On Learning Fairness and Accuracy on Multiple Subgroups

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose an analysis in fair learning that preserves the utility of the data while reducing prediction disparities under the criteria of group sufficiency. We focus on the scenario where the data contains multiple or even many subgroups, each with *limited number* of samples. As a result, we present a principled method for learning a fair predictor for all subgroups via formulating it as a bilevel objective. In the lower-level, the subgroup-specific predictors are learned through a small amount of data and the fair predictor. In the upper-level, the fair predictor is updated to be close to all subgroup specific predictors. We further prove that such a bilevel objective can effectively control the group sufficiency and generalization error. We evaluate the proposed framework on real-world datasets. Empirical evidence suggests the consistently improved fair predictions, as well as the comparable accuracy to the baselines.

1 Introduction

Machine learning has made rapid progress in sociotechnical systems such as automatic resume screening, video surveillance, and credit scoring for loan applications. Simultaneously, it has been observed that learning algorithms exhibited biased predictions on the *subgroups* of population [1, 2]. For example, the algorithm denies a loan application based on sensitive attributes [1] such as gender, race, or disability, which has heightened public concerns.

To this end, fair learning is recently highlighted to mitigate prediction disparities. The high-level idea is quite straightforward: adding fair constraints during the training [3]. As a result, fair learning principally gives rise to two desiderata. On the one hand, the fair predictor should be *informative* to ensure accurate predictions for the data. On the other hand, the predictor is required to guarantee fairness to avoid prediction disparities across subgroups. Therefore, it is crucial to understand the possibilities and then design provable approaches for achieving both *informative* and *fair* learning.

Clearly, achieving both objectives depends on predefined fair notations. Consider demographic parity [1] as the fair criteria, which necessitates the independence between the predictor's output $f(X)$ and the sensitive attribute A. Thus, if the sensitive attribute A and the ground-truth label Y are highly correlated, it is impossible to learn a both fair and informative predictor.

To avoid such intrinsic impossibilities, alternative fair notions have been developed. In this work, we focus on the criteria of *group sufficiency* [1, 4], which ensures that the conditional expectation of ground-truth label ($\mathbb{E}[Y|f(X), A]$) is identical across different subgroups, given the predictor's output. Notably, the risk of violating group sufficiency has arisen in a number of real-world scenarios. E.g., in medical artificial intelligence, the machine learning algorithm is used to assess the clinic risk,

[1] In this work, the *sensitive attribute* is referred to as the *subgroup index* such as different races.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

and guide decisions regarding initiating medical therapy. However, [5, 6] revealed a significant racial bias in such algorithms: when the algorithm predicts the same clinical risk score $f(X)$ for white and black patients, black patients are actually at a higher risk of severe illness: $\mathbb{E}[Y|f(X), A = \text{black}] \gg \mathbb{E}[Y|f(X), A = \text{white}]$. The deployed algorithms have resulted in more referrals of white patients to specialty healthcare services, resulting in both spending disparities and racial bias [5].

In summary, this work aims to propose a novel principled framework for ensuring group sufficiency, as well as preserving an informative prediction with a small generalization error. In particular, we focus on one challenge scenario: the *data includes multiple or even a large number of subgroups, some with only limited samples*, as often occurs in the real-world. For example, datasets for the self-driving car are collected from a wide range of geographical regions, each with a limited number of training samples [7]. How can we ensure group sufficiency, as well as accurate predictions? Specifically, our contributions are summarized as follows:

Controlling group sufficiency We introduce *group sufficiency gap* to measure fairness w.r.t. group sufficiency of a classifier f (Sec.2), and then derive an *upper bound* of the group sufficiency gap (Theorem 3.2). Under proper assumptions, the upper bound is controlled by the discrepancy between the classifier f and the subgroup Bayes predictors. Namely, minimizing the upper bound could also encourage an informative classifier.

Algorithmic contribution Motivated by the upper bound of the group sufficiency gap, we develop a principled algorithm. Concretely, we adopt a randomized algorithm that produces a predictive-distribution Q over the classifier ($f \sim Q$) to learn the informative and fair classification. We further formulate the problem as a bilevel optimization (Sec. 4.3), as shown in Fig.1. (1) In the lower-level, the subgroup specific dataset S_a and the fair predictive-distribution Q are used to learn the subgroup specific predictive-distribution $\overline{Q}_a^\star$, where Q is regarded as an *informative prior* for learning limited data within each subgroup. Theorem 4.1 formally demonstrates that under proper assumptions, the lower-level loss can effectively control the generalization error. (2) In the upper-level, the fair predictive-distribution Q is then updated to be close to all subgroup specific predictive-distributions, in order to minimize the upper bound of the group sufficiency gap.

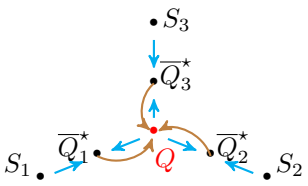


Figure 1: Illustration of the proposed algorithm. Consider three subgroups S_1, S_2, S_3, e.g., datasets for three different races. The proposed algorithm is then formulated as a bilevel optimization to learn an informative and fair predictive-distribution Q. In the lower-level (cyan), we learn the subgroup specific predictive-distribution $\overline{Q}_a^\star$ from dataset S_a (limited samples) and the prior Q. In the upper-level (brown), Q is then updated to be as close to all of the learned subgroup specific $\overline{Q}_a^\star$ as possible.

Empirical justifications The proposed algorithm is applicable to the general parametric and differentiable model, where we adopt the neural network in the implementation. We evaluate the proposed algorithm on two real-world NLP datasets that have shown prediction disparities w.r.t. group sufficiency. Compared with baselines, the results indicate that group sufficiency has been consistently improved, with almost no loss of accuracy.

2 Preliminaries

We assume the joint random variable (X, Y, A) follows an underlying distribution $\mathcal{D}(X, Y, A)$, where $X \in \mathcal{X}$ is the input, $Y \in \mathcal{Y}$ is the label, and the discrete random variable $A \in \mathcal{A}$ denotes the sensitive attribute or subgroup index. For instance, A represents gender, race, or age. We also denote $\mathbb{E}[Y|X]$ as the conditional expectation of Y, which is essentially a function of X. $\mathbb{E}_{A,X}[\cdot]$ is denoted as the expectation on the marginal distribution of $\mathcal{D}(A, X)$. Throughout the paper, we consider binary classification with $\mathcal{Y} = \{0, 1\}$. We further define the predictor as a scoring function $f : \mathcal{X} \to [0, 1]$ that maps the input into a real value in $[0, 1]$. It is worth mentioning that $f(X) \notin \mathcal{Y}$ since $f(X)$ is continuous. We then introduce group sufficiency and group sufficiency gap.

Definition 2.1 (Group sufficiency [4])**.** *A predictor f satisfies the group sufficiency with respect to the sensitive attribute A if* $\mathbb{E}[Y|f(X)] = \mathbb{E}[Y|f(X), A]$.

Intuitively, given a output score of the predictor $f(X) = \tau$, the conditional expectation of Y is invariant across different subgroups. Namely, conditioning on the specific subgroup $A = a$ does not

provide any additional information about the conditional expectation of Y. Then we could naturally define group sufficiency gap.

Definition 2.2 (Group sufficiency gap)**.** *The group sufficiency gap of a predictor f is defined as:* $\boldsymbol{Suf}_f = \mathbb{E}_{A,X}[|\mathbb{E}[Y|f(X)] - \mathbb{E}[Y|f(X), A]|]$

Specifically, $\mathbf{Suf}_f$ measures the extent of group sufficiency violation, induced by the predictor f, which is taken by the expectation over (X, A). For completeness, we also discuss other popular group fairness criteria: demographic parity and equalized odds.

Definition 2.3 (Demographic Parity (DP))**.** *A predictor f satisfies the demographic parity with respect to the sensitive attribute A if:* $\mathbb{E}[f(X)] = \mathbb{E}[f(X)|A]$

Demographic Parity (DP), also known as statistical parity or independence rule, emphasizes that the expectation of the output score $f(X)$ is independent of A. [1, 4] further revealed that if $A \not\perp\!\!\!\perp Y$, group sufficiency and demographic parity could not be simultaneously achieved.

Definition 2.4 (Equalized Odds (EO) [8])**.** *A predictor f satisfies the equalized odds with respect to A if:* $\mathbb{E}[f(X)|Y] = \mathbb{E}[f(X)|Y, A]$

Equalized odds (EO) emphasizes the conditional expectation of output f is invariant w.r.t. A, given the ground truth Y. [1, 9] reveal that if $\mathcal{D}(X, Y, A) > 0$ and $A \not\perp\!\!\!\perp Y$, group sufficiency and equalized odds can not both hold.

The analysis reveals a general *incompatibility* between group sufficiency and DP/EO when $A \not\perp\!\!\!\perp Y$, which often occurs in practice. Besides, DP/EO based criteria suffer the well-known fair accuracy trade-off [10]: enforcing the fair constraint degrades the prediction performance. This paper depicts that under the criteria of group sufficiency, these objectives could be both encouraged.

3 Upper bound of group sufficiency gap

To derive the theoretical results, we first introduce the group Bayes predictor.

Definition 3.1 (A-group Bayes predictor)**.** *The A-group Bayes predictor f^{Bayes} is defined as:* $f^{Bayes}(X, A) = \mathbb{E}[Y|X, A]$

The A-group Bayes predictor is associated with the underlying data distribution $\mathcal{D}(X, Y, A)$. Given the fixed realization $X = x, A = a$, we have $f^{\text{Bayes}}(x, a) = \mathbb{E}[Y|X = x, A = a]$. We could further demonstrate the following facts.

Proposition 3.1. *The A-group Bayes predictor f^{Bayes}: (1) satisfies the group sufficiency $\mathbb{E}[Y|f^{Bayes}(X, A)] = \mathbb{E}[Y|f^{Bayes}(X, A), A]$; (2) is the optimal predictor of subgroup $A = a$ under the binary cross-entropy loss function. Namely, $f^{Bayes}(x, a) = argmin_f \mathcal{L}_a^{BCE}(f)$, where $\mathcal{L}_a^{BCE}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}(x,y|A=a)} - [y \log(f(x)) + (1 - y) \log(1 - f(x))]$.*

The proof is delegated to the Appendix. Proposition 3.1 shows that the A-group Bayes predictor simultaneously meets both fairness (group sufficiency) and informative (optimal predictor under binary cross-entropy loss). Unfortunately, f^{Bayes} is impossible to estimate since it is related to the *underlying data distribution* $\mathcal{D}$, which is infeasible. Nevertheless, we can adopt A-group Bayes predictor f^{Bayes} to derive an upper bound of group sufficiency gap $\mathbf{Suf}_f$. The upper bound then holds for any predictor f that can be learned from the observed data.

Theorem 3.2. *The group sufficiency gap $\boldsymbol{Suf}_f$ is upper bounded by: $\boldsymbol{Suf}_f \leq 4\mathbb{E}_{A,X}[|f - f^{Bayes}|]$ Specifically, if A takes finite value ($|\mathcal{A}| < +\infty$) and follows uniform distribution with $\mathcal{D}(A = a) = 1/|\mathcal{A}|$. Then the group sufficiency gap is further simplified as:*

$$\boldsymbol{Suf}_f \leq \frac{4}{|\mathcal{A}|} \sum_a \mathbb{E}_X[|f - f^{Bayes}||A = a]$$

Theorem 3.2 reveals that the upper bound of the group sufficiency gap depends on the discrepancy between the predictor f and A-group Bayes predictor $f^{\text{Bayes}}(X, A)$. Namely, given different subgroups $A = a$, the optimal predictor f ought to be closed to all $f^{\text{Bayes}}(X, A = a), \forall a \in \mathcal{A}$.

Underlying assumption Theorem 3.2 also suggests underlying assumptions w.r.t. the data generation distribution $\mathcal{D}(X, Y, A)$ for achieving a small group sufficiency gap. If f^{Bayes} for each subgroup

$A = a$ are quite similar, then minimizing the upper bound yields a small group sufficiency gap $\mathbf{Suf}_f$. For example, consider the extreme scenario **if** the A-group Bayes predictors are identical w.r.t. $A : \mathbb{E}[Y|X, A = a] = \mathbb{E}[Y|X], \forall a \in \mathcal{A}$, where $\mathbb{E}[Y|X]$ is the conventional Bayes predictor defined on the marginalized distribution $\mathcal{D}(X, Y)$. The upper bound recovers the difference between the predictor f and standard Bayes predictor. If we use a probabilistic framework to approximate predictor $f(X) \approx \mathbb{E}[Y|X]$ (i.e, training the entire dataset without any fair constraint), both the group sufficiency gap and prediction error (since Bayes predictor is optimal) will be small, which is consistent with [11].
On the contrary, if A-group Bayes predictors are completely arbitrary with high variance for A, both group sufficiency gap and prediction error are large and impossible for the fair prediction.

4 Principled Approach

Based on the upper bound, we propose a principled approach to learn the predictor that achieves both small generalization error and group sufficiency gap.

4.1 Upper bound in randomized algorithm

To derive the theoretical result, we consider the randomized algorithm that learns a predictive-distribution Q over the scoring predictors from the data. For instance, if we consider Bayes framework, the predictor is drawn from the posterior distribution $\tilde{f} \sim Q$. In the inference, the predictor's output is formulated as the expectation of the learned predictive-distribution Q: $f(X) = \mathbb{E}_{\tilde{f} \sim Q} \tilde{f}(X)$.

In practice, it is infeasible to optimize over all the possible distributions. Then we should restrict the predictive-distribution Q within a distribution family $Q \in \mathcal{Q}$ such as Gaussian. We also denote $Q_a^\star \in \mathcal{Q}$ as the optimal prediction-distribution w.r.t. $A = a$ under binary cross-entropy loss within the distribution family $\mathcal{Q}$: $Q_a^\star := \text{argmin}_{Q_a \in \mathcal{Q}} \mathbb{E}_{\tilde{f}_a \sim Q_a} \mathcal{L}_a^{\text{BCE}}(\tilde{f}_a)$. In generally $Q_a^\star \neq f^{\text{Bayes}}(x, A = a)$, since the distribution family $\mathcal{Q}$ is only the subset of all possible distributions (shown in Fig. 2). We then extend the upper bound in the randomized algorithm.

Corollary 4.1. *The group sufficiency gap* $\boldsymbol{Suf}_f$ *in randomized algorithm w.r.t. learned predictive-distribution* Q *is upper bounded by:*

$$\boldsymbol{Suf}_f \leq \frac{2\sqrt{2}}{|\mathcal{A}|}[\sum_a \underbrace{\sqrt{KL(Q_a^\star \| Q)}}_{\textit{Optimization}} + \underbrace{\sqrt{KL(Q_a^\star \| \mathcal{D}(Y|X, A = a))}}_{\textit{Approximation}}]$$

Where KL is the Kullback–Leibler divergence. Corollary 4.1 further reveals that the upper bound is decomposed into two terms, showing in Fig.2.

Optimization term The optimization term is the average KL divergence between the learned distribution Q and optimal predictive-distribution $Q_a^\star$ for each subgroup $A = a$. Minimizing the optimization term implies that the learned distribution Q will be both fair and informative for the prediction, because it aims to minimize the upper bound of the group sufficiency gap $\mathbf{Suf}_f$ and be close to the optimal predictive-distribution w.r.t. each $A = a$.

Approximation term The approximation term is the average KL divergence between the optimal distribution $Q_a^\star$ and the underlying data generation distribution. Given the distribution family $\mathcal{Q}$, it is a unknown constant. Besides, if the distribution family $\mathcal{Q}$ has a rich expressive power such as deep neural-network, the approximation term will be small [12]. However, an extreme large distribution family $\mathcal{Q}$ could simultaneously yield a potential overfitting on finite samples. In this paper, the neural network is adopted and the approximation term is assumed to be a small constant. Thus, controlling $\mathbf{Suf}_f$ implies minimizing the optimization term.

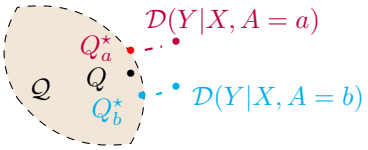


Figure 2: Illustration of optimization and approximation term. In the binary subgroup $\mathcal{A} = \{a, b\}$, the optimization term is to find $Q \in \mathcal{Q}$ to minimize the discrepancy between $(Q_a^\star, Q_b^\star)$. The approximation term is solely based on the distribution family $\mathcal{Q}$ (brown region). If the predefined $\mathcal{Q}$ has a rich expressive power, the approximation is treated as a small constant.

4.2 Challenge in learning limited samples

In practice, we only have access to finite or even limited samples, rather than the underlying distribution $\mathcal{D}$. We denote $S_a = \{(x_i^a, y_i^a)\}_{i=1}^m$ as the observed data w.r.t. subgroups $A = a$, which are i.i.d. samplings from the underlying distribution $\mathcal{D}(x, y|A = a)$. We also denote the **empirical** binary cross entropy loss w.r.t $A = a$ as: $\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}) = \frac{1}{m}\sum_{i=1}^m -[y_i^a \log(\tilde{f}(x_i^a)) + (1 - y_i^a)\log(1 - \tilde{f}(x_i^a))]$. Then a straight approach is to minimize the empirical term $\hat{Q}_a^\star$:

$$\hat{Q}_a^\star = \text{argmin}_{Q_a \in \mathcal{Q}} \, \mathbb{E}_{\tilde{f}_a \sim Q_a} \, \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a) \tag{1}$$

Then Q is updated through minimizing the average KL-divergence: $\sum_a \text{KL}(\hat{Q}_a^\star \| Q)$ from learned $\hat{Q}_a^\star$. However, this idea generally *does not work* in our setting, because each subgroup contains **limited** number of samples. Therefore, a direct minimization leads to overfitting for each subgroup and generalization error $\mathbb{E}_{\tilde{f} \sim \hat{Q}_a^\star} \, \mathcal{L}_a^{\text{BCE}}(\tilde{f})$ is quite large, showing in Fig. 5.

4.3 Q as an informative prior

We have demonstrated that Q can achieve both fair and informative prediction. Therefore, we regard Q as a *prior* information for minimizing the loss, yielding a bilevel objective.

$$\min_{Q \in \mathcal{Q}} \frac{1}{|\mathcal{A}|} \sum_a \text{KL}(\overline{Q}_a^\star \| Q) \tag{Upper-level}$$

$$\text{s.t. } \overline{Q}_a^\star = \text{argmin}_{Q_a \in \mathcal{Q}} \{\mathbb{E}_{\tilde{f}_a \sim Q_a} \, \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a) + \lambda \text{KL}(Q_a \| Q)\}, \forall a \in \mathcal{A} \tag{Lower-level}$$

Where $\lambda > 0$ is the hyper-parameter. The proposed loss is a typical bilevel optimization. (1) In the lower-level, we aim to learn $\overline{Q}_a^\star$ for each $a \in \mathcal{A}$. Different from Eq. (1), the loss in lower-level adds a regularization term $\text{KL}(Q_a \| Q)$ as an informative prior in learning $\overline{Q}_a^\star$, given a fixed predictive-distribution Q. Moreover, Theorem 4.1 formally justified that optimizing the lower-level loss is to minimize the upper bound of the generalization error. (2) In the upper-level, Q is updated through minimizing the average KL divergence between different $\overline{Q}_a^\star$, which controls the upper bound of $\textbf{Suf}_f$.

Theorem 4.1 (Generalization error bound)**.** *Supposing that datasets $\{S_a\}_{a=1}^{|\mathcal{A}|}$ with $S_a = \{(x_i^a, y_i^a)\}_{i=1}^m$ are i.i.d. sampled from $\mathcal{D}(x, y|A = a)$, the binary cross entropy (BCE) loss is upper bounded by L, $Q_a \in \mathcal{Q}$ is any learned distribution from dataset S_a and $Q \in \mathcal{Q}$ is any distribution. Then with high probability $\geq 1 - \delta$ with $\forall \delta \in (0, 1)$, we have:*

$$\frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f}_a \sim Q_a} \mathcal{L}_a^{BCE}(\tilde{f}_a) \leq \underbrace{\frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{\tilde{f}_a \sim Q_a} \hat{\mathcal{L}}_a^{BCE}(\tilde{f}_a)}_{(1)} + \underbrace{\frac{L}{\sqrt{|\mathcal{A}|m}} \sum_a \sqrt{KL(Q_a \| Q)}}_{(2)} + \underbrace{L\sqrt{\frac{\log(1/\delta)}{|\mathcal{A}|m}}}_{(3)}$$

Discussions Theorem 4.1 reveals the generalization error in the lower-level is upper bounded by three terms. (a) Term (1) is the average empirical prediction error, which corresponds to the first term in the lower-level loss. (b) Term (2) is the average KL-divergence between the learned subgroup distribution Q_a and the prior distribution Q, which corresponds to the second term in the lower-level loss. The combination of term (1-2) recovers the averaged lower-level loss w.r.t. A. [2] Thus optimizing the lower-level loss could control the generalization error. (c) When the confidence δ is fixed, term (3) will converge if $|\mathcal{A}|m \to +\infty$. Moreover, even if m (the sample size in each subgroup) is quite small, a sufficient large number of subgroups $|\mathcal{A}|$ can also ensure the convergence of term (3).

For the sake of simplicity, we assumed the identical samples size m in each subgroup S_a, while the theoretical result can be extended to subgroups with different samples m_a.

4.4 Practical Implementations

In this section, we develop a practical learning algorithm that can be applied to a wide range of differentiable and parametric models, including neural networks.

[2] In Theorem 4.1, the differences are in the square norm of KL divergence and setting the specific hyper-parameter: $\lambda = L\sqrt{|\mathcal{A}|/m}$.

Algorithm 1 Fair and Informative Learning for Multiple Subgroups

Input: Parameters w.r.t. distribution Q:$(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$, datasets $\{S_a\}$, $a \in \mathcal{A}$.
for Sampling a subset of $\{S_a\}$, where $a \in \mathcal{A}' \subseteq \mathcal{A}$ **do**
    ### Solving the lower-level ###
    Fix Q, optimizing the loss w.r.t. $Q_a = \mathcal{N}(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2)$ through SGD for each $a \in \mathcal{A}'$

$$\mathbb{E}_{\tilde{f}_{\mathbf{w}_a} \sim Q_a} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_{\mathbf{w}_a}) + \lambda \text{KL}(Q_a \| Q)$$

    Obtaining the solution $\overline{Q}_a^\star$, $a \in \mathcal{A}'$.
    ### Solving the upper-level ###
    Fix $\overline{Q}_a^\star$ with $a \in \mathcal{A}'$, optimizing the loss w.r.t. Q through SGD: $\frac{1}{|\mathcal{A}'|} \sum_a \text{KL}(\overline{Q}_a^\star \| Q)$
    Obtaining updated parameter $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ in Q
end for
Return: Parameter of distribution Q: $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$

Parametric models We choose the Isotropic Gaussian distribution (with diagonal covariance matrix) as the distribution family $\mathcal{Q}$, where the mean and covariance are set as d-dimensional parameter. Thus we need to learn the parameter $(\boldsymbol{\theta}, \boldsymbol{\sigma})$ for fair and informative $Q \in \mathcal{Q}$. As for the subgroup $A = a$, we learn parameters $(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a)$ for $\overline{Q}_a^\star \in \mathcal{Q}$. It is worth mentioning that the Isotropic Gaussian distribution is selected for its computational efficiency in the optimization. We can use any distribution as long as the density function is differentiable with respect to the parameters.

For the single predictor $\tilde{f}$, we use parametric neural-network models and assume $\tilde{f}$ is parameterized by a d-dimensional vector $\mathbf{w} \in \mathbb{R}^d$, denoted as $\tilde{f}_{\mathbf{w}}$. Then $\tilde{f}_{\mathbf{w}} \sim Q$ is equivalent to sampling the model parameter $\mathbf{w}$ from the predictive-distribution Q: $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^d \mathcal{N}(\boldsymbol{\theta}[i], \boldsymbol{\sigma}^2[i])$. Since Q is Isotropic Gaussian, each element i in the parameter $\mathbf{w}[i]$ follows a 1-dimensional Gaussian. Following the same line, $\tilde{f}_{\mathbf{w}_a} \sim \overline{Q}_a^\star$ can be modeled analogously: $\mathbf{w}_a \sim \mathcal{N}(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2) = \prod_{i=1}^d \mathcal{N}(\boldsymbol{\theta}_a[i], \boldsymbol{\sigma}_a^2[i])$. As a result, learning the distribution Q is equivalent to learning parameter $(\boldsymbol{\theta}, \boldsymbol{\sigma})$ in the bilevel objective.

Gradient Estimation Based on the previous setting, we aim to optimize the bilevel objective to obtain the parameter of Q: $(\boldsymbol{\theta}, \boldsymbol{\sigma})$. We use stochastic gradient descent (SGD) to optimize the parameters. In the lower-level, the loss in Sec. 4.3 is composed by the empirical prediction error and KL divergence term. The KL divergence has a closed form that can be differentiated efficiently. Specifically, since Q and the subgroup specific $\overline{Q}_a^\star$ are factorized Gaussian, the KL divergence takes a simple closed form and the gradient can be easily calculated: $\text{KL}(\overline{Q}_a^\star \| Q) = \frac{1}{2} \sum_{i=1}^d \left\{ \log \frac{\boldsymbol{\sigma}_a^2[i]}{\boldsymbol{\sigma}^2[i]} + \frac{\boldsymbol{\sigma}_a^2[i] + (\boldsymbol{\theta}_a[i] - \boldsymbol{\theta}[i])^2}{\boldsymbol{\sigma}^2[i]} - 1 \right\}$.

Re-parametrization trick As for the prediction error $\mathbb{E}_{\tilde{f}_{\mathbf{w}_a} \sim Q_a} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_{\mathbf{w}_a})$, the term $\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_{\mathbf{w}_a})$ is non-linear for $\mathbf{w}$, rendering the expectation intractable in the computation. To this end, we adopt the re-parameterization trick [13, 14] in computing the gradient w.r.t. the expectation term. The trick is based on describing the Gaussian distribution $\mathbf{w}_a \sim \mathcal{N}(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2)$ as first drawing $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and then applying the deterministic function $\mathbf{w}_a(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a) = \boldsymbol{\theta}_a + \boldsymbol{\sigma}_a \odot \boldsymbol{\epsilon}$ ($\odot$ is element-wise product) to approximate the sampling. Then the gradient term can be estimated as: $\nabla_{(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a)} \mathbb{E}_{\mathbf{w}_a \sim N(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a)} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_{\mathbf{w}_a}) = \nabla_{(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a)} \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_{\mathbf{w}_a(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a)})$, where the expectation can be approximated by Monte-Carlo sampling w.r.t. $\boldsymbol{\epsilon}$. For a fixed sample $\boldsymbol{\epsilon}$, the gradient can be computed through backpropagation. In the upper-level, the KL divergence has a closed form, thus it is easy to update the parameter of Q through backpropagation.

Proposed Algorithm Based on the analysis, the algorithm is shown in Algorithm. 1 for solving the bilevel objective in Sec. 4.3. Specifically, we adopt the alternating optimization. Namely, in the lower-level, we fix Q and optimize the subgroup specific predictor $\overline{Q}_a^\star$ through SGD. Then in the upper-level, we fix the learned $\overline{Q}_a^\star$ and update Q. Since we may face many subgroups, at each training epoch, we randomly sample a subset $\mathcal{A}'$ such that $|\mathcal{A}'| \ll |\mathcal{A}|$ for the memory saving.

Inference In the inference, we use the Monte-Carlo method to sample the weights of the neural network from distribution $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$, then averaging the output w.r.t. different sampled weights to approximate $f(x) = \mathbb{E}_{\tilde{f}_{\mathbf{w}} \sim Q} \tilde{f}_{\mathbf{w}}(x)$

5 Experiments

5.1 Experimental Setup

Dataset: Amazon review We adopt Amazon review dataset [15, 16], which aims to predict the sentiment (classification) from the review. The datasets consist of large-scale users. Each user has limited number of reviews, ranging from 75 to 400. The *user* is then treated as a subgroup, and it has been observed that standard training can lead to prediction disparities in several users.[17].The experiment is adapted from the protocol of [16]. Specifically, we convert the original review score (ranging from 1-5) to the binary label: the positive review (score ≥ 4) and negative review (score ≤ 3). We draw and then fix 200 users from the original dataset, which includes the training, validation, and test sets. In the implementation, we first adopt DistilBERT [18] to learn the embedding with dimension $\mathbb{R}^{768}$. Then we adopt $\tilde{f}_{\mathbf{w}}$ and $\tilde{f}_{\mathbf{w_a}}$ as the four-layer fully connected neural network, where $\mathbf{w} \sim Q$ and $\mathbf{w}_a \sim Q_a$. Additional experimental details are delegated to the Appendix.

Dataset: Toxic Comments We also use the toxic comment dataset [19] to predict the text comment being toxic or not, which has shown the significant performance degradation on specific sub-populations. Following [19], we select *race* as sensitive attribute, which includes Black, White, Asian and Latino & others (4 subgroups). We also follow the same setting as the original dataset [16], which has the separate training, validation, and test sets. Since toxic comments are marked by multiple annotators, we decide that the comment is toxic if it is marked by at least half of the annotators. In the implementation, we adopt the DistilBERT [18] as the embedding with output dimension $\mathbb{R}^{768}$. Then we also adopt $\tilde{f}_{\mathbf{w}}$ and $\tilde{f}_{\mathbf{w_a}}$ as the four-layer fully connected neural network, with the same network structure as Amazon. Additional details are delegated to the Appendix.

Baselines We compare with following baselines. (1) ERM: training a deep model without considering the sensitive attribute. (2) SNN. Since we adopt randomized algorithms in the paper, we additionally compare the stochastic neural network through the vanilla training from the whole dataset. Namely, we find a predictive-distribution Q to minimize $\frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{f \sim Q} \hat{\mathcal{L}}_a^{\text{BCE}}(f)$. (3) EIIL [20] proposed an IRM based approach to promote the group sufficiency. (4) FSCS [21] adopted the conditional mutual information constraint $I(A, Y | f(X))$ to promote the sufficiency. (5) DRO [22]. A re-weighting approach to assign the importance of the task. All the experiments are repeated five times.

Computing Suf$_f$ Since $f(X)$ is continuous, the group sufficiency gap is calculated by splitting the output of predictor into multiple intervals in $[0, 1]$ and computing the conditional expectation within each interval, as detailed in Appendix.

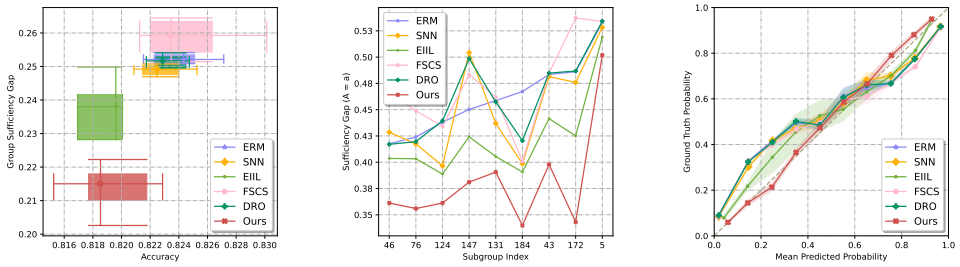
5.2 Experimental Results

We visualize the results in Fig. 3 for Amazon review product and Fig. 4 for toxic comment.

Accuracy and Fairness The accuracy and group sufficiency gap are depicted in Fig. 3(a) and Fig. 4(a). In Amazon review, the accuracy in the proposed approach has a slight decrease, compared with ERM. While the group sufficiency gap has improved by 3.0%, showing a significant improvement in the fairness. In toxic comments, the accuracy in proposed approach is nearly identical to the baseline, whereas group sufficiency gap has been significantly improved by 3.0-3.5%.

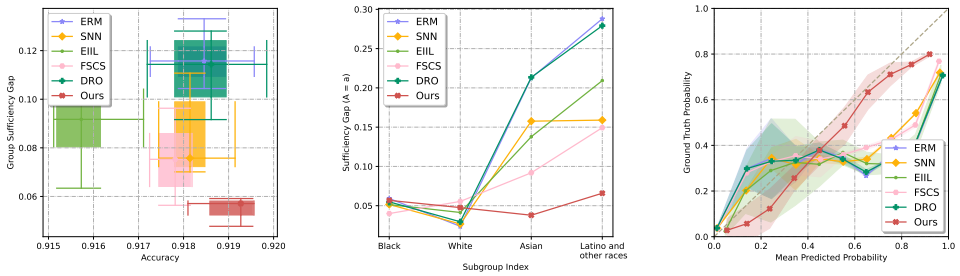
Group sufficiency gap on the specific subgroup To gain better understandings of group sufficiency gap, we visualize group sufficiency gap on specific subgroup $A = a$, i.e the discrepancy between the $\mathbb{E}[Y|f(X)]$ (conditional expectation on the entire data) and $\mathbb{E}[Y|f(X), A = a]$ (conditional expectation on subgroup $A = a$). I.e, $\mathbb{E}_X \left[|\mathbb{E}[Y|f(X)] - \mathbb{E}[Y|f(X), A = a]| \right]$.

In Amazon review dataset, we visualize the top-9 users' sufficiency gap in ERM, as shown in Fig. 3(b), where the gap of entire users is delegated to the Appendix. The proposed approach significantly reduces the group sufficiency gap of in most subgroups. The similar trend is also observed in Toxic dataset, as shown in Fig. 4(b), where the proposed approach has the nearly identical and small group



(a) Accuracy-$\mathbf{Suf}_f$ (b) $\mathbf{Suf}_f$ on each subgroup (c) Probability Calibration

Figure 3: Amazon Review dataset. (a) Boxplot of accuracy and group sufficiency gap $\mathbf{Suf}_f$ with 5 repeats: median, 75th percentile and minimum-maximum value. (b) Group sufficiency gap on subgroup $A = a$, which is the difference between $\mathbb{E}[Y|f(X)]$ and $\mathbb{E}[Y|f(X), A = a]$. We visualize the top-9 users' group sufficiency gap in ERM, whereas the result for all users is delegated to the Appendix. (c) Probability calibration curve over 5 repeats with mean and standard deviation. i.e $(f(X), \mathbb{E}[Y|f(X)])$. The proposed approach demonstrated a consistently improved probability calibration.



(a) Accuracy-$\mathbf{Suf}_f$ (b) $\mathbf{Suf}_f$ on each subgroup (c) Probability Calibration

Figure 4: Toxic dataset. (a) Boxplot of accuracy and group sufficiency gap $\mathbf{Suf}_f$ with 5 repeats: median, 75th percentile and minimum-maximum value. (b) Group sufficiency gap on the specific subgroup $A = a$, which is the difference between $\mathbb{E}[Y|f(X)]$ and $\mathbb{E}[Y|f(X), A = a]$. (c) Probability calibration over 5 repeats with mean and standard deviation. i.e $(f(X), \mathbb{E}[Y|f(X)])$, where the probability calibration for each subgroup is delegated to Appendix.

sufficiency gap for each race. In contrast, the baselines exhibit significant group sufficiency bias on the Asian and Latino & other races.

Probability Calibration A related concept to group sufficiency is the probability calibration [23], which is defined as $\mathbb{E}[Y|f(X)] = f(X)$ in the binary classification. We visualize the probability sufficiency of Amazon review in Fig. 3(c) and Toxic comments in Fig. 4(c). The results suggest that the proposed approach demonstrates a consistently better probability calibration on the whole data. We then visualize the probability calibration for each subgroup, as shown in the Appendix, where the results reveal the improved probability calibration for each subgroup.

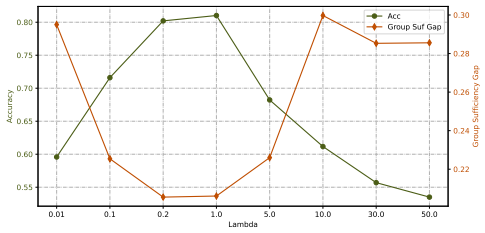


Figure 5: Analysis. Accuracy-$\mathbf{Suf}_f$ curve under different λ in Amazon review dataset.

Other sensitive attributes in Toxic comments
Apart from adopting race as sensitive attribute, we also consider other possible sensitive attributes

such as gender and religion, and the results are showed in the Appendix. The results in other sensitive attributes are similar to race, with improved fairness and no loss on accuracy.

Analysis: Influence of λ Theorem 4.1 suggests that there exists an optimal λ in the generalization error bound. Then we changed the value of λ in Amazon dataset, as shown in Fig. 5.

When $\lambda \to 0$, the subgroup specific parameters are simply learned from the limited samples within each subgroup. Then the fair predictor could not learn a proper prior from the subgroup specific predictor with a significant generalization error. Meanwhile, the group sufficiency gap is also large, which is consistent with [11]: overfitting generally degrades the group sufficiency. When λ is set between $[0.2, 1]$, the generalization error is small (with a high accuracy) and group sufficiency gap is kept small, implying that both fairness and accuracy can be achieved. In contrast, if we set a large value for $\lambda \gg 0$, the predictor is unable to learn from the data but from the random prior Q. The prediction will be completely random (accuracy $= 55\%$ when $\lambda = 50$). When the predictor outputs a random guess, different from demographic parity (DP) or equalized odds (EO), the group sufficiency gap is also large. The analysis reveals that there exists an optimal λ for simultaneously achieving accuracy and group sufficiency.

6 Related Work

Algorithmic fairness Fairness has been attached great importance and widely studied in various applications, such as natural language processing [24–26], natural language generation [27–29], computer vision [30, 31], and deep learning [32, 33]. Then various approaches have been proposed in algorithmic fairness. They typically add fair constraints during the training procedure, such as demographic parity or equalized odds [8, 34–38]. Apart from this, other fair notions are adopted such as accuracy parity [22, 39], which requires each subgroup to attain the same accuracy; small prediction variance [40, 41], which ensures small prediction variations among the subgroup; or small prediction loss for all the subgroups [42–45]. Furthermore, based on the concept of Independence (e.g. demographic parity $A \perp\!\!\!\perp f(X)$) or conditional independence (e.g. equalized odds $A \perp\!\!\!\perp f(X)|Y$ or group sufficiency $A \perp\!\!\!\perp Y|f(X)$), another popular line in fair learning is then naturally integrated with information theoretical framework through adding mutual information constraints such as [10, 46].

Understanding the fairness-accuracy trade-off As for the theoretical aspect, [47] further investigated the relation of fairness (demographic parity) and algorithmic stability. [48] formally justified the inherent trade-off between fairness (w.r.t. demographic parity and equalized odds) and accuracy, whereas the analysis is conducted for the binary sensitive attribute with the population loss. [49] studied the fair-accuracy trade-off in the multi-task learning.

Group sufficiency The notion of group sufficiency has recently been highlighted in various real-world scenarios such as health [6] and crime prediction [4, 9]. Specifically, [11] demonstrated that under proper assumptions, the group sufficiency can be controlled in the unconstraint learning. However, this conclusion may not necessarily always hold in the overparameterized models with limited samples, where [6, 16] essentially revealed the prediction disparities between the different subgroups in the unconstraint learning. [21] recently studied the fair selective classification w.r.t. group sufficiency through an information theoretical framework, while the theoretical guarantee is unknown. In contrast, the proposed lower-level loss within the paper can provably control the generalization error, and the upper-level loss controls the group sufficiency gap. Besides, a close notion to the group sufficiency is the *probability calibration* [23], which is defined as $\mathbb{E}[Y|f(X)] = f(X)$ in binary classification. We will empirically show the probability calibration could be consistently improved within our framework, whereas the analysis on finite samples and its theoretical relation with group sufficiency remains still opening [20].

7 Conclusion

We conducted a novel analysis by simultaneously learning an informative and fair classifier for multiple or even many subgroups. We derived a novel principled algorithm. We further theoretically justified the generalization error and fair guarantees of the proposed framework. The empirical results in two real-world datasets demonstrated the effectiveness in both preserving the accuracy, as well as group sufficiency.

References

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

[2] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

[3] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*, 2018.

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

[5] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 383 (9):874–882, 2020. doi: 10.1056/NEJMms2004740. URL https://doi.org/10.1056/NEJMms2004740.

[6] 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.

[7] Frank van Praat. So, here's the problem with self-driving cars. 2020.

[8] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

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

[10] 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.

[11] 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 2019. URL https://proceedings.mlr.press/v97/liu19f.html.

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

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

[14] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[15] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[16] 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.

[17] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[18] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[19] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

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

[21] 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.

[22] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[24] Shivashankar Subramanian, Afshin Rahimi, Timothy Baldwin, Trevor Cohn, and Lea Frermann. Fairness-aware class imbalanced learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2045–2051, 2021. doi: 10.18653/v1/2021.emnlp-main.155. URL `https://aclanthology.org/2021.emnlp-main.155`.

[25] Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *CoRR*, abs/2203.07228, 2022.

[26] Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online, 2020. Association for Computational Linguistics.

[27] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4275–4293. Association for Computational Linguistics, 2021.

[28] Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. *arXiv preprint arXiv:2203.12574*, 2022.

[29] Danielle Saunders, Rosie Sallis, and Bill Byrne. First the worst: Finding better gender translations during beam search. *CoRR*, abs/2104.07429, 2021.

[30] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12115–12124. Computer Vision Foundation / IEEE, 2021.

[31] Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14950–14959. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01470. URL `https://doi.org/10.1109/ICCV48922.2021.01470`.

[32] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14810–14820. IEEE, 2021.

[33] Xuan-Son Vu, Thanh-Son Nguyen, Duc-Trong Le, and Lili Jiang. Multimodal review generation with privacy and fairness awareness. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 414–425. International Committee on Computational Linguistics, 2020.

[34] 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.

[35] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[36] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

[37] Michael Feldman. *Computational fairness: Preventing machine-learned discrimination*. PhD thesis, 2015.

[38] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3995–4004, 2017.

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

[40] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

[41] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

[42] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[43] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Fairness with minimal harm: A pareto-optimal approach for healthcare. *arXiv preprint arXiv:1911.06935*, 2019.

[44] Ananth Balashankar, Alyssa Lees, Chris Welty, and Lakshminarayanan Subramanian. What is fair? exploring pareto-efficiency for fairness constrained classifiers. *arXiv preprint arXiv:1910.14120*, 2019.

[45] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

[46] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[47] Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR, 2019.

[48] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pages 2803–2813. PMLR, 2020.

[49] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. *arXiv preprint arXiv:2106.02705*, 2021.

[50] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3132674.

[51] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[52] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL `http://jmlr.org/papers/v18/17-468.html`.

[53] Eunbyung Park and Junier B Oliva. Meta-curvature. *Advances in Neural Information Processing Systems*, 32, 2019.

[54] Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8): 1889–1900, 2000.

[55] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[56] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[57] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021.

[58] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[59] Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR, 2014.

[60] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The proposed theoretical analysis is restricted in the randomized algorithm.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We study the algorithmic fairness. It is worth noting that ensuring group sufficiency may degrade other fair criteria such as DP/EO.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] In the paper and Appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provided the code for the reproduction.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the source code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We visualize the Boxplot of the results.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendices

A Rebuttal

A.1 Additional Discussions

Bi-level optimization Bi-level optimization seeks to solve problems with a hierarchical structure: that is, two levels of optimization problems in which one task is nested inside another [50]. Because many scenarios in machine learning can be decomposed into related sub-problems, bi-level optimization can be used. Several ideas related to bi-level optimization have been proposed in the context of fair-learning. We designed a min-max optimization to learn the fair representation when considering demographic parity (DP) or equalized odds (EO) [34, 10, 39]. In particular, the representation aims to minimize the loss caused by the discriminator in the lower-level optimization. Simultaneously, in the upper-level, we designed a discriminator to maximize the loss. Then fair representation could be enforced through the bi-level optimization. Besides, if we track the accuracy and its variants as the metrics for each subgroup [28], the bi-level objective could also be deployed in controlling the loss [51] or the prediction variance [41], where the lower-level's goal is to minimize the loss for each subgroup and the upper-level's goal is to estimate the prediction disparities. In our paper, we theoretically justified a novel bi-level optimization perspective: controlling group sufficiency and accuracy. Simultaneously, other bi-level optimization algorithms could be further considered in the fair learning such as recurrent based gradient updating [52], layer-wise transformation [53] or implicit gradient based approach [54].

Discussions on limitations We proposed the analysis on learning group sufficiency and informative predictors. and developed a principled approach for it. Simultaneously there are several limitations to the proposed theory and algorithm. (1) In general, group sufficiency and DP/EO are incompatible. Controlling group sufficiency, for example, would cause DP/EO degradation. This would be problematic if DP/EO were preferred in practice. (2) We also assumed that the ground truth A-Bayes predictors would be similar across groups. However, this assumption could be violated, resulting in a highly non-trivial scenario. Thus, in order to evaluate the conditional distribution shift, we need to consider a new setting by collect sufficient data per subgroup.

A.2 Results on different subgroup numbers

We visualize the results on different subgroup numbers, shown in Fig. 6. The results still suggest the consistently better results than baselines.

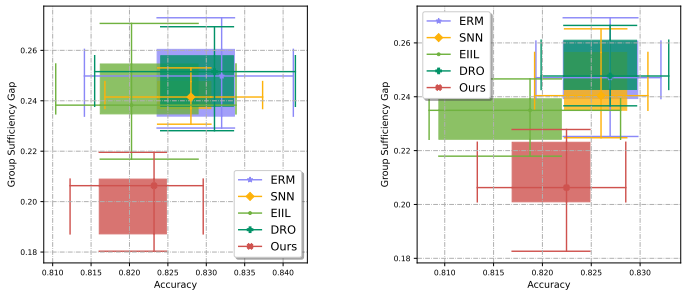
A.3 Additional Results on Adult dataset

We further evaluated the Adult dataset [55], which predicts salary being larger than 50K or not. We treat *gender* as the sensitive attribute and subsample 500 samples for each subgroup. We adopt $\tilde{f}_{\mathbf{w}}$ and $\tilde{f}_{\mathbf{w_a}}$ as the two-layer fully connected neural network, where $\mathbf{w} \sim Q$ and $\mathbf{w}_a \sim Q_a$. All results are repeated 4 times and illustrated in Tab. 1.

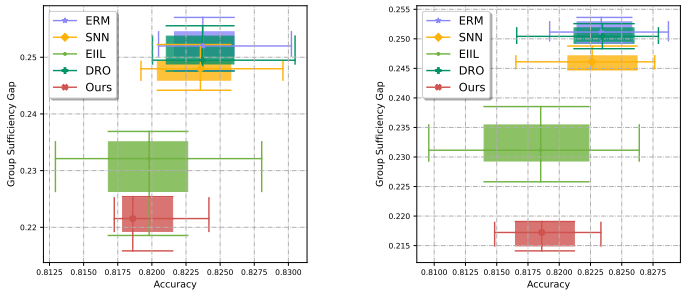
Table 1: Results (average and std) in Adult dataset (on %)

Method	Accuracy	Group sufficiency gap (smaller is better)
ERM	83.38 (0.421)	1.096 (0.182)
SNN	82.62 (0.478)	1.071 (0.220)
EIIL	83.27 (0.547)	1.018 (0.239)
FSCS	83.63 (0.704)	1.271 (0.316)
DRO	83.48 (0.377)	1.056 (0.151)
Ours	83.33 (0.308)	0.684 (0.008)

The result suggests a consistently better group sufficiency with comparable accuracy.



(a) Subgroup Number: 50 (b) Subgroup Number: 100



(c) Subgroup Number: 300 (d) Subgroup Number: 400

Figure 6: Results of different subgroups numbers on Amazon Review Dataset

A.4 Additional Results on vision dataset

We further consider CelebA dataset as a computer vision task [56]. We follow the protocol of [57], which predicts the wavy hair Y in the image X. We further regard gender as sensitive attribute A. We further adopted Res18 as the backbone and three layers fully-connected (randomized) layers. We sub-sample 200 instances per subgroup and fine tuning for maximum 20 epochs. All results are repeated 4 times and illustrated in Tab. 2.

Table 2: Results (average and std) in CelebA dataset (on %)

Method	Accuracy	Group sufficiency gap (smaller is better)
ERM	79.67 (0.40)	7.13 (0.96)
SNN	79.79 (0.33)	7.12 (0.89)
EIIL	79.90 (0.30)	6.23 (1.52)
FSCS	79.33 (0.62)	7.01 (1.88)
DRO	79.82 (0.27)	6.68 (1.14)
Ours	79.87 (0.17)	5.29 (0.67)

The result also suggests a consistently better group sufficiency with comparable accuracy.

A.5 Evolution of Q during the training

We visualize the test accuracy and group sufficiency gap of fair predictor Q during the training, shown in Tab. 3. The results in evaluated on the Toxic data with race as the sensitive attribute.

Table 3: Evolution of accuracy and group sufficiency gap during the training (on %)

Epoch	0	2	4	6	8	10	12	14
Suf gap	12.49	14.65	11.60	7.32	6.19	5.41	5.45	5.42
Accuracy	50.94	66.30	87.87	91.71	92.10	92.10	92.16	92.13

B Group sufficiency vs. demographic parity and equalized odds

For better understanding the properties of these three metrics, we consider the following scenario.

Consider the score predictor $f(X)$ uniformly outputs a value in $[0,1]$ for any $x \in \mathcal{X}$. i.e, $f(x) \sim \text{Unif}[0,1]$. Then it is easy to verify the demographic party and equalized odds both satisfy. Since the output of predictor is completely independent with the data. Thus we have

$$\mathbb{E}[f(X)] = \mathbb{E}[f(X)|A], \mathbb{E}[f(X)|Y] = \mathbb{E}[f(X)|A,Y]$$

In contrast, the group sufficiency does not necessarily hold. Specifically we have $\mathbb{E}[Y|f(X)] = \mathbb{E}[Y], \mathbb{E}[Y|f(X),A] = \mathbb{E}[Y|A]$, then we have

$$\textbf{Suf}_f = \mathbb{E}_A|\mathbb{E}[Y] - \mathbb{E}[Y|A]| > 0,$$

if $A \not\perp\!\!\!\perp Y$.

The aforementioned counterexample suggests that group sufficiency shows different behaviors, where the pure random prediction could trivially achieve the EO/DP.

Moreover, based on [1], if $\mathcal{D}(X,Y,A) > 0$ and $A \not\perp\!\!\!\perp Y$, the group sufficiency and demographic parity/equalized odds do not both hold. The aforementioned example also verifies this fact.

C Facts of A-Group Bayes predictor

C.1 Proof of Fact 1

We first introduce the generalized Tower rule of the conditional expectation.

Generalized tower rule of conditional expectation Let $(\Omega, \mathcal{F}, P)$ be the probability space and two sub σ-algebras $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$ are defined. Then we have

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \mathbb{E}[X|\mathcal{G}_1]$$

Proof. Based on the generalized tower rule, we have

$$\begin{aligned}\mathbb{E}[Y|f^{\text{Bayes}}(X,A),A] &= \mathbb{E}[\mathbb{E}[Y|X,A,f^{\text{Bayes}}(X,A)]|f^{\text{Bayes}}(X,A),A] \\ &= \mathbb{E}[\mathbb{E}[Y|X,A]|f^{\text{Bayes}}(X,A),A] = \mathbb{E}[f^{\text{Bayes}}|f^{\text{Bayes}}(X,A),A] = f^{\text{Bayes}}(X,A)\end{aligned}$$

Also we have:

$$\begin{aligned}\mathbb{E}[Y|f^{\text{Bayes}}(X,A)] &= \mathbb{E}[\mathbb{E}[Y|f^{\text{Bayes}}(X,A),A,X]|f^{\text{Bayes}}(X,A)] \\ &= \mathbb{E}[\mathbb{E}[Y|f^{\text{Bayes}}(X,A),A,X]|f^{\text{Bayes}}(X,A)] \\ &= \mathbb{E}[\mathbb{E}[Y|A,X]|f^{\text{Bayes}}(X,A)] = \mathbb{E}[f^{\text{Bayes}}(X,A)|f^{\text{Bayes}}(X,A)] = f^{\text{Bayes}}(X,A)\end{aligned}$$

Combining these two equations, we have the Fact 1:

$$\mathbb{E}[Y|f^{\text{Bayes}}(X,A),A] = \mathbb{E}[Y|f^{\text{Bayes}}(X,A)]$$

The aforementioned proof adopts the fact $\mathbb{E}[Y|f^{\text{Bayes}}(X,A),A,X] = \mathbb{E}[Y|A,X]$, since the conditional expectation is uninfluenced by the A-group Bayes predictor, given A, X. □

C.2 Proof of Fact 2

Following [58], we can compute the optimal predictor of attribute $A = a$ under the binary cross-entropy by taking the functional derivative w.r.t. f:

$$\frac{d\mathcal{L}_a^{\text{BCE}}(f)}{df} = 0$$

We have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}(x,y|a)}[f(x) - y] = 0 \\
\rightarrow &\mathbb{E}_{\mathcal{D}(x|a)}[f(x)\mathbb{E}_{\mathcal{D}(y|x,a)} - \mathbb{E}_{\mathcal{D}(y|x,a)}y] = 0 \\
\rightarrow &\mathbb{E}_{\mathcal{D}(x|a)}[f(x) - \mathbb{E}_{\mathcal{D}(y|x,a)}y] = 0
\end{aligned}$$

Therefore, we have the optimal predictor $f^\star = \mathbb{E}_{\mathcal{D}(y|x,a)}y = \mathbb{E}[Y = y|X = a, A = a]$, the A-group Bayes predictor. [58] further demonstrated the optimal is unique under the binary cross entropy loss.

D Upper bound of group sufficiency gap

Before deriving the theory, we need the following lemma.

Lemma D.1. *For any predictor f, we have*

$$\mathbb{E}[Y|f(X), A] = \mathbb{E}[f^{Bayes}(X, A)|f(X), A]$$

Proof.

$$\begin{aligned}
\mathbb{E}[Y|f(X), A] &= \mathbb{E}[\mathbb{E}[Y|f(X), A, X]|f(X), A] \\
&= \mathbb{E}[\mathbb{E}[Y|A, X]|f(X), A] \\
&= \mathbb{E}[f^{\text{Bayes}}(X, A)|f(X), A]
\end{aligned}$$

□

Based on Lemma B.1, we can derive the main Theorem.

Proof. For the simplicity, we denote $\mathbb{E}[Y|f, A] = \mathbb{E}[Y|f(X), A]$ and $\mathbb{E}[Y|f^{\text{Bayes}}, A] = \mathbb{E}[Y|f^{\text{Bayes}}(X, A), A]$. We first bound $\mathbb{E}_{A,X}[|\mathbb{E}[Y|f, A] - \mathbb{E}[Y|f^{\text{Bayes}}, A]|]$

$$\begin{aligned}
\mathbb{E}_{A,X}[|\mathbb{E}[Y|f, A] - \mathbb{E}[Y|f^{\text{Bayes}}, A]|] &= \mathbb{E}_{A,X}[|\mathbb{E}[f^{\text{Bayes}}|f, A] - f^{\text{Bayes}}|] \\
&= \mathbb{E}_{A,X}[|\mathbb{E}[f^{\text{Bayes}} - f|f, A] + f - f^{\text{Bayes}}|] \\
&\leq \mathbb{E}_{A,X}[|\mathbb{E}[f^{\text{Bayes}} - f|f, A]| + |f - f^{\text{Bayes}}|] \\
&= 2\mathbb{E}_{A,X}[|f - f^{\text{Bayes}}|]
\end{aligned}$$

Where $\mathbb{E}_{A,X}[|\mathbb{E}[f^{\text{Bayes}} - f|f, A]|] = \mathbb{E}_{A,X}[\mathbb{E}|f^{\text{Bayes}} - f|f, A, X|] = \mathbb{E}_{A,X}[|\mathbb{E}[f^{\text{Bayes}} - f|A, X]|] = \mathbb{E}_{A,X}[|f^{\text{Bayes}} - f|]$ is derived from the tower rule of conditional expectation.

Analogously, we can bound

$$\mathbb{E}_{A,X}[|\mathbb{E}[Y|f] - \mathbb{E}[Y|f^{\text{Bayes}}]|] \leq 2\mathbb{E}_{A,X}[|f - f^{\text{Bayes}}|]$$

Thus the group sufficiency gap is upper bounded by:

$$\begin{aligned}
\textbf{Suf}_f &= \mathbb{E}_{A,X}[|\mathbb{E}[Y|f] - \mathbb{E}[Y|f, A]|] \\
&= \mathbb{E}_{A,X}[|\mathbb{E}[Y|f] - \mathbb{E}[Y|f, A] - \mathbb{E}[Y|f^{\text{Bayes}}] - \mathbb{E}[Y|f^{\text{Bayes}}, A]|] \quad \text{(Using Fact 1)} \\
&\leq \mathbb{E}_{A,X}[|\mathbb{E}[Y|f] - \mathbb{E}[Y|f^{\text{Bayes}}]| + |\mathbb{E}[Y|f, A] - \mathbb{E}[Y|f^{\text{Bayes}}, A]|] \\
&\leq 4\mathbb{E}_{A,X}[|f - f^{\text{Bayes}}|]
\end{aligned}$$

Therefore, if A takes only finite value ($|\mathcal{A}| < +\infty$) and follows uniform distribution with $\mathcal{D}(A = a) = 1/|\mathcal{A}|$, then we have:

$$\textbf{Suf}_f \leq \frac{4}{|\mathcal{A}|}\sum_a \mathbb{E}_X[|f - f^{\text{Bayes}}||A = a]$$

□

E Upper bound of group sufficiency gap in randomized algorithm

According to the definition, we have:

$$\begin{aligned}
\mathbf{Suf}_f &\leq \frac{4}{|\mathcal{A}|}\sum_a \mathbb{E}_X[|\mathbb{E}_{\tilde{f}\sim Q}\tilde{f}(x) - \mathbb{E}_{\tilde{f}\sim Q_a^\star}\tilde{f}(x)| + |\mathbb{E}_{\tilde{f}\sim Q_a^\star}\tilde{f}(x) - \mathbb{E}_{\tilde{f}\sim \mathcal{D}(y|x,a)}\tilde{f}(x)|] \\
&\leq \frac{4}{|A|}\sum_a \mathrm{TV}(Q_a^\star \| Q) + \mathrm{TV}(Q_a^\star \| \mathcal{D}(y|x,a)) \quad \text{(Property of Total variation distance)} \\
&\leq \frac{2\sqrt{2}}{|A|}[\sum_a \underbrace{\sqrt{\mathrm{KL}(Q_a^\star \| Q)}}_{\text{Estimation Error}} + \underbrace{\sqrt{\mathrm{KL}(Q_a^\star \| \mathcal{D}(y|x,a))}}_{\text{Approximation Error}}] \quad \text{(Pinsker's inequality)}
\end{aligned}$$

The second line is derived from the property of Total variation distance. Note the scoring predictor ranges in $[0,1]$: $\tilde{f}(X=x) \in [0,1]$.

The third line is derived from Pinsker's inequality. i.e, $\mathrm{TV}(P\|Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)}$.

F Generalization upper bound

Step 1 We first demonstrate the following Lemma, which is based on [59, 60].

Lemma F.1. *Let f be a random variable taking value in A and let $X_1, \ldots, X_l$ be l independent variables with each X_k distributed to μ_k over the set A_k. For function $g_k : A \times A_k \to [a_k, b_k]$, $k = 1, \ldots, l$. Let $\zeta_k(f) = \mathbb{E}_{X_k\sim\mu_k} g_k(f, X_k)$ for any fixed value of f. Then for any fixed distribution π on A and any λ, $\delta > 0$, the following inequality holds with high probability $1-\delta$ over the sampling $X_1, \ldots, X_l$ for all distribution ρ over A.*

$$\mathbb{E}_{f\sim\rho}\sum_{k=1}^{l}\zeta_k(f) - \mathbb{E}_{f\sim\rho}\sum_{k=1}^{l} g_k(f, X_k) \leq \frac{1}{\lambda}\left(KL(\rho\|\pi) + \frac{\lambda^2}{8}\sum_{k=1}^{l}(b_k - a_k)^2 + \log\frac{1}{\delta}\right)$$

Proof. We consider a random variable $v(f) = \sum_{k=1}^{l}\zeta_k(f) - g_k(f, X_k)$. Then based on the Donsker-Varadhan (DV) formula such as [60], $\forall v(f)$ and $\forall\lambda > 0$, we have

$$\mathrm{KL}(\rho\|\pi) \geq \mathbb{E}_{f\sim\rho}\{\lambda v(f)\} - \log(\mathbb{E}_{f\sim\pi}(\exp\{\lambda v(f)\})),$$

Then we have

$$\lambda\mathbb{E}_{f\sim\rho}\{v(f)\} \leq \mathrm{KL}(\rho\|\pi) + \log(\mathbb{E}_{f\sim\pi}(\exp\{\lambda v(f)\}))$$

Note $v(f) = \sum_{k=1}^{l}\zeta_k(f) - g_k(f, X_k)$, then we have:

$$\begin{aligned}
\lambda\mathbb{E}_{f\sim\rho}\{v(f)\} &\leq \mathrm{KL}(\rho\|\pi) + \log(\mathbb{E}_{f\sim\pi}(\exp\{\lambda(\sum_{k=1}^{l}\zeta_k(f) - g_k(f, X_k)\}))) \\
&\leq \mathrm{KL}(\rho\|\pi) + \log(\mathbb{E}_{f\sim\pi}(\prod_{k=1}^{l}\exp\{\lambda(\zeta_k(f) - g_k(f, X_k)\})))
\end{aligned}$$

Note $\zeta_k(f) - g_k(f, X_k)$ is also bounded in $[a_k, b_k]$. Since $\zeta_k(f)$ is the expectation term of g_k w.r.t. X_k, then it is a sub-Gaussian random variable [60] such that:

$$\mathbb{E}_{X_k}[\exp\{\lambda\{\zeta_k(f) - g_k(f, X_k)\}\}] \leq \exp\{\frac{\lambda^2\sigma_k^2}{2}\},$$

where $\lambda > 0$ is a positive constant and $\sigma_k = \frac{b_k - a_k}{2} > 0$. Then we have:

$$\mathbb{E}_{f\sim\pi}\mathbb{E}_{X_1}\ldots\mathbb{E}_{X_l}(\prod_{k=1}^{l}\exp\{\lambda(\zeta_k(f) - g_k(f, X_k)\})) \leq \exp\{\lambda^2\sum_{k=1}^{l}\frac{\sigma_k^2}{2}\}$$

Since π is fixed but not depends on $X_1, \ldots, X_l$, thus we exchange the order of expectation. Then we apply Markov inequality w.r.t. the sampling of $X_1, \ldots, X_l$, with high probability $1-\delta$, we have:

$$\mathbb{E}_{f\sim\pi}(\prod_{k=1}^{l}\exp\{\lambda(\zeta_k(f)-g_k(f,X_k)\})) \leq \mathbb{E}_{f\sim\pi}\mathbb{E}_{X_1}\ldots\mathbb{E}_{X_l}(\prod_{k=1}^{l}\exp\{\lambda(\zeta_k(f)-g_k(f,X_k)\}))/\delta$$

Apply the of sub-Gaussian property, we have:

$$\log(\mathbb{E}_{f\sim\pi}(\prod_{k=1}^{l}\exp\{\lambda(\zeta_k(f)-g_k(f,X_k)\}))) \leq \log(\frac{1}{\delta}) + \lambda^2\sum_{k=1}^{l}\frac{\sigma_k^2}{2}$$

Plug-in all the elements with $\delta_k = \frac{b_k - a_k}{2}$, we have the results. □

Step 2 Then we could use the aforementioned Lemma to demonstrate the main theorem.

Proof. We adopt the lemma for the union of the whole training samples $S = \cup_{a\in\mathcal{A}} S_a$.

We set

$$\rho = \underbrace{(Q_1 \otimes Q_2 \otimes \cdots \otimes Q_{|\mathcal{A}|})}_{|\mathcal{A}| \text{ times}} \qquad \pi = \underbrace{(Q \otimes Q \otimes \cdots \otimes Q)}_{|\mathcal{A}| \text{ times}}$$

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:

$$\begin{aligned}\frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}_a\sim Q_a}\mathcal{L}_a^{\text{BCE}}(\tilde{f}_a) \leq \frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}_a\sim Q_a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a) \\ + \frac{1}{\lambda}(\text{KL}(Q_1\otimes\cdots\otimes Q_{|\mathcal{A}|}\|Q\otimes\cdots\otimes Q) + \log(\frac{1}{\delta})) + \frac{\lambda L}{8|\mathcal{A}|m}\end{aligned}$$

By setting $\lambda = \sqrt{8|\mathcal{A}|m(\text{KL}(Q_1\otimes\cdots\otimes Q_{|\mathcal{A}|}\|Q\otimes\cdots\otimes Q) + \log(1/\delta))/L}$, we have:

$$\begin{aligned}\frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}_a\sim Q_a}\mathcal{L}_a^{\text{BCE}}(\tilde{f}_a) \leq \frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}_a\sim Q_a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a) \\ + L\sqrt{\frac{1}{2|\mathcal{A}|m}(\text{KL}(Q_1\otimes\cdots\otimes Q_{|\mathcal{A}|}\|Q\otimes\cdots\otimes Q) + \log(\frac{1}{\delta}))}\end{aligned}$$

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

$$\begin{aligned}\frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}\sim Q_a}\mathcal{L}_a^{\text{BCE}}(\tilde{f}) &\leq \frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}\sim Q_a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}) + L\sqrt{\frac{1}{2|\mathcal{A}|m}(\sum_a \text{KL}(Q_a\|Q) + \log(\frac{1}{\delta}))} \\ &\leq \frac{1}{|\mathcal{A}|}\sum_a \mathbb{E}_{\tilde{f}\sim Q_a}\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}) + \frac{L}{\sqrt{|\mathcal{A}|m}}\sum_a\sqrt{\text{KL}(Q_a\|Q)} + L\sqrt{\frac{\log(1/\delta)}{|\mathcal{A}|m}}\end{aligned}$$

□

G Computing group sufficiency gap from the data

In this paper, we need to compute the conditional expectation from the data. i.e,

$$\mathbb{E}[Y|f(X)], \quad \mathbb{E}[Y|f(X), A=a],$$

where we have observed data $\{S_a\}, a \in \mathcal{A}$. Since $f(x)$ is a continuous value, ranging from $[0,1]$. Then we split $[0,1]$ into sperate interevals:

$$[0,\epsilon_1], [\epsilon_1,\epsilon_2], \ldots, [\epsilon_N, 1]$$

We compute the expectation and conditional expectation within each interval. i.e:

$$(\mathbb{E}[f(X)\mathbf{1}_{\{f(X)\in[\epsilon_i,\epsilon_{i+1}]\}}], \mathbb{E}[Y|f(X)\in[\epsilon_i,\epsilon_{i+1}]]) = (p_i, q_i)$$

$$(\mathbb{E}[f(X)\mathbf{1}_{\{f(X)\in[\epsilon_i,\epsilon_{i+1}],A=a\}}], \mathbb{E}[Y|f(X)\in[\epsilon_i,\epsilon_{i+1}], A=a]) = (p_i^a, q_i^a)$$

Then for each group $A = a$, the group sufficiency gap is computed as:

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

We use the linear interpolation if the average values in each interval are not equal. Then the group sufficiency can be formulated as:

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

We assign the $\mathcal{D}(A=a) = \frac{1}{|\mathcal{A}|}$ as uniform distribution for ensuring fairness for each subgroup.

Discussions In general:

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

The demonstration is straightforward. By using the Bayes rule, we have

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

Thus iff $\mathcal{D}(A=a|f(X)) = \frac{1}{|\mathcal{A}|}$, we have the equivalent form. Intuitively, $\mathcal{D}(A=a|f(X))$ refers the conditional probability of $A=a$, given the predicted score $f(X)$, which is related to the group membership inference such as [61]. If $\mathcal{D}(A=a|f(X))$ is large, the subgroup index can be easily revealed via the algorithm output. If the algorithm can fully preserve the privacy, then $\mathcal{D}(A=a|f(X)) = \frac{1}{|\mathcal{A}|}$.

H Experimental Details

In this part, we proposed a detailed description of the dataset and experiments settings.

H.1 Amazon review dataset

The experiment is adapted from the protocol of [16]. Specifically, we convert the original review score (ranging from 1-5) to the binary label: the positive review (score ≥ 4) and negative review (score ≤ 3). We sample and then fix 200 users from the original dataset, which contains the training (75-400 samples per user) , validation (75 samples per user), and test sets (75 samples per user).

The total training epoch is 100. In each training epoch, we sample a small subset of users ($N_{\text{user}} = 20$), then for each user we sample 50 samples with replacement. The early stopping strategy is also adopted.

We adopt 4-layers fully connected neural network as the model, where the weights of the model follows the Gaussian distribution Q_a or Q. The trade-off coefficient λ ranges from $[0.01, 50]$ and we fix $\lambda = 0.4$ in the evaluation. We set all the Monte-Carlo samples as 5. More implementation details of experiments and parameter settings can be found in the code.

H.2 Toxic comments

We adopt the toxic comment dataset [19] to predict whether the text comment is toxic or not, which has been observed the significant performance degradation on particular sub-populations.

Following [19, 16], we first choose the *race* as sensitive attribute, which includes Black, White, Asian and Latino & others (4 subgroups).

The total training epoch is 100. In each training epoch, we sample all the subgroups with the same sample size with replacement ($N = 50$). The early stopping strategy is also adopted. The training ($N = 33188$) validation ($N = 3438$) and test set ($N = 9744$) are following the protocol in [16].

We also consider the following sensitive attributes:

1. Religion. The religion includes Christian, Jewish, Muslin and others (such as Hindu, Buddhist, atheist).
2. Gender. The gender includes male, female and others (such as homosexual_gay_or_lesbian, bisexual, transgender, other_gender).

Because toxic comments are marked by multiple annotators, we determine that the comment is toxic if at least half of the annotators mark it. In the implementation, we also adopt the DistilBERT [18] to extract the embedding with dimension $\mathbb{R}^{768}$.

We adopt 4-layers fully connected neural network as the model, where the weights of the model follows the Gaussian distribution Q_a or Q. The trade-off coefficient λ ranges from $[0.01, 50]$ and we fix $\lambda = 0.6$ in the evaluation. We set all the Monte-Carlo samples as five $N = 5$. More implementation details of experiments and parameter settings can be found in the code.

I Additional Results in Amazon review

Additional results of each subgroup's fair performance and the probability calibration on the Amazon Review dataset are shown in Fig. 7 and Fig. 8.

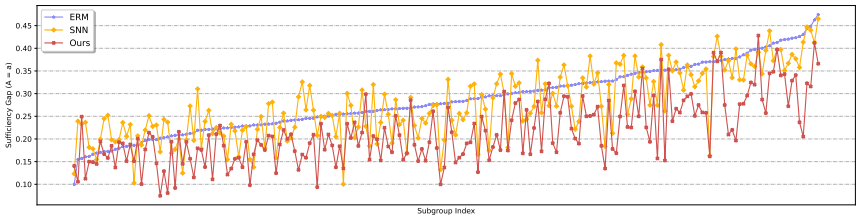


Figure 7: Group sufficiency gap on the specific subgroup $A = a$, which is the gap between $\mathbb{E}[Y|f(X)]$ and $\mathbb{E}[Y|f(X), A = a]$. We visualize the results for the entire users of Amazon Review dataset.

J Additional Results in Toxic Comments

J.1 Race as sensitive attribute: probability calibration

The additional results of the probability calibration on the Toxic Comments (Race) dataset is shown in Fig. 9.

J.2 Religion as sensitive attribute

The additional results of each subgroup's fair performance and the probability calibration on the Toxic Comments (Religion) dataset are shown in Fig. 10 and Fig. 11.

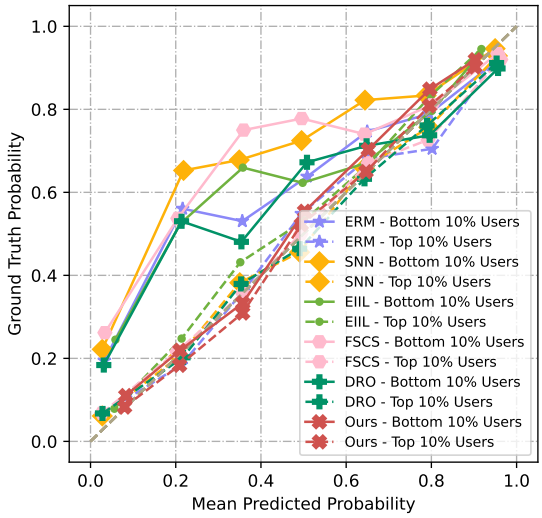
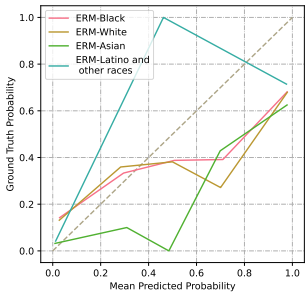


Figure 8: Probability calibration over aggregated users of Amazon Review dataset, where we sort the group sufficiency gap for each subgroup $A = a$, then we visualize the probability calibration curve for the top 10% and bottom 10% users. Since each user has quite limited labels, we aggregate the top 10% and bottom 10% and visualize the probability calibration. The results depict consistently better probability calibration of the proposed approach.

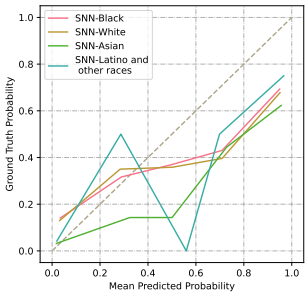
J.3 Gender as sensitive attribute

The additional results of each subgroup's fair performance and the probability calibration on the Toxic Comments (Gender) dataset are shown in Fig. 12 and Fig. 13.

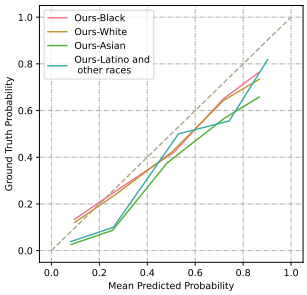
It is worth noting that although the group sufficiency in three approaches is quite similar. However, the proposed approach shows a significant better probability calibration than baselines.



(a) ERM

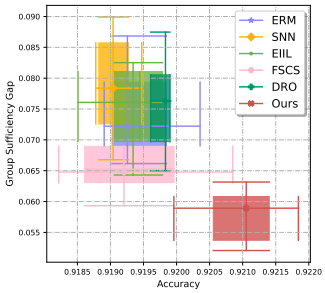


(b) SNN

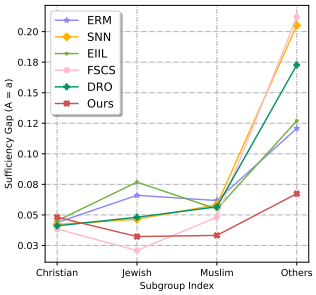


(c) Ours

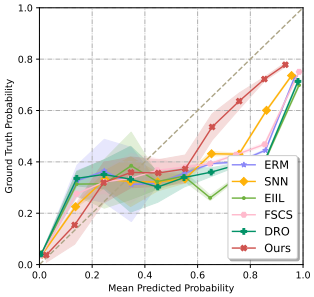
Figure 9: Ablation. Results in Toxic dataset (Race). Visualization of probability calibration for each subgroup $A = a$, i.e. $(f(X), \mathbb{E}[Y|f(X), A = a])$. The proposed approach shows significant improved probability calibration for each subgroup.



(a) Accuracy-$\mathbf{Suf}_f$

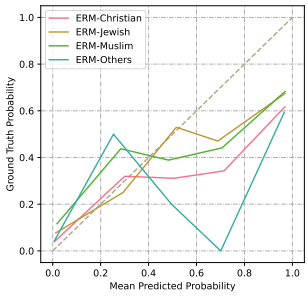


(b) $\mathbf{Suf}_f$ on each subgroup

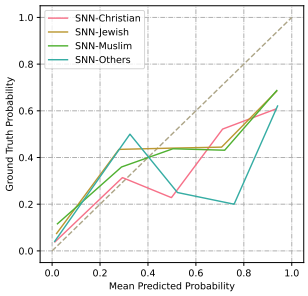


(c) Probability calibration

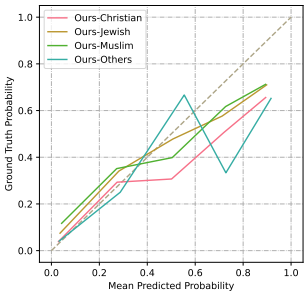
Figure 10: Results in Toxic dataset (Religion). Boxplot of accuracy and group sufficiency gap $\mathbf{Suf}_f$ with 5 repeats: median, 75th percentile and minimum-maximum value. (b) Group sufficiency gap on the specific subgroup $A = a$, which is the gap between $\mathbb{E}[Y|f(X)]$ and $\mathbb{E}[Y|f(X), A = a]$. (c) Visualization of probability calibration over 5 repeats with mean and standard deviation. i.e $(f(X), \mathbb{E}[Y|f(X)])$.



(a) ERM

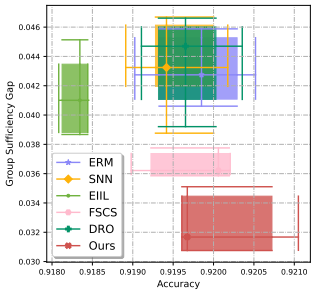


(b) SNN

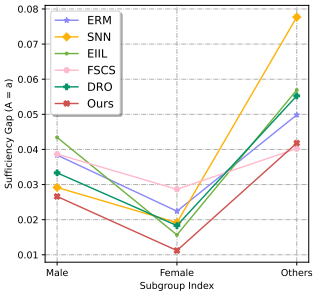


(c) Ours

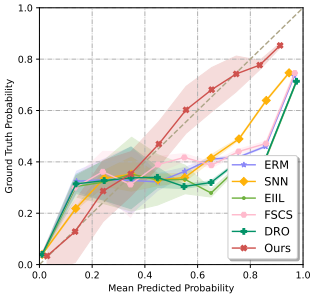
Figure 11: Ablation. Results in Toxic dataset (Religion). Visualization of probability calibration for each subgroup $A = a$, i.e. $(f(X), \mathbb{E}[Y|f(X), A = a])$. The proposed approach shows significant improved probability calibration for each subgroup.



(a) Accuracy-$\mathbf{Suf}_f$

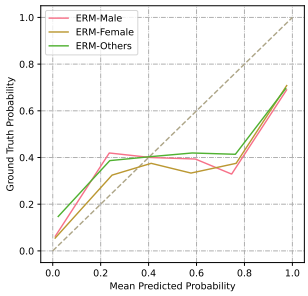


(b) $\mathbf{Suf}_f$ on each subgroup

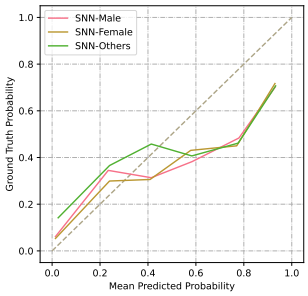


(c) Probability calibration

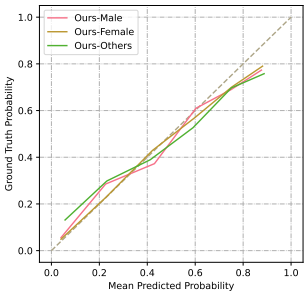
Figure 12: Results in Toxic dataset (Gender). Boxplot of accuracy and group sufficiency gap $\mathbf{Suf}_f$ with 5 repeats: median, 75th percentile and minimum-maximum value. (b) Group sufficiency gap on the specific subgroup $A = a$, which is the gap between $\mathbb{E}[Y|f(X)]$ and $\mathbb{E}[Y|f(X), A = a]$. (c) Visualization of probability calibration over 5 repeats with mean and standard deviation. i.e $(f(X), \mathbb{E}[Y|f(X)])$.



(a) ERM



(b) SNN



(c) Ours

Figure 13: Ablation. Results in Toxic dataset (Gender). Visualization of probability calibration for each subgroup $A = a$, i.e. $(f(X), \mathbb{E}[Y|f(X), A = a])$. The proposed approach shows significant improved probability calibration for each subgroup.