Gulf Journal of Mathematics
Vol 17, Issue 2 (2024) 327-336

ISSN: 2309-4966
https://doi.org/10.56947/gjom.v17i2.2343

ASYMPTOTIC BEHAVIOR OF SMOTE-GENERATED SAMPLES USING ORDER STATISTICS

FIRUZ KAMALOV[1*]

Abstract. Imbalanced datasets often lead to biased machine learning models that underperform on minority classes. The Synthetic Minority Over-sampling Technique (SMOTE) addresses this by generating synthetic samples for the minority class. Despite its empirical success, the theoretical properties of SMOTE remain underexplored. This paper investigates the asymptotic behavior of SMOTE-generated random variables using order statistics. We establish that, as the sample size n increases, the SMOTE-generated variable Z conditioned on the k-th order statistic $X_{(k)}$ converges in mean to $X_{(k)}$. Additionally, the expected value of Z converges to the expected value of the original random variable X as n approaches infinity. The results are derived under the assumption of left-bounded support for X. Our findings provide a theoretical foundation for SMOTE, demonstrating that it preserves key statistical properties of the original data distribution in large samples.

1. Introduction

Imbalanced datasets pose significant challenges in machine learning, particularly in classification tasks where one class significantly outnumbers others. Such imbalance can lead to biased predictive models that perform poorly on minority classes. The Synthetic Minority Over-sampling Technique (SMOTE) has emerged as a popular method to address this issue by generating synthetic samples of the minority class [3]. Despite its widespread application, the theoretical underpinnings of SMOTE remain less explored, especially concerning the statistical properties of the generated synthetic samples.

This paper delves into the theoretical aspects of the SMOTE-generated random variable. We aim to understand how the synthetic data points relate to the original data distribution, particularly as the sample size increases. Our investigation is grounded in the theory of order statistics, which provides a framework for analyzing the properties of sample minima, maxima, and intermediate order statistics derived from random samples.

We establish two principal results in this study:

(1) **Convergence of SMOTE Samples to Order Statistics**: We prove that the conditional expectation of a SMOTE-generated random variable

Date: Received: Aug 11, 2024; Accepted: Sep 21, 2024.

* Corresponding author.

2020 *Mathematics Subject Classification.* Primary 62D05; Secondary 62B86, 62B15, 62G30.
Key words and phrases. SMOTE, order statistics, sampling, convergence, imbalanced data.

Z given the k-th order statistic $X_{(k)}$ converges to $X_{(k)}$ in mean as the sample size n approaches infinity. Formally,

$$\lim_{n\to\infty} \mathbb{E}[|Z \mid X_{(k)} - X_{(k)}|] = 0.$$

It implies that, asymptotically, the synthetic samples generated via SMOTE become increasingly similar to the original data points from which they are derived.

(2) **Convergence of Expected SMOTE Samples to Population Mean**: We demonstrate that the expected value of the SMOTE-generated random variable Z converges to the expected value of the original random variable X as $n \to \infty$:

$$\lim_{n\to\infty} \mathbb{E}[Z] = \mathbb{E}[X].$$

The result suggests that, in large samples, the SMOTE augmentation does not distort the central tendency of the data distribution.

To establish the main results, we first develop key lemmas and theorems in order statistics. The left-boundedness of the support of X is a crucial condition in our analysis. We provide examples demonstrating that the convergence results may not hold if this condition is relaxed, highlighting the importance of the underlying distribution's characteristics in the behavior of SMOTE-generated samples.

Our theoretical contributions provide a deeper understanding of the statistical properties of SMOTE and its impact on data augmentation. By showing that SMOTE preserves certain asymptotic properties of the original data distribution, we offer theoretical support for its continued use in addressing class imbalance. Furthermore, our findings lay the groundwork for future research into the theoretical aspects of data augmentation techniques in machine learning.

2. SMOTE

In this section, we provide a detailed description of the Synthetic Minority Over-sampling Technique (SMOTE) and discuss the existing literature related to its theoretical aspects.

SMOTE is an over-sampling method introduced by Chawla et al. [3] to address the problem of imbalanced datasets in classification tasks. The key idea behind SMOTE is to generate synthetic samples of the minority class by interpolating between existing minority class samples, thereby creating new, plausible instances that contribute to balancing the class distribution.

2.1. **SMOTE Algorithm.**

Given a dataset with a minority class sample set $S = \{x_1, x_2, \ldots, x_n\}$, where each x_i is a feature vector in $\mathbb{R}^d$, the SMOTE algorithm proceeds as follows:

This process effectively increases the number of minority class samples by creating synthetic examples that are linear combinations of existing samples and their neighbors. By introducing new samples along the line segments connecting minority class samples, SMOTE helps the learning algorithm to better define the decision boundary between classes.

Algorithm 1 SMOTE Algorithm [3]

repeat
  Select a minority sample x_0 at random
  Identify the K-nearest neighbors of x_0 among the minority samples
  Randomly choose one of the K-nearest neighbors of x_0, denote it as x_k
  Generate a synthetic sample z by interpolating between x_0 and x_k:

$$z = x_0 + \theta(x_k - x_0), \quad \theta \sim \text{Uniform}(0, 1)$$

until M synthetic samples are generated

In our analysis, we consider a simplified version of SMOTE to facilitate mathematical tractability. Specifically, given a random sample $S = \{x_1, x_2, \ldots, x_n\}$, we randomly choose a point x_i in S. Then, for a fixed integer m, we select only the m-th nearest neighbor of x_i in S, denoted by $x_{(i,m)}$. Finally, we generate a synthetic sample z_m by randomly interpolating between x_i and $x_{(i,m)}$:

$$z_m = x_i + \theta(x_{(i,m)} - x_i), \quad \theta \sim \text{Uniform}(0, 1).$$

2.2. **Existing Literature on Theoretical Aspects of SMOTE.** Despite its widespread use in practice, the theoretical properties of SMOTE have been relatively underexplored. Most of the existing literature focuses on empirical evaluations and adaptations of SMOTE to specific domains or types of data [1].

Previous studies have investigated the impact of SMOTE on classifier performance and have proposed variations to improve its effectiveness. For instance, Han et al. [5] introduced Borderline-SMOTE, which generates synthetic samples only in the region near the decision boundary. He et al. [6] proposed ADASYN which is a similar method to SMOTE but it generates different number of samples depending on an estimate of the local distribution of the class to be oversampled. More recently, Kamalov & Denisov proposed an alternative to SMOTE based on the gamma probability distribution [8]. Several sampling methods using alternative techniques have been proposed to overcome the limitations of SMOTE including kernel density estimation [9], generative adversarial networks [12], variational autoencoder [7], and others. However, these studies primarily focus on empirical performance and do not delve deeply into the theoretical aspects of SMOTE. There is a notable gap in the literature regarding a rigorous mathematical analysis of how SMOTE affects the statistical properties of the data distribution, particularly in terms of convergence and asymptotic behavior.

While SMOTE has been thoroughly studied empirically, its theoretical aspects have received little attention in the literature. In [4], the authors describe the sampling distribution of SMOTE in the form of a double integral function:

$$f_Z(z) = (N-K)\binom{N-1}{K}\int_x f_X(x)\int_{r=||z-x||}^{\infty} f_X\left(x + \frac{(z-x)r}{||z-x||}\right)\left(\frac{r^{d-2}}{||z-x||^{d-1}}\right) \times B\left(1 - I_{B(x,r);N-K-1,K}\right) dr dx, \tag{2.1}$$

where $B\left(1 - I_{B(x,r);N-K-1,K}\right)$ is the incomplete beta function. Although the above result provides a general formulation of the sampling distribution of SMOTE, it is limited due to the complex integral term. In [10], the authors provide an alternative derivation of the result in Equation 2.1. The authors prove several other related results, albeit under a key condition of bounded support of X.

Our work aims to fill this gap by providing a theoretical analysis of SMOTE-generated random variables using the framework of order statistics. We investigate how the synthetic samples relate to the original data distribution as the sample size increases, and we establish convergence results that demonstrate the preservation of key statistical properties in large samples.

3. Main results

In this section, we present our main theoretical findings regarding the asymptotic behavior of SMOTE-generated random variables. Our analysis is grounded in the theory of order statistics, which plays a crucial role in understanding the distributional properties of sample data.

Order statistics are essential in our context because the SMOTE algorithm relies on selecting nearest neighbors, which inherently involves ordering the data points based on their distances. By studying the properties of order statistics, we can analyze how the distances between data points behave as the sample size increases, and consequently, how the SMOTE-generated samples relate to the original data distribution.

We begin by establishing key lemmas and theorems concerning the expected differences between order statistics. These results provide the foundation for quantifying the expected gaps between neighboring data points in an ordered sample. To proceed further, we require the following lemma, which relates to the expected difference of consecutive order statistics.

Lemma 3.1. [11] *Let X be a random variable with cumulative distribution function $F(x)$. Let $X_{(k)}$ denote the k^{th} order statistic in a random sample of size n. Then,*

$$\mathbb{E}\left(X_{(k+1)} - X_{(k)}\right) = \binom{n}{k} \int_{-\infty}^{\infty} F^k(x)(1 - F(x))^{n-k}\,\mathrm{d}x,$$

for all $k = 1, \ldots, n-1$.

The next result provides an upper bound for the integrand in Lemma 3.1.

Lemma 3.2. *Let $F(x)$ be the CDF of a random variable X and $1 \le k \le n$. Then*

$$\|F^k(x)(1 - F(x))^n\|_\infty = \left(\frac{k}{k+n}\right)^k \left(1 - \frac{k}{k+n}\right)^n.$$

Proof. Consider the derivative:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}F^k(x)(1 - F(x))^n &= kF^{k-1}(x) \cdot f(x) \cdot (1 - F(x))^n + F^k(x) \cdot n(1 - F(x))^{n-1} \cdot (-f(x)) \\ &= F^{k-1}(x) \cdot f(x) \cdot (1 - F(x))^{n-1}[k - (k+n)F(x)].\end{aligned} \tag{3.1}$$

By setting $k-(k+n)F(x)=0$, we see that the function $F^k(x)(1-F(x))^n$ attains its maximum value at the critical point $F^{-1}(\frac{k}{k+n})$. It follows that

$$\max F^k(x)(1-F(x))^n = \left(\frac{k}{k+n}\right)^k \left(1-\frac{k}{k+n}\right)^n.$$

□

Our goal is to consider the expected difference of two consecutive order statistics for a random variable with a bounded support on the left. We call a random variable X *left-bounded* if there exists a finite value $-\infty < L < \infty$ such that $F(x)=0$ for all, $x < L$.

We need the following observation before we proceed

$$\int_0^\infty (1-F(x))\,\mathrm{d}x = \int_0^\infty x\,f(x)\,\mathrm{d}x \le E(|X|) < \infty. \tag{3.2}$$

The above equation implies that

$$\int_L^\infty (1-F(x))\,\mathrm{d}x < \infty, \tag{3.3}$$

for any $-\infty < L < \infty$.

Theorem 3.3. *Let X be a random variable with left-bounded support. Let $S=\{x_1,x_2,\ldots,x_n\}$ be a random sample of X. Denote $X_{(k)}$ to be the k^{th} order statistic. Then*

$$\lim_{n\to\infty}\mathbb{E}\left[X_{(k+1)}-X_{(k)}\right]=0.$$

Proof. Let $\epsilon>0$ be given. Then, by Equation 3.3 there exists a compact set I such that

$$f(x)>0, \text{for all } x\in I$$

and

$$\int_{I^c}(1-F(x))\,\mathrm{d}x<\epsilon.$$

By Lemma 3.2, we have

$$\begin{aligned}\int_{I^c} F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x &= \int_{I^c}[1-F(x)]\cdot\left[F^k(x)(1-F(x))^{n-k-1}\right]\,\mathrm{d}x\\ &\le \|F^k(x)(1-F(x))^{n-k-1}\|_\infty \int_{I^c}[1-F(x)]\,\mathrm{d}x\\ &\le \left(\frac{k}{n-1}\right)^k\left(1-\frac{k}{n-1}\right)^{n-1}\left(1-\frac{k}{n-1}\right)^{-k}\cdot\epsilon\end{aligned} \tag{3.4}$$

It follows from Equation 3.4 that

$$\lim_{n\to\infty}\binom{n}{k}\int_{I^c}F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x \le C_k\cdot\epsilon, \tag{3.5}$$

where $C_k=\left(\frac{k}{e}\right)^k\cdot k!$. Next, we will consider $\int_I F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x$. Since $f(x)>0$ on the compact set I, there exists $\delta>0$ such that $f(x)>\delta$ for all $x\in I$.

Furthermore, assume without the loss of generality that $\sup_{x\in I}(1-F(x)) < 1$. Then, there exists $N > 0$ such that $(1-F(x))^N \leq \delta$ for all $x \in I$. Then

$$\begin{aligned}\int_I F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x &= \int_I \left[F^k(x)(1-F(x))^{n-k-N}\right]\cdot(1-F(x))^N\,\mathrm{d}x\\ &\leq \int_I \left[F^k(x)(1-F(x))^{n-k-N}\right]\cdot f(x)\,\mathrm{d}x\\ &\leq \int_0^1 u^k(1-u)^{n-k-N}\,\mathrm{d}u\\ &= \frac{k!(n-k-N)!}{(n-N+1)!},\end{aligned} \tag{3.6}$$

where the last equality is based on the Beta function. It follows from Equation 3.6 that

$$\lim_{n\to\infty}\binom{n}{k}\int_I F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x = 0. \tag{3.7}$$

Combining Equations 3.5 and 3.7 yields:

$$\lim_{n\to\infty}\binom{n}{k}\int_{-\infty}^{\infty} F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x = 0.$$

□

The result in Theorem 3.3 considers the expected difference between a pair of consecutive order statistics. This result can be extended to the difference between any fixed pair of order statistics as shown in the following corollary.

Corollary 3.4. *In the same scenario as in Theorem 3.3,*

$$\lim_{n\to\infty}\mathbb{E}\left[X_{(k+m)} - X_{(k)}\right] = 0,$$

where m is a fixed positive integer.

Proof. The case $m = 1$ is considered in Theorem 3.3. Assume $m \geq 2$, then

$$\mathbb{E}\left[X_{(k+m)} - X_{(k)}\right] = \sum_{j=2}^{m}\mathbb{E}\left[X_{(k+j)} - X_{(k+j-1)}\right] + \mathbb{E}\left[X_{(k+j-1)} - X_{(k+j-2)}\right].$$

The result follows by applying Theorem 3.3.

□

Let $S = \{x_1, x_2, \ldots, x_n\}$ be a random sample of X. Let Z_m denote the random variable that is generated as a random linear interpolation between x_i and its mth nearest neighbor in S . In other words, Z_m represents the SMOTE point between x_i and its mth nearest neighbor in the minority class sample. We will show that given $X_{(k)}$, Z_m converges to $X_{(k)}$ in mean.

Corollary 3.5. *In the same scenario as in Theorem 3.3,* $\lim_{n\to\infty}\mathbb{E}\left[|Z_m|X_{(k)} - X_{(k)}|\right] = 0$.

Proof. Assume, without the loss of generality, that $m < k < n - m$. Then

$$\mathbb{E}\left[|Z_m|X_{(k)} - X_{(k)}|\right] \leq \max\left\{\mathbb{E}\left[X_{(k+m)} - X_{(k)}\right], \mathbb{E}\left[X_{(k)} - X_{(k-m)}\right]\right\}$$

It follows from Corollary 3.4 that $\lim_{n\to\infty}\mathbb{E}\left[Z_m|X_{(k)} - X_{(k)}\right] = 0$. □

The left bound condition in Theorem 3.3 is critical. The next example illustrates that $\lim_{n\to\infty}\mathbb{E}\left(X_{(k+1)} - X_{(k)}\right)$ does not necessarily converge to zero if the support of the random variable is not bounded on the left. Consider the CDF $F(x) = e^x, -\infty < x \leq 0$. Then

$$\begin{aligned}\binom{n}{k}\int_{-\infty}^{0} F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x &= \binom{n}{k}\int_{-\infty}^{0}(e^x)^k(1-e^x)^{n-k}\,\mathrm{d}x \\ &= \binom{n}{k}\int_0^1 u^{k-1}(1-u)^{n-k}\,\mathrm{d}x \\ &= \binom{n}{k}\frac{(k-1)!(n-k)!}{n!} \\ &= \frac{1}{k}.\end{aligned} \tag{3.8}$$

Our next result shows that the expected value of Z converges to that of X unconditionally. We begin with the following auxiliary lemma.

Lemma 3.6. *Let $S = \{x_1, x_2, \ldots, x_n\}$ be an i.i.d. sample of a random variable X. Denote $X_{(k)}$ to be the k^{th} order statistic. Then,*

$$\lim_{n\to\infty}\frac{\mathbb{E}\left[X_{(k)}\right]}{n} = 0.$$

Proof. The result follows from the fact that $\mathbb{E}\left[X_{(n)}\right] \leq \mu + \sigma\sqrt{n-1}$ [2]. □

We are now in position to prove our next major result.

Theorem 3.7. *In the same scenario as in Lemma 3.6,* $\lim_{n\to\infty}\mathbb{E}\left[Z_m\right] = \mathbb{E}\left[X\right]$.

Proof. Let $Z_m|X_{(k)} = (1-\theta)X_{(k)} + \theta X_{(k,m)}$, where $X_{(k,m)}$ denotes the mth nearest neighbor of $X_{(k)}$. Note that $X_{(k,m)} \leq X_{(k+m)}$, for all $k \leq n - m$. Then

$$\begin{aligned}\mathbb{E}\left[Z_m|X_{(k)}\right] &\leq \mathbb{E}\left[(1-\theta)X_{(k)} + \theta X_{(k+m)}\right] \\ &= \frac{\mathbb{E}\left[X_{(k)}\right] + \mathbb{E}\left[X_{(k+m)}\right]}{2},\end{aligned} \tag{3.9}$$

for all $k \leq n - m$. Similarly, note that $X_{(k,m)} \leq X_{(n)}$, for all $n - m < k < n$. Then

$$\begin{aligned}\mathbb{E}\left[Z_m|X_{(k)}\right] &\leq \mathbb{E}\left[(1-\theta)X_{(k)} + \theta X_{(n)}\right] \\ &= \frac{\mathbb{E}\left[X_{(k)}\right] + \mathbb{E}\left[X_{(n)}\right]}{2}.\end{aligned} \tag{3.10}$$

for all $n - m < k < n$. Finally, note that $X_{(k,m)} \leq X_{(n-1)}$, for $k = n$.

To prove the main result we establish upper and lower bounds for $\mathbb{E}[Z_m]$. We begin by establishing the upper bound using Equations 3.9 and 3.10:

$$\begin{aligned}
\mathbb{E}[Z_m] &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[Z_m|X_{(k)}\right] \\
&\leq \frac{1}{n}\left[\sum_{k=1}^{n-m}\frac{\mathbb{E}\left[X_{(k)}\right]+\mathbb{E}\left[X_{(k+m)}\right]}{2}+\sum_{k=n-m+1}^{n-1}\frac{\mathbb{E}\left[X_{(k)}\right]+\mathbb{E}\left[X_{(n)}\right]}{2}\right]+\frac{\mathbb{E}\left[X_{(n)}\right]+\mathbb{E}\left[X_{(n-1)}\right]}{2} \\
&= \frac{1}{2n}\left[\sum_{k=1}^{m}\mathbb{E}\left[X_{(k)}\right]+\sum_{k=m+1}^{n}2\mathbb{E}\left[X_{(k)}\right]+(m-1)\mathbb{E}\left[X_{(n)}\right]+\mathbb{E}\left[X_{(n-1)}\right]\right] \\
&= \frac{1}{2n}\left[\sum_{k=1}^{n}2\mathbb{E}\left[X_{(k)}\right]-\sum_{k=1}^{m}\mathbb{E}\left[X_{(k)}\right]+(m-1)\mathbb{E}\left[X_{(n)}\right]+\mathbb{E}\left[X_{(n-1)}\right]\right] \\
&\leq \frac{1}{2n}\left[\sum_{k=1}^{n}2\mathbb{E}\left[X_{(k)}\right]+m\left(\mathbb{E}\left[X_{(n)}\right]-\mathbb{E}\left[X_{(1)}\right]\right)\right] \\
&= \mathbb{E}\left[\bar{X}\right]+\frac{m\left(\mathbb{E}\left[X_{(n)}\right]-\mathbb{E}\left[X_{(1)}\right]\right)}{2n}
\end{aligned}$$

Similarly, we need the following inequalities to establish the lower bound for $\mathbb{E}[Z_m]$:

$$\mathbb{E}\left[Z_m|X_{(k)}\right] \geq \frac{\mathbb{E}\left[X_{(k)}\right]+\mathbb{E}\left[X_{(k-m)}\right]}{2},\ m<k\leq n \tag{3.11}$$

and

$$\mathbb{E}\left[Z_m|X_{(k)}\right] \geq \frac{\mathbb{E}\left[X_{(k)}\right]+\mathbb{E}\left[X_{(1)}\right]}{2},\ 1<k\leq m. \tag{3.12}$$

Finally, note that $X_{(k,m)} \geq X_{(2)}$, for $k=1$.

We establish the lower bound using Equations 3.11 and 3.12:

$$\begin{aligned}
\mathbb{E}[Z_m] &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[Z_m|X_{(k)}\right] \\
&\geq \frac{1}{n}\left[\sum_{k=m+1}^{n}\frac{\mathbb{E}\left[X_{(k)}\right]+\mathbb{E}\left[X_{(k-m)}\right]}{2}+\sum_{k=2}^{m}\frac{\mathbb{E}\left[X_{(k)}\right]+\mathbb{E}\left[X_{(1)}\right]}{2}\right]+\frac{\mathbb{E}\left[X_{(1)}\right]+\mathbb{E}\left[X_{(2)}\right]}{2} \\
&= \frac{1}{2n}\left[\sum_{k=n-m+1}^{n}\mathbb{E}\left[X_{(k)}\right]+\sum_{k=1}^{n-m}2\mathbb{E}\left[X_{(k)}\right]+(m-1)\mathbb{E}\left[X_{(1)}\right]+\mathbb{E}\left[X_{(2)}\right]\right] \\
&= \frac{1}{2n}\left[\sum_{k=1}^{n}2\mathbb{E}\left[X_{(k)}\right]-\sum_{k=n-m+1}^{n}\mathbb{E}\left[X_{(k)}\right]+(m-1)\mathbb{E}\left[X_{(1)}\right]+\mathbb{E}\left[X_{(2)}\right]\right] \\
&\geq \frac{1}{2n}\left[\sum_{k=1}^{n}2\mathbb{E}\left[X_{(k)}\right]-m\left(\mathbb{E}\left[X_{(n)}\right]-\mathbb{E}\left[X_{(1)}\right]\right)\right] \\
&= \mathbb{E}\left[\bar{X}\right]-\frac{m\left(\mathbb{E}\left[X_{(n)}\right]-\mathbb{E}\left[X_{(1)}\right]\right)}{2n}
\end{aligned}$$

We obtain

$$\mathbb{E}\left[\bar{X}\right] - \frac{m\left(\mathbb{E}\left[X_{(n)}\right] - \mathbb{E}\left[X_{(1)}\right]\right)}{2n} \leq \mathbb{E}\left[Z_m\right] \leq \mathbb{E}\left[\bar{X}\right] + \frac{m\left(\mathbb{E}\left[X_{(n)}\right] - \mathbb{E}\left[X_{(1)}\right]\right)}{2n}$$

It follows from Lemma 3.6, that $\lim_{n\to\infty}\mathbb{E}\left[Z_m\right] = \lim_{n\to\infty}\mathbb{E}\left[\bar{X}\right] = \mathbb{E}\left[X\right]$. □

4. Conclusion

In this study, we conducted a theoretical examination of the Synthetic Minority Over-sampling Technique (SMOTE) by analyzing the asymptotic behavior of SMOTE-generated random variables through the lens of order statistics. Our investigation established two pivotal results: first, the conditional expectation of a SMOTE-generated random variable Z given the k-th order statistic $X_{(k)}$ converges in mean to $X_{(k)}$ as the sample size n approaches infinity, indicating that synthetic samples increasingly mirror the original minority class samples. Second, we demonstrated that the expected value of Z converges to the expected value of the original random variable X as n tends to infinity, thereby ensuring that SMOTE preserves the central tendency of the minority class distribution in large samples.

These theoretical insights provide a foundational understanding of SMOTE's effectiveness in addressing class imbalance, corroborating its empirical success in enhancing machine learning model performance. However, our analysis is contingent upon the assumption of left-bounded support for the original data distribution, highlighting a limitation that warrants further exploration. Future research should aim to extend this framework to accommodate distributions without left-bounded support and investigate the behavior of SMOTE in higher-dimensional feature spaces. Additionally, empirical validations across diverse datasets would reinforce the practical applicability of our theoretical findings. Overall, this study contributes to the theoretical underpinnings of data augmentation techniques in machine learning, offering a robust basis for the continued use and refinement of SMOTE in imbalanced learning scenarios.

References

1. Aguiar, G., Krawczyk, B., & Cano, A. (2024). A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework. Machine learning, 113(7), 4165-4243. https://doi.org/10.1007/s10994-023-06353-6
2. Aven, T. (1985). Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of Applied Probability*, 22, 723–728. https://doi.org/10.2307/3213876
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953
4. Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 113(7), 4903-4923. https://doi.org/10.1007/s10994-022-06296-4
5. Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing* (pp. 878–887). Springer. https://doi.org/10.1007/11538059_91

6. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)* (pp. 1322–1328). https://doi.org/10.1109/IJCNN.2008.4633969
7. Huang, K., & Wang, X. (2022). ADA-INCVAE: Improved data generation using variational autoencoder for imbalanced classification. Applied Intelligence, 52(3), 2838-2853. https://doi.org/10.1007/s10489-021-02566-1
8. Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems, 207, 106368. https://doi.org/10.1016/j.knosys.2020.106368
9. Kamalov, F. (2020). Kernel density estimation based sampling for imbalanced class distribution. Information Sciences, 512, 1192-1201. https://doi.org/10.1016/j.ins.2019.10.017
10. Sakho, A., Scornet, E., & Malherbe, E. (2024). Theoretical and experimental study of SMOTE: limitations and comparisons of rebalancing strategies. arXiv preprint arXiv:2402.03819. https://doi.org/10.48550/arXiv.2402.03819
11. Tsirelson, B. (2019). Expected difference of order statistics in terms of hazard rate. arXiv preprint arXiv:1911.06870. https://doi.org/10.48550/arXiv.1911.06870
12. Zhu, B., Pan, X., vanden Broucke, S., & Xiao, J. (2022). A GAN-based hybrid sampling method for imbalanced customer classification. Information Sciences, 609, 1397-1411. https://doi.org/10.1016/j.ins.2022.07.145

[1] Department of Electrical Engineering, Canadian University Dubai, Dubai, UAE.

Email address: firuz@cud.ac.ae