FAIR CONFORMAL CLASSIFICATION VIA LEARNING REPRESENTATION-BASED GROUPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction methods provide statistically rigorous marginal coverage guarantees for machine learning models, but such guarantees fail to account for algorithmic biases, thereby undermining fairness and trust. This paper introduces a fair conformal inference framework for classification tasks. The proposed method constructs prediction sets that guarantee conditional coverage on adaptively identified subgroups, which can be implicitly defined through nonlinear feature combinations. By balancing effectiveness and efficiency in producing compact, informative prediction sets and ensuring adaptive equalized coverage across unfairly treated subgroups, our approach paves a practical pathway toward trustworthy machine learning. Extensive experiments on both synthetic and real-world datasets demonstrate the effectiveness of the framework.

1 INTRODUCTION

The rapid advancement of modern machine learning models, especially deep neural networks, has enabled their deployment in high-stake decision-making situations such as medical diagnoses (Kaur et al., 2020), resume filtering (Deshpande et al., 2020), and financial fraud detection (Kamuangu, 2024). Despite their strong average performance, real-world deployment raises critical challenges, notably in uncertainty quantification (Guo et al., 2017; Ahmed et al., 2023) and algorithmic fairness (Berk et al., 2024; Almasoud & Idowu, 2025).

Ensuring reliable decision-making necessitates the development of unbiased uncertainty measures, as even highly accurate models are prone to producing over-confident and erroneous predictions (Ovadia et al., 2019). Conformal prediction (CP, (Vovk et al., 2005; Smith, 2024)) has emerged as a key framework for providing distribution-free, model-agnostic prediction sets with user-specified (marginal) coverage guarantees. These sets provide reliable uncertainty information for decision-makers especially when the set size is small (i.e., with high efficiency).

On the other hand, algorithmic biases often manifest as disproportionately poor performance on the subgroup defined by specific feature conditions (e.g., *Race=Black & Gender=Female*), which may arise from imbalanced data distribution or model inherent limitations (Hellman, 2020). These biases underscore the need for algorithmic fairness mechanisms that extend beyond average performance to ensure equitable treatment across all groups (Fabris et al., 2022; Das et al., 2023). However, there may exist tensions between the efficiency of CP and algorithmic fairness, because the former desires a small prediction set, while the latter may necessitate larger sets for equal conditional coverage across all subgroups (Gibbs et al., 2025).

Conformal prediction with *equalized coverage* (Romano et al., 2020a) provides a pragmatic approach to the efficiency–fairness trade-off. This approach ensures that the target coverage level (e.g., 90%) is satisfied not only marginally over the entire population, but also conditionally on each protected group of interest. However, acquiring prediction sets with equalized coverage is challenging, as the number of all plausible groups of interest is exponential in the number of features. A straightforward enumeration is practically infeasible both statistically and computationally, especially on multi-dimensional (continuous) features. Indeed, Romano et al. (2020a) only takes each single feature as the condition of groups (e.g., a group defined by *Gender=Female*), which is an arguably insufficient representation of the entire space of groups.

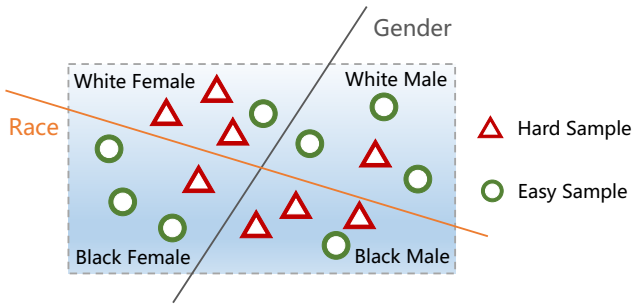


Figure 1: An illustrative example. The group space is divided into four parts by the feature *Race* and *Gender*. Hard samples (red triangles) are unfairly treated by the classifier, and easy samples (green circles) are normally treated. Note that a single feature (either Race or Gender) cannot discover the unfair subgroups (both have four triangles and four circles). Stronger expressiveness is desirable to capture the unfair subgroup "*White Female* or *Black Male*".

Later, Zhou & Sesia (2024) observe that algorithmic biases often concentrate on a minority of subgroups, and propose adaptively fair conformal prediction (AFCP) to identify these potentially disadvantaged subgroups. In a nutshell, AFCP computes the conditional coverage score for each discrete feature and selects the top-k sensitive features with a greedy strategy (where k is a hyperparameter). However, this group identification method still has limited expressiveness. For example, it cannot capture groups defined by a nonlinear combination of features, such as Exclusive OR (see the subgroup "*White Female* or *Black Male*" in Fig. 1). Additionally, AFCP is based on Naïve Bayes, which incurs a high computational cost and restricts its applicability to continuous features.

1.1 Our Contributions

In this paper, we propose a new group-fair conformal prediction method, **fair conformal prediction for representation-based groups** (FAREG), which accommodates both group expressiveness and time efficiency. Different from existing work (Romano et al., 2020a; Zhou & Sesia, 2024) which directly extracts groups from the raw input feature X, our approach encodes X into a latent representation Z via a mapping $Z = f(X)$, and learns unfair groups characterized by the low group coverage based on Z. The introduction of Z as a high-level representation of features strengthens the expressiveness of models, allowing a thorough exploration of groups. Meanwhile, we can enhance the interpretability by reconstructing input X from the encoding Z. To this end, we carefully design an encoder-decoder architecture and the optimization objective, based on the principle of variational inference.

In addition, we propose a *nonlinear* version of the conditional coverage metric WSC (Cauchois et al., 2021), namely WSC$^+$, aiming to evaluate the conditional coverage of unfairly treated groups more precisely. This allows users to check a conformal procedure and to compare multiple alternative conformal procedures.

The main contributions of this paper are summarized as follows. First, we propose a new conformal prediction method to enhance the expressiveness of unfair group identification. Second, we extend the traditional conditional coverage metric WSC to a nonlinear version WSC$^+$ for more accurate evaluation. Comprehensive experiments on both synthetic and real-world datasets confirm the effectiveness and efficiency of our proposed method.

2 Preliminary

For any natural number n, we write $[n] := \{1, \ldots, n\}$. We work with the most widely-used version of conformal prediction, i.e., *split* conformal prediction, where we assume a calibration set $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$ of i.i.d. (or simply exchangeable) observations sampled from an (unknown) distribution P_{XY}. In standard classification, $X_i \in \mathcal{X}$ represents the input feature from a feature space $\mathcal{X} \subseteq \mathbb{R}^d$ and $Y_i \in [L]$ is a categorical label. A given classifier $\hat{f}$ is trained (on a training set) to predict the conditional distribution $P(Y \mid X)$. Furthermore, X_{N+1} is a test instance with an unknown label

Y_{N+1} sampled by P_{XY}. CP constructs a prediction set $C(X_{N+1})$ for Y_{N+1} based on $\mathcal{D}$. The output $C(X_{N+1})$ guarantees marginal coverage at a user-specified level, i.e.,

$$\mathbb{P}[Y_{N+1} \in C(X_{N+1})] \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$ is a predefined miscoverage rate.

Typically, CP proceeds in three steps: (1) computing the predefined conformity score $V(x_i, y_i)$ for each sample $(x_i, y_i) \in \mathcal{D}$ using the predictive results of the classifier $\hat{f}$; (2) setting $(1-\alpha)(1+1/N)$-quantile score of $\mathcal{D}$ as a threshold $\hat{\eta}$; (3) constructing the prediction set $C_{\mathfrak{m}}(X_{N+1}, \mathcal{D}) := \{y \in [K] \,|\, V(X_i, y) \geq \hat{\eta}\}$, which is used as $C(X_{N+1})$ for X_{N+1}.

It can be shown that $C_{\mathfrak{m}}(X_{N+1}, \mathcal{D})$ meets the desirable *marginal coverage*. Intuitively, marginal coverage implies that the prediction set is guaranteed to contain the true label with the *average* $1-\alpha$ probability over the population. However, this guarantee is deemed to be insufficient, especially when miscoverage exhibits systematic bias, disproportionately affecting individuals belonging to groups characterized by certain features.

By contrast, *conditional coverage* requires $\mathbb{P}[Y_{N+1} \in C(X_{N+1}) \,|\, X_{N+1} = x] \geq 1 - \alpha$ for each $x \in \mathcal{X}$. This is much stronger as it demands correct coverage across all regions of the feature space, not just on average. However, achieving conditional coverage is impossible without imposing extra assumptions on the underlying distribution P_{XY} (such as the smoothness of P_{XY} (Cai et al., 2014; Lei & Wasserman, 2014) and strictly limiting the size of feature space $\mathcal{X}$ (Lee & Barber, 2021)). As these strong assumptions are often violated, conditional coverage is less meaningful in practice.

Equalized coverage (Romano et al., 2020a) represents a pragmatic compromise to ensure validity across *predefined* sample groups that need to be protected. Given a group $\mathcal{G} \subseteq \mathcal{X}$, it is required that

$$\mathbb{P}[Y_{N+1} \in C(X_{N+1}) \,|\, X_{N+1} \in \mathcal{G}] \geq 1 - \alpha \tag{1}$$

for all $\mathcal{G}$ of interest. In particular, these groups are typically related to some specific features called sensitive features.

However, the requirement for rigorous equalized coverage is localized, as algorithmic biases disproportionately affect only a minority of subgroups (Zhou & Sesia, 2024), as mentioned in Seciton 1. Therefore, AFCP further proposes adaptive equalized coverage based on equalized coverage, formalized by

$$\mathbb{P}[Y_{N+1} \in C(X_{N+1}) \,|\, X_{N+1} \in \hat{\mathcal{G}}] \geq 1 - \alpha, \tag{2}$$

where $\hat{\mathcal{G}}$ is adaptively selected corresponding to sensitive features. Eq. 2 indicates that $C(X_{N+1})$ is well-calibrated for the selected group $\hat{\mathcal{G}}$ defined by these sensitive features.

3 METHODOLOGY

This section presents FAREG, a learning-based method that adaptively identifies groups affected by algorithmic bias and adjusts their prediction sets to achieve equalized coverage while preserving high informativeness.

3.1 LEARNING REPRESENTATION-BASED GROUPS

Optimization Objective. For any feature $x \in \mathcal{X}$, we write its encoding $z = f(x) \in \mathcal{Z}$, where $\mathcal{Z}$ is a latent representation space. Intuitively, z denotes the latent representations of feature combinations of x. We introduce a random binary variable S and Z taking values in $\mathcal{Z}$ to formalize the membership of a group. Naturally, we consider a conditional distribution $P(S \,|\, Z)$ such that the probability of $x \in \hat{\mathcal{G}}$ for a group $\hat{\mathcal{G}}$ is equal to $\mathbb{P}(S = 1 \,|\, Z = f(x))$. Our goal is twofold: (1) to learn an encoding $Z = f(X)$ that is maximally informative about S and X, while (2) Z does not reveal the identity of any individual i in the sample (e.g., the calibration set).

We apply the *deep variational information bottleneck* (Deep VIB) method (Alemi et al., 2017). Specifically, for two random variables X and Y with the joint pdf (parameterized by θ), $p_\theta(x, y)$, $I(X, Y; \theta) = \int p_\theta(x, y) \log \frac{p_\theta(x,y)}{p_\theta(x)p_\theta(y)} \, dxdy$ denotes their mutual information. The optimization objective can be formalized as

$$\max I(Z, S; \theta) + I(Z, X; \theta) - \beta I(Z, i; \theta),$$

where i is a random variable to take any instance from the sample (e.g., in this paper, the calibration set $\mathcal{D}$) with a uniform distribution, θ is the model parameter, and β is a weight hyperparameter. (We abbreviate $I(Z,S), I(Z,X), I(Z,i)$ as I_1, I_2, I_3 for convenience.)

By introducing $q_\phi(s|z), q_\varphi(x|z), r(z)$ as the variational approximation to $p_\theta(s|z), p_\theta(x|z), p(z)$ in respective terms, we perform variational inference and obtain

$$I_1 + I_2 - \beta I_3 \geq \int p_\theta(x)p_\theta(s\,|\,x)p_\theta(z\,|\,x)\log q_\phi(s\,|\,z)\,\mathrm{d}x\,\mathrm{d}s\,\mathrm{d}z + \int p_\theta(x)p_\theta(z\,|\,x)\log q_\varphi(x\,|\,z)\,\mathrm{d}x\,\mathrm{d}z - \frac{\beta}{N}\sum_i \int p_\theta(z\,|\,x_i)\log\frac{p_\theta(z\,|\,x_i)}{r(z)}\,\mathrm{d}z.$$

(The details are given in Appendix A.1.)

In practice, we can approximate $p_\theta(x,s) = p_\theta(x)p_\theta(s|x)$ and $p_\theta(x)$ using the empirical distribution on the observations (e.g., the calibration set $\mathcal{D}$). As for $p_\theta(z|x)$, the reparameterization trick (Kingma & Welling, 2013) forces z to conform to a normal distribution which relies on x_i, and hence its deterministic function can be rewritten as $z = f(x,\epsilon)$ with an (auxiliary) noise variable ϵ.

Substituting all of these into the above equation, we obtain the following loss function

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{\tilde{z}\sim f(x_i,\epsilon)}[\log q_\phi(s_i\,|\,\tilde{z}) + \log q_\varphi(x_i\,|\,\tilde{z})] - \beta D_{\mathrm{KL}}(p_\theta(z\,|\,x_i)\|r(z))\right). \tag{3}$$

Intuitively, the expected log-likelihood $\mathbb{E}_{\tilde{z}\sim f(x_i,\epsilon)}[\log q_\phi(s_i\,|\,\tilde{z}) + \log q_\varphi(x_i\,|\,\tilde{z})]$ allows the encoding $\tilde{z}$ to predict s_i and regenerate x_i simultaneously, whereas the Kullback-Leibler (KL) divergence aims to compress the remaining useless information of $\tilde{z}$.

Instantiation. Eq. 3 suggests a natural design of the Encoder-Decoder architecture. In our method, the stochastic encoder with parameter θ has the form $p_\theta(z\,|\,x) = \mathcal{N}(z\,|\,f_\mu(x), f_\sigma(x))$, where $f_\mu(x)$ and $f_\Sigma(x)$ are two MLP networks to output the mean and variance of a normal distribution. We set $r(z)$ as a standard normal distribution $\mathcal{N}(0,1)$ and directly minimize the KL divergence term in Eq. 3 using the reparameterization trick.

We now concentrate on two decoders with parameters ϕ and φ. The instantiation of decoder with parameter φ is trivial. For the expected log-likelihood $\mathbb{E}_{\tilde{z}\sim f(x_i,\epsilon)}[\log q_\varphi(x_i\,|\,\tilde{z})]$ in Eq. 3, we utilize the standard Mean Squared Error (MSE) as the reconstruction loss (Kingma & Welling, 2013).

Decoder with parameter ϕ aims at predicting S, which indicates whether the sample X belongs to group $\hat{\mathcal{G}}$ or not. Assume a set of observations, e.g., the calibration set $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$. The distribution $P(S\,|\,X)$ can be viewed as a binary classifier h comprising an encoder with parameter θ and a decoder with the parameter ϕ. The result of h on $\mathcal{D}$ is a vector $\mathbf{s} = [s_1, \ldots, s_N] \in \{0,1\}^N$. Let $\hat{\mathcal{G}}_\mathbf{s} \subseteq \mathcal{D}$ denote the group determined by $\mathbf{s}$ on $\mathcal{D}$ and $\mathcal{H}$ be the family of all plausible h. We extend an inequality (Cauchois et al., 2021) to measure the deviation between the empirical coverage probability $\mathbb{P}_n$ on $\mathcal{D}$ and the oracle coverage probability $\mathbb{P}$.

Proposition 1. *Let the VC-dimension $VC(\mathcal{H}) \leq R$ and $\delta = |\hat{\mathcal{G}}_\mathbf{s}|/N$ be the proportion of $\hat{\mathcal{G}}_\mathbf{s}$ to the entire dataset. Then the gap between the empirical coverage probability $\mathbb{P}_n$ on the observations and the oracle coverage probability $\mathbb{P}$ is upper bounded, i.e., there exists some constant C_1 for all $\tau > 0$*

$$\sup_{h\in\mathcal{H}}\left\{|\mathbb{P}_n[Y \in C(X)\,|\,X \in \hat{\mathcal{G}}_\mathbf{s}] - \mathbb{P}[Y \in C(X)\,|\,X \in \hat{\mathcal{G}}_\mathbf{s}]|\right\} \leq C_1\sqrt{\frac{R\log N + \tau}{\delta N}}$$

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

Proposition 1 (cf. Appendix A.2 for proof) highlights two key directions for reducing the discrepancy between $\mathbb{P}_n$ and $\mathbb{P}$. First, a lower VC-dimension $VC(h)$ leads to a more precise estimation $\mathbb{P}_n$, implying that the classifier h should exhibit limited complexity. Second, the selected group must be sufficiently large to ensure reliable estimation.

We maximize the expected log-likelihood $\mathbb{E}_{\tilde{z}\sim f(x_i,\epsilon)}[\log q_\phi(s_i\,|\,\tilde{z})]$ in Eq. 3 via minimizing the expected empirical conditional coverage of the selected group $\hat{\mathcal{G}}$. The group $\hat{\mathcal{G}}$ on $\mathcal{D}$ is determined by

a random vector $\mathbf{S}$, sampled from a joint Bernoulli distribution $B = \prod_{i=1}^{N} \text{Bernoulli}(q_\phi(S_i = 1 \mid \tilde{z}))$. Hence, given $\mathcal{D}$, we formulate the following optimization problem :

$$\min_{\phi} \mathbb{E}_{\mathbf{S}\sim B}[\mathbb{P}_n[Y \in C(X) \mid X \in \hat{\mathcal{G}}_{\mathbf{S}}]] \quad \text{s.t.} \ \frac{1}{N}\sum_{i=1}^{N} q_\phi(S_i = 1 \mid \tilde{z}) \geq \delta. \tag{4}$$

In the above minimization problem, $\delta = |\hat{\mathcal{G}}_{\mathbf{s}}|/N$ denotes the the proportion of the selected group size to the whole dataset $\mathcal{D}$, and the decoder with parameter ϕ is a simple logistic regression model of the form $q_\phi(s \mid \tilde{z}) = \sigma(s \mid f_m(\tilde{z}))$, where σ is the sigmoid function and f_m is a MLP network.

To solve the constrained optimization problem, we employ the Projected Gradient Descent (PGD), an iterative optimization algorithm (Madry et al., 2017), to optimize the parameter ϕ. In each training step, PGD performs a gradient descent update and then projects the new point onto the feasible set to ensure all constraints are satisfied. Specifically, when the predictive distribution $q_\phi(s \mid \tilde{z})$ does not meet the constraint $\frac{1}{N}\sum_{i=1}^{N} q_\phi(S_i = 1 \mid \tilde{z}) \geq \delta$ after one back propagation process, we project it back onto the constraint-friendly space. Such a projection is equivalent to an ℓ_2 distance minimization problem. Let $q^*_\phi(s_1 \mid \tilde{z}) \geq \cdots \geq q^*_\phi(s_N \mid \tilde{z})$ be the descending order of $\{q_\phi(s_i \mid \tilde{z})\}_{i=1}^{N}$, and the projection results in

$$q'_\phi(s_i \mid \tilde{z}) = \min\left(1, q_\phi(s_i \mid \tilde{z}) + \frac{\omega}{2}\right), \tag{5}$$

where $\omega = (\delta - k - \sum_{i=k+1}^{N} q^*_\phi(s_i \mid \tilde{z}))/(N-k) \geq 0$, $k \in [N]$ is the greatest index to satisfy $q^*_\phi(s_k \mid \tilde{z}) + \omega/2 \geq 1$ and $q^*_\phi(s_{k+1} \mid \tilde{z}) + \omega/2 < 1$. (The details are given in Appendix A.3.)

Overall, we employ the empirical conditional coverage loss $\mathcal{L}_{\text{CC}}$, the reconstruction loss $\mathcal{L}_{\text{MSE}}$, and the KL divergence loss $\mathcal{L}_{\text{KL}}$ to replace the corresponding terms in Eq. 3, resulting in

$$\mathcal{L} = \mathcal{L}_{\text{CC}} + \mathcal{L}_{\text{MSE}} - \beta\mathcal{L}_{\text{KL}}. \tag{6}$$

3.2 Constructing the Adaptive Prediction Sets

After selecting the unfair group $\hat{\mathcal{G}}$, we proceed to construct the final prediction set with $\hat{\mathcal{G}}$. First, a standard conformal prediction set $C_{\text{m}}(X_{N+1}, \mathcal{D})$ is constructed using classic adaptive conformal prediction. Then, we perform T sampling of the vector $\mathbf{s}_t$ $(t \in [T])$ from the joint Bernoulli distribution B learned by models in Eq. 4. Each $\mathbf{s}_t$ defines a group $\hat{\mathcal{G}}_{\mathbf{s}_t}$, and such group is used as a calibration set to build a prediction set $C_{\text{m}}(X_{N+1}, \hat{\mathcal{G}}_{\mathbf{s}_t})$ as mentioned in Section 2. The final prediction set for Y_{N+1} is given by the union of all these sets:

$$C(X_{N+1}) = C_{\text{m}}(X_{N+1}, \mathcal{D}) \cup \bigcup_{t=1}^{T} C_{\text{m}}(X_{N+1}, \hat{\mathcal{G}}_{\mathbf{s}_t}). \tag{7}$$

Our approach FAREG is summarized in Algorithm 1. To analyze its time complexity, assume we have M test instances and the complexity of conducting classic conformal prediction is $\mathcal{O}(N+M)$. Then, training the model to select groups is $\mathcal{O}(EN(|\theta| + |\phi| + |\varphi|))$, where E is the number of epochs. For all M test instances, the time of selecting groups and constructing prediction sets is $\mathcal{O}(TN + TM)$. The overall complexity of our FAREG is $\mathcal{O}(EN(|\theta| + |\phi| + |\varphi|) + T(N + M))$, which is $\mathcal{O}(N + M)$, disregarding constant multiplicative factors. In contrast, the complexity of AFCP is $\mathcal{O}(N \log N + NM)$ (Zhou & Sesia, 2024).

The following result, proved in Appendix A.4, ensures that the prediction set $C(X_{N+1})$ generated by FAREG achieves adaptive equalized coverage (Eq. 2) over the selected group set $\{\hat{\mathcal{G}}_{\mathbf{s}_t}\}_{t=1}^{T}$.

Theorem 1. *If $\{(X_i, Y_i)\}_{i=1}^{N+1}$ are exchangeable, the prediction set $C(X_{N+1})$ and the selected group set $\{\hat{\mathcal{G}}_{\mathbf{s}_t}\}_{t=1}^{T}$ output by Algorithm 1 satisfy the adaptive equalized coverage defined in Eq. 2, and this guarantee still holds when the selected groups are defined by a more complex combination of features (e.g., non-linear) compared to AFCP.*

4 Experiments

4.1 Experimental Setup

Baselines. We select the classic CP method Marginal (Romano et al., 2020b) for classification, the initial CP method Partial (Romano et al., 2020a) considering equalized coverage, and the state-

Algorithm 1 The overall framework of FAREG.

Input: calibration dataset $\mathcal{D} = \{X_i, Y_i\}_{i=1}^N$; test instance with feature X_{N+1}; list of K sensitive features; pre-trained classifier $\hat{f}$; fixed rule to compute nonconformity scores; level $\alpha \in (0,1)$; selected group size proportion δ; hyperparameter β; sampling times T;

Output: prediction set $C(X_{N+1})$; selected group set $\{\hat{\mathcal{G}}_{\mathbf{s}_t}\}_{t=1}^T$.

```
Construct classic conformal prediction set C_m(X_{N+1}, D) based on the output of f̂;
for each batch do
    Calculate KL divergence loss L_KL with reparameterization trick;
    Sample z̃ ~ f(x, ε);
    Calculate conditional coverage loss L_CC and reconstruction loss L_MSE using z̃;
    Put all losses together in L as defined in Eq. 6;
    Update parameters θ, φ and ϕ via the gradient descent of L;
    if Σ_{i=1}^N q_φ(S_i = 1 | z̃) < δ · N then
        Project each q_φ(S_i = 1 | z̃) to satisfy minimum set constraint using Eq. 5;
    end if
end for
for t ∈ [T] do
    Sample s_t ~ B;                    ▷ B is a joint Bernoulli distribution mentioned in Eq. 4
    Construct C_m(X_{N+1}, Ĝ_{s_t});
end for
Construct prediction set C(X_{N+1}) following Eq. 7.
```

of-the-art method AFCP (Zhou & Sesia, 2024) as our baselines. The vanilla version of AFCP is designed to pick at most one sensitive feature (referred to as AFCP1). We also extend AFCP1 to select two sensitive features (referred to as AFCP2), given unreal, strong prior knowledge. Note that in real-world applications, it is typically unknown exactly how many features the unfair group may correspond to.

Evaluation Metrics. To evaluate the prediction sets $C(X_{N+1})$ produced by different CP methods, we use the coverage conditional on a specific group (referred to as Group Coverage), Average Coverage (viz., marginal coverage), and Average Size (viz., efficiency) as the metrics.

Additionally, we propose a new conditional coverage metric, viz., WSC$^+$, to capture groups defined by complicated (nonlinear) feature relationships. Traditional conditional coverage metric (Cauchois et al., 2021) considers the worst coverage over all slabs containing δ mass on the observations, which is defined as

$$\text{WSC}_n(C, \mathbf{v}) := \inf_{a<b} \left\{ \mathbb{P}_n(Y \in C(X) \,|\, a \le \mathbf{v}^T X \le b) \;\text{ s.t. }\; \mathbb{P}_n(a \le \mathbf{v}^T X \le b) \ge \delta \right\},$$

where $\mathbf{v} \in \mathbb{R}^d$ and $a < b \in \mathbb{R}$.

To strengthen the WSC metric, we replace the linear mapping $\mathbf{v}^T$ in the above equation with an arbitrary non-linear function π, giving rise to WSC$^+$, i.e.,

$$\text{WSC}_n^+(C, \pi) := \inf_{a<b} \left\{ \mathbb{P}_n(Y \in C(X) \,|\, a \le \pi(X) \le b) \;\text{ s.t. }\; \mathbb{P}_n(a \le \pi(X) \le b) \ge \delta \right\}. \tag{8}$$

Assume a quadratic function $\pi(\mathbf{x}) = \mathbf{x}^T\mathbf{W}\mathbf{x} + \mathbf{v}^T\mathbf{x}$, where $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{v} \in \mathbb{R}^d$. We uniformly draw 1,000 samples $\pi_j = \{\mathbf{W}_j, \mathbf{v}_j\}$ to compute the worst-slab coverage for each π_j on the test instances. Following Cauchois et al. (2021), we use the grid search to achieve the optimal a, b satisfying the desiderata as well. In this case, we have a lower bound for our metric WSC$^+$.

Proposition 2. *Let $\pi(\mathbf{x}) = \mathbf{x}^T\mathbf{W}\mathbf{x} + \mathbf{v}^T\mathbf{x}$ be a quadratic function and Π be a parameter space of π. Then, if C effectively provides conditional coverage at level $1-\alpha$, we have*

$$\textit{WSC}_n^+ = \inf_{\pi\in\Pi} \textit{WSC}_n^+(C, \pi) \ge 1 - \alpha - \mathcal{O}(1)\sqrt{\frac{\mathcal{O}(d^2)\log N}{\delta N}}. \tag{9}$$

The proof is given in Appendix A.5.

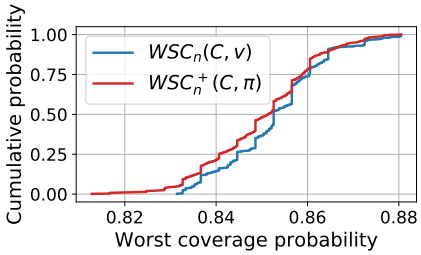


Figure 2: CDF of Conditional Coverage ($\delta = 0.5$), which plots the respective cumulative probability curves of different worst-slab coverage discovered by $\text{WSC}_n(C, \mathbf{v})$ and $\text{WSC}_n^+(C, \pi)$ over 1,000 samplings. The red curve is always above the blue curve, indicating that our $\text{WSC}_n^+(C, \pi)$ finds more groups with the poor coverage than $\text{WSC}_n(C, \mathbf{v})$.

METRIC	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.5$
WSC_n	0.616 (0.053)	0.748 (0.037)	0.793 (0.025)	0.822 (0.023)	0.842 (0.023)
WSC_n^+	0.582 (0.047)	0.674 (0.048)	0.750 (0.034)	0.800 (0.028)	0.829 (0.024)
IMP.	-5.52%	-9.89%	-5.42%	-2.68%	-1.54%

Table 1: The performance of WSC_n and WSC_n^+ metrics w.s.t. different δ. We repeat the experiment 10 times, and report the average results (the value in () is the standard deviation). Smaller coverage is better. Our metric WSC_n^+ performs better than WSC_n by up to 9.89% to mine the group with the minimum worst-slab coverage (defined in Eq. 9).

To demonstrate the advantages of the new metric WSC^+, we randomly draw the features $X \in [0, 1]^{10}$ from a uniform distribution and create a simple dataset for classification as described in Appendix B.1. Note that we define the group needed to be protected to satisfy $(X[0] \geq 0.1) \oplus (X[1] \geq 0.1)$ = True. We respectively plot the Cumulative Distribution Functions (CDF) of $\text{WSC}_n(C, v)$ and $\text{WSC}_n^+(C, \pi)$ over 1,000 samples π_j when $\delta = 0.5$ in Fig. 2, and observe that our $\text{WSC}_n^+(C, \pi)$ always reveals the groups with the worse coverage than that of $\text{WSC}_n(C, v)$, which can be attributed to representational capability of the nonlinear function π in $\text{WSC}_n^+(C, \pi)$.

Moreover, we also list the average results of two metrics, WSC_n and WSC_n^+, as δ increases over 10 repeated experiments in Table 1. Similar to Fig. 2, the minimum worst-slab coverage found by our WSC_n^+ is smaller than that found by WSC_n by up to 9.89%. As δ increases, the condition coverage tends to the marginal coverage (0.9), and the gap between WSC_n and WSC_n^+ narrows, as expected.

Implementations.[1] All the experiments are carried out on NVIDIA GeForce RTX 3090. We repeat each experiment 10 times and report the average to suppress randomness. We set $\delta = 0.5$ for WSC_n^+ by default. More implementation details, such as hyperparameters and training settings, are presented in Appendix B.1.

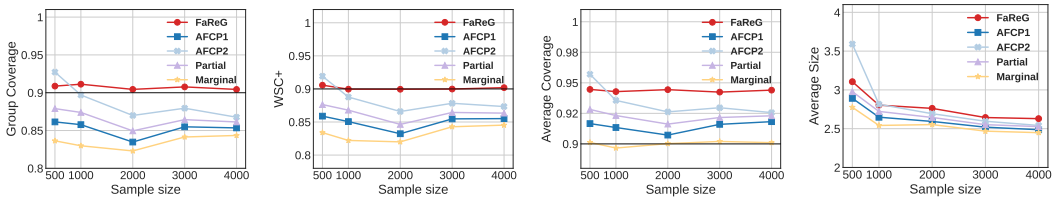
4.2 SYNTHETIC DATA

We evaluate our method on synthetic data designed to mimic a mental illness diagnosis scenario. The dataset includes six possible labels: Depression, Anxiety Disorders, Bipolar Disorder, Schizophrenia, Anorexia, and Post-Traumatic Stress Disorder (PTSD). Each sample contains four sensitive features—Age Group, Region, Gender, and Color—along with six non-sensitive features independently sampled from a uniform distribution within a value range $[0, 1]$. The sensitive features are generated as follows: (1) Gender is uniformly drawn from {*Female, Male* }; (2) Color is uniformly drawn from {*Red, Blue*}; (3) Age Group is drawn from {*Child, Youth, Middle, Elder*} with equal probability; (4) Region follows a fixed cyclical sequence: Asia, Europe, Africa, America, Oceania.

We then generate true labels Y for the dataset, where diagnosis is more challenging for a specific subgroup defined by the Exclusive NOR (XNOR) operation (cf. Appendix B.1). Specifically, we assume $X[0]$ is Color, $X[1]$ is Gender, and $X[2]$ is any non-sensitive feature, and define Y based solely on these three attributes. Through the label generation, we have the following subgroup $X[0] \odot X[1]$ = *True*: *Color=Red (True) & Gender=Female (True)* or *Color=Blue (False) & Gender=Male (False)*, simulating a real-world situation that algorithmic biases occur on this subgroup.

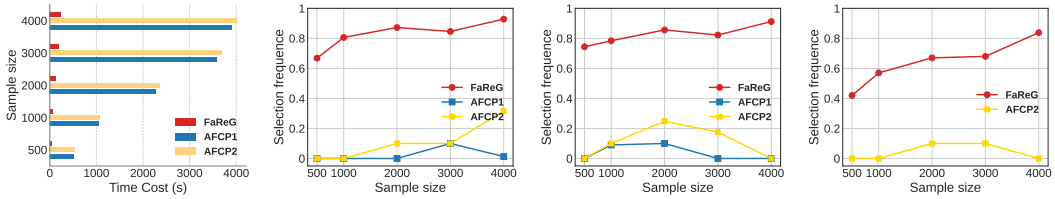
Fig. 3 depicts the results of conditional coverage, average coverage (marginal coverage), and average prediction set size (efficiency), respectively. For conditional coverage, Group Coverage is the coverage on the subgroups defined by XNOR operation as mentioned in data construction, and we

[1] An implementation of our approach can be accessed at the following anonymous link: https://github.com/Anonymity67543/FaReG.



(a) Group Coverage (b) WSC$_n^+$ (c) Average Coverage (d) Average Size

Figure 3: Performance of prediction sets produced by different CP methods on synthetic data w.r.t. the total number of training and calibration data instances. Only our FAREG achieves the ideal conditional coverage (0.9), and meanwhile, does not sacrifice too much information (set sizes) compared to baselines.



(a) Time Cost (b) Selects $X[0]$ (c) Selects $X[1]$ (d) Selects $X[0]\&X[1]$

Figure 4: Fig. (a) reports the running time of different CP methods with the increasing total number of training and calibration data instances. Fig. (b)–(d) are the results of the selection frequency of target features $X[0]$ and $X[1]$. As the sample size increases, our method becomes more consistent with target features.

compute WSC$_n^+$ on four predefined sensitive features. In Fig. 3a and 3b, our FAREG is the only one that always achieves valid coverage (greater than 0.9) for the targeted group with varying sample sizes. Although the conditional coverage of AFCP2 also exceeds 0.9 when the sample size reaches 500, as shown in Fig. 3d, it produces considerably larger prediction sets, which is less informative for decision-making.

In Fig. 4a, we compare the average running time of different CP methods over 10 repeated experiments, and FAREG significantly reduces the time cost, whose time complexity is linear in the number of data instances. This result is consistent with the analysis in Section 3.2.

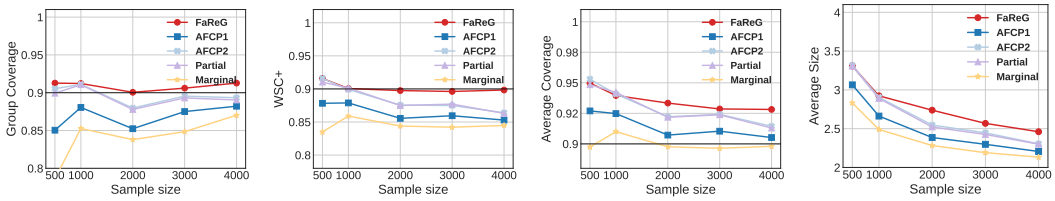
To determine which features are selected by our method, we analyze the predictive variable S and the reconstructed feature $\hat{X}$ by perturbing the latent representation Z, following the Beta-VAE approach (Higgins et al., 2017). Specifically, we impose a slight perturbation (e.g., ± 0.001) on each dimension of Z and identify the dimension that most influences S. Given this influential dimension and prior knowledge (as in AFCP2) that there are exactly two target features, we compute the change ratios for each dimension of $\hat{X}$ before and after perturbation, and select the two features with the top-2 maximum change ratios.

Figures 4b, 4c, and 4d respectively report the frequency of selecting $X[0]$ or $X[1]$ individually, and that of selecting both $X[0]$ and $X[1]$ simultaneously. The results demonstrate that our approach captures more target features than the baselines, and this advantage becomes more pronounced as the sample size increases.

Additionally, we present the results of parameter sensitivity and group visualization in Appendix B.2 and B.3, respectively.

4.3 NURSERY DATA

We evaluate our FAREG and baseline methods on the publicly available Nursery data (Rajkovic, 1989), originally constructed from a hierarchical decision model developed to rank applications for nursery schools. The dataset comprises 12,960 instances, each described by eight categorical features: Parents' occupation (3 levels, *Parent:*={*usual, pretentious, great-pret*}), Child's nursery (5



(a) Group Coverage (b) WSC_n^+ (c) Average Coverage (d) Average Size

Figure 5: Performance of prediction sets produced by different CP methods on the Nursery data w.r.t. the total number of training and calibration data instances. Only our FAREG achieves the ideal conditional coverage (0.9) and keeps most of the uncertainty information of sets.

levels), Family form (4 levels), Number of children (4 levels), Housing conditions (3 levels), Financial standing (2 levels, *Finance:*={*convenient, inconv*}), Social conditions (3 levels) and Health status (3 levels). The task is to classify applications into one of five priority ranks. We take all features into account (as sensitive features) except Housing conditions.

In data preprocessing, we strictly follow Zhou & Sesia (2024), and consider a group defined by *Parent=usual* & *Finance=inconv* or *Parent=pretentious* & *Finance=inconv*. To make the issue more interesting and control the degree of algorithmic bias, we corrupt the labels of instances in such a group by adding independent, uniform noise and rounding to the nearest integer (label) as similar as Zhou & Sesia (2024). This perturbation amplifies the intrinsic unpredictability of the group defined before, thereby increasing its vulnerability to algorithmic bias.

Fig. 5 presents the results. Our method consistently achieves the valid coverage under both conditional coverage metrics, i.e., Group Coverage and WSC_n^+, outperforming all baselines. Partial and AFCP2 perform better than the other CP methods, but FAREG still achieves superior results.

5 Related Work

Conformal Prediction (CP) has seen vigorous development in recent years (Vovk et al., 2005; Smith, 2024). Its applications span diverse domains, from image classification (Sadinle et al., 2019) and object detection (Teng et al., 2023) to large language models (Kumar et al., 2023).

Some CP work, building on the split conformal framework (Papadopoulos et al., 2002; Lei et al., 2018), introduces advanced nonconformity scores to ensure valid marginal coverage on the empirical data distribution. For example, Romano et al. (2019) gives a nonconformity score based on quantile regression, while Romano et al. (2020b) and Angelopoulos et al. (2020) design nonconformity scores for classification. Additionally, Hoff (2023) proposes a nonconformity score to achieve Bayes optimal coverage.

Another line of work has explored various notions of equalized coverage (Romano et al., 2020a) and empirically evaluated the corresponding conformal predictors in real-world applications (Lu et al., 2022). For regression tasks, Wang et al. (2023) guarantees equal coverage rates across more fine-grained groups on continuous features, and Liu et al. (2022) propose to learn a real-valued quantile function with respect to sensitive features. They address a distinct notion of equalized coverage tailored to continuous outcomes. In classification, label-conditional coverage is a common alternative to equalized coverage (Vovk et al., 2003; Löfström et al., 2015; Ding et al., 2023). This work defines the groups to be protected based on the label Y_{N+1}, instead of the features X_{N+1}. Jung et al. (2022) and Gibbs et al. (2025) adopt group-conditional coverage, which is analogous to equalized coverage, to improve prediction sets. Different from the previous work, our approach FAREG can adaptively identify unfairly treated groups without the assumption that such groups are pre-defined. AFCP (Zhou & Sesia, 2024) develops an algorithm to construct CP sets with valid equalized coverage for adaptively selected groups, which establishes the current state-of-the-art for equalized coverage tasks.

6 CONCLUSION

In this paper, we propose FAREG, a fair conformal prediction method that learns latent groups to achieve adaptive equalized coverage. By leveraging a variational encoder-decoder to discover subgroups with poor coverage in a high-level feature space, our approach captures complex algorithmic biases that linear methods may neglect. We also propose WSC^+, a nonlinear metric for evaluating the conditional coverage of unfair groups more accurately. Extensive experiments confirm that FAREG efficiently offers stronger fairness guarantees, showing a more expressive and practical path toward fair, reliable conformal inference.

Limitations. The enhanced expressivity of representation-based groups may sacrifice model interpretability partially, compared to groups explicitly defined on manifest features. However, the encoder-decoder structure compensates this shortcoming well via reconstructing the input X, which is empirically confirmed by Section 4.2 and Appendix B.3.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of our work, we have made our source code publicly available through a public anonymous repository. All experimental details, including dataset partitions, hyperparameter configurations, and model details, are fully documented in Appendix B.1. We are confident that these materials provide the necessary information to replicate our findings.

USAGE OF LLMS

Large Language Models (LLMs) were utilized exclusively as writing assistants to enhance the linguistic quality of this manuscript, focusing on improving clarity, grammar, and readability. Their involvement was strictly limited to this editorial function. LLMs played no role in any substantive research components, including conceptualization, experimental design, data analysis, interpretation of results, or scientific content creation. All intellectual contributions, methodological developments, findings, and conclusions originate solely from the authors.

REFERENCES

Shams Forruque Ahmed, Md Sakib Bin Alam, Maruf Hassan, Mahtabin Rodela Rozbu, Taoseef Ishtiak, Nazifa Rafa, M Mofijur, ABM Shawkat Ali, and Amir H Gandomi. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11):13521–13617, 2023.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017. URL `https://arxiv.org/abs/1612.00410`. arXiv preprint arXiv:1612.00410, accepted to ICLR 2017.

Ahmed S Almasoud and Jamiu Adekunle Idowu. Algorithmic fairness in predictive policing. *AI and Ethics*, 5(3):2323–2337, 2025.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Richard A Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *Sociological Methods & Research*, 53(4):1629–1675, 2024.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

T Tony Cai, Mark Low, and Zongming Ma. Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association*, 109(507):1054–1070, 2014.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of machine learning research*, 22 (81):1–42, 2021.

Sanjiv Das, Richard Stanton, and Nancy Wallace. Algorithmic fairness. *Annual Review of Financial Economics*, 15(1):565–593, 2023.

Ketki V Deshpande, Shimei Pan, and James R Foulds. Mitigating demographic bias in ai-based resume filtering. In *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization*, pp. 268–275, 2020.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf008, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Peter Hoff. Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2):901–928, 2023.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Paulin Kamuangu. A review on financial fraud detection using ai and machine learning. *Journal of Economics, Finance, and Accounting Studies*, 6(1):67, 2024.

Simarjeet Kaur, Jimmy Singla, Lewis Nkenyereye, Sudan Jha, Deepak Prashar, Gyanendra Prasad Joshi, Shaker El-Sappagh, Md Saiful Islam, and SM Riazul Islam. Medical diagnostic systems using artificial intelligence (ai) algorithms: principles and perspectives. *Ieee Access*, 8:228049–228069, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Yonghoon Lee and Rina Barber. Distribution-free inference for regression: discrete, continuous, and in between. *Advances in Neural Information Processing Systems*, 34:7448–7459, 2021.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Meichen Liu, Lei Ding, Dengdeng Yu, Wulong Liu, Linglong Kong, and Bei Jiang. Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, 35:11561–11572, 2022.

Tuve Löfström, Henrik Boström, Henrik Linusson, and Ulf Johansson. Bias reduction through conditional conformal prediction. *Intelligent Data Analysis*, 19(6):1355–1375, 2015.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12008–12016, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

Vladislav Rajkovic. Nursery. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5P88W.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2020a.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020b.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Ralph C Smith. *Uncertainty quantification: theory, implementation, and applications*. SIAM, 2024.

Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive inference with feature conformal prediction. In *The Eleventh International Conference on Learning Representations*, 2023.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Fangxin Wang, Lu Cheng, Ruocheng Guo, Kay Liu, and Philip S Yu. Equal opportunity of coverage in fair regression. *Advances in Neural Information Processing Systems*, 36:7743–7755, 2023.

Yanfei Zhou and Matteo Sesia. Conformal classification with equalized coverage for adaptively selected groups. *Advances in Neural Information Processing Systems*, 37:108760–108823, 2024.

A TECHNICAL PROOFS

A.1 VARIATIONAL INFERENCE

As mentioned in Section 3.1, our optimization objective is as follows,

$$\max I(Z,S) + I(Z,X) - \beta I(Z,i).$$

First of all, we consider $I(Z,S)$ and

$$I(Z,S) = \int p_\theta(s,z) \log \frac{p_\theta(s,z)}{p_\theta(s)p_\theta(z)} \,\mathrm{d}s\,\mathrm{d}z = \int p_\theta(s,z) \log \frac{p_\theta(s\,|\,z)}{p_\theta(s)} \,\mathrm{d}s\,\mathrm{d}z. \tag{10}$$

Since the KL divergence between two conditional probability distribution $p_\theta(s\,|\,z)$ and $q_\phi(s\,|\,z)$ is non-negative, we have

$$D_{\mathrm{KL}}(p_\theta(s\,|\,z)\|q_\phi(s\,|\,z)) \geq 0 \Rightarrow \int p_\theta(s,z) \log p_\theta(s\,|\,z)\,\mathrm{d}s \geq \int p_\theta(s,z) \log q_\phi(s\,|\,z)\,\mathrm{d}s,$$

where $q_\phi(s\,|\,z)$ is a variational approximation to the intractable distribution $p_\theta(s\,|\,z)$.

Plugging the above inequality into Eq. 10, we obtain

$$\begin{aligned} I(Z,S) &\geq \int p_\theta(s,z) \log \frac{q_\phi(s\,|\,z)}{p_\theta(s)} \,\mathrm{d}s\,\mathrm{d}z \\ &= \int p_\theta(s,z) \log q_\phi(s\,|\,z) \,\mathrm{d}s\,\mathrm{d}z + \int p_\theta(s) \log p_\theta(s)\,\mathrm{d}s \\ &\geq \int p_\theta(s,z) \log q_\phi(s\,|\,z) \,\mathrm{d}s\,\mathrm{d}z, \end{aligned} \tag{11}$$

where the second inequality is derived by the non-negativity of entropy.

Since $S \perp\!\!\!\perp Z\,|\,X$ holds, we have

$$p_\theta(s,z) = \int p_\theta(x,s,z)\,\mathrm{d}x = \int p_\theta(x)p_\theta(s\,|\,x)p_\theta(z\,|\,x)\,\mathrm{d}x.$$

Hence, we get

$$I(Z,S) \geq \int p_\theta(x)p_\theta(s\,|\,x)p_\theta(z\,|\,x) \log q_\phi(s\,|\,z)\,\mathrm{d}x\,\mathrm{d}s\,\mathrm{d}z. \tag{12}$$

Similar to Eq. 11, we also have

$$\begin{aligned} I(Z,X) &\geq \int p_\theta(x,z) \log q_\varphi(x\,|\,z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \int p_\theta(x)p_\theta(z\,|\,x) \log q_\varphi(x\,|\,z)\,\mathrm{d}x\,\mathrm{d}z. \end{aligned} \tag{13}$$

As for $I(Z,i)$, we have

$$\begin{aligned} I(Z,i) &= \sum_i \int p_\theta(z\,|\,i)p_\theta(i) \log \frac{p_\theta(z\,|\,i)}{p_\theta(z)} \,\mathrm{d}z \\ &= \frac{1}{N} \sum_i \int p_\theta(z\,|\,x_i) \log \frac{p_\theta(z\,|\,x_i)}{p_\theta(z)} \,\mathrm{d}z \\ &\leq \frac{1}{N} \sum_i \int p_\theta(z\,|\,x_i) \log \frac{p_\theta(z\,|\,x_i)}{r(z)} \,\mathrm{d}z, \end{aligned} \tag{14}$$

where $r(z)$ is a variational approximation to the posterior distribution $p_\theta(z)$. We usually set $r(z)$ as a standard normal distribution $\mathcal{N}(0,1)$ in practice.

Combining Eq. 12 with Eq. 13 and Eq. 14, we obtain

$$\begin{aligned} I(Z,S) + I(Z,X) - \beta I(Z,i) \geq & \int p_\theta(x)p_\theta(s\,|\,x)p_\theta(z\,|\,x) \log q_\phi(s\,|\,z)\;\mathrm{d}x\,\mathrm{d}s\,\mathrm{d}z \\ & + \int p_\theta(x)p_\theta(z\,|\,x) \log q_\varphi(x\,|\,z)\;\mathrm{d}x\,\mathrm{d}z - \frac{\beta}{N} \sum_i \int p_\theta(z\,|\,x_i) \log \frac{p_\theta(z\,|\,x_i)}{r(z)} \,\mathrm{d}z. \end{aligned}$$

With Monte Carlo sampling, we use the empirical dataset on $\{X_i, S_i, Y_i\}_{i=1}^N$ to estimate $p_\theta(x)p_\theta(s\,|\,x)$ and $p_\theta(x)$, where S_i is computed by minimizing the conditional coverage of groups defined by $\mathbf{S} = \{s_1, \ldots, s_N\}$, i.e., $\mathbb{P}_n[Y_i \in C(X_i)\,|\,X_i \in \hat{\mathcal{G}}_\mathbf{S}]$ on $\{X_i, S_i, Y_i\}_{i=1}^N$. We leverage the reparameterization trick (Kingma & Welling, 2013) as mentioned in Section 3.1, and finally obtain

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{\tilde{z}\sim f(x_i,\epsilon)}[\log q_\phi(s_i\,|\,\tilde{z}) + \log q_\varphi(x_i\,|\,\tilde{z})] - \beta D_{\mathrm{KL}}(p_\theta(z\,|\,x_i)\|r(z))\right).$$

A.2 Proof of Proposition 1

Proof. We first present a technical lemma, where $P_n h = \frac{1}{N}\sum_{i=1}^N h(X_i)$ and $Ph = \int h(x)dP(x)$, given an observed dataset $\{X_i, Y_i\}_{i=1}^N$.

Lemma 1 (Boucheron et al. (2005))**.** *There exists a numerical constant C_1 such that for any $\tau > 0$,*

$$|P_n h - Ph| \le C_1\left[\sqrt{\min\{P_n h, Ph\}\frac{VC(h)\log N + \tau}{N}} + \frac{VC(h)\log N + \tau}{N}\right]$$

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

By this Lemma, we have

$$\begin{aligned}&|P_n(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s}) - P(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s})| \\ &\quad\le C_1\left[\sqrt{\min\{P_n(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s}), P(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s})\}\tfrac{VC(h)\log N+\tau}{N}} + \tfrac{VC(h)\log N+\tau}{N}\right].\end{aligned} \tag{15}$$

Similarly, we get

$$\begin{aligned}&|P_n(X \in \hat{\mathcal{G}}_\mathbf{s}) - P(X \in \hat{\mathcal{G}}_\mathbf{s})| \\ &\quad\le C_2\left[\sqrt{\min\{P_n(X \in \hat{\mathcal{G}}_\mathbf{s}), P(X \in \hat{\mathcal{G}}_\mathbf{s})\}\tfrac{VC(h)\log N+\tau}{N}} + \tfrac{VC(h)\log N+\tau}{N}\right].\end{aligned} \tag{16}$$

Then, it remains to show that

$$\begin{aligned}&|P_n(Y \in C(X)\,|\,X \in \hat{\mathcal{G}}_\mathbf{s}) - P(Y \in C(X)\,|\,X \in \hat{\mathcal{G}}_\mathbf{s})| \\ &\quad= \left|\frac{P_n(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s})}{P_n(X \in \hat{\mathcal{G}}_\mathbf{s})} - \frac{P(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s})}{P(X \in \hat{\mathcal{G}}_\mathbf{s})}\right|.\end{aligned}$$

Let $a = P_n(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s}), b = P(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s}), c = (P_n - P)(X \in \hat{\mathcal{G}}_\mathbf{s})$ and $d = P(X \in \hat{\mathcal{G}}_\mathbf{s})$. We can derive $b \le d$, and observe that

$$\left|\frac{a}{c+d} - \frac{b}{d}\right| \le \left|\frac{a}{c+d} - \frac{b-c}{c+d}\right| \le \frac{|a-b|}{c+d} + \frac{|c|}{c+d}. \tag{17}$$

Substitute Eq. 15 and Eq. 16 into Eq. 17, and use $\delta = P_n(X \in \hat{\mathcal{G}}_\mathbf{s})$, we obtain

$$\begin{aligned}&\left|\frac{P_n(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s})}{P_n(X \in \hat{\mathcal{G}}_\mathbf{s})} - \frac{P(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s})}{P(X \in \hat{\mathcal{G}}_\mathbf{s})}\right| \\ &\le \frac{|P_n(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s}) - P(Y \in C(X), X \in \hat{\mathcal{G}}_\mathbf{s})|}{P_n(X \in \hat{\mathcal{G}}_\mathbf{s})} - \frac{|P_n(X \in \hat{\mathcal{G}}_\mathbf{s}) - P(X \in \hat{\mathcal{G}}_\mathbf{s})|}{P_n(X \in \hat{\mathcal{G}}_\mathbf{s})} \\ &\le C_3\left[\sqrt{\frac{VC(h)\log N + \tau}{\delta N}} + \frac{VC(h)\log N + \tau}{\delta N}\right],\end{aligned}$$

which completes the proof. □

A.3 Optimization Process of Eq. 5

The projection operation of the PGD algorithm described in Section 3.1 requires solving the following optimization to minimize the ℓ_2 distance:

$$\min_{v_1,\ldots,v_n} \sum_{i=1}^{N}(v_i - u_i)^2 \quad \text{s.t. } \sum_{i=1}^{N} v_i \geq \delta, \quad v_i \in [0,1] \;\; i = 1,\ldots,N, \tag{18}$$

where $u_1,\ldots,u_N$ are given and $u_i \in [0,1]$ holds for each $i \in [N]$.

With the above constraints, we compute the Lagrangian as

$$\mathcal{L}(v_i;\lambda_i,\mu_i,\omega) = \sum_{i=1}^{N}(v_i - u_i)^2 + \sum_{i=1}^{N}\lambda_i(-v_i) + \sum_{i=1}^{N}\mu_i(v_i - 1) + \omega(\delta - \sum_{i=1}^{N} v_i),$$

where $\{\lambda_i\}_{i=1}^N, \{\mu_i\}_{i=1}^N$ and ω are the Lagrange multipliers. Let the partial derivatives vanish, and we have

$$\frac{\partial \mathcal{L}}{\partial v_i} = 2(v_i - u_i) - \lambda_i + \mu_i - \omega = 0 \Rightarrow 2(v_i - u_i) = \lambda_i - \mu_i + \omega$$

For the complementary relaxation conditions, there are four different cases:

- If $v_i = 0$, constraint $v_i \geq 0$ is activated and we have $\lambda_i \geq 0, \mu_i = 0$;
- If $v_i = 1$, constraint $v_i \leq 1$ is activated and we have $\mu_i \geq 0, \lambda_i = 0$;
- If $0 < v_i < 1$, we have $\mu_i = \lambda_i = 0$ and then $v_i = u_i + \omega/2$;
- If $\sum v_i > \delta$, constraint $\sum v_i \geq \delta$ is not activated and then $\omega = 0$; otherwise, $\omega \geq 0$.

When $\sum u_i \geq \delta$, we have $v_i = u_i$, which is an optimal solution to the minimization problem in Eq. 18.

When $\sum u_i < \delta$, let $v_i = \min(1, u_i + \omega/2)$, where $\omega \geq 0$ and $\sum_{i=1}^N \min(1, u_i + \omega/2) \geq \delta$. In this case, we resort $\{v_i\}_{i=1}^N$ in descending order, i.e., $v_{(1)} \geq v_{(2)} \geq \cdots \geq v_{(N)}$. Let $k \in [N]$ is the greatest index to satisfy $v_{(k)} + \omega/2 \geq 1$ and $v_{(k+1)} + \omega/2 < 1$. Then, constraint $\sum v_i = \delta$ can be written as

$$k \cdot 1 + \sum_{i=k+1}^{N} (v_{(i)} + \omega/2) = \delta.$$

Hence, we obtain

$$\omega = \frac{2(\delta - k - \sum_{i=k+1}^{N} v_{(i)})}{N - k}.$$

In practice, we can compute k and ω via traversing the value of k from maximum N to minimum 1.

A.4 Proof of Theorem 1

Proof. When making the similar assumption as Theorem 1 in AFCP (Zhou & Sesia, 2024), for each group $\hat{\mathcal{G}}_{\mathbf{s}} \in \{\hat{\mathcal{G}}_{\mathbf{s}_t}\}_{t=1}^T$, we can substitute $X_{N+1} \in \hat{\mathcal{G}}_{\mathbf{s}}$ for $\phi(X_{N+1}, \hat{A}(X_{N+1}))$ and $X_{N+1} \in \hat{\mathcal{G}}_{\mathbf{s}}^o$ for $\phi(X_{N+1}, \hat{A}^o(X_{N+1}))$ as conditions, where $\hat{\mathcal{G}}_{\mathbf{s}}^o$ is an imaginary oracle group. Then, according to Theorem 1 (Zhou & Sesia, 2024), we have

$$\mathbb{P}[Y_{N+1} \in C(X_{N+1}) \,|\, X_{N+1} \in \hat{\mathcal{G}}_{\mathbf{s}}] \geq 1 - \alpha.$$

AFCP assumes that the group selection algorithm can always achieve the oracle group $\hat{\mathcal{G}}_{\mathbf{s}}^o$, which means that the algorithm must have enough expressiveness to include $\hat{\mathcal{G}}_{\mathbf{s}}^o$ into the candidate group space. However, this necessary condition could be violated, as AFCP's candidate group space is limited to linear groups defined by individual features. In contrast, our method, FAREG, employs a more expressive model that extends its candidate group space into the nonlinear realm. Consequently, the guarantee for FAREG remains valid for groups defined by complex, nonlinear feature combinations.

Next, we formally analyze the expressiveness of AFCP and our FAREG based on the VC-dimension. As described in Section 1, AFCP computes the group coverage scores for each feature and greedily

picks the most sensitive feature with the lowest group coverage score. The essence of such a process is a decision stump dividing all features into two parts (sensitive or not sensitive) using a threshold, and thus its VC-dimension is 2. In contrast, based on established theory (Shalev-Shwartz & Ben-David, 2014), the VC-dimension of FAREG scales with its parameter size M, i.e.,

$$\mathrm{VC}(\mathrm{AFCP}) = 2, \quad \mathrm{VC}(\mathrm{FAREG}) = \mathcal{O}(M).$$

Hence, the VC-dimension of our FAREG is typically far larger than that of AFCP, indicating the stronger expressiveness of our method, i.e., our candidate group space serves as a superset of AFCP's candidate group space.

□

A.5 Proof of Proposition 2

Proof. According to Proposition 1 and the definition of WSC_n^+ (Eq. 8), we obtain

$$\sup_{\pi\in\Pi}\{|\mathrm{WSC}_n^+(C,\pi) - \mathbb{P}(Y \in C(X) \,|\, a \le \pi(X) \le b)|\} \le \mathcal{O}(1)\sqrt{\frac{VC(\Pi)\log N}{\delta N}}$$

by omitting τ. Then, we eliminate the absolute value as

$$-\mathcal{O}(1)\sqrt{\frac{VC(\pi)\log N}{\delta N}} \le \mathrm{WSC}_n^+(C,\pi) - \mathbb{P}(Y \in C(X) \,|\, a \le \pi(X) \le b) \le \mathcal{O}(1)\sqrt{\frac{VC(\pi)\log N}{\delta N}},$$

which holds for all $\pi \in \Pi$. Hence, if $\mathbb{P}(Y \in C(X) \,|\, a \le \pi(X) \le b) = 1-\alpha$, we can observe

$$\mathrm{WSC}_n^+(C,\pi) \ge \mathbb{P}(Y \in C(X) \,|\, a \le \pi(X) \le b) - \mathcal{O}(1)\sqrt{\frac{VC(\pi)\log N}{\delta N}}$$

for any $\pi \in \Pi$.

Next, we only need to prove $\mathcal{O}(d^2) \ge VC(\pi)$. Recall that $VC(\pi)$ denotes the VC-dimension of the binary classifier π, and $\pi = \mathbf{x}^T\mathbf{W}\mathbf{x} + \mathbf{v}^T\mathbf{x}$ is a quadratic function, where $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{v} \in \mathbb{R}^d$. Therefore, the VC-dimension of π is equal to the dimension of its expanded feature space $\mathcal{M} = d(d+1)/2 + d$, i.e., $\mathcal{O}(d^2)$, which completes the proof. □

B Further Experiment Details

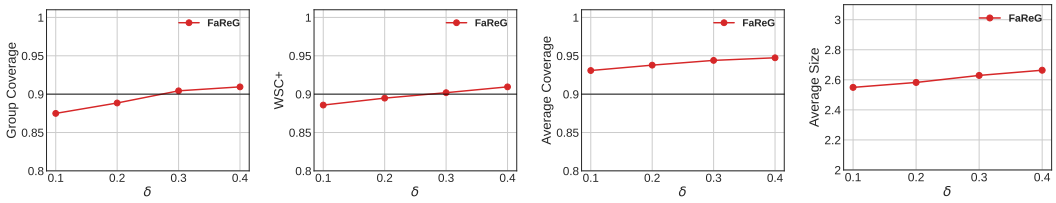
B.1 Dataset Construction and Hyperparameters

Table 2: Hyperparameters of FAREG.

Dataset	Synthetic Data	Nursery Data
Model	MLP	MLP
Number of Layers	3	3
Hidden dimension	[64,32]	[64,32]
Epoch	2000	800
Batch Size	500	500
Learning Rate	0.001	0.01
β	2.0	0.1
δ	0.3	0.1
T	20	100

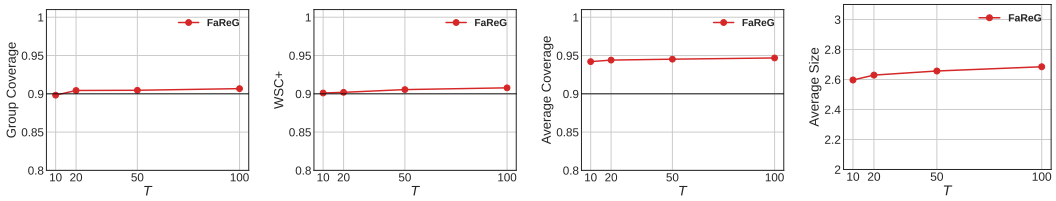
For the dataset we use to evaluate two metrics in Section 4.1, only $X[0], X[1]$, and $X[2]$ influence the label Y and we define the conditional distribution $P(Y \,|\, X)$ as

$$P(Y \,|\, X) = \begin{cases} \left(\frac{1}{3},\frac{1}{3},\frac{1}{3},0,0,0\right), & \text{if } (X[0]\ge 0.1) \oplus (X[1]\ge 0.1) \text{ and } X[2] < 0.5, \\ \left(0,0,0,\frac{1}{3},\frac{1}{3},\frac{1}{3}\right), & \text{if } (X[0]\ge 0.1) \oplus (X[1]\ge 0.1) \text{ and } X[2] \ge 0.5, \\ (1,0,0,0,0,0), & \text{if not } (X[0]\ge 0.1) \oplus (X[1]\ge 0.1) \text{ and } X[2] < \frac{1}{6}, \\ (0,1,0,0,0,0), & \text{if not } (X[0]\ge 0.1) \oplus (X[1]\ge 0.1) \text{ and } \frac{1}{6} \le X[2] \le \frac{2}{6}, \\ \vdots & \\ (0,0,0,0,0,1), & \text{if not } X[0] = (X[0]\ge 0.1) \oplus (X[1]\ge 0.1) \text{ and } \frac{5}{6} \le X[2] \le 1. \end{cases}$$



(a) Group Coverage (b) WSC_n^+ (c) Average Coverage (d) Average Size

Figure 6: Performance of prediction sets produced by our FAREG on synthetic data w.r.t. the selected group size proportion δ.



(a) Group Coverage (b) WSC_n^+ (c) Average Coverage (d) Average Size

Figure 7: Performance of prediction sets produced by our FAREG on synthetic data w.r.t. the sampling times T.

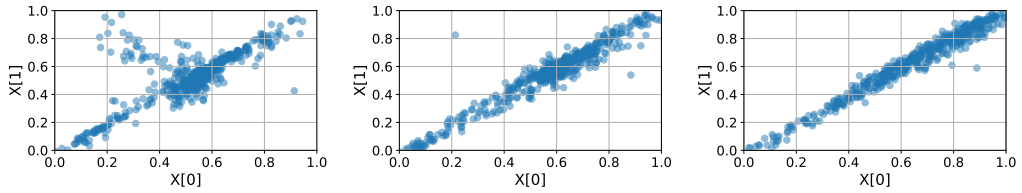
For the classification models as the input of conformal prediction, we strictly follow the settings in (Zhou & Sesia, 2024) on both synthetic and real-world data. To train FAREG to mine unfair groups, we randomly split the calibration set $\mathcal{D}$ into the training set and validation set with the ratio 5:5. We list the hyperparameters of FAREG in Table 2. Note that we use the same network structure for encoders and decoders, i.e., a simple 3-layer MLP, which is consistent with Proposition 1. Since there are three optimization objectives in Eq. 6, which may conflict with each other to some extent, we divide the training into two stages. At the first stage, we train the encoder with the parameter θ and the decoder with the parameter ϕ by fixing the decoder with the parameter φ in practice, i.e., the first term $\mathcal{L}_{CC}$ and third term $\mathcal{L}_{KL}$ in Eq. 6. Then, we use $\mathcal{L}_{MSE}$ to reconstruct X based on Z at the second stage.

Recall from Section 4.2 that Color is denoted as $X[0]$, Gender is denoted as $X[1]$, and the first standard feature is denoted as $X[2]$. The conditional distribution of $Y \mid X$ is determined by a simple decision tree, where only $X[0]$, $X[1]$, and $X[2]$ provide valuable predictive information for Y, formulated as follows,

$$P(Y \mid X) = \begin{cases} \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0\right), & \text{if } X[0] = \textit{Red} \text{ and } X[1] = \textit{Female} \text{ and } X[2] < 0.5, \\ \left(0, 0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), & \text{if } X[0] = \textit{Red} \text{ and } X[1] = \textit{Female} \text{ and } X[2] \geq 0.5, \\ \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0\right), & \text{if } X[0] = \textit{Blue} \text{ and } X[1] = \textit{Male} \text{ and } X[2] < 0.5, \\ \left(0, 0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), & \text{if } X[0] = \textit{Blue} \text{ and } X[1] = \textit{Male} \text{ and } X[2] \geq 0.5, \\ (1, 0, 0, 0, 0, 0), & \text{if } X[0] = \textit{Red} \text{ and } X[1] = \textit{Male} \text{ and } X[2] < \frac{1}{6}, \\ (0, 1, 0, 0, 0, 0), & \text{if } X[0] = \textit{Red} \text{ and } X[1] = \textit{Male} \text{ and } \frac{1}{6} \leq X[2] \leq \frac{2}{6}, \\ \vdots & \\ (0, 0, 0, 0, 0, 1), & \text{if } X[0] = \textit{Red} \text{ and } X[1] = \textit{Male} \text{ and } \frac{5}{6} \leq X[2] \leq 1, \\ (1, 0, 0, 0, 0, 0), & \text{if } X[0] = \textit{Blue} \text{ and } X[1] = \textit{Female} \text{ and } X[2] < \frac{1}{6}, \\ \vdots & \\ (0, 0, 0, 0, 0, 1), & \text{if } X[0] = \textit{Blue} \text{ and } X[1] = \textit{Female} \text{ and } \frac{5}{6} \leq X[2] \leq 1. \end{cases}$$

B.2 PARAMETER SENSITIVITY

In this section, we investigate the sensitivity of two key parameters: the selected group size proportion δ and the number of sampling iterations T. The results for δ and T are presented in Fig. 6



(a) $|\hat{\mathcal{G}}_{\mathbf{s}}|/N = 0.31$ (b) $|\hat{\mathcal{G}}_{\mathbf{s}}|/N = 0.39$ (c) $|\hat{\mathcal{G}}_{\mathbf{s}}|/N = 0.47$

Figure 8: The visualization results of reconstruction $\hat{X}$ when $|\hat{\mathcal{G}}_{\mathbf{s}}|/N$ increases. The latent representation z generally captures an XNOR relationship between $X[0]$ and $X[1]$.

and Fig. 7, respectively. Overall, the metrics Group Coverage and WSC_n^+ show relative insensitivity to the number of sampling iterations T, as illustrated in Fig. 7a and 7b. In contrast, both Group Coverage and WSC_n^+ increase with the proportion δ of the selected group size relative to the entire dataset. This trend empirically provides implicit support for Proposition 1.

B.3 Group Visualization

To analyze how features $X[0]$ and $X[1]$ contribute to group membership in $\hat{\mathcal{G}}_{\mathbf{s}}$, we perturb the encoding z and examine the resulting reconstructions $\hat{X}$. Taking the sample size of 4000 on the synthetic dataset as an example, we randomly select one run from 10 repeated trials and add perturbations of $+0.003$ and $+0.006$ to the fourth dimension of z, respectively. The reconstructed features $\hat{X}$ are visualized in Fig. 8.

Fig. 8a shows that, without perturbation, the latent representation z generally captures an XNOR relationship between $X[0]$ and $X[1]$, indicating that the encoder effectively filters out irrelevant feature information. After applying perturbations (see Fig. 8b and 8c), the XNOR pattern becomes more pronounced as $|\hat{\mathcal{G}}_{\mathbf{s}}|/N$ increases, revealing a positive correlation between $X[0] \odot X[1]$ and membership in $\hat{\mathcal{G}}_{\mathbf{s}}$. This result strengthens the interpretability of our approach by demonstrating that the representation-based groups reflect meaningful feature interactions.