Fairness-Aware Class Imbalanced Learning on Multiple Subgroups

Davoud Ataee Tarzanagh[1] **Bojian Hou**[1] **Boning Tong**[1] **Qi Long**[1] **Li Shen**[1]

[1] University of Pennsylvania

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 ($\mathbb{E}[Y|f(X), A]$) across different subgroups $\mathcal{A} = 1, \ldots, A$ 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?

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

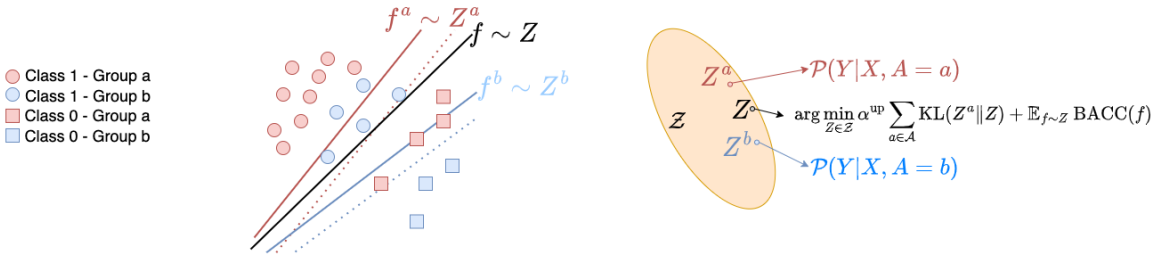


Figure 1: An illustration of the FACIMS model defined in (7). f^a and f^b maximize the margin for minority classes for groups a and b in (7b). In the upper level problem (7a), FACIMS finds $\mathbf{Z} \in \mathcal{Z}$ to achieve a small balanced accuracy while minimizing the discrepancy between $(\mathbf{Z}^{a,\star}, \mathbf{Z}^{b,\star})$. The approximation term $\text{KL}(\mathbf{Z}^{a,\star}|\mathbf{Z})$ is based on the distribution family $\mathcal{Z}$ (orange region). If the predefined $\mathcal{Z}$ has good expressive power, the approximation is treated as a small constant.

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 $\mathcal{O}(1/\sqrt{T})$ convergence rate of our proposed three-level optimization framework, corresponding to a $\mathcal{O}(\epsilon^{-2})$ 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 $\mathcal{P}(X, Y, A)$, 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 $\mathcal{P}(A, X)$.

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

$$\text{Population risk: } \mathcal{L}(f; \mathcal{P}) := \mathbb{E}_{\mathcal{P}}[\ell(y, \hat{y}_f(\mathbf{x}))], \tag{1a}$$

$$\text{Empirical risk: } \mathcal{L}(f; \mathcal{S}) := \tfrac{1}{n} \textstyle\sum_{i=1}^n \ell(y_i, \hat{y}_f(\mathbf{x}_i)). \tag{1b}$$

We denote the frequency of the k'th class via $\pi_k = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{P}}(y = k)$. 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

$$\text{Class-conditional risk: } f_k := \mathbb{E}_{\mathcal{P}_k}[\ell(y, \hat{y}_f(\mathbf{x}))], \tag{2a}$$

$$\text{Balanced risk: } \text{BACC}(f) := \tfrac{1}{K} \textstyle\sum_{k=1}^K f_k. \tag{2b}$$

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 Δ_j as the margin for each class, defined as follows:

$$\ell_\Delta(f; \mathbf{x}, y) = -\log \frac{e^{f(\mathbf{x})_y - \Delta_y}}{e^{f(\mathbf{x})_y - \Delta_y} + \sum_{j \neq y} e^{f(\mathbf{x})_j}}, \quad \text{where } \Delta_j = \frac{C}{n_j^{1/4}} \text{ for } j \in \{1, \ldots, K\}. \tag{LDAM}$$

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:

$$\ell_u(f;\mathbf{x},y) = -u_y \log \frac{e^{f(\mathbf{x})_y}}{\sum_{j=1}^{K} e^{f(\mathbf{x})_y}}, \quad \text{where } u_j = \frac{1}{1+(n_j-1)\mathbb{1}_{\text{epoch}\geq K}}. \tag{DRW}$$

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:

$$\ell(f,\mathbf{v};\mathbf{x},y) = -u_y \log \frac{e^{\gamma_y f(\mathbf{x})_y+\Delta_y}}{\sum_{j=1}^{k} e^{\gamma_j f(\mathbf{x})_j+\Delta_j}}. \tag{VS}$$

Here, $\mathbf{v} := (u_j, \Delta_j, \gamma_j)$ 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 $\mathcal{P}(X, Y, A) > 0$ 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:

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

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

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

$\mathbf{SGap}_f$ measures the extent of group sufficiency violation, induced by the predictor f, which is taken by the expectation over (X, A). Hence, $\mathbf{SGap}_f = 0$ 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 $\mathbf{SGap}_f$ 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 $f^{A,\texttt{Bayes}}$ associated with distribution $\mathcal{P}(X, Y, A)$ is defined as: $f^{A,\texttt{Bayes}}(X) = \mathbb{E}[Y|X, A]$.*

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.$|\mathcal{A}| < +\infty$ and follows uniform distribution with $p(A = a) = 1/|\mathcal{A}|$, then*

$$\textbf{\textit{SGap}}_f(A) \leq \frac{4}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \mathbb{E}_X \left[|f - f^{A,\texttt{Bayes}}| \big| A = a\right]. \tag{4}$$

Hence, $\mathbf{SGap}f(A)$ depends on the discrepancy between the predictor f and the A-group Bayes predictor $f^{A,\texttt{Bayes}}$. In other words, when considering different subgroups $A = a$, the optimal predictor f should closely align with all the group Bayes predictors $f^{A=a,\texttt{Bayes}}$, 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 $\mathbf{Z}$, which assigns higher scores to predictors that are favorable based on the available data. In the context of the Bayes framework, the predictor is sampled from the posterior distribution, represented as $\tilde{f} \sim \mathbf{Z}$. During the inference stage, the predictor's output is computed as the expectation of the learned predictive distribution $\mathbf{Z}$: $f(X) = \mathbb{E}_{\tilde{f}\sim\mathbf{Z}}\tilde{f}(X)$.

In practical scenarios, it is infeasible to optimize over the entire space of possible distributions. Therefore, we constrain the predictive distribution $\mathcal{Z}$ to a specific distribution family $\mathbf{Z} \in \mathcal{Z}$, such as the Gaussian distribution. Additionally, we denote $\mathbf{Z}^{a,\star} \in \mathcal{Z}$ as the optimal prediction distribution with respect to the subgroup $A = a$ within the distribution family $\mathcal{Z}$:

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

In general, $\mathbf{Z}^{a,\star} \neq \tilde{f}^{a,\texttt{Bayes}}$, since the distribution family $\mathcal{Z}$ 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* $\mathbf{Z}$ *is bounded as follows:*

$$\textbf{\textit{SGap}}_f \leq \mathcal{O}\left(\texttt{Optim} + \texttt{Approx}\right), \tag{5}$$

where

$$\texttt{Optim} := \frac{1}{|\mathcal{A}|} \sum_a \sqrt{KL(\mathbf{Z}^{a,\star} \| \mathbf{Z})},$$

$$\texttt{Approx} := \frac{1}{|\mathcal{A}|} \sum_a \sqrt{KL(\mathbf{Z}^{a,\star} \| \mathcal{P}(Y|X, A=a))}.$$

Minimizing the `Optim` term ensures that the learned distribution $\mathbf{Z}$ is both fair and informative for making predictions. On the other hand, the `Approx` term represents the KL divergence between the optimal distribution $\mathbf{Z}^{a,\star}$ and $\mathcal{P}$. If the distribution family $\mathcal{Z}$ 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 sub-groups with potentially improved generalization bound and $\textbf{SGap}_f$. We begin with formulating the loss function design as a bilevel optimization over hyperparameters $\mathbf{v}$ and a distribution $\mathcal{Z}$. Assume each group $a \in \mathcal{A}$ has a fine-tuning training set $\mathcal{T}^a = \cup_{i=1}^n \{(x_i^a, y_i^a)\}$ and a separate validation set $\mathcal{V}^a = \cup_{i=1}^n \{(x_i^a, y_i^a)\}$, where data are independently and identically distributed (i.i.d.) and drawn from the per-task data distribution $\mathcal{P}^a$. Following [Li et al., 2021], define the *empirical risk* and the balanced *empirical risk* over a finite-sample dataset $\mathcal{S}$ as $\mathcal{L}_{\text{vs}}(\tilde{f}, \mathbf{v}; \mathcal{S}) := 1/n \sum_{i=1}^n \ell(\tilde{f}, \mathbf{v}; \mathbf{x}_i, y_i)$ and $\mathcal{L}_{\text{bal}}(\tilde{f}; \mathcal{S}) := 1/n \sum_{i=1}^n \ell(\tilde{f}, \bar{\mathbf{v}}; \mathbf{x}_i, y_i)$. Here, $\bar{\mathbf{v}}$ can be manually adjusted using (DRW), (LDAM), and (VS).

Let $\mathbf{Z}$ 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:

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

$$\text{s.t. } \mathbf{Z}^{a,*} = \arg\min_{\mathbf{Z}^a \in \mathcal{Z}} \ \alpha^{\text{low}} \text{KL}(\mathbf{Z}^a \| \mathbf{Z}) \tag{6b}$$

$$+ \mathbb{E}_{\tilde{f}^a \sim \mathbf{Z}^a} \mathcal{L}_{\text{vs}}(\tilde{f}^a, \mathbf{v}; \mathcal{T}^a), \ \ \forall a \in \mathcal{A}.$$

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

In the upper-level problem (6a), we update $\mathbf{Z}$ by minimizing the average KL-divergence between different $\mathbf{Z}^{a,\star}$, controlling the upper bound of $\textbf{SGap}_f$ 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 $\mathcal{P}[Y|X, A]$ 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 $\mathcal{Z}$ to learn global informative $\mathbf{Z}$ with parameters $(\boldsymbol{\theta}, \boldsymbol{\sigma})$. For each subgroup $A = a$, we also learn group-specific parameters $(\boldsymbol{\theta}^a, \boldsymbol{\sigma}^a)$ for $\mathbf{Z}^a$ in $\mathcal{Z}$. The Isotropic Gaussian distribution is selected for computational efficiency in optimization, but other differentiable distributions can also be used for parameter density functions.

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

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

$$\mathbf{w}^a \sim \mathcal{N}(\boldsymbol{\theta}^a, \boldsymbol{\sigma}^a) = \prod_{i=1}^d \mathcal{N}(\boldsymbol{\theta}^a[i], \boldsymbol{\sigma}^a[i]), \ \ \forall \in \mathcal{A}.$$

To enhance the convergence to a flat minimum and effectively avoid saddle points for minority classes, we integrate the sharpness-aware minimization (SAM) algorithm [Foret et al., 2020] into (6b). SAM is a recently introduced technique that improves generalization performance by jointly minimizing the loss value and the loss sharpness, leveraging the geometry of the loss landscape. Given a perturbation parameter $\beta > 0$ and the empirical risk $\mathcal{L}(\tilde{f}_{\mathbf{w}}; \mathcal{S})$, the goal of training is to choose $\mathbf{w}$ having low population loss

FACIMS

$$\min_{\mathbf{v},\mathbf{Z}\sim\mathcal{Z}} \sum_{a\in\mathcal{A}} \alpha^{\text{up}}\text{KL}\left(\mathbf{Z}^{a,*}\|\mathbf{Z}\right) + \mathbb{E}_{\mathbf{w}\sim\mathbf{Z}}\,\mathcal{L}_{\text{bal}}(\tilde{f}_{\mathbf{w}};\mathcal{V}^a), \tag{7a}$$

$$\text{s.t.}\quad \mathbf{Z}^{a,*} \in \arg\min_{\mathbf{Z}^a\in\mathcal{Z}} \max_{\|\boldsymbol{\epsilon}^a\|\le\beta^a} \alpha^{\text{low}}\,\text{KL}(\mathbf{Z}^a\|\mathbf{Z}) + \mathbb{E}_{\mathbf{w}^a\sim\mathbf{Z}^a}\,\mathcal{L}_{\text{vs}}(\tilde{f}_{\mathbf{w}^a+\boldsymbol{\epsilon}^a},\mathbf{v};\mathcal{T}^a), \quad \forall a\in\mathcal{A}. \tag{7b}$$

Algorithm 1 An Alternating Stochastic Gradient Method for FACIMS

1: **Input:** Parameters w.r.t. distribution $\mathbf{Z}$:$(\boldsymbol{\theta},\boldsymbol{\sigma})$; regularizations $(\alpha^{\text{low}},\alpha^{\text{up}})$; constants $\{\beta^a\}$, stepsizes $(\gamma^{\text{up}},\gamma^{\text{low}})$.
2: **for** $t=0\ldots T-1$ **do**
3: Sample dataset $S_t^a=(\mathcal{T}_t^a,\mathcal{V}_t^a)$, where $a\in\mathcal{A}'\subseteq\mathcal{A}$.
4: **for** $a\in\mathcal{A}$ **do**
5: Update $\boldsymbol{\epsilon}^a$ in the lower level:

$$\boldsymbol{\epsilon}_{t+1}^a \longleftarrow \arg\max_{\|\boldsymbol{\epsilon}^a\|\le\beta^a} \alpha^{\text{low}}\,\text{KL}(\mathbf{Z}_t^a\|\mathbf{Z}_t) + \mathbb{E}_{\mathbf{w}^a\sim\mathbf{Z}_t^a}\mathcal{L}_{\text{vs}}(\tilde{f}_{\mathbf{w}^a+\boldsymbol{\epsilon}^a},\mathbf{v};\mathcal{T}_t^a) \tag{8}$$

6: Update $\mathbf{Z}^a=\mathcal{N}(\boldsymbol{\theta}^a,\boldsymbol{\sigma}^a)$ using SGD (with step size γ^{low}) in the middle level:

$$\mathbf{Z}_{t+1}^a \longleftarrow \arg\min_{\mathbf{Z}^a\in\mathcal{Z}} \alpha^{\text{low}}\,\text{KL}(\mathbf{Z}^a\|\mathbf{Z}_t) + \mathbb{E}_{\mathbf{w}^a\sim\mathbf{Z}^a}\mathcal{L}_{\text{vs}}(\tilde{f}_{\mathbf{w}^a+\boldsymbol{\epsilon}_{t+1}^a},\mathbf{v};\mathcal{T}_t^a). \tag{9}$$

7: **end for**
8: Update $(\mathbf{Z},\mathbf{v})$ using SGD (with step size γ^{up}) in the upper level:

$$(\mathbf{Z}_{t+1},\mathbf{v}_{t+1}) \longleftarrow \arg\min_{\mathbf{Z}\sim\mathcal{Z},\mathbf{v}} \sum_{a\in\mathcal{A}} \alpha^{\text{up}}\text{KL}(\mathbf{Z}_{t+1}^a\|\mathbf{Z}) + \mathbb{E}_{\mathbf{w}\sim\mathbf{Z}}\mathcal{L}_{\text{bal}}(\tilde{f}_{\mathbf{w}};\mathcal{V}_t^a). \tag{10}$$

9: **end for**
10: **Return:** Parameter of distribution $\mathbf{Z}_T$: $(\boldsymbol{\theta}_T,\boldsymbol{\sigma}_T)$

$\mathcal{L}(\tilde{f}_{\mathbf{w}};\mathcal{P})$. SAM achieves this via the problem

$$\min_{\tilde{f}_{\mathbf{w}}} \max_{\|\epsilon\|\le\beta} \mathcal{L}(\tilde{f}_{\mathbf{w}+\boldsymbol{\epsilon}};\mathcal{S}). \tag{SAM}$$

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

$$\max_{\|\epsilon\|\le\beta} \left[\mathcal{L}(\tilde{f}_{\mathbf{w}+\boldsymbol{\epsilon}};\mathcal{S}) - \mathcal{L}(\tilde{f}_{\mathbf{w}};\mathcal{S})\right]$$

then (SAM) essentially minimizes the sum of the sharpness and the empirical loss $\mathcal{L}(\tilde{f}_{\mathbf{w}};\mathcal{S})$.

We incorporate (SAM) into (6b) and propose (7) by introducing a set of positive constants $\{\beta^a\}_{a\in\mathcal{A}}$. 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 $\mathcal{A}'$ such that $|\mathcal{A}'| << |\mathcal{A}|$ for memory saving.

4 THEORETICAL ANALYSIS OF FACIMS

Next, we analyze the performance of the FACIMS method.

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

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

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

Table 1: Statistical summary of the datasets including class and sensitive feature information.

Dataset	#Instance	#Features	Class	Class Distr.	Sensitive Feature	Sensitive Feature Distr.
Alzheimer's Disease	5137	17	AD / MCI	21% / 79%	Race	93.75% / 3.20% / 1.88% / 1.17%
Credit Card	30,000	22	Credible / Not Credible	22% / 77%	Education Level	46.77% / 35.28% / 16.39% / 0.93% / 0.41% / 0.17% / 0.05%
Drug Consumption	1885	9	Never used / Not used in the past year / Used in the past year / Used in the past day	1.81% / 5.41% / 65.98% / 26.80%	Education Level	6.74% / 6.90% / 26.86% / 14.28% / 25.48% / 15.02% / 4.72%

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 stepsizes* $\gamma^{\text{up}}, \gamma^{\text{low}}$ *and sharpness parameter* $\beta = \mathcal{O}(1)$*, with some proper constants, we can get that the iterates* $\{(\boldsymbol{\theta}_T, \boldsymbol{\sigma}_T, \mathbf{v}_T)\}$ *generated by Algorithm 1 satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(\boldsymbol{\theta}_T, \boldsymbol{\sigma}_T, , \mathbf{v}_T)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \quad (11)$$

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

Next, we establish the generalization performance.

Theorem 6. *Assume the function* $\mathcal{L}(\cdot)$ *is bounded for any i.i.d* $\mathcal{S}$*. Let* $F(\cdot;\mathcal{P}) = \mathbb{E}_{\mathcal{S}\sim\mathcal{P}}\left[F(\cdot;\mathcal{S})\right]$*. Assume* $F(\boldsymbol{\theta}, \boldsymbol{\sigma}, \mathbf{v}; \mathcal{P}) \leq \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\beta^2\mathbb{I})}\left[F(\boldsymbol{\theta}+\epsilon, \boldsymbol{\sigma}+\epsilon, \mathbf{v}; \mathcal{P})\right]$ *at the stationary point of* (7) *denoted by* $(\hat{\boldsymbol{\theta}}+\epsilon, \hat{\boldsymbol{\sigma}}+\epsilon, \hat{\mathbf{v}})$*. 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*

$$F(\boldsymbol{\Theta};\mathcal{P}) \leq \max_{\|\epsilon\|\leq\beta}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\beta^2\mathbb{I})}\left[F(\boldsymbol{\Theta}+\epsilon, \boldsymbol{\sigma}+\epsilon, \mathbf{v}; \mathcal{D})\right]$$
$$\leq \sqrt{\frac{(p+K)\ln\left(1+\frac{\|\hat{\boldsymbol{\Theta}}\|_2^2}{\beta^2}\left(1+\sqrt{\frac{\ln(n|\mathcal{A}|)}{p}}\right)^2\right)}{4n|\mathcal{A}|}}$$
$$+5\sqrt{\frac{\ln\frac{1}{\delta}+\ln(n|\mathcal{A}|)}{4n|\mathcal{A}|}}. \quad (12)$$

Theorem 6 shows that the difference between the population loss and the empirical loss of FACIMS is bounded by $\tilde{\mathcal{O}}((p+K)/n|\mathcal{A}|)$ that vanishes as the number of group-specific training data goes to infinity. Note that the bound in (12) is a function of β. 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, Certifi-

[1] http://adni.loni.usc.edu

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

[3] https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

Table 2: Numerical results (mean ± standard deviation) for 5 repeats of different methods on Alzheimer's disease (AD) and Credit Card (CC) datasets regarding six measurements. Time is in the format of "hours:minutes:seconds". FACIMS-I means FACIMS ($\beta = 0$), and FACIMS-II means FACIMS ($\beta = 0$, $v = \bar{v}$). "↑" indicates the larger the better while "↓" indicates the smaller the better. The best one in each column is bold.

Data	Method	Balanced Accuracy ↑	Demographic Parity ↓	Equalized Odds ↓	Sufficiency Gap ↓	Recall 0 ↑	Recall 1 ↑	Time ↓
AD	EIIL	.8639±.0199	.0764±.0176	.1015±.0529	.1193±.0206	.9288±.0119	.7991±.0409	0:03:32
	FSCS	.8498±.0485	.0711±.0287	.1650±.1008	.1254±.0528	.9504±.0426	.7493±.1018	0:08:05
	FAMS	.8369±.0136	**.0431±.0210**	.1444±.0435	.1328±.0273	.7624±.0077	.9114±.0096	0:09:51
	ERM	.8687±.0136	.0550±.0196	.1143±.0390	.1701±.0387	**.9883±.0053**	.7491±.0430	**0:00:51**
	BERM	.8886±.0042	.0869±.0204	.0813±.0129	.1456±.0330	.9854±.0043	.7918±.0520	0:02:24
	FACIMS-II	.8839±.0079	.0747±.0182	.0868±.0130	.1167±.0139	.8456±.0148	**.9222±.0043**	0:09:58
	FACIMS-I	.8887±.0066	.0893±.0080	**.0450±.0049**	.1059±.0060	.8780±.0104	.8994±.0148	0:13:38
	FACIMS	**.8897±.0098**	.0765±.0208	.0616±.0142	**.1052±.0197**	.8832±.0072	.8962±.0054	0:15:26
CC	EIIL	.6357±.0267	.0834±.0200	.1723±.0515	.1266±.023	.7897±.0176	.4817±.0448	0:03:30
	FSCS	.5976±.0277	.0850±.0137	.2000±.0456	.2007±.0039	.8953±.0130	.3000±.0685	0:42:10
	FAMS	.6542±.0098	.0746±.0066	.1859±.0368	.1352±.0106	.8194±.0374	.4890±.0270	0:10:21
	ERM	.6104±.0111	.0599±.0173	**.1577±.0175**	.2760±.0710	**.9919±.0233**	.2289±.0820	**0:02:07**
	BERM	.6570±.0106	.1060±.0125	.1631±.0304	.2315±.0623	.8717±.0191	.4423±.0146	0:02:09
	FACIMS-II	.6446±.0163	.0707±.0073	.1973±.0358	.1340±.0147	.8002±.0374	.4890±.0270	0:10:03
	FACIMS-I	.6768±.0040	.0750±.0105	.1951±.0524	.1396±.0081	.8081±.0114	.5455±.0098	0:14:07
	FACIMS	**.6799±.0374**	**.0593±.0070**	.1567±.0230	**.1264±.0145**	.8136±.0054	**.5462±.0017**	0:14:18

cate diploma, University degree, Masters, Doctorate). The data information is summarized in Table 1 below. As can be seen, the class distribution shows that the dataset suffers from heavy label imbalance.

Baselines To validate the effectiveness of our method, FACIMS, we compare it with seven baseline methods.

- EIIL [Creager et al., 2021]: An Invariant Risk Minimization (IRM) based approach that promotes group sufficiency.
- FSCS [Lee et al., 2021]: An approach that adopts the conditional mutual information constraint to improve group sufficiency.
- FAMS [Shui et al., 2022c]: A bilevel framework that considers maintaining both the accuracy and group sufficiency gap for multiple subgroups.
- ERM: Empirical Risk Minimization using a four-layer fully connected neural network trained with cross-entropy loss.
- BERM: ERM with a balanced cross-entropy loss, incorporating class proportions as weights similar to [Cao et al., 2019].
- FACIMS ($\beta = 0$, $\mathbf{v} = \bar{\mathbf{v}}$): Our method without the lower level. Besides, in the upper level, we manually adjust the logits using the proportion of class [Menon et al., 2020, Kini et al., 2021] instead of learning the hyperparameter for logits adjustment.
- FACIMS ($\beta = 0$): Our method without the lower level which aims to flatten the sharp landscape of the objective in the middle level.

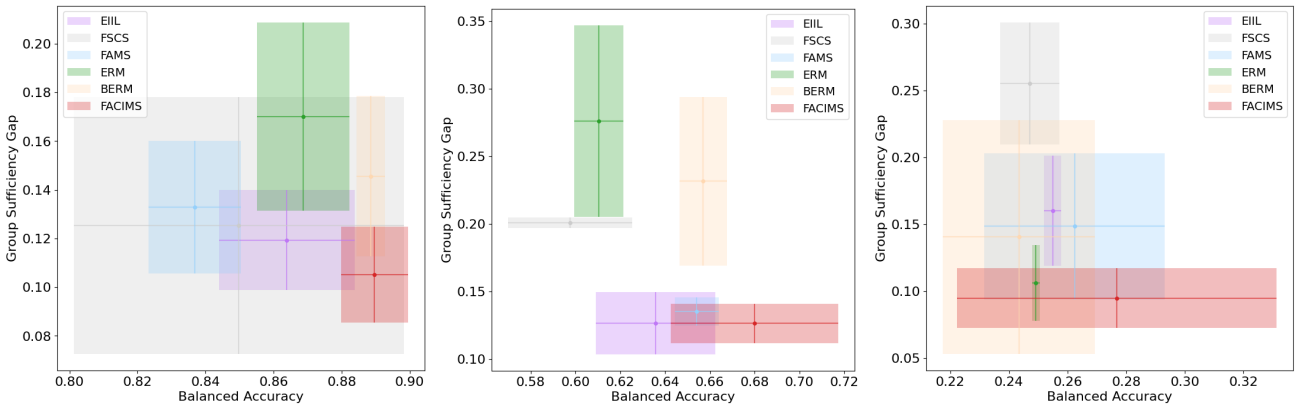
We set α^{up} and α^{low} to be 0.7. We use the grid of [0.1, 0.01, 0.001] to search the learning rate for global model and local models and report the results over five independent repeats.

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 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 6.5% and 4.0% respectively, compared to ERM and BERM which do not address this issue. Although EIIL, FSCS, and FAMS specifically target the group sufficiency problem and achieve lower sufficiency gaps than ERM and BERM, our method still outperforms these three baseline methods by improving the 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-



(a) Alzheimer's Disease (b) Credit Card (c) Drug Consumption

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.

ing the hyperparameters (as in [Menon et al., 2020]) further decreases the balanced accuracy and group sufficiency gap. However, our bilevel structure for addressing fairness ensures that the group sufficiency gap remains good despite these drops.

On credit card dataset, we have similar results. As for balanced accuracy, our method FACIMS improves the performance by 4.4%, 8.2%, 2.6%, 7.0%, and 2.3% comapred to EIIL, FSCS, FAMS, ERM and BERM. When it comes to the group sufficiency gap, the performance of our method is improved by 0.2%, 7.4%, 0.8%, 15%, and 11% comapred to the same baseline methods as metioned above. The performances of FACIMS ($\beta = 0$) and FACIMS ($\beta = 0, \mathbf{v} = \bar{\mathbf{v}}$) drop slightly regarding both measurements.

To provide a more intuitive visualization of the results, we present boxplots in Figure 2. Each axis represents a measurement, where the mean value is represented by the middle of the box and the box width corresponds to twice the length of the standard deviation. The model's performance is reflected by the position of the box, with improved performance observed towards the bottom right corner, indicating higher balanced accuracy and lower group sufficiency gap. For clarity, we have excluded two variants of our method from Figure 2. The complete figures can be found in the appendix. Figure 2 highlights that our method is positioned towards the bottom right corner, indicating improved performance compared to other methods.

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 $13\,\text{min} \times 5 = 65\,\text{min}$, which is around four times longer than our differentiable three-level FACIMS approach.

Influence of α^{low} In the middle level, the parameter α^{low} determines the attention given to $\text{KL}(\mathbf{Z}^a|\mathbf{Z})$. A higher value of α^{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 α^{low}: 0.01, 0.1, 0.2, and 1. Figure 3 illustrates the Accuracy-**SGap**$_f$ curve under varying α^{low} on the Alzheimer's disease dataset. The figure demonstrates a clear

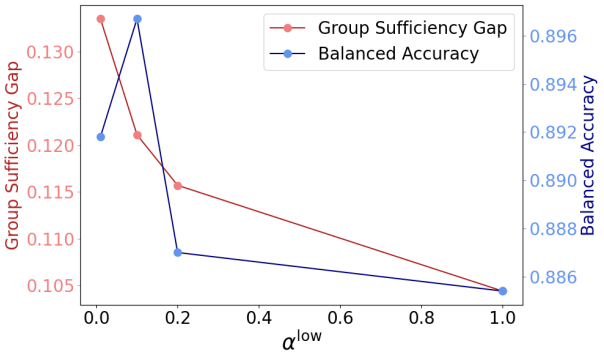


Figure 3: Accuracy-**SGap**$_f$ curve under different α^{low} in Alzheimer's disease dataset.

trend: as α^{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 OPTIMZIATION

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

Momin Abbas, Quan Xiao, Lisha Chen, Pin-Yu Chen, and Tianyi Chen. Sharp-maml: Sharpness-aware model-agnostic meta learning. In *International Conference on Machine Learning*, pages 10–32. PMLR, 2022.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

Tina Behnia, Ganesh Ramachandra Kini, Vala Vakilian, and Christos Thrampoulidis. On the implicit geometry of cross-entropy parameterizations for label-imbalanced data. In *International Conference on Artificial Intelligence and Statistics*, pages 10815–10838. PMLR, 2023.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106: 249–259, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153:235–256, 2007.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Hongsheng Hu, Zoran Salcic, Gillian Dobbie, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *arXiv preprint arXiv:2103.07853*, 2021.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34: 18970–18983, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Joshua K Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W Wornell. Fair selective classification via sufficiency. In *International Conference on Machine Learning*, pages 6076–6086. PMLR, 2021.

Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems*, 34:3163–3177, 2021.

Mingquan Lin, Yuyun Xiao, Bojian Hou, Tingyi Wanyan, Mohit Manoj Sharma, Zhangyang Wang, Fei Wang, Sarah Van Tassel, and Yifan Peng. Evaluate underdiagnosis and overdiagnosis bias of deep learning model on primary open-angle glaucoma diagnosis in under-served patient populations. *arXiv preprint arXiv:2301.11315*, 2023.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019a.

Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4051–4060. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/liu19f.html.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/abs/10.1126/science.aax2342.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, and R Venkatesh Babu. Escaping saddle points for effective generalization on class-imbalanced data. *arXiv preprint arXiv:2212.13827*, 2022.

Li Shen, Paul M Thompson, Steven G Potkin, et al. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging Behav*, 8 (2):183–207, 2014.

Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In *ICML*, 2022a.

Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. *arXiv preprint arXiv:2205.13316*, 2022b.

Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *arXiv preprint arXiv:2210.10837*, 2022c.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019.

Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. Fairness-aware class imbalanced learning. *arXiv preprint arXiv:2109.10444*, 2021.

Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179. PMLR, 2022.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Michael W Weiner, Dallas P Veitch, Paul S Aisen, et al. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia*, 13 (4):e1–e85, 2017.

Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019.

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 ϕ obtained by concatenating Θ^a for all a, given by $\phi = [\Theta^1, \cdots, \Theta^A]^\top$. We then define

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

Then, from the properties of implicit functions, we get

$$\nabla F(\Theta) = -\nabla_{\theta\phi} g(\theta, \phi) \nabla^{-1}_{\phi\phi} g(\theta, \phi) \nabla_\phi f(\phi). \tag{13}$$

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 $\xi^a := (x, y) \sim \mathcal{S}^a, \psi^a := (x, y) \sim \mathcal{S}^a, \xi^{a,\prime} := (x, y) \sim \mathcal{S}^{a'}$, so the corresponding κ-batch gradient and Hessian estimators for bilvel-type methods can be written as

$$\begin{aligned}
\nabla\mathcal{L}(\theta; \mathcal{S}^a, \xi^a) &= \frac{1}{\kappa} \sum_{\xi^a\sim\mathcal{S}^a} \nabla l(\theta, x, y), \\
\nabla^2\mathcal{L}(\theta; \mathcal{S}^a, \psi^a) &= \frac{1}{\kappa} \sum_{\psi^a\sim\mathcal{S}^a} \nabla^2 l(\theta, x, y), \\
\nabla\mathcal{L}(\theta; \mathcal{D}'^a, \xi'^a) &= \frac{1}{\kappa} \sum_{\xi^{a\prime}\sim\mathcal{S}^{a\prime}} \nabla l(\theta, x, y).
\end{aligned}$$

We rewrite the updates of Algorithm 1 as follows:

$$\begin{aligned}
\Theta_{t+1} &= \Theta_t - \gamma^{\text{up}} \sum_{a\in\mathcal{A}} (I - \gamma^{\text{low}} \nabla^2\mathcal{L}(\Theta_t + \epsilon^a(\Theta_t); \mathcal{S}^a, \psi^a)) \nabla\mathcal{L}(\hat{\Theta}^a_{t+1}; \mathcal{S}^a, \xi^a); \\
\hat{\Theta}^a_{t+1} &= \Theta^a_t - \gamma^{\text{low}} \nabla\mathcal{L}(\Theta_t + \epsilon^a(\Theta_t); \mathcal{S}^a, \xi^a) \\
&= \Theta^a_t - \gamma^{\text{low}} \left(\nabla\mathcal{L}(\Theta_t + \epsilon^a(\Theta_t); \mathcal{S}^a, \xi^a) - \frac{\epsilon(\Theta_t)}{\gamma^{\text{low}}} \right).
\end{aligned}$$

Since $\nabla\mathcal{L}(\Theta; \mathcal{S}^a), \nabla^2\mathcal{L}(\Theta; \mathcal{S}^a)$ are Lipschitz continuous with ℓ_1, ℓ_2 according to Assumption A, then we have that

$$\begin{aligned}
\|\nabla\mathcal{L}(\Theta_t + \epsilon^a(\Theta_t); \mathcal{S}^a) - \frac{\epsilon(\Theta_t)}{\gamma^{\text{low}}} - \nabla\mathcal{L}(\Theta_t; \mathcal{S}^a)\| &\le \ell_1\beta \\
\|\nabla^2\mathcal{L}(\Theta_t + \epsilon^a(\Theta_t); \mathcal{S}^a) - \nabla^2\mathcal{L}(\Theta_t; \mathcal{S}^a)\| &\le \ell_2\beta.
\end{aligned}$$

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_* as

$$\hat{\mathcal{R}}_i(\mathcal{F}) = \frac{1}{n_i} \mathbb{E}_\xi [\sup_{f\in\mathcal{F}} \sum_{j\in S_i} \xi_j [f(x_j)_i - \max_{i'\neq i} f(x_j)_{i'}]],$$

$$\hat{\mathcal{R}}_{i,a}(\mathcal{F}) = \frac{1}{n_{i,a}}\mathbb{E}_\xi[\sup_{f\in\mathcal{F}}\sum_{j\in S_{i,a}}\xi_j[f(x_j)_i - \max_{i'\neq i} f(x_j)_{i'}]],$$

where ξ_j is i.i.d. drawn from a uniform distribution $\{-1, 1\}$.

Lemma 7. *Let*

$$\hat{\mathcal{L}}_{\gamma,i}[f] = \mathbb{P}_{x\sim\hat{\mathcal{P}}_i}(\max_{j\neq i} f(x)_j > f(x)_i - \gamma), \tag{14a}$$

$$\hat{\mathcal{L}}_{\gamma,(i,a)}[f] = \mathbb{P}_{x\sim\hat{\mathcal{P}}_{i,a}}(\max_{j\neq i} f(x)_j > f(x)_i - \gamma). \tag{14b}$$

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

$$\mathcal{L}_*[f] \lesssim \hat{\mathcal{L}}_{\gamma,*}[f] + \frac{1}{\gamma}\hat{\mathcal{R}}_*(\mathcal{F}) + \epsilon_*(n_*, \delta, \gamma_*). \tag{15}$$

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

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

Proof. Let

$$\rho := (\mathbf{Z}_1 \otimes \mathbf{Z}_2 \otimes \cdots \otimes \mathbf{Z}_{|\mathcal{A}|}) \qquad \text{and} \qquad \pi := (\mathbf{Z} \otimes \mathbf{Z} \otimes \cdots \otimes \mathbf{Z})_{|\mathcal{A}| \text{ times}}$$

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

$$\begin{aligned}\frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}_a\sim\mathbf{Z}^a}\mathcal{L}_a^{\text{BCE}}(\tilde{f}_a) \le \frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}_a\sim\mathbf{Z}^a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a) \\ + \frac{1}{\lambda}(\text{KL}(\mathbf{Z}_1 \otimes \cdots \otimes \mathbf{Z}_{|\mathcal{A}|} \| \mathbf{Z} \otimes \cdots \otimes \mathbf{Z}) + \log(\frac{1}{\delta})) + \frac{\lambda L}{8|\mathcal{A}|n}\end{aligned}$$

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

$$\begin{aligned}\frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}\sim\mathbf{Z}^a}\mathcal{L}_a^{\text{BCE}}(\tilde{f}) &\le \frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}\sim\mathbf{Z}^a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}) + L\sqrt{\frac{1}{2|\mathcal{A}|n}(\sum_a \text{KL}(\mathbf{Z}^a\|\mathbf{Z}) + \log(\frac{1}{\delta}))} \\ &\le \frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}\sim\mathbf{Z}^a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}) + \frac{L}{\sqrt{|\mathcal{A}|n}}\sum_a\sqrt{\text{KL}(\mathbf{Z}^a\|\mathbf{Z})} + L\sqrt{\frac{\log(1/\delta)}{|\mathcal{A}|n}}.\end{aligned} \tag{16}$$

Now, let $P_{\cdot,0} = P = \mathcal{N}(\mathbf{0}, \sigma_P^2\mathbf{I})$, $P_{\cdot} = \mathbf{Z} = \mathcal{N}(\cdot, \alpha^2\mathbf{I})$, then

$$\begin{aligned}D_{KL}(\mathbf{Z}\|P) &= \frac{1}{2}\left\{\text{tr}\left(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_\mathbf{Z}\right) + \left(\boldsymbol{\mu}_P - \boldsymbol{\mu}_\mathbf{Z}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_P - \boldsymbol{\mu}_\mathbf{Z}) - k + \ln\frac{|\boldsymbol{\Sigma}_P|}{|\boldsymbol{\Sigma}_\mathbf{Z}|}\right\} \\ &= \frac{1}{2}\left[\frac{k\alpha^2 + \|\cdot\|_2^2}{\sigma_P^2} - k + k\ln\left(\frac{\sigma_P^2}{\alpha^2}\right)\right].\end{aligned}$$

Let $T = \{c\exp((1-j)/k) \mid j \in \mathbb{N}\}$ be the set of values for σ_P^2. If for any $j \in \mathbb{N}$, the PAC-Bayesian bound in (16) holds for $\sigma_P^2 = c\exp((1-j)/k)$ with probability $1-\delta_j$ with $\delta_j = \frac{6\delta}{\pi^2 j^2}$, then by the union bound, all bounds w.r.t. all $\sigma_P^2 \in T$ hold simultaneously with probability at least $1 - \sum_{j=1}^\infty \frac{6\delta}{\pi^2 j^2} = 1 - \delta$.

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 $j = \left\lfloor 1 - k\ln\left(\left(\beta^2 + \|\cdot\|_2^2/k\right)/c\right)\right\rfloor$. By setting $c = \beta^2(1+\exp(4n/k))$, then $\ln\left(\left(\beta^2 + \|\cdot\|_2^2/k\right)/c\right) < 0$, thus we can ensure that $j \in \mathbb{N}$. Furthermore, for $\sigma_P^2 = c\exp((1-j)/k)$, we have:

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

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

The KL-divergence term can be further bounded as

$$\begin{aligned}
D_{KL}(\mathbf{Z}\|P) &= \frac{1}{2}\left[\frac{k\beta^2+\|\cdot\|_2^2}{\sigma_P^2} - k + k\ln\left(\frac{\sigma_P^2}{\beta^2}\right)\right] \\
&\leq \frac{1}{2}\left[\frac{k\beta^2+\|\cdot\|_2^2}{\beta^2+\|\cdot\|_2^2/k} - k + k\ln\left(\frac{\exp(1/k)\left(\beta^2+\|\cdot\|_2^2/k\right)}{\beta^2}\right)\right] \\
&= \frac{1}{2}\left[k\ln\left(\frac{\exp(1/k)\left(\beta^2+\|\cdot\|_2^2/k\right)}{\beta^2}\right)\right] \\
&= \frac{1}{2}\left[1 + k\ln\left(1+\frac{\|\cdot\|_2^2}{k\beta^2}\right)\right].
\end{aligned}$$

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

$$\begin{aligned}
\ln\frac{1}{\delta_j} &= \ln\frac{1}{\delta} + \ln\frac{\pi^2 j^2}{6} \\
&\leq \ln\frac{1}{\delta} + \ln\frac{\pi^2 k^2 \ln^2\left(c/\left(\beta^2+\|\cdot\|_2^2/k\right)\right)}{6} \\
&\leq \ln\frac{1}{\delta} + \ln\frac{\pi^2 k^2 \ln^2\left(c/\beta^2\right)}{6} \\
&= \ln\frac{1}{\delta} + \ln\frac{\pi^2 k^2 \ln^2(1+\exp(4n/k))}{6} \\
&\leq \ln\frac{1}{\delta} + \ln\frac{\pi^2 k^2 (4n/k)^2}{6} \leq \ln\frac{1}{\delta} + 2\ln(6n).
\end{aligned}$$

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

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

■

C COMPUTING GAPS FROM THE DATA

Recall from (3),

$$\mathbf{SGap}_f(A) = \mathbb{E}[|\mathbb{E}[Y \mid f(X)] - \mathbb{E}[Y \mid f(X), A]|], \tag{17a}$$

$$\mathbf{PGap}_f(A) = \mathbb{E}[\mathbb{E}[f(X)] - \mathbb{E}[f(X)|A]], \tag{17b}$$

$$\mathbf{OGap}_f(A) = \mathbb{E}[\mathbb{E}[f(X)|Y] - \mathbb{E}[f(X)|Y, A]]. \tag{17c}$$

C.1 COMPUTING SUFFICENCY GAP

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

$$[0, \delta_1], [\delta_1, \delta_2], \ldots, [\delta_N, 1]$$

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

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

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

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

$$\mathbf{SGap}_f(A = a) = \sum_i \left| q_i - \underbrace{\left(q_i^a + \frac{p_i - p_i^a}{p_{i+1}^a - p_i^a}(q_i^{a+1} - q_i^a) \right)}_{\text{Linear Interpolation}} \right|.$$

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

Finally, we set

$$\mathbf{SGap}_f = \frac{1}{|\mathcal{A}|} \sum_a \mathbf{SGap}_f(A = a) \tag{19}$$

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:*

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

By using the Bayes rule, we get

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

Hence, iff $\mathcal{P}(A = a|f(X)) = \frac{1}{|\mathcal{A}|}$, we have the equivalent form. Specifically, $\mathcal{P}(A = a|f(X))$ 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 $\mathcal{P}(A = a|f(X))$ is large, the subgroup index can be easily revealed via the algorithm output. If the algorithm can fully preserve the privacy, then $\mathcal{P}(A = a|f(X)) = \frac{1}{|\mathcal{A}|}$.

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 $\mathbb{E}[f(X)|A]]$. Then we calculate the expectation of the prediction over all instances. Finally, we get the $\mathbf{PGap}_f$ by calculating $\mathbb{E}[\mathbb{E}[f(X)] - \mathbb{E}[f(X)|A]]$.

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:

$$\mathbf{PGap}_f(A = a) = \left| \frac{p}{N} - \frac{p_a}{N_a} \right|.$$

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

$$\mathbf{PGap}_f = \frac{1}{\mathcal{A}} \sum_a \mathbf{PGap}_f(A = a) \tag{20}$$

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:

$$\textbf{OGap}_f(A = a) = |TPR - TPR_a|\,.$$

The equalized odds is finally calculated by

$$\textbf{OGap}_f = \frac{1}{\mathcal{A}} \sum_a \textbf{OGap}_f(A = a). \tag{21}$$

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 (β) 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., $A \not\perp Y$.

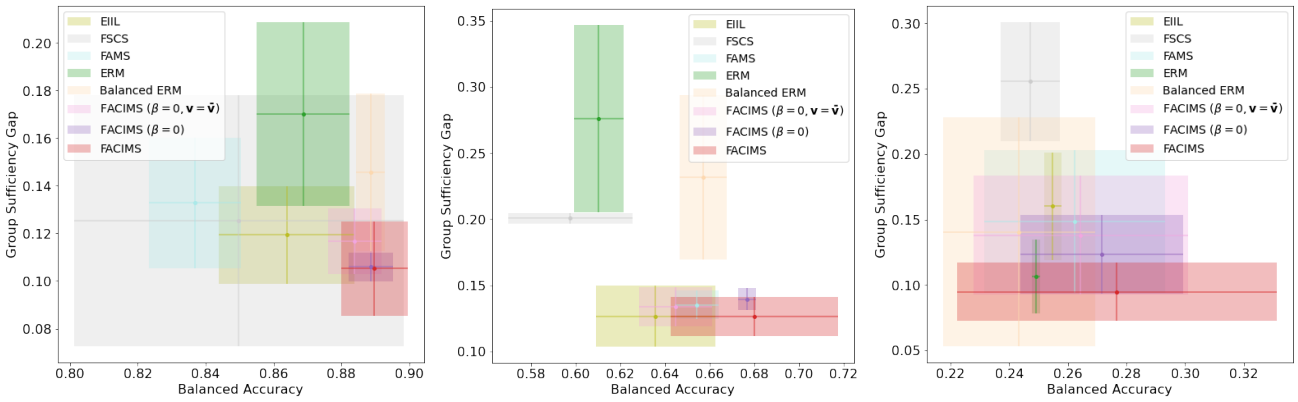
Table 5 provides the performance of FACIMS on the above synthetic imbalanced dataset. From this table, we can see that when there is a correlation between groups and labels ($\pi = 7$), there is a slight improvement in both accuracy and balanced accuracy. Further, the degeneration in demographic parity 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.

Method	Balanced Accuracy ↑	Demographic Parity ↓	Equalized Odds ↓	Sufficiency Gap ↓
EIIL	0.2549±0.0029	0.0199±0.0079	0.0673±0.0279	0.1602±0.0411
FSCS	0.2471±0.0101	0.0141±0.006	0.0399±0.0402	0.2555±0.0456
FAMS	0.2624±0.0308	0.0087±0.0021	0.2972±0.0323	0.1485±0.0546
ERM	0.2492±0.0013	0.0158±0.0052	**0.0061±0.0105**	0.1062±0.0283
BERM	0.2434±0.0259	0.0197±0.0016	0.1626±0.0259	0.1406±0.0875
FACIMS ($\beta = 0, v = \bar{v}$)	0.2644±0.0366	**0.0044±0.0016**	0.2547±0.0610	0.1380±0.0454
FACIMS ($\beta = 0$)	0.2716±0.0277	0.0192±0.0027	0.2158±0.0410	0.1231±0.0301
FACIMS	**0.2767±0.0545**	0.0257±0.0136	0.1469±0.1115	**0.0948±0.0224**

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.

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



(a) Alzheimer's Disease (b) Credit Card (c) Drug Consumption

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.

Table 5: The performance of FACIMS on the synthetic imbalanced dataset with different π.

π	**1**	**7**
Accuracy	0.9000±0.0083	0.9240±0.0026
Balanced Accuracy	0.8767±0.0196	0.8885±0.0121
Demographic Parity	0.0381±0.0131	0.2544±0.0117
Equalized Odds	0.0560±0.0131	0.0904±0.0080
Sufficiency Gap	0.1256±0.0444	0.1866±0.0107