi_m^u(x) f(x|y=m) = f_X(x)\sum_{m=1}^M \pi_m^u(x)\eta_m(x)/\pi_m$ yields $(1/(2M\overline{\pi}))f_X(x) \leq f_X^u(x) \leq (2/(M\underline{\pi}))f_X(x)$.

(iii) For any $1 \leq m \leq M$ and $x, x' \in \mathcal{X}$, by (13) and (29), we have

$$|\eta_m^u(x') - \eta_m^u(x)| \leq c_L\|x' - x\|^\alpha \cdot \frac{n/n_{(m)}}{\sum_{m=1}^M n\eta_m(x)/n_{(m)}}.$$

On the event E, for $n > n_1$, there holds

$$\begin{aligned} |\eta_m^u(x') - \eta_m^u(x)| &\leq c_L\|x' - x\|^\alpha \big(1 - (2/\underline{\pi})\sqrt{\log n/n}\big)^{-1} \bigg(\sum_{m=1}^M n\eta_m(x)/n_{(m)} \bigg)^{-1} \\ &\leq c_L\|x' - x\|^\alpha \big(1 - (2/\underline{\pi})\sqrt{\log n/n}\big)^{-1} \big(1 + (2/\overline{\pi})\sqrt{\log n/n}\big) \leq 4c_L\|x' - x\|^\alpha. \end{aligned}$$

Thus, (50) implies that (46) and (47) hold with probability P^n at least $1 - 2M/n^3$ if $n > n_1$, which completes the proof. ■

The following technical lemma is needed in the proof of Proposition 8.

Lemma 22 *Let $Z_1, \ldots, Z_n$ be a sequence of independent zero-mean real-valued random variables with $|Z_i| \leq C$ for some constant $C > 0$. Let $(v_1, \ldots, v_n)$ be a weight vector, with $v_{\max} = \max_i |v_i| > 0$. Then for all $\varepsilon > 0$, we have*

$$\mathrm{P}\Big(\sum_{i=1}^n v_i Z_i \geq \varepsilon\Big) \leq 2\exp\Big(-\frac{\varepsilon^2}{2C^2 v_{\max}\sum_{i=1}^n |v_i|}\Big).$$

Proof [of Lemma 22] For $1 \leq i \leq n$, we have $|v_i Z_i| \leq C v_i$. Applying Hoeffding's inequality in Lemma 17, we get $\mathrm{P}\big(\sum_{i=1}^n v_i Z_i \geq \varepsilon\big) \leq 2\exp\big(-2\varepsilon^2/\sum_{i=1}^n (2Cv_i)^2\big)$. This together with the inequality $\sum_{i=1}^n (2Cv_i)^2 = 4C^2 \sum_{i=1}^n v_i^2 \leq 4C^2 v_{\max} \sum_{i=1}^n |v_i|$ yields the assertion. ■

Proof [of Proposition 8] By the definition of $\widehat{\eta}^{k,u}$ and $\overline{\eta}^{k,u}$, we have

$$\widehat{\eta}_m^{k,u}(x) - \overline{\eta}_m^{k,u}(x) = \frac{1}{k}\sum_{i=1}^k \big(\mathbf{1}\{Y_{(i)}^u(x) = m\} - \eta_m^u(X_{(i)}^u(x))\big).$$

Conditional on D_b^u, the random variables $\mathbf{1}\{Y_{(1)}^u(x) = m\} - \eta_m^u(X_{(1)}^u(x)), \ldots, \mathbf{1}\{Y_{(s_u)}^u(x) = m\} - \eta_m^u(X_{(s_u)}^u(x))$ are independent with zero mean and $|\mathbf{1}\{Y_{(1)}^u(x) = m\} - \eta_m^u(X_{(1)}^u(x))| \leq 1$. Applying Lemma 22, we get $(\mathrm{P}_{Y|X}^u)^{s_u}\big(|\widehat{\eta}_m^{k,u}(x) - \overline{\eta}_m^{k,u}(x)| \geq \varepsilon | D_n^u\big) \leq 2\exp(-\varepsilon^2 k/2)$. Setting $\varepsilon := \sqrt{2(2d+3)\log s_u / k}$, we get

$$(\mathrm{P}_{Y|X}^u)^{s_u}\big(|\widehat{\eta}_m^{k,u}(x) - \overline{\eta}_m^{k,u}(x)| \geq \varepsilon | D_n^u\big) \leq 2 s_u^{-(2d+3)}. \tag{51}$$

Note that this inequality holds only for a fixed x. To derive the uniform upper bound over $\mathcal{X}$, let $\mathcal{S} := \big\{(\sigma_1, \ldots, \sigma_{s_u}) : \text{all permutations of } (1, \ldots, s_u) \text{ obtainable by moving } x \in \mathbb{R}^d\big\}$. Then we have

$$\begin{aligned}
&(\mathrm{P}_{Y|X}^u)^{s_u}\Big(\sup_{x \in \mathbb{R}^d}\big(|\widehat{\eta}_m^{k,u}(x) - \overline{\eta}_m^{k,u}(x)| - \varepsilon\big) > 0 \Big| D_n^u\Big) \\
&\leq (\mathrm{P}_{Y|X}^u)^{s_u}\Big(\bigcup_{(\sigma_1, \ldots, \sigma_{s_u}) \in \mathcal{S}} \Big|\sum_{i=1}^k k^{-1}(\mathbf{1}\{Y_{\sigma_i} = m\} - \eta_m^u(X_{\sigma_i}))\Big| > \varepsilon \Big| D_n^u\Big) \\
&\leq \sum_{(\sigma_1, \ldots, \sigma_{s_u}) \in \mathcal{S}} (\mathrm{P}_{Y|X}^u)^{s_u}\Big(\Big|\sum_{i=1}^k k^{-1}(\mathbf{1}\{Y_{\sigma_i} = m\} - \eta_m^u(X_{\sigma_i}))\Big| > \varepsilon \Big| D_n^u\Big).
\end{aligned}$$

For any $(\sigma_1, \ldots, \sigma_{s_u}) \in \mathcal{S}$, by (51), we have

$$(\mathrm{P}_{Y|X}^u)^{s_u}\Big(\Big|\sum_{i=1}^k k^{-1}\big(\mathbf{1}\{Y_{\sigma_i} = m\} - \eta_m^u(X_{\sigma_i})\big)\Big| > \varepsilon \Big| D_n^u\Big) \leq 2/s_u^{2d+3}.$$

This together with Lemma 20 implies

$$(\mathrm{P}_{Y|X}^u)^{s_u}\Big(\sup_{x \in \mathbb{R}^d}(|\widehat{\eta}_m(x) - \overline{\eta}_m(x)| - \varepsilon) > 0 \Big| D_n^u\Big) \leq 2(25/d)^d / s_u^3.$$

when $s_u \geq 2d$. Then a union bound argument with $c_1 := 2(2d+3)$ yields

$$(\mathrm{P}^u_{Y|X})^{s_u}\big(\|\widehat{\eta}^{k,u} - \overline{\eta}^{k,u}\|_\infty \leq \sqrt{c_1 \log s_u / k}\big| D^u_n\big) \geq 1 - 2M(25/d)^d / s_u^3.$$

By (46) in Lemma 21, if $n \geq n_1$, then we have $s_u \geq n_{(1)} \geq n\underline{\pi}/2$ with probability P^n at least $1 - 2M/n^3$. Consequently, if $n > \max\{n_1, \lceil 4d/\underline{\pi}\rceil\}$, there holds $\|\widehat{\eta}^{k,u}(x) - \overline{\eta}^{k,u}(x)\|_\infty \leq \sqrt{c_1 \log s_u / k}$ with probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - (2M + 16M(25/d)^d / \underline{\pi}^3)/n^3$. Therefore, if $n \geq N_1 := \max\{n_1, \lceil 4d/\underline{\pi}\rceil, \lceil 2M + 16M(25/d)^d/\underline{\pi}\rceil\}$, then we have $\|\widehat{\eta}^{k,u}(x) - \overline{\eta}^{k,u}(x)\|_\infty \leq \sqrt{c_1 \log s_u / k}$ with probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - 1/n^2$, which completes the proof. ■

7.1.2 Proofs Related to Section 5.1.2

To conduct our analysis, we first need to recall the definitions of *VC dimension* (*VC index*) and *covering number*, which are frequently used in capacity-involved arguments and measure the complexity of the underlying function class (van der Vaart and Wellner, 1996; Kosorok, 2008; Giné and Nickl, 2021).

Definition 23 (VC dimension) *Let $\mathcal{B}$ be a class of subsets of $\mathcal{X}$ and $A \subset \mathcal{X}$ be a finite set. The trace of $\mathcal{B}$ on A is defined by $\{B \cap A : B \subset \mathcal{B}\}$. Its cardinality is denoted by $\Delta^{\mathcal{B}}(A)$. We say that $\mathcal{B}$ shatters A if $\Delta^{\mathcal{B}}(A) = 2^{\#(A)}$, that is, if for every $A' \subset A$, there exists a $B \subset \mathcal{B}$ such that $A' = B \cap A$. For $n \in \mathrm{N}$, let*

$$m^{\mathcal{B}}(n) := \sup_{A \subset \mathcal{X},\, \#(A) = n} \Delta^{\mathcal{B}}(A). \tag{52}$$

Then, the set $\mathcal{B}$ is a Vapnik-Chervonenkis class if there exists $n < \infty$ such that $m^{\mathcal{B}}(n) < 2^n$ and the minimal of such n is called the VC dimension of $\mathcal{B}$, and abbreviate as $\mathrm{VC}(\mathcal{B})$.

Since an arbitrary set of n points $\{x_1, \ldots, x_n\}$ possess 2^n subsets, we say that $\mathcal{B}$ *picks out* a certain subset from $\{x_1, \ldots, x_n\}$ if this can be formed as a set of the form $B \cap \{x_1, \ldots, x_n\}$ for a $B \in \mathcal{B}$. The collection $\mathcal{B}$ *shatters* $\{x_1, \ldots, x_n\}$ if each of its 2^n subsets can be picked out in this manner. From Definition 23 we see that the VC dimension of the class $\mathcal{B}$ is the smallest n for which no set of size n is shattered by $\mathcal{B}$, that is,

$$\mathrm{VC}(\mathcal{B}) = \inf\Big\{n : \max_{x_1, \ldots, x_n} \Delta^{\mathcal{B}}(\{x_1, \ldots, x_n\}) \leq 2^n\Big\},$$

where $\Delta^{\mathcal{B}}(\{x_1, \ldots, x_n\}) = \#\{B \cap \{x_1, \ldots, x_n\} : B \in \mathcal{B}\}$. Clearly, the more refined $\mathcal{B}$ is, the larger is its index.

Definition 24 (Covering Number) *Let $(\mathcal{X}, d)$ be a metric space and $A \subset \mathcal{X}$. For $\varepsilon > 0$, the ε-covering number of A is denoted as*

$$\mathcal{N}(A, d, \varepsilon) := \min\Big\{n \geq 1 : \exists x_1, \ldots, x_n \in \mathcal{X} \text{ such that } A \subset \bigcup_{i=1}^n B(x_i, \varepsilon)\Big\},$$

where $B(x, \varepsilon) := \{x' \in \mathcal{X} : d(x, x') \leq \varepsilon\}$.

The following Lemma, which is needed in the proof of Lemma 26, provides the covering number of the indicator functions on the collection of the balls in $\mathbb{R}^d$.

Lemma 25 *Let* $\mathcal{B} := \{B(x,r) : x \in \mathbb{R}^d, r > 0\}$ *and* $\mathbf{1}_{\mathcal{B}} := \{\mathbf{1}_B : B \in \mathcal{B}\}$. *Then for any* $\varepsilon \in (0,1)$*, there exists a universal constant* C *such that*

$$\mathcal{N}(\mathbf{1}_{\mathcal{B}}, \|\cdot\|_{L_1(\mathrm{Q})}, \varepsilon) \leq C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}$$

holds for any probability measure Q.

Proof [of Lemma 25] For the collection $\mathcal{B} := \{B(x,r) : x \in \mathbb{R}^d, r > 0\}$, Dudley (1979) shows that for any set $A \in \mathbb{R}^d$ of $d+2$ points, not all subsets of A can be formed as a set of the form $B \cap A$ for a $B \in \mathcal{B}$. In other words, $\mathcal{B}$ can not pick out all subsets from $A \in \mathbb{R}^d$ of $d+2$ points. Therefore, the collection $\mathcal{B}$ fails to shatter A. Consequently, according to Definition 23, we have $\mathrm{VC}(\mathcal{B}) = d+2$. Using Theorem 9.2 in Kosorok (2008), we immediately obtain the assertion. ■

To prove Proposition 9 and Lemma 31, we need the following lemma, which provides a high probability uniform bound on the distance between any point and its k-th nearest neighbor. When we consider the approximation of the under-sampling k-NN classifier, the sub-samples need to follow the under-sampling distribution P^u whose density satisfies $f_X^u(x) > 1/(2M\overline{\pi})$ with high probability as shown in (46). Therefore, in Lemma 26 it is necessary to assume that the density is bounded below by a positive constant.

Lemma 26 *Let* $R_{(k)}(x) := \|X_{(k)}(x) - x\|$ *denote the distance from* x *to its* k*-th nearest neighbor,* $1 \leq k \leq n$. *Moreover, let* f_X *be the density function of* P_X *and suppose that there exists a constant* $\underline{c} > 0$ *such that* $f_X(x) \geq \underline{c}$. *Then there exists an* $n_2 \in \mathbb{N}$ *and* $c_0 = 2/\underline{c} > 0$ *such that for all* $n > n_2$*, there holds*

$$\sup_{x\in\mathcal{X}} \sup_{k \geq 48(2d+9)\log n} R_{(k)}(x)^{\alpha} \leq (c_0 k/n)^{\alpha/d}$$

with probability P^n *at least* $1 - 1/n^3$.

Proof [of Lemma 26] For $x \in \mathcal{X}$ and $\eta \in [0,1]$, we define the η-quantile diameter

$$\rho_x(\eta) := \inf\bigl\{r : \mathrm{P}(B(x,r)) \geq \eta\bigr\}.$$

Let us first consider the set $\mathcal{B}_k^+ := \bigl\{B\bigl(x, \rho_x\bigl((k+\sqrt{3\tau k})/n\bigr)\bigr) : x \in \mathcal{X}\bigr\} \subset \mathcal{B}$. Lemma 25 implies that for any probability Q, there holds

$$\mathcal{N}(\mathbf{1}_{\mathcal{B}_k^+}, \|\cdot\|_{L_1(\mathrm{Q})}, \varepsilon) \leq \mathcal{N}(\mathbf{1}_{\mathcal{B}}, \|\cdot\|_{L_1(\mathrm{Q})}, \varepsilon) \leq C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}. \tag{53}$$

By the definition of the covering number, there exists an ε-net $\{A_j\}_{j=1}^J \subset \mathcal{B}_k^+$ with $J := \lfloor C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}\rfloor$ and for any $x \in \mathcal{X}$, there exists some $j \in \{1,\ldots,J\}$ such that

$$\bigl\|\mathbf{1}\bigl\{B\bigl(x, \rho_x\bigl((k+\sqrt{3\tau k})/n\bigr)\bigr)\bigr\} - \mathbf{1}_{A_j}\bigr\|_{L_1(D)} \leq \varepsilon. \tag{54}$$

For any $i = 1, \ldots, n$, let the random variables ξ_i be defined by $\xi_i = \mathbf{1}_{A_j}(X_i) - (k + \sqrt{3\tau k})/n$. Then we have $\|\xi_i\|_\infty \leq 1$, $\mathbb{E}_{\mathrm{P}}\xi_i = 0$ and $\mathbb{E}_{\mathrm{P}}\xi_i^2 \leq \mathbb{E}_{\mathrm{P}}\xi_i = (k + \sqrt{3\tau k})/n$. Applying Bernstein's inequality in Lemma 18, we obtain

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{A_j}(X_i) - (k + \sqrt{3\tau k})/n \geq -\sqrt{2\tau(k + \sqrt{3\tau k})/n} - 2\tau/(3n)$$

with probability P^n at least $1 - e^{-\tau}$. Then the union bound together with the covering number estimate (53) implies that for any A_j, $j = 1, \cdots, J$, there holds

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{A_j}(X_i) - (k + \sqrt{3(\tau + \log J)k})/n \\ &\geq -\sqrt{2(\tau + \log J)\big(k + \sqrt{3(\tau + \log J)k}\big)/n} - 2(\tau + \log J)/(3n)\end{aligned}$$

with probability P^n at least $1 - e^{-\tau}$. This together with (54) yields that for any $x \in \mathcal{X}$, there holds

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbf{1}\big\{B\big(x, \rho_x\big((k + \sqrt{3\tau k})/n\big)\big)\big\}(X_i) - (k + \sqrt{3(\tau + \log J)k})/n \\ &\geq -\sqrt{2(\tau + \log J)\big(k + \sqrt{3(\tau + \log J)k}\big)/n} - 2(\tau + \log J)/(3n) - \varepsilon\end{aligned}$$

with probability P^n at least $1 - e^{-\tau}$.

Now, if we take $\varepsilon = 1/n$, then for any $n > n_2 := \max\{4e, d + 2, C\}$, there holds

$$\log J = \log C + \log(d + 2) + (d + 2)\log(4e) + (d + 1)\log n \leq (2d + 5)\log n.$$

Let $\tau := \log(n^4)$. A simple calculation yields that if $k \geq 48(2d + 9)\log n$, then we have

$$\sqrt{2(\tau + \log J)\big(k + \sqrt{3(\tau + \log J)k}\big)/n} \leq \sqrt{5(\tau + \log J)k/2/n}.$$

Consequently, for all $n > n_2$, there holds

$$\sqrt{2(\tau + \log J)\big(k + \sqrt{3(\tau + \log J)k}\big)/n} + 2(\tau + \log J)/(3n) + 1/n \leq \sqrt{3(\tau + \log J)k/n}.$$

Consequently for all $x \in \mathcal{X}$, there holds $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\big\{B\big(x, \rho_x\big((k + \sqrt{3\tau k})/n\big)\big)\big\}(X_i) \geq k/n$ with probability P^n at least $1 - 1/n^4$. By the definition of $R_{(k)}(x)$, there holds

$$R_{(k)}(x) \leq \rho_x\big((k + \sqrt{3\tau k})/n\big) \tag{55}$$

with probability P^n at least $1 - 1/n^4$. For any $x \in \mathcal{X}$, we have $\mathrm{P}_X\big(B\big(x, \rho_x\big((k+\sqrt{3\tau k})/n\big)\big)\big) = (k + \sqrt{3\tau k})/n$. Since the density $f_X(x)$ satisfies $\underline{c} \leq f_X(x) \leq \overline{c}$, we have $(k + \sqrt{3\tau k})/n \geq \underline{c}\rho_x^d\big((k + \sqrt{3\tau k})/n\big)$ and consequently

$$\rho_x\big((k + \sqrt{3\tau k})/n\big) \leq \big((k + \sqrt{3\tau k})/(\underline{c}n)\big)^{1/d} \leq \big((k + 3\sqrt{k\log n})/(\underline{c}n)\big)^{1/d}. \tag{56}$$

Combining (55) with (56), we obtain that $R_{(k)}(x) \leq \big((k + 3\sqrt{k \log n})/(\underline{c}n)\big)^{1/d}$ holds for all $x \in \mathcal{X}$ with probability P^n at least $1 - 1/n^4$. Therefore, a union bound argument yields that for all $x \in \mathcal{X}$, there holds

$$\sup_{k \geq 48(2d+9)\log n} R_{(k)}(x) \leq \big((k + 3\sqrt{k \log n})/(\underline{c}n)\big)^{1/d} \leq \big(2k/(\underline{c}n)\big)^{1/d}$$

with probability P^n at least $1 - 1/n^3$ for all $n \geq n_2$, which yields to the assertion with $c_0 = 2/\underline{c}$. ■

Proof [of Proposition 9] Let $n_1 \in \mathbb{N}_+$ satisfy $\log n_1/n_1 \leq \min\{\overline{\pi}^2/4, \underline{\pi}^2/16\}$. By Lemma 21, we see that for $1 \leq m \leq M$, if $n > n_1$, there hold

$$n\pi_m/2 \leq n_{(m)}, \qquad 1/(2M\overline{\pi}) \leq f_X^u(x) \leq 2/(M\underline{\pi}), \tag{57}$$

and

$$|\eta_m^u(x') - \eta_m^u(x)| \leq 4c_L\|x - x'\|^\alpha, \tag{58}$$

with probability P^n at least $1 - 2M/n^3$, since (57) implies that P^u satisfies the assumptions of Lemma 26, we find that for all $x \in \mathcal{X}$ and $s_u \geq n_2$, there holds

$$\sup_{k \geq 48(2d+9)\log s_u} \|X_{(k)}^u(x) - x\|^\alpha \leq (4M\overline{\pi}k/s_u)^{\alpha/d} \tag{59}$$

with probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - 1/s_u^3$. By using $s_u \geq n_{(1)} \geq n\pi_1/2$ in (57), we find that if $n \geq N_2 := \max\{n_1, \lceil 2n_2/\underline{\pi}\rceil, \lceil 8/\underline{\pi}^3 + 2M\rceil\}$, then (59) holds with probability $\mathrm{P}^n \otimes \mathrm{P}_Z$ at least $1 - 1/s_u^3 - 2M/n^3 \geq 1 - (8/\underline{\pi}^3 + 2M)/n^3 \geq 1 - 1/n^2$.

Let $R_{(i)}^u(x) := \|X_{(i)}^u(x) - x\|$ and $a_n^u := \lceil 48(2d+9)\log s_u\rceil$. Then (58) implies that for any $1 \leq m \leq M$,

$$\begin{aligned}
\sum_{i=1}^k k^{-1}\big|\eta_m^u(X_{(i)}^u(x)) - \eta_m^u(x)\big| &\leq (4/k)\sum_{i=1}^{a_n^u} c_L(R_{(i)}^u(x))^\alpha + (4/k)\sum_{i=a_n^u+1}^{k} c_L(R_{(i)}^u(x))^\alpha \\
&\leq (4/k)\sum_{i=1}^{a_n^u} c_L(R_{(a_n^u+1)}^u(x))^\alpha + (4/k)\sum_{i=a_n^u+1}^{k} c_L(R_{(i)}^u(x))^\alpha \\
&\leq (4c_La_n^u/k)\big(4M\overline{\pi}(a_n^u+1)/s_u\big)^{\alpha/d} + \sum_{i=1}^k (4c_L/k)(4M\overline{\pi}i/s_u)^{\alpha/d} \\
&\leq 4c_L(8M\overline{\pi}k/s_u)^{\alpha/d} + \sum_{i=1}^k (4c_L/k)(4M\overline{\pi}i/s_u)^{\alpha/d}.
\end{aligned}$$

Since the function $g(t) := t^{\alpha/d}$ is increasing in $(0, \infty)$, we have

$$\sum_{i=1}^k k^{-1}(i/s_u)^{\alpha/d} \leq (s_u/k)\int_0^{(k+1)/s_u} g(t)\,dt$$

$$\leq (s_u/k)(d/(\alpha+d))\big((k+1)/s_u\big)^{(\alpha+d)/d} \leq 2^{(\alpha+d)/d}(d/(\alpha+d))(k/s_u)^{\alpha/d},$$

which yields $\|\overline{\eta}^{k,u}(x)-\eta^u(x)\|_\infty \leq c_2(k/s_u)^{\alpha/d}$ with constant $c_2 := 4c_L(8M\overline{\pi})^{\alpha/d}+4c_L(4M\overline{\pi})^{\alpha/d}\cdot 2^{(\alpha+d)/d}d/(\alpha+d)$. Thus we finish the proof. ■

7.1.3 Proofs Related to Section 5.1.3

Proof [of Proposition 10] Let the event E be defined by

$$E := \Big\{1-(2/\underline{\pi})\sqrt{\log n/n} \leq n_{(m)}/(n\pi_m) \leq 1+(2/\overline{\pi})\sqrt{\log n/n},\ \forall 1\leq m\leq M\Big\}$$

and the number $n_1 \in \mathbb{N}_+$ satisfy $\log n_1/n_1 \leq \min\{\overline{\pi}^2/4, \underline{\pi}^2/16\}$. The following arguments will be made on the event E and for $n > n_1$. An elementary calculation implies that for any $1 \leq m \leq M$ and $x \in \mathcal{X}$, there holds

$$\begin{aligned}\big|n\eta_m(x)/n_{(m)} - \eta_m(x)/\pi_m\big| &\leq \big|n/n_{(m)} - 1/\pi_m\big| = \big|(n\pi_m - n_{(m)})/(n_i\pi_m)\big| \\ &= |1-n_{(m)}/(n\pi_m)|/\big(\pi_m(n_{(m)}/(n\pi_m)-1)+\pi_m\big).\end{aligned}$$

Thus, on the event E, there holds

$$\big|n\eta_m(x)/n_{(m)} - \eta_m(x)/\pi_m\big| \leq (2/\overline{\pi})\sqrt{\log n/n}/\big(\pi_m - (2/\underline{\pi})\sqrt{\log n/n}\pi_m\big).$$

Therefore, for $n > n_1$, we have

$$\big|n\eta_m(x)/n_{(m)} - \eta_m(x)/\pi_m\big| \leq (4/\pi_m\overline{\pi})\sqrt{\log n/n} \leq (4/(\overline{\pi}\underline{\pi}))\sqrt{\log n/n}. \tag{60}$$

Consequently, for any $1 \leq m \leq M$ and $x \in \mathcal{X}$, there holds

$$\begin{aligned}\big|\eta_m^w(x)-\eta_m^u(x)\big| &\leq \bigg|\frac{n\eta_m(x)/n_{(m)}}{\sum_{i=1}^M n\eta_i(x)/n_i} - \frac{\eta_m(x)/\pi_m}{\sum_{i=1}^M \eta_i(x)/\pi_i}\bigg| \\ &\leq \bigg|\frac{n\eta_m(x)/n_{(m)}}{\sum_{i=1}^M n\eta_i(x)/n_i} - \frac{n\eta_m(x)/n_{(m)}}{\sum_{i=1}^M \eta_i(x)/\pi_i}\bigg| + \bigg|\frac{n\eta_m(x)/n_{(m)}}{\sum_{i=1}^M \eta_i(x)/\pi_i} - \frac{\eta_m(x)/\pi_m}{\sum_{i=1}^M \eta_i(x)/\pi_i}\bigg| \\ &\leq \frac{|n\eta_m(x)/n_{(m)} - \eta_m(x)/\pi_m|}{\sum_{i=1}^M \eta_i(x)/\pi_i} + \frac{\sum_{m=1}^M |n\eta_m(x)/n_{(m)} - \eta_m(x)/\pi_m|}{\sum_{i=1}^M \eta_i(x)/\pi_i}.\end{aligned} \tag{61}$$

Obviously, we have $\sum_{i=1}^M \eta_i(x)/\pi_i \geq \sum_{i=1}^M \eta_i(x)/\overline{\pi} = 1/\overline{\pi}$. This together with (60) and (61) yields that $\|\eta_m^w - \eta_m^u\|_\infty \leq (4(1+M)/\underline{\pi})\sqrt{\log n/n}$ holds for $1 \leq m \leq M$, which together with (50) implies that for all $n > n_1$, (34) holds with probability P^n at least $1-2M/n^3$. Therefore, for all $n \geq N_3 := \max\{n_1, 2M\}$, there holds (34) with probability P^n at least $1-1/n^2$. Thus we complete the proof of Proposition 10. ■

7.1.4 Proofs Related to Section 4.1

Proof [of Theorem 1] Propositions 8 and 9 imply that if $n \geq \max\{N_1, N_2\}$, there hold $\|\overline{\eta}^{k,u}(x) - \eta^u(x)\|_\infty \lesssim (k/s_u)^{\alpha/d}$ and $\|\widehat{\eta}^{k,u}(x) - \overline{\eta}^{k,u}(x)\|_\infty \lesssim \sqrt{\log s_u/k}$ with probability $\mathrm{P}_Z \otimes \mathrm{P}^\mathrm{n}$ at least $1 - 2/n^2$. Consequently, we have

$$\begin{aligned}\|\widehat{\eta}^{k,u}(x) - \eta^u(x)\|_\infty &\leq \|\widehat{\eta}^{k,u}(x) - \overline{\eta}^{k,u}(x)\|_\infty + \|\overline{\eta}^{k,u}(x) - \eta^u(x)\|_\infty \\ &\lesssim (k/s_u)^{\alpha/d} + \sqrt{\log s_u/k} \lesssim (\log s_u/s_u)^{\alpha/(2\alpha+d)}.\end{aligned}$$

According to (46) in Lemma 21, we have $s_u \geq n_{(1)} \geq n\pi_1/2$ with probability P^n at least $1 - 2M/n^3 \geq 1 - 1/n^2$. Since $g(x) := \log(x)/x$ is decreasing on $[e, \infty)$, we have

$$\|\widehat{\eta}^{k,u}(x) - \eta^u(x)\|_\infty \lesssim (\log(n\pi_1/2)/(n\pi_1/2))^{\alpha/(2\alpha+d)} \lesssim (\log n/n)^{\alpha/(2\alpha+d)}.$$

This together with Proposition 10 yield that if $n \geq N_1^* := \max\{N_1, N_2, N_3\}$, there holds

$$\begin{aligned}\|\widehat{\eta}^{k,u}(x) - \eta^w(x)\|_\infty &= \|\widehat{\eta}^{k,u}(x) - \eta^u(x)\|_\infty + \|\eta^u(x) - \eta^w(x)\|_\infty \\ &\lesssim (\log n/n)^{\alpha/(2\alpha+d)} + \sqrt{\log n/n} \lesssim (\log n/n)^{\alpha/(2\alpha+d)}.\end{aligned}$$

with probability P^n at least $1 - 4/n^2$. Consequently, Lemma 15 yields that

$$\|\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) - \eta^w_{\widehat{\psi}^{k,u}(x)}(x)\|_\infty \lesssim (\log n/n)^{\alpha/(2\alpha+d)}$$

holds with probability $\mathrm{P}_Z \otimes \mathrm{P}^\mathrm{n}$ at least $1 - 4/n^2$, where $\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) = \eta^w_{\psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x)}(x)$, i.e, the Bayes classifier w.r.t. the classification loss L_{cl} and the balanced distribution P^w. Using Lemma 16, we obtain

$$\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\widehat{\psi}^{k,u}) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} \lesssim (\log n/n)^{\alpha(\beta+1)/(2\alpha+d)}$$

with probability $\mathrm{P}_Z \otimes \mathrm{P}^\mathrm{n}$ at least $1 - 4/n^2$. This together with (24) and Theorem 6 implies $\mathfrak{R}_{\mathrm{AM}}(\widehat{\psi}^{k,u}) \lesssim (\log n/n)^{\alpha(\beta+1)/(2\alpha+d)}$, which completes the proof. ■

Proof [of Theorem 2] We use Theorem 3.5 in (Audibert and Tsybakov, 2007) to prove Theorem 2. First of all, we verify as follows that for any probability distribution $\mathrm{P} \in \mathcal{P}$, the balanced version of it, P^w belongs to a certain class of probability distributions $\mathcal{P}_\Sigma$ in Definition 3.1 in Audibert and Tsybakov (2007).

(i) Condition *(i)* in Assumption 1 implies that P^w satisfies the margin assumption.

(ii) For any $1 \leq m \leq M$ and $x, x' \in \mathcal{X}$, by (2), we have

$$|\eta^w_m(x') - \eta^w_m(x)| = \frac{|\eta_m(x') - \eta_m(x)|}{\pi_m \sum_{m=1}^M \eta_m(x)/\pi_m} \leq (\overline{\pi}/\underline{\pi}) \cdot |\eta_m(x') - \eta_m(x)|.$$

Condition (ii) in Assumption 1 then yields $|\eta^w_m(x') - \eta^w_m(x)| \leq 4c_L(\overline{\pi}/\underline{\pi}) \cdot \|x' - x\|^\alpha$. Therefore, the posterior probability function η^w belongs to the Hölder class.

(iii) Since P_X is the uniform distribution on $[0,1]^d$, (27) yields

$$1/(M\overline{\pi}) \leq f^w_X(x) = f_X(x) \sum_{m=1}^M \eta_m(x)/(M\pi_m) \leq 1/(M\underline{\pi})$$

for $x \in [0,1]^d$ and $f_X^w(x) = 0$ otherwise, which implies that the strong density assumption on P_X^w is satisfied.

Therefore, applying Theorem 3.5 in (Audibert and Tsybakov, 2007), there exists a constant $C > 0$ such that for any $\psi_n \in \mathcal{F}$, there holds

$$\sup_{\mathrm{P}^w \in \mathcal{P}_\Sigma} \mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\psi_n) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} \geq C n^{-\alpha(\beta+1)/(2\alpha+d)}.$$

This together with (24) and Theorem 6 implies

$$\sup_{\mathrm{P} \in \mathcal{P}} r^*_{\mathrm{AM}} - r_{\mathrm{AM}}(\psi_n) \geq C n^{-\alpha(\beta+1)/(2\alpha+d)}.$$

Thus the assertion is proved since $\psi_n \in \mathcal{F}$ is an arbitrary classifier. ■

7.2 Proofs Related to the Under-bagging k-NN Classifier

In this section, we first present in Sections 7.2.1-7.2.3 the proofs of the theoretical results on bounding the bagging error in Section 5.2.1, the bagged approximation error in Section 5.2.2, and the bagged sample error in Section 5.2.3, respectively. Then in Section 7.1.4, we prove the main results on the convergence rates of the under-bagging k-NN classifier, i.e., Theorem 3 and Corollary 4 in Section 4.2.

7.2.1 Proofs Related to Section 5.2.1

Proof [of Proposition 11] By the definition of $\widehat{\eta}^{B,u}(x)$ and $\widetilde{\eta}^{B,u}(x)$, we have

$$|\widehat{\eta}_m^{B,u}(x) - \widetilde{\eta}_m^{B,u}(x)| = \biggl| \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n V_i^{b,u}(x) \mathbf{1}\{Y_{(i)}(x) = m\} - \sum_{i=1}^n \overline{V}_i^u(x) \mathbf{1}\{Y_{(i)}(x) = m\} \biggr|.$$

Let $\xi_b := \sum_{i=1}^n (V_i^{b,u}(x) - \overline{V}_i^u(x)) \mathbf{1}\{Y_{(i)}(x) = m\}$. Then we have

$$\|\xi_b\|_\infty \leq \max \biggl\{ \sum_{i=1}^n V_i^{b,u}(x), \sum_{i=1}^n \overline{V}_i^u(x) \biggr\} \leq 1.$$

This yields that $\mathbb{E}[\xi_b^2 | D_n] \leq 1$. Applying Bernstein's inequality in Lemma 18, we obtain that for every $\tau > 0$,

$$\mathrm{P}_Z^B \biggl(|\widehat{\eta}_m^{B,u}(x) - \widetilde{\eta}_m^{B,u}(x)| \geq \sqrt{\frac{2\tau}{B}} + \frac{2\tau}{3B} \biggm| D_n \biggr) \leq e^{-\tau}.$$

Then with $\tau := (2d+3)\log n$, we have

$$\mathrm{P}_Z^B \biggl(|\widehat{\eta}_m^{B,u}(x) - \widetilde{\eta}_m^{B,u}(x)| \leq \sqrt{\frac{2(2d+3)\log n}{B}} + \frac{2(2d+3)\log n}{3B} \biggm| D_n \biggr) \geq 1 - 1/n^{2d+3}.$$

Setting $\varepsilon := \sqrt{8(2d+3)\log n / B}$, then the condition $9B \geq 2(2d+3)\log n$ implies that

$$\mathrm{P}_Z^B (|\widehat{\eta}_m^{B,u}(x) - \widetilde{\eta}_m^{B,u}(x)| \leq \varepsilon | D_n) \geq 1 - 1/n^{2d+3}. \tag{62}$$

In order to derive the uniform upper bound over $\mathcal{X}$, let

$$\mathcal{S} := \big\{(\sigma_1, \ldots, \sigma_n) : \text{all permutations of } (1, \ldots, n) \text{ obtainable by moving } x \in \mathbb{R}^d\big\}.$$

Then we have

$$\begin{aligned}
&\mathrm{P}_Z^B\bigg(\sup_{x\in\mathbb{R}^d}\Big(|\widehat{\eta}_m^{B,u}(x) - \widetilde{\eta}_m^{B,u}(x)| - \varepsilon\Big) > 0\bigg| D_n\bigg) \\
&\leq \mathrm{P}_Z^B\bigg(\bigcup_{(\sigma_1,\ldots,\sigma_n)\in\mathcal{S}}\bigg|\frac{1}{B}\sum_{b=1}^B\sum_{i=1}^n V_{i,\sigma}^{b,u}(x)\mathbf{1}\{Y_{\sigma_i}(x) = m\} - \sum_{i=1}^n \overline{V}_{i,\sigma}^u(x)\mathbf{1}\{Y_{\sigma_i}(x) = m\}\bigg| > \varepsilon\bigg| D_n\bigg) \\
&\leq \sum_{(\sigma_1,\ldots,\sigma_n)\in\mathcal{S}} \mathrm{P}_Z^B\bigg(\bigg|\frac{1}{B}\sum_{b=1}^B\sum_{i=1}^n V_{i,\sigma}^{b,u}(x)\mathbf{1}\{Y_{\sigma_i}(x) = m\} - \sum_{i=1}^n \overline{V}_{i,\sigma}^u(x)\mathbf{1}\{Y_{\sigma_i}(x) = m\}\bigg| > \varepsilon\bigg| D_n\bigg),
\end{aligned}$$

where $V_{i,\sigma}^{b,u}(x)$ equals $1/k$ if $\sum_{j=1}^i Z^b(X_{\sigma_j}(x), Y_{\sigma_j}(x)) \leq k$, and 0 otherwise, and $\overline{V}_{i,\sigma}^u(x) = k^{-1}\mathrm{P}_Z(\sum_{j=1}^i Z^b(X_{\sigma_j}(x), Y_{\sigma_j}(x)) \leq k|\{X_i, Y_i\}_{i=1}^n)$. For any $(\sigma_1, \ldots, \sigma_n) \in \mathcal{S}$, (62) implies

$$\mathrm{P}_Z^B\bigg(\bigg|\frac{1}{B}\sum_{b=1}^B\sum_{i=1}^n V_{i,\sigma}^{b,u}(x)\mathbf{1}\{Y_{\sigma_i}(x) = m\} - \sum_{i=1}^n \overline{V}_{i,\sigma}^u(x)\mathbf{1}\{Y_{\sigma_i}(x) = m\}\bigg| > \varepsilon\bigg| D_n\bigg) \leq 1/n^{2d+3}.$$

This together with Lemma 20 yields

$$\mathrm{P}_Z^B\bigg(\sup_{x\in\mathbb{R}^d}(|\widehat{\eta}_m^{B,u}(x) - \widetilde{\eta}_m^{B,u}(x)| - \varepsilon) > 0\bigg| D_n\bigg) \leq (25/d)^d/n^3$$

for all $n \geq 2d$. Then a union bound argument implies

$$\mathrm{P}_Z^B\big(\|\widehat{\eta}^{B,u}(x) - \widetilde{\eta}^{B,u}(x)\|_\infty \leq \sqrt{8(2d+3)\log n/B}\big|D_n\big) \geq 1 - M(25/d)^d/n^3.$$

Consequently, if $n \geq N_4 := \max\{\lceil M(25/d)^d\rceil, 2d\}$, then we have

$$\mathrm{P}_Z^B \otimes \mathrm{P}^n\big(\|\widetilde{\eta}^{B,u}(x) - \overline{\eta}^{B,u}(x)\|_\infty \leq \sqrt{8(2d+3)\log n/B}\big) \geq 1 - M(25/d)^d n^3 \geq 1 - 1/n^2,$$

which completes the proof. ■

7.2.2 Proofs Related to Section 5.2.2

To bound the bagged approximation error $\|\overline{\eta}^{B,u}(x) - \eta^u(x)\|_\infty$, the theoretical property of $\overline{V}_i^u(x)$ plays a crucial role. In fact, by the definition of $\overline{\eta}^{B,u}$ and η^u, we have

$$\begin{aligned}
\|\overline{\eta}^{B,u}(x) - \eta^u(x)\|_\infty &= \bigg\|\sum_{i=1}^n \overline{V}_i^u(x)\eta_m^u(X_{(i)}(x)) - \eta^u(x)\bigg\|_\infty \\
&\leq \bigg\|\sum_{i=1}^n \overline{V}_i^u(x)\big(\eta_m^u(X_{(i)}(x)) - \eta^u(x)\big)\bigg\|_\infty + \bigg\|\sum_{i=1}^n \overline{V}_i^u(x) - 1\bigg\|_\infty.
\end{aligned}$$

Let $\xi_i := Z^b(X_{(i)}(x), Y_{(i)}(x)), i \geq 1$. According to (37), $\overline{V}_i^u(x)$ can be re-expressed as

$$\overline{V}_i^u(x) = \frac{1}{k} \mathrm{P}_Z\biggl(\sum_{l=1}^{i} \xi_l \leq k, \xi_i = 1 \bigg| \{(X_i, Y_i)\}_{i=1}^n\biggr) = \frac{1}{k} \sum_{j=1}^{k} p_{i,j}^u(x). \tag{63}$$

where

$$p_{i,j}^u(x) = \mathrm{P}_Z\biggl(\sum_{l=1}^{i} \xi_l = j,\ \xi_i = 1 \bigg| \{(X_i, Y_i)\}_{i=1}^n\biggr). \tag{64}$$

As a result, it is necessary to investigate the properties of $p_{i,j}^u(x)$ for all $x \in \mathcal{X}$. Note that $\{\xi_i, i \geq 1\}$ can be regarded as a sequence of independent Bernoulli trials with the probabilities of success $a(X_{(i)}(x), Y_{(i)}(x))$, that is,

$$\mathrm{P}(\xi_i = 1) = a(X_{(i)}(x), Y_{(i)}(x)), \tag{65}$$

$$\mathrm{P}(\xi_i = 0) = 1 - a(X_{(i)}(x), Y_{(i)}(x)). \tag{66}$$

Then by (64), $p_{i,j}^u(x)$ represents the probability when we observing the sequence $\{\xi_i, 1 \leq i \leq n\}$ until j successes have occurred, the total number of trials equals to i.

Recall that the classical Pascal distribution models the number of successes in a sequence of i.i.d. Bernoulli trials before a specified number of failures occurs. We refer the readers to Spiegel (1992); DeGroot (2012) for more details. However, (65) implies that the probabilities of success for the Bernoulli trials $\{\xi_i, i \geq 1\}$ are not the same. Therefore, it is necessary to consider a *Generalized Pascal* (GP) distribution where the probabilities of success depend on the location of x in $\mathbb{R}^d$. To this end, we introduce some basic notations. Suppose that $\{\xi_i, i \geq 1\}$ is a sequence of independent Bernoulli trials. In each trial, we have the probability of success $\mathrm{P}(\xi_i = 1) = p_i$ and the probability of failure $\mathrm{P}(\xi_i = 0) = 1 - p_i$. The total number of trials we have seen until j successes have occurred, namely X, are said to have the Generalized Pascal distribution with parameters j and $p = \{p_i\}_{i=1}^\infty$, that is, $X \sim \mathrm{GP}(j, p)$. For $i \geq j$, let $\Omega(j, i) := \{\omega = \{\omega_1, \ldots, \omega_j\} : 1 \leq \omega_1 < \omega_2 < \cdots < \omega_{j-1} < \omega_j = i\}$. Then the probability mass function of the Generalized Pascal distribution is

$$\mathrm{P}_{\mathrm{GP}}(X = i) = f_{\mathrm{GP}}(i; j, p) = \sum_{\omega \in \Omega(j,i)} p_\ell \prod_{i=1}^{i-1} p_i^{\boldsymbol{1}\{i \in \omega\}} (1 - p_i)^{\boldsymbol{1}\{i \notin \omega\}}, \quad i \geq j,\, i \in \mathbb{N}_+. \tag{67}$$

In fact, according to (64), $p_{i,j}^u(x)$ can be re-expressed as

$$p_{i,j}^u(x) = f_{\mathrm{GP}}(i, j, p(x)), \tag{68}$$

where $p(x) = (p_1(x), \ldots, p_n(x), \ldots)$ with elements defined by $p_i(x) = a(X_{(i)}(x), Y_{(i)}(x))$, $1 \leq i \leq n$. According to the acceptance probability (9) in Section 3.2, it is easy to see that for all $x \in \mathcal{X}$, we have

$$p(x) \in S_\nu := \biggl\{p = \{p_i\}_{i=1}^\infty : (p_1, \ldots, p_n) = (\nu_{\sigma_1}, \ldots, \nu_{\sigma_n}) \text{ and } p_i = \nu_n \text{ for } i > n, \\ \text{where } \{\sigma_1, \ldots, \sigma_n\} \text{ is a permutation of } \{1, \ldots, n\}\biggr\}. \tag{69}$$

where $\nu = (\nu_1, \ldots, \nu_n, \ldots)$ is an infinite-dimensional vector with elements defined by

$$\nu_i = \begin{cases} \frac{s}{Mn_{(M)}}, & \text{if } 1 \leq i \leq n_{(M)}, \\ \frac{s}{Mn_{(m)}}, & \text{if } \sum_{\ell=m+1}^{M} n_\ell < i \leq \sum_{\ell=m}^{M} n_\ell \text{ and } 1 \leq m \leq M-1, \\ \frac{s}{Mn_{(1)}}, & \text{if } i > n. \end{cases} \tag{70}$$

Therefore, in our analysis on $p_{i,j}^u(x)$, it suffices to study the property of the distribution function $f_{\mathrm{GP}}(i, j, p)$ with p restricted on the set S_ν.

In the following, we present some results on the Generalized Pascal distribution, which is later crucial for the proof of Lemmas 31 and 32. The first lemma gives a uniform upper bound of the tail probability of the Generalized Pascal distribution.

Lemma 27 *Let $p \in S_\nu$ and suppose that $\sum_{i=1}^{\ell} p_i \geq j$. Then we have*

$$\sum_{i=\ell+1}^{\infty} f_{\mathrm{GP}}(i; j, p) \leq \exp\biggl(-\frac{1}{2\ell}\biggl(\sum_{i=1}^{\ell} p_i - j\biggr)^2\biggr).$$

Proof [of Lemma 27] Let $\{\xi_i,\ i \geq 1\}$ be a sequence of independent Bernoulli trials such that $\mathrm{P}(\xi_i = 1) = p_i$ and $\mathrm{P}(\xi_i = 0) = 1 - p_i$. By the definition of the Generalized Pascal distribution, we have $\sum_{i=\ell+1}^{\infty} f_{\mathrm{GP}}(i; j, p) = \mathrm{P}\bigl(\sum_{i=1}^{\ell} \xi_i < j\bigr)$. Let $\xi_i' = 1 - \xi_i$ for $1 \leq i \leq \ell$. Then we have

$$\sum_{i=\ell+1}^{\infty} f_{\mathrm{GP}}(i; j, p) = \mathrm{P}\biggl(\sum_{i=1}^{\ell} \xi_i' > \ell - j\biggr) = \mathrm{P}\biggl(\frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i' > 1 - \frac{j}{\ell}\biggr). \tag{71}$$

Moreover, there hold $\mathrm{P}(\xi_i' = 1) = 1 - p_i$ and $\mathrm{P}(\xi_i' = 0) = p_i$. Consequently we have $\mathbb{E}[\xi_i'] = 1 - p_i$ and $\mathrm{Var}[\xi_i'] \leq 1/4$. Using Bernstein's inequality in Lemma 18, for any $\tau > 0$, there holds

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i' \geq \sqrt{\frac{\tau}{2\ell}} + \frac{2\tau}{3\ell} + 1 - \frac{1}{\ell}\sum_{i=1}^{\ell} p_i \tag{72}$$

with probability at most $e^{-\tau}$. Let $\tau := \frac{1}{2\ell}(\sum_{i=1}^{\ell} p_i - j)^2$. Then we have $\tau \leq (\ell - j)^2/(2\ell) \leq \ell/2$ and consequently (72) implies

$$\frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i' \geq \sqrt{\frac{2\tau}{\ell}} + 1 - \frac{1}{\ell}\sum_{i=1}^{\ell} p_i \geq \frac{1}{\ell}\biggl(\sum_{i=1}^{\ell} p_i - j\biggr) + 1 - \frac{1}{\ell}\sum_{i=1}^{\ell} p_i = 1 - \frac{j}{\ell}.$$

Therefore, we have

$$\mathrm{P}\biggl(\frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i' > 1 - \frac{j}{\ell}\biggr) \leq e^{-\tau}. \tag{73}$$

Combining (71) with (73), we obtain the assertion. ■

The next lemma provides the effect of changing the position of entries of $p \in S_\nu$ on the Generalized Pascal distribution.

Lemma 28 *Given $1 \leq q \leq n-1$, let $p \in S_\nu$ such that $p_q \geq p_{q+1}$. We define a map $h_q : S_\nu \to S_\nu$ such that the elements of $p' = h_q(p)$ satisfying*

$$p'_i := \begin{cases} p_i, & \text{if } i < q \text{ or } i > q+1, \\ p_{q+1}, & \text{if } i = q, \\ p_q, & \text{if } i = q+1. \end{cases} \tag{74}$$

Then for any $i \geq j$, the following statements hold:

(i) If $i < q$ or $i > q+1$, then we have

$$f_{\mathrm{GP}}(i; j, p) = f_{\mathrm{GP}}(i; j, p'); \tag{75}$$

(ii) If $i = q$, then we have

$$f_{\mathrm{GP}}(i; j, p) \geq f_{\mathrm{GP}}(i; j, p'); \tag{76}$$

(iii) If $i = q+1$, then we have

$$f_{\mathrm{GP}}(i; j, p) \leq f_{\mathrm{GP}}(i; j, p'). \tag{77}$$

Proof [of Lemma 28] *(i)* If $i < q$, by the definition of p' in (74), we have $p'_i = p_i$ for $i \leq i$. Thus, by (67), there holds

$$\begin{aligned} f_{\mathrm{GP}}(i; j, p) &= \sum_{\omega \in \Omega(j,i)} p_i \prod_{\ell=1}^{i-1} p_i^{\mathbf{1}\{\ell \in \omega\}} (1-p_i)^{\mathbf{1}\{\ell \notin \omega\}} \\ &= \sum_{\omega \in \Omega(j,i)} p'_i \prod_{\ell=1}^{i-1} p_i'^{\mathbf{1}\{\ell \in \omega\}} (1-p'_i)^{\mathbf{1}\{\ell \notin \omega\}} = \mathrm{P}_{\mathrm{GP}}(X' = i). \end{aligned}$$

If $i > q+1$, then we define the map $g : \Omega(j,i) \to \Omega(j,i)$ by

$$g(\omega) := \begin{cases} \omega, & \text{if } \{q, q+1\} \subset \omega \text{ or } \{q, q+1\} \not\subseteq \omega, \\ \omega \setminus \{q\} \cup \{q+1\}, & \text{if } q \in \omega \text{ and } q+1 \notin \omega, \\ \omega \setminus \{q+1\} \cup \{q\}, & \text{if } q+1 \in \omega \text{ and } q \notin \omega. \end{cases}$$

By the definition of p' in (74), for every $\omega \in \Omega(j,i)$, there holds

$$p_i \prod_{\ell=1}^{i-1} p_i^{\mathbf{1}\{\ell \in \omega\}} (1-p_i)^{\mathbf{1}\{\ell \notin \omega\}} = p'_i \prod_{\ell=1}^{i-1} p_i'^{\mathbf{1}\{\ell \in g(\omega)\}} (1-p'_i)^{\mathbf{1}\{\ell \notin g(\omega)\}}.$$

Taking the sum over all possible elements in $\Omega(j,i)$, we obtain

$$f_{\mathrm{GP}}(i; j, p) = \sum_{\omega \in \Omega(j,i)} p_i \prod_{\ell=1}^{i-1} p_i^{\mathbf{1}\{\ell \in \omega\}} (1-p_i)^{\mathbf{1}\{\ell \notin \omega\}} = \sum_{\omega \in \Omega(j,i)} p'_i \prod_{\ell=1}^{i-1} p_i'^{\mathbf{1}\{\ell \in g(\omega)\}} (1-p'_i)^{\mathbf{1}\{\ell \notin g(\omega)\}}.$$

It can be verified that $g : \Omega(j,i) \to \Omega(j,i)$ is a one-to-one map. Therefore, we have

$$f_{\text{GP}}(i;j,p) = \sum_{\omega \in \Omega(j,i)} p'_i \prod_{\ell=1}^{i-1} {p'_i}^{\mathbf{1}\{\ell \in \omega\}} (1-p'_i)^{\mathbf{1}\{\ell \notin \omega\}} = f_{\text{GP}}(i;j,p').$$

(ii) If $i = q$, then by (74), we have $p'_q = p_{q+1} \le p_q$ and $p'_i = p_i$ for $\ell < q$. Consequently we obtain

$$\begin{aligned} f_{\text{GP}}(q;j,p) &= \sum_{\omega \in \Omega(j,q)} p_q \prod_{\ell=1}^{q-1} p_i^{\mathbf{1}\{\ell \in \omega\}} (1-p_i)^{\mathbf{1}\{\ell \notin \omega\}} \\ &\ge \sum_{\omega \in \Omega(j,q)} p'_q \prod_{\ell=1}^{q-1} {p'_i}^{\mathbf{1}\{\ell \in \omega\}} (1-p'_i)^{\mathbf{1}\{\ell \notin \omega\}} = f_{\text{GP}}(q;j,p'). \end{aligned}$$

(iii) If $i = q+1$, we consider two specific cases: $i = j$ and $i > j$. Suppose that $i = q+1 = j$. Obviously, there holds

$$1 = \sum_{i=j}^{\infty} f_{\text{GP}}(i;j,p) = \sum_{i=j}^{\infty} f_{\text{GP}}(i;j,p'). \tag{78}$$

Then (78) together with (75) yields $f_{\text{GP}}(q+1;j,p) = f_{\text{GP}}(q+1;j,p')$. If $i = q+1 > j$, then we have $q \ge j$. Combining (75), (76), and (78), we obtain $f_{\text{GP}}(q+1;j,p) \ge f_{\text{GP}}(q+1;j,p')$, which completes the proof. ■

Lemma 29 *Given $1 \le q \le n-1$, let $p \in S_\nu$ such that $p_q \ge p_{q+1}$. Moreover, let $p' = h_q(p)$ be as in (74). Then we have*

$$\sum_{i=j}^{n} i f_{\text{GP}}(i;j,p) \le \sum_{i=j}^{n} i f_{\text{GP}}(i;j,p').$$

Proof [of Lemma 29] The proof can be divided into the following three cases.

If $j \le q$, then Lemma 28 together with (78) yields

$$f_{\text{GP}}(q;j,p) + f_{\text{GP}}(q+1;j,p) = f_{\text{GP}}(q;j,p') + f_{\text{GP}}(q+1;j,p'). \tag{79}$$

By (75), we have

$$\begin{aligned} &\sum_{i=j}^{n} i f_{\text{GP}}(i;j,p) - \sum_{i=j}^{n} i f_{\text{GP}}(i;j,p') \\ &= (q+1)(f_{\text{GP}}(q+1;j,p) - f_{\text{GP}}(q+1;j,p')) + q(f_{\text{GP}}(q;j,p) - f_{\text{GP}}(q;j,p'))). \end{aligned}$$

This together with (79) implies

$$\sum_{i=j}^{n} i f_{\text{GP}}(i;j,p) - \sum_{i=j}^{n} i f_{\text{GP}}(i;j,p') = f_{\text{GP}}(q;j,p') - f_{\text{GP}}(q;j,p) \le 0,$$

where the last inequality follows from (76).

If $j = q+1$, (75) together with (77) yields

$$\sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,p) - \sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,p') = (q+1)(f_{\mathrm{GP}}(q+1;j,p) - f_{\mathrm{GP}}(q+1;j,p')) \leq 0.$$

Otherwise if $j > q+1$, then (75) yields $\sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,p) = \sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,p')$, which completes the proof. ■

The next theorem provides the upper bound of the expectation of the truncated Generalized Pascal distribution, which is needed to prove Lemma 31.

Theorem 30 *Let $p \in S_\nu$. Then for any $j \leq u \leq n$ satisfying $\sum_{\ell=1}^{u} \nu_\ell > j$, there holds*

$$\sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,p) \leq \frac{j}{\nu_1}\Big(\frac{\nu_u}{\nu_1}\Big)^j + n \exp\Big(-\frac{1}{2u}\Big(\sum_{\ell=1}^{u} \nu_\ell - j\Big)^2\Big).$$

Proof [of Theorem 30] For any $p \in S_\nu$, we can exchange the positions of two adjacent entries of $(p_1, \ldots, p_n)$ successively to arrange the entry with larger value behind. It is easy to see that a finite number of such operations can change p to ν, that is, $(p_1, \ldots, p_n)$ is rearranged in ascending order. Therefore, there exists a series of probability sequences $\{p^t, 1 \leq t \leq T\}$ such that $p^{(1)} = p$, $p^{(T)} = \nu$, and $p^{(t+1)} = h_{q_t}(p^{(t)})$ for $2 \leq t \leq T-1$, that is, we exchange $p_{q_t}^{(t)}$ and $p_{q_t+1}^{(t)}$ in the t-th operation. By Lemma 29, we have

$$\sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,p) \leq \sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,\nu).$$

(67) implies that for $j \leq i \leq u$, there holds

$$\begin{aligned}
\sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,\nu) &\leq \sum_{\omega \in \Omega(j,i)} \nu_i \prod_{\ell=1}^{i-1} \nu_i^{\mathbf{1}\{\ell \in \omega\}} (1-\nu_i)^{\mathbf{1}\{\ell \notin \omega\}} \leq \sum_{\omega \in \Omega(j,i)} \nu_u \prod_{\ell=1}^{i-1} \nu_u^{\mathbf{1}\{\ell \in \omega\}} (1-\nu_1)^{\mathbf{1}\{\ell \notin \omega\}} \\
&\leq \Big(\frac{\nu_u}{\nu_1}\Big)^j \sum_{\omega \in \Omega(j,i)} \nu_1 \prod_{\ell=1}^{i-1} \nu_1^{\mathbf{1}\{\ell \in \omega\}} (1-\nu_1)^{\mathbf{1}\{\ell \notin \omega\}} = \binom{i-1}{j-1} \Big(\frac{\nu_u}{\nu_1}\Big)^j \nu_1^j (1-\nu_1)^{i-j}
\end{aligned}$$

and consequently we have

$$\begin{aligned}
\sum_{i=j}^{u} i f_{\mathrm{GP}}(i;j,\nu) &\leq \Big(\frac{\nu_u}{\nu_1}\Big)^j \sum_{i=j}^{u} i \binom{i-1}{j-1} \nu_1^j (1-\nu_1)^{i-j} \\
&\leq \Big(\frac{\nu_u}{\nu_1}\Big)^j \sum_{i=j}^{\infty} i \binom{i-1}{j-1} \nu_1^j (1-\nu_1)^{i-j} = \frac{j}{\nu_1}\Big(\frac{\nu_u}{\nu_1}\Big)^j.
\end{aligned} \tag{80}$$

Under the assumption $\sum_{\ell=1}^{u} \nu_\ell > j$, Theorem 27 yields

$$\sum_{i=u+1}^{\infty} f_{\mathrm{GP}}(i;j,\nu) < \exp\Big(-\frac{1}{2u}\Big(\sum_{\ell=1}^{u} \nu_\ell - j\Big)^2\Big).$$

Consequently we have

$$\sum_{i=u+1}^{n} i f_{\mathrm{GP}}(i;j,\nu) \leq n \sum_{i=u+1}^{n} f_{\mathrm{GP}}(i;j,\nu) \leq n \sum_{i=u+1}^{\infty} f_{\mathrm{GP}}(i;j,\nu) \leq n \exp\biggl(-\frac{1}{2u}\biggl(\sum_{\ell=1}^{u} \nu_\ell - j\biggr)^2\biggr). \tag{81}$$

Combining (80) and (81), we obtain

$$\sum_{i=j}^{n} i f_{\mathrm{GP}}(i;j,p) \leq \frac{j}{\nu_1}\biggl(\frac{\nu_u}{\nu_1}\biggr)^j + n \exp\biggl(-\frac{1}{2u}\biggl(\sum_{\ell=1}^{u} \nu_\ell - j\biggr)^2\biggr),$$

which finishes the proof. ∎

To prove Proposition 12, we need the following two lemmas. Lemma 31 provides the uniform upper bound of the weighted sum of the i-th nearest neighbor distance $R_{(i)}(x)$, which supplies the key to the proof of the bagged approximation error term. Lemma 32 bounds the sum of the bagged weights $\overline{V}_i^u(x)$ uniformly.

Lemma 31 *Let $\overline{V}_i^u(x)$ be defined as in (37) and $R_{(i)}(x) := \|X_{(i)}(x) - x\|$. Suppose that $s\exp(-(s/M-k)^2/(2n)) \leq M\underline{\pi}/2$. Then there exists a constant $c_3 > 0$ and an $n_3 \in \mathbb{N}$ such that for all $n \geq n_3$, with probability P^n at least $1-(2M+1)/n^3$, for all $x \in \mathcal{X}$, there holds*

$$\sum_{i=1}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) \leq c_3 (k/s)^{\alpha/d}.$$

Proof [of Lemma 31] Let $a_n := \lceil 48(2d+9)\log n \rceil$. Lemma 26 implies that if $n > n_2$, then for all $x \in \mathcal{X}$, there holds $\sup_{i \geq a_n} R_{(i)}(x) \leq (2i/n)^{1/d}$ with probability P^n at least $1 - 1/n^3$. Then we have

$$\begin{aligned}
\sum_{i=1}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) &= \sum_{i=1}^{a_n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) \\
&\leq R_{(a_n)}^\alpha(x) + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) \leq (2a_n/n)^{\alpha/d} + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) \\
&\leq (2k/n)^{\alpha/d} + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) \leq (2k/s)^{\alpha/d} + \sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x).
\end{aligned} \tag{82}$$

For the second term in (82), there holds

$$\begin{aligned}
\sum_{i=a_n}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) &= \sum_{i=a_n}^{n} k^{-1} \sum_{j=1}^{k} p_{i,j}^u(x) R_{(i)}^\alpha(x) \\
&\leq \sum_{i=1}^{n} k^{-1} \sum_{j=1}^{k} p_{i,j}^u(x) (2i/n)^{\alpha/d} = k^{-1}(2/n)^{\alpha/d} \sum_{j=1}^{k} \sum_{i=1}^{n} i^{\alpha/d} p_{i,j}^u(x).
\end{aligned} \tag{83}$$

By Jensen's inequality, we have

$$\sum_{i=1}^{n} i^{\alpha/d} p_{i,j}^u(x) \leq \biggl(\sum_{i=1}^{n} p_{i,j}^u(x)\biggr)^{1-\alpha/d} \biggl(\sum_{i=1}^{n} i p_{i,j}^u(x)\biggr)^{\alpha/d} \leq \biggl(\sum_{i=1}^{n} i p_{i,j}^u(x)\biggr)^{\alpha/d}. \tag{84}$$

Since for all $x \in \mathcal{X}$, we have $p_{i,j}^u(x) = f_{\mathrm{GP}}(i,j,p(x))$ by (68), where $p(x) \in S_\nu$ with S_ν defined by (70). Applying Theorem 30 with $u = n_{(M)}$, we have

$$\sum_{i=1}^{n} i p_{i,j}^u(x) = \sum_{i=1}^{n} i f_{\mathrm{GP}}(i;j,p(x)) \leq Mjn_{(M)}/s + n \exp\bigl(-(s/M-j)^2/(2n_{(M)})\bigr)$$

for all $x \in \mathcal{X}$. According to (46) in Lemma 21 for all $n > n_1$, with probability P^n at least $1 - 2M/n^3$, there holds $n_{(M)} \geq n\pi_m/2$. This together with the condition $s \exp(-(s/M - k)^2/(2n)) \leq M\underline{\pi}/2$ and $n_{(M)} \geq n\pi_M/2$ yields that for all $n \geq n_3 := \max\{n_1, n_2\}$, there holds $\sum_{i=1}^n i p_{i,j}^u(x) \leq 2Mjn_{(M)}/s$ with probability P^n at least $1-(2M+1)/n^3$. Combining this with (84), we obtain $\sum_{i=1}^n i^{\alpha/d} p_{i,j}^u(x) \leq (2Mjn_{(M)}/s)^{\alpha/d}$, which together with (83) implies

$$\sum_{i=n_1}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) \leq k^{-1}(4M/n)^{\alpha/d} \sum_{j=1}^{k} (jn_{(M)}/s)^{\alpha/d}.$$

Since $g(t) = t^{\alpha/d}$ is increasing in $[0,1]$, we have $k^{-1}\sum_{j=1}^k (j/k)^{\alpha/d} \leq 2\int_0^1 x^{\alpha/d}\,dx = 2(1+\alpha/d)$. Consequently we obtain

$$\sum_{i=n_1}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) = k^{-1}\bigl(4Mn_{(M)}k/(ns)\bigr)^{\alpha/d} \sum_{j=1}^{k} (j/k)^{\alpha/d} \leq 2(1+\alpha/d)(4Mk/s)^{\alpha/d}.$$

Combining this with (82), we find that for all $x \in \mathbb{R}^d$, there holds

$$\sum_{i=1}^{n} \overline{V}_i^u(x) R_{(i)}^\alpha(x) \leq (2k/s)^{\alpha/d} + 2(1+\alpha/d)(4Mk/s)^{\alpha/d} \leq c_3(k/s)^{\alpha/d},$$

where the constant $c_3 := 2^{\alpha/d} + 2(1+\alpha/d)(4M)^{\alpha/d}$. Thus, we finish the proof. ■

The following lemma is needed in the proof of Proposition 12.

Lemma 32 *Let $\overline{V}_i^u(x)$ be defined by* (37) *and suppose $k \leq s$. Then for all $x \in \mathcal{X}$, we have*

$$1 - \sum_{i=1}^{n} \overline{V}_i^u(x) \leq \exp\bigl(-(s-k)^2/(2n)\bigr).$$

Proof [of Lemma 32] By (37), we have

$$\sum_{i=1}^{n} \overline{V}_i^u(x) = \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{k} p_{i,j}^u(x) = \frac{1}{k} \sum_{j=1}^{k} \sum_{i=1}^{n} p_{i,j}^u(x). \tag{85}$$

By (68), we have $\sum_{i=1}^n p_{i,j}^u(x) = \sum_{i=1}^n f_{\mathrm{GP}}(i;j,p(x))$ where $p(x) \in S_\nu$ with S_ν defined as in (70). Since $\sum_{i=1}^n p_i(x) = \sum_{m=1}^M n_{(m)} \cdot \frac{s}{Mn_{(M)}} = s$, Theorem 27 implies

$$\sum_{i=1}^n p_{i,j}^u(x) = \sum_{i=1}^n f_{\mathrm{GP}}(i;j,p(x)) \geq 1 - \exp\bigl(-(s-j)^2/(2n)\bigr) \geq 1 - \exp\bigl(-(s-k)^2/(2n)\bigr).$$

Combining this with (85), we obtain $\sum_{i=1}^n \overline{V}_i^u(x) \geq 1 - \exp\bigl(-(s-k)^2/(2n)\bigr)$, which yields the assertion. ■

Proof [of Proposition 12] By the definition of $\overline{\eta}^{B,u}$ and η^u, we have

$$\begin{aligned}\|\overline{\eta}^{B,u}(x) - \eta^u(x)\|_\infty &= \biggl\|\sum_{i=1}^n \overline{V}_i^u(x)\eta_m^u(X_{(i)}(x)) - \eta^u(x)\biggr\|_\infty \\ &\leq \biggl\|\sum_{i=1}^n \overline{V}_i^u(x)\bigl(\eta_m^u(X_{(i)}(x)) - \eta^u(x)\bigr)\biggr\|_\infty + \biggl\|\sum_{i=1}^n \overline{V}_i^u(x) - 1\biggr\|_\infty.\end{aligned}$$

Lemma 21 implies that for all $n \geq n_1$, with probability at least $1 - 2M/n^3$, there holds

$$\|\overline{\eta}^{B,u}(x) - \eta^u(x)\|_\infty \leq 4c_L \sup_{x\in\mathcal{X}}\biggl(\sum_{i=1}^n \overline{V}_i^u(x)\|X_{(i)}(x) - x\|^\alpha\biggr) + \biggl\|\sum_{i=1}^n \overline{V}_i^u(x) - 1\biggr\|_\infty.$$

Applying Lemma 31, we obtain

$$\|\overline{\eta}^{B,u}(x) - \eta^u(x)\|_\infty \leq c_3(k/s)^{\alpha/d} + \exp\bigl(-(s-k)^2/(2n)\bigr) \tag{86}$$

for all $n > n_3$ with probability P^n at least $1 - (4M+1)/n^3$. Consequently, if $n \geq N_5 := \max\{4M+1, n_3\}$, then (86) holds with probability P^n at least $1 - 1/n^2$. This completes the proof. ■

7.2.3 Proofs Related to Section 5.2.3

To prove Proposition 13, we need the following lemma, which bounds the maximum value of the bagged weights $\overline{V}_i^u(x)$ defined by (37).

Lemma 33 *Let $\overline{V}_i^u(x)$ be defined by* (37)*. Then for any $x \in \mathbb{R}^d$, there holds*

$$\max_{1\leq i\leq n} \overline{V}_i^u(x) \leq s/(kMn_{(1)}).$$

Proof [of Lemma 33] By (63) and (64), we have

$$\overline{V}_i^u(x) = \frac{1}{k}\sum_{j=1}^k p_{i,j}^u(x)$$

$$= \frac{1}{k} \sum_{j=1}^{k} \mathrm{P}_Z \Big(\sum_{\ell=1}^{i} Z^b(X_{(\ell)}(x), Y_{(\ell)}(x)) = j, Z^b(X_{(i)}(x), Y_{(i)}(x)) = 1 \Big| \{(X_i, Y_i)\}_{i=1}^n \Big)$$
$$\leq k^{-1} \mathrm{P}_Z \big(Z^b(X_{(i)}(x), Y_{(i)}(x)) = 1 \big| \{(X_i, Y_i)\}_{i=1}^n \big)$$
$$= k^{-1} a(X_{(i)}(x), Y_{(i)}(x)) \leq s/(Mkn_{(1)}),$$

which finishes the proof. ■

Proof [of Proposition 13] By the definition of $\widetilde{\eta}^{B,u}$ and $\overline{\eta}^{B,u}$, we have

$$\big| \widetilde{\eta}_m^{B,u} - \overline{\eta}_m^{B,u} \big| = \sum_{i=1}^{n} \overline{V}_i^u(x) \big(\mathbf{1}\{Y_{(i)}(x) = m\} - \eta_m^u(X_{(i)}(x)) \big).$$

For any fixed $x \in \mathcal{X}$, Lemmas 22 and 33 yield

$$(\mathrm{P}^u)_{Y|X}^n \big(\big| \widetilde{\eta}_m^{B,u}(x) - \overline{\eta}_m^{B,u}(x) \big| \geq \varepsilon \big| D_n \big) \leq 2 \exp\big(-\varepsilon^2 k M n_{(1)} / (2s) \big).$$

Setting $\varepsilon := \sqrt{2(2d+3) s \log n / (kMn_{(1)})}$, we get

$$(\mathrm{P}^u)_{Y|X}^n \big(\big| \widetilde{\eta}_m^{B,u}(x) - \overline{\eta}_m^{B,u}(x) \big| \geq \varepsilon \big| D_n \big) \leq 2 n^{-(2d+3)}. \tag{87}$$

Note that this inequality holds only for fixed x. In order to derive the uniform upper bound over $\mathcal{X}$, let $\mathcal{S} := \big\{ (\sigma_1, \ldots, \sigma_n) : \text{all permutations of } (1, \ldots, n) \text{ obtainable by moving } x \in \mathbb{R}^d \big\}$. Then we have

$$(\mathrm{P}^u)_{Y|X}^n \Big(\sup_{x \in \mathbb{R}^d} \big(| \widetilde{\eta}_m^{B,u}(x) - \overline{\eta}_m^{B,u}(x) | - \varepsilon \big) > 0 \Big| D_n \Big)$$
$$\leq (\mathrm{P}^u)_{Y|X}^n \Big(\bigcup_{(\sigma_1, \ldots, \sigma_n) \in \mathcal{S}} \Big| \sum_{i=1}^{n} \overline{V}_{i,\sigma}^u (\mathbf{1}\{Y_{\sigma_i} = m\} - \eta_m^u(X_{\sigma_i})) \Big| > \varepsilon \Big| D_n \Big)$$
$$\leq \sum_{(\sigma_1, \ldots, \sigma_n) \in \mathcal{S}} (\mathrm{P}^u)_{Y|X}^n \Big(\Big| \sum_{i=1}^{n} \overline{V}_{i,\sigma}^u (\mathbf{1}\{Y_{\sigma_i} = m\} - \eta_m^u(X_{\sigma_i})) \Big| > \varepsilon \Big| D_n \Big),$$

where $\overline{V}_{i,\sigma}^u(x) = k^{-1} \mathrm{P}_Z (\sum_{j=1}^{i} Z^b(X_{\sigma_j}(x), Y_{\sigma_j}(x)) \leq k | \{X_i, Y_i\}_{i=1}^n)$. For any $(\sigma_1, \ldots, \sigma_n) \in \mathcal{S}$, (87) implies

$$(\mathrm{P}^u)_{Y|X}^n \Big(\Big| \sum_{i=1}^{n} \overline{V}_{i,\sigma}^u (\mathbf{1}\{Y_{\sigma_i} = m\} - \eta_m^u(X_{\sigma_i})) \Big| > \varepsilon \Big| D_n \Big) \leq 2/n^{2d+3}.$$

Combining this with Lemma 20, we obtain

$$(\mathrm{P}^u)_{Y|X}^n \Big(\sup_{x \in \mathbb{R}^d} (| \widetilde{\eta}_m^{B,u}(x) - \overline{\eta}_m^{B,u}(x) | - \varepsilon) > 0 \Big| D_n \Big) \leq 2 (25/d)^d / n^3$$

for $n \geq 2d$. Then a union bound argument yields

$$(\mathrm{P}^u)_{Y|X}^n \big(\| \widetilde{\eta}^{B,u}(x) - \overline{\eta}^{B,u}(x) \|_\infty \leq \sqrt{2(2d+3) s \log n / (kMn_{(1)})} \big| D_n \big) \geq 1 - 2M (25/d)^d / n^3.$$

Consequently, by the law of total probability, we have

$$\mathrm{P}_Z^B \otimes \mathrm{P}^n\big(\|\widetilde{\eta}^{B,u}(x) - \overline{\eta}^{B,u}(x)\|_\infty \leq \sqrt{2(2d+3)s\log n/(kMn_{(1)})}\big) \geq 1 - 2M(25/d)^d/n^3.$$

Therefore, if $n \geq N_6 := \max\{\lceil 2M(25/d)^d \rceil\}$, there holds

$$\mathrm{P}_Z^B \otimes \mathrm{P}^n\big(\|\widetilde{\eta}^{B,u}(x) - \overline{\eta}^{B,u}(x)\|_\infty \leq \sqrt{2(2d+3)s\log n/(kMn_{(1)})}\big) \geq 1 - 1/n^2.$$

Thus we complete the proof of Proposition 13. ■

7.2.4 Proofs Related to Section 4.2

Proof [of Theorem 3] Choosing s, B, and k according to (17), (19), and (18), respectively, Propositions 11 , 12, 13 yield that if $n \geq \max\{N_4, N_5, N_6\} = \max\{N_5, N_6\}$, there holds

$$\begin{aligned}\|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty &\lesssim \sqrt{\log n/B} + (k/s)^{\alpha/d} + \exp\big(-(s-k)^2/(2n)\big) + \sqrt{s\log n/(kMn_{(1)})} \\ &\lesssim \big(\log(Mn_{(1)})/(Mn_{(1)})\big)^{\alpha/(2\alpha+d)}\end{aligned}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 3/n^2$. According to (46) in Lemma 21, we have $Mn_{(1)} \geq Mn\pi_1/2$. Note that $g(x) := \log(x)/x$ is decreasing on $[e, \infty)$. Consequently, if $n \geq \max\{N_5, N_6, 2M, \lceil 2e/(M\underline{\pi})\rceil\}$, there holds

$$\|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty \lesssim \big(\log(Mn\pi_1/2)/(Mn\pi_1/2)\big)^{\alpha/(2\alpha+d)} \lesssim (\log n/n)^{\alpha/(2\alpha+d)}.$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 4/n^2$. This together with Proposition 10 yields that for all $n \geq N_2^* := \max\{N_3, N_5, N_6, 2M, \lceil 2e/(M\underline{\pi})\rceil\}$, there holds

$$\begin{aligned}\|\widehat{\eta}^{B,u}(x) - \eta^w(x)\|_\infty &= \|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty + \|\eta^u(x) - \eta^w(x)\|_\infty \\ &\lesssim (\log n/n)^{\alpha/(2\alpha+d)} + \sqrt{\log n/n} \lesssim (\log n/n)^{\alpha/(2\alpha+d)}\end{aligned}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 5/n^2$. Lemma 15 yields that

$$\|\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) - \eta^w_{\widehat{\psi}^{B,u}(x)}(x)\|_\infty \lesssim (\log n/n)^{\alpha/(2\alpha+d)}$$

holds with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 5/n^2$, where $\eta^{w,*}_{L_{\mathrm{cl}},\mathrm{P}^w}(x) = \eta_{\psi^*_{L_{\mathrm{cl}},\mathrm{P}^w}(x)}(x)$, i.e, the Bayes classifier w.r.t. the classification loss L_{cl} and the balanced distribution P^w. Consequently, Lemma 16 implies that

$$\mathcal{R}_{L_{\mathrm{cl}},\mathrm{P}^w}(\widehat{\psi}^{B,u}) - \mathcal{R}^*_{L_{\mathrm{cl}},\mathrm{P}^w} \lesssim (\log n/n)^{\alpha(\beta+1)/(2\alpha+d)}$$

holds with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^{\mathrm{n}}$ at least $1 - 5/n^2$. By (24) and Theorem 6, we have $\mathfrak{R}_{\mathrm{AM}}(\widehat{\psi}^{B,u}) \lesssim (\log n/n)^{\alpha(\beta+1)/(2\alpha+d)}$, which finishes the proof. ■

Proof [of Corollary 4] Taking $k = 1$ in Proposition 12, we get

$$\begin{aligned}\|\overline{\eta}^{b,u}(x) - \eta^u(x)\|_\infty &\le \sum_{i=1}^n \overline{V}_i^u(x) R_{(i)}^\alpha(x) + \exp\bigl(-(s-1)^2/(2n)\bigr)\\ &\le (2n_1/n)^{\alpha/d} + c_3 s^{-\alpha/d} + \exp\bigl(-(s-1)^2/(2n)\bigr)\\ &\le c_3'(\log s/s)^{\alpha/d} + \exp\bigl(-(s-1)^2/(2n)\bigr),\end{aligned}$$

where the constant $c_3' := (12d+32)^{\alpha/d} + c_3$. This together with Propositions 11 and 13 yields that if $n \ge \max\{N_4, N_5, N_6\} = \max\{N_5, N_6\}$, there holds

$$\|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty \lesssim \sqrt{\log n/B} + (\log s/s)^{\alpha/d} + \exp\bigl(-(s-1)^2/(2n)\bigr) + \sqrt{s\log n/(Mn_{(1)})}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 3/n^2$.

If $d > 2\alpha$, with $s = (Mn_{(1)})^{\frac{d}{2\alpha+d}} (\log(Mn_{(1)}))^{\frac{2\alpha-d}{2\alpha+d}}$, $B = (Mn_{(1)})^{\frac{2\alpha}{2\alpha+d}} (\log(Mn_{(1)}))^{\frac{d-2\alpha}{2\alpha+d}}$ we get

$$\|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty \lesssim (\log^2(Mn_{(1)})/Mn_{(1)})^{\alpha/(2\alpha+d)}.$$

According to (46) in Lemma 21, we have $Mn_{(1)} \ge Mn\pi_1/2$. Note that $g(x) := \log^2(x)/x$ is decreasing on $[e^2, \infty)$. Consequently, if $n \ge \max\{N_5, N_6, 2M, \lceil 2e^2/(M\underline{\pi})\rceil\}$, there holds

$$\|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty \lesssim \bigl(\log^2(Mn\pi_1/2)/(Mn\pi_1/2)\bigr)^{\alpha/(2\alpha+d)} \lesssim (\log n/n)^{\alpha/(2\alpha+d)}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 4/n^2$. This together with Proposition 10 yields that for all $n \ge N_3^* := \max\{N_3, N_5, N_6, 2M, \lceil 2e^3/(M\underline{\pi})\rceil\}$, there holds

$$\begin{aligned}\|\widehat{\eta}^{B,u}(x) - \eta^w(x)\|_\infty &= \|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty + \|\eta^u(x) - \eta^w(x)\|_\infty\\ &\lesssim (\log n/n)^{\alpha/(2\alpha+d)} + \sqrt{\log n/n} \lesssim (\log n/n)^{\alpha/(2\alpha+d)}\end{aligned}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 5/n^2$.

Otherwise if $d \le 2\alpha$, with $s = (Mn_{(1)}\log(Mn_{(1)}))^{1/2}$ and $B = (Mn_{(1)}/\log(Mn_{(1)}))^{1/2}$, by similar arguments as above, for all $n \ge N_3^*$, there holds

$$\begin{aligned}\|\widehat{\eta}^{B,u}(x) - \eta^w(x)\|_\infty &= \|\widehat{\eta}^{B,u}(x) - \eta^u(x)\|_\infty + \|\eta^u(x) - \eta^w(x)\|_\infty\\ &\lesssim \max\bigl\{(\log n/n)^{\alpha/(2d)}, (\log^3 n/n)^{1/4}\bigr\}\end{aligned}$$

with probability $\mathrm{P}_Z^B \otimes \mathrm{P}^n$ at least $1 - 5/n^2$. By exploiting similar arguments as that in the proof of Theorem 3, we obtain the assertion. ■

Acknowledgments

The authors would like to thank the action editor and reviewers for their constructive comments, which lead to a significant improvement of this work. Hanfang Yang was supported by the National Key R&D Program of China (Grant No. 2018YFC2000302). Zhouchen Lin was supported by the NSF China (No. 61731018), NSFC Tianyuan Fund for Mathematics (No. 12026606), Project 2020BD006 supported by PKU-Baidu Fund, and Qualcomm. Resources supporting this work were provided by High-performance Computing Platform of Renmin University of China. The corresponding author is Hanfang Yang.

References

Rishabh Ahuja, Arun Solanki, and Anand Nayyar. Movie recommender system using k-means clustering and k-nearest neighbor. In *Proceedings of the 9th International Conference on Cloud Computing, Data Science & Engineering*, pages 263–268, 2019.

Cuneyt G. Akcora, Yitao Li, Yulia R. Gel, and Murat Kantarcioglu. Bitcoinheist: Topological data analysis for ransomware prediction on the bitcoin blockchain. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4439–4445, 2020.

Mohanad Al-Ghobari, Amgad Muneer, and Suliman Mohamed Fati. Location-aware personalized traveler recommender system using collaborative filtering k-NN. *Computers, Materials & Continua*, 69(2):1553–1570, 2021.

Samaher Al_Janabi and Fatma Razaq. Intelligent big data analysis to design smart predictor for customer churn in telecommunication industry. In Yousef Farhaoui and Laila Moussaid, editors, *Proceedings of the International Conference on Big Data and Smart Digital Environment*, pages 246–272, 2019.

Md Adnan Arefeen, Sumaiya Tabassum Nimi, and M Sohel Rahman. Neural network-based undersampling techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020.

Micheal Olaolu Arowolo, Marion Adebiyi, Ayodele Adebiyi, and Olatunji Okesola. PCA model for RNA-Seq malaria vector data classification using k-NN and decision tree algorithm. In *International Conference in Mathematics, Computer Engineering and Computer Science*, pages 1–8, 2020.

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

Sarah M. Ayyad, Ahmed I. Saleh, and Labib M. Labib. Gene expression cancer classification using modified k-nearest neighbors technique. *Biosystems*, 176:41–51, 2019.

S. Babu and N.R. Ananthanarayanan. Enhanced prediction model for customer churn in telecommunication using EMOTE. In *Proceedings of the International Conference on Intelligent Computing and Applications*, pages 465–475, 2018.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Gustavo EAPA Batista, Ana LC Bazzan, Maria Carolina Monard, et al. Balancing training data for automated annotation of keywords: a case study. In *WOB*, pages 10–18, 2003.

Nabil Belacel, Guangze Wei, and Yassine Bouslimani. The k-closest resemblance classifier for Amazon products recommender system. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 873–880, 2020.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

Sergei N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*, volume 246. Springer, 2015.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *The Journal of Machine Learning Research*, 11(2):687–712, 2010.

Jonas Biteus and Tony Lindgren. Planning flexible maintenance for heavy trucks using machine learning models, constraint programming, and route optimization. *SAE International Journal of Materials and Manufacturing*, 10(3):306–315, 2017.

Leandro A. Bugnon, Cristian Yones, Diego H. Milone, and Georgina Stegmayer. Deep neural architectures for highly imbalanced data in bioinformatics. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2857–2867, 2019.

Luis M. Candanedo and Véronique Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016.

Philip K. Chan and Salvatore J. Stolfo. Learning with non-uniform class and cost distributions: Effects and a distributed multi-classifier approach. In *Workshop Notes KDD-98 Workshop on Distributed Data Mining*, pages 1–9, 1998.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 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.

Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

S.A. Danziger, S.J. Swamidass, Jue Zeng, L.R. Dearth, Qiang Lu, J.H. Chen, J. Cheng, V.P. Hoang, H. Saigo, R. Luo, P. Baldi, R.K. Brachmann, and R.H. Lathrop. Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):114–125, 2006.

Morris H. DeGroot. *Probability and Statistics*. Pearson, 2012.

Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, and Shichao Zhang. Efficient k-NN classification algorithm for big data. *Neurocomputing*, 195:143–148, 2016.

Zonglin Di, Qi Kang, Daogang Peng, and Mengchu Zhou. Density peak-based pre-clustering support vector machine for multi-class imbalanced classification. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 27–32. IEEE, 2019.

Maik Döring, László Györfi, and Harro Walk. Rate of convergence of k-nearest-neighbor classification rule. *The Journal of Machine Learning Research*, 18(1):8485–8500, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Harshit Dubey and Vikram Pudi. Class based weighted k-nearest neighbor over imbalance dataset. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 305–316, 2013.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. 2nd edition, 2001.

Richard M. Dudley. Balls in $\mathbb{R}^k$ do not cut all subsets of $k+2$ points. *Advances in Mathematics*, 31 (3):306–308, 1979.

Peter Flach. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9808–9814, 2019.

Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226, 1977.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2021.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

Haixiang Guo, Yijing Li, Yanan Li, Xiao Liu, and Jinling Li. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. *Engineering Applications of Artificial Intelligence*, 49:176–193, 2016.

Haixiang Guo, Yijing Li, Jennifer Shang, Mingyun Gu, Yuanyue Huang, and Bing Gong. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

Peter Hall and Richard J. Samworth. Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society: Series B*, 67(3):363–379, 2005.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

Hanyuan Hang and Ingo Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *The Annals of Statistics*, 45(2):708–743, 2017.

Hanyuan Hang, Yunlong Feng, Ingo Steinwart, and Johan A.K. Suykens. Learning theory estimates with observations from general stationary stochastic processes. *Neural computation*, 28(12):2853–2889, 2016.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Haibo He and Yunqian Ma. Imbalanced learning: foundations, algorithms, and applications. 2013.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56, pages 111–117. Citeseer, 2000.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Justin M. Johnson and Taghi M. Khoshgoftaar. The effects of data sampling with deep learning and highly imbalanced big data. *Information Systems Frontiers*, 22(5):1113–1131, 2020.

François Kawala, Ahlame Douzal-Chouakria, Eric Gaussier, and Eustache Dimert. Prédictions d'activité dans les réseaux sociaux en ligne. In *4ième conférence sur les modèles et l'analyse des réseaux: Approches mathématiques et informatiques*, page 16, 2013.

Justin Khim, Ziyu Xu, and Shashank Singh. Multiclass classification via class-weighted nearest neighbors. *arXiv preprint arXiv:2004.04715*, 2020.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York, 2008.

Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning*, pages 1113–1120, 2011.

Michał Koziarski, Bartosz Krawczyk, and Michał Woźniak. Radial-based oversampling for noisy imbalanced data classification. *Neurocomputing*, 343:19–33, 2019.

Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, page 179. Citeseer, 1997.

Yulia Ery Kurniawati, Adhistya Erna Permanasari, and Silmi Fauziati. Adaptive synthetic-nominal and adaptive synthetic-k-NN for multiclass imbalance learning on laboratory test data. In *Proceedings of the 4th International Conference on Science and Technology*, pages 1–6, 2018.

Joffrey L. Leevy, Taghi M. Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018.

Yuxuan Li and Xiuzhen Zhang. Improving k-nearest neighbor with exemplar generalization for imbalanced classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 321–332, 2011.

Zhenchuan Li, Mian Huang, Guanjun Liu, and Changjun Jiang. A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. *Expert Systems with Applications*, 175:114750, 2021.

Shenglan Liu, Ping Zhu, and Sujuan Qin. An improved weighted k-NN algorithm for imbalanced data classification. In *Proceedings of the 4th International Conference on Computer and Communications*, pages 1814–1819, 2018.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

N. Malini and M. Pushpa. Analysis on credit card fraud identification techniques based on k-NN and outlier detection. In *Proceedings of the Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics*, pages 255–258, 2017.

Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, pages 1–7. ICML, 2003.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

Jiri Matousek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2013.

Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28(1):92–122, 2014.

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 603–611, 2013.

Aiman Moldagulova and Rosnafisah Bte Sulaiman. Using k-NN algorithm for classification of textual documents. In *Proceedings of the 8th International Conference on Information Technology*, pages 665–671, 2017.

Ehsan Montahaei, Mahsa Ghorbani, Mahdieh Soleymani Baghshah, and Hamid R Rabiee. Adversarial classifier for imbalanced problems. *arXiv preprint arXiv:1811.08812*, 2018.

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5713–5725, 2018.

Harikrishna Narasimhan, Harish Ramaswamy, Aadirupa Saha, and Shivani Agarwal. Consistent multiclass algorithms for complex performance measures. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2398–2407, 2015.

Mar Mar Nwe and Khin Thidar Lynn. k-NN-based overlapping samples filter approach for classification of imbalanced data. In *Proceedings of the International Conference on Software Engineering Research, Management and Applications*, pages 55–73, 2019.

Juri Opitz and Sebastian Burst. Macro F1 and macro F1. *arXiv preprint arXiv:1911.03347*, 2019.

Sunkuru Gopal Krishna Patro, Brojo Kishore Mishra, Sanjaya Kumar Panda, Raghvendra Kumar, Hoang Viet Long, David Taniar, and Ishaani Priyadarshini. A hybrid action-related k-nearest neighbour approach for recommendation systems. *IEEE Access*, 8:90978–90991, 2020.

Derek A. Pisner and David M. Schnyer. Support vector machine. In *Machine Learning*, pages 101–121. 2020.

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. Optimizing f-measures by cost-sensitive classification. *Advances in neural information processing systems*, 27, 2014.

Bhavani Raskutti and Adam Kowalczyk. Extreme re-balancing for SVMs: A case study. *ACM SIGKDD Explorations Newsletter*, 6(1):60–69, 2004.

Mojdeh Rastgoo, Guillaume Lemaitre, Joan Massich, Olivier Morel, Franck Marzani, Rafael Garcia, and Fabrice Meriaudeau. Tackling the problem of data imbalancing for melanoma classification. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 32–39, 2016.

Jaime Salvador-Meneses, Zoila Ruiz-Chavez, and Jose Garcia-Rodriguez. Compressed k-NN: k-nearest neighbors with data compression. *Entropy*, 21(3):234, 2019.

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

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

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2009.

Ahmad Abu Shanab and Taghi M. Khoshgoftaar. How to optimally combine univariate and multivariate feature selection with data sampling for classifying noisy, high dimensional and class imbalanced DNA microarray data. In *Reuse in Intelligent Systems*, pages 33–61. 2020.

Roberto L. Shinmoto Torres, Damith C. Ranasinghe, Qinfeng Shi, and Alanson P. Sample. Sensor enabled wearable RFID technology for mitigating the risk of falls near beds. In *Proceedings of the International Conference on RFID*, pages 191–198, 2013.

Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine learning*, 95(2):225–256, 2014.

Ishan Sohony, Rameshwar Pratap, and Ullas Nambiar. Ensemble learning for credit card fraud detection. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 289–294, 2018.

Akila Somasundaram and Srinivasulu Reddy. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. *Neural Computing and Applications*, 31(1): 3–14, 2019.

Murray Ralph Spiegel. *Theory and Problems of Probability and Statistics*. New York, NY: McGraw-Hill, 2nd edition, 1992.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009.

Antonio J. Tallón-Ballesteros and José C Riquelme. Data mining methods applied to a digital forensics task for supervised machine learning. In *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications*, pages 413–428. 2014.

Songbo Tan. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, 28(4):667–671, 2005.

Ivan Tomek. Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772, 1976.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

Chen Wang, Chengyuan Deng, and Suzhen Wang. Imbalance-xgboost: Leveraging weighted and focal losses for binary label-imbalanced classification with xgboost. *Pattern Recognition Letters*, 136:190–197, 2020.

Le Wang, Meng Han, Xiaojuan Li, Ni Zhang, and Haodong Cheng. Review of classification methods on unbalanced data sets. *IEEE Access*, 9:64606–64628, 2021.

Lei Wang, Latifur Khan, and Bhavani Thuraisingham. An effective evidence theory based k-nearest neighbor classification. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 797–801, 2008.

Gary M. Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004.

Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.

Jianxin Wu, S. Charles Brubaker, Matthew D. Mullin, and James M. Rehg. Fast asymmetric learning for cascade face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 (3):369–382, 2008a.

Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, and S. Yu Philip. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008b.

Lirong Xue and Samory Kpotufe. Achieving the time of 1-NN, but the accuracy of k-NN. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1628–1636, 2018.

Yuguang Yan, Mingkui Tan, Yanwu Xu, Jiezhang Cao, Michael Ng, Huaqing Min, and Qingyao Wu. Oversampling for imbalanced data via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5605–5612, 2019.

Qiang Yang and Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(4):597–604, 2006.

Yingze Yang, Pengcheng Xiao, Yijun Cheng, Weirong Liu, and Zhiwu Huang. Ensemble strategy for hard classifying samples in class-imbalanced data set. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 170–175. IEEE, 2018.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480, 2009.

Bo-Wen Yuan, Xing-Gang Luo, Zhong-Liang Zhang, Yang Yu, Hong-Wei Huo, Johannes Tretter, and Xiao-Dong Zou. A novel density-based adaptive k-nearest neighbor method for dealing with overlapping problem in imbalanced datasets. *Neural Computing and Applications*, 33(9):4457–4481, 2021.

Chong Zhang, Kay Chen Tan, Haizhou Li, and Geok Soon Hong. A cost-sensitive deep belief network for imbalanced classification. *IEEE transactions on neural networks and learning systems*, 30(1): 109–122, 2018.

Jianping Zhang and Inderjeet Mani. k-NN approach to unbalanced data distributions: A case study involving information extraction. In *Proceeding of International Conference on Machine Learning, Workshop on Learning from Imbalanced Data Sets*, 2003.

Shichao Zhang. Cost-sensitive k-NN classification. *Neurocomputing*, 391:234–242, 2020.

Xiuzhen Zhang, Yuxuan Li, Ramamohanarao Kotagiri, Lifang Wu, Zahir Tari, and Mohamed Cheriet. KRNN: k rare-class nearest neighbour classification. *Pattern Recognition*, 62:33–44, 2017.

Bo Zhao, Xianmin Zhang, Hai Li, and Zhuobo Yang. Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions. *Knowledge-Based Systems*, 199:105971, 2020.

Puning Zhao and Lifeng Lai. Minimax rate optimal adaptive nearest neighbor classification and regression. *IEEE Transactions on Information Theory*, 67(5):3155–3182, 2021.