Abstract

We present a novel Bayesian-based optimization framework that addresses the challenge of generalization in overparameterized models when dealing with imbalanced subgroups and limited samples per subgroup. Our proposed tri-level optimization framework utilizes local predictors, which are trained on a small amount of data, as well as a fair and class-balanced predictor at the middle and lower levels. To effectively overcome saddle points for minority classes, our lower-level formulation incorporates sharpness-aware minimization. Meanwhile, at the upper level, the framework dynamically adjusts the loss function based on validation loss, ensuring a close alignment between the global predictor and local predictors. Theoretical analysis demonstrates the framework’s ability to enhance classification and fairness generalization, potentially resulting in improvements in the generalization bound. Empirical results validate the superior performance of our tri-level framework compared to existing state-of-the-art approaches. The source code can be found at https://github.com/PennShenLab/FACIMS.

1 INTRODUCTION

Machine learning has achieved exceptional performance through overparameterization and advanced techniques. This progress is supported by high-quality datasets with sufficient samples for each data class and subgroup. However, real-world datasets frequently exhibit imbalances of different types and magnitudes, reflecting the significance and diversity of the underlying domains [Barocas et al., 2023]. Two common imbalances are observed in label-imbalanced and group-sensitive classification scenarios.

Label-imbalanced classification (LIC) suffers from a significant discrepancy in the number of examples across classes, requiring the use of balanced accuracy as a more suitable metric than conventional misclassification error. To improve model performance and balanced accuracy, various methods have been developed, including [Buda et al., 2018] and loss re-weighting [He and Garcia, 2009]. Weighted cross-entropy (wCE) loss, a classical approach, amplifies the contribution of minority examples in proportion to the imbalance level. However, wCE may not effectively handle the imbalance in overparameterized models [Cao et al., 2019], which can result in poor generalization. Recent studies propose alternative loss functions, such as logit-adjusted loss [Menon et al., 2020, Cao et al., 2019], class-dependent temperature loss [Ye et al., 2020], and vector-scaling loss [Kini et al., 2021], aiming to address the challenges associated with overparameterization. Nonetheless, there is still a risk of overfitting on minority class samples despite these advancements [Rangwani et al., 2022].

In group-sensitive classification (GSC), the goal is to ensure fairness concerning protected attributes like gender or race, addressing the issue of stereotyping where certain target labels are more frequently associated with specific groups [Mehrabi et al., 2021]. For instance, the occupation of "nurse" being commonly associated with females. While there is no universal fairness metric [Kleinberg et al., 2016], one suggestion is group sufficiency, which aims to maintain identical conditional expectations of the ground-truth label $E[Y|f(X), A]$ across different subgroups $A = 1, \ldots, A$ given the predictor’s output $f(X)$. However, in overparameterized models with limited samples per subgroup, this control of group sufficiency may not always hold, despite its effectiveness under certain assumptions in unconstrained learning [Liu et al., 2019a, Shui et al., 2022c].

Given the aforementioned challenges regarding the performance of LIC and GSC in overparameterized models, we pose the following question:

Q: How can a classifier be designed to effectively generalize on imbalanced subgroups with limited samples?
To address Q, we establish a link between LIC and GSC and propose a novel Bayesian framework that maintains informative predictions for imbalanced data while minimizing generalization error. Our contributions can be summarized as follows.

- We design a Bayesian-based tri-level optimization framework called Fairness-Aware Class Imbalanced Learning on Multiple Subgroups (FACIMS). In FACIMS, local predictors are learned using a small amount of training data and a fair, class-balanced predictor. The lower-level formulation utilizes the sharpness-aware minimization [Foret et al., 2020] to encourage convergence to a flat minimum and effectively avoid saddle points for minority classes. The upper-level problem dynamically adjusts the loss function by monitoring the validation loss, following a similar approach to [Li et al., 2021], and updates the global predictor to align with all subgroup-specific predictors.
- We establish the $O(1/\sqrt{T})$ convergence rate of our proposed three-level optimization framework, corresponding to a $O(\epsilon^{-2})$ sample complexity with a fixed number of samples used per iteration.
- We quantify the generalization performance of the models trained using our proposed tri-level FACIMS approach. The generalization bound analysis demonstrates that our method can achieve superior generalization performance compared to bilevel variants, such as [Rangwani et al., 2022], for fair learning on multiple subgroups.
- We conduct experiments on synthetic and real-world datasets to evaluate the balanced accuracy, demographic parity, equalized odds, and group sufficiency. The results showcase the effectiveness of our proposed method.

2 PRELIMINARIES

We consider a joint random variable $(X, Y, A)$ that follows an underlying distribution $P(X, Y, A)$, where $X \in \mathcal{X} \subset \mathbb{R}^d$ represents the input, $Y \in \mathcal{Y} = \{1, \ldots, K\}$ represents the label, $A \in \mathcal{A} = \{1, \ldots, A\}$ is a scalar discrete random variable that denotes the sensitive attribute or subgroup index. For instance, $A$ could represent gender or race. Throughout, $\mathbb{E}[Y|X]$ denotes the conditional expectation of $Y$, which can be seen as a function of $X$. $\mathbb{E}_{A,X}[\cdot]$ represents the expectation over the marginal distribution of $P(A, X)$.

Suppose we have a dataset $S = (x_i, y_i)_{i=1}^n$ sampled i.i.d. from a distribution $P$ with input space $\mathcal{X}$ and $K$ classes. Let $f : \mathcal{X} \rightarrow \mathbb{R}^K$ be a model that outputs a distribution over classes and let $h_f(x) = \arg \max_{k \in [K]} f(x)$. The standard classification error is denoted by $\text{ACC} = P_S[y \neq \hat{y}_f(x)]$. For a loss function $\ell(y, \hat{y})$, we similarly denote

$$\text{Population risk: } \mathcal{L}(f; P) := \mathbb{E}_P[\ell(y, \hat{y}_f(x))], \quad (1a)$$
$$\text{Empirical risk: } \mathcal{L}(f; S) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{y}_f(x_i)). \quad (1b)$$

We denote the frequency of the $k$-th class via $\pi_k = P_{(x,y) \sim P}(y = k)$. Label/class-imbalance occurs when the class frequencies differ substantially, i.e., $\max_{i \in [K]} \pi_i >> \min_{i \in [K]} \pi_i$. We define

$$\text{Class-conditional risk: } f_k := \mathbb{E}_{P_x}[\ell(y, \hat{y}_f(x))], \quad (2a)$$
$$\text{Balanced risk: } \text{BACC}(f) := \frac{1}{K} \sum_{k=1}^K f_k. \quad (2b)$$

2.1 PARAMETRIC LOSSES

We review some of the SOTA re-weighting methods for training on imbalanced data with distribution shifts.

Label-Distribution-Aware Margin (LDAM) [Cao et al., 2019] determines optimal margins for each class by minimizing errors using a generalization bound. It utilizes $\Delta_j$ as the margin for each class, defined as follows:

$$\ell_{\Delta}(f; x, y) = -\log \frac{e^{f(x)_{y}} - \Delta_j}{e^{f(x)_{y}} - \Delta_j + \sum_{j \neq y} e^{f(x)_{y}}} \quad \text{(LDAM)}$$

where $\Delta_j = \frac{C}{n_j^{1/2}}$ for $j \in \{1, \ldots, K\}$.

LDAM prioritizes classes with low sample sizes ($n_j$) over those with high frequencies. Deferred Re-Weighting (DRW) [Cao et al., 2019] involves training the model with an average loss until a certain epoch, then applying weights
proportional to the inverse of the sample size to the loss term for each class. The loss function for DRW is as follows:

$$\ell_u(f; x, y) = -u_y \log \frac{e^{f(x)_y}}{\sum_{j=1}^K e^{f(x)_j}}, \quad \text{(DRW)}$$

where $u_j = \frac{1}{1 + (n_j - 1)\bar{\gamma}_j}$.

This way of re-weighting has been shown to be effective for improving generalization performance when combined with various losses.

Vector Scaling (VS) [Kini et al., 2021] loss is a recently proposed loss function that unifies the idea of multiplicative shift [Ye et al., 2020], additive shift [Menon et al., 2020], and loss re-weighting. It has the following form:

$$\ell(f, v; x, y) = -u_y \log \frac{e^{\gamma_y f(x)_y + \Delta_x}}{\sum_{j=1}^K e^{\gamma_j f(x)_j + \Delta_x}}. \quad \text{(VS)}$$

Here, $v := (u_j, \Delta_j, \gamma_j)$ are some logit hyperparameters.

Throughout, our main focus is on VS loss, but our framework can also accommodate other loss functions.

### 2.2 FAIRNESS NOTIONS

We next review some fairness notions and the corresponding gaps.

**Definition 1.** Let $f$ be a score function that maps the random variable $X$ to a real number.

- **Group Sufficiency (GS):** We say that $f$ is sufficient with respect to attribute $A$ if $\mathbb{E}[Y|f(X)] = \mathbb{E}[Y|f(X), A]$.

- **Demographic Parity (DP):** $f$ satisfies demographic parity with respect to $A$ if $\mathbb{E}[f(X)] = \mathbb{E}[f(X)|A]$.

- **Equalized Odds (EO):** $f$ satisfies equalized odds with respect to $A$ if $\mathbb{E}[f(X)|Y] = \mathbb{E}[f(X)|Y, A]$.

GS means that the score function $f$ captures all the information about the label $Y$ that is relevant for prediction, regardless of the attribute $A$. DP ensures that the expected score $f(X)$ remains constant, regardless of the attribute $A$. This principle guarantees that the distribution of scores remains unaffected by the sensitive attribute, thereby promoting fairness in the decision-making process. EO dictates that the expected score $f(X)$ remains consistent across all combinations of labels $Y$ and attributes $A$. It ensures that individuals sharing the same label but differing attributes are treated equally in terms of their predicted scores, irrespective of the sensitive attribute.

The impossibility theorem of fairness asserts that, in general cases, it is impossible to simultaneously achieve all common and intuitive definitions of fairness. Notably, [Barocas et al., 2019, Chouldechova, 2017] demonstrate that if $A \not\perp Y$, it is not feasible to achieve both group sufficiency and demographic parity. Moreover, [Barocas et al., 2019, Pleiss et al., 2017] reveal that when $P(X, Y, A) > 0$ and $A \not\perp Y$, it is not possible for both group sufficiency and demographic parity to hold simultaneously.

Definition 1 leads to a notion of the **group sufficiency gap**, **demographic parity gap**, and **equalized odds gap** defined, respectively, as:

$$\text{SGap}_f(A) = \mathbb{E}[|f(Y)| - \mathbb{E}[Y|f(X), A]], \quad \text{(3a)}$$

$$\text{PGap}_f(A) = \mathbb{E}[f(X)] - \mathbb{E}[f(X)|A], \quad \text{(3b)}$$

$$\text{OGap}_f(A) = \mathbb{E}[f(X)|Y] - \mathbb{E}[f(X)|Y, A]. \quad \text{(3c)}$$

$\text{SGap}_f$ measures the extent of group sufficiency violation, induced by the predictor $f$, which is taken by the expectation over $(X, A)$. Hence, $\text{SGap}_f = 0$ suggests that $f$ satisfies group sufficiency and vice versa. For completeness, we also discuss computing these gaps in Appendix C.

To conclude this section, we provide Group $A$-Bayes predictor and an upper bound for $\text{SGap}_f$ from [Shui et al., 2022c]. These findings serve as the foundation for our Bayesian-based tri-level optimization framework.

**Definition 2 (A-group Bayes predictor).** The $A$-group Bayes predictor $f^{A,\text{Bayes}}_A$ associated with distribution $P(X, Y, A)$ is defined as: $f^{A,\text{Bayes}}_A(X) = \mathbb{E}[Y|X, A]$.

The following Theorem provides the upper bound of group sufficiency gap w.r.t. any predictor $f$:

**Theorem 3.** If $A$ takes finite value, i.e., $|A| < +\infty$ and follows uniform distribution with $p(A = a) = 1/|A|$, then

$$\text{SGap}_f(A) \leq \frac{4}{|A|} \sum_{a \in \mathcal{A}} \mathbb{E}_X |f - f^{A,\text{Bayes}}_A| |A = a|. \quad \text{(4)}$$

Hence, $\text{SGap}_f(A)$ depends on the discrepancy between the predictor $f$ and the $A$-group Bayes predictor $f^{A,\text{Bayes}}_A$. In other words, when considering different subgroups $A = a$, the optimal predictor $f$ should closely align with all the group Bayes predictors $f^{A=a,\text{Bayes}}_a$, for all $a \in \mathcal{A}$.

### 3 PROPOSED FRAMEWORK

In this section, we present the formulation of FACIMS, which is a framework designed to promote both classification accuracy and fairness through a randomized algorithm. FACIMS achieves this by learning a predictive distribution $Z$, which assigns higher scores to predictors that are favorable based on the available data. In the context of the FACIMS framework, the predictor is sampled from the posterior distribution, represented as $f \sim Z$. During the inference stage, the predictor’s output is computed as the expectation of the learned predictive distribution $Z$: $f(X) = \mathbb{E}_f \sim Z f(X)$. 
In practical scenarios, it is infeasible to optimize over the entire space of possible distributions. Therefore, we constrain the predictive distribution $Z$ to a specific distribution family $Z \in \mathcal{Z}$, such as the Gaussian distribution. Additionally, we denote $Z^{a,*} \in \mathcal{Z}$ as the optimal prediction distribution with respect to the subgroup $A = a$ within the distribution family $Z$:

$$Z^{a,*} = \arg \min_{Z \in \mathcal{Z}} \mathbb{E}_{f_{a} \sim Z} \mathcal{L}(\tilde{f}^a, v; S^a).$$

In general, $Z^{a,*} \neq \tilde{f}^a_{\text{Bayes}}$, since the distribution family $Z$ is only the subset of all possible distributions.

**Corollary 4** (Shui et al. [2022c]). The group sufficiency gap in a randomized algorithm w.r.t. the learned predictive distribution $Z$ is bounded as follows:

$$\text{SGap}_f \leq \mathcal{O}(\text{Optim} + \text{Approx}),$$

where

$$\text{Optim} := \frac{1}{|A|} \sum_{a} \sqrt{\text{KL}(Z^{a,*}||Z)},$$

$$\text{Approx} := \frac{1}{|A|} \sum_{a} \sqrt{\text{KL}(Z^{a,*}||P(Y|X, A = a))}.$$ 

Minimizing the Optim term ensures that the learned distribution $Z$ is both fair and informative for making predictions. On the other hand, the Approx term represents the KL divergence between the optimal distribution $Z^{a,*}$ and $P$. If the distribution family $Z$ has a rich expressive power, like that of a deep neural network, the Approx term will be small. See Figure 1 for a visual representation.

Now, we provide a framework for fairness-aware class imbalanced learning on multiple subgroups with potentially improved generalization bound and SGap$_f$. We begin by formulating the loss function design as a bilevel optimization over hyperparameters $\nu$ and a distribution $Z$. Assume each group $a \in A$ has a fine-tuning training set $T^a = \cup_{i=1}^{n_a} \{(x^a_i, y^a_i)\}$ and a separate validation set $V^a = \cup_{i=1}^{\nu_a} \{(x^a_i, y^a_i)\}$, where data are independently and identically distributed (i.i.d.) and drawn from the per-task data distribution $P^a$. Following [Li et al., 2021], define the empirical risk and the balanced empirical risk over a finite-sample dataset $S$ as $\mathcal{L}_{\nu}(f, v; S) := 1/n \sum_{i=1}^{n_a} \ell(f, v; x_i, y_i)$ and $\mathcal{L}_{\text{bal}}(\tilde{f}; S) := 1/n \sum_{i=1}^{\nu_a} \ell(\tilde{f}, v; x_i, y_i)$. Here, $\nu$ can be manually adjusted using (DRW), (LDAM), and (VS).

Let $Z$ stand for both fair and informative prediction. Building on [Kini et al., 2021, Li et al., 2021, Shui et al., 2022c], we design the following objective:

$$\min_{Z \sim Z, \nu} \sum_{a \in A} \alpha^{\text{up}} \text{KL}(Z^{a,*}||Z) + \mathbb{E}_{f_{a} \sim Z} \mathcal{L}_{\text{bal}}(\tilde{f}; V^a), \quad (6a)$$

s.t.

$$Z^{a,*} = \arg \min_{Z \in \mathcal{Z}} \alpha^{\text{low}} \text{KL}(Z^{a,*}||Z)$$

$$+ \mathbb{E}_{f_{a} \sim Z} \mathcal{L}_{\nu}(\tilde{f}^a, v; T^a), \quad \forall a \in A. \quad (6b)$$

Here, the lower-level problem (6b) includes a regularization term $\text{KL}(Z^{a,*}||Z)$ as an informative prior for learning local predictor $Z^{a,*}$. This optimization reduces the upper bound of the group-specific generalization error.

In the upper-level problem (6a), we update $Z$ by minimizing the average KL-divergence between different $Z^{a,*}$, controlling the upper bound of SGap$_f$ according to (5), as well as the balanced empirical risk. However, directly minimizing (6b) in a single-level approach does not work well in our setting due to the limited number of samples in each subgroup. This leads to overfitting and large generalization error for each subgroup. To address this, we consider additional assumptions, such as the similarity in the data generation distribution $P[Y|X, A]$ for each subgroup. With these assumptions, we can learn shared and fair models that are informative and sufficient for a large number of subgroups.

### 3.1 Parametric Models and FACIMS

In this section, we propose a practical learning algorithm applicable to various differentiable and parametric models, including neural networks.

We utilize the Isotropic Gaussian distribution as $Z$ to learn global informative $Z$ with parameters $(\theta, \sigma)$. For each subgroup $A = a$, we also learn group-specific parameters $(\theta^a, \sigma^a)$ for $Z^a$. The Isotropic Gaussian distribution is selected for computational efficiency in optimization, but other differentiable distributions can be also used for parameter density functions.

Given a training set, we learn $\tilde{f}_w : X \mapsto Y$ parameterized by $w \in \mathbb{R}^d$. Then $f_{\tilde{w}} \sim Z$ is equivalent to sampling the model parameter $w$ from the predictive-distribution $Z$. Hence, learning the distribution $Z$ is equivalent to learning parameter $(\theta, \sigma)$. Note that for each subgroup $A = a$, $f_{\tilde{w}}^a \sim Z^a$ can be modeled similarly. Both procedures can be formulated as follows:

$$w \sim \mathcal{N}(\theta, \sigma) = \prod_{i=1}^{d} \mathcal{N}(\theta[i], \sigma[i]),$$

$$w^a \sim \mathcal{N}(\theta^a, \sigma^a) = \prod_{i=1}^{d} \mathcal{N}(\theta^a[i], \sigma^a[i]), \quad \forall a \in A.$$
We incorporate (SAM) into (6b) and propose (7) by introducing weight perturbation $\epsilon$.

Given $w$, the maximization in (SAM) seeks to find the weight perturbation $\epsilon$ in the Euclidean ball that maximizes the empirical loss. If we define the sharpness as

$$\max_{\|\epsilon\| \leq \beta^a} \left[ \mathcal{L}(\hat{f}_{w+\epsilon}; S) - \mathcal{L}(\hat{f}_w; S) \right]$$

then (SAM) essentially minimizes the sum of the sharpness and the empirical loss $\mathcal{L}(\hat{f}_w; S)$.

We incorporate (SAM) into (6b) and propose (7) by introducing a set of positive constants $\{\beta^a\}_{a \in A}$. The FACIMS framework, combined with SAM, promotes convergence to a flat minimum and aids in escaping saddle points for minority classes [Rangwani et al., 2022]. We empirically demonstrate the superiority of integrating SAM into FACIMS over popular baselines and provide theoretical evidence suggesting improved generalization bounds. Despite the tri-level problem formulation in (7), our algorithm design efficiently approximates the maximization step, making the computational cost comparable to that of (6).

Based on the analysis and (7), we provide an alternating optimization algorithm for solving (7) in Algorithm 1. Line 3 provides a partial group setting, i.e., for many subgroups, we can randomly sample a subset $A'$ such that $|A'| << |A|$ for memory saving.

### 4 THEORETICAL ANALYSIS OF FACIMS

Next, we analyze the performance of the FACIMS method.

For simplicity, we replace $\mathcal{L}_{bal}$ and $\mathcal{L}_{vs}$ with $\mathcal{L}$. We also use $\nabla \mathcal{L}$ to denote the stochastic gradients of $\mathcal{L}$. Define $\Theta := (\theta, \sigma, \nu)$ and let $F$ and $F^0$ denote the objective of (7a) and (7b), respectively.

**Assumption A** (Lipschitz continuity). Assume that $\nabla \mathcal{L}(\cdot; \nu^a), \nabla^2 \mathcal{L}(\cdot; \nu^a), \nabla \mathcal{L}(\cdot; \nu^a), \nabla^2 \mathcal{L}(\cdot; \nu^a), \forall \alpha \in A$ are Lipschitz continuous with constant $\ell_0, \ell_1, \ell_1, \ell_2$.

**Assumption B** (Stochastic derivatives). Assume that $\nabla \mathcal{L}(\cdot; \nu^a), \nabla^2 \mathcal{L}(\cdot; \nu^a), \nabla \mathcal{L}(\cdot; \nu^a), \nabla^2 \mathcal{L}(\cdot; \nu^a), \nabla \mathcal{L}(\cdot; \nu^a)$ are unbiased estimators of $\nabla \mathcal{L}(\cdot; \nu^a), \nabla^2 \mathcal{L}(\cdot; \nu^a), \nabla \mathcal{L}(\cdot; \nu^a), \nabla^2 \mathcal{L}(\cdot; \nu^a)$ respectively and their variances are bounded by $\sigma^2$. 

\[
\min_{\mathbf{Z} \sim \mathcal{Z}} \sum_{a \in A} \alpha^u \mathbb{KL}(\mathbf{Z}^a; \mathbf{Z}) + \mathbb{E}_{w \sim \mathcal{Z}} \mathcal{L}_{bal}(\hat{f}_w; V^a), \tag{7a}
\]

s.t. $\mathbf{Z}^a \in \arg\min_{\mathbf{Z}^a \in \mathcal{Z}} \max_{\|\epsilon^a\| \leq \beta^a} \alpha^l \mathbb{KL}(\mathbf{Z}^a; \mathbf{Z}) + \mathbb{E}_{w \sim \mathcal{Z}} \mathcal{L}_{vs}(\hat{f}_w + \epsilon^a, \nu^a; T^a), \quad \forall a \in A. \tag{7b}
\]
Assumptions A–B also appear similarly in the convergence analysis of and bilevel optimization [Chen et al., 2021, Tarzanagh et al., 2022]. With the above assumptions, we get the following theorem. The proof is deferred in Appendix A.

**Theorem 5.** Under Assumption A–B, and choosing step-sizes \( \gamma^\text{up}, \gamma^\text{low} \) and sharpness parameter \( \beta = O(1) \), with some proper constants, we can get that the iterates \( \{ (\theta_T, \sigma_T, v_T) \} \) generated by Algorithm 1 satisfy

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \| \nabla F(\theta_T, \sigma_T, v_T) \|^2 \right] = O \left( \frac{1}{\sqrt{T}} \right). \tag{11}
\]

Theorem 5 implies that under some standard assumption, Algorithm 1 can find \( \epsilon \) stationary points for FACIMS objective (6) with \( O(\epsilon^{-2}) \) iterations and \( O(\epsilon^{-2}) \) samples.

Next, we establish the generalization performance.

**Theorem 6.** Assume the function \( \mathcal{L}(\cdot) \) is bounded for any i.i.d \( S \). Let \( F(\cdot; P) = \mathbb{E}_{S \sim P} [F(\cdot; S)] \). Assume \( F(\Theta, \sigma, v; P) \leq \mathbb{E}_{S \sim N(0, \beta^2 I)} [F(\Theta + \epsilon, \sigma + \epsilon, v; P)] \) at the stationary point of (7) denoted by \( (\Theta + \epsilon, \hat{\sigma} + \epsilon, \hat{v}) \). Then, with probability \( 1 - \delta \) over the choice of the training set \( \mathcal{S} \sim \mathcal{P} \), with \( |\mathcal{S}| = n|\mathcal{A}| \), then generalization error is bounded by

\[
F(\Theta; P) \leq \max_{|\epsilon| \leq \beta} \mathbb{E}_{S \sim N(0, \beta^2 I)} [F(\Theta + \epsilon, \sigma + \epsilon, v; D)]
\]

\[
\leq \sqrt{\left( \frac{(p + K) \ln \left( 1 + \frac{\|\Theta\|^2}{\beta^2} \left( 1 + \frac{\ln(n|\mathcal{A}|)}{p} \right)^2 \right)}{4n|\mathcal{A}|} \right) + 5 \frac{\ln \frac{1}{\delta} + \ln(n|\mathcal{A}|)}{4n|\mathcal{A}|}}. \tag{12}
\]

Theorem 6 shows that the difference between the population loss and the empirical loss of FACIMS is bounded by \( \mathcal{O}((p + K)/n|\mathcal{A}|) \) that vanishes as the number of group-specific training data goes to infinity. Note that the bound in (12) is a function of \( \beta \). Hence, for a choice of \( \beta \to 0 \), the bound (12) is not optimal. This suggests that our three-level FACIMS can have better generalization performance than that from bilevel variants such as [Rangwani et al., 2022].

## 5 EXPERIMENTS

### 5.1 EXPERIMENTAL SETUP

**Datasets.** We applied our model to the Alzheimer’s disease (AD), credit card and drug consumption datasets, and the data information is summarized in Table 1.

- **Alzheimer’s Disease dataset**\(^1\) were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database [Weiner et al., 2017, Shen et al., 2014]. We included 5137 instances, including 4080 mild cognitive impairment (MCI, a prodromal stage of AD) and 1057 AD instances, to conduct the binary classification. Moreover, we chose race as the sensitive feature and divided the participants into four subgroups, where white subjects exceeding 90%. Our features included 17 AD-related biomarkers, including cognitive scores, volumes of brain regions extracted from the magnetic resonance imaging (MRI) scans, amyloid and tau measurements from positron emission tomography (PET) scans and cerebrospinal fluid (CSF), and risk factors like APOE4 carriers and age.

- **Credit Card dataset**\(^2\) contains 22 attributes like clients’ basic information, history of payments, and bill statement amount to classify whether the clients are credible or not. We included 30000 instances with 6636 credible and 23365 not credible clients. We chose the education level as the sensitive feature where we observed more clients who graduated from university than other six levels.

- **Drug Consumption dataset**\(^3\) contains demographic information such as age, gender, and education level, as well as measures of personality traits thought to influence drug use for 1885 respondents. The task is to predict alcohol use with \( K = 4 \) categories (never used, not used in the past year, used in the past year, and used in the past day) for multi-class outcomes. The sensitive feature is education level (Left school before or at 16, Left school at 17-18, Some college, Certificate, Diploma, Higher degree).

---

\(^1\) [http://adni.loni.usc.edu](http://adni.loni.usc.edu)

\(^2\) [https://archive.ics.uci.edu/ml/datasets/credit+approval](https://archive.ics.uci.edu/ml/datasets/credit+approval)

\(^3\) [https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified](https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified)
We set \( \alpha \) and \( \beta \) to be 0.7. We use the grid of \( [0.1, 0.01, 0.001] \) to search the learning rate for global model and local models and report the results over five independent repeats.

### 5.2 EXPERIMENTAL RESULTS

In this section, we analyze Alzheimer’s disease and credit card datasets. The numerical results of the multi-class dataset drug consumption are included in the appendix due to page limits.

#### Balanced Accuracy and Sufficiency Gap

We primarily focus on balanced accuracy and group sufficiency gap as our main goals. Table 2 shows that on the Alzheimer’s disease dataset, our method FACIMS outperforms EIL, FSCS, FAMS, and ERM in terms of balanced accuracy, with improvements of 2.6%, 4.0%, 5.3%, and 2.1% respectively. While BERM addresses the class imbalance issue and demonstrates a significant improvement over ERM by nearly 2%, our method still achieves a higher balanced accuracy than BERM. Our method significantly improves the group sufficiency gap by 1.4%, 2.0%, and 2.8% respectively.

Although EIL, FSCS, and FAMS specifically target the group sufficiency gap, our method still outperforms these three baseline methods by improving the sufficiency gap as our main goals. Table 2 shows that on the Alzheimer’s disease dataset, our method FACIMS outperforms EIIL, FSCS, FAMS, and ERM in terms of balanced accuracy, with improvements of 2.6%, 4.0%, 5.3%, and 2.1% respectively. While BERM addresses the class imbalance issue and demonstrates a significant improvement over ERM by nearly 2%, our method still achieves a higher balanced accuracy than BERM. Our method significantly improves the group sufficiency gap by 1.4%, 2.0%, and 2.8% respectively. While BERM addresses the class imbalance issue and demonstrates a significant improvement over ERM by nearly 2%, our method still achieves a higher balanced accuracy than BERM. Our method significantly improves the group sufficiency gap by 1.4%, 2.0%, and 2.8% respectively.

Removing the lower level \((\beta = 0)\) leads to a slight decrease in balanced accuracy and group sufficiency gap as the objective landscape is not flattened in the middle level. Additionally, manually adjusting the logits instead of learn-
Figure 2: Boxplot comparing balanced accuracy and group sufficiency gap for three real datasets with 5 repeats. The mean is represented by the middle of each box, while the box width represents twice the standard deviation. Better performance is indicated by boxes located towards the bottom right (higher balanced accuracy and lower group sufficiency). Two FACIMS variants are excluded for clarity, with complete results available in the appendix.

Results on Other Metrics In addition to the result analysis on balanced accuracy and group sufficiency, we also report demographic parity, equalized odds, recall, and time in Table 2. The results show that our method achieves competitive results despite not outperforming all baselines in terms of demographic parity and equalized odds gaps. We emphasize that our method primarily addresses the group sufficiency gap for fairness, and it is challenging to optimize all three fairness measurements simultaneously, as discussed in Section 2. When assessing a classifier’s performance, it is important to achieve a high recall for each class. However, the average recall across all classes determines the balanced accuracy, highlighting the need for a balanced recall quantity across all classes. Our approach and its variations demonstrate a more balanced recall for each class, as illustrated in Table 2.

Comparing the time aspect, despite employing a complex tri-level optimization framework for training our model, the total runtime is not significantly longer than other fairness baselines. Indeed, utilizing differentiable bilevel methods in the large hyperparameter search provides substantial cost reduction and speedup compared to traditional approaches like grid search or random search. For instance, the first variant of our method, FACIMS \((\beta = 0, v = \bar{v})\), runs in approximately 13 minutes. However, employing grid search or random search to tune the parametric loss would require significantly more time. For example, if we perform a search with five different settings to enhance the accuracy of FACIMS \((\beta = 0, v = \bar{v})\), the total time would be 13 min \times 5 = 65 min, which is around four times longer than our differentiable three-level FACIMS approach.

Influence of \(\alpha_{\text{low}}\) In the middle level, the parameter \(\alpha_{\text{low}}\) determines the attention given to \(\text{KL}(Z' | Z)\). A higher value of \(\alpha_{\text{low}}\) brings the local model closer to the global model, leading to improved group sufficiency gap but potentially worse balanced accuracy. We experimented with four different values of \(\alpha_{\text{low}}\): 0.01, 0.1, 0.2, and 1. Figure 3 illustrates the Accuracy-SGap curve under varying \(\alpha_{\text{low}}\) on the Alzheimer’s disease dataset. The figure demonstrates a clear
Figure 3: Accuracy-SGap curve under different $\alpha^{low}$ in Alzheimer’s disease dataset.

trend: as $\alpha^{low}$ increases, both the balanced accuracy and group sufficiency gap decrease, aligning with our expectations. This analysis provides insight into how the KL divergence in the middle level influences the group sufficiency gap and balanced accuracy, enhancing our understanding of the framework’s mechanism.

6 RELATED WORK

6.1 LONG-TAILED LEARNING

Re-sampling [Buda et al., 2018] and Re-weighting [He and Garcia, 2009] are commonly used methods for training on imbalanced datasets. Recent studies focus on optimizing loss landscapes for class-imbalanced datasets [Khan et al., 2017, Cao et al., 2019, Menon et al., 2020, Ye et al., 2020, Li et al., 2021, Kini et al., 2021, Behnia et al., 2023, Thrampoulidis et al., 2022]. Our work is related to the long-tail learning literature [Cao et al., 2019, Menon et al., 2020, Ye et al., 2020, Kini et al., 2021], where authors propose refined class-balanced loss functions that better adapt to training data. These include the logit-adjusted loss [Menon et al., 2020, Cao et al., 2019], the class-dependent temperature loss [Ye et al., 2020], and the VS loss [Kini et al., 2021], which unifies the concepts of multiplicative shift, additive shift, and loss re-weighting.

6.2 NESTED OPTIMIZATION

Nested optimization involves solving hierarchical problems with multiple levels of optimization, where one task is nested within another [Colson et al., 2007, Tarzanagh et al., 2022, Chen et al., 2021, Ji et al., 2021]. Min-max nested optimization is commonly used to learn fair representations in the context of demographic parity or equalized odds [Zemel et al., 2013, Song et al., 2019, Zhao et al., 2019]. Bi-level optimization and meta-learning algorithms have also been explored in the context of fair learning and classification [Shui et al., 2022b, Abbas et al., 2022]. Recent advancements in differentiable algorithms have led to faster bilevel algorithms for learning hyperparameters and classification [Li et al., 2021, Lorraine et al., 2020, Tarzanagh et al., 2022, Chen et al., 2021, Ji et al., 2021]. Building on [Li et al., 2021], we propose a theoretically justified tri-level optimization perspective to control the group sufficiency gap and improve generalization performance across multiple subgroups with limited samples.

6.3 FAIRNESS

Group-sensitive learning aims to ensure fairness in the presence of under-represented groups [Lin et al., 2023]. Our work mainly focuses on the fair notion of group sufficiency. This notion has recently been studied in the health of populations [Obermeyer et al., 2019] and crime prediction [Chouldechova, 2017, Pleiss et al., 2017]. Liu et al. [2019b] show that under some assumptions, the group sufficiency can be controlled in unconstraint learning. On the other hand, Obermeyer et al. [2019], Shui et al. [2022a], Koh et al. [2021] claim that this conclusion may not always hold in the overparameterized models with limited samples per group. Subramanian et al. [2021] provided a method for fair and class-imbalanced learning. Lee et al. [2021] recently presented a method for learning a fair predictor for all groups via formulating it as a bilevel objective. In contrast, our tri-level algorithm uses a Bayesian framework for imbalanced learning so that margins depend not just on class imbalance, but also on the subgroup distribution within each class. Besides, it provides a SAM-type nested optimization to effectively escape saddle points for minority classes.

7 CONCLUSIONS

We studied fairness-aware class imbalanced learning on multiple subgroups (FACIMS) using a Bayesian-based optimization framework. Through extensive empirical and theoretical analysis, we demonstrated that FACIMS enhances the generalization performance of overparameterized models when dealing with limited samples per subgroup.

ACKNOWLEDGEMENTS

This work was supported in part by the NIH grants U01 AG066833, RF1 AG063481, U01 AG068057, R01 LM013463 P30 AG073105, and U01 CA274576, and the NSF grant IIS 1837964. Data used in this study were obtained from the Alzheimer’s Disease Neuroimaging Initiative database (adni.loni.usc.edu), which was funded by NIH U01 AG024904. The authors Davoud Ataee Tarzanagh, Bojian Hou and Boning Tong have contributed equally to this paper.
References


Supplementary Material for “Fairness-Aware Class Imbalanced Learning on Multiple Subgroups”

Roadmap. The appendix is structured as follows: In Section A, we present the proof of Theorem 5. Section B contains the proof of Theorem 6. Section C describes the computation of fairness gaps. Lastly, Section D presents additional experimental results.

A PROOF OF THEOREM 5

To establish the relationship between FACIMS and bilevel optimization [Chen et al., 2021], we introduce a new vector \( \phi \) obtained by concatenating \( \Theta^a \) for all \( a \), given by \( \phi = [\Theta^1, \cdots, \Theta^A]^{\top} \). We then define

\[
F(\Theta) = f(\phi'(\Theta)), \quad f(\phi) = \sum_{a \in A} \mathcal{L}(\Theta^a; T^a), \quad g(\Theta, \phi) = \sum_{a \in A} \mathcal{L}(\theta; \mathcal{Y}^a).
\]

Then, from the properties of implicit functions, we get

\[
\nabla F(\Theta) = -\nabla_{\Theta} g(\Theta, \phi) \nabla_{\phi} g(\Theta, \phi) \nabla_{\phi} f(\phi).
\] (13)

Then, for notational simplicity, we consider the single-sample case with \( \kappa = 1 \) and define three independent samples for stochastic gradient and Hessian computation as \( \xi^a := (x, y) \sim S^a, \psi^a := (x, y) \sim S^a, \xi^{a, i} := (x, y) \sim S^a \), so the corresponding \( \kappa \)-batch gradient and Hessian estimators for bilvel-type methods can be written as

\[
\nabla \mathcal{L}(\Theta; S^a, \xi^a) = \frac{1}{\kappa} \sum_{\xi^a \sim S^a} \nabla l(\theta, x, y),
\]

\[
\nabla^2 \mathcal{L}(\Theta; S^a, \psi^a) = \frac{1}{\kappa} \sum_{\psi^a \sim S^a} \nabla^2 l(\theta, x, y),
\]

\[
\nabla \mathcal{L}(\Theta; S^a, \xi^{a, i}) = \frac{1}{\kappa} \sum_{\xi^{a, i} \sim S^{a, i}} \nabla l(\theta, x, y).
\]

We rewrite the updates of Algorithm 1 as follows:

\[
\Theta_{t+1} = \Theta_t - \gamma_{\text{up}}^{\text{up}} \sum_{a \in A} (I - \gamma_{\text{low}} \nabla^2 \mathcal{L}(\Theta_t + e^a(\Theta_t); S^a, \psi^a)) \nabla \mathcal{L}(\Theta_{t+1}; S^a, \xi^a);
\]

\[
\hat{\Theta}_{t+1} = \Theta_t - \gamma_{\text{low}} \nabla \mathcal{L}(\Theta_t + e^a(\Theta_t); S^a, \xi^a)
\]

\[
= \Theta_t - \gamma_{\text{low}} \left( \nabla \mathcal{L}(\Theta_t + e^a(\Theta_t); S^a, \xi^a) - \frac{e^a(\Theta_t)}{\gamma_{\text{low}}} \right).
\]

Since \( \nabla \mathcal{L}(\Theta; S^a), \nabla^2 \mathcal{L}(\Theta; S^a) \) are Lipschitz continuous with \( \ell_1, \ell_2 \) according to Assumption A, then we have that

\[
\| \nabla \mathcal{L}(\Theta_t + e^a(\Theta_t); S^a) - \frac{e^a(\Theta_t)}{\gamma_{\text{low}}} - \nabla \mathcal{L}(\Theta_t; S^a) \| \leq \ell_1 \beta
\]

\[
\| \nabla^2 \mathcal{L}(\Theta_t + e^a(\Theta_t); S^a) - \nabla^2 \mathcal{L}(\Theta_t; S^a) \| \leq \ell_2 \beta.
\]

Now, if we set \( \gamma_{\text{up}} = \gamma_{\text{low}} = \mathcal{O}(\frac{1}{\sqrt{T}}) \), \( \beta = \mathcal{O}(1) \), then the reminder of the proof follows from [Chen et al., 2021, Theorem 1]

B PROOF OF THEOREM 6

Let us define the empirical Rademacher complexity of \( \mathcal{F} \) of subgroup/class margin on \( S_a \) as

\[
\hat{R}_t(F) = \frac{1}{n_t} \mathbb{E}_\xi[\sup_{f \in \mathcal{F}} \sum_{j \in S_a} \xi_j f(x_j)_1 - \max_{i \neq 1} f(x_j)_i],
\]
\[
\hat{R}_{i,a}(F) = \frac{1}{n_{i,a}} \mathbb{E}_x \left[ \sup_{f \in F} \sum_{j \in S_{i,a}} \xi_j [f(x_j) - \max_{i' \neq i} f(x_{j})] \right],
\]

where \( \xi_j \) is i.i.d. drawn from a uniform distribution \{-1, 1\}.

**Lemma 7.** Let

\[
\hat{\ell}_{\gamma,i}[f] = \mathbb{P}_{x \sim \hat{p}_i}(\max_{j \neq i} f(x)_j > f(x)_i - \gamma),
\]

\[
\hat{\ell}_{\gamma,i,a}[f] = \mathbb{P}_{x \sim \hat{p}_{i,a}}(\max_{j \neq i} f(x)_j > f(x)_i - \gamma).
\]

With probability at least \(1 - \delta\) over the randomness of the training data, for any proper complexity measure of class \( F \), for any \( f \in F, * \in \{i, (i,a)\} \in \mathcal{Y}, a \in A \), and all margins \( \gamma > 0 \)

\[
\mathcal{L}_*[f] \leq \hat{\ell}_{\gamma,*}[f] + \frac{1}{\gamma} \hat{R}_*[F] + \epsilon_*(n_*, \delta, \gamma_*) =: \hat{R}_*[F] + \epsilon_*(n_*, \delta, \gamma_*) + \hat{R}_*[F].
\]

Here, \( \hat{R}_*[F] \) is the empirical Rademacher complexity of \( F \) of subgroup/class margin on training dataset corresponding to index set \( S_\ast \), which can be further upper bounded by \( \sqrt{\frac{C(F)}{n_\ast}} \). Also, \( \epsilon_*(n_\ast, \delta, \gamma_\ast) \) is usually a lower-order term in \( n_\ast \).

**Proof.** This is a direct application of the generalization bound in [Kakade et al., 2008].

**Proof.** Let

\[
\rho := (Z_1 \otimes Z_2 \otimes \cdots \otimes Z_{|A|}) \text{ and } \pi := (Z \otimes Z \otimes \cdots \otimes Z)_{|A| \text{ times}}
\]

We also set \( X_k = (x_k^1, y_k^1), l = |A|m, f = (\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_{|A|}) \), \( g_k(f, X_k) = \frac{1}{|A|m} \mathbb{E}_{\tilde{BCE}}(\tilde{f}_a(x_k^1), y_k^1) \). Since we adopt the binary cross entropy loss, \( a_k = 0 \) and \( b_k = L / (|A|m) \), then with high probability \( 1 - \delta \), we have:

\[
\frac{1}{|A|} \sum_a \mathbb{E}_{j \sim Z_a} \mathcal{L}_a^{BCE}(\tilde{f}_a) \leq \frac{1}{|A|} \sum_a \mathbb{E}_{j \sim Z_a} \mathcal{L}_a^{BCE}(\tilde{f}_a) + \frac{1}{\lambda} (\text{KL}(Z_1 \otimes \cdots \otimes Z_{|A|} || Z \otimes \cdots \otimes Z) + \log(1/\delta)) + \frac{\lambda L}{8 |A|^2 n}.
\]

Through the decomposition property of KL divergence, we finally have:

\[
\frac{1}{|A|} \sum_a \mathbb{E}_{j \sim Z_a} \mathcal{L}_a^{BCE}(\tilde{f}) \leq \frac{1}{|A|} \sum_a \mathbb{E}_{j \sim Z_a} \mathcal{L}_a^{BCE}(\tilde{f}) + L \sqrt{\frac{1}{2 |A|^2 n} \left( \sum_a \text{KL}(Z_a || Z) + \log(1/\delta) \right)} \leq \frac{1}{|A|} \sum_a \mathbb{E}_{j \sim Z_a} \mathcal{L}_a^{BCE}(\tilde{f}) + \frac{L}{\sqrt{|A|^2 n}} \sum_a \sqrt{\text{KL}(Z_a || Z) + L \sqrt{\log(1/\delta)}}.
\]

Now, let \( P_0 = P = \mathcal{N}(0, \sigma^2 P I), P = Z = \mathcal{N}(\cdot, \alpha^2 I) \), then

\[
D_{KL}(Z||P) = \frac{1}{2} \left\{ \text{tr} \left( \Sigma^{-1}_P \Sigma_Z \right) + (\mu_P - \mu_Z)^T \Sigma^{-1}_P (\mu_P - \mu_Z) - k + \ln \left( \frac{||\Sigma_P||}{\Sigma_Z} \right) \right\} = \frac{1}{2} \left[ k\alpha^2 + \frac{2}{\sigma^2} \right] - k + \ln \left( \frac{\sigma^2 P}{\alpha^2} \right).
\]

Let \( T = \{ c \exp((1 - j)/k) | j \in \mathbb{N} \} \) be the set of values for \( \sigma^2 P \). If for any \( j \in \mathbb{N} \), the PAC-Bayesian bound in (16) holds for \( \sigma^2_P = c \exp((1 - j)/k) \) with probability \( 1 - \delta_j \), with \( \delta_j = \frac{6\delta}{N \pi^2 j^2} \), then by the union bound, all bounds w.r.t. all \( \sigma^2_P \in T \) hold simultaneously with probability at least \( 1 - \sum_{j=1}^{\infty} \frac{6\delta}{N \pi^2 j^2} = 1 - \delta \).
First consider $\| \cdot \|^2 \leq \beta^2(\exp(4n/k) - 1)$, then $k\beta^2 + \| \cdot \|^2_2 \leq k\beta^2(\exp(4n/k) + 1)$. Now set $j = \lceil 1 - k \ln \left((\beta^2 + \| \cdot \|^2_2)/c\right) \rceil$. By setting $c = \beta^2(1 + \exp(4n/k))$, then $\ln \left((\beta^2 + \| \cdot \|^2_2)/c\right) < 0$, thus we can ensure that $j \in \mathbb{N}$. Furthermore, for $\sigma^2_n = c \exp((1 - j)/k)$, we have:

$$\beta^2 + \| \cdot \|^2_2/k \leq \sigma^2_n \leq \exp(1/k)(\beta^2 + \| \cdot \|^2_2/k)$$

where the first inequality is derived from $1 - j = \lceil k \ln((\beta^2 + \| \cdot \|^2_2)/c) \rceil \geq k \ln((\beta^2 + \| \cdot \|^2_2)/c)\), the second inequality is derived from $1 - j = \lceil k \ln((\beta^2 + \| \cdot \|^2_2)/c) \rceil \leq k \ln((\beta^2 + \| \cdot \|^2_2)/c) + 1$.

The KL-divergence term can be further bounded as

$$D_{KL}(Z \| \mathcal{P}) = \frac{1}{2} \left[ \frac{k\beta^2 + \| \cdot \|^2_2}{\sigma^2_n} - k + k \ln \left(\frac{\sigma^2_n}{\beta^2}\right) \right]$$

$$\leq \frac{1}{2} \left[ \frac{k\beta^2 + \| \cdot \|^2_2}{\beta^2 + \| \cdot \|^2_2/k} - k + k \ln \left(\frac{\exp(1/k)}{\beta^2 + \| \cdot \|^2_2/k}\right) \right]$$

$$= \frac{1}{2} \left[ k \ln \left(\frac{\exp(1/k)}{\beta^2 + \| \cdot \|^2_2/k}\right) \right]$$

$$= \frac{1}{2} \left[ 1 + k \ln \left(1 + \frac{\| \cdot \|^2_2}{\beta^2}\right) \right].$$

Given the bound that corresponds to $j$ holds with probability $1 - \delta_j$ for $\delta_j = \frac{6 \delta}{\pi^2 \delta}$, the ln term above can be written as:

$$\ln \frac{1}{\delta_j} = \ln \frac{1}{\delta} + \ln \frac{\pi^2 \delta^2}{6}$$

$$\leq \ln \frac{1}{\delta} + \ln \frac{\pi^2 \delta^2 \ln(2(\beta^2 + \| \cdot \|^2_2/k))}{6}$$

$$\leq \ln \frac{1}{\delta} + \ln \frac{\pi^2 \delta^2 \ln(2c/\beta^2)}{6}$$

$$= \ln \frac{1}{\delta} + \ln \frac{\pi^2 \delta^2 \ln(1 + \exp(4n/k))}{6}$$

$$\leq \ln \frac{1}{\delta} + \ln \frac{\pi^2 \delta^2 (4n/k)^2}{6} \leq \ln \frac{1}{\delta} + 2 \ln(6n).$$

From [Laurent and Massart, 2000, Lemma 1], we have that for $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and any positive $t$:

$$P \left( \| \epsilon \|^2_2 - k\sigma^2 \geq 2\sigma^2 \sqrt{kt} + 2t\sigma^2 \right) \leq \exp(-t).$$

Therefore, with probability $1 - 1/\sqrt{n}$ we have that:

$$\| \epsilon \|^2_2 \leq \sigma^2 \left(2 \ln(\sqrt{n}) + k + 2\sqrt{k \ln(\sqrt{n})}\right) \leq \sigma^2 k \left(1 + \sqrt{\frac{\ln n}{k}}\right)^2 = \beta^2.$$

\[\blacksquare\]

C COMPUTING GAPS FROM THE DATA

Recall from (3),

$$S \text{Gap}_f(A) = \mathbb{E}[\mathbb{E}[Y \mid f(X)] - \mathbb{E}[Y \mid f(X), A]],$$

$$P \text{Gap}_f(A) = \mathbb{E}[\mathbb{E}[f(X)] - \mathbb{E}[f(X)|A]],$$

$$O \text{Gap}_f(A) = \mathbb{E}[\mathbb{E}[f(X)|Y] - \mathbb{E}[f(X)|Y, A]].$$

(17a)
C.1 COMPUTING SUFFICIENCY GAP

Note that \( \{S_a\}, a \in \mathcal{A} \) denotes the observed data, and \( f(x) \) is a continuous value, ranging from \([0, 1]\). We split \([0, 1]\) into separate intervals:

\[
[0, \delta], [\delta, \delta_1], \ldots, [\delta_N, 1]
\]

Now, we compute the conditional expectation from the data, i.e., \( \mathbb{E}[Y \mid f(X)] \) and \( \mathbb{E}[Y \mid f(X), A = a] \) within each interval:

\[
(p_i, q_i) := (\mathbb{E}[f(X)1_{\{f(X) \in [\delta_i, \delta_{i+1}]\}}], \mathbb{E}[Y \mid f(X) \in [\delta_i, \delta_{i+1}]]) \tag{18a}
\]

\[
(p_i^a, q_i^a) := (\mathbb{E}[f(X)1_{\{f(X) \in [\delta_i, \delta_{i+1}], A = a\}}], \mathbb{E}[Y \mid f(X) \in [\delta_i, \delta_{i+1}], A = a]) \tag{18b}
\]

Now, from (18a) and (18b), for each group \( A = a \), the group sufficiency gap is computed as:

\[
SGap_f(A = a) = \sum_i \left| q_i - \frac{p_i^a - p_i^a}{p_{i+1}^a - p_i^a} (q_i^a + 1 - q_i^a) \right| \quad \text{Linear Interpolation}
\]

Note that in the case when the average values in each interval are not equal, we apply linear interpolation.

Finally, we set

\[
SGap_f = \frac{1}{|\mathcal{A}|} \sum_a SGap_f(A = a) \tag{19}
\]

We set \( \mathcal{P}(A = a) = \frac{1}{|\mathcal{A}|} \) as uniform distribution for ensuring fairness for each subgroup.

**Remark 8.** It can be easily seen that, in general:

\[
\mathbb{E}[Y \mid f(X)] \neq \frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}[Y \mid f(X), A = a].
\]

By using the Bayes rule, we get

\[
\mathbb{E}[Y \mid f(X)] = \sum_a \mathcal{P}(A = a \mid f(X)) \mathbb{E}[Y \mid f(X), A = a].
\]

Hence, if \( \mathcal{P}(A = a \mid f(X)) = \frac{1}{|\mathcal{A}|} \), we have the equivalent form. Specifically, \( \mathcal{P}(A = a \mid f(X)) \) refers the conditional probability of \( A = a \), given the predicted score \( f(X) \), which is related to the group membership inference [Hu et al., 2021]. If \( \mathcal{P}(A = a \mid f(X)) \) is large, the subgroup index can be easily revealed via the algorithm output. If the algorithm can fully preserve the privacy, then \( \mathcal{P}(A = a \mid f(X)) = \frac{1}{|\mathcal{A}|} \).

C.2 COMPUTING DEMOGRAPHIC PARITY GAP

It is straightforward to compute demographic parity gap according to the definition. Specifically, we first calculate the expectation of the prediction over each group \( \mathbb{E}[f(X) \mid A] \). Then we calculate the expectation of the prediction over all instances. Finally, we get the \( PGap_f \) by calculating \( \mathbb{E}[\mathbb{E}[f(X)] - \mathbb{E}[f(X) \mid A]] \).

In practical, we can count the number of positive predictions \( p_a \) for each group. This can be done by counting the number of true positives and false positives for each group in the dataset. Then we calculate the proportion of positive predictions \( \frac{p_a}{N_a} \) for each group \( a \) where \( N_a \) is the total number of predictions for each group. Similarly, for the whole dataset, we have \( \frac{p}{N} \) where \( p \) is the number of positive predictions for the whole dataset and \( N \) is the number of all the instances. Then we calculate the absolute difference between the proportion of positive predictions for each group and the whole dataset:

\[
PGap_f(A = a) = \left| \frac{p}{N} - \frac{p_a}{N_a} \right|
\]

Finally, we can obtain the demographic parity gap by taking the average over all the absolute differences:

\[
PGap_f = \frac{1}{|\mathcal{A}|} \sum_a PGap_f(A = a) \tag{20}
\]
C.3 COMPUTING EQUALIZED ODDS GAP

Equalized odds is not only conditioned on $A$ but also conditioned on $Y$. We focus on the positive class. Thus for each group $a$, we first count the number of true positive ($TP_a$) and false negative ($FN_a$) predictions. Then we calculate the true positive rate ($TPR_a$) for group $a$ by

$$TPR_a = \frac{TP_a}{TP_a + FN_a}.$$ 

Similarly, we have the $TPR$ for the whole dataset:

$$TPR = \frac{TP}{TP + FN}.$$ 

Then the equalized odds for each group is:

$$OGap_f(A = a) = |TPR - TPR_a|.$$ 

The equalized odds is finally calculated by

$$OGap_f = \frac{1}{A} \sum_a OGap_f(A = a).$$ (21)

D ADDITIONAL EXPERIMENTS

In Table 3 and Table 4, we report the numerical results of the multi-class dataset drug consumption regarding the six measurements including balanced accuracy, demographic parity, equalized odds, sufficiency gap, recall and time. Similar to the results of Alzheimer’s disease and credit card, our method FACIMS can outperform the baselines methods in terms of balanced accuracy and group sufficiency gap. The performance of the two variants of our method FACIMS ($\beta = 0, v = \bar{v}$) and FACIMS ($\beta$) drops slightly. Our method has more balanced recall over all four classes. In Figure 4, we include all the methods including the two variants of our FACIMS. It is a bit messy but still we can see the superiority of our method.

Next, we investigate the performance of our method when $A \not\perp Y$ and the classes are imbalanced. Our synthetic dataset is generated as follows: We consider two classes ($Y = 0$ or $Y = 1$) and two groups ($A = a$ or $A = b$). We set $n_1 = 5 \times n_0$, where $n_i$ is the number of instances in $i$-th class so that the data is class imbalanced. The ratio of group $a$ to group $b$ in class 1 is set to $\pi : 1$ for some $\pi > 0$. On the contrary, the ratio of group $a$ to group $b$ in class 0 is set to $1 : \pi$.

Let $n_0 = 80$ and $n_1 = 400$.

- If $\pi = 1$, the ratio of group $a$ to group $b$ in class 0 is $40 : 40 = 1 : 1$ and in class 1 is $200 : 200 = 1 : 1$. Also, the ratio of group $a$ to group $b$ in the whole population is $240 : 240 = 1 : 1$. Hence, $Y$ and $A$ are independent.

- If $\pi = 7$, the ratio of group $a$ to group $b$ in class 0 is $10 : 70 = 1 : 7$ and in class 1 is $350 : 50 = 7 : 1$. But, the ratio of group $a$ to group $b$ in the whole population is $360 : 120 = 3 : 1$. Hence, $Y$ and $A$ are dependent, i.e., $A \not\perp Y$.

Table 5 provides the performance of FACIMS on the above synthetic imbalanced dataset. From this table, we can see that when there is a correlation between groups and labels ($\pi = 7$), there is a slight improvement in both accuracy and balanced accuracy. Further, the degeneration in demographic parity is significant in comparison with group sufficiency and equalized odds. This is consistent with [Barocas et al., 2019] and our discussion in Section 2; that is, if $A \not\perp Y$, then equalized odds, demographic parity, and group sufficiency could not be simultaneously achieved.
Table 3: Classification results (mean ± standard deviation) for 5 repeats of different methods on Drug Consumption dataset. "↑" indicates the larger the better while "↓" indicates the smaller the better. The best one in each column is bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Balanced Accuracy ↑</th>
<th>Demographic Parity ↓</th>
<th>Equalized Odds ↓</th>
<th>Sufficiency Gap ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIIL</td>
<td>0.2549±0.0029</td>
<td>0.0199±0.0079</td>
<td>0.0673±0.0279</td>
<td>0.1602±0.0411</td>
</tr>
<tr>
<td>FSCS</td>
<td>0.2471±0.0101</td>
<td>0.0141±0.0086</td>
<td>0.0399±0.0402</td>
<td>0.2555±0.0456</td>
</tr>
<tr>
<td>FAMS</td>
<td>0.2624±0.0308</td>
<td>0.0087±0.0021</td>
<td>0.2972±0.0323</td>
<td>0.1485±0.0546</td>
</tr>
<tr>
<td>ERM</td>
<td>0.2492±0.0013</td>
<td>0.0158±0.0052</td>
<td>0.0061±0.0105</td>
<td>0.1062±0.0283</td>
</tr>
<tr>
<td>BERM</td>
<td>0.2434±0.0259</td>
<td>0.0197±0.0016</td>
<td>0.1626±0.0259</td>
<td>0.1406±0.0875</td>
</tr>
<tr>
<td>FACIMS (β = 0, v = ¯v)</td>
<td>0.2644±0.0366</td>
<td><strong>0.4968±0.0724</strong></td>
<td>0.2198±0.1071</td>
<td>0.1329±0.0509</td>
</tr>
<tr>
<td>FACIMS (β = 0)</td>
<td>0.2716±0.0277</td>
<td>0.0192±0.0027</td>
<td>0.2158±0.0410</td>
<td>0.1231±0.0301</td>
</tr>
<tr>
<td>FACIMS</td>
<td><strong>0.2767±0.0545</strong></td>
<td>0.0257±0.0136</td>
<td>0.1469±0.1115</td>
<td><strong>0.0948±0.0224</strong></td>
</tr>
</tbody>
</table>

Table 4: Recall (mean ± standard deviation) and timing (hours:minutes:seconds) results for 5 repeats of different methods on the Drug Consumption dataset. "↑" indicates the larger the better while "↓" indicates the smaller the better. The best one in each column is bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall 0 ↑</th>
<th>Recall 1 ↑</th>
<th>Recall 2 ↑</th>
<th>Recall 3 ↑</th>
<th>Time ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIIL</td>
<td>0.0000±0.0000</td>
<td>0.0000±0.0000</td>
<td>0.9256±0.0345</td>
<td>0.0938±0.0417</td>
<td>0:04:27</td>
</tr>
<tr>
<td>FSCS</td>
<td>0.0000±0.0000</td>
<td>0.0000±0.0000</td>
<td>0.9690±0.0043</td>
<td>0.0192±0.0385</td>
<td>0:02:39</td>
</tr>
<tr>
<td>FAMS</td>
<td>0.2000±0.0632</td>
<td><strong>0.4968±0.0724</strong></td>
<td>0.2198±0.1071</td>
<td>0.1329±0.0509</td>
<td>0:12:37</td>
</tr>
<tr>
<td>ERM</td>
<td>0.0000±0.0000</td>
<td>0.0000±0.0000</td>
<td><strong>0.9968±0.0000</strong></td>
<td>0.0000±0.0000</td>
<td>0:00:31</td>
</tr>
<tr>
<td>BERM</td>
<td>0.0000±0.0000</td>
<td>0.2839±0.0899</td>
<td>0.6054±0.0978</td>
<td>0.0842±0.0455</td>
<td><strong>0:00:25</strong></td>
</tr>
<tr>
<td>FACIMS (β = 0, v = ¯v)</td>
<td>0.2200±0.1720</td>
<td>0.3226±0.1274</td>
<td>0.2783±0.0584</td>
<td>0.2368±0.0968</td>
<td>0:12:55</td>
</tr>
<tr>
<td>FACIMS (β = 0)</td>
<td>0.1000±0.0632</td>
<td>0.2710±0.0483</td>
<td>0.4327±0.0399</td>
<td><strong>0.2829±0.0186</strong></td>
<td>0:14:21</td>
</tr>
<tr>
<td>FACIMS</td>
<td><strong>0.6200±0.3709</strong></td>
<td>0.3355±0.4154</td>
<td>0.0960±0.1275</td>
<td>0.0553±0.1105</td>
<td>0:15:15</td>
</tr>
</tbody>
</table>

Figure 4: Boxplot of balanced accuracy and group sufficiency gap with 5 repeats for both Alzheimer's disease, credit card, and drug consumption datasets. Along each axis, the middle of each box is the mean and the box width is twice the length of the standard deviation. The more the box is on the right bottom (bigger balanced accuracy and smaller group sufficiency), the better the performance is. Here we include all the methods including the two variants of our method.

(a) Alzheimer’s Disease    (b) Credit Card    (c) Drug Consumption
Table 5: The performance of FACIMS on the synthetic imbalanced dataset with different $\pi$.

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>1</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.9000±0.0083</td>
<td>0.9240±0.0026</td>
</tr>
<tr>
<td>Balanced Accuracy</td>
<td>0.8767±0.0196</td>
<td>0.8885±0.0121</td>
</tr>
<tr>
<td>Demographic Parity</td>
<td>0.0381±0.0131</td>
<td>0.2544±0.0117</td>
</tr>
<tr>
<td>Equalized Odds</td>
<td>0.0560±0.0131</td>
<td>0.0904±0.0080</td>
</tr>
<tr>
<td>Sufficiency Gap</td>
<td>0.1256±0.0444</td>
<td>0.1866±0.0107</td>
</tr>
</tbody>
</table>