Theoretical and experimental study of SMOTE: limitations and comparisons of rebalancing strategies

Abdoulaye Sakho, Erwan Scornet, Emmanuel Malherbe

Theoretical and experimental study of SMOTE: limitations and comparisons of rebalancing strategies

Abdoulaye SAKHO *[1,2], Erwan SCORNET[2], and Emmanuel MALHERBE[1]

[1]Artefact Research Center, Paris, France
[2]LPSM Sorbonne Université, Paris, France

Abstract

Synthetic Minority Oversampling Technique (SMOTE) is a common rebalancing strategy for handling imbalanced data sets. Asymptotically, we prove that SMOTE (with default parameter) regenerates the original distribution by simply copying the original minority samples. We also prove that SMOTE density vanishes near the boundary of the support of the minority distribution, therefore justifying the common BorderLine SMOTE strategy. Then we introduce two new SMOTE-related strategies, and compare them with state-of-the-art rebalancing procedures. We show that rebalancing strategies are only required when the data set is highly imbalanced. For such data sets, SMOTE, our proposals, or undersampling procedures are the best strategies.

1 Introduction

Imbalanced data sets are a typical problem encountered practically in several applications (He and Garcia, 2009), such as fraud detection (Hassan and Abraham, 2016), medical diagnosis (Khalilia et al., 2011) and even churn detection (Nguyen and Duong, 2021). In presence of imbalanced data sets, most machine learning algorithms have a tendency to predict the majority class, therefore leading to biased predictions. Several strategies have been developed in order to handle this issue, as explained by Krawczyk (2016) and Ramyachitra and Manikandan (2014). All of these strategies can be split into two categories: the model-level approaches and the data-level approaches.

Model-level approaches deal with this problem by acting directly on machine learning algorithms. Such approaches contain the class weighting of the samples: higher weights are allocated to the minority samples. Zhu et al. (2018) introduce a Random Forest based algorithm that have different weights allocated to the classes for each tree of the forest. However, model-level approaches are not agnostic to the model, which can be restrictive. In this paper, we focus on data-level approaches.

Data-level approaches can be split into two groups. All the sampling strategies that do not generate new observations in the initial set belong to this group. Mani and Zhang (2003) explain that the most used under sampling strategies are Random Under Sampling and Nearmiss under sampling. The Random Under Sampling technique produce the desired balance between classes by dropping majoring samples without any notion of order between these samples. The Nearmiss1 strategies (Mani and Zhang, 2003) include distinction between minority samples by ranking them using their mean distance to their nearest neighbor of the minority class. Then the ranking is used to drop the samples (from the bottom) until a given balancing ratio is reached. The main default of these under sampling strategies is the fact that they remove samples from the data set, which results in a loss of information. The second group of data-level approaches consists in all strategies that generate new synthetic samples in the minority class, therefore called synthetic approaches. The most famous strategy in this group is probably *Synthetic Minority Oversampling Technique* (SMOTE, see Chawla et al., 2002). In SMOTE, new minority samples are generated via linear interpolation between an original sample and one of its

*abdoulaye.sakho@artefact.com

nearest neighbor, with both of them belonging to the minority class. Other approaches are based on Generative Adversial Networks (GAN Islam and Zhang, 2020), which are computationally expensive and mostly designed for specific data structures such as images. The main difficulty of these strategies is to synthesize relevant new samples, which should not be outliers or simple copies of original points.

Several papers study some specificities of the sampling strategies for imbalanced data sets and the impact of hyperparameter tuning. For example, Kamalov et al. (2022) study the optimal sampling ratios for imbalanced data set when using synthetic approaches. Aguiar et al. (2023) realize a survey on imbalance data sets in the context of online learning and propose a standardized framework in order to compare rebalancing strategies in this context. Furthermore, Wongvorachan et al. (2023) aim at comparing the synthetic approaches (Random Over Sampling, Random Under Sampling and SMOTE) in purpose of application on educational data.

Contributions We place ourselves in the framework of imbalanced classification with continuous input variables, since synthetic procedures such as SMOTE are originally designed to handle continuous variables. We first show that the points generated via SMOTE procedure with default parameters are asymptotically distributed as the minority class, when the number of minority samples increases. We also prove that, without tuning the hyperparameter K (usually set to 5), SMOTE asymptotically copies the original minority samples, therefore lacking the intrinsic variability required in any synthetic generative procedure. This highlights the importance of hyperparameter tuning in SMOTE, when the number of samples in the minority class is large enough. As a product of our analysis, we establish that SMOTE density vanishes near the boundary of the support of the minority distribution, therefore justifying the introduction of SMOTE variants such as BorderLine SMOTE strategy. Our theoretical analysis naturally leads us to introduce two SMOTE alternatives (CV-SMOTE and Multivariate Gaussian SMOTE). We compare the impact on predictive performances of several rebalancing strategies (including our proposals) on simulated and real-world data sets. We show that rebalancing strategies are required only for strongly imbalanced data sets. When such approaches are necessary, SMOTE, our proposals or undersampling strategies appear to be the best procedures.

2 Related works

In the synthetic approaches group, SMOTE is the central algorithm from which most of the others algorithms derive. Indeed, except the Neural-Networks based algorithms, most of the others strategies always use a variant of the linear interpolation introduced inside SMOTE procedure. Several variants try to focus on the generation of the synthetic samples near the boundary of the minority class support. The most common one is *ADASYN* (He et al., 2008) whose main idea is to produce more synthetic samples via linear interpolation between samples from the minority class which are mostly surrounded by majority class samples. Borderline SMOTE approaches (Han et al., 2005) aim at generating new synthetics samples on the frontier of both classes. Another SMOTE variant focusing on borders is SVM-SMOTE (Nguyen et al., 2011) whose the idea is to begin by applying a Support Vector Machine classifier to the imbalanced data. Then the linear interpolation is done on the support vector from the minority class.

Imb-learn (see Lemaître et al., 2017) is an open source package containing python implementations of : SMOTE, ADASYN, Borderline SMOTE and SVM-SMOTE.

Others variants of SMOTE does not focus on the boundary of the minority class support, but more generally on the way of generating the new samples. For example, Pan et al. (2020) introduce a SMOTE variant that select 3 points and interpolate inside the triangle formed by these 3 minority samples. Two-steps procedures such as *DBSMOTE* (Bunkhumpornpat et al., 2012) generates synthetic samples based on a preliminary procedure (DBSCAN method) are also introduced in the community. Chawla et al. (2003) introduce SMOTEBoost, a strategy that apply SMOTE before fitting a new weak classifier inside a boosting procedure.

Several works study theoretically the rebalancing strategies. The class weighting method is studied theoretically by King and Zeng (2001). King and Zeng (2001) study the Random Under Sampling strategy effect on a logistic regression classifier. To the best of our knowledge, there are only few theoretical works dissecting the intrinsic machinery in SMOTE algorithm. For example, Elreedy et al. (2023) establish the distribution of SMOTE samples based on the distribution of the original minority samples. Before that, Elreedy and Atiya (2019) derive the expectation and covariance matrix of the data generated by SMOTE. Elreedy and Atiya (2019) also highlights the effects of the input dimension, the hyperparameter K and the number of minority samples on SMOTE procedure. Results on simulated and real-world data sets show that the predictive performance of the classifier applied after SMOTE increases with the number of minority samples. Increasing the number of input

Algorithm 1 SMOTE iteration.

Input: Minority class samples $X_1, \dots, X_n$, number K of nearest-neighbors
Select uniformly X_c (called **central point**) among $\{X_1, \dots, X_n\}$.
Denote by $I = X_{(1)}(X_c,) \dots, X_{(K)}(X_c)$, the K nearest-neighbor of X_c (with respect to the L_2 norm).
Select $X_k \in I$ uniformly.
$w \leftarrow \mathcal{U}([0,1])$
$Z \leftarrow X_c + w(X_k - X_c)$
Return Z

variables decreases SMOTE ability of regenerating the minority distribution, in terms of Total Variance Difference, Kullback–Leibler divergence and Frobenius norm. A guideline is also given : choosing $K \in \{5, 6, \dots, 10\}$ is a good trade-off in order to be free of the high errors of large K values and limit the correlation phenomenon of small K values.

3 A study of SMOTE

Notations We denote by $\mathcal{U}([a,b])$ the uniform distribution over $[a,b]$. We denote by $\mathcal{N}(\mu, \Sigma)$ the multivariate normal distribution centered on μ and of covariance matrix Σ. For any set A, we denote by $Vol(A)$, the Lebesgue measure of A. For any $z \in \mathbb{R}^d$ and $r > 0$, we let $B(z,r)$ be the ball centered at z of radius r. We let $c_d = Vol(B(0,1))$ the volume of the unit ball in $\mathbb{R}^d$. For any $p, q \in \mathbb{N}$, and any $z \in [0,1]$, we denote by $\mathcal{B}(p,q;z) = \int_{t=0}^{z} t^{p-1}(1-t)^{q-1}\mathrm{d}t$ the incomplete beta function.

3.1 SMOTE algorithm

We assume to be given a training sample composed of N independent and identically distributed pairs (X_i, Y_i), where $X_i \in \mathbb{R}^d$ and $Y \in \{0,1\}$. We consider an imbalanced problem, in which the class $Y = 1$ is under-represented, compared to the class $Y = 0$, and thus called the minority class. We assume that we have n minority samples in our training set. In this paper, we place our theoretical and experimental studies in the context of SMOTE applied to data sets containing only *continuous* explanatory variables.
In this section, we study the SMOTE procedure, which generates synthetic data through linear interpolations between two pairs of original samples of the minority class. SMOTE algorithm has a single hyperparameter, K which stands for the number of nearest neighbors considered when interpolating. A single SMOTE iteration is detailed in Algorithm 1. In a classic machine learning pipeline, SMOTE procedure is repeated in order to obtain a prespecified ration between the both classes before training a classifier.

3.2 Theoretical results on SMOTE

SMOTE has been shown to exhibit good performances when combined to standard classification algorithms (see for instance Mohammed et al., 2020). However, there exist only few works that aim at understanding theoretically SMOTE behavior. In this section, we assume that $X_1, \dots, X_n$ are i.i.d samples from the minority class (that is, $Y_i = 1$ for all $i \in [n]$), with a common density f_X with bounded support, denoted by $\mathcal{X}$.

Lemma 1 (Convexity)**.** *Given f_X the distribution density of the minority class, with support $\mathcal{X}$, for all K, n, the associated SMOTE density $f_{Z_{K,n}}$ satisfies*

$$Supp(f_{Z_{K,n}}) \subseteq Conv(\mathcal{X}). \tag{1}$$

By construction, synthetic observations generated by SMOTE cannot fall outside the convex hull of $\mathcal{X}$. Equation (1) is not an equality, as SMOTE samples are the convex combination of only two original samples. For example, in dimension two, if $\mathcal{X}$ is concentrated near the vertices of square, then SMOTE samples are distributed near the square edges, whereas $\mathrm{Conv}(\mathcal{X})$ is the whole square.

SMOTE distribution has been derived in Elreedy et al. (2023). We build on these works and provide, in the following theorem, a different expression for the density of the data generated by SMOTE, denoted by $f_{Z_{K,n}}$. When no confusion is possible, we denote $f_{Z_{K,n}}$ simply by f_Z.

Theorem 3.1. *Assume that X_c is the central point chosen in a SMOTE iteration. Then, for all $x_c \in \mathcal{X}$, the random variable Z generated by SMOTE has a conditional density $f_{Z_{K,n}}(.|X_c = x_c)$ which satisfies*

$$f_{Z_{K,n}}(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mathcal{B}\left(n-K-1, K; 1-\beta_{x_c,z,w}\right)\mathrm{d}w, \tag{2}$$

where $\beta_{x_c,z,w} = \mu_X\left(B\left(x_c, ||z-x_c||/w\right)\right)$ and μ_X is the probability measure associated to f_X. Consequently, the density $f_{Z_{K,n}}$ of the data generated by SMOTE is

$$\begin{aligned} f_{Z_{K,n}}(z) = (n-K-1)\binom{n-1}{K}\int\limits_{x_c\in\mathcal{X}}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \\ \times \mathcal{B}\left(n-K-1, K; 1-\beta_{x_c,z,w}\right) f_X(x_c)\mathrm{d}w\mathrm{d}x_c. \end{aligned} \tag{3}$$

Using the following substitution $w = \|z-x_c\|/r$, we have,

$$\begin{aligned} f_{Z_{K,n}}(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_{r=\|z-x_c\|}^{\infty} f_X\left(x_c + \frac{(z-x_c)r}{\|z-x_c\|}\right) \\ \times \frac{r^{d-2}\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, r\right)\right)\right)}{\|z-x_c\|^{d-1}}\mathrm{d}r. \end{aligned} \tag{4}$$

Theorem 3.1 provides the expression of the density of SMOTE synthetic data unconditionally, and conditional on the central point used to generate the new observation. The expressions established in Theorem 3.1 are very similar to Theorem 1 and Lemma 1 in Elreedy et al. (2023). Although our proof shares the same structure as that of Elreedy et al. (2023), our starting point is different, as we consider random variables instead of geometrical arguments. The proof of Theorem 3.1 can be found in Section 8.2.

SMOTE algorithm has only one hyperparameter K, which is the number of nearest neighbors taken into account for building the linear interpolation. By default, this parameter is set to 5. The following theorem describes the behavior of SMOTE distribution asymptotically, as $K/n \to 0$.

Theorem 3.2. *For all Borel sets $B \subset \mathbb{R}^d$, if $K/n \to 0$, as n tends to infinity, we have*

$$\lim_{n\to\infty} \mathbb{P}[Z_{K,n} \in B] = \mathbb{P}[X \in B]. \tag{5}$$

The proof of Theorem 3.2 can be found in 8.3. Theorem 3.2 suggests that any choice of K such that K/n tends to zero is asymptotically correct, thus corroborating the usual practice of choosing $K = 5$. A close inspection of Theorem 3.1 allows us to derive more precise bounds about the behavior of SMOTE, as established in Theorem 3.3.

Theorem 3.3. *Assume that there exists $R > 0$, such that $\mathcal{X} \subset B(0, R)$. Suppose that there exists C_2 such that, for all $x \in \mathbb{R}^d$, $f_X(x) \leq C_2 \mathbb{1}_{x\in\mathcal{X}}$. Then, for all $n \geq K \geq 1$, for all $x_c \in \mathcal{X}$ and for all $\alpha > 0$, we have*

$$\mathbb{P}(|Z_{K,n} - X_c| \geq \alpha | X_c = x_c) \leq \varepsilon(n, \alpha, K, x_c), \tag{6}$$

where

$$\varepsilon(n, K, x_c, \alpha) = c_d R^d \eta(\alpha, R)\exp\left[n\left(3\sqrt{\frac{e}{1-\beta_{x_c,\alpha}}}\sqrt{\frac{K}{n}} + \ln\left(1-\beta_{x_c,\alpha}\right)\right)\right], \tag{7}$$

with $\beta_{x_c,\alpha} = \mu_X\left(B\left(x_c, \alpha\right)\right) > 0$ and

$$\eta(\alpha, R) = \begin{cases} C_2 \ln\left(\frac{2R}{\alpha}\right) & \text{if } d = 1, \\ \frac{C_2}{d-1}\left(\left(\frac{2R}{\alpha}\right)^{d-1} - 1\right) & \text{if } d > 1, \\ 0 & \text{if } \alpha > 2R. \end{cases}$$

Consequently, if $\lim_{n\to\infty} K/n = 0$, we have, for all $x_c \in \mathcal{X}$, $Z_{K,n}|X_c = x_c \to x_c$ in probability.

According to Theorem 3.3, SMOTE regenerates the original distribution at the cost of copying the original minority samples. Indeed, the probability of a SMOTE sample to be at a distance superior to $\alpha > 0$ from his central point tends to 0 when $K/n \to 0$. This means that when $K/n \to 0$, SMOTE samples are closer to their associated central point, leading the procedure to generate new samples by copying the original ones. The proof of Theorem 3.3 can be found in 8.4.

Corollary 3.3.1. *Assume that there exists $R > 0$, such that $\mathcal{X} \subset B(0, R)$. Suppose that there exist C_1, C_2 such that, for all $x \in \mathbb{R}^d$, $C_1 \mathbb{1}_{x \in \mathcal{X}} \leq f_X(x) \leq C_2 \mathbb{1}_{x \in \mathcal{X}}$. Let $x_c \in \overset{o}{\mathcal{X}}$. Let $\gamma \in \mathbb{R}$. For all $n/K \geq R$, we have*

For $d = 1$ and γ such that

$$\max\left(0, \frac{\ln(4C_1)}{\ln\left(\frac{n}{K}\right)}\right) < \gamma < \frac{1}{2} + \min\left(0, \frac{\ln\left(3\sqrt{2e}/C_1\right)}{\ln(K/n)}\right), \tag{8}$$

we have

$$\mathbb{P}\left(|Z_{K,n} - X_c| \geq (K/n)^\gamma | X_c = x_c\right) \leq \frac{4C_2 e^{-1}(\gamma + 1)}{3K\sqrt{2e}} \exp\left[-3\sqrt{\frac{enK}{2}}\right] 2R.$$

For $d > 1$ and all γ such that,

$$\max\left(0, \frac{\ln(2C_1 c_d)}{d\ln\left(\frac{n}{K}\right)}\right) < \gamma < \frac{1}{2d} + \min\left(0, \frac{\ln\left(\frac{6\sqrt{2e}}{C_1 c_d}\right)}{d\ln(K/n)}\right),$$

we have

$$\mathbb{P}\left(|Z_{K,n} - X_c| \geq (K/n)^\gamma | X_c = x_c\right) \leq \frac{2C_2}{d}\left(\frac{8d+2}{3\sqrt{2e}K}\right)^{4d+1} \exp\left[\frac{-3\sqrt{2enK}}{2}\right] c_d R^d.$$

Through Corollary 3.3.1 the result of Theorem 3.3 is illustrated for a given value of α depending on K and n. Corollary 3.3.1 provides a characteristic distance of the synthetic samples and their associated central point. The proof of Corollary 3.3.1 can be found in 8.5. The next step of our work is to study SMOTE near the boundary of the support of the minority class. This is the purpose of Theorem 3.4

Theorem 3.4. *Assume that there exists $R > 0$, such that $\mathcal{X} = B(0, R)$. Suppose that there exist C_2 such that, for all $x \in \mathbb{R}^d$, $f_X(x) \leq C_2 \mathbb{1}_{x \in \mathcal{X}}$. We consider $0 < \varepsilon < R$. Then, for all K, n, and all $z \in B(0, R) \backslash B(0, R - \varepsilon)$, and for $d > 1$, we have*

$$f_{Z_{K,n}}(z) \leq C(K, n, d, C_2) R^{(1/4)-d+1} \varepsilon^{1/4}. \tag{9}$$

with,

$$C(K, n, d, C_2) = \frac{2C_2}{d-1}\frac{(n-K-1)}{K}\binom{n-1}{K}\left(\frac{\Gamma(d/2)(2^{d-1}-1)\sqrt{2}}{2C_2\pi^{d/2}}\right)^{1/2}.$$

Theorem 3.4 highlights the fact that SMOTE density vanishes near the boundary of the minority class support. The proof of Theorem 3.4 can be found in 8.6. Theorem 3.4 justifies why the community introduced variants of SMOTE that tends to generate synthetic samples on the boundary of the minority class. An example is Borderline SMOTE which was introduced by Han et al. (2005).

4 Numerical illustrations

Through Section 3, we highlighted the fact that SMOTE asymptotically regenerate the distribution of the minority class, by copying the minority samples. The purpose of this section is to numerically illustrate the theoretical limitations of SMOTE procedure. We show in particular that with the default value $K = 5$, the SMOTE procedure generates data points that are very similar to the original data set.

4.1 Simulated data

In order to measure the similarity between any generated data set $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$ and the original data set $\mathbf{X} = \{X_1, \ldots, X_n\}$, we compute

$$C(\mathbf{Z}, \mathbf{X}) = \frac{1}{m} \sum_{i=1}^{m} \|Z_i - X_{(1)}(Z_i)\|_2, \tag{10}$$

where $X_{(1)}(Z_i)$ is the nearest neighbor of Z_i among $X_1, \ldots, X_n$. Intuitively, this quantity measures how far the generated data set is from the original observations: if the new data are copies of the original ones, this measure equals zero. We apply the following protocol: for each value of n,

1. We generate a data set $\mathbf{X}$ composed of n i.i.d samples $X_1, \ldots, X_n$ following a bivariate uniform distribution $\mathcal{U}([-3, 3]^2)$ (as chosen by Elreedy et al., 2023).
2. We generate a data set $\mathbf{Z}$ composed of $m = 1000$ i.i.d new observations $Z_1, \ldots, Z_m$ by applying SMOTE procedure on the original data set $\mathbf{X}$, with different values of K. We compute $C(\mathbf{Z}, \mathbf{X})$.
3. We generate a data set $\tilde{\mathbf{X}}$ composed of $m = 1000$ i.i.d new samples $\tilde{X}_1, \ldots, \tilde{X}_m$, distributed as $\mathcal{U}([-3, 3]^2)$. We compute $C(\tilde{\mathbf{X}}, \mathbf{X})$, which is a reference value in the ideal case of new points sampled from the same distribution.

Steps 1-3 are repeated $B = 75$ times. The average of $C(\mathbf{Z}, \mathbf{X})$ (resp. $C(\tilde{\mathbf{X}}, \mathbf{X})$) over these B repetitions is computed and denoted by $\bar{C}(\mathbf{Z}, \mathbf{X})$ (resp. $\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$). We consider two different metrics: $\bar{C}(\mathbf{Z}, \mathbf{X})$ and $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$. The results are depicted in Figure 1 and Figure 2.

Results. Figure 1 depicts the quantity $\bar{C}(\mathbf{Z}, \mathbf{X})$ as a function of the size of the minority class, for different values of K. The metric $\bar{C}(\mathbf{Z}, \mathbf{X})$ is consistently smaller for $K = 5$ than for other values of K, therefore highlighting that data generated by SMOTE with $K = 5$ are closer to the original data sample. This phenomenon is strengthened as n increases. This is an artifact of the simulation setting as the original data samples fill the input space as n increases.

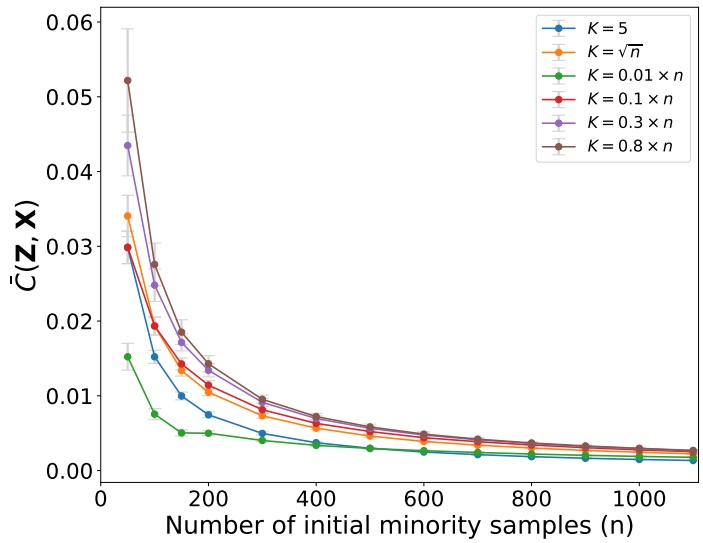

Figure 1: Average distance $\bar{C}(\mathbf{Z}, \mathbf{X})$ of SMOTE samples to their nearest neighbors in the original sample, distributed as $\mathcal{U}([-3, 3]^2)$.

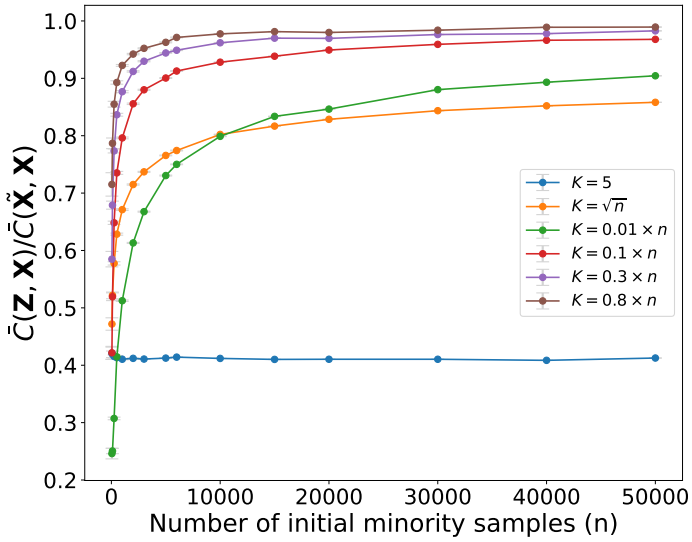

Figure 2: Average normalized distance $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$ of SMOTE samples to their nearest neighbor in the original sample, distributed as $\mathcal{U}([-3, 3]^2)$.

Figure 2 is similar to Figure 1, except that the renormalized quantity $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$ is plotted as a function of n. We notice that the asymptotic for $K = 5$ is different since it is the only one where the distance between SMOTE data points and original data points seems not to vary with n. Besides, this distance is smaller than the other ones, thus stressing out that the SMOTE data points are very close to the original distribution for $K = 5$. Note that, for the other asymptotics in K, the diversity of SMOTE observations increases with n, meaning $\bar{C}(\mathbf{Z}, \mathbf{X})$ gets closer from $\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$. Besides, this diversity is asymptotically more important for $K = 0.1n$ and $K = 0.01n$. This corroborates our theoretical findings (Theorem 3.2) as this asymptotics do not satisfy $K/n \to 0$. Indeed, when K is set to a fraction of n, the SMOTE distribution does not converge to the original distribution anymore, therefore generating data points that are not simple copies of the original uniform samples. By construction SMOTE data points are close to central points which may explain why the quantity of interest in Figure 2 is smaller than 1.

4.2 Extension to real-world data sets.

In this section, we apply SMOTE on real-world data and compare the distribution of the generated data points to the original distribution, using the metric $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$.

For each value of n, we subsample n data points from the minority class. Then,

1. We uniformly split the data set into $X_1, \ldots, X_{n/2}$ (denoted by $\mathbf{X}$) and $\tilde{X}_1, \ldots, \tilde{X}_{n/2}$ (denoted by $\tilde{\mathbf{X}}$).
2. We generate a data set $\mathbf{Z}$ composed of $m = n/2$ i.i.d new observations $Z_1, \ldots, Z_m$ by applying SMOTE procedure on the original data set $\mathbf{X}$, with different values of K. We compute $C(\mathbf{Z}, \mathbf{X})$.
3. We use $\tilde{\mathbf{X}}$ in order to compute $C(\tilde{\mathbf{X}}, \mathbf{X})$.

This procedure is repeated $B = 100$ times to compute averages values as in Section 4.1.

Results. We apply the above protocol to two real-world data sets (GA4 and Phoneme), described in Table 1. Figure 3 and Figure 4 display the quantity $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}, \mathbf{X})$ as a function of the size n of the minority class. As in Section 4.1, we observe in Figure 3 and Figure 4 that for the strategies, Average normalized distance $\bar{C}(\mathbf{Z}, \mathbf{X})/\bar{C}(\tilde{\mathbf{X}}$ increases except for SMOTE $K = 5$. The strategies using a value of hyperparameter K such that $K/n \to 0$ tends to converge to a value smaller than all the strategies with K such that $K/n \not\to 0$.

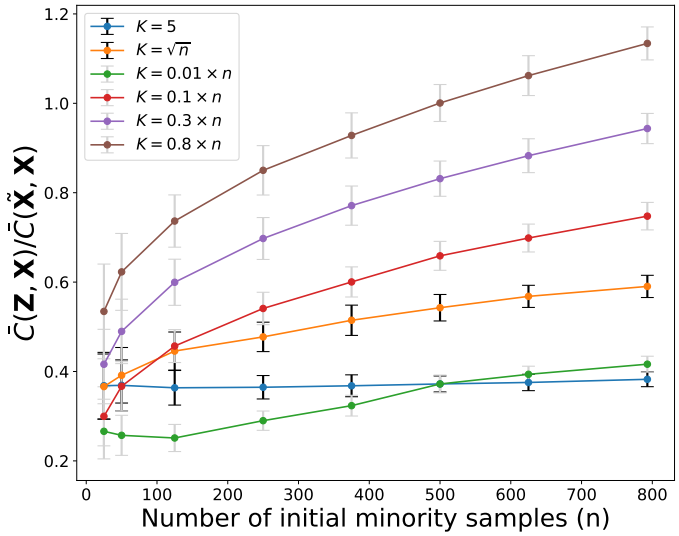

Figure 3: Protocol real-data on Phoneme data set.

Algorithm 2 Multivariate Gaussian SMOTE iteration.

Input: Minority class samples $X_c, \ldots, X_n$, number of nearest-neighbor K.
Select uniformly X_c (called **central point**), a random instance between $X_1, \ldots, X_n$.
Denote by $I = X_{(1)}(X_c) \ldots, X_{(K)}(X_c)$, the K nearest-neighbor of X_c (minimizing L_2 norm).
$\hat{\mu}(X_c) \leftarrow \frac{1}{K+1} \sum_{X_k \in I \cup \{X_c\}} X_k$
$\hat{\Sigma}(X_c) \leftarrow \frac{1}{K+1} \sum_{X_k \in I \cup \{X_c\}} (X_k - \hat{\mu})^T (X_k - \hat{\mu})$
Sample $Z \sim \mathcal{N}\left(\hat{\mu}(x_c), \hat{\Sigma}(x_c)\right)$
Return Z

5 Predictive evaluation on real-world data sets

In this section, we first describe the different rebalancing strategies and two new ones we propose, and then describe our experimental protocol. The data sets we consider are described in Table 1. They are open source data sets for imbalanced cases from UCI Irvine (see Dua and Graff, 2017), except for GA4, Phoneme and Credit Card.
Recall that in this paper, we focus on applying data sets containing only continuous explanatory variables. To this aim, we have removed all categorical variables in each data set.

5.1 Rebalancing strategies

Over/Under Sampling strategies Random Under Sampling (RUS) acts on the majority class by selecting uniformly without replacement several samples in order to obtain a prespecified size for the majority class. Similarly, Random Over Sampling (ROS) acts on the minority class by selecting uniformly with replacement several samples to be copied in order to obtain a prespecified size for the minority class.

Class weight The class weighting strategy assigns the same weight (hyperparameter of the procedure) to each minority samples. The default setting for this strategy is to choose a weight ρ such that $\rho n = N - n$, where n and $N - n$ are respectively the number of minority and majority samples in the data set.

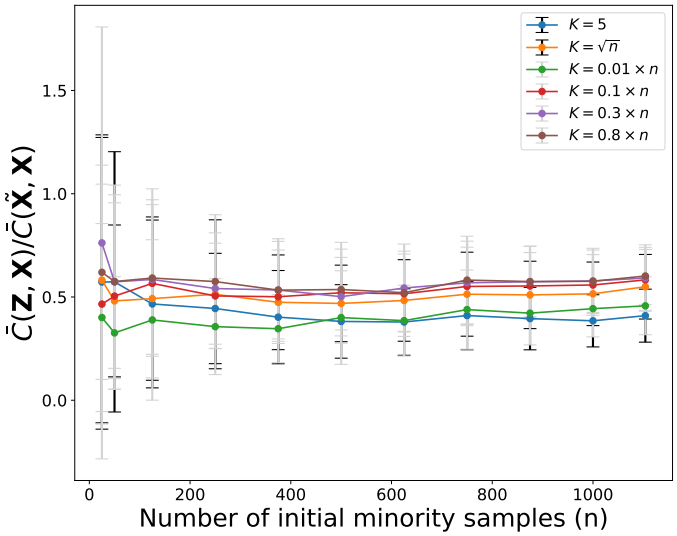

Figure 4: Protocol real-data on GA4 data set.

NearMissOne NearMissOne is an undersampling procedure. For each sample X_i in the majority class, the averaged distance of X_i to its K nearest neighbors in the minority class is computed. Then, the samples X_i are ordered (decreasingly) according to this averaged distance. Finally, from bottom to top, the X_i are dropped until the given number/ratio is reached. Consequently, the X_i with the smallest mean distance are dropped firstly.

Borderline SMOTE Borderline SMOTE 1 Han et al. (2005) procedure works as follows. For each individual X_i in the minority class, let $m_-(X_i)$ be the number of samples of the majority class among the m nearest neighbors of X_i, where m is a hyperparameter. For all X_i in the minority class such that $m/2 \leq m_-(X_i) < m$, do the following:

- Select uniformly X_k among the K nearest-neighbors of X_i in the minority class, and build a linear interpolation between X_i and this nearest neighbor as

$$Z = WX_i + (1 - W)X_k, \tag{11}$$

  where $W \sim \mathcal{U}([0, 1])$.

- Repeat Step 1 q times.

Borderline SMOTE 2 Han et al. (2005) works exactly as Borderline SMOTE 1 except that the selected point in Step 1 is uniformly chosen among the K nearest neighbors of x_i in the full original sample (including positive and negative examples). Then, the new observation Z is built as in Equation 11 with $W \sim \mathcal{U}([0, 0.5])$.

5.2 Introducing new oversampling strategies

The limitations of SMOTE highlighted in Section 3 drive us to two new rebalancing strategies.

CV SMOTE We introduce a new algorithm, called CV SMOTE, that finds the best hyperparameter K among a prespecified grid via a 5-fold cross-validation procedure. The grid is composed of the set $\{1, 2, \ldots, 15\}$ extended with the values $\lfloor 0.01 n_{train} \rfloor$, $\lfloor 0.1 n_{train} \rfloor$ and $\lfloor \sqrt{n_{train}} \rfloor$, where n_{train} is the number of minority samples in the training set.

Table 1: Description of the data sets, where n is the number of samples, and d the number of features.

	Total N	minority samples n/N	d
GA4	319 066	0.7%	7
CreditCard	284 315	0.2%	29
Abalone	4 177	1%	8
Phoneme	5 404	29%	5
Yeast	1 462	11%	8
Pima	768	35%	8
Wine	4 974	4%	11
Vehicule	846	23%	18
Ionosphere	351	36%	32
Haberman	306	26%	3
Breast cancer	630	36%	9

Recall that through Theorem 3.3, we show that SMOTE procedure with the default value of hyperparameter $K = 5$ asymptotically copy the original samples. The idea of CV SMOTE is then to try several values of K in order to avoid copying samples and probably get better improvement of the used classifier at the following step of the machine learning pipeline. CV SMOTE is the simplest idea that the theorems drive us as solution of SMOTE limitations.

Multivariate Gaussian SMOTE(K) Now, we introduce a new oversampling strategy named Multivariate Gaussian SMOTE (MGS). In this procedure, we generate new samples from the distribution $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$, where the empirical mean and covariance $\hat{\mu}, \hat{\Sigma}$ are estimated using the K neighbors and the central point. We detail one MGS iteration in Algorithm 2.

The idea behind MGS, is to exploit for the maximum the neighborhood of the central point. Using a multivariate gaussian distribution, which support is unbounded, reduces the risk of simply copying the original samples when $K/n \to 0$.

5.3 Experiments

Table 2: Random Forest ROC AUC for different rebalancing strategies and different data sets. Only datasets such that the None strategy is not among the best ones (displayed in bold) are displayed. Data sets artificially undersampled for minority class are in italics. Other data sets are presented in Table 3. Mean standard deviations are computed.

Resampling Strategy	None	Class weight	RUS	ROS	Near Miss1	BS1	BS2	Smote	CV Smote	MGS
GA4 (1%) (±0.004)	0.660	0.472	**0.866**	0.500	0.848	0.652	0.695	0.506	0.720	0.650
CreditCard (0.2%) (±0.003)	0.939	0.938	**0.975**	0.941	0.906	0.945	0.945	0.954	0.954	0.950
Abalone (1%) (±0.015)	0.697	0.702	0.719	0.712	0.570	0.712	0.742	0.756	0.750	**0.799**
Phoneme (1%) (±0.021)	0.819	0.821	0.851	0.814	0.575	0.847	0.861	0.876	0.877	**0.899**
Yeast (1%) (±0.019)	0.906	0.928	0.931	0.929	0.806	0.946	0.940	0.967	**0.968**	0.944
Wine (4%)(±0.008)	0.819	0.815	**0.846**	0.810	0.748	0.827	0.780	0.828	0.822	0.822
Pima (10%) (±0.012)	0.797	0.804	0.802	0.800	0.680	0.814	0.812	0.807	0.806	**0.821**
Haberman (26%) (±0.013)	0.684	0.680	0.670	0.673	**0.704**	0.652	0.660	0.687	0.675	0.664
Haberman (10%) (±0.037)	0.580	0.580	0.599	0.582	**0.634**	0.609	0.624	0.617	0.598	0.619

Protocol We compare the different rebalancing strategies on 11 real-world data sets, described in Table 1. We use 80%/20% (train/test) stratified split of the data, and apply each rebalancing strategy on the training set, in order to obtain a balanced data set. A learning procedure (Logistic regression or Random Forest) with

default hyperparameters is trained on the rebalanced training set. The performance is evaluated on the test set via the ROC AUC. This procedure is repeated 100 times and the averaged results are computed. We use the implementation of *imb-learn* (Lemaître et al., 2017) for the state-of-the-art strategies (see Appendix 7.1 for more details).

For 6 data sets out of 11, the None strategy is the best, probably highlighting that the imbalance ratio is not high enough or the learning task not difficult enough to require a tailored rebalancing strategy. To analyze what could happen for more imbalanced data sets, we use the following protocol. We subsample the minority class for each one of the 6 data sets mentioned above, so that the resulting imbalanced ratio is set to either 10% or 1%. This subsampling strategy is applied once for each data set and each imbalance ratio in a nested fashion, so that the minority class of the 1% data set is included in the 10% data set. Data sets such that the None strategy is not the best are displayed in Table 2. Other results are presented in Table 3 in Appendix 7.2.

Rebalancing methods Note that all data sets presented in Table 2 are highly imbalanced, with a ratio lower than 1% (10%, 26% and 10% for Pima, Haberman and Haberman(10%) data sets respectively). Whilst in the vast majority of scenarios, None is among the best approaches to deal with imbalanced data (see Table 3), it seems to be outperformed by dedicated rebalancing strategies for highly imbalanced data sets, presented in Table 2. Similar observations are extracted from Table 5 when using Logistic Regression. Therefore, considering only continuous input variables, and measuring the predictive performance with ROC AUC, we observe that rebalancing strategy are required only in some specific settings in which the minority class is highly under-represented. Besides, we see that RUS strategy shows an advantage for very large data-sets, that we expect to be less sensitive to the loss of information from undersampling.

Several seminal papers already noticed that the None strategy was competitive in terms of predictive performances. He et al. (2008) compare the None strategy, ADASYN and SMOTE, followed by a Decision tree classifier on 5 data sets (including Vehicle, Pima, Ionosphere and Abalone). In terms of Precision and F1 score, the None strategy is on par with the two other rebalancing methods. Han et al. (2005) study the impact of Borderline SMOTE and others SMOTE variant on 4 data sets (including Pima and Haberman). The None strategy is competitive (in terms of F1 score) on two of these data sets.

SMOTE We remark that the performances of CV SMOTE are comparable to that of SMOTE with the default hyperparameter setting ($K = 5$). This could be explained by our grid choice (which could be expanded) or by the data set characteristics. Indeed, the only data set for which we note that CV SMOTE is notably better than SMOTE is GA4, which contains the highest number of minority samples. This corresponds to our theoretical analysis (Theorem 3.3) that highlights that SMOTE, by default, tends to copy original minority samples, when the number of minority samples is large enough. Therefore, more analyses should be carried out to analyze the potential efficiency of CV SMOTE when the number of minority samples is large enough.

Multivariate Gaussian SMOTE(K) This new strategy exhibits good predictive performances. Indeed, as shown in Table 2, MGS has the best improvement on 3 data sets. This could be explained by the Gaussian sampling of synthetic observations that allows generated data points to fall outside the convex hull of the minority class, therefore limiting the border phenomenon, established in Theorem 3.4. MGS is potentially a promising new strategy, which will be available in an open source package.

6 Conclusion

Our work in this paper is both theoretical and experimental. We first proved that SMOTE (with default parameter) regenerates the original distribution by simply copying the original minority samples. We also established that SMOTE density vanishes near the boundary of the support of the minority distribution, therefore justifying the introduction of SMOTE variants focusing on the border. Our experiments show that for most data sets, the None strategy seems to be competitive, at least when the minority class is well represented. While our CV SMOTE approach is not efficient in general, our MGS proposal appears to be promising, by circumventing the borderline issue of SMOTE, exhibited in Theorem 3.4.

More experiments should be carried out to understand the surprising performances of RUS, which consistently outperforms ROS, whereas the two methods are really close, as they both rely on resampling. Besides, in order to analyze MGS(K) in more details, we would like to study the impact of a renormalizing factor λ in the covariance matrix estimation, such that the last step in Algorithm 2 would turn into $Z \sim \mathcal{N}(\hat{\mu}, \lambda\hat{\Sigma})$.

References

Aguiar, G., B. Krawczyk, and A. Cano (2023). A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. *Machine learning*, 1–79.

Berrett, T. B. (2017). Modern k-nearest neighbour methods in entropy estimation, independence testing and classification.

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

Bunkhumpornpat, C., K. Sinapiromsaran, and C. Lursinsap (2012). Dbsmote: density-based synthetic minority over-sampling technique. *Applied Intelligence 36*, 664–684.

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

Chawla, N. V., A. Lazarevic, L. O. Hall, and K. W. Bowyer (2003). Smoteboost: Improving prediction of the minority class in boosting. In *Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7*, pp. 107–119. Springer.

Dua, D. and C. Graff (2017). Uci machine learning repository.

Elreedy, D. and A. F. Atiya (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences 505*, 32–64.

Elreedy, D., A. F. Atiya, and F. Kamalov (2023, January). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*.

Han, H., W.-Y. Wang, and B.-H. Mao (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer.

Hassan, A. K. I. and A. Abraham (2016). Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*, pp. 117–127. Springer.

He, H., Y. Bai, E. A. Garcia, and S. Li (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328. Ieee.

He, H. and E. A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering 21*(9), 1263–1284.

Islam, J. and Y. Zhang (2020). Gan-based synthetic brain pet image generation. *Brain informatics 7*, 1–12.

Kamalov, F., A. F. Atiya, and D. Elreedy (2022). Partial resampling of imbalanced data. *arXiv preprint arXiv:2207.04631*.

Khalilia, M., S. Chakraborty, and M. Popescu (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making 11*, 1–13.

King, G. and L. Zeng (2001). Logistic regression in rare events data. *Political Analysis 9*(2), 137–163.

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

Lemaître, G., F. Nogueira, and C. K. Aridas (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research 18*(17), 1–5.

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

Mohammed, A. J., M. M. Hassan, and D. H. Kadir (2020). Improving classification performance for a novel imbalanced medical dataset using smote method. *International Journal of Advanced Trends in Computer Science and Engineering 9*(3), 3161–3172.

Nguyen, H. M., E. W. Cooper, and K. Kamei (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms 3*(1), 4–21.

Nguyen, N. N. and A. T. Duong (2021). Comparison of two main approaches for handling imbalanced data in churn prediction problem. *Journal of advances in information technology 12*(1).

Pan, T., J. Zhao, W. Wu, and J. Yang (2020). Learning imbalanced datasets based on smote and gaussian distribution. *Information Sciences 512*, 1214–1233.

Ramyachitra, D. and P. Manikandan (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR) 5*(4), 1–29.

Wadsworth, G. P., J. G. Bryan, and A. C. Eringen (1961). Introduction to probability and random variables. *Journal of Applied Mechanics 28*(2), 319.

Wongvorachan, T., S. He, and O. Bulut (2023). A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. *Information 14*(1), 54.

Zhu, M., J. Xia, X. Jin, M. Yan, G. Cai, J. Yan, and G. Ning (2018). Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access 6*, 4641–4652.

7 Experiments

7.1 Binary classification protocol

For each data set, the ratio hyperparameters of each rebalancing strategy are chosen so that the minority and majority class have the same weights in the training phase.The main purpose is to apply the strategies exactly on the same data points (X_{train}), then train the chosen classifier and evaluate the strategies on the same X_{test}. This objective is achieved by selecting each time 4 fold for the training, apply each of the strategies to these 4 exact same fold.

The state-of-the-art rebalancing strategies(see Lemaître et al., 2017) are used with their default hyperparameter values.

GA4 and CreditCard For the GA4 data set, a Time Series split is performed instead of a Stratified 5−fold, because of the temporality of the data. Furthermore, a group out is applied on the different client ID. This means that a client that appears in the training set (composed each time of 4 folds) is systemically deleted from the test set if it appears inside. For the CreditCard data set, a 5 fold Time Series split is performed instead of a Stratified 5−fold, because of the temporality of the data.

7.2 Additional experiments

Table 3: Random Forest ROC AUC for different rebalancing strategies and different data sets. Only datasets such that the None strategy is on par with the best strategies are displayed. Other data sets are presented in Table 2. Mean standard deviations are computed. The best strategy is highlighted in bold for each data set. CV SMOTE is not applied to data sets artifically undersampled Ionosphere(1%) and Breast Cancer (1%) due to the small number of minority class samples.

Strategy	None	Class weight	RUS	ROS	Near Miss1	BS1	BS2	Smote	CV Smote	MGS
Phoneme (±0.0007)	0.961	0.961	0.950	**0.962**	0.900	0.960	0.961	**0.962**	**0.962**	0.961
Phoneme (10%) (±0.003)	0.957	0.957	0.936	0.957	0.649	0.957	0.957	0.958	0.959	**0.961**
Pima (±0.004)	**0.826**	0.826	0.823	0.824	0.803	0.813	0.812	0.823	0.823	0.826
Yeast (±0.003)	0.960	0.964	**0.968**	0.964	0.906	0.962	0.965	0.964	0.966	0.965
Vehicle (±0.001)	0.994	0.994	0.990	0.995	0.924	0.995	0.994	0.994	0.994	**0.994**
Vehicle (10%) (±0.003)	0.992	0.993	0.978	**0.994**	0.755	0.991	0.988	0.992	0.991	0.992
Ionosphere (±0.002)	**0.974**	0.973	0.967	0.973	0.944	0.971	0.972	0.971	0.972	0.970
Ionosphere (10%) (±0.008)	**0.969**	0.965	0.932	0.965	0.822	0.949	0.946	0.949	0.957	0.950
Ionosphere (1%) (±0.006)	**1.0**	1.0	0.962	1.0	0.980	1.0	0.985	1.0		0.999
Breast Cancer (±0.001)	**0.992**	0.991	0.991	0.991	0.990	0.991	0.990	0.991	0.991	0.992
Breast Cancer (10%) (±0.002)	0.993	0.993	0.992	0.993	0.990	0.992	0.989	0.993	0.992	**0.994**
Breast Cancer (1%) (±0.001)	**1.0**	0.999	0.999	0.999	0.999	0.999	0.999	0.999		0.999

Table 4: Random Forest ROC AUC SMOTE strategies. The best strategy is highlighted in bold for each data set.

SMOTE Strategy	$K = 5$	$K = \sqrt{n}$	$K = 0.01n$	$K = 0.1n$	CV Smote
GA4 (±0.004)	0.506	0.639	0.567	0.701	**0.720**
CreditCard (±0.003)	0.954	0.961	**0.962**	0.965	0.954
Abalone (±0.015)	**0.756**	**0.756**	0.719	0.725	0.750
Phoneme (±0.001)	**0.962**	0.961	**0.962**	0.960	**0.962**
Pima (±0.003)	**0.826**	0.823	0.823	0.823	0.823
Wine (±0.007)	**0.828**	0.825	**0.828**	0.825	0.822
Yeast (±0.003)	**0.964**	0.965	0.964	0.965	**0.966**
Vehicle (±0.001)	**0.994**	**0.994**	**0.994**	**0.994**	**0.994**
Ionosphere (±0.002)	0.971	0.971	**0.972**	**0.972**	**0.972**
Haberman (±0.013)	**0.675**	**0.675**	0.672	**0.677**	0.664
Breast cancer (±0.001)	**0.991**	0.991	0.991	0.991	0.990

Table 5: Logistic Regression ROC AUC. The best strategy is highlighted in bold for each data set. CV SMOTE is not applied to data sets artifically undersampled Ionosphere(1%) and Breast Cancer (1%) due to the small number of minority class samples.

Strategy	None	Class weight	RUS	ROS	Near Miss1	BS1	BS2	Smote	CV Smote	MGS
GA4 (±0.001)	0.831	0.866	0.865	0.866	**0.896**	0.862	0.870	0.859	0.862	0.855
CreditCard (±0.001)	0.939	0.938	**0.975**	0.941	0.906	0.945	0.945	0.954	0.954	0.950
Abalone (±0.002)	0.767	0.855	0.734	0.866	0.777	0.868	**0.873**	0.866	0.866	0.848
Phoneme (±0.001)	**0.812**	0.811	0.811	0.811	0.571	0.801	0.805	0.811	0.811	0.810
Phoneme (10%) (±0.001)	**0.809**	0.806	0.805	0.806	0.431	0.801	0.805	0.805	0.805	0.805
Phoneme (1%)(±0.009)	0.798	0.803	0.787	0.803	0.435	0.771	0.767	0.802	**0.805**	0.802
Yeast (±0.001)	0.960	0.964	**0.968**	0.965	0.906	0.962	0.965	0.964	0.966	0.965
Yeast (1%) (±0.015)	0.893	0.947	0.867	**0.949**	0.733	0.926	0.931	0.947	0.947	0.941
Pima (±0.004)	**0.830**	0.832	0.823	0.826	0.812	0.819	0.819	0.825	0.825	0.826
Pima (10%) (±0.008)	0.785	0.776	0.774	0.782	0.684	0.787	**0.796**	0.782	0.778	0.787
Wine (±0.004)	0.827	**0.846**	0.831	0.845	0.722	0.845	0.805	0.845	0.845	0.842
Vehicle (±0.001)	**0.993**	0.991	0.991	0.993	0.994	0.992	0.988	0.993	0.993	0.991
Vehicle (10%) (±0.003)	**0.994**	0.995	0.989	0.994	0.983	0.994	0.989	0.994	0.994	0.994
Ionosphere (±0.006)	**0.878**	0.875	0.879	0.870	0.842	0.854	0.858	0.868	0.871	0.865
Ionosphere (10%) (±0.019)	**0.950**	0.947	0.925	0.943	0.848	0.936	0.938	0.946	0.951	0.932
Ionosphere (1%) (±0.008)	**0.997**	0.997	0.947	0.997	0.915	0.997	0.994	0.997		0.997
Breast Cancer (±0.001)	0.610	0.721	0.681	0.693	**0.757**	0.623	0.622	0.680	0.648	0.676
Breast Cancer (10%)(±0.069)	0.576	0.657	0.652	0.659	0.589	**0.669**	0.647	0.658	0.665	0.654
Breast Cancer (1%) (±0.059)	0.850	0.855	**0.881**	0.852	0.335	0.850	0.859	0.857		0.869
Haberman (±0.013)	0.694	0.698	0.688	0.693	**0.720**	0.664	0.652	0.685	0.681	0.687
Haberman (10%)(±0.003)	0.633	0.634	0.611	0.632	0.668	0.598	0.605	0.618	0.615	0.598

Table 6: Logistic regression ROC AUC SMOTE strategies. The best strategy is highlighted in bold for each data set.

SMOTE Strategy	$K = 5$	$K = \sqrt{n}$	$K = 0.01n$	$K = 0.1n$	CV Smote
GA4 (±0.001)	0.859	0.857	0.857	0.860	**0.862**
CreditCard (±0.001)	0.954	0.961	0.961	**0.965**	0.954
Abalone (±0.001)	0.866	0.866	0.867	**0.868**	0.824
Phoneme (±0.001)	**0.811**	0.811	0.811	0.811	0.811
Pima (±0.004)	**0.825**	0.824	0.824	**0.825**	**0.825**
Wine (±0.003)	**0.845**	0.845	0.843	0.845	0.845
Yeast (±0.003)	**0.966**	0.965	0.966	0.965	0.966
Vehicle (±0.001)	0.992	0.992	0.993	0.992	**0.993**
Ionosphere (±0.007)	0.868	0.871	0.868	0.871	**0.871**
Breast cancer(±0.001)	0.680	0.668	**0.682**	0.674	0.648
Haberman (±0.013)	0.685	0.687	0.688	**0.689**	0.681

8 Main proofs

8.1 Proof of Lemma 1

Proof of Lemma 1. Let $\mathcal{X}$ be the support of P_X. SMOTE generates new points by linear interpolation of the original minority sample. This means that for all x, y in the minority samples or generated by SMOTE procedure, we have $(1-t)x + ty \in Conv(\mathcal{X})$ by definition of $Conv(\mathcal{X})$. This leads to the fact that precisely, all the new SMOTE samples are contained in $Conv(\mathcal{X})$. This implies $Supp(P_Z) \subseteq Conv(\mathcal{X})$.

□

8.2 Proof of Theorem 3.1

Proof of Theorem 3.1. We consider a single SMOTE iteration. Recall that the central point X_c (see Algorithm 1) is fixed, and thus denoted by x_c.

The random variables $X_{(1)}(x_c), \ldots, X_{(n-1)}(x_c)$ denote a reordering of the initial observations $X-1, X_2, \ldots, X_n$ such that

$$||X_{(1)}(x_c) - x_c|| \leq ||X_{(2)}(x_c) - x_c|| \leq \ldots \leq ||X_{(n-1)}(x_c) - x_c||.$$

For clarity, we remove the explicit dependence on x_c. Recall that SMOTE builds a linear interpolation between x_c and one of its K nearest neighbors chosen uniformly. Then the newly generated point Z satisfies

$$Z = (1-W)x_c + W \sum_{k=1}^{K} X_{(k)} \mathbb{1}_{\{I=k\}}, \tag{12}$$

where W is a uniform random variable over $[0,1]$, independent of $I, X_1, \ldots, X_n$, with I distributed as $\mathcal{U}(\{1, \ldots, K\})$.

From now, consider that the k-th nearest neighbor of x_c, $X_{(k)}(x_c)$, has been chosen (that is $I = k$). Then Z satisfies

$$\begin{aligned} Z &= (1-W)x_c + WX_{(k)} && (13)\\ &= x_c - Wx_c + WX_{(k)}, && (14) \end{aligned}$$

which implies

$$Z - x_c = W(X_{(k)} - x_c). \tag{15}$$

Let f_{Z-x_c}, f_W and $f_{X_{(k)}-x_c}$ be respectively the density functions of $Z - x_c$, W and $X_{(k)} - x_c$. Let $z, z_1, z_2 \in \mathbb{R}^d$. Recall that $z \leq z_1$ means that each component of z is lower than the corresponding component of z_1. Since W and $X_{(k)} - x_c$ are independent, we have,

$$\begin{aligned} \mathbb{P}(z_1 \leq Z - x_c \leq z_2) &= \int_{w\in\mathbb{R}} \int_{x\in\mathbb{R}^d} f_{W, X_{(k)}-x_c}(w,x) \mathbb{1}_{\{z_1 \leq wx \leq z_2\}} \mathrm{d}w\mathrm{d}x && (16)\\ &= \int_{w\in\mathbb{R}} \int_{x\in\mathbb{R}^d} f_W(w) f_{X_{(k)}-x_c}(x) \mathbb{1}_{\{z_1 \leq wx \leq z_2\}} \mathrm{d}w\mathrm{d}x && (17)\\ &= \int_{w\in\mathbb{R}} f_W(w) \left(\int_{x\in\mathbb{R}^d} f_{X_{(k)}-x_c}(x) \mathbb{1}_{\{z_1 \leq wx \leq z_2\}} \mathrm{d}x \right) \mathrm{d}w. && (18) \end{aligned}$$

Besides, let $u = wx$. Then $x = (\frac{u_1}{w}, \ldots, \frac{u_d}{w})^T$.The Jacobian of such transformation equals:

$$\begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial u_1} & \cdots & \frac{\partial x_d}{\partial u_d} \end{vmatrix} = \begin{vmatrix} \frac{1}{w} & & 0 \\ & \ddots & \\ 0 & \cdots & \frac{1}{w} \end{vmatrix} = \frac{1}{w^d} \tag{19}$$

Therefore, we have $x = u/w$ and $\mathrm{d}x = \mathrm{d}u/w^d$, which leads to

$$\begin{aligned} &\mathbb{P}(z_1 \leq Z - x_c \leq z_2) && (20)\\ &= \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) \left(\int_{u\in\mathbb{R}^d} f_{X_{(k)}-x_c} \left(\frac{u}{w}\right) \mathbb{1}_{\{z_1 \leq u \leq z_2\}} \mathrm{d}u \right) \mathrm{d}w. && (21) \end{aligned}$$

Note that a random variable Z' with density function

$$f_{Z'}(z') = \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) f_{X_{(k)}-x_c}\left(\frac{z'}{w}\right) \mathrm{d}w \tag{22}$$

satisfies, for all $z_1, z_2 \in \mathbb{R}^d$,

$$\mathbb{P}(z_1 \leq Z - x_c \leq z_2) = \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) \left(\int_{u\in\mathbb{R}^d} f_{X_{(k)}-x_c}\left(\frac{u}{w}\right) \mathbb{1}_{\{z_1\leq u\leq z_2\}} \mathrm{d}u \right) \mathrm{d}w. \tag{23}$$

Therefore, the variable $Z - x_c$ admits the following density

$$f_{Z-x_c}(z'|X_c = x_c, I = k) = \int_{w\in\mathbb{R}} \frac{1}{w^d} f_W(w) f_{X_{(k)}-x_c}\left(\frac{z'}{w}\right) \mathrm{d}w. \tag{24}$$

Since W follows a uniform distribution on $[0, 1]$, we have

$$f_{Z-x_c}(z'|X_c = x_c, I = k) = \int_0^1 \frac{1}{w^d} f_{X_{(k)}-x_c}\Big(\frac{z'}{w}\Big) \mathrm{d}w. \tag{25}$$

The density $f_{X_{(k)}-x_c}$ of the k-th nearest neighbor of x_c can be computed exactly (see, Lemma 6.1 in Berrett, 2017), that is

$$\begin{aligned} f_{X_{(k)}-x_c}(u) = (n-1)\binom{n-2}{k-1} f_X(x_c+u) \big[\mu_X\left(B(x_c, ||u||)\right)\big]^{k-1} \\ \times \big[1 - \mu_X\left(B(x_c, ||u||)\right)\big]^{n-k-1}, \end{aligned} \tag{26}$$

where

$$\mu_X\left(B(x_c, ||u||)\right) = \int_{B(x_c,\|u\|)} f_X(x) \mathrm{d}x. \tag{27}$$

We recall that $B(x_c, \|u\|)$ is the ball centered on x_c and of radius $\|u\|$. Hence we have

$$f_{X_{(k)}-x_c}(u) = (n-1)\binom{n-2}{k-1} f_X(x_c+u) \mu_X\left(B(x_c, ||u||)\right)^{k-1} \left[1 - \mu_X\left(B(x_c, ||u||)\right)\right]^{n-k-1}. \tag{28}$$

Since $Z - x_c$ is a translation of the random variable Z, we have

$$f_Z(z|X_c = x_c, I = k) = f_{Z-x_c}(z - x_c|X_c = x_c, I = k). \tag{29}$$

Injecting Equation (28) in Equation (25), we obtain

$$\begin{aligned} & f_Z(z|X_c = x_c, I = k) && (30)\\ & = f_{Z-x_c}(z - x_c|X_c = x_c, I = k) && (31)\\ & = \int_0^1 \frac{1}{w^d} f_{X_{(k)}-x_c}\Big(\frac{z-x_c}{w}\Big) \mathrm{d}w && (32)\\ & = (n-1)\binom{n-2}{k-1} \int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k-1} && (33)\\ & \quad \times \left[1 - \mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-1} \mathrm{d}w && (34) \end{aligned}$$

Recall that in SMOTE, k is chosen at random in $\{1, \dots, K\}$ through the uniform random variable I. So far, we have considered I fixed. Taking the expectation with respect to I, we have

$$
\begin{aligned}
& f_Z(z|X_c = x_c) && (35)\\
&= \sum_{k=1}^{K} f_Z(z|X_c = x_c, I = k)\mathbb{P}[I = k] && (36)\\
&= \frac{1}{K}\sum_{k=1}^{K}\int_0^1 \frac{1}{w^d} f_{X_{(k)}-x_c}\Big(\frac{z-x_c}{w}\Big)\mathrm{d}w && (37)\\
&= \frac{1}{K}\sum_{k=1}^{K}(n-1)\binom{n-2}{k-1}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k-1} && (38)\\
&\quad \times \left[1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-1}\mathrm{d}w && (39)\\
&= \frac{(n-1)}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\sum_{k=1}^{K}\binom{n-2}{k-1}\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k-1} && (40)\\
&\quad \times \left[1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-1}\mathrm{d}w && (41)\\
&= \frac{(n-1)}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\sum_{k=0}^{K-1}\binom{n-2}{k}\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k} && (42)\\
&\quad \times \left[1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-2}\mathrm{d}w. && (43)
\end{aligned}
$$

Note that the sum can be expressed as the cumulative distribution function of a Binomial distribution parameterized by $n-2$ and $\mu_X(B(x_c, \|z-x_c\|/w))$, so that

$$
\begin{aligned}
& \sum_{k=0}^{K-1}\binom{n-2}{k}\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)^{k}\left[1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-k-2} && (44)\\
&= (n-K-1)\binom{n-2}{K-1}\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right), && (45)
\end{aligned}
$$

(see Technical Lemma 2 for details). We inject Equation (45) in Equation (35)

$$
\begin{aligned}
f_Z(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)& \\
\times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w.& \quad (46)
\end{aligned}
$$

We know that

$$
f_Z(z) = \int_{x_c \in \mathcal{X}} f_Z(z|X_c = x_c) f_X(x_c)\mathrm{d}x_c.
$$

Combining this remark with the result of Equation (46) we get

$$
\begin{aligned}
f_Z(z) = (n-K-1)\binom{n-1}{K}\int_{x_c\in\mathcal{X}}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)& \\
\times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) f_X(x_c)\mathrm{d}w\mathrm{d}x_c.& \quad (47)
\end{aligned}
$$

Link with Elreedy's formula According to the Elreedy formula

$$
\begin{aligned}
f_Z(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_{r=\|z-x_c\|}^{\infty} f_X\left(x_c + \frac{(z-x_c)r}{\|z-x_c\|}\right)\frac{r^{d-2}}{\|z-x_c\|^{d-1}}& \\
\times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, r\right)\right)\right)\mathrm{d}r.& \quad (48)
\end{aligned}
$$

Now, let $r = \|z - x_c\|/w$ so that $\mathrm{d}r = -\|z - x_c\|\mathrm{d}w/w^2$. Thus,

$$
\begin{aligned}
&f_Z(z|X_c = x_c) \\
&= (n-K-1)\binom{n-1}{K}\int_0^1 f_X\left(x_c + \frac{z-x_c}{w}\right)\frac{1}{w^{d-2}}\frac{1}{\|z-x_c\|} && (49)\\
&\quad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{z-x_c}{w}\right)\right)\right)\frac{\|z-x_c\|}{w^2}\mathrm{d}w && (50)\\
&= (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{z-x_c}{w}\right)\right)\right)\mathrm{d}w. && (51)
\end{aligned}
$$

□

8.3 Proof of Theorem 3.2

Proof of Theorem 3.2. For any event A, B, we have

$$1 - \mathbb{P}[A \cap B] = \mathbb{P}[A^c \cup B^c] \leq \mathbb{P}[A^c] + \mathbb{P}[B^c], \tag{52}$$

which leads to

$$
\begin{aligned}
\mathbb{P}[A \cap B] &\geq 1 - \mathbb{P}[A^c] - \mathbb{P}[B^c] && (53)\\
&= \mathbb{P}[A] - \mathbb{P}[B^c]. && (54)
\end{aligned}
$$

By construction,

$$\|X_c - Z\| \leq \|X_c - X_{(K)}(X_c)\|. \tag{55}$$

Let $x \in \mathcal{X}$ and $\eta > 0$. Let $\alpha, \varepsilon > 0$. We have,

$$
\begin{aligned}
&\mathbb{P}[X_c \in B(x, \alpha - \varepsilon)] - \mathbb{P}[\|X_c - X_{(K)}(X_c)\| > \varepsilon] && (56)\\
&\leq \mathbb{P}[X_c \in B(x, \alpha - \varepsilon), \|X_c - X_{(K)}(X_c)\| \leq \varepsilon] && (57)\\
&\leq \mathbb{P}[X_c \in B(x, \alpha - \varepsilon), \|X_c - Z\| \leq \varepsilon] && (58)\\
&\leq \mathbb{P}[Z \in B(x, \alpha)]. && (59)
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\mathbb{P}[Z \in B(x, \alpha)] - \mathbb{P}[\|X_c - X_{(K)}(X_c)\| > \varepsilon] && (60)\\
&\leq \mathbb{P}[Z \in B(x, \alpha), \|X_c - X_{(K)}(X_c)\| \leq \varepsilon] && (61)\\
&\leq \mathbb{P}[Z \in B(x, \alpha), \|X_c - Z\| \leq \varepsilon] && (62)\\
&\leq \mathbb{P}[X_c \in B(x, \alpha + \varepsilon)]. && (63)
\end{aligned}
$$

Since X_c admits a density, for all $\varepsilon > 0$ small enough

$$\mathbb{P}[X_c \in B(x, \alpha + \varepsilon)] \leq \mathbb{P}[X_c \in B(x, \alpha)] + \eta, \tag{64}$$

and

$$\mathbb{P}[X_c \in B(x, \alpha)] - \eta \leq \mathbb{P}[X_c \in B(x, \alpha - \varepsilon)]. \tag{65}$$

Let ε such that (64) and (65) are verified. According to Lemma 2.3 in Biau and Devroye (2015), since $X_1, \ldots, X_n$ are i.i.d., if K/n tends to zero as $n \to \infty$, we have

$$\mathbb{P}[\|X_c - X_{(K)}(X_c)\| > \varepsilon] \to 0. \tag{66}$$

Thus, for all n large enough,

$$\mathbb{P}[X_c \in B(x, \alpha)] - 2\eta \leq \mathbb{P}[Z \in B(x, \alpha)] \tag{67}$$

and

$$\mathbb{P}[Z \in B(x,\alpha)] \leq 2\eta + \mathbb{P}[X_c \in B(x,\alpha)]. \tag{68}$$

Finally, for all $\eta > 0$, for all n large enough, we obtain

$$\mathbb{P}[X_c \in B(x,\alpha)] - 2\eta \leq \mathbb{P}[Z \in B(x,\alpha)] \leq 2\eta + \mathbb{P}[X_c \in B(x,\alpha)], \tag{69}$$

which proves that

$$\mathbb{P}[Z \in B(x,\alpha)] \to \mathbb{P}[X_c \in B(x,\alpha)]. \tag{70}$$

Therefore, by the Monotone convergence theorem, for all Borel sets $B \subset \mathbb{R}^d$,

$$\mathbb{P}[Z \in B] \to \mathbb{P}[X_c \in B]. \tag{71}$$

□

8.4 Proof of Theorem 3.3

Proof of Theorem 3.3. Let $x_c \in \mathcal{X}$ be a central point in a SMOTE iteration. From Theorem 3.1, we have,

$$\begin{aligned}
&f_Z(z|X_c = x_c) \\
&= (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w \qquad (72)\\
&= (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mathbb{1}_{\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\}} \\
&\qquad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w. \qquad (73)
\end{aligned}$$

Let $R \in \mathbb{R}$ such that $\mathcal{X} \subset \mathcal{B}(0,R)$. For all $u = x_c + \frac{z-x_c}{w}$, we have

$$w = \frac{||z-x_c||}{||u-x_c||}. \tag{74}$$

If $u \in \mathcal{X}$, then $u \in \mathcal{B}(0,R)$. Besides, since $x_c \in \mathcal{X} \subset B(0,R)$, we have $||u-x_c|| < 2R$ and

$$w > \frac{||z-x_c||}{2R}. \tag{75}$$

Consequently,

$$\mathbb{1}_{\left\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\right\}} \leq \mathbb{1}_{\left\{w>\frac{||z-x_c||}{2R}\right\}}. \tag{76}$$

So finally

$$\mathbb{1}_{\left\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\right\}} = \mathbb{1}_{\left\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\right\}}\mathbb{1}_{\left\{w>\frac{||z-x_c||}{2R}\right\}}. \tag{77}$$

Hence,

$$\begin{aligned}
f_Z(z|X_c = x_c) &= (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mathbb{1}_{\left\{x_c+\frac{z-x_c}{w}\in\mathcal{X}\right\}}\mathbb{1}_{\left\{w>\frac{||z-x_c||}{2R}\right\}} \\
&\qquad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w \qquad (78)\\
&= (n-K-1)\binom{n-1}{K}\int_{\frac{||z-x_c||}{2R}}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \\
&\qquad \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w. \qquad (79)
\end{aligned}$$

Now, let $0 < \alpha \leq 2R$ and $z \in \mathbb{R}^d$ such that $||z - x_c|| > \alpha$. In such a case, $w > \frac{\alpha}{2R}$ and:

$$f_Z(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) \mathrm{d}w. \tag{80}$$

Using Lemma 3, we have

$$f_Z(z|X_c = x_c)$$
$$= (n-K-1)\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right) \mathrm{d}w \tag{81}$$
$$\leq \frac{(n-K-1)}{K}\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \times \left[1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^{n-K-2}\left[1-\left[\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right]^K\right] \mathrm{d}w \tag{82}$$
$$\leq \frac{(n-K-1)}{K}\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \times \left[1-\mu_X\left(B\left(x_c, \frac{\alpha}{w}\right)\right)\right]^{n-K-2}\left[1-\left[\mu_X\left(B\left(x_c, \frac{\alpha}{w}\right)\right)\right]^K\right] \mathrm{d}w. \tag{83}$$

Recall that there is $C_2 \in \mathbb{R}$ such that $f_X \leq C_2$. Hence, for all $z \in \mathbb{R}^d$ such that $||z - x_c|| > \alpha$,

$$f_Z(z|X_c = x_c) \tag{84}$$
$$\leq \frac{C_2(n-K-1)}{K}\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d}\left[1-\mu_X\left(B\left(x_c, \frac{\alpha}{w}\right)\right)\right]^{n-K-2}\left[1-\left[\mu_X\left(B\left(x_c, \frac{\alpha}{w}\right)\right)\right]^K\right] \mathrm{d}w \tag{85}$$
$$\leq \frac{C_2(n-K-1)}{K}\binom{n-1}{K}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d}\left[1-\mu_X\left(B\left(x_c, \alpha\right)\right)\right]^{n-K-2}\left[1-\left[\mu_X\left(B\left(x_c, \alpha\right)\right)\right]^K\right] \mathrm{d}w \tag{86}$$
$$\leq \frac{C_2(n-K-1)}{K}\binom{n-1}{K}\left[1-\mu_X\left(B\left(x_c, \alpha\right)\right)\right]^{n-K-2}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d}\mathrm{d}w \tag{87}$$
$$= C_2\frac{n}{K}\left(1-\frac{K-1}{n}\right)\binom{n-1}{K}\left[1-\mu_X\left(B\left(x_c, \alpha\right)\right)\right]^{n-K-2}\int_{\frac{\alpha}{2R}}^1 \frac{1}{w^d}\mathrm{d}w \tag{88}$$
$$= \eta(\alpha, R)\frac{C_2 n}{K}\left(1-\frac{K-1}{n}\right)\binom{n-1}{K}\left[1-\mu_X\left(B\left(x_c, \alpha\right)\right)\right]^{n-K-2} \tag{89}$$

with

$$\eta(\alpha, R) = \begin{cases} \ln\left(\frac{2R}{\alpha}\right) & \text{if } d = 1 \\ \frac{1}{d-1}\left(\left(\frac{2R}{\alpha}\right)^{d-1} - 1\right) & \text{otherwise} \end{cases}.$$

Using Lemma 4, and letting $\beta_{x_c,\alpha} = \mu_X\left(B\left(x_c, \alpha\right)\right)$, we have

$$f_Z(z|X_c = x_c) \tag{90}$$
$$\leq \eta(\alpha, R)\frac{C_2 n}{K}\left(1-\frac{K-1}{n}\right)\left(\frac{e(n-1)}{K}\right)^K\left(1-\beta_{x_c,\alpha}\right)^{n-K-2} \tag{91}$$
$$\leq \eta(\alpha, R)C_2\left(\frac{n}{K}\right)^{K+1} e^K\left(1-\frac{K-1}{n}\right)\left(1-\beta_{x_c,\alpha}\right)^{n-K-2} \tag{92}$$
$$\leq \eta(\alpha, R)C_2\left(\frac{n}{K}\right)^{K+1} e^K\left(1-\beta_{x_c,\alpha}\right)^{n-K-2} \tag{93}$$
$$= C_2\eta(\alpha, R)\exp\left[n\left(\frac{K}{n} + \frac{(K+1)}{n}\ln\left(\frac{n}{K}\right) - \frac{K+2}{n}\ln(1-\beta_{x_c,\alpha}) + \ln(1-\beta_{x_c,\alpha})\right)\right]. \tag{94}$$

So,

$$\frac{K}{n}+\frac{(K+1)}{n}\ln\left(\frac{n}{K}\right)-\frac{K+2}{n}\ln(1-\beta_{x_c,\alpha})\leq\frac{K+2}{n}+\frac{K+2}{n}\ln\left(\frac{n}{K}\right)+\frac{K+2}{n}\ln\left(\frac{1}{1-\beta_{x_c,\alpha}}\right) \tag{95}$$

$$=\frac{K+2}{n}\left[1+\ln\left(\frac{n}{K}\right)+\ln\left(\frac{1}{1-\beta_{x_c,\alpha}}\right)\right] \tag{96}$$

$$=\frac{K+2}{n}\left[\ln(e)+\ln\left(\frac{n}{K}\right)+\ln\left(\frac{1}{1-\beta_{x_c,\alpha}}\right)\right] \tag{97}$$

$$=\frac{K+2}{n}\ln\left(\frac{en}{K(1-\beta_{x_c,\alpha})}\right). \tag{98}$$

Since $K\geq 1$, we have,

$$\frac{K+2}{n}\ln\left(\frac{en}{K(1-\beta_{x_c,\alpha})}\right)\leq\frac{K+2K}{n}\ln\left(\frac{en}{K(1-\beta_{x_c,\alpha})}\right) \tag{99}$$

$$=\frac{3K}{n}\ln\left(\frac{en}{K(1-\beta_{x_c,\alpha})}\right) \tag{100}$$

$$=\frac{K(1-\beta_{x_c,\alpha})}{en}\ln\left(\frac{en}{K(1-\beta_{x_c,\alpha})}\right)\frac{3e}{1-\beta_{x_c,\alpha}} \tag{101}$$

Using Lemma 5, we have

$$\frac{K(1-\beta_{x_c,\alpha})}{en}\ln\left(\frac{en}{K(1-\beta_{x_c,\alpha})}\right)\frac{3e}{1-\beta_{x_c,\alpha}}\leq\sqrt{\frac{K(1-\beta_{x_c,\alpha})}{en}}\frac{3e}{1-\beta_{x_c,\alpha}} \tag{102}$$

$$=3\sqrt{\frac{e}{1-\beta_{x_c,\alpha}}}\sqrt{\frac{K}{n}} \tag{103}$$

So finally,

$$\frac{K}{n}+\frac{(K+1)}{n}\ln\left(\frac{n}{K}\right)-\frac{K+2}{n}\ln(1-\beta_{x_c,\alpha})\leq 3\sqrt{\frac{e}{1-\beta_{x_c,\alpha}}}\sqrt{\frac{K}{n}}. \tag{104}$$

Hence, for all $z\notin B(x_c,\alpha)$,

$$f_Z(z|X_c=x_c)\leq C_2\eta(\alpha,R)\exp\left[n\left(3\sqrt{\frac{e}{1-\beta_{x_c,\alpha}}}\sqrt{\frac{K}{n}}+\ln\left(1-\beta_{x_c,\alpha}\right)\right)\right] \tag{105}$$

$$=\epsilon(n,\alpha,K,x_c). \tag{106}$$

Consequently,

$$\mathbb{P}\left(|Z-X_c|\geq\alpha|X_c=x_c\right)=\int_{z\notin\mathcal{B}(x_c,\alpha),z\in\mathcal{X}}f_Z(z|X_c=x_c)\mathrm{d}z \tag{107}$$

$$\leq\int_{z\notin\mathcal{B}(x_c,\alpha),z\in\mathcal{X}}\varepsilon(n,\alpha,K,x_c)\mathrm{d}z \tag{108}$$

$$=\varepsilon(n,\alpha,K,x_c)\int_{z\notin\mathcal{B}(x_c,\alpha),z\in\mathcal{X}}\mathrm{d}z \tag{109}$$

$$\leq\varepsilon(n,\alpha,K,x_c)\times\mathrm{Vol}(\mathrm{Conv}(\mathcal{X})), \tag{110}$$

where $\mathrm{Vol}(\mathrm{Conv}(\mathcal{X}))$ is the volume of the convex hull of the support of the distribution of X. Since $x_c\in\mathcal{X}$, by definition of the support (see), we know that for all $\rho>0$, $\mu_X(B(x_c,\rho))>0$. Thus, $\beta_{x_c,\alpha}=\mu_X\left(B\left(x_c,\alpha\right)\right)>0$. Consequently,

$$\varepsilon(n,\alpha,K,x_c)=C_2\eta(\alpha,R)\exp\left[n\left(3\sqrt{\frac{e}{1-\beta_{x_c,\alpha}}}\sqrt{\frac{K}{n}}+\ln\left(1-\beta_{x_c,\alpha}\right)\right)\right] \tag{111}$$

tends to zero, as K/n tends to infinity. □

8.5 Proof of Corollary 3.3.1

Proof of Corollary 3.3.1. Let $x \in \overset{o}{\mathcal{X}}$. By assumption, according to Theorem 3.3, for all $n \geq K \geq 1$, for all $\alpha > 0$ such that $\alpha \leq 2R$,

$$\mathbb{P}(|Z - X_c| \geq \alpha | X_c = x_c) \leq \epsilon(n, \alpha, K, x_c) \times \text{Vol}(\text{Conv}(\mathcal{X})), \tag{112}$$

where $\text{Vol}(\text{Conv}(\mathcal{X}))$ is the volume of the convex hull of $\mathcal{X}$, and

$$\varepsilon(n, K, x_c, \alpha) = C_2 \eta(\alpha, R) \exp\left[n \left(3\sqrt{\frac{e}{1 - \beta_{x_c,\alpha}}} \sqrt{\frac{K}{n}} + \ln\left(1 - \beta_{x_c,\alpha}\right) \right) \right], \tag{113}$$

with $\beta_{x_c,\alpha} = \mu_X\left(B\left(x_c, \alpha\right)\right)$ and

$$\eta(\alpha, R) = \begin{cases} \ln\left(\frac{2R}{\alpha}\right) & \text{if } d = 1 \\ \frac{1}{d-1}\left(\left(\frac{2R}{\alpha}\right)^{d-1} - 1\right) & \text{otherwise.} \end{cases}$$

Let $\alpha_n = (K/n)^\gamma$, with $\gamma > 0$. By assumption, K/n tends to zero as n tends to infinity, so that $\lim_{n\to\infty} \alpha_n = 0$. Since $x_c \in \mathcal{X}$, for all n large enough, $B(x_c, \alpha_n) \subset \overset{o}{\mathcal{X}}$. Therefore, for all n large enough,

$$C_1 c_d \alpha_n^d \leq \beta_{x_c,\alpha_n} = \int_{x \in B(x_c,\alpha_n)} f_X(x) \mathrm{d}x \leq C_2 c_d \alpha_n^d, \tag{114}$$

with $c_d = \pi^{d/2} / \Gamma\left(\frac{d}{2} + 1\right)$. Besides, we have for all $x \in [-\infty, 1], \ln(1 - x) \leq -x$. Then, according to (113) and (114),

$$\varepsilon(n, K, x_c, \alpha_n) \leq C_2 \eta(\alpha_n, R) \exp\left[n \left(3\sqrt{\frac{e}{1 - C_1 c_d \alpha_n^d}} \sqrt{\frac{K}{n}} - C_1 c_d \alpha_n^d \right) \right]. \tag{115}$$

Case $d = 1$. If $d = 1$, then

$$\varepsilon(n, K, x_c, \alpha_n) \leq C_2 \ln\left(\frac{2R}{\alpha_n}\right) \exp\left[n \left(3\sqrt{\frac{e}{1 - C_1 c_1 \alpha_n^d}} \sqrt{\frac{K}{n}} - C_1 c_1 \alpha_n^d \right) \right] \tag{116}$$

$$\leq C_2 \ln\left(2R\left(\frac{n}{K}\right)^\gamma\right) \exp\left[n \left(3\sqrt{\frac{e}{1 - C_1 c_1 \left(\frac{K}{n}\right)^\gamma}} \sqrt{\frac{K}{n}} - C_1 c_1 \left(\frac{K}{n}\right)^\gamma \right) \right] \tag{117}$$

$$\leq C_2 \ln\left(2R\left(\frac{n}{K}\right)^\gamma\right) \exp\left[n\sqrt{\frac{K}{n}} \left(3\sqrt{\frac{e}{1 - C_1 c_1 \left(\frac{K}{n}\right)^\gamma}} - C_1 c_1 \left(\frac{K}{n}\right)^{\gamma - \frac{1}{2}} \right) \right]. \tag{118}$$

For this upper to tend to zero, we need to have $0 < \gamma < 1/2$. Note that

$$C_1 c_1 \left(\frac{K}{n}\right)^\gamma \leq \frac{1}{2}, \tag{119}$$

is equivalent to

$$\gamma \geq \frac{\ln(2 C_1 c_1)}{\ln\left(\frac{n}{K}\right)}. \tag{120}$$

Thus, assuming that

$$\max\left(0, \frac{\ln(2 C_1 c_1)}{\ln\left(\frac{n}{K}\right)}\right) < \gamma < 1/2, \tag{121}$$

we have

$$\varepsilon(n,K,x_c,\alpha_n) \leq C_2 \ln\left(2R\left(\frac{n}{K}\right)^{\gamma}\right) \exp\left[n\sqrt{\frac{K}{n}}\left(3\sqrt{2e} - C_1 c_d \left(\frac{K}{n}\right)^{\gamma}\right)\right]. \tag{122}$$

Additionally, assuming that

$$C_1 c_1 \left(\frac{K}{n}\right)^{\gamma-\frac{1}{2}} \geq 6\sqrt{2e}, \tag{123}$$

which is equivalent to

$$\gamma \leq \frac{1}{2} + \frac{\ln\left(6\sqrt{2e}/C_1 c_1\right)}{\ln(K/n)}, \tag{124}$$

leads to

$$\varepsilon(n,K,x_c,\alpha_n) \leq C_2 \ln\left(2R\left(\frac{n}{K}\right)^{\gamma}\right) \exp\left[-3\sqrt{2enK}\right]. \tag{125}$$

So far, we have assumed that

$$\max\left(0, \frac{\ln(2C_1c_1)}{\ln\left(\frac{n}{K}\right)}\right) < \gamma < \frac{1}{2} + \min\left(0, \frac{\ln\left(6\sqrt{2e}/C_1c_1\right)}{\ln(K/n)}\right). \tag{126}$$

Besides, we assume that $n/K \geq R$, which leads to

$$\varepsilon(n,K,x_c,\alpha_n) \leq (\gamma+1) C_2 \ln\left(\frac{n}{K}\right) \exp\left[-3\sqrt{2enK}\right]. \tag{127}$$

Letting $v = \sqrt{n/K}$, we have

$$\varepsilon(n,K,x_c,\alpha_n) \leq 2C_2 v \exp\left[-\frac{3\sqrt{2e}Kv}{2}\right] \exp\left[-\frac{3\sqrt{2e}Kv}{2}\right]. \tag{128}$$

Maximizing $f : v \mapsto 2A_1 v \exp\left[-A_2 v/2\right]$ over $v \in (0,\infty)$ with $A_1 = (\gamma+1)C_2$ and $A_2 = 3K\sqrt{2e}$ leads to

$$\varepsilon(n,K,x_c,\alpha_n) \leq \frac{4A_1}{eA_2} \exp\left[-\frac{3\sqrt{2e}Kv}{2}\right] \tag{129}$$

Thus, for all $n/K \geq R$ and all γ such that

$$\max\left(0, \frac{\ln(2C_1c_1)}{\ln\left(\frac{n}{K}\right)}\right) < \gamma < \frac{1}{2} + \min\left(0, \frac{\ln\left(6\sqrt{2e}/C_1c_1\right)}{\ln(K/n)}\right), \tag{130}$$

we have

$$\varepsilon(n,K,x_c,\alpha_n) \leq \frac{4C_2 e^{-1}(\gamma+1)}{3K\sqrt{2e}} \exp\left[-3\sqrt{\frac{enK}{2}}\right]. \tag{131}$$

Since $c_1 = 2$, for all $n/K \geq R$, and all γ such that

$$\max\left(0, \frac{\ln(4C_1)}{\ln\left(\frac{n}{K}\right)}\right) < \gamma < \frac{1}{2} + \min\left(0, \frac{\ln\left(3\sqrt{2e}/C_1\right)}{\ln(K/n)}\right), \tag{132}$$

that is, for all $R \leq n/K$.

Case $d > 1$. If $d > 1$, then

$$\varepsilon(n, K, x_c, \alpha_n) \tag{133}$$

$$\leq C_2 \frac{1}{d-1} \left(\left(\frac{2R}{\alpha_n} \right)^{d-1} - 1 \right) \exp \left[n \left(3 \sqrt{\frac{e}{1 - C_1 c_d \alpha_n^d}} \sqrt{\frac{K}{n}} - C_1 c_d \alpha_n^d \right) \right] \tag{134}$$

$$\leq C_2 \frac{1}{d-1} \left(\left(2R \left(\frac{n}{K} \right)^{\gamma} \right)^{d-1} - 1 \right) \exp \left[n \left(3 \sqrt{\frac{e}{1 - C_1 c_d \left(\frac{K}{n} \right)^{\gamma d}}} \sqrt{\frac{K}{n}} - C_1 c_d \left(\frac{K}{n} \right)^{\gamma d} \right) \right] \tag{135}$$

$$\leq C_2 \frac{1}{d-1} \left(2R \left(\frac{n}{K} \right)^{\gamma} \right)^{d-1} \exp \left[n \left(3 \sqrt{\frac{e}{1 - C_1 c_d \left(\frac{K}{n} \right)^{\gamma d}}} \sqrt{\frac{K}{n}} - C_1 c_d \left(\frac{K}{n} \right)^{\gamma d} \right) \right] \tag{136}$$

$$= C_2 \frac{1}{d-1} \exp \left[(d-1) \ln \left(2R \left(\frac{n}{K} \right)^{\gamma} \right) + n \left(3 \sqrt{\frac{e}{1 - C_1 c_d \left(\frac{K}{n} \right)^{\gamma d}}} \sqrt{\frac{K}{n}} - C_1 c_d \left(\frac{K}{n} \right)^{\gamma d} \right) \right] \tag{137}$$

Let $\varepsilon > 0$ and from now $\gamma d = 1/2 + \lambda$. Then,

$$\varepsilon(n, K, x_c, \alpha_n) \tag{138}$$

$$\leq C_2 \frac{1}{d-1} \exp \left[(d-1) \ln \left(2R \left(\frac{n}{K} \right)^{\frac{1}{d}\left(\frac{1}{2}+\lambda\right)} \right) + n \left(3 \sqrt{\frac{e}{1 - C_1 c_d \left(\frac{K}{n} \right)^{(1/2)+\lambda}}} \sqrt{\frac{K}{n}} - C_1 c_d \left(\frac{K}{n} \right)^{(1/2)+\lambda} \right) \right] \tag{139}$$

$$\leq C_2 \frac{1}{d-1} \exp \left[(d-1) \ln \left(2R \left(\frac{n}{K} \right)^{\frac{1}{d}\left(\frac{1}{2}+\lambda\right)} \right) + n \sqrt{\frac{K}{n}} \left(3 \sqrt{\frac{e}{1 - C_1 c_d \left(\frac{K}{n} \right)^{(1/2)+\lambda}}} - C_1 c_d \left(\frac{K}{n} \right)^{\lambda} \right) \right]. \tag{140}$$

For this upper to tend to zero, we need to have $\frac{-1}{2} < \lambda < 0$. Note that,

$$C_1 c_d \left(\frac{K}{n} \right)^{(1/2)+\lambda} \leq \frac{1}{2}, \tag{141}$$

is equivalent to

$$\lambda \geq \frac{\ln(2 C_1 c_d)}{\ln\left(\frac{n}{K}\right)} - \frac{1}{2}. \tag{142}$$

Thus, assuming that

$$\max \left(0, \frac{\ln(2 C_1 c_d)}{\ln\left(\frac{n}{K}\right)} \right) - \frac{1}{2} < \lambda < 0, \tag{143}$$

we have

$$\varepsilon(n, K, x_c, \alpha_n) \tag{144}$$

$$\leq C_2 \frac{1}{d-1} \exp \left[(d-1) \ln \left(2R \left(\frac{n}{K} \right)^{\frac{1}{d}\left(\frac{1}{2}+\lambda\right)} \right) + n \sqrt{\frac{K}{n}} \left(3\sqrt{2e} - C_1 c_d \left(\frac{K}{n} \right)^{\lambda} \right) \right]. \tag{145}$$

Additionally, assuming that

$$C_1 c_1 \left(\frac{K}{n} \right)^{\lambda} \geq 6\sqrt{2e}, \tag{146}$$

which is equivalent to

$$\lambda \leq \frac{\ln\left(6\sqrt{2e}/C_1 c_d\right)}{\ln(K/n)}, \tag{147}$$

leads to

$$\varepsilon(n,K,x_c,\alpha_n) \leq C_2 \frac{1}{d-1} \exp\left[(d-1)\ln\left(2R\left(\frac{n}{K}\right)^{\frac{1}{d}\left(\frac{1}{2}+\lambda\right)}\right) - n\sqrt{\frac{K}{n}}\left(3\sqrt{2e}\right)\right] \tag{148}$$

So far, we have assumed that

$$\max\left(0, \frac{\ln(2C_1c_d)}{\ln\left(\frac{n}{K}\right)}\right) - \frac{1}{2} < \lambda < \min\left(0, \frac{\ln\left(6\sqrt{2e}/C_1c_d\right)}{\ln(K/n)}\right). \tag{149}$$

Besides, we assume that $n/K \geq R$, which leads to

$$\varepsilon(n,K,x_c,\alpha_n) \leq C_2 \frac{1}{d-1} \exp\left[(d-1)\ln\left(\left(\frac{n}{K}\right)^{\frac{1}{d}\left(\frac{1}{2}+\lambda\right)+1}\right) - n\sqrt{\frac{K}{n}}\left(3\sqrt{2e}\right)\right] \tag{150}$$

$$= C_2 \frac{1}{d-1} \exp\left[(d-1)\left(\frac{1}{d}\left(\frac{1}{2}+\lambda\right)+1\right)\ln\left(\frac{n}{K}\right) - n\sqrt{\frac{K}{n}}\left(3\sqrt{2e}\right)\right] \tag{151}$$

Letting $v = \sqrt{n/K}$, we have

$$\varepsilon(n,K,x_c,\alpha_n) \leq C_2 \frac{1}{d-1} \exp\left[(d-1)\left(\frac{1}{d}\left(\frac{1}{2}+\lambda\right)+1\right)\ln\left(v^2\right) - vK\left(3\sqrt{2e}\right)\right]. \tag{152}$$

Maximizing $g : v \mapsto B_1 x^{\lceil 2B_2 \rceil} \exp\left[\frac{-3\sqrt{2e}Kx}{2}\right]$ with $B_1 = C_2 \frac{1}{d-1}$ and $B_2 = (d-1)\left(\frac{1}{d}\left(\frac{1}{2}+\lambda\right)+1\right)$ leads to

$$\varepsilon(n,K,x_c,\alpha_n) \leq B_1 \left(\frac{4B_2+2}{3\sqrt{2e}K}\right)^{2B_2+1} \exp\left[\frac{-3\sqrt{2enK}}{2}\right]. \tag{153}$$

Thus, for all $n/K \geq R$ and all γ such that

$$\max\left(0, \frac{\ln(2C_1c_d)}{\ln\left(\frac{n}{K}\right)}\right) - \frac{1}{2} < \lambda < \min\left(0, \frac{\ln\left(6\sqrt{2e}/C_1c_d\right)}{\ln(K/n)}\right), \tag{154}$$

which is equivalent to,

$$\frac{1}{d}\max\left(0, \frac{\ln(2C_1c_d)}{\ln\left(\frac{n}{K}\right)}\right) < \gamma < \frac{1}{d}\left(\frac{1}{2} + \min\left(0, \frac{\ln\left(6\sqrt{2e}/C_1c_d\right)}{\ln(K/n)}\right)\right). \tag{155}$$

Remark that, $\gamma < 1/2$ which implies,

$$B_2 = (d-1)(\gamma+1) \tag{156}$$

$$< (d-1)\left(\frac{1}{2}+\gamma\right) \tag{157}$$

$$\leq \frac{3}{2}(d-1) \tag{158}$$

$$\leq 2d. \tag{159}$$

On the other side, we have for all $d > 1 : 1/d - 1 \leq 2/d$, which leads to

$$B_1 = C_2 \frac{1}{d-1} \tag{160}$$

$$\leq \frac{2C_2}{d}. \tag{161}$$

Finally, for all $n/K \geq R$ and all γ such that

$$\frac{1}{d}\max\left(0, \frac{\ln(2C_1c_d)}{\ln\left(\frac{n}{K}\right)}\right) < \gamma < \frac{1}{d}\left(\frac{1}{2} + \min\left(0, \frac{\ln\left(6\sqrt{2e}/C_1c_d\right)}{\ln(K/n)}\right)\right), \tag{162}$$

we have,

$$\varepsilon(n, K, x_c, \alpha_n) \leq \frac{2C_2}{d}\left(\frac{8d+2}{3\sqrt{2e}K}\right)^{4d+1}\exp\left[\frac{-3\sqrt{2enK}}{2}\right] \tag{163}$$

□

8.6 Proof of Theorem 3.4

Proof of Theorem 3.4. Let $\varepsilon > 0$ and $z \in B(0,R)$ such that $\|z\| \geq R-\varepsilon$. Let $A_\varepsilon = \{x \in B(0,R), \langle x-z, z\rangle \leq 0\}$. Let $0 < \alpha < 2R$ and $\tilde{A}_{\alpha,\varepsilon} = A_\varepsilon \cap \{x, \|z-x\| \geq \alpha\}$. An illustration is displayed in Figure 5.

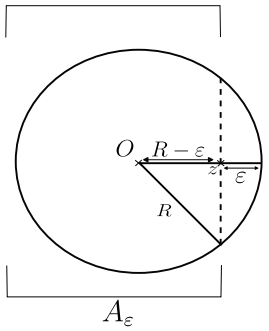

Figure 5: Illustration of Theorem 3.4.

We have

$$f_Z(z) = \int_{x_c \in \tilde{A}_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c)\mathrm{d}x_c + \int_{x_c \in \tilde{A}^c_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c)\mathrm{d}x_c \tag{164}$$

First term Let $x_c \in \tilde{A}_{\alpha,\varepsilon}$. In order to have $x_c + \frac{z-x_c}{w} = z + \left(-1+\frac{1}{w}\right)(z-x_c) \in B(0,R)$, it is necessary that

$$\left(-1+\frac{1}{w}\right)\|z-x_c\| \leq \sqrt{2\varepsilon R} \tag{165}$$

which leads to

$$w \geq \frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\|z-x_c\|}} \tag{166}$$

Since $x_c \in \tilde{A}_{\alpha,\varepsilon}$, we have $\|x_c - z\| \geq \alpha$. Thus, according to inequality (166), $x_c + \frac{z-x_c}{w} \in B(0,R)$ implies

$$w \geq \frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\alpha}}. \tag{167}$$

Recall that $x_c + \frac{z-x_c}{w} \in \mathcal{X}$. Consequently, according to Theorem 3.1, for all $x_c \in \tilde{A}_{\alpha,\varepsilon}$,

$$f_Z(z|X_c = x_c) \tag{168}$$

$$= (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)\mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w \tag{169}$$

$$\leq \frac{C_2(n-K-1)}{K}\binom{n-1}{K}\int_{\frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\alpha}}}^1 \frac{1}{w^d}\left[1-\mu_X\left(B\left(x_c, \frac{\alpha}{w}\right)\right)\right]^{n-K-2}\left[1-\left[\mu_X\left(B\left(x_c, \frac{\alpha}{w}\right)\right)\right]^K\right]\mathrm{d}w \tag{170}$$

$$\leq \frac{C_2 n}{K}\left(1-\frac{K-1}{n}\right)\binom{n-1}{K}\int_{\frac{1}{1+\frac{\sqrt{2\varepsilon R}}{\alpha}}}^1 \frac{1}{w^d}\mathrm{d}w \tag{171}$$

$$\leq \eta(\alpha, R)\frac{C_2 n}{K}\left(1-\frac{K-1}{n}\right)\binom{n-1}{K}, \tag{172}$$

with

$$\eta(\alpha, R) = \begin{cases} \ln\left(1+\frac{\sqrt{2\varepsilon R}}{\alpha}\right) & \text{if } d=1 \\ \frac{1}{d-1}\left(\left(1+\frac{\sqrt{2\varepsilon R}}{\alpha}\right)^{d-1}-1\right) & \text{otherwise} \end{cases}.$$

Second term According to Theorem 3.1, we have

$$f_Z(z|X_c = x_c) = (n-K-1)\binom{n-1}{K}\int_0^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right) \times \mathcal{B}\left(n-K-1, K; 1-\mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)\right)\mathrm{d}w. \tag{173}$$

Letting $\beta_{x_c,\alpha}(w) = \mu_X\left(B\left(x_c, \frac{||z-x_c||}{w}\right)\right)$, according to Lemma 3,

$$\mathcal{B}\left(n-K-1, K; 1-\beta_{x_c,\alpha}(w)\right) \leq \frac{(1-\beta_{x_c,\alpha}(w))^{n-K-2}}{K}\left(1-\beta_{x_c,\alpha}(w)^K\right) \tag{174}$$

$$\leq \frac{(1-\beta_{x_c,\alpha}(w))^{n-K-2}}{K}, \tag{175}$$

since $\beta_{x_c,\alpha}(w) \in [0,1]$. Note that, since $\mathcal{X} \subset B(0,R)$, all points $x, z \in \mathcal{X}$ satisfy $\|x-z\| \leq 2R$. Consequently, if $\|z-x_c\|/w \geq 2R$,

$$B(0,R) \subset B(x_c, \|z-x_c\|/w). \tag{176}$$

Hence, for all $w \leq \|z-x_c\|/2R$, $\beta_{x_c,\alpha}(w) = 1$. Plugging this equality into (173), we have

$$f_Z(z|X_c = x_c) \tag{177}$$

$$= \frac{(n-K-1)}{K}\binom{n-1}{K}\int_0^{\|z-x_c\|/2R} \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)(1-\beta_{x_c,\alpha}(w))^{n-K-2}\mathrm{d}w \tag{178}$$

$$+ \frac{(n-K-1)}{K}\binom{n-1}{K}\int_{\|z-x_c\|/2R}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)(1-\beta_{x_c,\alpha}(w))^{n-K-2}\mathrm{d}w \tag{179}$$

$$= \frac{(n-K-1)}{K}\binom{n-1}{K}\int_{\|z-x_c\|/2R}^1 \frac{1}{w^d} f_X\left(x_c + \frac{z-x_c}{w}\right)(1-\beta_{x_c,\alpha}(w))^{n-K-2}\mathrm{d}w \tag{180}$$

$$\leq C_2\frac{(n-K-1)}{K}\binom{n-1}{K}\int_{\|z-x_c\|/2R}^1 \frac{1}{w^d}\mathrm{d}w \tag{181}$$

$$\leq C_2\frac{(n-K-1)}{K}\binom{n-1}{K}\left[-\frac{1}{d-1}w^{-d+1}\right]_{\|z-x_c\|/2R}^1 \tag{182}$$

$$\leq \frac{C_2(2R)^{d-1}}{d-1}\frac{(n-K-1)}{K}\binom{n-1}{K}\frac{1}{\|z-x_c\|^{d-1}}. \tag{183}$$

Besides, note that, for all $\alpha > 0$, we have

$$\int_{B(z,\alpha)} \frac{1}{\|z - x_c\|^{d-1}} f_X(x_c) \mathrm{d}x_c \tag{184}$$

$$\leq C_2 \int_{B(0,\alpha)} \frac{1}{r^{d-1}} r^{d-1} \sin^{d-2}(\varphi_1) \sin^{d-3}(\varphi_2) \dots \sin(\varphi_{d-2}) \mathrm{d}r \mathrm{d}\varphi_1 \dots \mathrm{d}\varphi_{d-2}, \tag{185}$$

where $r, \varphi_1, \dots, \varphi_{d-2}$ are the spherical coordinates. A direct calculation leads to

$$\int_{B(z,\alpha)} \frac{1}{\|z - x_c\|^{d-1}} f_X(x_c) \mathrm{d}x_c \leq C_2 \int_0^\alpha \mathrm{d}r \int_{S(0,\alpha)} \sin^{d-2}(\varphi_1) \sin^{d-3}(\varphi_2) \dots \sin(\varphi_{d-2}) \mathrm{d}\varphi_1 \dots \mathrm{d}\varphi_{d-2} \tag{186}$$

$$\leq \frac{2C_2 \pi^{d/2}}{\Gamma(d/2)} \alpha, \tag{187}$$

as

$$\int_{S(0,\alpha)} \sin^{d-2}(\varphi_1) \sin^{d-3}(\varphi_2) \dots \sin(\varphi_{d-2}) \mathrm{d}\varphi_1 \dots \mathrm{d}\varphi_{d-2} \tag{188}$$

is the surface of the S^{d-1} sphere. Finally, for all $z \in \mathcal{X}$, for all $\alpha > 0$, and for all K, N such that $1 \leq K \leq N$, we have

$$\int_{B(z,\alpha)} f_Z(z|X_c = x_c) f_X(x_c) \mathrm{d}x_c \leq \frac{2C_2^2 (2R)^{d-1} \pi^{d/2}}{(d-1)\Gamma(d/2)} \frac{(n-K-1)}{K} \binom{n-1}{K} \alpha. \tag{189}$$

Final result Using Figure 5 and Pythagore's Theorem, we have $a^2 \leq \sqrt{2\varepsilon R}$. Let $d > 1$ and $\epsilon > 0$. Then we have for all α such that $\alpha > a$.

$$f_Z(z) \tag{190}$$

$$= \int_{x_c \in \tilde{A}_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c) \mathrm{d}x_c + \int_{x_c \in \tilde{A}^c_{\alpha,\varepsilon}} f_Z(z|X_c = x_c) f_X(x_c) \mathrm{d}x_c \tag{191}$$

$$\leq \frac{C_2}{d-1} \left(\left(1 + \frac{\sqrt{2\varepsilon R}}{\alpha} \right)^{d-1} - 1 \right) \frac{(n-K-1)}{K} \binom{n-1}{K} + \frac{2C_2^2 (2R)^{d-1} \pi^{d/2}}{(d-1)\Gamma(d/2)} \frac{(n-K-1)}{K} \binom{n-1}{K} \alpha \tag{192}$$

$$= \frac{C_2}{d-1} \frac{(n-K-1)}{K} \binom{n-1}{K} \left[\left(\left(1 + \frac{\sqrt{2\varepsilon R}}{\alpha} \right)^{d-1} - 1 \right) + \frac{2C_2 (2R)^{d-1} \pi^{d/2}}{\Gamma(d/2)} \alpha \right], \tag{193}$$

But this inequality is true if $\alpha \geq a$. We know that $(1+x)^{d-1} \leq (2^{d-1} - 1)x + 1$ for $x \in [0, 1]$ and $d - 1 \geq 0$. Then, for α such that $\frac{\sqrt{2\varepsilon R}}{\alpha} \leq 1$,

$$f_Z(z) \tag{194}$$

$$\leq \frac{C_2}{d-1} \frac{(n-K-1)}{K} \binom{n-1}{K} \left[\left(\left((2^{d-1} - 1) \frac{\sqrt{2\varepsilon R}}{\alpha} + 1 \right) - 1 \right) + \frac{2C_2 (2R)^{d-1} \pi^{d/2}}{\Gamma(d/2)} \alpha \right] \tag{195}$$

$$\leq \frac{C_2}{d-1} \frac{(n-K-1)}{K} \binom{n-1}{K} \left[\left((2^{d-1} - 1) \frac{\sqrt{2\varepsilon R}}{\alpha} \right) + \frac{2C_2 (2R)^{d-1} \pi^{d/2}}{\Gamma(d/2)} \alpha \right]. \tag{196}$$

Since $\frac{\sqrt{2\varepsilon R}}{\alpha} \leq 1$, then $\alpha \geq \sqrt{2\varepsilon R} \geq a$. So our initial condition on α to get the upper bound of the second term is still true. Now, we choose α such that,

$$(2^{d-1} - 1) \frac{\sqrt{2\varepsilon R}}{\alpha} \leq \frac{2C_2 (2R)^{d-1} \pi^{d/2}}{\Gamma(d/2)} \alpha, \tag{197}$$

which leads to the following condition

$$\alpha \geq \left(\frac{\Gamma(d/2)(2^{d-1} - 1)\sqrt{2\varepsilon R}}{2C_2 (2R)^{d-1} \pi^{d/2}} \right)^{1/2}. \tag{198}$$

Finally, for

$$\alpha = \left(\frac{\Gamma(d/2)(2^{d-1}-1)\sqrt{2\varepsilon R}}{2C_2(2R)^{d-1}\pi^{d/2}} \right)^{1/2}, \tag{199}$$

we have,

$$f_Z(z) \leq \frac{C_2}{d-1} \frac{(n-K-1)}{K} \binom{n-1}{K} \left[2 \frac{2C_2(2R)^{d-1}\pi^{d/2}}{\Gamma(d/2)} \alpha \right] \tag{200}$$

$$= 2 \frac{C_2}{d-1} \frac{(n-K-1)}{K} \binom{n-1}{K} \left(\frac{\Gamma(d/2)(2^{d-1}-1)\sqrt{2\varepsilon R}}{2C_2(2R)^{d-1}\pi^{d/2}} \right)^{1/2}. \tag{201}$$

□

8.7 Technical lemmas

8.7.1 Cumulative distribution function of a binomial law

Lemma 2 (Cumulative distribution function of a binomial distribution)**.** *Let X be a random variable following a binomial law of parameter $n \in \mathbf{N}$ and $p \in [0,1]$. The cumulative distribution function F of X can be expressed as Wadsworth et al. (1961):*

(i)

$$F(k;n,p) = \mathbb{P}(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i},$$

(ii)

$$F(k;n,p) = (n-k)\binom{n}{k} \int_0^{1-p} t^{n-k-1}(1-t)^k \mathrm{d}t$$

$$= (n-k)\binom{n}{k} \mathcal{B}(n-k, k+1; 1-p),$$

with $\mathcal{B}(a,b;x) = \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t$, the incomplete beta function.

Proof. see Wadsworth et al. (1961). □

8.7.2 Upper bounds for the incomplete beta function

Lemma 3. *Let $B(a,b;x) = \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t$, be the incomplete beta function. Then we have*

$$\frac{x^a}{a} \leq B(a,b;x) \leq x^{a-1}\left(\frac{1-(1-x)^b}{b} \right),$$

for $a > 0$.

Proof. We have

$$\begin{aligned} B(a,b;x) &= \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t \\ &\leq \int_{t=0}^{x} x^{a-1}(1-t)^{b-1}\mathrm{d}t \\ &= x^{a-1}\int_{t=0}^{x}(1-t)^{b-1}\mathrm{d}t \\ &= x^{a-1}\left[(-1)\frac{(1-t)^b}{b}\right]_0^x \\ &= x^{a-1}\left[-\frac{(1-x)^b}{b}+\frac{1}{b}\right] \\ &= x^{a-1}\frac{1-(1-x)^b}{b}. \end{aligned}$$

On the other hand,

$$\begin{aligned} B(a,b;x) &= \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\mathrm{d}t \\ &\geq \int_{t=0}^{t} x^{a-1}\mathrm{d}t \\ &= \left[\frac{t^a}{a}\right]_0^x \\ &= \frac{x^a}{a}-\frac{0^a}{a} \\ &= \frac{x^a}{a}. \end{aligned}$$

□

8.7.3 Upper bounds for binomial coefficient

Lemma 4. *For $k, n \in \mathbb{N}$ such that $k < n$, we have*

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k. \tag{202}$$

Proof. We have,

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k!} \leq \frac{n^k}{k!}. \tag{203}$$

Besides,

$$e^k = \sum_{i=0}^{+\infty}\frac{k^i}{i!} \Longrightarrow e^k \geq \frac{k^k}{k!} \Longrightarrow \frac{e^k}{k^k} \geq \frac{1}{k!}. \tag{204}$$

Hence,

$$\binom{n}{k} = \frac{n(n-1)\dots(n-k+1)}{k!} \leq \frac{n^k}{k!} \leq \left(\frac{en}{k}\right)^k. \tag{205}$$

□

8.7.4 Inequality $x \ln\left(\frac{1}{x}\right) \leq \sqrt{x}$

Lemma 5. *For $x \in]0, +\infty[$,*

$$x \ln\left(\frac{1}{x}\right) \leq \sqrt{x}. \tag{206}$$

Proof. Let,

$$f(x) = \sqrt{x} - x \ln\left(\frac{1}{x}\right) \tag{207}$$

$$= \sqrt{x} + x \ln(x). \tag{208}$$

Then,

$$f^{'}(x) = \frac{1}{2\sqrt{x}} + \ln x + 1. \tag{209}$$

And,

$$f^{''}(x) = \frac{1}{x} - \frac{1}{4x^{3/2}}. \tag{210}$$

We have,

$$\begin{aligned} f^{''}(x) \geq 0 &\Longrightarrow \frac{1}{x} - \frac{1}{4x^{3/2}} \geq 0 \\ &\Longrightarrow \frac{1}{x} \geq \frac{1}{4x^{3/2}} \end{aligned} \tag{211}$$

Since $x \in]0, +\infty[$,

$$Equation\ (211) \Longrightarrow \frac{x^{3/2}}{x} \geq \frac{1}{4} \tag{212}$$

$$\Longrightarrow \sqrt{x} \geq \frac{1}{4} \tag{213}$$

$$\Longrightarrow x \geq \frac{1}{16}. \tag{214}$$

This result leads to,

x	0		$\frac{1}{16}$		$+\infty$
f''		$-$	0	$+$	
f'	$\searrow$		$2 + \ln(\frac{1}{16}) + 1$		$\nearrow$

(215)

We have $2 + \ln(\frac{1}{16}) + 1 > 0$. So $f^{'}(x) > 0$ for all $x \in]0, \infty[$. Furthermore $\lim_{x \to 0^+} f(x) = 0$, hence $f(x) > 0$ for all $x \in]0, \infty[$, therefore $\sqrt{x} > x \ln\left(\frac{1}{x}\right)$ for all $x \in]0, \infty[$.

□