Gulf Journal of Mathematics
Vol 19, Issue 1 (2025) 400-415

ISSN: 2309-4966
https://doi.org/10.56947/gjom.v19i1.2639

ANALYTICAL FORMULATION OF SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) FOR IMBALANCED LEARNING

FIRUZ KAMALOV[1*], SALAH EDDINE CHOUTRI[2] and AMIR F. ATIYA[3]

Abstract. Imbalanced data is an issue that affects various applications in machine learning and data science. Synthetic minority oversampling technique (SMOTE) is a common method used to artificially balance the data. Despite the popularity of SMOTE, there is limited information about its analytical properties. In this paper, we develop a precise theoretical formulation of the sampling distribution of SMOTE in several important cases. We also examine the convergence of the SMOTE distribution to the underlying distribution in mean. The results provide a better understanding of SMOTE and other sampling algorithms. In addition, we uncover surprising connections to other fields such as information theory, Euler's constant, and compound distributions. Finally, we show that the SMOTE-generated distribution Z converges to that of the true underlying distribution X in mean.

1. Introduction

Imbalanced data refers to skewed distribution of classes in the dataset. It is an issue that affects various applications in machine learning and data science including medicine, network security, sentiment analysis, finance, and others [9, 10, 20, 21]. Sampling is a frequently used approach to deal with imbalanced data. The most popular existing sampling method is the synthetic minority oversampling technique (SMOTE) [2]. While the SMOTE algorithm is widely used and has multiple extensions, there is limited research regarding its sampling distribution. In this paper, we attempt to partially fill the gap in the literature by deriving the exact analytical density distribution of the SMOTE algorithm in several special cases. Our results reveal surprising connections to other fields such as information theory, Euler's constant, and the compound distributions.

There are two types of sampling techniques: 1) oversampling - increasing the number of the minority points, and 2) undersampling - decreasing the number of the majority points. Oversampling, or simply sampling, is the more commonly used technique in the literature. There exists a plethora of sampling algorithms that attempt to artificially increase the number of minority points in the data

Date: Received: Dec 21, 2024; Accepted: Jan 18, 2025.

* Corresponding author.

2020 *Mathematics Subject Classification.* Primary 62D05; Secondary 62B86, 62B15, 62G30.

Key words and phrases. imbalanced data, sampling, SMOTE, uniform distribution, Gaussian distribution.

[11, 12, 15]. While studies have shown the efficacy of sampling methods in various practical applications, there is little information about the theoretical aspects of these methods [5]. The lack of theoretical analysis of the distributions of the synthetic minority data restricts our understanding of the sampling methods and their robustness.

Our goal in this paper is to shed light on the theoretical probability density functions for the distributions of the synthetic minority samples generated by SMOTE algorithm for key special cases. In particular, we derive the explicit probability density functions in cases where the true distributions of the minority data are the following:

- Uniform
- Exponential
- Gaussian
- Cauchy

While we focus on univariate distributions in this paper, the results can be straightforwardly generalized to multivariate distributions when the features are independent. Even when the features are not independent, decorrelation techniques such as principal component analysis can be used to achieve acceptable levels of independence.

Finally, we examine the convergence of the SMOTE distribution to the underlying distribution in mean. In particular, we prove that $\lim_{n\to\infty} \mathbb{E}\,(Z - X) = 0$. It shows that as the sample size increases the SMOTE algorithm generates samples that are close to the true underlying distribution.

The paper is structures as follows. Section 2 describes the SMOTE algorithm and related literature. Section 3 contains the main results regarding the theoretical distributions. In Section 4, we study the distribution of the synthetic points as the sample size increases. Section 5 conclude the paper with closing remarks.

2. SMOTE

The SMOTE algorithm was originally proposed by Chawla et al. (2002) in [2]. It generates new synthetic samples of the minority class according to the following procedure. First, SMOTE randomly selects a minority instance x_i and finds its k nearest minority class neighbors. The synthetic instance is created by choosing one of the k nearest neighbors x_j at random and connecting x_i and x_j to form a line segment in the feature space. The synthetic instance z is then generated as a convex combination of the two chosen instances x_i and x_j [7]. To simplify the notation, in the remainder of the paper, we will denote the first chosen instance x_1 and the second instance x_2.

There exists numerous extensions of SMOTE that attempt to modify the original algorithm in order to improve the sampling outcome. The Adaptive Synthetic (ADASYN) algorithm is similar to SMOTE but it generates different number of samples depending on an estimate of the local distribution of the class to be oversampled [8]. Borderline SMOTE is another popular variant of the original SMOTE algorithm proposed in [6], where the borderline samples are detected and used to generate new synthetic samples. Support vector machines were used in

[18] to develop SVM-SMOTE. More recently several other extensions including Center Point SMOTE, Inner and Outer SMOTE [1], Deep Attention SMOTE [14], and others have been proposed.

Despite the immense popularity of SMOTE the sampling distribution of its synthetically generated points is not fully known. The concept of "no free lunch" in machine learning was applied to imbalanced learning in [17]. A comparative empirical study by the authors showed that any two resampling strategies would have the same classification performance given no a priori knowledge or data assumptions. In [3], the authors derive the expectation and covariance matrix of the SMOTE generated patterns. At the same time, in [13], the authors investigate the asymptotic behavior of SMOTE-generated samples using order statistics. Furthermore, in [4], the authors describe the sampling distribution of SMOTE in the form of a double integral function:

Theorem 2.1. [4] *Let x be a random sample of a random variable X. Let Z be a random variable defined as a random linear interpolation between $K-$nearest neighbors of x_k as given by SMOTE algorithm in* [2]*. Then, the probability density of Z is given by:*

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

where $B\left(1-I_{B(x,r);N-K-1,K}\right)$ is the incomplete beta function.

While the above result provides a general formulation of the sampling distribution of SMOTE, it is limited due to the complex integral term. Our goal is to consider several major cases of the distribution of x and obtain a simplified formulation of the density function $f_Z(z)$ of the of the random variable Z representing the synthetic instance. First, we consider the case when the underlying distribution of the minority class is uniform. Solving this case yields an unexpected connection to information entropy. Second, we consider the case when the underlying distribution is exponential which leads to a connection with Euler's constant. Third, we consider the Gaussian distribution which leads to a connection with compound distributions. Finally, we study the Cauchy distribution. Studying the density characteristics of the SMOTE patterns could stimulate developing density oriented over-sampling methods [16, 23].

3. Main results

In this paper, we consider the following simplified version of the SMOTE algorithm. Let (x_1, x_2) be a random sample drawn from a real-valued random variable X. Let z be a randomly chosen point between x_1 and x_2. In other words, z is a random convex combination of x_1 and x_2. In this procedure, x_1 and x_2 represent the minority class instances and z represents a synthetic instance.

Our goal is to develop analytic expressions for the density of the random variable Z when the distribution of minority data is known. We focus on four major special cases: uniform, exponential, Gaussian, and Cauchy distributions.

3.1. **Uniform distribution.** In this subsection, we discuss the distribution of the synthetically generated points, in the case when the true distribution of the minority class points is uniform. In other words, we consider the scenario where the minority class sample is randomly taken from a uniformly distributed source, and the SMOTE algorithm is applied to the selected sample to generate new synthetic points.

We formalize the above scenario in the following fashion. Suppose that a random sample of size 2 is randomly taken from a uniform distribution over the interval $[0, 1]$. Let us denote the sample points by X_1 and X_2. Given the points x_1 and x_2, a new point z is chosen on the line segment joining x_1 and x_2, as illustrated in Figure 1. Let Z be the corresponding random variable. Our goal is to derive the probability density function of Z.

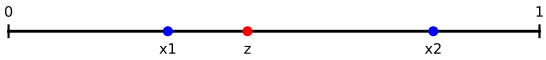

FIGURE 1. Synthetic point z is randomly generated between neighboring minority points x_1 and x_2.

Let us first assume that $X_1 < X_2$. Then, the conditional probability density of Z is

$$f(z|x_1, x_2) = \frac{1}{x_2 - x_1}. \tag{3.1}$$

To obtain the marginal density function of Z, we integrate over the range of possible values of X_1 and X_2:

$$\begin{aligned} f(z|x_1 < x_2) &= \int_0^z \int_z^1 \frac{1}{x_2 - x_1}\, dx_2\, dx_1 \\ &= \int_0^z \big[\ln(x_2 - x_1)\big|_z^1\big]\, dx_1 \\ &= \int_0^z \ln(1 - x_1) - \ln(z - x_1)\, dx_1 \\ &= -(1 - x_1)\ln(1 - x_1) + (1 - x_1) + (z - x_1)\ln(z - x_1) - (z - x_1)\big|_0^z \\ &= -(1 - z)\ln(1 - z) + (1 - z) - (1 + z\ln z - z) \\ &= -z\ln z - (1 - z)\ln(1 - z) \end{aligned} \tag{3.2}$$

Note that in Equation 3.2, we used the fact that

$$\lim_{z \to 0} z \ln z = 0.$$

Similarly, the marginal density function of Z when $x_2 < x_1$ is given by the expression $-z\ln z - (1 - z)\ln(1 - z)$. Thus, the complete marginal distribution function is given by

$$f(z) = -2\big(z\ln z + (1 - z)\ln(1 - z)\big) \tag{3.3}$$

To verify the above theoretical result, we simulate the distribution of Z based on 10^7 randomly generated points. As shown in Figure 2, the histogram of the simulated values of Z matches with the graph of the function $f(z)$.

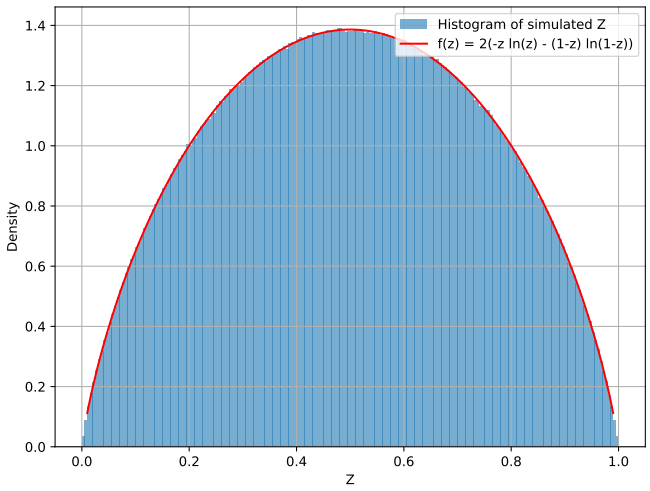

FIGURE 2. The histogram of 10^7 simulated values of Z based on the uniform distribution together with the graph of $f(z)$.

There exists a surprising connection between the distribution of Z presented in Equation 3.3 and information entropy. Recall that the entropy of a Bernoulli random variable X is given by the equation

$$H(X) = -q \ln q - (1-q) \ln(1-q), \tag{3.4}$$

where q is the probability of the positive outcome. If we view the expression in Equation 3.4 as the function of z, i.e., $H(z) = -z \ln z - (1-z) \ln(1-z)$, then we get that the distribution of Z equals twice the entropy of a Bernoulli random variable

$$f(z) = 2H(z). \tag{3.5}$$

3.2. **Exponential distribution.** In this subsection, we explore the distribution when the true distribution of the minority points is exponential. In particular, assume that $X \sim \exp(\lambda = 1)$. We will employ a similar strategy as in the case of the uniform distribution to set up our analysis.

$$
\begin{aligned}
f(z|x_1 < x_2) &= \int_0^z \int_z^\infty \frac{1}{x_2 - x_1} e^{-x_2} e^{-x_1}\, dx_2\, dx_1 \\
&= \int_0^z e^{-x_1} \left[\int_z^\infty \frac{1}{x_2 - x_1} e^{-x_2}\, dx_2 \right] dx_1 \\
&= \int_0^z e^{-2x_1} \int_{z-x_1}^\infty \frac{e^{-t}}{t}\, dt\, dx_1 \\
&= -\frac{1}{2} e^{-2x_1} \int_{z-x_1}^\infty \frac{e^{-t}}{t}\, dt \Big|_0^z - \int_0^z -\frac{1}{2} e^{-2x_1} \cdot \frac{e^{-(z-x_1)}}{z - x_1}\, dx_1 \\
&= -\frac{1}{2} e^{-2z} \int_0^\infty \frac{e^{-t}}{t}\, dt + \frac{1}{2} \int_z^\infty \frac{e^{-t}}{t}\, dt + \frac{1}{2} \int_0^z \frac{e^{-z-x_1}}{z - x_1}\, dx_1 \\
&= -\frac{1}{2} e^{-2z} \int_0^\infty \frac{e^{-t}}{t}\, dt + \frac{1}{2} \int_z^\infty \frac{e^{-t}}{t}\, dt + \frac{1}{2} e^{-2z} \int_0^z \frac{e^t}{t}\, dt \\
&= \frac{1}{2} e^{-2z} \int_0^z \frac{e^t - e^{-t}}{t}\, dt + \frac{1}{2}(1 - e^{-2z}) \int_z^\infty \frac{e^{-t}}{t}\, dt.
\end{aligned}
\tag{3.6}
$$

Note that in line 2 of Equation 3.6, we made a substitution $t = x_2 - x_1$ and in line 3 we employed the integration by parts formula with $u = \int_{z-x_1}^\infty \frac{e^{-t}}{t}\, dt$ and $dv = e^{-2x_1}$. In line 5, we made another substitution $t = z - x_1$. In line 6, we split the first integral and grouped similar integrals.

Next, we analyze each integral obtained in the end of Equation 3.6 using the Taylor series expansion of the exponential function.

$$
\begin{aligned}
\frac{e^t - e^{-t}}{t} &= \frac{1}{t} \left[(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \dots) - (1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \dots) \right] \\
&= 2(1 + \frac{t^2}{3!} + \frac{t^4}{5!} + \dots) \\
&= 2 \sum_{k=0}^\infty \frac{t^{2k}}{(2k+1)!}.
\end{aligned}
\tag{3.7}
$$

It follows from Equation 3.7 that

$$
\int_0^z \frac{e^t - e^{-t}}{t}\, dt = 2 \sum_{k=0}^\infty \frac{z^{2k+1}}{(2k+1)(2k+1)!}. \tag{3.8}
$$

Similarly,

$$\begin{aligned}\int_z^\infty \frac{e^{-t}}{t}\,dt &= \int_z^\infty \frac{1}{t}\left(1+\sum_{k=1}^{\infty}(-1)^k\frac{t^k}{k!}\right)dt \\ &= \int_z^\infty \frac{1}{t}+\sum_{k=1}^{\infty}(-1)^k\frac{t^{k-1}}{k!}\,dt \\ &= \ln z+\sum_{k=1}^{\infty}(-1)^k\frac{t^k}{k\cdot k!}\Bigg|_z^\infty \\ &= -\gamma-\ln z-\sum_{k=1}^{\infty}(-1)^k\frac{z^k}{k\cdot k!},\end{aligned} \tag{3.9}$$

where γ is Euler's constant. Combining the results in Equations 3.6, 3.8, and 3.9, we get the following

$$f(z|x_1<x_2)=e^{-2z}\sum_{k=0}^{\infty}\frac{z^{2k+1}}{(2k+1)(2k+1)!}-\frac{1}{2}(1-e^{-2z})\left(\gamma+\ln z+\sum_{k=1}^{\infty}(-1)^k\frac{z^k}{k\cdot k!}\right). \tag{3.10}$$

Finally, we observe that

$$f(z)=2\cdot f(z|x_1<x_2). \tag{3.11}$$

To verify our results, we simulated the values of Z with 10^7 random samples. As shown in Figure 3, the histogram of the simulated values matches almost perfectly with the graph of $f(z)$.

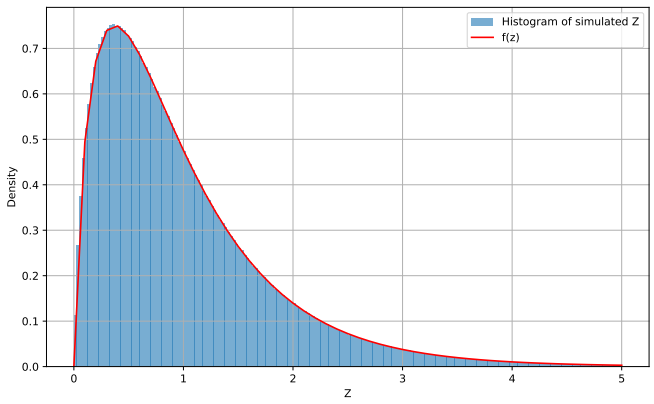

FIGURE 3. The histogram of 10^7 simulated values of Z based on the exponential distribution together with the graph of $f(z)$ in Equation 3.11.

3.3. **Gaussian distribution.** In this subsection, we explore the sampling distribution when the true distribution of the minority points is Gaussian. Since z is randomly chosen between X_1 and X_2, it can be viewed as a linear interpolation of the minority points. Let us define the random variable Z as

$$Z = \Theta X_1 + (1 - \Theta)X_2,$$

where

- $X_1, X_2 \sim N(0, 1)$ are independent,
- $\Theta \sim U(0, 1)$ is independent of X_1 and X_2.

Next, we attempt to derive the density function of the random variable Z, from the joint density of Z and Θ using Bayes' theorem for densities:

$$f_{Z|\Theta}(z|\theta) = \frac{f_{Z,\Theta}(z, \theta)}{f_\Theta(\theta)}, \tag{3.12}$$

where

- $f_{Z|\Theta}(z|\theta)$ is the conditional density of Z given $\Theta = \theta$,
- $f_{Z,\Theta}(z, \theta)$ is the joint density of of Z and Θ,
- $f_\Theta(\theta)$ is the marginal density of Θ.

For a fixed value $\Theta = \theta$, the random variable $Z|\Theta$ is normally distributed since it is a combination of two normally distributed random variables X_1 and X_2. Its parameters are computed as follows:

- Conditional mean $E[Z|\theta]$: Since $E[X_1] = E[X_2] = 0$, $E[Z|\theta] = \theta \cdot 0 + (1 - \theta) \cdot 0 = 0$.
- Conditional variance $Var[Z|\theta]$: The variance of Z given θ is

$$Var[Z|\theta] = \theta^2 \cdot Var(X_1) + (1 - \theta)^2 \cdot Var(X_2) = \theta^2 \cdot 1 + (1 - \theta)^2 \cdot 1 = \theta^2 + (1 - \theta)^2$$

$$(\text{since } Var(X_1) = Var(X_2) = 1 \text{ and } X_1, X_2 \text{ are independent}).$$

Thus, the distribution of $Z|\Theta = \theta$ is $N(0, \theta^2 + (1 - \theta)^2)$ and its density is given by

$$f_{Z|\Theta}(z|\theta) = \frac{1}{\sqrt{2\pi(\theta^2 + (1 - \theta)^2)}} \exp\left(-\frac{z^2}{2(\theta^2 + (1 - \theta)^2)}\right).$$

Using Bayes' formula in (3.12), the joint density of Z and Θ is given by

$$f_{Z,\Theta}(z, \theta) = \frac{1}{\sqrt{2\pi(\theta^2 + (1 - \theta)^2)}} \exp\left(-\frac{z^2}{2(\theta^2 + (1 - \theta)^2)}\right) \cdot 1.$$

To find the marginal density of Z, we integrate the joint density over all possible values of Θ:

$$f_Z(z) = \int_0^1 f_{Z,\Theta}(z, \theta)d\theta = \int_0^1 \frac{1}{\sqrt{2\pi(\theta^2 + (1 - \theta)^2)}} \exp\left(-\frac{z^2}{2(\theta^2 + (1 - \theta)^2)}\right) d\theta.$$

This integral is challenging and might not have an explicit form requiring numerical integration methods. However, given the form of this integral, the linearity of Z as a combination of normal variables and the "mixing" weights coming from Θ, the distribution of Z can be regarded as a mixture of Gaussian distributions with smoothly different variances.

To highlight the influence of the variance of the compound distribution $\Theta^2 + (1 - \Theta)^2$ on the density of Z, we express the marginal density of Z as follows.

Let $\Lambda = \Theta^2 + (1-\Theta)^2$, be the random variance with the support $[0.5, 1]$, its cumulative distribution function is given by

$$\begin{aligned} F_\Lambda(\lambda) &= P(\Lambda \leq \lambda) \\ &= P(\theta \leq \Theta \leq (1-\theta)) \\ &= 1 - 2\theta \\ &= 1 - (1 - \sqrt{2\lambda - 1}) \\ &= \sqrt{2\lambda - 1}. \end{aligned}$$

Taking the derivative of $F_\lambda(\Lambda)$ yields the density of Λ,

$$\frac{1}{\sqrt{2\lambda - 1}}, \ \lambda \in [0.5, 1]. \tag{3.13}$$

Therefore, the marginal density of Z can be expressed as

$$\begin{aligned} f_Z(z) &= \int_{0.5}^{1} f_{Z|\Lambda}(z|\lambda) f_\Lambda(\lambda) d\lambda \\ &= \int_{0.5}^{1} \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{z^2}{2\lambda}\right) \frac{1}{\sqrt{2\lambda - 1}} \, d\lambda. \end{aligned} \tag{3.14}$$

The formulation in Equation 3.14 shows that $f_Z(z)$ can be viewed as a compound probability distribution [19]. In particular, Z is distributed according to the Gaussian distribution with variance λ that is again distributed according to the distribution in Equation 3.13.

In the absence of an analytical approach of computing the integrals describing the density of the random variable Z, one can analyze the properties of the distribution such as the expectation and the variance. It is clear that the expectation $E[Z] = 0$. The variance is given by:

$$\mathrm{Var}(Z) = E[\Theta^2]\mathrm{Var}(X_1) + E[(1-\Theta)^2]\mathrm{Var}(X_2).$$

Since $E[\Theta^2] = \int_0^1 \theta^2 d\theta = 1/3$ (and similarly for $E[(1-\Theta)^2]$), and noting the independence and identical distribution of X_1 and X_2, we find:

$$\mathrm{Var}(Z) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$

Finally, while the distribution of Z is given by the compound density function in Equation 3.14, it can be well approximated via Gaussian density function with the above parameters. As shown in Figure 4, the true distribution of Z is very close to the Gaussian distribution with mean 0 and variance $\frac{2}{3}$.

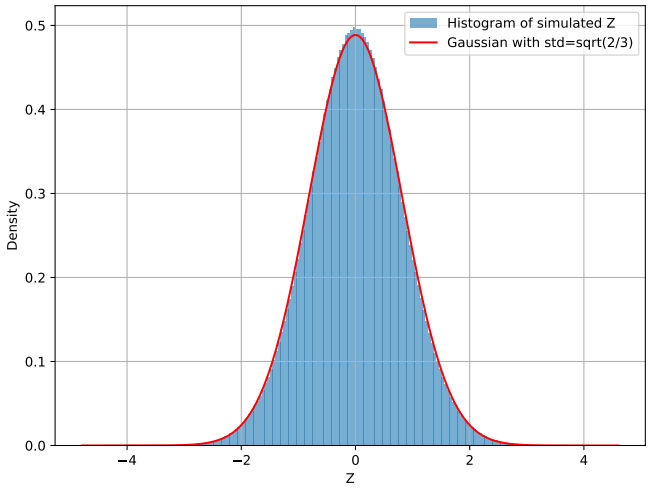

FIGURE 4. The histogram of 10^7 simulated values of Z together with the density of $\mathcal{N}(0, 2/3)$.

3.4. **Cauchy distribution.** In this subsection, we assume that our sample is drawn from a standard Cauchy distribution

$$Z = \Theta X_1 + (1 - \Theta)X_2,$$

where

- $X_1, X_2 \sim Cauchy(0, 1)$ are independent,
- $\Theta \sim U(0, 1)$ is independent of X_1 and X_2.

Note that for a fixed value θ, the random variables θX_1 and $(1-\theta)X_2$ both follow Cauchy distribution. Specifically

- $\theta X_1 \sim$ Cauchy$(0,\ \theta)$,
- $(1-\theta)X_2 \sim$ Cauchy$(0,\ 1-\theta)$.

Since $Z = \theta X_1 + (1-\theta)X_2$ is the sum of two independent Cauchy random variables with different scale parameters, the conditional distribution of Z, given $\Theta = \theta$ is also Cauchy with density $f_{Z|\Theta}(z|\theta) = \frac{1}{\pi(1+z^2)}$. We derive the density function of the random variable Z, by integrating the joint density of Z and Θ over all the possible values of Θ,

$$\begin{aligned} f_Z(z) &= \int_0^1 f_{Z|\Theta}(z|\theta) \cdot f_\Theta(\theta) d\theta \\ &= \int_0^1 \frac{1}{\pi(1+z^2)}\, d\theta \\ &= \frac{1}{\pi(1+z^2)}. \end{aligned} \tag{3.15}$$

It follows from Equation 3.15 that the marginal distribution of Z is also Cauchy. The generalization to different parameters is straightforward since Cauchy is a stable distribution.

To verify the above theoretical result, we simulate the distribution of Z based on 10^7 randomly generated points. As shown in Figure 5, the histogram of the simulated values of Z matches with the graph of the function $f_Z f(z)$.

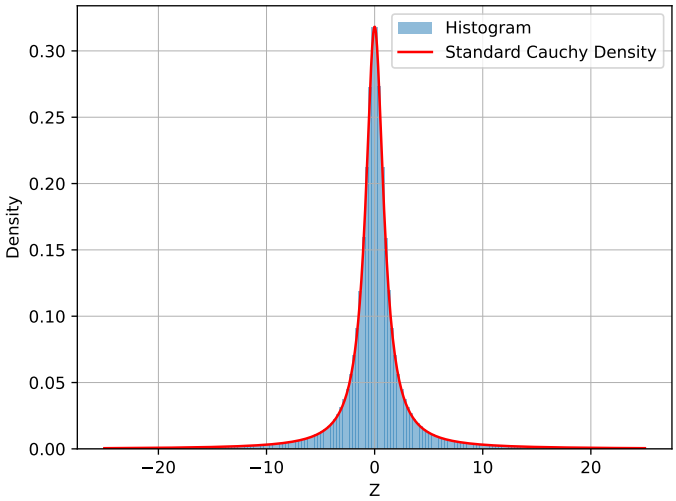

FIGURE 5. The histogram of 10^7 simulated values of Z together with the density of $Cauchy(0, 1)$.

4. Convergence in mean

It was shown in Section 3 that the distribution of Z does not match that of X in case of the uniform and exponential random variables, when the sample size is 2. However, as shown in Figures 6 and 7, the distribution of Z approaches the distribution of X as the sample size n increases. In this section, we show analytically in special cases that Z converges to X in mean, as n increases.

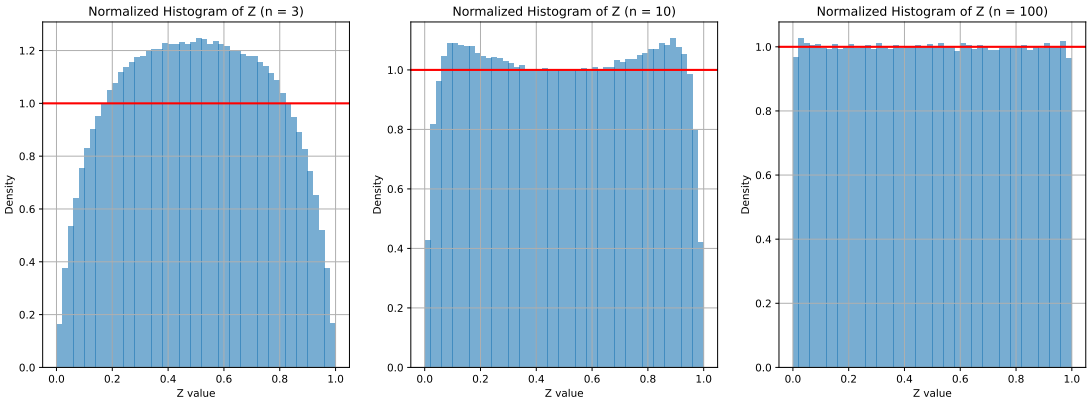

FIGURE 6. Numerical simulation of Z for different values of n, where $X \sim U(0, 1)$.

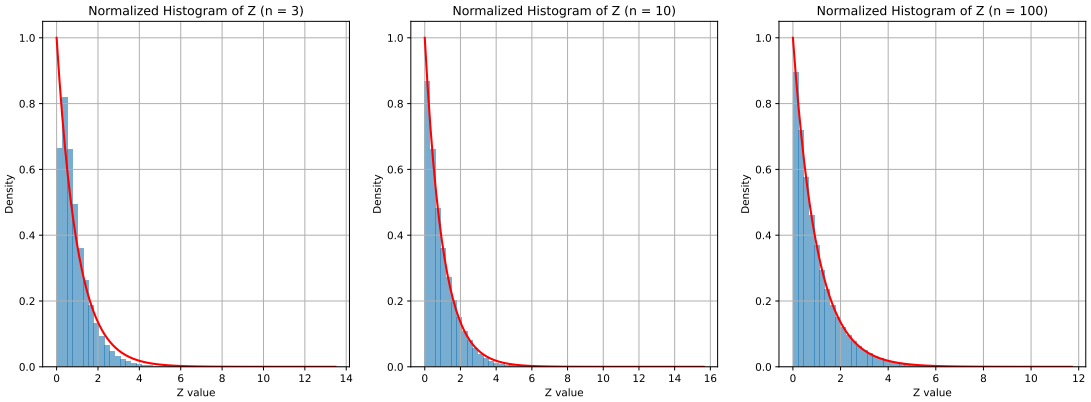

FIGURE 7. Numerical simulation of Z for different values of n, where $X \sim Exp(\lambda = 1)$.

We consider the following version of SMOTE for a sample of size n. Let X be a random variable with a probability density function $f_X(x)$. Let $n \geq 2$ be a fixed positive integer. Apply the following procedure:

(1) Select a random sample of X: $S = \{x_1, x_2, \ldots, x_n\}$.
(2) Select a random point $x_i \in S$.
(3) Let $x_j \in S$ be the nearest neighbor of x_i.
(4) Select a random point z on the line segment connecting x_i and x_j.

We require the following lemma to obtain our results.

Lemma 4.1. [22] *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 $k = 1, \ldots, n-1$.

First, we will show that for the uniform and exponential distributions the distance between neighboring points in a sample converges to zero as $n \to \infty$.

Proposition 4.2. *In the same scenario as in Lemma 4.1, the following are true*

a. *If $f(x) = 1, [0, 1]$, then $\mathbb{E}\left(X_{(k+1)} - X_{(k)}\right) = \frac{1}{n}$*
b. *If $f(x) = \alpha e^{-\alpha x}, [0, \infty)$, then $\mathbb{E}\left(X_{(k+1)} - X_{(k)}\right) = \frac{1}{a(n-k)}$*

Proof. If $f(x) = 1, [0, 1]$, then

$$\begin{aligned}
\mathbb{E}\left(X_{(k+1)} - X_{(k)}\right) &= \binom{n}{k} \int_0^1 x^k (1-x)^{n-k}\, \mathrm{d}x \\
&= \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} \\
&= \frac{1}{n}.
\end{aligned}$$

If $f(x) = \alpha e^{-\alpha x}, [0, \infty)$, then

$$\begin{aligned}
\mathbb{E}\left(X_{(k+1)} - X_{(k)}\right) &= \binom{n}{k} \int_0^\infty (1 - e^{-\alpha x})^k (e^{-\alpha x})^{n-k} \, \mathrm{d}x \\
&= \binom{n}{k} \int_0^1 (1-u)^k (u)^{n-k-1} \frac{\mathrm{d}u}{a} \\
&= \frac{1}{a} \binom{n}{k} \frac{k!(n-k-1)!}{n!} \\
&= \frac{1}{a(n-k)}.
\end{aligned}$$

□

Next, we will use the above proposition to show that Z converges to X in mean.

Corollary 4.3. *Let X be a uniform or exponential random variable as in Proposition 4.2. Then $\lim_{n\to\infty} \mathbb{E}\,(Z - X) = 0$.*

Proof. First, assume that X is a uniform random variable. We begin by showing the following conditional convergence:

$$\lim_{n\to\infty} \mathbb{E}\left(Z|X_{(k)} - X_{(k)}\right) = 0,$$

for any fixed k. Since $Z|X_{(k)}$ lies between $X_{(k)}$ and its m-th closest neighbor, then

$$\left|Z|X_{(k)} - X_{(k)}\right| \leq \max\left\{X_{(k+m)} - X_{(k)}, X_{(k)} - X_{(k-m)}\right\}$$

Assume, without the loss of generality, that

$$\mathbb{E}\left(\left|Z|X_{(k)} - X_{(k)}\right|\right) \leq \mathbb{E}\left(X_{(k+m)} - X_{(k)}\right)$$

Then,

$$\begin{aligned}
\mathbb{E}\left(\left|Z|X_{(k)} - X_{(k)}\right|\right) &\leq \mathbb{E}\left(X_{(k+m)} - X_{(k)}\right) \\
&= \sum_{i=0}^{m-1} \mathbb{E}\left(X_{(k+i+1)} - X_{(k+i)}\right) \\
&= m \cdot \frac{1}{n} \\
&= 0.
\end{aligned}$$

To find the unconditional expectation, we need to average over all $X_{(k)}$:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\,(Z - X) &= \lim_{n\to\infty} \mathbb{E}_{X_{(k)}}\left(\mathbb{E}(Z|X_{(k)} - X_{(k)})\right) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}(Z|X_{(k)} - X_{(k)}) \\
&\leq \lim_{n\to\infty} \frac{m}{n} \\
&= 0
\end{aligned}$$

Next, assume that X is an exponential random variable. Following a similar argument as above, we get

$$\begin{aligned}\mathbb{E}\left(\left|Z|X_{(k)}-X_{(k)}\right|\right) &\leq \sum_{i=0}^{m-1}\mathbb{E}\left(X_{(k+i+1)}-X_{(k+i)}\right)\\ &= \sum_{i=0}^{m-1}\frac{1}{n-k-i}\end{aligned}$$

To find the unconditional expectation, we need to average over all $X_{(k)}$:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left(Z-X\right) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Z|X_{(k)}-X_{(k)})\\ &\leq \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n-1}\sum_{i=0}^{m-1}\frac{1}{n-k-i}\\ &= 0\end{aligned}$$

□

5. Conclusion

Sampling is a popular approach to dealing with imbalanced data. In this paper, we studied the sampling distribution of the SMOTE-generated instances under various scenarios. We obtained analytical expressions in important cases where the underlying distribution of the minority instances is uniform, exponential, Gaussian, and Cauchy. The results provide a better understanding of sampling methods. In addition, we uncovered surprising connections with other fields of mathematics.

While our study sheds light on analytical aspects of sampling methods, it has a limited scope. We considered a simplified version of SMOTE. Further analysis is required to obtain more general results. It is an interesting and fruitful avenue for research - one which we hope to pursue in the future.

References

1. Bao, Y., & Yang, S. (2023). Two novel SMOTE methods for solving imbalanced classification problems. IEEE Access, 11, 5816-5823. https://doi.org/10.1109/ACCESS.2023.3236794
2. 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
3. Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences, 505, 32-64. https://doi.org/10.1016/j.ins.2019.07.070
4. Elreedy, D., Atiya, A. F., & Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 1-21. https://doi.org/10.1007/s10994-022-06296-4
5. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, No. 2018). Cham: Springer. https://doi.org/10.1007/978-3-319-98074-4

6. Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11538059_91
7. He, H., & Ma, Y. (Eds.). (2013). Imbalanced learning: foundations, algorithms, and applications. https://doi.org/10.1002/9781118646106
8. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). 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. https://doi.org/10.1109/IJCNN.2008.4633969
9. Iqbal, S., Qureshi, A. N., Li, J., Choudhry, I. A., & Mahmood, T. (2023). Dynamic learning for imbalanced data in learning chest X-ray and CT images. Heliyon, 9(6). https://doi.org/10.1016/j.heliyon.2023.e16807
10. Jiang, C., Lu, W., Wang, Z., & Ding, Y. (2023). Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring. Expert systems with applications, 213, 118878. https://doi.org/10.1016/j.eswa.2022.118878
11. 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
12. Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems, 207, 106368. Chicago https://doi.org/10.1016/j.knosys.2020.106368
13. Kamalov, F. (2024). Asymptotic behavior of SMOTE-generated samples using order statistics. Gulf Journal of Mathematics, 17(2), 327-336. https://doi.org/10.56947/gjom.v17i2.2343
14. Liu, D., Zhong, S., Lin, L., Zhao, M., Fu, X., & Liu, X. (2023). Deep attention SMOTE: Data augmentation with a learnable interpolation factor for imbalanced anomaly detection of gas turbines. Computers in Industry, 151, 103972. https://doi.org/10.1016/j.compind.2023.103972
15. Liu, Y., Liu, Y., Bruce, X. B., Zhong, S., & Hu, Z. (2023). Noise-robust oversampling for imbalanced data classification. Pattern Recognition, 133, 109008. https://doi.org/10.1016/j.patcog.2022.109008
16. Mayabadi, S., & Saadatfar, H. (2022). Two density-based sampling approaches for imbalanced and overlapping data. Knowledge-Based Systems, 241, 108217. https://doi.org/10.1016/j.knosys.2022.108217
17. Moniz, N., & Monteiro, H. (2021). No Free Lunch in imbalanced learning. Knowledge-Based Systems, 227, 107222. https://doi.org/10.1016/j.knosys.2021.107222
18. Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), 4-21. https://doi.org/10.1504/IJKESDP.2011.039875
19. Olofsson, T., & Ahlén, A. (2018, May). Computing probability density functions of compound distributions: A comparative investigation. In 2018 IEEE International Conference on Communications (ICC) (pp. 1-7). IEEE. https://doi.org/10.1109/ICC.2018.8423041
20. Rao, Y. N., & Suresh Babu, K. (2023). An imbalanced generative adversarial network-based approach for network intrusion detection in an imbalanced dataset. Sensors, 23(1), 550. https://doi.org/10.3390/s23010550
21. Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429-441. https://doi.org/10.1016/j.ins.2019.11.004
22. 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

23. Yan, Y., Jiang, Y., Zheng, Z., Yu, C., Zhang, Y., & Zhang, Y. (2022). LDAS: Local density-based adaptive sampling for imbalanced data classification. Expert Systems with Applications, 191, 116213. https://doi.org/10.1016/j.eswa.2021.116213

[1] Department of Electrical Engineering, Canadian University Dubai, Dubai, UAE
Email address: firuz@cud.ac.ae

[2] NYUAD Research Institute, New York University, Abu Dhabi, UAE
Email address: Sc8101@nyu.edu

[3] Computer Engineering Department, Cairo University, Giza, 12613, Egypt
Email address: amir@alumni.caltech.edu