Enhancing Fairness through Reweighting: A Path to Attain the Sufficiency Rule

Xuan Zhao[a,*], **Klaus Broelemann**[a], **Salvatore Ruggieri**[b] **and Gjergji Kasneci**[c]

[a]SCHUFA Holding AG
[b]University of Pisa
[c]Technical University of Munich

Abstract. We introduce an innovative approach to enhancing the empirical risk minimization (ERM) process in model training through a refined reweighting scheme of the training data to enhance fairness. This scheme aims to uphold the sufficiency rule in fairness by ensuring that optimal predictors maintain consistency across diverse sub-groups. We employ a bilevel formulation to address this challenge, wherein we explore sample reweighting strategies. Unlike conventional methods that hinge on model size, our formulation bases generalization complexity on the space of sample weights. We discretize the weights to improve training speed. Empirical validation of our method showcases its effectiveness and robustness, revealing a consistent improvement in the balance between prediction performance and fairness metrics across various experiments. Code is available at https://github.com/zhaoxuan00707/Reweighting_for_sufficiency.

1 Introduction

Machine learning has found extensive application in real-world decision-making processes, including areas such as health care systems [1]. Algorithmic fairness has garnered significant attention as a means to mitigate predictive bias linked to protected features such as ethnicity, gender, or age. Consequently, numerous fairness notions catering to diverse objectives have been proposed. While many existing approaches in classification or regression adhere to independence or separation rules (refer to Section 2 and related references) [29, 39, 13], it's worth noting that these rules may not always be suitable in various applications. In such cases, alternative fairness notions, such as the sufficiency rule [12], are favored. In simple terms, the sufficiency rule, detailed in Section 2, ensures that the conditional expectation of $\mathbb{E}[Y|\hat{Y}]$ remains consistent across different subgroups, providing a more nuanced approach to fairness.

In practical terms, neglecting the sufficiency rule can result in significant biases within intelligent healthcare systems. For instance, many health systems utilize algorithms to identify and support patients with complex health requirements. These algorithms generate a score indicating the level of healthcare needs, with higher scores suggesting greater sickness and the need for more care. Notably, a study by [32] uncovers a widely used industry algorithm affecting millions of patients, which exhibits pronounced racial bias. It was found that for a given predicted score $\hat{Y} = s$, black patients tend to be considerably sicker than white patients ($\mathbb{E}_{black}[Y|\hat{Y} = s] > \mathbb{E}_{white}[Y|\hat{Y} = s]$). Moreover, the study highlights that rectifying this disparity could significantly increase the percentage of black patients receiving additional care from 17.7% to 46.5%. From an algorithmic perspective, the sufficiency rule is generally incompatible with concepts such as independence or separation, as demonstrated in Section 2 and the Appendix. This suggests that existing fair algorithms designed for independence or separation may not enhance or could even exacerbate issues related to the sufficiency rule. Notably, recent work on Invariant Risk Minimization (IRM) proposed by [2, 4] has potential to address this challenge. IRM seeks to maintain invariant correlations between the embedding (or representation) and the true label by incorporating regularization techniques into Deep Neural Network (DNN) training. The criteria of the sufficiency rule and IRM are intrinsically consistent (see more details in Section 3.1.1); the idea being that if the correlations between the embedding (or representation) and the true label remain robust and unaffected by specific sub-groups, the resulting representation can be considered fair. To better understand this concept, IRM addresses the challenge of ensuring that *a cow is correctly classified as a cow in a picture, regardless of whether the background is Grass or Desert* [2]. On the other hand, the sufficiency rule aims to ensure that *a patient predicted to be high-risk is truly high-risk, regardless of whether the patient is Black or White*. IRM approaches have attracted attention due to their promising performance on modest models and datasets [2] and their simplicity in facilitating end-to-end training. Nevertheless, recent studies have indicated diminished effectiveness of the regularization terms when applied to overparameterized DNNs [9, 25]. For instance, CelebA comprises only 200k training data, whereas ResNet-18 boasts 11.4 million parameters. Overparameterized DNNs can easily diminish the regularization term of IRM to zero during training while still depending on spurious features. In such scenarios, applying IRM methods directly for fairness to uphold the sufficiency rule in relatively larger models is deemed inappropriate.

This paper introduces a novel approach to address the aforementioned limitation by proposing a model-agnostic sample reweighting method. Our method transforms the parameter search space of the model into one of sample weights by formalizing the learning of sample reweighting as a bilevel optimization problem. Within the **inner loop**, we train DNN on the weighted training samples. In the **outer loop**, we employ the IRM criterion as the outer objective to guide the learning process of the sample weights, thereby enforcing the sufficiency rule. We iteratively alternate between the inner and outer loops, ultimately obtaining a set of weights w with an advanta-

[*] Corresponding Author. Email: zhaoxuan00707@gmail.com

geous characteristic: utilizing only learned sample weights on training samples, we can conduct weighted empirical risk minimization (ERM) training to achieve superior fairness.

Our contributions are summarized as follows:

1. We introduce a model-agnostic sample reweighting approach rooted in bilevel optimization for IRM learning to promote fairness. This method offers notable advantages, particularly in transforming the optimization problem from the parameter space of DNNs to the space of sample weights. This shift effectively mitigates the overfitting issues commonly encountered by IRM regularization-based methods.
2. Our method formulate the fairness issue as a bilevel optimization and does not impose specific fairness constraints, thus avoiding the issue of determining critical hyperparameters for fairness regularization.
3. We substantiate the superior performance of our approach through empirical evaluations across diverse tasks, showcasing its effectiveness compared to state-of-the-art methods.

The structure of this paper is as follows. In Section 2, we present a comprehensive review of the notation and background related to fairness notions, IRM, and reweighting methods. Section 3 outlines our sample reweighting method in detail. In Section 4, we conduct experiments to compare the accuracy and fairness across four datasets against state-of-the-art methods, illustrating the robustness and effectiveness of our framework.

2 Preliminaries and Related Work

2.1 *Sufficiency Rule in Fairness*

We denote the predictive features as $X \in \mathcal{X}$, the ground truth label as $Y \in \mathcal{Y}$, and the algorithm's output as $\hat{Y} \in \mathcal{Y}$. We consider a binary protected feature or two sub-groups $\mathcal{D}_0$ and $\mathcal{D}_1$. Then, in accordance with [26], the sufficiency rule is defined as follows:

$$\mathbb{E}_{\mathcal{D}_0}[Y|\hat{Y}=s] = \mathbb{E}_{\mathcal{D}_1}[Y|\hat{Y}=s], \forall s \in \mathcal{Y} \tag{1}$$

2.1.1 *Sufficiency Gap*

Eq. (1) indicates that the conditional expectation of the ground truth label Y is consistent across both $\mathcal{D}_0$ and $\mathcal{D}_1$, given the same prediction output s. In [37], the sufficiency gap is proposed as a metric for fairness measurement. In binary classification, the sufficiency gap is naturally defined as follows:

$$\Delta\text{Suf} = \frac{1}{2} \sum_{y\in\{0,1\}} |P_{\mathcal{D}_0}(Y=y|\hat{Y}=y) - P_{\mathcal{D}_1}(Y=y|\hat{Y}=y)| \tag{2}$$

The sufficiency gap $\Delta\text{Suf} \in [0,1]$. A value close to 0 indicates equality between two sub-groups, which have close Positive Predictive Values (PPV) and Negative Predictive Values (NPV). To grasp the significance of this metric, consider a healthcare system that only outputs binary scores: High Risk or Low Risk. As highlighted in [32], if $P_{\mathcal{D}_{\text{black}}}(Y = \text{High Risk}|\hat{Y} = \text{Low Risk}) \gg P_{\mathcal{D}_{\text{white}}}(Y = \text{High Risk}|\hat{Y} = \text{Low Risk})$, then the severity of illness is underestimated more for black patients than for white patients. Therefore, a small value of ΔSuf indicates that racial discrimination is addressed.

2.1.2 *Relation to Other Fairness Notions*

We briefly contrast the Sufficiency rule with the commonly employed Independence and Separation rules in binary classification. For comprehensive justifications and comparisons, please consult Appendix.

The *Independence rule* is:

$$\mathbb{E}_{\mathcal{D}_0}[\hat{Y}] = \mathbb{E}_{\mathcal{D}_1}[\hat{Y}] \tag{3}$$

In binary classification, the Independence rule is often referred to as demographic parity (DP) [42]. Furthermore, it can be argued that if $P_{\mathcal{D}_0}(Y=y) \neq P_{\mathcal{D}_1}(Y=y)$ (indicating distinct label distributions in the sub-groups), it is impossible for both the Sufficiency and Independence rules to hold simultaneously [5].

Separation Rule is:

$$\mathbb{E}_{\mathcal{D}_0}[\hat{Y}|Y=s] = \mathbb{E}_{\mathcal{D}_1}[\hat{Y}|Y=s], \forall t \in Y \tag{4}$$

In binary classification, the Separation rule is also referred to as Equalized Odds (EO) [18]. Additionally, [3] have further illustrated that if $P_{\mathcal{D}_0}(Y=y) \neq P_{\mathcal{D}_1}(Y=y)$ and the joint distribution of $(Y, \hat{Y})$ has a positive probability in $\mathcal{D}_0$ and $\mathcal{D}_1$, then it is impossible for both the Sufficiency and Separation rules to coexist [5] (please refer to Appendix for further details).

2.2 *Invariant Risk Minimization*

IRM operates under the assumption that there are multiple environments $\mathcal{E} := \{e_1, e_2, ..., e_E\}$ within the sample space $\mathcal{X} \times \mathcal{Y}$, each characterized by distinct joint distributions. Furthermore, it assumes that the correlation between the spurious features and labels varies inconsistently across these environments. The predictor $f(\cdot;\theta)$ in IRM is expressed as a composite function of a representation $\phi(\cdot;\Phi)$ and a classifier $h(\cdot;v)$, formulated as $f(\cdot;\theta) = h(\phi(\cdot;\Phi);v)$, where $\theta = \{v, \Phi\}$ represents the trainable parameters. The fundamental idea is that if a predictor $f(\cdot;\theta)$ performs effectively across all environments, it suggests that the correlation between the spurious features and labels is not accurately captured[35, 2]. In these cases, the data representation function ϕ elicits an invariant predictor across environments $\mathcal{E}$ if and only if for all latent z in the intersection of the supports of $\phi(X^e)$ we have $\mathbb{E}[Y^e|\phi(X^e)=z] = \mathbb{E}[Y^{e'}|\phi(X^{e'})=z]$, for all $e, e' \in \mathcal{E}$ (For loss functions such as the mean squared error and the cross-entropy, optimal classifiers can be written as conditional expectations). Please refer to Appendix for more details. Consequently, IRM aims to minimize a specific IRM risk to identify such a robust predictor. Several approaches have been proposed to enhance IRM: [23, 40] advocate for penalizing the variance of risks across different environments, while [7, 41] attempt to estimate the violation of invariance by training neural networks. Moreover, theoretical guarantees for IRM on linear models with adequate training environments are provided by [2, 36, 8].

Two popular risks are:

$$\mathcal{R}^{\text{IRMv1}}(\mathcal{D},\theta) := \sum_e \mathcal{L}(\mathcal{D}^e,\theta) + \lambda\|\nabla_v \mathcal{L}(\mathcal{D}^e,\theta)\|_2^2 \tag{5}$$

$$\mathcal{R}^{\text{REx}}(\mathcal{D},\theta) := \sum_e \mathcal{L}(\mathcal{D}^e,\theta) + \lambda\mathbb{V}_e[\mathcal{L}(\mathcal{D}^e,\theta)] \tag{6}$$

where $\mathcal{D} = \bigcup^e \mathcal{D}^e$ denotes the data drawn from all environments, where $\mathcal{D}^e$ represents the data from environment e. The expression $\mathbb{V}_e[\mathcal{L}(\mathcal{D}^e,\theta)]$ signifies the variance of the loss across various environments.

However, it has been observed that IRM exhibits diminished efficacy when applied to overparameterized neural networks [17, 11]. [25] elucidates that this limitation can largely be attributed to the problem of overfitting. Consequently, utilizing these methods directly for addressing fairness concerns is not straightforward.

2.3 Reweighting

Sample reweighting constitutes a classical approach for addressing various tasks such as distribution shifts, imbalanced classification, and fairness concerns. Here, we specifically delve into reweighting methodologies associated with fairness considerations. Fairness with Adaptive Weights [6] imposes constraints on the sum of weights across sensitive groups to ensure equality, assigning weights to each sample based on its likelihood of misclassification. Adaptive Sensitive Reweighting to Mitigate Bias [22] assigns weights to samples based on their alignment with the unobserved true labeling. [24] intricately models the impact of each training sample on fairness-related metrics and predictive utility. Additionally, [43] utilizes Neural Networks to reweigh samples, aiming to achieve causal fairness. To the best of our knowledge, no reweighting method has been specifically applied to achieve the sufficiency rule in fairness. Furthermore, our method stands apart from heuristic reweighting methods, as it does not necessitate complex hyper-parameter selection processes.

3 Reweighting to Achieve Sufficiency Rule

3.1 Bilevel Formulation of Reweighting

Given a dataset $\mathcal{D}$ constituted as a set $\{(x_i, y_i)\}_{i=1}^n$, where each (x_i, y_i) is drawn from $\mathcal{X} \times \mathcal{Y}$, the weighted empirical loss is defined as $\mathcal{L}(\mathcal{D}, \theta; w) := \frac{1}{n}\sum_{i=1}^n w_i l(f(x_i;\theta), y_i)$, with $f(\cdot;\theta)$ representing a neural network parameterized by θ, $l(\cdot,\cdot)$ indicating the loss function (e.g., cross-entropy or least squares loss), and $w_i \in \mathbb{R}^+$ denoting the non-negative weight assigned to each sample.

We formulate the objective of learning sample weights to mitigate reliance on sensitive features as the subsequent bilevel optimization problem:

$$\min_{w\in\mathcal{W}} \mathcal{R}(\mathcal{D}, \theta^*(w)), \qquad (7)$$
$$\text{s.t. } \theta^*(w) \in \arg\min_{\theta} \mathcal{L}(\mathcal{D}, \theta; w)$$

Here, w denotes a vector of sample weights with a length of n, indicating the importance of each training sample, where each component w_i of w satisfies $w_i \geq 0$. Any IRM Risk $\mathcal{R}(\mathcal{D}, \theta)$ discussed in Section 2 can function as the outer objective. In our subsequent experiments, we employ the risk (5), denoted as IRMv1. Within the inner loop, we minimize the weighted ERM loss on the training samples to derive a model $\theta^*(w)$, while within the outer loop, we evaluate the learned model's reliance on sensitive features through IRM Risk and adjust the sample weights accordingly. By iteratively alternating between the inner and outer loops, the sample weights gradually adjust to a state where they can yield satisfactory IRM/fairness performance via straightforward ERM training. It's worth noting that, instead of different environmental settings as in the IRM scenario, the fairness problem involves distinct sensitive groups, such as $\mathcal{D}_0$ and $\mathcal{D}_1$, as depicted in Section 2. Although we showcase our approach within the context of binary sensitive groups in this section, it can be readily extended to scenarios involving multi-categorical sensitive groups (refer to the experimental details on the toxic comments dataset and COMPAS dataset in Section 4).

Our approach provides the following benefits: 1) by establishing an implicit mapping from the sample weight space to the model parameter space in the outer loop, where the former consistently remains smaller than the latter in deep learning tasks (as detailed in Section 1), we effectively address overfitting issues typically associated with IRM regularization-based methods (the objective of the outer loop); 2) our approach avoids the need to impose specific fairness constraints, thereby circumventing the challenge of determining critical hyperparameters for fairness regularization to achieve a better trade-off between fairness and accuracy.

3.1.1 Connection to the Sufficiency Rule:

We elaborate the connection between the outer loop of our bilevel objective and the Sufficiency Rule [37].

Proposition 1. *In a classification task, minimizing the loss in the outer loop as illustrated in details in Section 2.2 is tantamount to:*

$$\mathbb{E}_{\mathcal{D}_0}[Y|Z=z] = \mathbb{E}_{\mathcal{D}_1}[Y|Z=z], \textit{ (IRM definition)} \qquad (8)$$
$$\mathbb{E}_{\mathcal{D}_0}[Y|\hat{Y}=h^*(z)] = \mathbb{E}_{\mathcal{D}_1}[Y|\hat{Y}=h^*(z)] \qquad (9)$$

where h_0^*, h_1^* are the optimal predictor for each sub-group, $h^* = h_1^* = h_0^*$ and $z = \phi(x)$.

Proposition 1 illustrates that the objective of the outer loop loss aligns with the sufficiency rule in binary classification.

3.2 Enhance Reweighting by Sparsity and Continuation

We discretize the optimization method [45] here,

$$\min_{m\in\mathcal{C}} \mathcal{R}(\mathcal{D}, \theta^*(m)), \qquad (10)$$
$$\text{s.t. } \theta^*(m) \in \arg\min_{\theta} \mathcal{L}(\mathcal{D}, \theta; m)$$

where the mask $m \in \{0,1\}_n$ represents a binary vector, and $m_i = 1$ denotes that sample i is included in the training set, otherwise it is excluded. K is a positive integer that determines the size of the selected set, and $\mathcal{C} = \{m : m_i \in \{0,1\}, \|m\|_0 \leq K\}$ denotes the feasible region of m. Essentially, the inner loop trains the network to converge on the selected set to obtain the model $\theta^*(m)$, while the outer loop assesses the loss of $\theta^*(m)$ on the entire set and optimizes it to guide the learning of m.

The distinction between our discrete bilevel formulation (10) and the original bilevel formulation (7) lies in the absence of individual weights w_i for each sample in the sparse formulation (10). We opt for this sparse formulation for several reasons: 1) empirical results demonstrate satisfactory performance even without these weights; 2) it simplifies the development of an efficient training algorithm; 3) excluding noisy data enhances the robustness of the model.

Given the discrete nature of the mask m, directly solving the bilevel optimization problem (10) is intractable due to its NP-hard nature. Hence, we adopt a continualization approach [45] via probabilistic reparameterization to render gradient-based optimization feasible. We treat each mask m_i as an independent binary random variable and transform the problem into the continuous probability space. Specifically, we reparameterize m_i as a Bernoulli random variable with probability s_i for being 1 and $1 - s_i$ for being 0, i.e., $m_i \sim \text{Bern}(s_i)$, where $s_i \in [0,1]$. Assuming independence among the variables m_i, the distribution function of m

becomes $p(m|s) = \prod_{i=1}^{n}(s_i)^{m_i}(1-s_i)^{(1-m_i)}$. Thus, we control the selected size through the sum of probabilities s_i, since $\mathbb{E}_{m\sim p(m|s)}[\|m\|_0] = \sum_{i=1}^{n} s_i$. Consequently, $\mathcal{C}$ can be relaxed into $\tilde{\mathcal{C}} = \{s_i : 0 \le s_i \le 1, \|s\|_1 \le K\}$. Finally, problem (10) naturally relaxes into the following:

$$\min_{s\in\tilde{\mathcal{C}}} \Psi(s) = \mathbb{E}_{p(m|s)}[\mathcal{R}(\mathcal{D}, \theta^*(m))], \quad \text{s.t. } \theta^*(m) \in \arg\min_{\theta} \mathcal{L}(\mathcal{D}, \theta; m) \tag{11}$$

where $\tilde{\mathcal{C}} = \{s_i : 0 \le s_i \le 1, \|s\|_1 \le K\}$ is the domain.

Several beneficial aspects of our formulation (11) include:

1. Our formulation serves as a close relaxation (though not equivalent) of Problem (10). This is evident for the following reasons:
   (a) It is apparent that $\min_{s\in\tilde{\mathcal{C}}}\Psi(s) \le \min_{m\in\mathcal{C}}\Psi(m)$ since any deterministic binary mask m can be represented as a specific stochastic one by setting s_i to either 0 or 1.
   (b) Our constraint $\tilde{\mathcal{C}}$ induces sparsity on s through the l_1-norm and the range $[0, 1]$, resulting in most components of the optimal s being either 0 or 1. Therefore, our eventually learned stochastic weight is nearly deterministic.
2. Due to the sparsity constraint, the size of the selected set in the inner loop, remains small, which greatly enhances the efficiency of optimizing θ^* (refer to details in Appendix).
3. As indicated in Eq. (13), our outer objective $\Psi(s)$ is differentiable, enabling the utilization of general gradient-based methods for optimization.

3.3 Optimization Method

Current bilevel optimization algorithms [34, 16] typically incur high computational costs owing to the resource-intensive implicit differentiation inherent in their chain-rule-based gradient estimator. Specifically, if employed in our context, they commonly approximate the gradient in the following manner:

$$\nabla_s\Psi(s) \approx \nabla_s\theta^*(m)\nabla_\theta\mathcal{R}(\mathcal{D}, \theta^*(m)) \tag{12}$$

Hence, they need to compute the implicit differentiation of the inner loop optimum, i.e, $\nabla_s\theta^*(m)$, which is expensive since they have to compute the inverse of a huge hessian matrix or unroll the backward propagation for multiple steps. Even though some efficient bilevel optimization algorithms have been proposed to alleviate the computational burden (for instance, [28] adopted Neumann series to approximate the hessian inverse), the approximation is nevertheless time-consuming.

The probabilistic formulation (11) of the bilevel problem allows us to circumvent costly computations by computing the gradient using forward propagation instead of backward propagation. This can be illustrated by the following equations:

$$\begin{aligned}
\nabla_s\Psi(s) &= \nabla_s\mathbb{E}_{p(m|s)}[\mathcal{R}(\mathcal{D}, \theta^*(m))] \\
&= \nabla_s \int \mathcal{R}(\mathcal{D}, \theta^*(m))p(m|s)dm \\
&= \int \mathcal{R}(\mathcal{D}, \theta^*(m))\frac{\nabla_s p(m|s)}{p(m|s)}p(m|s)dm \\
&= \int \mathcal{R}(\mathcal{D}, \theta^*(m))\nabla_s \ln p(m|s)p(m|s)dm \\
&= \mathbb{E}_{p(m|s)}[\mathcal{R}(\mathcal{D}, \theta^*(m))\nabla_s \ln p(m|s)]
\end{aligned} \tag{13}$$

This indicates that $\mathcal{R}(\mathcal{D}, \theta^*(m))\nabla_s \ln p(m|s)$ serves as an unbiased stochastic gradient of $\nabla_s\Psi(s)$. Consequently, with the inner loop optimum $\theta^*(m)$ at hand, we can update s (probability) via projected stochastic gradient descent:

$$s \leftarrow \mathcal{P}_{\tilde{\mathcal{C}}}(s - \eta\mathcal{R}(\mathcal{D}, \theta^*(m))\nabla_s \ln p(m|s)) \tag{14}$$

It's evident that this approach does not entail any implicit differentiation, and its component $\mathcal{R}(\mathcal{D}, \theta^*(m))$ can be computed through forward propagation. Additionally, $\ln p(m|s)$ exhibits a straightforward form, and the projection possesses a closed-form solution [45] given the simplicity of the constraint $\tilde{\mathcal{C}}$. Consequently, we can efficiently update s.

Thus, we can tackle our bilevel optimization problem (11) by alternately: 1) sampling m, i.e., a selected set, from $p(m|s)$ for the inner loop and training the model on this selected set to obtain $\theta^*(m)$; 2) updating the probability s. The details are shown in Algorithm 1.

Algorithm 1 Reweighting for the Sufficiency Rule

```
Require: a neural network parameterized by $\theta$, a dataset $\mathcal{D}$, and a
    selected set size $K$.
1: Initiate probabilities $s^1$ as $\frac{K}{|\mathcal{D}|}\mathbf{1}$.
2: for iteration $t$ of training, where $t$ is from 1 to $T$. do
3:     Sample $m$ based on the probability vector $s^t$.
4:     Continue training the inner loop until convergence achieved:
       $\theta^*(m) \leftarrow \underset{\theta}{\text{argmin}}\hat{\mathcal{L}}(\theta; m)$
5:     Sample a mini-batch $\mathcal{K}$ from the dataset $\mathcal{D}$ :
       $\mathcal{K} = \{(x_1, y_1), ..., (x_\mathcal{K}, y_\mathcal{K})\}$
6:     Update $s$ according to $\theta^*(m)$ and $\mathcal{K}$.
       $s^{t+1} \leftarrow \mathcal{P}_{\tilde{\mathcal{C}}}(s^t - \eta\mathcal{R}_\mathcal{K}(\mathcal{D}, \theta^*(m))\nabla_s \ln p(m|s^t))$
7: end for
8: Output: The selected set $\{(x_i, y_i) : m_i = 1 \text{ and } (x_i, y_i) \in \mathcal{D}\}$
    where $m$ is sampled from $p(m|s^{T+1})$.
```

4 Experiments

We adopt the aforementioned sufficiency gap as the fair metric and accuracy as the metric for utility. Our neural network models are trained on an Intel(r) Core(TM) i7-8700 CPU. The networks in our experiments are built using the Pytorch package [33].

4.1 Baselines

We compare our method with (I) Empirical Risk Minimization (**ERM**) which trains the model without considering fairness; (II) No Utility-Cost Fairness via Data Reweighing (**NUF**) [24]; (III) Fair Representation Learning through Implicit Path Alignment (**IPA**) [37], an approach in the fair representation learning to achieve also the sufficiency rule; (IV) Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation (**AR**) [44]. Notably, the baseline (IV) is grounded in Demographic Parity (DP), illustrating their general incompatibility with addressing the sufficiency rule. Additionally, we include the original Invariant Risk Minimization regularization [2], denoted as IRMv1, which incorporates a gradient penalty to encourage invariance across different groups. Even though it is designed for another purpose, as shown earlier in Section 3, it has potential to address fairness to reach the sufficiency. Results are averaged over five repetitions. Further experimental results are provided in Appendix.

Table 1: Accuracy and ΔSuf in Toxic comments (left) and CelebA datasets (right)

Toxic comments	Accuracy(↑)	ΔSuf(↓)
ERM(I)	0.768±0.004	0.173±0.008
NUF(II)	0.762±0.007	0.190±0.008
IPA(III)	0.745±0.007	0.091±0.012
AR(IV)	0.756±0.006	0.128±0.097
Ours(V)	**0.763±0.004**	**0.028±0.004**
IRMv1(VI)	0.753±0.004	0.068±0.008

CelebA	Accuracy(↑)	ΔSuf(↓)
ERM(I)	0.956±0.005	0.210 ±0.094
NUF(II)	0.947±0.007	0.104±0.004
IPA(III)	0.938±0.103	0.092±0.161
AR(IV)	0.950±0.012	0.197±0.007
Ours(V)	**0.953±0.094**	**0.045±0.004**
IRMv1(VI)	0.946±0.009	0.088±0.007

Table 2: Accuracy and ΔSuf in Adult (left) and COMPAS datasets (right)

Adult	Accuracy(↑)	ΔSuf(↓)
ERM(I)	0.831±0.014	0.160±0.007
NUF(II)	0.815±0.017	0.068±0.015
IPA(III)	0.810±0.004	0.058±0.024
AR(IV)	0.820±0.023	0.230±0.014
Ours(V)	**0.827±0.016**	0.036±0.007
IRMv1(VI)	0.825±0.018	**0.032±0.012**

COMPAS	Accuracy(↑)	ΔSuf(↓)
ERM(I)	0.652±0.024	0.276±0.094
NUF(II)	0.633±0.032	0.156±0.008
IPA(III)	0.647±0.017	0.097±0.009
AR(IV)	**0.659±0.019**	0.285±0.018
Ours(V)	0.647±0.004	**0.068±0.015**
IRMv1(VI)	0.645±0.008	0.078±0.017

4.2 Datasets and Experiment Setups

The **toxic comments dataset** [20] presents a binary classification challenge in natural language processing (NLP), aiming to determine whether a comment exhibits toxicity. Originally, the labeling process for this dataset is not binary due to involvement from multiple annotators, leading to potential discrepancies. To address this, we adopt a straightforward strategy where a comment is classified as toxic if at least one annotator marks it as such, similar to the approach in [37]. Notably, some comments in this dataset are annotated with identity attributes such as gender and race. It has been observed that the race attribute correlates with the toxicity label, posing a risk of predictive discrimination. Therefore, we designate race as the protected feature and specifically focus on two sub-groups: Black and Asian. For computational efficiency, we begin by leveraging a pre-trained BERT model [15] to extract word embeddings, resulting in vectors of 748 dimensions.

The **CelebA dataset** [27] comprises approximately 200K images featuring celebrity faces, each associated with 40 human-annotated binary attributes such as gender, hair color, and age. For our experiment, we randomly partitioned the dataset, selecting approximately 82K images for training and 18K for validation. We employed the ResNet-18 architecture [19], pre-trained on ImageNet [14], omitting the final fully-connected layer to obtain embeddings of 512 dimensions for simplicity. Within the CelebA dataset, our specific task involves predicting hair color ({blond, dark}) based on the image input. Notably, the gender attribute ({male, female}) is correlated with hair color.

For experiments on tabular data, we use the **Adult dataset** [21] and the **COMPAS dataset** [30] (For more details of the datasets, please refer to Appendix). Adult dataset used personal information such as education level and working hours per week to predict whether an individual earns more or less than $50,000 per year. We use gender as the sensitive feature in Adult dataset. COMPAS dataset is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending. We use race as the sensitive feature. Here we report the sufficiency gap between two sub-groups of African American and Caucasian even though that ethnicity group is a multi-categorical feature. Please refer to Appendix for more data and training details.

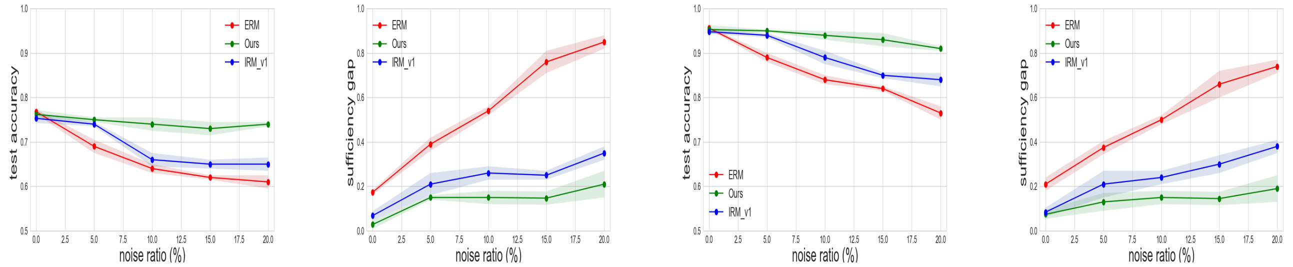
4.3 Analysis

4.3.1 Performance Comparison

In Table 1 and Table 2, we present the accuracy and sufficiency gap metrics. Notably, we observe that the Demographic Parity (DP) based fair method (IV) is incompatible with the sufficiency rule, as evidenced by its tendency to increase ΔSuf even surpassing that of ERM. On the other hand, baselines (III, VI), which aim to track the sufficiency rule, exhibit improved sufficiency gap ΔSuf with comparable accuracy, albeit inferior to our approach in Table 1. This discrepancy may stem from an overparameterization issue, as previously discussed. Our method consistently demonstrates a superior Accuracy-Fairness trade-off, significantly enhancing sufficiency without substantial accuracy loss. We observe a similar performance pattern on tabular datasets (Table 2). However, the performance drop of baselines (III, VI) is less pronounced. This discrepancy may be attributed to the comparatively smaller DNN models utilized in training on tabular datasets, which are less susceptible to overparameterization compared to the toxic comments dataset and CelebA dataset.

4.3.2 Robustness with Noisy Data

We extend our experimentation to scenarios where the dataset incorporates corrupted labels, aiming to demonstrate the robustness of our approach. Following the model configuration outlined in Section 4.2, we introduce symmetric noise [38] into the dataset. Notably, as illustrated in Figure 1, our method exhibits robustness towards variations in the dataset's label quality, as evidenced by consistent performance in both accuracy and sufficiency gap metrics. This robustness can be attributed to the comprehensive information assimilated through iterative sampling, leading to the construction of the final weight vector w. Essentially, the sparsity induced by our method facilitates the



(a) Accuracy vs. Noise ratio on Toxic comments

(b) Sufficiency gap vs. Noise ratio on toxic comments

(c) Accuracy vs. Noise ratio on CelabA

(d) Sufficiency gap vs. Noise ratio on CelabA

Figure 1: Change of accuracy and sufficiency gap under different noise ratios on Toxic comments and CelebA datasets, which shows that our method is robust when the data label is noisy.

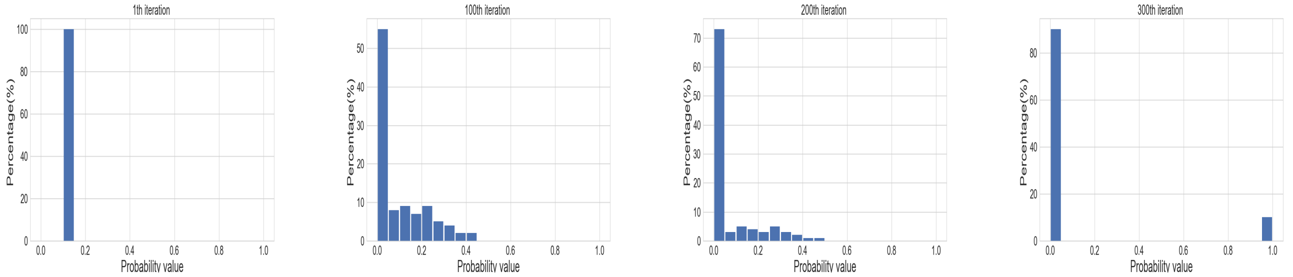


Figure 2: The evolution of probability score distribution during the search process reveals a trend where most probabilities tend to converge towards either 0 or 1. This convergence ultimately leads to deterministic weights and convergence of the algorithm.

elimination of noisy data samples, thus preserving the model's effectiveness.

4.3.3 *Sensitivity to Choices of K*

The selected sizes for the Toxic comments and CelebA experiments are 5000 and 10000, respectively, as shown in Figure 3. As the selected sizes increase, we observe an improvement in both accuracy and sufficiency gap performance. Yet, beyond a certain threshold, this improvement plateaus, aligning with the corset concept [31]. Corset theory suggests that there exists a small subset capable of summarizing the larger dataset effectively. Training exclusively on this condensed set yields competitive performance compared to training on the entire dataset.

4.3.4 *Convergence of Probabilities during Search*

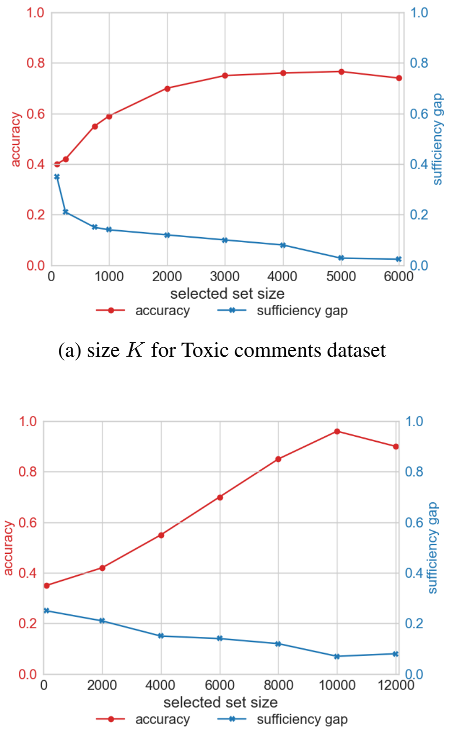
A simplified approach is taken by selecting 1000 samples from a larger pool of 10000 training data instances (CelebA). Figure 2 illustrates the evolution of probability distributions throughout the search process. Initially, all sample probabilities are uniformly distributed at 0.1. Over the course of the search, most of these probabilities tend to converge towards either 0 or 1, indicative of diminishing uncertainty. Consequently, a sparse mask with minimal variance is formed, reflecting a nearly deterministic pattern in weight assignment. This trend ultimately leads to the establishment of deterministic weights, signifying algorithmic convergence.

4.3.5 *Gradual Change of Group Weights*

In Figure 4, we depict the training dynamics of sample weight fractions from our CelebA experiment. Initially, all sample weights are uniformly set to 1. The weight fraction of the (Male, Blond Hair) group begins at a mere 0.085%. Following 100 iterations of updates, this fraction gradually increases to approximately 20%. Simultaneously, the weight fraction of the (Male, Dark Hair) group decreases to approximately 20%, while both the (Female, Dark Hair) and (Female, Blond Hair) groups stabilize at approximately 30%. Interestingly, in [6], the assumption is that bias is introduced due to under-representation of the minority groups, hence, they upweight/-downweight sensitive groups to the same importance level. Figure 4 demonstrates that even though we do not constrain on the group level importance, somehow, our method possesses the ability to dynamically adjust the weight fraction of (sub)-groups automatically.

5 Discussion and Conclusion

We presented a model agnostic sample reweighing method to achieve the sufficiency rule of fairness. We formulated this problem as a bilevel optimization to learn sample weights. We further enhance our method with sparsity constraints to improve training speed. Then, we analyzed the sufficiency gap and prediction accuracy of the reweighting algorithm, demonstrating its superior performance over state-of-the-art approaches. The empirical results also show that our method is robust towards noisy labels. One limitation of our framework is that the overall performance of IRMv1, in terms of test accuracy, consistently improves when there is a significant difference between the



(a) size K for Toxic comments dataset

(b) size K for CelebA dataset

Figure 3: Choices of K (selected set size). The size is set to 5000 for Toxic comments and 10000 for CelebA.

training environments [10]. This suggests that, in terms of fairness, IRMv1 is more effective when there is greater disparity between the sensitive sub-groups.

References

[1] M. A. Ahmad, A. Patel, C. Eckert, V. Kumar, and A. Teredesai. Fairness in machine learning for healthcare. In *KDD*, pages 3529–3530. ACM, 2020.

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

[3] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[4] P. Bühlmann. Invariance, causality and robustness, 2018.

[5] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12, Mar. 2022. ISSN 2045-2322.

[6] J. Chai and X. Wang. Fairness with Adaptive Weights. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2853–2866. PMLR, 2022.

[7] S. Chang, Y. Zhang, M. Yu, and T. S. Jaakkola. Invariant rationalization. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR, 2020.

[8] Y. Chen, E. Rosenfeld, M. Sellke, T. Ma, and A. Risteski. Iterative feature matching: Toward provable domain generalization with logarithmic environments. In *NeurIPS*, 2022.

[9] V. Cherepanova, V. Nanda, M. Goldblum, J. P. Dickerson, and T. Goldstein. Technical challenges for training fair neural networks. *CoRR*, abs/2102.06764, 2021.

[10] Y. J. Choe, J. Ham, and K. Park. An empirical study of invariant risk minimization. *CoRR*, abs/2004.05007, 2020.

[11] Y. J. Choe, J. Ham, and K. Park. An empirical study of invariant risk minimization, 2020.

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

[13] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. In *NeurIPS*, 2020.

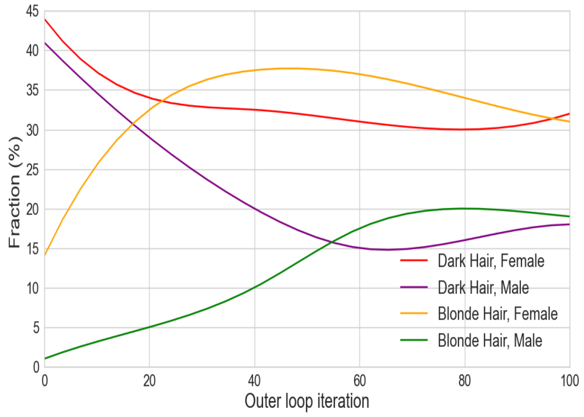


Figure 4: The fluctuation in the distribution of group weight fractions for ResNet-18 on the CelebA dataset is notable. Specifically, there's a shift to 20% for both the (Male, Blond Hair) and (Male, Dark Hair) gourps. Similarly, the (Female, Dark Hair) and (Female, Blond Hair) groups see their fractions adjusted to 30%. These observations suggest that our methodology is capable of autonomously adapting weight fractions across various (sub)-groups.

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[16] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3748–3758. PMLR, 2020.

[17] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *ICLR*. OpenReview.net, 2021.

[18] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[20] Jigsaw. Toxic comment classification challenge, 2018. URL https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[21] R. Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 202–207, Portland, Oregon, Aug. 1996. AAAI Press.

[22] E. Krasanakis, E. S. Xioufis, S. Papadopoulos, and Y. Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *WWW*, pages 853–862. ACM, 2018.

[23] D. Krueger, E. Caballero, J. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 2021.

[24] P. Li and H. Liu. Achieving fairness at no utility cost via data reweighing with influence. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12917–12930. PMLR, 2022.

[25] Y. Lin, H. Dong, H. Wang, and T. Zhang. Bayesian invariant risk minimization. In *CVPR*, pages 16000–16009. IEEE, 2022.

[26] L. T. Liu, M. Simchowitz, and M. Hardt. The implicit fairness criterion of unconstrained learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 4051–4060. PMLR, 2019.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[28] J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 1540–1552. PMLR, 2020.

[29] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel. Learning adversarially fair and transferable representations. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3381–3390. PMLR, 2018.

[30] Mattu, J. Angwin, L. Kirchner, Surya, and J. Larson. How We Analyzed the COMPAS Recidivism Algorithm, 2016.

[31] B. Mirzasoleiman, J. A. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 6950–6960. PMLR, 2020.

[32] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 2019.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

[34] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 737–746. JMLR.org, 2016.

[35] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, 2015.

[36] E. Rosenfeld, P. K. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *ICLR*. OpenReview.net, 2021.

[37] C. Shui, Q. Chen, J. Li, B. Wang, and C. Gagné. Fair representation learning through implicit path alignment. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 20156–20175. PMLR, 2022.

[38] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023.

[39] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon. Learning controllable fair representations. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173. PMLR, 2019.

[40] C. Xie, F. Chen, Y. Liu, and Z. Li. Risk variance penalization: From distributional robustness to causality. *CoRR*, abs/2006.07544, 2020.

[41] Y. Xu and T. S. Jaakkola. Learning representations that support robust transfer of predictors. *CoRR*, abs/2110.09940, 2021.

[42] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org, 2013.

[43] X. Zhao, K. Broelemann, S. Ruggieri, and G. Kasneci. Causal fairness-guided dataset reweighting using neural networks. In *IEEE Big Data*, pages 1386–1394. IEEE, 2023.

[44] X. Zhao, S. Fabbrizzi, P. R. Lobo, S. Ghodsi, K. Broelemann, S. Staab, and G. Kasneci. Adversarial reweighting guided by wasserstein distance for bias mitigation, 2023.

[45] X. Zhou, R. Pi, W. Zhang, Y. Lin, Z. Chen, and T. Zhang. Probabilistic bilevel coreset selection. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 27287–27302. PMLR, 2022.

A Related Work

A.1 Sufficiency Rule

The Sufficiency principle adopts a perspective where individuals receiving the same model decision are expected to experience similar outcomes regardless of their sensitive attributes. In this context, A represents the sensitive attribute.

On the other hand, Separation deals with error rates concerning the proportion of errors relative to the ground truth. For instance, it considers the number of individuals whose loan application would have been approved but were actually denied. Sufficiency, however, considers the number of individuals who would default on their loan among those who were granted it.

From a mathematical standpoint, this distinction resembles that between recall (or true positive rate) and precision, denoted as $P[\hat{Y} = 1|Y = 1]$ and $P[Y = 1|\hat{Y} = 1]$, respectively. A fairness principle that focuses on this type of error rate is known as Predictive Parity, also termed as the outcome test:

$$P[Y = 1|A = a, \hat{Y} = 1] = P[Y = 1|A = b, \hat{Y} = 1], \forall a, b \in \mathcal{A} \tag{15}$$

In other words, the model's precision should remain consistent across different sensitive groups. If we extend this requirement to apply to the scenario where $Y = 0$, we obtain the following statement of conditional independence:

$$Y \perp A | \hat{Y} \tag{16}$$

This concept is known as sufficiency.

Predictive Parity, including its broader form of sufficiency, focuses on ensuring parity in errors among individuals receiving the same decision. Predictive Parity, in particular, considers the viewpoint of the decision-maker, as they categorize individuals based on decisions rather than actual outcomes. For instance, in the context of credit lending, sufficiency is more under the control of the decision-maker than separation. This is because achieving parity given a decision is directly observable, whereas parity given truth is only known afterward. Furthermore, the group of individuals granted a loan ($\hat{Y} = 1$) is less susceptible to selection bias compared to the group of loan repayments ($Y = 1$). We have information only on repayment for the $\hat{Y} = 1$ group, while nothing is known about the others ($\hat{Y} = 0$).

Similarly, other group metrics can be defined, such as Equality of Accuracy across groups: $P[\hat{Y} = Y|A = a] = P[\hat{Y} = Y|A = b]$, for all $a, b \in \mathcal{A}$, focusing on unconditional errors, among others.

A.1.1 Incompatibility

In most cases, even in classification setting, the actual output of a model is not a binary value, but rather a score $S \in \mathbb{R}$.

1. if $Y \not\perp A$, then it's impossible for sufficiency and independence to coexist. Consequently, if there's an imbalance in base rates among groups identified by A, enforcing both sufficiency and independence simultaneously is unfeasible.
2. if $Y \not\perp A$ and the distribution (A, S, Y) is strictly positive, then separation and sufficiency cannot both be achieved simultaneously. This implies that separation and sufficiency can coexist only under two conditions: when there's no imbalance in sensitive groups (indicating independence of the target from sensitive attributes), or when the joint probability (A, S, Y) is degenerate. In the case of binary targets, this degeneracy occurs when there are specific values of A and S for which only $Y = 1$ (or $Y = 0$) holds. In other words, separation and sufficiency coincide when the score perfectly resolves the uncertainty in the target. For instance, the ideal classifier where $S = Y$ effortlessly satisfies both sufficiency and separation.

A.2 Invariant Risk Minimization

Minimizing training error leads machines into recklessly absorbing all the correlations found in training data. Understanding which patterns are useful has been previously studied as a correlation-versus-causation dilemma, since spurious correlations stemming from data biases are unrelated to the causal explanation of interest. Following this line, IRM leverage tools from causation to develop the mathematics of spurious and invariant correlations, in order to alleviate the excessive reliance of machine learning systems on data biases, allowing them to generalize to new test distributions.

B Experiment Details and Results

B.1 Datasets

Adult dataset The Adult dataset was drawn from the 1994 United States Census Bureau data. It used personal information such as education level and working hours per week to predict whether an individual earns more or less than \$50,000 per year. The dataset is imbalanced – the instances made less than \$50,000 constitute 25% of the dataset, and the instances made more than \$50,000 include 75% of the dataset. As for gender, it is also imbalanced. We use age, years of education, capital gain, capital loss, hours-per-week, etc., as continuous features, and education level, gender, etc., as categorical features.

COMPAS dataset COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending (recidivism). The COMPAS dataset includes the processed COMPAS data between 2013-2014. The data cleaning process followed the guidance in the original COMPAS repo. It Contains 6172 observations and 14 features. In our causal graph, we use 7 features. Due to the limited size of COMPAS dataset, it does not perform so well on NN based tasks. Note that we need more pre-processing on the tabular datasets. We normalize the continuous features and use one-hot encoding to deal with the categorical features.

B.2 Training Details

In our experiments, we set the following hyperparameters for optimization. In the inner loop, the model undergoes training for 100 epochs using Stochastic Gradient Descent (SGD) with a learning rate of 0.1 and momentum of 0.9. In the outer loop, the probabilities s are optimized using Adam with a learning rate of 2.5 and a cosine scheduler. The outer loop is updated iteratively for 500 to 2000 times. It's worth mentioning that we employ architectures with fully connected layers for the classifier.

Toxic comments We split the training, validation and testing set as 70%, 10% and 20%. The mini-batch-size is set as 500.

CelebA The training/validation/test set are around 82K, 18K and 18K. The batch-size is set as 1000.

Effectiveness of Sparsity in Promoting Training Speed An experiment is conducted on the CelebA dataset to compare the training speed between our method with and without a sparsity constraint on sample sizes. The inclusion of the constraint significantly reduces the inner loop computation time, decreasing it from 9.24 to 5.72 hours.

Repetition We repeat experiments on each dataset five times. Before each repetition, we randomly split data into training data and test data for the computation of the standard errors of the metrics.