Supplementary of "Deep Generative Model for Robust Imbalance Classification"

Proof of Theorem 1 and 2

In this supplementary, we will theoretically analyze the generalization error bounds of the proposed model. Motivated by [3], we can define the generalization of input feature and label generation process by measuring the difference between the population real data distribution ($\mathcal{P}_{real}$) and the corresponding generated data distribution ($\mathcal{P}_G$). The generalization error will be acceptable if this population distance is close to the empirical distance between the observed real data distribution ($\tilde{\mathcal{P}}_{real}$) and the corresponding generated data distribution ($\tilde{\mathcal{P}}_G$). In the proposed deep latent variable model, given latent code Z, the input feature information X is independent to input label Y, i.e., $(X \perp Y|Z)$. In this case, the data distribution $(X, Y) \sim \mathcal{P}_{real}$ can be factorized into two parts $X \sim \mathcal{X}_{real}$ and $Y \sim \mathcal{Y}_{real}$, similar to $\mathcal{P}_G$, $\tilde{\mathcal{P}}_{real}$ and $\tilde{\mathcal{P}}_G$.

Definition 2. *For the empirical real distribution ($\tilde{\mathcal{X}}_{real}$, $\tilde{\mathcal{Y}}_{real}$) with N training examples, a generated distribution ($\tilde{\mathcal{X}}_G$, $\tilde{\mathcal{Y}}_G$) generalizes under the distribution distance $d(\cdot,\cdot)$ with generalization error $\delta_1, \delta_2 > 0$ if the following holds with high probability,*

$$|E(X) - E(\tilde{X})| \leq \delta_1 \tag{8}$$

$$|E(Y) - E(\tilde{Y})| \leq \delta_2 \tag{9}$$

where

$$\begin{aligned}
E(X) &= \mathbb{E}_{\mathbf{x}^{(real)} \sim \mathcal{X}_{real}, \mathbf{x}^{(G)} \sim \mathcal{X}_G} d(\mathbf{x}^{(real)}, \mathbf{x}^{(G)}), \\
E(\tilde{X}) &= \frac{1}{N} \sum_{i=1}^{N} d(\tilde{\mathbf{x}}_i^{(real)}, \tilde{\mathbf{x}}_i^{(G)}) | \tilde{\mathbf{x}}_i^{(real)} \sim \tilde{\mathcal{X}}_{real}, \tilde{\mathbf{x}}_i^{(G)} \sim \tilde{\mathcal{X}}_G \\
E(Y) &= \mathbb{E}_{\mathbf{y}^{(real)} \sim \mathcal{Y}_{real}, \mathbf{y}^{(G)} \sim \mathcal{Y}_G} d(\mathbf{y}^{(real)}, \mathbf{y}^{(G)}), \\
E(\hat{Y}) &= \frac{1}{N} \sum_{i=1}^{N} d(\tilde{\mathbf{y}}_i^{(real)}, \tilde{\mathbf{y}}_i^{(G)}) | \tilde{\mathbf{y}}_i^{(real)} \sim \tilde{\mathcal{Y}}_{real}, \tilde{\mathbf{y}}_i^{(G)} \sim \tilde{\mathcal{Y}}_G
\end{aligned}$$

$E(\mathcal{X})$ and $E(\mathcal{Y})$ indicate the population distance between the real and generated distributions on feature and label information respectively. $E(\tilde{\mathcal{X}})$ and $E(\tilde{\mathcal{Y}})$ are the corresponding empirical distances.

Theorem 3. *For any $X \in \mathbb{R}^{D \times N}$ $(N, D > 0)$,*

$$E(X) \leq E(\tilde{X}) + \sqrt{\frac{\log \delta_1}{-2N} (\max_i d_i)^2}$$

holds with probability at least $1 - \delta_1 (\delta_1 > 0)$ over uniformly choosing an empirical version $(\tilde{X})$ of X. Here, $d_i = d(\tilde{\mathbf{x}}_i^{(real)}, \tilde{\mathbf{x}}_i^{(G)}) = ||\tilde{\mathbf{x}}_i^{(real)} - \tilde{\mathbf{x}}_i^{(G)}||_2^2$.

Proof 1. *Since the entries in X are chosen independently and uniformly, it is reasonable to assume each $d_i = d(\tilde{\mathbf{x}}_i^{(real)}, \tilde{\mathbf{x}}_i^{(G)})$ is a random variable and satisfies $p(\zeta \geq d_i \geq 0) = 1$, where $\zeta = \max_i d_i$. Hence, based on the Hoeffding Inequality, we have $p(|E(X) - E(\tilde{X})| \geq \epsilon) \leq \exp(\frac{-2N\epsilon^2}{\zeta^2})$. By setting $\epsilon = \sqrt{\frac{\log \delta_1}{-2N}\zeta^2}$, we have $p\Big(|E(X) - E(\tilde{X})| \leq \sqrt{\frac{\log \delta_1}{-2N}\zeta^2}\Big) \geq 1 - \delta_1$, i.e.,*

$$p\Big(E(\mathbf{X}) \leq E(\tilde{\mathbf{X}})| + \sqrt{\frac{\log \delta_1}{-2N} (\max_i d_i)^2}\Big) \geq 1 - \delta_1$$

Therefore, the error of generative model for feature information is bounded.

Theorem 4. *Given prior label probabilities $\{p_1, \ldots, p_c, \ldots, p_C\}$ (where $p_c = P(y = c)$ and $\sum_{c=1}^{C} p_c = 1$) and the conditional latent variable densities $\{f_1, f_2, \ldots, f_C\}$ (where $f_c = f(\mathbf{z}|y = c)$) , following [30], the error rate of generative classifier can be formulated and bounded:*

$$\epsilon^C = 1 - \int \max\{p_1 f_1(\mathbf{z}), \ldots, p_C f_C(\mathbf{z})\} d\mathbf{z} \leq \delta_2$$

Proof 2. *According to the definition of ϵ^C, we can get*

$$\epsilon^C = \mathbb{E}_{\mathbf{z}}\big[1 - \max_{c=1,\ldots,C} p(y = c|\mathbf{z})\big] \tag{10}$$

Denote $a_u = p(y = u|\mathbf{z})$ and they are sorted in an ascending order, i.e. $a_C = \max\{a_u|_{u=1}^C\}$, we have $\sum_{v=1}^{C} a_v = 1$ and

$$\begin{aligned}
1 - a_C &= \sum_{u=1}^{C-1} a_u = \sum_{u=1}^{C-1} \sum_{v=1}^{C} a_u a_v \\
&= 2 \sum_{u=1}^{C-2} \sum_{v=u+1}^{C-1} a_u a_v + \sum_{u=1}^{C-1} a_u a_C + \sum_{u=1}^{C-1} a_u^2 \\
&\leq 2 \sum_{u=1}^{C-2} \sum_{v=u+1}^{C-1} a_u a_v + \sum_{u=1}^{C-1} a_u a_C + \sum_{u=1}^{C-1} a_u a_C
\end{aligned} \tag{11}$$

Let denote the marginal distribution of Z as

$$f^{(C)}(\mathbf{z}) = \sum_{c=1}^{C} p_c f_c(\mathbf{z}) = \sum_{c=1}^{C} p_c f(\mathbf{z}|y = c)$$

$$\delta_{u,v}^C := \int \frac{p_u f_u(\mathbf{z}) p_v f_v(\mathbf{z}\big)}{f^{(C)}(\mathbf{z})} d\mathbf{z} = \mathbb{E}_{\mathbf{z}}[a_u a_v]$$

According to the definition of Expectation, we know

$$\begin{aligned}
\delta_{u,v}^C &= \mathbb{E}_{\mathbf{z}}[a_u a_v] = 2 \sum_{u=1}^{C-1} \sum_{v=u+1}^{C} a_u a_v \\
&= 2 \sum_{u=1}^{C-2} \sum_{v=u+1}^{C-1} a_u a_v + 2 \sum_{u=1}^{C-1} a_u a_C.
\end{aligned} \tag{12}$$

Obviously, $1 - a_C \leq \delta_{u,v}^C$. By setting $\delta_2 = \delta_{u,v}^C$, $\epsilon^C \leq \delta_2$ holds.