
On Group Sufficiency Under Label Bias

Haoran Zhang Olawale Salaudeen Marzyeh Ghassemi

Massachusetts Institute of Technology

{haoranz, olawale, mghassem}@mit.edu

Abstract

Real-world classification datasets often contain label bias, where observed labels differ systematically from the true labels at different rates for different demographic groups. Machine learning models trained on such datasets may then exhibit disparities in predictive performance across these groups. In this work, we characterize the problem of learning fair classification models with respect to the underlying ground truth labels when given only label biased data. We focus on the particular fairness definition of group sufficiency, i.e. equal calibration of risk scores across protected groups. We theoretically show that enforcing fairness with respect to label biased data necessarily results in group miscalibration with respect to the true labels. We then propose a regularizer which minimizes an upper bound on the sufficiency gap by penalizing a conditional mutual information term. Across experiments on eight tabular, image, and text datasets with both synthetic and real label noise, we find that our method reduces the sufficiency gap by up to 7.2% with no significant decrease in overall accuracy.

1 Introduction

Machine learning systems are increasingly being deployed in safety critical settings, from parole recommendations to medical triage and financial lending [1, 2, 3]. However, prior work has shown that such systems often exhibit biases in the form of performance discrepancies between demographic groups [4, 5, 6]. There exists a range of work which propose algorithmic approaches to correct these disparities, spanning from pre-processing [7, 8, 9], to in-processing [10, 11, 12, 13], to post-processing [14, 15, 16]. However, these methods rarely consider the *source* of these biases. Prior work has shown that naively enforcing these fairness definitions can have unintended side effects, leading to worse performance for all [17, 18, 19, 20], and exacerbating other types of biases [21, 22, 16].

In this work, we take a different approach and tackle fair classification from a data-centric perspective [23], focusing on datasets that contain significant differences in label errors between protected groups, also known as *label bias* [24]. For example, in the medical diagnosis setting, women with heart disease are underdiagnosed at higher rates than men with heart disease, and are instead more likely to be misdiagnosed with mental health conditions [25]. Similarly, Black patients are less likely to be prescribed opioid painkillers than White patients in emergency departments, even after adjusting for level of pain [26]. Finally, when screening resumes for job applications, applicants with African American names are much less likely to receive an interview than those with White names [27].

When downstream models are trained on these label-biased datasets (i.e. to diagnose, prescribe, or hire), the model may learn to predict noisy labels and propagate disparities in the historical data [28, 29], leading to the observed performance disparities. When the training dataset contains label bias, our goal is to learn a fair and performant model with respect to the true unobserved labels.

Learning under label noise is a well-studied problem [30, 31, 32, 33, 34], where methods aim to learn the best model with respect to the true labels, without any consideration for per-group performance (i.e. without considering label bias). Though there has been limited work looking at fairness under

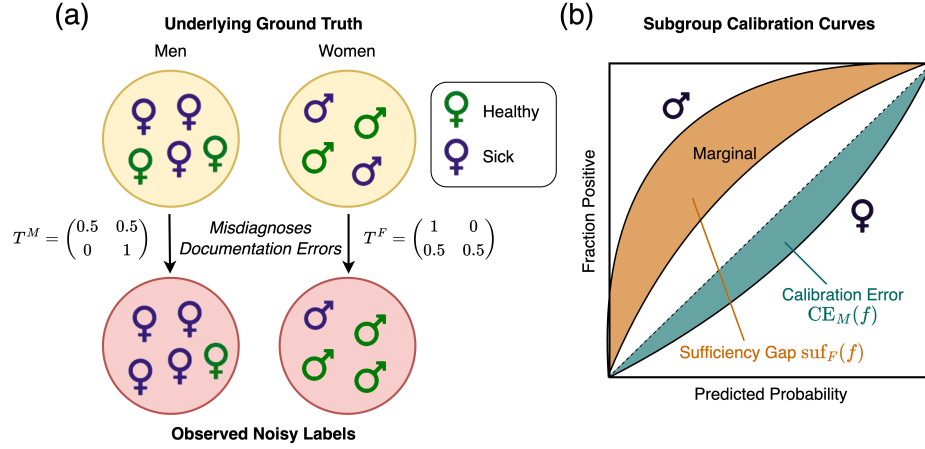


Figure 1: Our Problem Setup. (a) Observed labels are corrupted from unobserved true labels at different rates for each gender, which may be caused by misdiagnoses [25] or noisy automated processes used to extract labels [42]. (b) Training a model on these noisy labels and evaluating calibration curves with respect to true labels results in different calibration curves for each group. Specifically, each group’s calibration curve differs from the marginal curve. We propose a method to achieve group sufficiency (zero sufficiency gap for all groups) under these conditions.

label bias [35, 36, 37], these works all focus on the fairness definitions of demographic parity and equal odds [15].

Here, we study the problem of learning classification models under label bias which satisfy a different fairness property — *group sufficiency*, which corresponds to equal calibration of the risk score across groups [21, 38]. A group sufficient score can then be post-processed to achieve perfect calibration for all groups via group-agnostic post-processing [21]. Importantly, group sufficiency is incompatible with demographic parity and equal odds on the risk score under different group base rates, so it is not possible to achieve all fairness definitions simultaneously [22]. Here, we focus on group sufficiency, as prior work have shown that it is a requirement to prevent discrimination or maximize utility in many contexts [39, 40, 41]. For example, Loi and Heitz [40] argue, from a moral philosophy perspective, that risk scores being miscalibrated for a particular demographic group may constitute discrimination against that group.

We start by theoretically characterizing how class-attribute conditioned label noise distorts per-group calibration and derive necessary and sufficient conditions for group insufficiency under label bias (Section 4). This demonstrates that naive models trained on noisy labeled data are necessarily *unfair* with respect to the true underlying labels, and motivates the need for specialized methods. Building on this theory, we propose a regularizer term, CMI-REG, based on penalizing conditional mutual information (Section 5). CMI-REG can be added to any existing off-the-shelf noise-robust loss. We evaluate CMI-REG with 7 base losses on 8 datasets spanning tabular, text, and image domains, finding that we effectively reduce the sufficiency gap while maintaining, and often improving, overall performance over the base losses. To the best of our knowledge, our work is the first to study this problem setting and propose a method to achieve this fairness definition.

Contributions. Our work makes the following contributions¹:

1. We theoretically establish necessary and sufficient conditions for which label bias causes a lack of group sufficiency under class-attribute conditioned noise.
2. We propose a method to achieve sufficiency under label bias, assuming class-attribute conditioned noise. We theoretically prove that our method minimizes an upper bound on the sufficiency gap. Our method is a regularizer term that can be flexibly added to any existing loss-based noisy label learning method.
3. We empirically evaluate our method on 8 datasets with both synthetic and real-world noise, showing that it achieves sufficiency with respect to the true labels without a significant decrease in accuracy over the baselines.

¹Code: https://github.com/MLforHealth/sufficiency_label_bias

2 Related Work

2.1 Learning Under Label Noise

There has been a long line of work on the problem of learning classification models under label noise. Under uniform label noise, symmetric loss functions such as mean absolute error (MAE) are theoretically noise robust [43]. This then motivated a line of work which adapts MAE to have better optimization properties and empirical performance, including improved MAE [44], symmetric cross entropy [45], generalized cross entropy (GCE) [30], and generalized Jensen-Shannon divergence [46].

Under class conditioned noise (where the probability of observing a noisy label depends only on the true label), one direction focuses on estimating the class transition matrix. When the class transition matrix is known, a simple weighting approach is sufficient [47, 48]. Methods to estimate the transition matrix have been proposed, given a small dataset with clean labels [49, 50], anchor points for each class [51, 52], or by examining the transition matrix of an ERM model trained on noisy labels [53, 54]. Liu and Guo [31] proposes a peer loss which, in the binary case, only requires knowledge of the sum of the off-diagonal transition matrix elements.

Under instance dependent label noise, there are no theoretical guarantees in general. Methods that have done well empirically in this area include those based on training dynamics, e.g. area under the margin [55], DataMap [56], first-time k-epoch learning [57], early regularization [58], as well as methods based on training multiple simultaneous models such as Co-teaching [59] and DivideMix [32].

However, given that all methods in this category are agnostic of the group attribute and do not explicitly enforce a fairness condition, they do not explicitly satisfy any fairness properties, and thus do not succeed in our problem setup as we will later show empirically.

2.2 Sufficiency as a Fairness Definition

Sufficiency of the risk score is a commonly used group fairness definition in the classification setting, which corresponds to equal calibration curves for all groups. It is a slightly weaker definition than calibration by group (i.e. all groups being perfectly calibrated) [21]. Prior work has shown that under different per-group prevalences of the label, it is impossible to simultaneously satisfy sufficiency and probabilistic equal odds [22, 60]. Liu et al. [38] show that ERM models inherently prefer sufficiency, and that intervening to achieve probabilistic equal odds necessarily worsens this condition. Empirically, this has been observed in images [19, 61] and tabular data [62, 41]. Raghavan [39] discuss the implications of this impossibility, making a case that the choice of fairness metric is highly context dependent. Similarly, Loi and Heitz [40] discuss, from a philosophical view, when group calibration is a required condition to avoid discrimination.

In order to achieve sufficiency, prior methods have been proposed based on mutual information [63] and multi-objective optimization [64, 65]. One related concept is multicalibration [66], which seeks to achieve approximate calibration for all computationally identifiable subgroups. However, multicalibration has been found to be far less sample efficient than standard post-hoc calibration methods [67]. Both multicalibration (when applied to provided attributes) and standard post-hoc calibration methods require access to the group attribute during inference, which our method will not assume. Regardless, no prior work has studied the setting of group sufficiency under label noise, which is the focus of this work.

2.3 Fairness Under Label Bias

Prior works on fairness under label bias have only focused on demographic parity and equal odds. Wick et al. [68] use simulated data to show that enforcing statistical parity on label-biased data may already achieve the same fairness definition on clean data, with increased accuracy. Wang et al. [35] study the problem of equal odds under class-attribute conditioned label noise, and propose two solutions based on an alternate constraint using a surrogate loss, and a group-weighted peer loss [31]. Zhang et al. [37] learn fair representations under class-conditional noise for the same definitions. Wu et al. [36] study fairness under instance-dependent noise but targets counterfactual fairness [69], which has been shown to have similar drawbacks as equal odds [70]. None of these works consider group sufficiency, nor do they provide guarantees for calibration error under noise.

Finally, Mhasawade et al. [24] propose various candidate causal graphs to model the label bias scenario and derive cases when sufficiency is satisfied without further intervention, but do not propose any algorithmic solutions in the remaining cases. Under their framework, we study the $Y \rightarrow Y_i$ and $Y_i \not\perp A$ setting, under which they provide no guarantees. In contrast, we provide a theoretical characterization of this setting and propose a method to achieve group sufficiency under label bias.

3 Preliminaries

Let $X \in \mathcal{X}$ denote an input feature, $Y \in \{1, 2, \dots, K\}$ the true (clean) label, and $\tilde{Y} \in \{1, 2, \dots, K\}$ the observed (noisy) label. Additionally, let $G \in \mathcal{G}$ represent a group attribute. We have a dataset $\mathcal{D} = \{(x_i, g_i, y_i, \tilde{y}_i)\}_{i=1}^N$ with $(x_i, g_i, y_i, \tilde{y}_i) \sim P_{X,G,Y,\tilde{Y}}$, where the true labels y_i are unobserved, and we only have access to the noisy labels $\tilde{y}_i$.

In practice, we want classifiers whose predicted probabilities accurately reflect true real-world outcomes, both overall and consistently across different demographic or group attributes. To quantify how well a model achieves these goals, we introduce the standard definition of group sufficiency, which quantifies whether predictions provide equal information about outcomes, regardless of group membership.

Definition 3.1 (Group Sufficiency). A multiclass classifier $f : \mathcal{X} \rightarrow \Delta^{K-1}$ is *group sufficient* with respect to a label Y and group G if it satisfies:

$$Y \perp\!\!\!\perp G \mid f(X).$$

This condition implies that the prediction $f(X)$ captures information about Y that renders the group attribute G irrelevant given the model’s output.

Definition 3.2 (Group Sufficiency Gap). The per-group sufficiency gap, as well as the overall sufficiency gap for a model f , are defined as [38]:

$$\text{Suf}_g(f) = \left(\sum_{k=1}^K \frac{1}{K} \mathbb{E}[\|\mathbb{E}[\mathbf{1}_{Y=k} \mid f(X)_k] - \mathbb{E}[\mathbf{1}_{Y=k} \mid f(X)_k, G=g]\|^p] \right)^{1/p}$$

$$\text{Suf}(f) = \mathbb{E}_g[\text{Suf}_g(f)]$$

Similarly, let $\widetilde{\text{Suf}}$ denote the sufficiency gap with respect to $\tilde{Y}$. The *group sufficiency gap* quantifies the extent to which a model’s predictions provide unequal information about the true label Y across different groups G ; it measures deviations from the ideal group sufficiency condition $Y \perp\!\!\!\perp G \mid f(X)$. This gap is zero when $f(X)$ is group sufficient—Lemma A.1.

To formally analyze the effect of label noise and group membership on calibration and sufficiency, we introduce explicit assumptions about the data-generating process. These assumptions clarify the relationship between noisy and true labels, conditional independence, and the structure of the noise across different groups.

Assumption 3.1 (Class-Attribute Conditioned Noise). The noisy label $\tilde{Y}$ is conditionally independent of the input X given the true label Y and group attribute G :

$$\tilde{Y} \perp\!\!\!\perp X \mid (Y, G).$$

Definition 3.3 (Group-Conditional Transition Matrix). For each group $g \in \mathcal{G}$, define a row-stochastic noise transition matrix $T^g \in [0, 1]^{K \times K}$ where

$$T_{i,j}^g = \mathbb{P}(\tilde{Y} = j \mid Y = i, G = g).$$

Assumption 3.2 (Invertible Group-Conditional Transition Matrix). T^g is invertible for all groups ($g \in \mathcal{G}$).

These assumptions are standard in the literature on learning under label noise [48, 71]. Importantly, the set of singular matrices has Lebesgue measure zero within the space of all row-stochastic matrices [72], and strictly diagonally dominant matrices—which naturally arise when noise rates are sufficiently small—are always invertible [73].

Definition 3.4 (Label Bias). Let $T \in [0, 1]^{K \times K}$ where $T_{i,j} = P(\tilde{Y} = j \mid Y = i) = \sum_{g \in \mathcal{G}} T_{i,j}^g P(G = g)$. If $\exists g \in \mathcal{G}$ such that $T^g \neq T$, then we say that the observed labels $\tilde{Y}$ are *label biased* relative to Y .

4 Main Result and Problem Statement

First, we characterize the necessary and sufficient conditions for which we would expect a non-zero sufficiency gap.

Theorem 4.1 (Closed-Form Decomposition of the Sufficiency Gap). Let $s = f(X) \in \Delta^{K-1}$ be the model score, $\tilde{\pi}_g(s) = P(\tilde{Y} = \cdot \mid f(X) = s, G = g)$ the group-conditional noisy label posterior, and $\tilde{\pi}(s) = P(\tilde{Y} = \cdot \mid f(X) = s)$ the marginal counterpart. Then under Assumptions 3.1 and 3.2, for any $p \geq 1$,

$$\text{Suf}_g(f) = \left(\frac{1}{K} \mathbb{E} \left[\left\| \underbrace{((T^g)^\top)^{-1} \Delta_{\tilde{\pi}}(f(X))}_{\text{posterior shift}} + \underbrace{\tilde{\pi}(f(X)) \Delta_{T^g}}_{\text{noise matrix mismatch}} \right\|_p^p \right] \right)^{1/p}, \quad (1)$$

where $\Delta_{\tilde{\pi}}(f(X)) = (\tilde{\pi}_g(f(X)) - \tilde{\pi}(f(X)))$ and $\Delta_{T^g} = (((T^g)^\top)^{-1} - (T^\top)^{-1})$.

Corollary 4.1.1 (Necessary Condition for Sufficiency Gap). Under Assumptions 3.1 and 3.2, suppose that f satisfies group sufficiency with respect to noisy labels (i.e. $\tilde{Y} \perp\!\!\!\perp G \mid f(X)$). Then, if $\exists g \in \mathcal{G}$ such that $\text{Suf}_g(f) > 0$, it must be that $T^g \neq T$.

Thus, when f is fair with respect to the noisy labels (which ERM models trained on noisy data have a tendency to be, as the sufficiency gap is upper bounded by its excess risk [38]), label bias is *necessary* for a sufficient gap. The contrapositive of this corollary also shows that under uniform label noise which is group agnostic, a group sufficient classifier with respect to noisy labels (implying $\text{Suf}(f) = 0$) will also be group sufficient to the true labels. This implies that if we know apriori that the noise rates do not differ across groups (i.e. there is no label bias), enforcing group sufficiency with respect to the noisy labels is a valid mitigation strategy. Thus, for the rest of this work, we will deal with the case of non-group-agnostic noise.

Remark 4.2 (Sufficient Condition for Sufficiency Gap). Theorem 4.1 implies that under label bias (i.e. $T^g \neq T$) and a non-trivial classifier ($f(X) \neq 0$), we have that $\text{Suf}_g(f) > 0$, so long as the model is group-sufficient for the noisy labels ($\Delta_{\tilde{\pi}}(f(X)) = 0$) and $\tilde{\pi}(f(X))$ is not in the null space of Δ_{T^g} . Specifically, even though T and T^g are different, if the marginal $\tilde{\pi}$ does not have sufficient mass in the directions that they are different, then the sufficiency gap is unaffected since the inverses of T 's project $\tilde{\pi}$ to the same output and cancel out, resulting in $\text{Suf}_g(f) = 0$.

We define and provide similar bounds for the per-group calibration error in Appendix A. Further, we characterize the behavior of Suf_g under perturbations to the transition matrix T^g , which may happen when T^g is estimated from observational data. All proofs can be found in Appendix A.

Problem Statement. Our goal is to learn a risk score $f_\theta : \mathcal{X} \rightarrow \Delta^{K-1}$ satisfying Definition 3.1. Specifically, given a loss function $l((x, y), f) \rightarrow \mathbb{R}$. We would additionally like to achieve:

$$\begin{aligned} & \min_{\theta \in \Theta} \mathbb{E}_{(X, Y) \sim p_{X, Y}} [l((X, Y), f_\theta)] \\ & \text{s.t. } Y \perp\!\!\!\perp G \mid f_\theta(X) \end{aligned} \quad (2)$$

5 Methodology

To guide the design of our method, we derive an upper bound on the group sufficiency gap in terms of the conditional mutual information between the label and group, given the model's prediction. This result highlights that reducing the dependence between Y and G conditioned on $f(X)$ is sufficient to control the sufficiency gap:

Proposition 5.1 (Conditional Mutual Information Bound on Sufficiency Gap).

$$\text{Suf}_g(f) \leq \frac{\sqrt{2}}{K} \sqrt{\frac{I(Y; G \mid f(X))}{P(G = g)}}; \quad \text{Suf}(f) \leq \frac{\sqrt{2}}{K} \sqrt{I(Y; G \mid f(X))}.$$

This bound motivates our core objective: to minimize the sufficiency gap by minimizing the conditional mutual information $I(Y; G \mid f(X))$. However, since the true labels Y are unobserved in

practice, directly estimating this quantity is non-trivial. In the following section, we describe how we approximate this objective using the noisy labels $\tilde{Y}$. Formal proofs are provided in Appendix A.

Under Assumptions 3.1 and 3.2, we propose a differentiable regularizer that approximates and minimizes the conditional mutual information upper bound using the observed noisy labels.

$$\min_{\theta} \sum_{i=1}^n \ell_{pred}(f_{\theta}, x_i, \tilde{y}_i) + \lambda R(f_{\theta}), \quad (3)$$

where $f_{\theta} : \mathcal{X} \rightarrow \mathcal{Y}$, $\lambda > 0$, and ℓ_{pred} is a pointwise loss function, potentially selected for robustness to label noise, e.g., Peer Loss [31] or GCE [30].

CMI Regularizer We can decompose $I(Y; G | f(X))$ into the difference of two entropy terms:

$$\begin{aligned} I(Y; G | f(X)) &= H(Y | f_{\theta}(X)) - H(Y | f_{\theta}(X), G) \\ &= \mathbb{E}_{X,Y}[-\log \mathbb{P}(Y | f_{\theta}(X))] - \mathbb{E}_{X,Y,G}[-\log \mathbb{P}(Y | f_{\theta}(X), G)] \end{aligned}$$

For a mini-batch $\mathcal{B}$ of size B , each entropy term can be estimated using a Monte Carlo estimator:

$$\begin{aligned} \hat{H}(Y | f_{\theta}(X), G) &= \frac{1}{B} \sum_{i \in \mathcal{B}} (-\log \mathbb{P}(Y = y_i | f_{\theta}(X) = f_{\theta}(x_i), G = g_i)) \\ &= \frac{1}{B} \sum_{i \in \mathcal{B}} (-\log \mathbf{p}_Y(\cdot | f_{\theta}(X) = f_{\theta}(x_i), G = g_i))_{\tilde{y}_i} \end{aligned}$$

where $\mathbf{p}_Y(\cdot | f_{\theta}(X) = f_{\theta}(x_i), G = g_i) := [\mathbb{P}(Y = 1 | f_{\theta}(X) = f_{\theta}(x_i), G = g_i), \dots, \mathbb{P}(Y = K | f_{\theta}(X) = f_{\theta}(x_i), G = g_i)]$, and $(\cdot)_{\tilde{y}_i}$ denotes taking the $\tilde{y}_i$ -th element of the K -dimensional vector.

As we do not observe y_i , we can apply the backward correction [48] to this estimator, where log is applied elementwise:

$$\hat{H}(Y | f_{\theta}(X), G) = \frac{1}{B} \sum_{i \in \mathcal{B}} (-T_{g_i}^{-1} \log \mathbf{p}_Y(\cdot | f_{\theta}(X) = f_{\theta}(x_i), G = g_i))_{\tilde{y}_i} \quad (4)$$

and similarly

$$\hat{H}(Y | f_{\theta}(X)) = \frac{1}{B} \sum_{i \in \mathcal{B}} (-T_{g_i}^{-1} \log \mathbf{p}_Y(\cdot | f_{\theta}(X) = f_{\theta}(x_i)))_{\tilde{y}_i} \quad (5)$$

Note that the population entropy $H(Y | f(X))$ does not depend on G . However, as we only observe $\tilde{Y}$, our estimator must invert the noise process that generated $\tilde{Y}$. Under Assumption 3.1, the unbiased backward-correction of Y requires the group-specific matrix $T_{g_i}^{-1}$ for the i -th sample. Using the marginal matrix T^{-1} would bias the estimator when there is label bias.

Heads for Entropy Estimation To estimate $\mathbf{p}_Y$ given only noisy data, we parameterize two functions $h_{\phi} : \Delta^{K-1} \rightarrow \Delta^{K-1}$ and $h_{\psi} : \Delta^{K-1} \times \mathcal{G} \rightarrow \Delta^{K-1}$ each by a single linear layer, which we then learn by using a standard cross entropy loss with the forward correction [48]:

$$h_{\phi} = \arg \min_{\phi} \frac{1}{B} \sum_{i \in \mathcal{B}} -\log(T_{g_i}^{\top} h_{\phi}(f_{\theta}(x_i)))_{\tilde{y}_i}, \quad (6)$$

$$h_{\psi} = \arg \min_{\psi} \frac{1}{B} \sum_{i \in \mathcal{B}} -\log(T_{g_i}^{\top} h_{\psi}(f_{\theta}(x_i), g_i))_{\tilde{y}_i} \quad (7)$$

Thus, we have that

$$R(f_{\theta}) = \hat{I}(Y; G | f(X)) = \frac{1}{B} \sum_{i \in \mathcal{B}} (-T_{g_i}^{-1} \log h_{\phi}(f_{\theta}(x_i)))_{\tilde{y}_i} - (-T_{g_i}^{-1} \log h_{\psi}(f_{\theta}(x_i), g_i))_{\tilde{y}_i} \quad (8)$$

A single linear layer is each used to parameterize h_{ϕ} and h_{ψ} , as similar models are frequently learned for post-hoc calibration [74]. In practice, h_{ϕ} and h_{ψ} operate on the logits of f instead of the output

Algorithm 1: Single Step Update of CMI-REG

Input: Batch $\mathcal{B} = \{(\mathbf{x}_i, \mathbf{g}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^B$;
Per-group transition matrices $\{T^g \in [0, 1]^{K \times K}\}_{g \in \mathcal{G}}$, $\lambda > 0$, $C \geq 1$;
Learning rates $\alpha_f, \alpha_h > 0$;
Current classifier and auxiliary heads: f_θ, h_ϕ, h_ψ ;
 $\ell_{\text{pred}}(z_i, \tilde{y}_i)$;
Output: Updated f_θ, h_ϕ, h_ψ

```
1  $z_i \leftarrow f_\theta(x_i) \forall i \in \mathcal{B}$ ; // logits
2 for  $j = 1$  to  $C$  do // update heads, freeze  $\theta$ 
3    $\phi \leftarrow \phi - \alpha_h \nabla_\phi \frac{1}{B} \sum_{i \in \mathcal{B}} -\log(T_{g_i}^\top h_\phi(z_i))_{\tilde{y}_i}$ ; // Eq. (6)
4    $\psi \leftarrow \psi - \alpha_h \nabla_\psi \frac{1}{B} \sum_{i \in \mathcal{B}} -\log(T_{g_i}^\top h_\psi(z_i, g_i))_{\tilde{y}_i}$ ; // Eq. (7)
5  $\mathcal{L}_{\text{pred}} = \frac{1}{B} \sum_{i \in \mathcal{B}} \ell_{\text{pred}}(z_i, \tilde{y}_i)$ ;
6 Compute  $R_B$  using Equation (8);
7  $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda R_B$ 
8  $\theta \leftarrow \theta - \alpha_f \nabla_\theta \mathcal{L}$ ;
9 return  $f_\theta, h_\phi, h_\psi$ ;
```

probabilities. Given a batch B , we first update h_ϕ and h_ψ using gradient descent on Equations (6) and (7) for C iterations, and then update f_θ using Equation (3) once. As the backward corrected entropy estimators (Equations 4 and 5) can produce entropies outside of $[0, \log K]$, we clip estimated entropies to this range. The full algorithm can be found in Algorithm 1. When the two heads are perfectly learned, $R(f_\theta)$ is an unbiased estimator of the conditional mutual information (see Lemma A.5).

Directly Regularizing Sufficiency (SUF-REG) In parallel, we consider a direct penalty on the sufficiency gap by plugging in estimates of $\tilde{\pi}$ and $\tilde{\pi}_g$ into Theorem 4.1. Concretely, $\tilde{\pi}$ and $\tilde{\pi}_g$ are estimated by two linear heads and trained in an alternating manner similar to CMI-REG. We refer to this regularizer term as SUF-REG, and it can be combined with any ℓ_{pred} similar to CMI-REG. We emphasize that SUF-REG is also a novel contribution that has not been explored in prior work. This variant avoids mutual-information estimation but can be higher-variance. An algorithm outline for SUF-REG can be found in Appendix C.

6 Experiments

Datasets and Noise Types We evaluate our method using the 8 datasets shown in Table 1, which span tabular, image, and text modalities, and a range of number of classes. For datasets with synthetic noise, we experiment with two synthetic noise types: *Group Uniform* (where uniform noise is added only for one group), and *Group Asymmetric* (where class-specific noise is only added for one group). Additional details of how the per-group transition matrices are constructed can be found in Appendix B.2. Datasets with real noise have label noise derived from multiple annotators (in the case of `civilcomments` and `cifar10ns`) or from human annotators of noisy web data (for `clothing1m`). Further details of dataset construction can be found in Appendix B.1.

Training We train MLP, ImageNet-pretrained ResNet-18 [75], and BERT-base [76] models for tabular, image, and text modalities respectively, optimized using the Adam optimizer [77]. All datasets are divided into 60%/20%/20% training/validation/test splits. For synthetic noise, to mitigate the confounding effect of different marginal noise rates depending on group proportions, we balance the groups by subsampling the majority group during both training and testing. Hyperparameters were selected using a random hyperparameter search [78] with 20 runs. For the full hyperparameter grid for each method, see Appendix B.3. Each hyperparameter setting was repeated three times with different random seeds, which affects the dataset split, model initialization, and random noise (in the case of synthetic noise).

Table 1: Datasets used this paper. All attributes contain two groups. n : number of samples, K : number of classes. Data processing details can be found in Appendix B.1.

Dataset	Modality	Task	Attribute	n	K	Noise Type
adult [80]	Tabular	Income $\geq 50k$	Gender	32,561	2	Synthetic
lsac [81]	Tabular	Student passes the bar	Race	18,337	2	Synthetic
crime [82, 83]	Tabular	Binned rate of violent crime	Primary ethnicity	1,994	5	Synthetic
income [84]	Tabular	Binned income	Race	1,445,699	3	Synthetic
grades [85]	Tabular	Student passes exam	Gender	856	2	Real
civilcomments [86]	Text	Comment is toxic	Contains identity	448,000	2	Real
clothing1m [87]	Image	Type of clothing	Contains face	1,072,409	14	Real
cifar10ns [88, 89]	Image	Image classification	Image is grayscale	60,000	10	Real

To select the optimal hyperparameters, we experiment with three model selection strategies: (1) *Noisy Brier*: We select the model with the lowest overall Brier score on the noisy validation set. The Brier score is chosen as the selection criteria as it accounts for both performance and calibration. (2) *Theory*: We select the model with the smallest value from Theorem 4.1 on the noisy validation set, which requires knowledge of the transition matrices but not any clean labels, (3) *Clean Brier*: We select the model with the lowest overall Brier score on the *clean* validation set, which requires a validation set with clean labels. For our method (and baselines) which require transition matrices, we experiment with two settings: (1) *Known*: we use the true T^g , (2) *Estimated*: we estimate T^g using Confident Learning [54].

Baselines We consider the following off-the-shelf loss functions for ℓ_{pred} : Cross Entropy (CE), Backward [48], Forward [48], Generalized Cross Entropy (GCE) [30], α -weighted peer loss (Peer) [31], Early Learning Regularization (ELR) [58], and Determinant Mutual Information (DMI) [79]. We compare our methods (CMI-REG and SUF-REG) against each of the baseline losses. We also compare against the group-weighted peer loss for achieving equal odds from Wang et al. [35] (“GroupPeer”). Note that this method is only applicable when $K = 2$. We additionally compare against applying multicalibration (“MC”) post-processing using Algorithm 1 from Hébert-Johnson et al. [66] (specifically, using the implementation from Hansen et al. [67]), using the single attribute to define $\mathcal{C}$, and defining the calibration set to be a 40% subset of the original training set. This post-processing is done to the scores of an ERM model trained on labels produced by per-group stochastic flipping according to the row-normalized $(T_g)^{-1}$. Finally, as oracle, we train an ERM model on the clean training data.

Metrics All metrics shown are evaluated on the cleanly labeled test set. To evaluate overall performance, we use the one-vs-rest AUROC macro-averaged across classes. To evaluate fairness, we compute the sufficiency gap $\text{Suf}(f)$ as in Definition 3.2 with $p = 1$, by using a binning estimator with 10 equally sized bins.

7 Results

We present results for the four tabular datasets with synthetic group asymmetric noise, clean Brier selection, and known transition matrices in the main paper. Additional experimental results for group uniform noise, real-world noise, other selection criteria, estimated transition matrices, and additional synthetic noise levels can be found in Appendix D.

On Performance and Fairness In Table 2, we show the overall performance and sufficiency gap of SUF-REG and CMI-REG when added to seven different base losses. We find that in the large majority of cases, the addition of either regularizer results in a lower sufficiency gap, with CMI-REG exhibiting marginally better fairness. However, the confidence intervals are generally large as calibration error estimation has high variance [90]. Regardless, the model which achieves the lowest sufficiency gap is consistently regularized by CMI-REG. The base losses which function best with these regularization terms are GCE and Peer loss. Interestingly, in several cases, adding the regularizer term actually improves overall performance. Similar results for group uniform noise and real-world noise can be found in Appendix D. When the estimated transition matrix is used instead (Appendix D.5), we find that regularizers still generally improve fairness, though the performance improvements become limited.

Table 2: Clean test-set AUROC and sufficiency gap (Suf) of our regularizers when added to off-the-shelf loss functions, for tabular datasets with synthetic group asymmetric noise and known transition matrices. Models are selected using the clean Brier score. The best method within each base loss is **bolded**, and the best model overall is also underlined.

	income		adult		lsac		crime	
	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$
CE	0.731 (0.000)	0.040 (0.003)	0.899 (0.000)	0.029 (0.003)	0.581 (0.049)	0.108 (0.003)	0.741 (0.025)	0.075 (0.011)
CE + SUF-REG	0.732 (0.000)	0.035 (0.007)	0.894 (0.000)	0.029 (0.004)	0.712 (0.022)	0.089 (0.007)	0.804 (0.021)	0.057 (0.015)
CE + CMI-REG	0.736 (0.001)	0.033 (0.001)	0.903 (0.003)	0.016 (0.004)	0.708 (0.077)	0.072 (0.026)	0.748 (0.030)	0.065 (0.007)
Forward	0.801 (0.002)	0.038 (0.002)	0.887 (0.001)	0.024 (0.004)	0.436 (0.006)	0.097 (0.001)	0.788 (0.011)	0.056 (0.010)
Forward + SUF-REG	0.812 (0.002)	0.059 (0.003)	0.898 (0.000)	0.032 (0.001)	0.690 (0.011)	0.091 (0.016)	0.789 (0.006)	0.053 (0.006)
Forward + CMI-REG	0.802 (0.003)	0.019 (0.003)	0.887 (0.001)	0.023 (0.006)	0.664 (0.247)	0.075 (0.052)	0.797 (0.000)	0.052 (0.016)
Backward	0.794 (0.002)	0.042 (0.002)	0.887 (0.001)	0.021 (0.003)	0.473 (0.042)	0.100 (0.003)	0.711 (0.007)	0.080 (0.014)
Backward + SUF-REG	0.800 (0.002)	0.032 (0.003)	0.888 (0.001)	0.025 (0.004)	0.631 (0.028)	0.097 (0.017)	0.721 (0.023)	0.073 (0.002)
Backward + CMI-REG	0.797 (0.003)	0.036 (0.005)	0.888 (0.001)	0.018 (0.007)	0.452 (0.032)	0.091 (0.009)	0.739 (0.065)	0.066 (0.005)
DMI	0.590 (0.115)	0.047 (0.016)	0.887 (0.001)	0.029 (0.004)	0.476 (0.184)	0.088 (0.000)	0.406 (0.033)	0.087 (0.033)
DMI + SUF-REG	0.591 (0.111)	0.045 (0.016)	0.877 (0.003)	0.028 (0.006)	0.478 (0.167)	0.090 (0.000)	0.395 (0.033)	0.088 (0.033)
DMI + CMI-REG	0.590 (0.119)	0.048 (0.017)	0.885 (0.002)	0.020 (0.006)	0.478 (0.170)	0.090 (0.001)	0.484 (0.042)	0.107 (0.023)
ELR	0.731 (0.001)	0.040 (0.003)	0.898 (0.001)	0.030 (0.005)	0.364 (0.266)	0.061 (0.050)	0.662 (0.091)	0.092 (0.012)
ELR + SUF-REG	0.754 (0.000)	0.025 (0.003)	0.873 (0.005)	0.022 (0.002)	0.465 (0.054)	0.102 (0.001)	0.781 (0.054)	0.054 (0.011)
ELR + CMI-REG	0.738 (0.003)	0.033 (0.002)	0.860 (0.010)	0.019 (0.001)	0.319 (0.040)	0.089 (0.014)	0.664 (0.070)	0.088 (0.001)
GCE	0.730 (0.002)	0.041 (0.004)	0.894 (0.002)	0.022 (0.001)	0.462 (0.028)	0.099 (0.002)	0.739 (0.025)	0.077 (0.014)
GCE + SUF-REG	0.748 (0.001)	0.041 (0.001)	0.884 (0.003)	0.027 (0.000)	0.590 (0.082)	0.097 (0.004)	0.771 (0.047)	0.061 (0.008)
GCE + CMI-REG	0.735 (0.000)	0.036 (0.002)	0.894 (0.008)	0.022 (0.002)	0.257 (0.039)	0.050 (0.018)	0.771 (0.014)	0.062 (0.000)
Peer	0.730 (0.022)	0.030 (0.001)	0.899 (0.000)	0.029 (0.004)	0.553 (0.063)	0.110 (0.003)	0.684 (0.015)	0.069 (0.003)
Peer + SUF-REG	0.732 (0.002)	0.035 (0.005)	0.890 (0.000)	0.031 (0.003)	0.590 (0.032)	0.098 (0.001)	0.734 (0.032)	0.077 (0.000)
Peer + CMI-REG	0.686 (0.002)	0.054 (0.003)	0.901 (0.003)	0.028 (0.002)	0.551 (0.058)	0.096 (0.006)	0.630 (0.055)	0.070 (0.021)
MC	0.561 (0.019)	0.078 (0.010)	0.787 (0.004)	0.017 (0.001)	0.500 (0.000)	0.107 (0.007)	0.500 (0.000)	0.104 (0.012)
GroupPeer	—	—	0.898 (0.001)	0.035 (0.004)	0.664 (0.010)	0.095 (0.023)	—	—
Oracle	0.854 (0.001)	0.003 (0.000)	0.919 (0.002)	0.013 (0.001)	0.838 (0.011)	0.018 (0.009)	0.825 (0.013)	0.046 (0.010)

We find that naively adapting a post-hoc calibration method (multicalibration) results in generally worse performance and limited fairness gains, especially on `lsac` and `crime`. We hypothesize this stems from the limited effectiveness of the noise-aware label-flipping procedure used to construct training labels, and MC’s sample-inefficient nature [67] combined with the smaller, class-imbalanced datasets.

Pareto Fronts In Figure 2, we show the Pareto front of AUROC vs. sufficiency gap for two datasets. We find that CMI-REG and SUF-REG Pareto-dominate the base losses in almost all cases. We also find that these two regularizers consistently achieve the best AUROC-sufficiency trade-off when comparing across all base losses (Appendix D.3).

On Model Selection In Appendix D.4, we examine the impact of varying the model selection criteria. As expected, selecting using clean Brier gives the best performance and lowest sufficiency gap. Selecting using the theoretical expression (Theorem 4.1) on the noisy validation set gives similarly low sufficiency gaps, but compromises significantly on performance. Selecting using the noisy Brier gives mediocre performance with slightly worse sufficiency. Across all selection criteria, selected models with regularization terms have lower sufficiency gaps than the base model in almost all cases.

8 Discussion

On Performance and Fairness Adding either CMI-REG or SUF-REG to seven noise-robust base losses rarely harmed AUROC and, in a surprising number of cases, improved AUROC, by up to 25.4% in one instance. We attribute this to the fact that both regularizers bias optimization toward decision boundaries that are locally calibrated with respect to the true labels across groups, and we speculate such decision boundaries are also more likely to coincide with high performing solutions with respect to the clean data.

Comparing CMI-REG and SUF-REG Even though CMI-REG minimizes an upper bound while SUF-REG minimizes the direct sufficiency gap, we observe that CMI-REG performs better than SUF-REG. Across eight combinations of datasets and noise types, CMI-REG achieves the lowest sufficiency gap seven times, and SUF-REG once. We attribute this to optimization and finite sample effects. First, SUF-REG must backpropagate through an inverted T inside of a p -norm, which may be

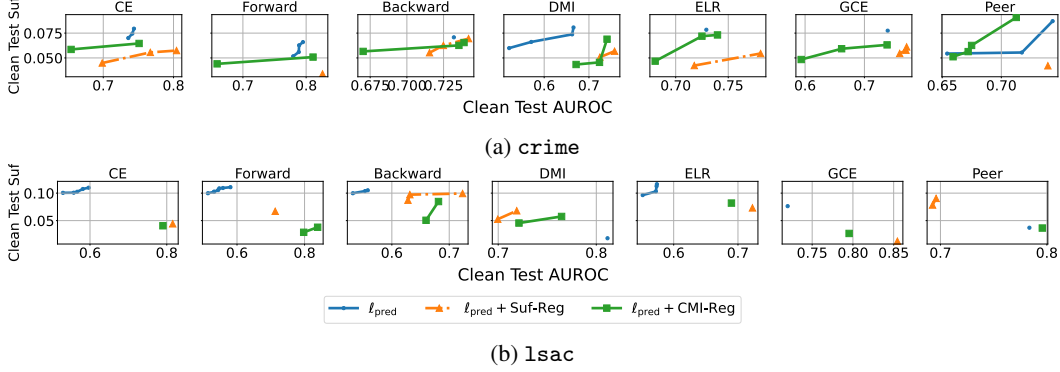


Figure 2: Pareto fronts of AUROC versus Suf of CMI-REG, SUF-REG, and baseline loss for tabular datasets with synthetic group asymmetric noise and known transition matrices. A curve towards the bottom right of the plot is desirable as it achieves the best fairness-accuracy trade-off. We observe that CMI-REG and SUF-REG Pareto-dominate the base loss for most losses. Pareto curves for remaining datasets can be found in Appendix D.3.

more numerically challenging than simply backpropagating through an inverted T multiplied by the log-likelihood, leading to abnormal gradient behavior. Second, CMI-REG consists of two convex cross-entropy terms, which may lead to a smoother optimization landscape than the piecewise-linear absolute error terms (when $p = 1$) inside SUF-REG.

9 Conclusion

In this work, we studied the problem of achieving group sufficiency under label bias. We have shown theoretically that under certain conditions, label bias is necessary and sufficient to induce a sufficiency gap. To address this, we introduced CMI-REG and SUF-REG, two drop-in regularizers that provably upper bound the sufficiency gap, and in practice, consistently reduce the sufficiency gap across eight datasets.

Limitations Our analysis rests on the class-attribute conditioned noise model (Assumption 3.1). When noise is instead instance-dependent [36, 91, 92], the regularizers SUF-REG and CMI-REG will become biased. Theoretically studying the effect of instance-dependent label bias is an area of future work. Next, our method assumes knowledge (or estimation) of the transition matrices T^g , and this estimation problem is impossible in general without further information or assumptions [93, 52]. These assumptions may be violated in real-world datasets, and applying our method using an imperfect transition matrix will result in worse performance as empirically observed. Finally, we measure fairness with respect to a single protected attribute at a time. Studying label bias between intersections of (potentially many) attributes is another area of future work.

Acknowledgments and Disclosure of Funding

This research was supported in part by an Optum Labs Research Award, a National Science Foundation (NSF) 22-586 Faculty Early Career Development Award (#2339381), a Gordon & Betty Moore Foundation award, a Google Research Scholar award, and the AI2050 Program at Schmidt Sciences.

References

- [1] Richard Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13:193–216, 2017.
- [2] Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature biomedical engineering*, 6(12):1330–1345, 2022.
- [3] Patrick Hall, Benjamin Cox, Steven Dickerson, Arjun Ravi Kannan, Raghu Kulkarni, and Nicholas Schmidt. A united states fair lending perspective on machine learning. *Frontiers in Artificial Intelligence*, 4:695301, 2021.
- [4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.
- [5] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.
- [6] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.
- [7] Sumon Biswas and Hridesh Rajan. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 981–993, 2021.
- [8] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019.
- [9] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.
- [10] Vishnu Suresh Lokhande, Aditya Kumar Akash, Sathya N Ravi, and Vikas Singh. Fairalm: Augmented lagrangian method for training fair models with little regret. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020.
- [11] 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.
- [12] Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. *Advances in neural information processing systems*, 33:361–371, 2020.
- [13] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.
- [14] André F Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. *arXiv preprint arXiv:2306.07261*, 2023.
- [15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
- [16] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.
- [17] Harvineet Singh, Matthäus Kleindessner, Volkan Cevher, Rumi Chunara, and Chris Russell. When do minimax-fair learning and empirical risk minimization coincide? In *International Conference on Machine Learning*, pages 31969–31989. PMLR, 2023.

- [18] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10421, 2022.
- [19] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on health, inference, and learning*, pages 204–233. PMLR, 2022.
- [20] Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117, 2023.
- [21] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- [22] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores, November 2016.
- [23] Haoran Zhang, Walter Gerych, and Marzyeh Ghassemi. A data-centric perspective to fair machine learning for healthcare. *Nature Reviews Methods Primers*, 4(1):86, 2024.
- [24] Vishwali Mhasawade, Alexander D’Amour, and Stephen R Pfohl. A causal perspective on label bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1282–1294, 2024.
- [25] Nancy N Maserejian, Carol L Link, Karen L Lutfey, Lisa D Marceau, and John B McKinlay. Disparities in physicians’ interpretations of heart disease symptoms by patient gender: results of a video vignette factorial experiment. *Journal of women’s health*, 18(10):1661–1667, 2009.
- [26] Joshua H Tamayo-Sarver, Susan W Hinze, Rita K Cydulka, and David W Baker. Racial and ethnic disparities in emergency department analgesic prescription. *American journal of public health*, 93(12):2067–2073, 2003.
- [27] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94 (4):991–1013, 2004.
- [28] Yiqiao Liao and Parinaz Naghizadeh. The impacts of labeling biases on fairness criteria. In *Proceedings of the 10th International Conference on Learning Representations, ICLR, Virtually*, pages 25–29, 2022.
- [29] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. How to be fair? a study of label and selection bias. *Machine Learning*, 112(12):5081–5104, 2023.
- [30] Zhilu Zhang and Mert R. Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels, November 2018.
- [31] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, pages 6226–6236. PMLR, 2020.
- [32] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.
- [33] Suyeon Kim, Dongha Lee, SeongKu Kang, Sukang Chae, Sanghwan Jang, and Hwanjo Yu. Learning discriminative dynamics with label corruption for noisy label detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22477–22487, 2024.
- [34] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1405–1413, 2021.

- [35] Jialu Wang, Yang Liu, and Caleb Levy. Fair Classification with Group-Dependent Label Noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 526–536, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445915.
- [36] Songhua Wu, Mingming Gong, Bo Han, Yang Liu, and Tongliang Liu. Fair Classification with Instance-dependent Label Noise. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 927–943. PMLR, June 2022.
- [37] Yixuan Zhang, Feng Zhou, Zhidong Li, Yang Wang, and Fang Chen. Fair Representation Learning with Unreliable Labels. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 4655–4667. PMLR, April 2023.
- [38] Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning, January 2019.
- [39] Manish Raghavan. What Should We Do when Our Ideas of Fairness Conflict? *Communications of the ACM*, 67(1):88–97, January 2024. ISSN 0001-0782, 1557-7317. doi: 10.1145/3587930.
- [40] Michele Loi and Christoph Heitz. Is calibration a fairness requirement?: An argument from the point of view of moral philosophy and decision theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2026–2034, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533245.
- [41] Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1039–1052, 2022.
- [42] Kimberly J O’malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40 (5p2):1620–1639, 2005.
- [43] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust Loss Functions under Label Noise for Deep Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), February 2017. ISSN 2374-3468. doi: 10.1609/aaai.v31i1.10894.
- [44] Xinshao Wang, Yang Hua, Elyor Kodirov, David A Clifton, and Neil M Robertson. Imae for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude’s variance matters. *arXiv preprint arXiv:1903.12141*, 2019.
- [45] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019.
- [46] Erik Engleson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.
- [47] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [48] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach, March 2017.
- [49] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018.
- [50] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, 2(3):4, 2014.
- [51] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

- [52] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are Anchor Points Really Indispensable in Label-Noise Learning? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [53] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International conference on machine learning*, pages 1062–1070. PMLR, 2019.
- [54] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, April 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12125.
- [55] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying Mislabeled Data using the Area Under the Margin Ranking. In *Advances in Neural Information Processing Systems*, volume 33, pages 17044–17056. Curran Associates, Inc., 2020.
- [56] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics, October 2020.
- [57] Suqin Yuan, Lei Feng, and Tongliang Liu. Late Stopping: Avoiding Confidently Learning from Mislabeled Examples. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16033–16042, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.01473.
- [58] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.
- [59] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels, October 2018.
- [60] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [61] Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, pages 1–11, 2024.
- [62] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621, 2021.
- [63] 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.
- [64] Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135, 2022.
- [65] Davoud Ataee Tarzanagh, Bojian Hou, Boning Tong, Qi Long, and Li Shen. Fairness-aware class imbalanced learning on multiple subgroups. In *Uncertainty in Artificial Intelligence*, pages 2123–2133. PMLR, 2023.
- [66] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.
- [67] Dutch Hansen, Siddhartha Devic, Preetum Nakkiran, and Vatsal Sharan. When is multicalibration post-processing necessary? *Advances in Neural Information Processing Systems*, 37:38383–38455, 2024.

- [68] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. Unlocking Fairness: A Trade-off Revisited. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [69] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.
- [70] Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.
- [71] Mingyuan Zhang, Jane Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *International conference on machine learning*, pages 12468–12478. PMLR, 2021.
- [72] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.
- [73] Maria Sofia Bucarelli, Lucas Cassano, Federico Siciliano, Amin Mantrach, and Fabrizio Silvestri. Leveraging inter-rater agreement for classification in the presence of noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3439–3448, 2023.
- [74] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.
- [75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [76] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [77] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [78] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1):281–305, 2012.
- [79] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_DMI: An Information-theoretic Noise-robust Loss Function, November 2019.
- [80] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: <https://doi.org/10.24432/C5XW20>.
- [81] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.
- [82] Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.
- [83] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pages 71–80. IEEE, 2013.
- [84] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.
- [85] Daphne Lenders and Toon Calders. Real-life performance of fairness interventions-introducing a new benchmarking dataset for fair ml. In *Proceedings of the 38th ACM/SIGAPP symposium on applied computing*, pages 350–357, 2023.

- [86] 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.
- [87] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.
- [88] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *International Conference on Learning Representations.*, 2022.
- [89] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.
- [90] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.
- [91] Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Instance-dependent noisy label learning via graphical modelling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2288–2298, 2023.
- [92] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*, 2020.
- [93] Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, pages 21475–21496. PMLR, 2023.
- [94] Teodora Popordanoska, Gorjan Radevski, Tinne Tuytelaars, and Matthew B Blaschko. Estimating calibration error under label shift without labels. *arXiv preprint arXiv:2312.08586*, 2023.
- [95] Mark D Reid and Robert C Williamson. Generalised pinsker inequalities. *arXiv preprint arXiv:0906.1244*, 2009.
- [96] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*, 2021.
- [97] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus 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.
- [98] Siyao Qi, Xinyu Zuo, Weijia Feng, and IG Naveen. Face recognition model based on mtcnn and facenet. In *2022 IEEE 2nd international conference on mobile networks and wireless communications (ICMNWC)*, pages 1–5. IEEE, 2022.
- [99] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *International conference on machine learning*, pages 27412–27427. PMLR, 2022.

A Proof of Lemmas and Theorems

A.1 Definitions

Definition A.1 (Calibration Error). The marginal calibration error of a multiclass classifier $f : \mathcal{X} \rightarrow \Delta^{K-1}$ with respect to Y and a p -norm ($p \geq 1$) is defined as [90, 94]:

$$\text{CE}(f) = \left(\sum_{k=1}^K \frac{1}{K} \mathbb{E}[|f(X)_k - \mathbb{E}[\mathbf{1}_{Y=k} | f(X)_k]|^p] \right)^{1/p}$$

When $K = 2$ and $p = 1$, this is equivalent to the definition of Expected Calibration Error (ECE).

Definition A.2 (Per-Group Calibration Error). Let g denote a particular group:

$$\text{CE}_g(f) = \left(\sum_{k=1}^K \frac{1}{K} \mathbb{E}[|f(X)_k - \mathbb{E}[\mathbf{1}_{Y=k} | f(X)_k, G = g]|^p | G = g] \right)^{1/p}$$

$\widetilde{\text{CE}}(f)$ and $\widetilde{\text{CE}}_g(f)$ are defined similarly with respect to $\tilde{Y}$.

A.2 $\text{Suf}_g(f)$ and Group Sufficiency

Lemma A.1. When $K = 2$, f is group sufficient $\iff \forall g \in \mathcal{G} : \text{Suf}_g(f) = 0$

Proof. ($\implies$). Write $S := f(X) \in [0, 1]$. Because $K = 2$, the prediction is completely characterised by S : $P(Y = 1 | S) = S$, $P(Y = 0 | S) = 1 - S$.

Fix any group $g \in \mathcal{G}$ and any score value $s \in [0, 1]$ with $\Pr(S = s, G = g) > 0$. Group sufficiency gives the conditional independence

$$P(Y = y | S = s, G = g) = P(Y = y | S = s), \quad y \in \{0, 1\}.$$

Hence

$$\mathbb{E}[\mathbf{1}_{Y=1} | S = s, G = g] = P(Y = 1 | S = s, G = g) = P(Y = 1 | S = s) = \mathbb{E}[\mathbf{1}_{Y=1} | S = s].$$

Substituting this equality into the definition of the sufficiency gap (Definition 3.2) yields $\text{Suf}_g(f) = 0$.

($\impliedby$) **for $K = 2$.** Assume $K = 2$ and that $\text{Suf}_g(f) = 0$ for every group g . Let s be any score value taken by $f(X)$ and let $g \in \mathcal{G}$ be fixed. Write $p_g(s) := P(Y = 1 | f(X) = s, G = g)$ and $p(s) := P(Y = 1 | f(X) = s)$. Because $\text{Suf}_g(f) = 0$, the absolute difference in Definition 3.2 is zero, so $p_g(s) = p(s)$ almost surely. For binary labels, $P(Y = 0 | \cdot) = 1 - P(Y = 1 | \cdot)$, hence

$$P(Y = 0 | f(X) = s, G = g) = 1 - p_g(s) = 1 - p(s) = P(Y = 0 | f(X) = s),$$

so the conditional distributions of Y given $f(X) = s$ are identical for every group. Therefore $P(Y = y, G = g | f(X) = s) = P(Y = y | f(X) = s)P(G = g | f(X) = s)$, i.e. $Y \perp\!\!\!\perp G | f(X)$, establishing group sufficiency. $\square$

A.3 Bound on CE

Theorem A.2 (Lower Bound on $\text{CE}_g(f)$). Let $A^g = ((T^g)^\top)^{-1}$, and $s_k^g(X) = \sum_{k=1}^K (\mathbf{1}_{k=j} - A_{kj}^g) \mathbb{E}[\mathbf{1}_{\tilde{Y}=k} | f(X)_k, G = g]$. Then, under Assumption 3.1,

$$\text{CE}_g(f) \geq \left| \widetilde{\text{CE}}_g(f) - \left(\frac{1}{K} \sum_{k=1}^K \mathbb{E}[|s_k^g(X)|^p | G = g] \right)^{1/p} \right|$$

For non-zero noise rate (e.g. $(T_{(0,1)}^g, T_{(1,0)}^g) \neq (0, 0)$), this gives that $\text{CE}_g(\tilde{f}_g^*(X)) > 0$, where $\tilde{f}_g^*$ is the Bayes optimal classifier with respect to noisy labels. Thus, learning the best model solely on the noisy labels will give us an uncalibrated model with respect to the true labels.

Proof. Fix a group $g \in \mathcal{G}$ and abbreviate all expectations and probabilities *conditionally* on $G = g$ (we omit the conditioning in the notation for clarity). For every $k \in \{1, \dots, K\}$ define

$$\tilde{\pi}_k(X) := \mathbb{E}[\mathbf{1}_{\tilde{Y}=k} \mid f(X)_k], \quad r_k(X) := \mathbb{E}[\mathbf{1}_{Y=k} \mid f(X)_k].$$

Under Assumptions 3.1 and 3.2, Bayes' rule gives

$$r_k(X) = \sum_{\ell=1}^K (A^g)_{k\ell} \tilde{\pi}_\ell(X), \quad A^g := ((T^g)^\top)^{-1}.$$

Hence

$$f(X)_k - r_k(X) = f(X)_k - \tilde{\pi}_k(X) + \underbrace{\sum_{\ell=1}^K (\mathbf{1}_{k=\ell} - A_{k\ell}^g) \tilde{\pi}_\ell(X)}_{=: s_k^g(X)}.$$

By definition of $s_k^g(X)$ we have the decomposition

$$f(X)_k - r_k(X) = a_k(X) + s_k^g(X), \quad a_k(X) := f(X)_k - \tilde{\pi}_k(X).$$

Next, introduce the L^p -norm over the product space (X, k) (conditionally on $G = g$):

$$\|v\|_p := \left(\frac{1}{K} \sum_{k=1}^K \mathbb{E}[|v_k(X)|^p] \right)^{1/p}.$$

With this notation

$$\|a\|_p = \widetilde{\text{CE}}_g(f), \quad \|s^g\|_p = \left(\frac{1}{K} \sum_{k=1}^K \mathbb{E}[|s_k^g(X)|^p] \right)^{1/p}, \quad \|a + s^g\|_p = \text{CE}_g(f).$$

The L^p space ($p \geq 1$) satisfies the Minkowski inequality:

$$\|a + s^g\|_p \geq \left| \|a\|_p - \|s^g\|_p \right|.$$

Substituting the three norms identified above yields

$$\text{CE}_g(f) \geq \left| \widetilde{\text{CE}}_g(f) - \|s^g\|_p \right| = \left| \widetilde{\text{CE}}_g(f) - \left(\frac{1}{K} \sum_{k=1}^K \mathbb{E}[|s_k^g(X)|^p] \right)^{1/p} \right|,$$

which is exactly the claimed lower bound. $\square$

A.4 Proof of Theorem 4.1

Closed-Form Decomposition of the Sufficiency Gap. Let $s = f(X) \in \Delta^{K-1}$ be the model score, $\tilde{\pi}_g(s) = \mathbb{P}(\tilde{Y} = \cdot \mid f(X) = s, G = g)$ the group-conditional noisy label posterior, and $\tilde{\pi}(s) = \mathbb{P}(\tilde{Y} = \cdot \mid f(X) = s)$ the marginal counterpart. Let $T \in [0, 1]^{K \times K}$ where $T_{i,j} = P(\tilde{Y} = j \mid Y = i) = \sum_{g \in \mathcal{G}} T_{i,j}^g P(G = g)$. Then under Assumptions 3.1 and 3.2, for any $p \geq 1$,

$$\text{Suf}_g(f) = \left(\frac{1}{K} \mathbb{E} \left[\left\| \underbrace{((T^g)^\top)^{-1}(\tilde{\pi}_g(f(X)) - \tilde{\pi}(f(X)))}_{\text{posterior shift}} + \underbrace{(((T^g)^\top)^{-1} - (T^\top)^{-1}) \tilde{\pi}(f(X))}_{\text{noise matrix mismatch}} \right\|_p^p \right] \right)^{1/p}.$$

Proof. Let $s = f(X) \in \Delta^{K-1}$ and set

$$\begin{aligned} \pi_g(s) &= P(Y = \cdot \mid s, G = g), \quad \pi(s) = P(Y = \cdot \mid s), \\ \tilde{\pi}_g(s) &= P(\tilde{Y} = \cdot \mid s, G = g), \quad \tilde{\pi}(s) = P(\tilde{Y} = \cdot \mid s). \end{aligned}$$

From Definition 3.1, we have

$$\text{Suf}_g(f)^p = \frac{1}{K} \mathbb{E}[\|\pi_g(s) - \pi(s)\|_p^p \mid G = g]. \quad (9)$$

Assumption 3.1 implies

$$\tilde{\pi}_g(s) = (T^g)^\top \pi_g(s) \quad \text{and} \quad \tilde{\pi}(s) = T^\top \pi(s) \quad (10)$$

and Assumption 3.2 gives

$$\pi_g(s) = ((T^g)^\top)^{-1} \tilde{\pi}_g(s) \quad \text{and} \quad \pi(s) = (T^\top)^{-1} \tilde{\pi}(s). \quad (11)$$

Subtracting the two identities in (11),

$$\pi_g(s) - \pi(s) = ((T^g)^\top)^{-1} (\tilde{\pi}_g(s) - \tilde{\pi}(s)) + (((T^g)^\top)^{-1} - (T^\top)^{-1}) \tilde{\pi}(s). \quad (12)$$

Substituting (12) into (9) yields

$$\text{Suf}_g(f) = \left(\frac{1}{K} \mathbb{E}[\|((T^g)^\top)^{-1} (\tilde{\pi}_g(s) - \tilde{\pi}(s)) + (((T^g)^\top)^{-1} - (T^\top)^{-1}) \tilde{\pi}(s)\|_p^p \mid G = g] \right)^{1/p}. \quad (13)$$

Equation (13) is exactly the closed-form decomposition stated in Theorem 4.1. $\square$

A.5 Errors in Estimation of T^g

Finally, estimating T from finite samples can be difficult. Proposition A.4 bounds the true sufficiency gap error with respect to the empirical sufficiency gap, mediated by estimating T, T^g from finite samples.

Lemma A.3 (Concentration of transition matrix estimates). Let $\hat{T}^g$ and $\hat{T}$ be the empirical frequency-count estimates of the true transition matrices T^g and T from n_g and n i.i.d. samples, respectively. Assume each class appears at least n_g/K (resp. n/K) times. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|T^g - \hat{T}^g\|_2 \leq \eta_g, \quad \|T - \hat{T}\|_2 \leq \eta,$$

where

$$\eta_g = \sqrt{\frac{2K^2 \log(2K/\delta)}{n_g}}, \quad \eta = \sqrt{\frac{2K^2 \log(2K/\delta)}{n}}.$$

Proof. Fix $i \in [K]$ and let $m_i \geq n_g/K$ be the number of samples with $Y = i$. By a vector-valued Hoeffding bound,

$$\mathbb{P}(\|\hat{T}_{i,\cdot}^g - T_{i,\cdot}^g\|_2 \geq \varepsilon) \leq 2 \exp\left(-\frac{m_i \varepsilon^2}{2}\right) \leq 2 \exp\left(-\frac{n_g \varepsilon^2}{2K}\right).$$

Taking a union bound over $i \in [K]$ gives

$$\mathbb{P}\left(\max_i \|\hat{T}_{i,\cdot}^g - T_{i,\cdot}^g\|_2 \geq \varepsilon\right) \leq 2K \exp\left(-\frac{n_g \varepsilon^2}{2K}\right).$$

Setting $\varepsilon = \sqrt{\frac{2K \log(2K/\delta)}{n_g}}$ ensures the RHS is at most δ . Then using $\|T^g - \hat{T}^g\|_2 \leq \|T^g - \hat{T}^g\|_F \leq \sqrt{K} \max_i \|\hat{T}_{i,\cdot}^g - T_{i,\cdot}^g\|_2$, we have

$$\|T^g - \hat{T}^g\|_2 \leq \sqrt{K} \cdot \sqrt{\frac{2K \log(2K/\delta)}{n_g}} = \sqrt{\frac{2K^2 \log(2K/\delta)}{n_g}}.$$

The same argument applies to $\|T - \hat{T}\|_2$, completing the proof. $\square$

Proposition A.4. Suppose Assumptions 3.1 and 3.2 hold. Let $\hat{T}^g$ and $\hat{T}$ be estimates of T^g and T , respectively, such that with probability at least $1 - \delta$,

$$\|T^g - \hat{T}^g\|_2 \leq \eta_g \quad \text{and} \quad \|T - \hat{T}\|_2 \leq \eta.$$

Assume the estimates are well-conditioned, i.e., $\alpha_g \eta_g < 1$ and $\alpha \eta < 1$, where $\alpha_g = \|((T^g)^\top)^{-1}\|_2$ and $\alpha = \|(T^\top)^{-1}\|_2$. Define

$$C_g = \frac{\alpha_g^2}{1 - \alpha_g \eta_g} \quad \text{and} \quad C = \frac{\alpha^2}{1 - \alpha \eta}.$$

Then for some constant M_p depending only on p (from the sufficiency gap definition), with probability at least $1 - \delta$,

$$\left| \text{Suf}_g(f) - \widehat{\text{Suf}}_g(f) \right| \leq (M_p C_g \eta_g + C \eta)_+, \quad \text{where} \quad (x)_+ = \max\{0, x\}.$$

Proof. Let $\mathcal{E} = \{\|T^g - \hat{T}^g\|_2 \leq \eta_g, \|T - \hat{T}\|_2 \leq \eta\}$. By Lemma A.3, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$, and we work on this event.

Let $A^g = ((T^g)^\top)^{-1}$ and $B^g = A^g - (T^\top)^{-1}$, and define their empirical analogues $\hat{A}^g = ((\hat{T}^g)^\top)^{-1}$ and $\hat{B}^g = \hat{A}^g - (\hat{T}^\top)^{-1}$. By the Neumann-series lemma, for any matrix A and perturbation Δ such that $\|A^{-1}\|_2 \|\Delta\|_2 < 1$, we have

$$\|(A + \Delta)^{-1} - A^{-1}\|_2 \leq \frac{\|A^{-1}\|_2^2 \|\Delta\|_2}{1 - \|A^{-1}\|_2 \|\Delta\|_2}.$$

Applying this to T^g and T , and using the definitions $\alpha_g = \|A^g\|_2$ and $\alpha = \|(T^\top)^{-1}\|_2$, we obtain

$$\|\hat{A}^g - A^g\|_2 \leq \frac{\alpha_g^2 \eta_g}{1 - \alpha_g \eta_g} = C_g \eta_g \quad \text{and} \quad \|(\hat{T}^\top)^{-1} - (T^\top)^{-1}\|_2 \leq \frac{\alpha^2 \eta}{1 - \alpha \eta} = C \eta.$$

Now define the vector-valued sufficiency for group g as

$$\mathbf{z}_g = A^g \Delta_{\tilde{\pi}} + \tilde{\pi} B^g,$$

and similarly, define the empirical estimate

$$\hat{\mathbf{z}}_g = \hat{A}^g \Delta_{\tilde{\pi}} + \tilde{\pi} \hat{B}^g.$$

By triangle inequality and sub-multiplicativity of norms,

$$\|\hat{\mathbf{z}}_g - \mathbf{z}_g\|_p \leq \|\hat{A}^g - A^g\|_2 \|\Delta_{\tilde{\pi}}\|_p + \|\tilde{\pi}\|_p \|\hat{B}^g - B^g\|_2.$$

Note that $\|\Delta_{\tilde{\pi}}\|_p \leq 2$ and $\|\tilde{\pi}\|_p \leq 1$ since both are probability distributions. Moreover, by triangle inequality on the matrices,

$$\|\hat{B}^g - B^g\|_2 = \|\hat{A}^g - A^g + (T^\top)^{-1} - (\hat{T}^\top)^{-1}\|_2 \leq \|\hat{A}^g - A^g\|_2 + \|(\hat{T}^\top)^{-1} - (T^\top)^{-1}\|_2.$$

Combining the bounds above, we obtain

$$\|\hat{\mathbf{z}}_g - \mathbf{z}_g\|_p \leq 2C_g \eta_g + (C_g \eta_g + C \eta) = (2 + 1)C_g \eta_g + C \eta = M_p C_g \eta_g + C \eta,$$

for $M_p = 3$ (or more generally any $M_p \geq 2$).

We now translate this into a bound on the sufficiency gap. By Minkowski's inequality,

$$|\text{Suf}_g(f) - \widehat{\text{Suf}}_g(f)| \leq \left(\frac{1}{K} \mathbb{E} \|\hat{\mathbf{z}}_g - \mathbf{z}_g\|_p^p \right)^{1/p} \leq (M_p C_g \eta_g + C \eta)_+.$$

This bound holds on event $\mathcal{E}$, which occurs with probability at least $1 - \delta$. This completes the proof. $\square$

A.6 Proof of Corollary 4.1.1

Here, we prove the contrapositive of the corollary: Under Assumptions 3.1 and 3.2, if $\tilde{Y} \perp\!\!\!\perp G \mid f(X)$ and $T^g = T$ for all $g \in \mathcal{G}$, then $\text{Suf}_g(f) = 0$ for all $g \in \mathcal{G}$.

Proof. Applying Theorem 4.1, $T^g = T$ implies that $\Delta_{T^g} = 0$, and $\tilde{Y} \perp\!\!\!\perp G \mid f(X)$ implies $\tilde{\pi}_g(s) = \tilde{\pi}(s)$ for all s and g , and so $\Delta_{\tilde{\pi}}(f(X)) = 0$. Thus, $\text{Suf}_g(f) = 0$. $\square$

A.7 Proof of Proposition 5.1

Proof. For clarity we give the proof for $p = 1$; the extension to $p > 1$ follows by replacing absolute values with $|\cdot|^p$ and using the generalized Pinsker inequality [95].

Fix a group $g \in \mathcal{G}$ and let $Z := f(X) \in \Delta^{K-1}$ be the (predicted) score vector. For every score value z , define the conditional label distributions

$$P_{Y|Z=z} \quad \text{and} \quad P_{Y|Z=z, G=g}.$$

Because $\sum_{k=1}^K \mathbf{1}_{Y=k} = 1$, the L_1 -form of the per-group sufficiency gap (Definition 3.2 with $p = 1$) can be written as

$$\text{Suf}_g(f) = \frac{1}{K} \mathbb{E}_{Z|G=g} [\|P_{Y|Z} - P_{Y|Z, G=g}\|_1] = \frac{2}{K} \mathbb{E}_{Z|G=g} [\text{TV}(P_{Y|Z}, P_{Y|Z, G=g})],$$

where $\text{TV}(P, Q) = \frac{1}{2} \|P - Q\|_1$ denotes total variation.

For any two discrete distributions, $\text{TV}(P, Q) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(P||Q)}$. Applying this point-wise and Jensen's inequality,

$$\text{Suf}_g(f) \leq \frac{\sqrt{2}}{K} \sqrt{\mathbb{E}_{Z|G=g} [D_{\text{KL}}(P_{Y|Z, G=g} || P_{Y|Z})]}. \quad (\text{A})$$

By the chain rule for KL divergences,

$$I(Y; G | Z) = \mathbb{E}_Z \left[\sum_{g'} P(G = g' | Z) D_{\text{KL}}(P_{Y|Z, G=g'} || P_{Y|Z}) \right]. \quad (\text{B})$$

Because $P(G = g | Z) \leq 1$ for every Z ,

$$I(Y; G | Z) \geq P(G = g) \mathbb{E}_{Z|G=g} [D_{\text{KL}}(P_{Y|Z, G=g} || P_{Y|Z})].$$

Plugging this equation into (A) gives

$$\text{Suf}_g(f) \leq \frac{\sqrt{2}}{K} \sqrt{\frac{I(Y; G | Z)}{P(G = g)}},$$

which is the first inequality.

Let $p_g := P(G = g)$, and $a_g := \mathbb{E}_{Z|G=g} [D_{\text{KL}}(P_{Y|Z, G=g} || P_{Y|Z})]$. We have, for each group,

$$\text{Suf}_g(f) \leq \frac{\sqrt{2}}{K} \sqrt{a_g}.$$

Hence

$$\text{Suf}(f) = \sum_g p_g \text{Suf}_g(f) \leq \frac{\sqrt{2}}{K} \sum_g p_g \sqrt{a_g}. \quad (\text{S1})$$

Write $u_g := \sqrt{p_g}$ and $v_g := \sqrt{p_g a_g}$. Then $u_g v_g = p_g \sqrt{a_g}$ and

$$\sum_g p_g \sqrt{a_g} = \sum_g u_g v_g.$$

By the Cauchy-Schwarz inequality,

$$\left(\sum_g u_g v_g \right)^2 \leq \left(\sum_g u_g^2 \right) \left(\sum_g v_g^2 \right) = \left(\sum_g p_g \right) \left(\sum_g p_g a_g \right) = I(Y; G | Z).$$

Taking square roots yields

$$\sum_g p_g \sqrt{a_g} \leq \sqrt{I(Y; G | Z)}. \quad (\text{S2})$$

Substituting (S2) into (S1) gives

$$\text{Suf}(f) \leq \frac{\sqrt{2}}{K} \sqrt{I(Y; G | f(X))}.$$

□

A.8 Unbiasedness and Convergence of the CMI Estimator

Lemma A.5 (Unbiasedness of $R(f_\theta)$). Under the assumption that $\forall x_i \in \mathcal{B} : h_\phi(f_\theta(x_i)) = \mathbb{P}(Y \mid f_\theta(x_i))$ and $h_\psi(f_\theta(x_i)) = \mathbb{P}(Y \mid f_\theta(x_i), g_i)$, we have that

$$\mathbb{E}[R(f_\theta)] = I(Y; G \mid f(X))$$

Proof. Let $\mathbf{h}_Y(\cdot \mid s_i) := [-\log p_Y(1 \mid s_i), \dots, -\log p_Y(K \mid s_i)]$, and $\mathbf{h}_{\tilde{Y}}(\cdot \mid s_i)$ similarly. Let $[\cdot]_{\tilde{y}_i}$ denote taking the $\tilde{y}_i$ -th element of this K -dimensional vector.

To clarify the problem, we start with the definition of entropy:

$$H(Y \mid f(X)) = \mathbb{E}_{X,Y}[-\log p_Y(Y \mid f(X))],$$

which we can estimate given a minibatch $\{(x_i, g_i, \tilde{y}_i)_{i=1}^B\}$ by applying the (unbiased) Monte Carlo estimator, defining $s_i = f(x_i)$ for convenience:

$$\hat{H}(Y \mid f(X)) = \frac{1}{B} \sum_{i=1}^B -\log p_Y(y_i \mid s_i) = \frac{1}{B} \sum_{i=1}^B [\mathbf{h}_Y(\cdot \mid s_i)]_{y_i}.$$

Since we do not observe y_i , we propose instead to use the backward-correction estimator shown in Equation (4):

$$\hat{H}'(Y \mid f(X)) = \frac{1}{B} \sum_{i=1}^B [T_{g_i}^{-1} \mathbf{h}_Y(\cdot \mid s_i)]_{\tilde{y}_i}.$$

We want to show that this estimator is unbiased, i.e. $\mathbb{E}_{X,G,Y,\tilde{Y}}[\hat{H}'(Y \mid f(X))] = H(Y \mid f(X))$.

First, observe that

$$\begin{aligned} \mathbb{E}_{\tilde{Y}|Y=y_i, G=g_i}[(T_{g_i}^{-1} \mathbf{h}_Y(\cdot \mid s_i))_{\tilde{Y}}] &= \sum_{c=1}^K \mathbb{P}(\tilde{Y} = c \mid Y = y_i, G = g_i) [T_{g_i}^{-1} \mathbf{h}_Y(\cdot \mid s_i)]_c \\ &= \sum_{c=1}^K T_{y_i, c}^{g_i} [T_{g_i}^{-1} \mathbf{h}_Y(\cdot \mid s_i)]_c \\ &= e_{y_i}^\top T_{g_i} T_{g_i}^{-1} \mathbf{h}_Y(\cdot \mid s_i) \\ &= [\mathbf{h}_Y(\cdot \mid s_i)]_{y_i} \\ &= -\log p_Y(y_i \mid s_i), \end{aligned} \tag{1}$$

which is no longer a function of g_i . Similarly,

$$\mathbb{E}_{\tilde{Y}|Y,G}[T_G^{-1} \mathbf{h}_Y(\cdot \mid s_i)] = -\log p_Y(Y \mid s_i).$$

Then,

$$\begin{aligned} \mathbb{E}_{X,G,Y,\tilde{Y}}[\hat{H}'(Y \mid f(X))] &= \mathbb{E}_{X,G,Y}[\mathbb{E}_{\tilde{Y}|X,G,Y}[\hat{H}'(Y \mid f(X))]] \\ &= \mathbb{E}_{X,G,Y}[\mathbb{E}_{\tilde{Y}|G,Y}[\hat{H}'(Y \mid f(X))]] && \text{(Assumption 3.1)} \\ &= \mathbb{E}_{X,G,Y}[-\log p_Y(Y \mid f(X))] && \text{(Substituting (1))} \\ &= \mathbb{E}_{X,Y}[-\log p_Y(Y \mid f(X))] \\ &= H(Y \mid f(X)). \end{aligned}$$

Thus, $\hat{H}'$ is an unbiased estimator of $H(Y \mid f(X))$. □

Lemma A.6 (Finite-sample convergence of the CMI estimator). Let $\hat{I}_n$ be the CMI estimator defined in Equation 8, estimated with IID samples $(x_i, g_i, \tilde{y}_i)_{i=1}^n$. Assume that $\forall x_i \in \mathcal{B} : h_\phi(f_\theta(x_i)) = \mathbb{P}(Y \mid f_\theta(x_i))$ and $h_\psi(f_\theta(x_i)) = \mathbb{P}(Y \mid f_\theta(x_i), g_i)$. Further assume that there exists $\epsilon \in (0, 1/2)$ such that for all (s, y, g) ,

$$p_{Y|S}(y \mid s) \geq \epsilon, \quad p_{G|S}(g \mid s) \geq \epsilon, \quad p_{Y,G|S}(y, g \mid s) \geq \epsilon.$$

Then, under Assumptions 3.1 and 3.2, we have that, with probability $1 - \delta$,

$$|\hat{I}_n - I(Y; G \mid f(X))| \leq 3\sqrt{2} \log(1/\epsilon) \sqrt{\frac{\log(6/\delta)}{n}} = \mathcal{O}(n^{-1/2}).$$

Further, combining with Proposition 5.1, we have that

$$\text{Suf}_g(f) \leq \sqrt{\frac{2}{K}} (\hat{I}_n + \mathcal{O}(n^{-1/2}))^{1/2}.$$

Proof. Let $S = f(X)$ and suppose the conditional probability heads are *oracle*, i.e., $\hat{p}_{Y|S} = p_{Y|S}$, $\hat{p}_{G|S} = p_{G|S}$, and $\hat{p}_{Y,G|S} = p_{Y,G|S}$ (approximation error = 0). Define the bounded random variables

$$\phi_Y := -\log p(Y \mid S), \quad \phi_G := -\log p(G \mid S), \quad \phi_{YG} := -\log p(Y, G \mid S).$$

By the ϵ -lower bound, each $\phi \in [0, \log(1/\epsilon)]$ almost surely. With $\mathbb{E}$ the population expectation and $\mathbb{E}_n$ the empirical mean over an i.i.d. sample of size n , the plug-in estimator and the population CMI can be written as

$$\hat{I}_n = \mathbb{E}_n[\phi_Y] + \mathbb{E}_n[\phi_G] - \mathbb{E}_n[\phi_{YG}], \quad I(Y; G \mid S) = \mathbb{E}[\phi_Y] + \mathbb{E}[\phi_G] - \mathbb{E}[\phi_{YG}],$$

because $H(Y \mid S) = \mathbb{E}[\phi_Y]$, $H(G \mid S) = \mathbb{E}[\phi_G]$ and $H(Y, G \mid S) = \mathbb{E}[\phi_{YG}]$.

Therefore

$$|\hat{I}_n - I(Y; G \mid S)| \leq |(\mathbb{E}_n - \mathbb{E})[\phi_Y]| + |(\mathbb{E}_n - \mathbb{E})[\phi_G]| + |(\mathbb{E}_n - \mathbb{E})[\phi_{YG}]|.$$

Each term is a deviation of the empirical mean of a bounded random variable. By Hoeffding's inequality, for any $t > 0$ and any bounded $Z \in [0, B]$,

$$\Pr(|(\mathbb{E}_n - \mathbb{E})[Z]| \geq t) \leq 2 \exp\left(-\frac{2nt^2}{B^2}\right).$$

Applying this with $B = \log(1/\epsilon)$ and a union bound over the three variables $\{\phi_Y, \phi_G, \phi_{YG}\}$, we obtain that with probability at least $1 - \delta$,

$$|\hat{I}_n - I(Y; G \mid S)| \leq 3 \log \frac{1}{\epsilon} \sqrt{\frac{1}{2n} \log \frac{6}{\delta}}.$$

□

B Experimental Details

B.1 Data Processing

income We use the ACS income dataset from Ding et al. [84], using data from all 50 states and Puerto Rico for 2018. Gender is used as the protected attribute, with women being the group with noise added. We create a 3-class classification problem from the dataset by defining class 1 to be less than \$30,000, class 2 to be between \$30,000 and \$50,000, and class 3 to be greater than \$50,000. This dataset is governed by the terms of use provided by the US Census Bureau.

adult We use the Adult dataset from the UCI Machine Learning repository [80]. Gender is used as the protected attribute, with women being the group with noise added. This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

crime We use the Communities and Crime dataset from the UCI Machine Learning repository [80]. Following Denis et al. [96], we create a 5-class classification problem by dividing the continuous crime rate into 5 equally sized bins. Following Denis et al. [96], we define the binary attribute to be whether a community is majority Black, with the non-Black group being noised. This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

lsac We use the law school admissions dataset from Wightman [81], where the goal is to predict whether someone passes the bar given their academic history. We use race as the protected attribute, and add noise to White applicants.

grades We use the dataset from Lenders and Calders [85], where the goal is to predict whether a high school student will pass an exam. We use the biased human labels from the ranking strategy as observed labels, and the eight categorical features from the Kaggle version of the dataset as features.

cifar10ns We combine the CIFAR-10N dataset from Wei et al. [88], with the CIFAR-10S dataset from Wang et al. [89]. Specifically, we use the human-derived noisy labels from CIFAR-10N, and augment the images to create an attribute by randomly turning 5% of images belonging to the second class to be grayscale, and the remaining classes to be 95% grayscale. The attribute is then whether the image is grayscale or not. Note that CIFAR-10N is released under the Creative Commons Attribution-NonCommercial 4.0 license.

civilcomments [86] We use the CivilComments dataset from the WILDS library [97], where the goal is to classify toxicity given a string of text. We choose toxicity as the binary target, and the presence of any identity as the protected attribute. Note that each sample in CivilComments has been labeled by multiple annotators. For the noisy labels, we observe labels with $Ber(p)$, where p is the proportion of labelers who said that the sample is *severely toxic*. Note that severe toxicity is a different label than the toxicity true label.

clothing1m [87] The objective to predict the type of clothing from an image scraped from the web. The noisy label is derived from keyword scraping, while the true label is derived from human labeling. Note that only the validation and test sets contain clean labels. To create an attribute for this dataset, we run an MTCNN [98] from the `facenet-pytorch` repo to detect the presence of a face in the image. We found that 64.8% of images in the dataset contain a face. The dataset is released for academic use under an unknown license.

B.2 Synthetic Noise

For both synthetic noise types, we add noise to the labels for only one of the groups.

B.2.1 Group Uniform

Given noise rate $\eta \in (0, 1)$, We noise one group with the uniform noise transition matrix:

$$T = \begin{pmatrix} 1 - \eta & \frac{\eta}{K - 1} & \cdots & \frac{\eta}{K - 1} \\ \frac{\eta}{K - 1} & 1 - \eta & \cdots & \frac{\eta}{K - 1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\eta}{K - 1} & \frac{\eta}{K - 1} & \cdots & 1 - \eta \end{pmatrix}$$

B.2.2 Group Asymmetric

Given noise rate $\eta \in (0, 1)$, for $K = 2$, we noise one group with the following transition matrix:

$$T = \begin{pmatrix} 1 & 0 \\ \eta & 1 - \eta \end{pmatrix}$$

For $K > 2$, we use the pairflip noise type from Zhu et al. [99]:

$$P = \begin{pmatrix} 1-\eta & \eta & 0 & \cdots & 0 \\ 0 & 1-\eta & \eta & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1-\eta & \eta \\ \eta & 0 & \cdots & 0 & 1-\eta \end{pmatrix}$$

For both noise types, we use $\eta = 0.75$ for all datasets, except for lsac where we use $\eta = 0.4$. We choose these noise rates as they produce ERM models with better-than-chance performance, while having significant sufficiency gaps.

B.3 Hyperparameter Grids

For all datasets and methods, we use a random hyperparameter search with 20 iterations.

We use the following hyperparameter grids which are common across all methods:

- Batch size: $2^{Rand(6,11)}$ for tabular datasets, $2^{Rand(4,6)}$ for text and image datasets
- Number of steps: $Rand(750, 2000)$ for tabular datasets, $Rand(2000, 4000)$ for text and image datasets.
- Learning rate: $10^{Unif(-5, -2)}$

For the tabular datasets, we use an MLP varying the following architecture hyperparameters:

- Depth: $Rand(2, 5)$
- Width: $Rand(32, 256)$
- Dropout: $Unif(0, 0.5)$

We use the following hyperparameter grids for specific base losses:

- GCE q : $Unif(0, 1)$
- ELR λ : $Unif(1.0, 5.0)$
- ELR β : $Unif(0.5, 0.9)$
- Peer loss α : $Unif(0, 2)$

We use the following hyperparameter grids for the MC baseline:

- α : $10^{Unif(-3,0)}$
- λ : $10^{Unif(-3,0)}$
- Number of iterations: $Rand(50, 200)$

We use the following hyperparameter grids for SUF-REG and CMI-REG:

- Learning rate for h_ϕ and h_ψ : $10^{Unif(-3, -2)}$
- C : $Rand(1, 8)$
- λ : $10^{Unif(-0.5, 1)}$
- Warm up steps: $Rand(20, 100)$

C SUF-REG Method

In order to directly minimize Theorem 4.1, we parameterize $\tilde{\pi}(s)$ by a model $h_\phi(f_\theta(x_i))$, and $\tilde{\pi}_g(s)$ by a model $h_\psi(f_\theta(x_i), g)$, with both being linear layers operating on the logits. We then alternate learning ψ and ϕ with learning θ , similar to CMI-REG. The algorithm is found in Algorithm 2.

Algorithm 2: Sufficiency–Regularised Learning under Class–Attribute Label Noise

Input: Batch $\mathcal{B} = \{(\mathbf{x}_i, \mathbf{g}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^B$;
Per-group transition matrices $\{T^g \in [0, 1]^{K \times K}\}_{g \in \mathcal{G}}$, $\lambda > 0$, $C \geq 1$;
Learning rates $\alpha_f, \alpha_h > 0$;
Current classifier and auxiliary heads: f_θ, h_ϕ, h_ψ ;
 $\ell_{\text{pred}}(z_i, \tilde{y}_i)$;
Output: Updated f_θ, h_ϕ, h_ψ

```
1  $z_i \leftarrow f_\theta(x_i) \forall i \in \mathcal{B}$ ; // logits
2 for  $j = 1$  to  $C$  do // update heads, freeze  $\theta$ 
3    $\phi \leftarrow \phi - \alpha_h \nabla_\phi \frac{1}{B} \sum_{i \in \mathcal{B}} (\log h_\phi(z_i))_{\tilde{y}_i}$ ;
4    $\psi \leftarrow \psi - \alpha_h \nabla_\psi \frac{1}{B} \sum_{i \in \mathcal{B}} (\log h_\psi(z_i, g_i))_{\tilde{y}_i}$ ;
5  $\mathcal{L}_{\text{pred}} = \frac{1}{B} \sum_{i \in \mathcal{B}} \ell_{\text{pred}}(z_i, \tilde{y}_i)$ ;
6  $R_B = \frac{1}{K} \sum_{i \in \mathcal{B}} |((T^{g_i})^\top)^{-1} h_\psi(z_i, g_i) - (T^\top)^{-1} h_\phi(z_i)|$ ;
7  $\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda R_B$ 
8  $\theta \leftarrow \theta - \alpha_f \nabla_\theta \mathcal{L}$ ;
9 return  $f_\theta, h_\phi, h_\psi$ ;
```

D Additional Experimental Results

D.1 Varying Synthetic Noise Type

In Tables 2 and 3, we present results for group-asymmetric and group-uniform synthetic noise respectively at the default noise levels. Across eight combinations of dataset and noise types, we find that our regularizer achieves the lowest sufficiency gap in all cases (with CMI-REG being the best in seven of eight cases).

Table 3: Clean test-set AUROC and sufficiency gap (Suf) of our regularizers when added to off-the-shelf loss functions, for tabular datasets with synthetic **group-uniform** noise and known transition matrices. Models are selected using the clean Brier score. The best method within each base loss is **bolded**, and the best model overall is also underlined.

	income		adult		lsac		crime	
	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$
CE	0.762 (0.002)	0.046 (0.000)	0.634 (0.014)	0.112 (0.003)	0.587 (0.049)	0.108 (0.009)	0.613 (0.019)	0.089 (0.001)
CE + SUF-REG	0.670 (0.000)	0.025 (0.003)	0.755 (0.050)	0.036 (0.010)	0.563 (0.008)	0.114 (0.006)	0.641 (0.013)	0.096 (0.006)
CE + CMI-REG	0.795 (0.007)	0.058 (0.004)	0.698 (0.137)	0.073 (0.018)	0.596 (0.005)	0.101 (0.008)	0.625 (0.020)	0.085 (0.010)
Forward	0.761 (0.003)	0.043 (0.001)	0.539 (0.082)	0.100 (0.013)	0.457 (0.093)	0.095 (0.004)	0.602 (0.040)	0.081 (0.018)
Forward + SUF-REG	0.677 (0.001)	0.024 (0.003)	0.848 (0.006)	0.039 (0.015)	0.500 (0.000)	0.107 (0.007)	0.639 (0.004)	0.086 (0.002)
Forward + CMI-REG	0.744 (0.021)	0.031 (0.005)	0.846 (0.009)	0.029 (0.010)	0.496 (0.006)	0.107 (0.007)	0.618 (0.010)	0.093 (0.004)
Backward	0.764 (0.000)	0.042 (0.002)	0.540 (0.081)	0.100 (0.013)	0.476 (0.106)	0.096 (0.011)	0.605 (0.040)	0.079 (0.008)
Backward + SUF-REG	0.766 (0.001)	0.048 (0.000)	0.696 (0.013)	0.045 (0.002)	0.459 (0.097)	0.092 (0.010)	0.609 (0.003)	0.074 (0.012)
Backward + CMI-REG	0.768 (0.005)	0.035 (0.001)	0.878 (0.023)	0.028 (0.008)	0.523 (0.113)	0.093 (0.013)	0.630 (0.049)	0.082 (0.010)
DMI	0.617 (0.032)	0.039 (0.013)	0.386 (0.114)	0.098 (0.010)	0.474 (0.183)	0.088 (0.002)	0.419 (0.007)	0.102 (0.003)
DMI + SUF-REG	0.617 (0.032)	0.038 (0.014)	0.707 (0.040)	0.106 (0.003)	0.477 (0.175)	0.088 (0.000)	0.444 (0.038)	0.097 (0.004)
DMI + CMI-REG	0.619 (0.033)	0.039 (0.010)	0.757 (0.042)	0.099 (0.008)	0.478 (0.178)	0.088 (0.000)	0.477 (0.057)	0.102 (0.017)
ELR	0.767 (0.001)	0.048 (0.002)	0.555 (0.110)	0.099 (0.017)	0.363 (0.265)	0.061 (0.051)	0.610 (0.035)	0.090 (0.005)
ELR + SUF-REG	0.764 (0.003)	0.045 (0.002)	0.685 (0.157)	0.079 (0.027)	0.340 (0.007)	0.096 (0.004)	0.616 (0.027)	0.092 (0.016)
ELR + CMI-REG	0.784 (0.021)	0.031 (0.014)	0.806 (0.067)	0.060 (0.019)	0.230 (0.026)	0.041 (0.003)	0.599 (0.042)	0.096 (0.005)
GCE	0.768 (0.002)	0.047 (0.004)	0.540 (0.085)	0.100 (0.013)	0.459 (0.043)	0.098 (0.003)	0.604 (0.039)	0.085 (0.014)
GCE + SUF-REG	0.722 (0.056)	0.019 (0.003)	0.701 (0.010)	0.061 (0.011)	0.582 (0.013)	0.103 (0.005)	0.637 (0.009)	0.081 (0.015)
GCE + CMI-REG	0.747 (0.007)	0.011 (0.001)	0.848 (0.044)	0.043 (0.020)	0.704 (0.000)	0.086 (0.005)	0.609 (0.015)	0.072 (0.001)
Peer	0.767 (0.000)	0.045 (0.004)	0.510 (0.003)	0.092 (0.004)	0.556 (0.066)	0.110 (0.008)	0.616 (0.001)	0.077 (0.023)
Peer + SUF-REG	0.745 (0.039)	0.024 (0.011)	0.509 (0.003)	0.092 (0.004)	0.640 (0.175)	0.082 (0.025)	0.670 (0.025)	0.061 (0.008)
Peer + CMI-REG	0.742 (0.013)	0.019 (0.002)	0.756 (0.021)	0.098 (0.003)	0.590 (0.052)	0.114 (0.011)	0.606 (0.031)	0.102 (0.005)
MC	0.805 (0.002)	0.030 (0.002)	0.861 (0.015)	0.045 (0.011)	0.500 (0.000)	0.107 (0.007)	0.508 (0.013)	0.101 (0.008)
GroupPeer	–	–	0.572 (0.062)	0.076 (0.004)	0.635 (0.056)	0.100 (0.006)	–	–
Oracle	0.854 (0.001)	0.003 (0.000)	0.919 (0.002)	0.013 (0.001)	0.838 (0.011)	0.018 (0.009)	0.825 (0.013)	0.046 (0.010)

D.2 Real-World Noise

In Table 4, we present results for datasets with real noise. We find that in the grades dataset, CMI-REG does show significant improvements in sufficiency gap without much decrease in performance. However, in the other 3 datasets, the sufficiency gap is negligible when training with just cross-entropy on noisy data. In addition, almost none of the other loss functions outperform simple cross-entropy with significance on overall AUROC. This indicates that real-world noise is highly non-class-conditional, and existing learning-under-label-noise methods fail as a result. Further, it indicates that there the group attribute or noisy label have been selected when processing the datasets in a way such that there is actually *minimal label bias*. Regardless, in these scenarios, we would hope that applying our method does not have a negative effect, i.e. that it leaves AUROC unchanged while maintaining a negligible sufficiency gap. We observe that this is the case for all three datasets.

Table 4: Clean test-set AUROC and sufficiency gap (Suf) for four datasets with real-world label noise. Models are selected using the clean Brier score. Within each base loss, the best value for every metric/dataset is **bolded**; the overall best non-Oracle value for each metric/dataset is additionally underlined.

	grades		civilcomments		cifar10ns		clothing1m	
	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$
CE	0.681 (0.043)	0.050 (0.028)	0.929 (0.001)	0.004 (0.001)	0.984 (0.001)	0.009 (0.003)	0.961 (0.001)	0.007 (0.001)
CE + CMI-REG	0.683 (0.049)	0.047 (0.030)	0.929 (0.001)	0.005 (0.001)	0.984 (0.000)	0.007 (0.001)	0.952 (0.002)	0.007 (0.001)
Forward	0.689 (0.048)	0.055 (0.021)	0.906 (0.008)	0.010 (0.009)	0.988 (0.000)	0.003 (0.000)	0.943 (0.001)	0.009 (0.000)
Forward + CMI-REG	0.674 (0.044)	0.054 (0.014)	0.904 (0.010)	0.011 (0.008)	0.987 (0.001)	0.002 (0.000)	0.944 (0.003)	0.009 (0.002)
Backward	0.692 (0.032)	0.078 (0.002)	0.883 (0.028)	0.013 (0.011)	0.976 (0.000)	0.006 (0.000)	0.931 (0.002)	0.007 (0.000)
Backward + CMI-REG	0.688 (0.058)	0.051 (0.007)	0.920 (0.002)	0.004 (0.001)	0.976 (0.001)	0.006 (0.000)	0.931 (0.002)	0.007 (0.000)
ELR	0.667 (0.035)	0.060 (0.012)	0.928 (0.002)	0.004 (0.001)	0.978 (0.004)	0.003 (0.001)	0.949 (0.000)	0.008 (0.001)
ELR + CMI-REG	0.666 (0.040)	0.057 (0.009)	0.928 (0.000)	0.004 (0.000)	0.976 (0.002)	0.003 (0.001)	0.942 (0.000)	0.007 (0.000)
GCE	0.683 (0.026)	0.073 (0.017)	0.915 (0.001)	0.006 (0.000)	0.986 (0.001)	0.008 (0.002)	0.942 (0.001)	0.007 (0.000)
GCE + CMI-REG	0.681 (0.053)	0.048 (0.009)	0.914 (0.002)	0.009 (0.005)	0.987 (0.000)	0.004 (0.000)	0.940 (0.001)	0.007 (0.000)
Peer	0.681 (0.035)	0.067 (0.020)	0.925 (0.004)	0.005 (0.003)	0.961 (0.002)	0.004 (0.000)	0.895 (0.000)	0.009 (0.000)
Peer + CMI-REG	0.692 (0.043)	0.053 (0.018)	0.924 (0.005)	0.007 (0.004)	0.962 (0.002)	0.003 (0.000)	0.894 (0.001)	0.009 (0.000)
Oracle	0.713 (0.075)	0.050 (0.009)	0.936 (0.001)	0.004 (0.001)	—	—	—	—

D.3 Pareto Plots

To examine the trade-off between AUROC and sufficiency gap, we plot Pareto plots for these two metrics for each dataset with synthetic noise. We present two styles of Pareto plots. First, in Figure 3, we show each base loss in a subplot, in order to examine whether our methods consistently improve the baseline for each individual loss. We note that this is a fairly strict criteria, as there may be base losses that are incompatible with our loss. Regardless, we find that this occurs in most cases. Next, in Figure 4, we aggregate all losses in a single plot, in order to show the best trade-off in the state-of-the-art before our work (i.e. base losses only) versus the two regularizers we contribute. We find that our regularizers advance the state-of-the-art by having either one Pareto-dominating the base loss in all cases.

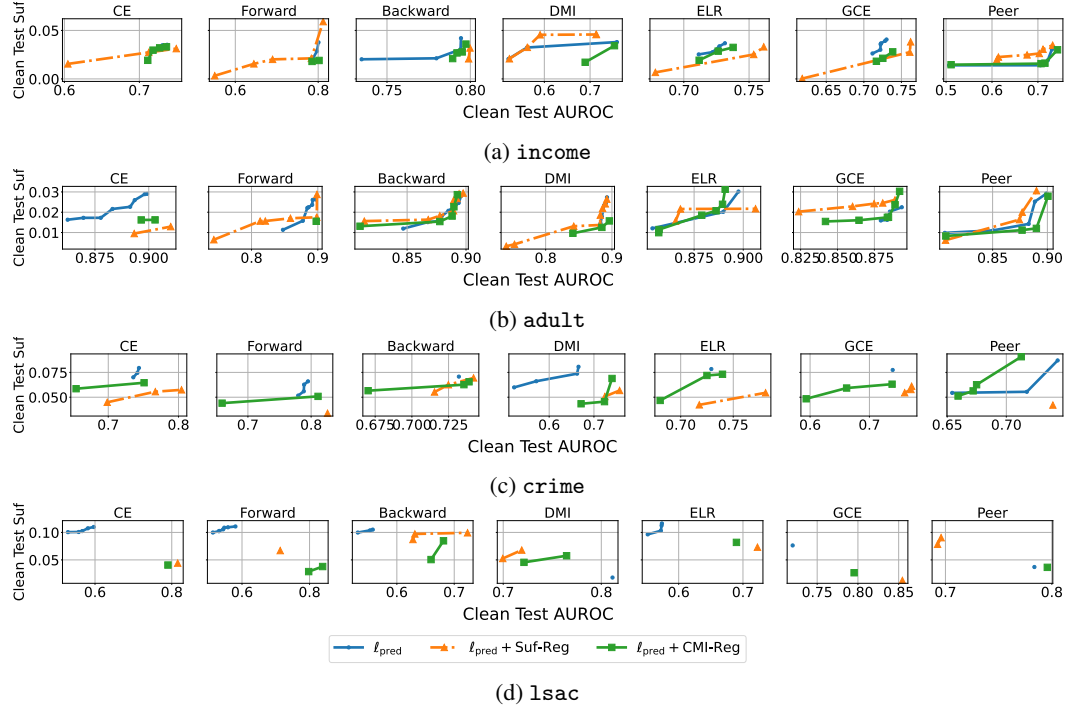


Figure 3: Pareto fronts of AUROC versus Suf of CMI-REG, SUF-REG, and baseline loss for synthetic group asymmetric noise and known transition matrices. A curve towards the bottom right of the plot is desirable as it achieves the best fairness-accuracy trade-off. We observe that CMI-REG and SUF-REG Pareto-dominate the base loss for most losses.

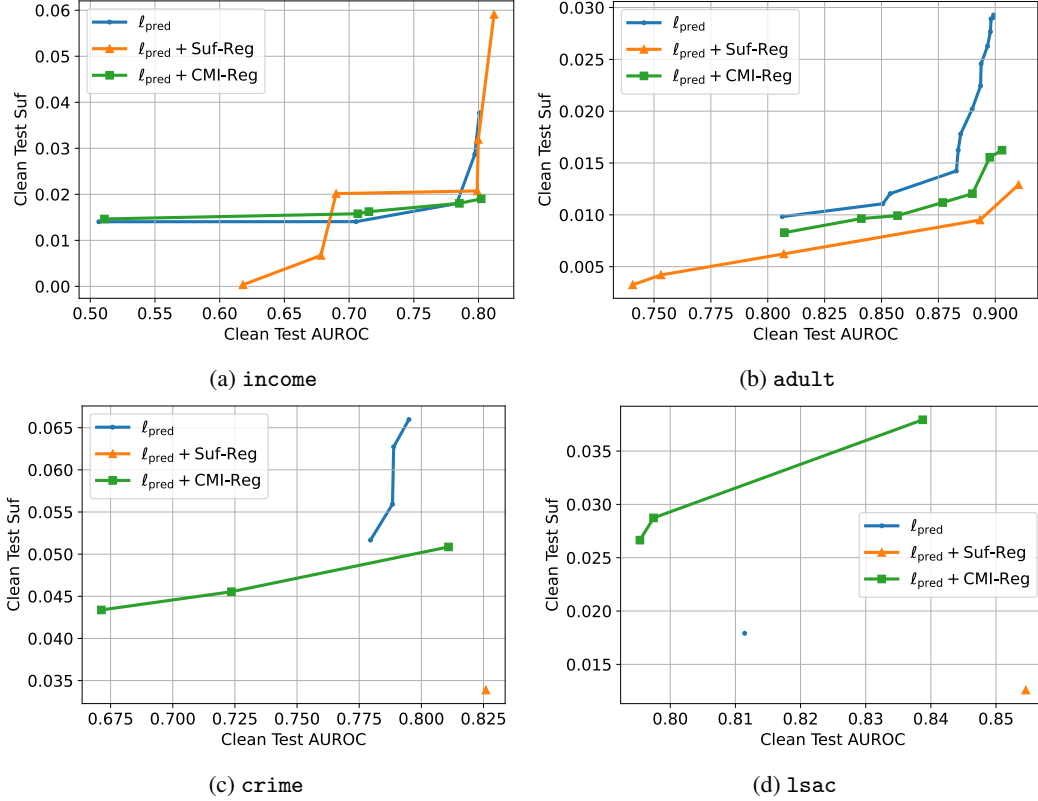


Figure 4: Pareto fronts of AUROC versus Suf of CMI-REG, Suf-REG, and baseline loss for synthetic group asymmetric noise and known transition matrices. A curve towards the bottom right of the plot is desirable as it achieves the best fairness-accuracy trade-off.

D.4 Impact of Model Selection

In Table 5, we compare the performance of different model selection criteria. Selection criteria are: (1) *Clean Brier*: We select the model with the lowest overall Brier score on the *clean* validation set, which requires a validation set with clean labels. (2) *Default*: We use default hyperparameter values reported in the papers for hyperparameters in ℓ_{pred} , and use $\lambda = 1, C = 3$ for Suf-REG and CMI-REG. (3) *Theory*: We select the model with the smallest value from Theorem 4.1 on the noisy validation set, which requires knowledge of the transition matrices but not any clean labels (4) *Noisy Brier*: We select the model with the lowest overall Brier score on the noisy validation set.

As expected, selecting using clean Brier gives the best performance and lowest sufficiency gap. Selecting using the theoretical expression (Theorem 4.1) on the noisy validation set gives similarly low sufficiency gaps, but compromises significantly on performance. Selecting using the noisy Brier gives mediocre performance with slightly worse sufficiency. Across all selection criteria, selected models with regularization terms have lower sufficiency gaps than the base model in almost all cases.

D.5 Estimated Transition Matrix at Varying Noise Types

In Tables 6 and 7, we compare our method versus the base losses when estimated transition matrices are used at two other synthetic noise strengths for group-uniform noise ($\eta = 0.6$ and $\eta = 0.8$ respectively), with clean Brier as the selection criteria. We find that with the estimated transition matrix, regularizers still generally improve fairness, though the performance improvements become limited relative to when the real transition matrix is used.

In Tables 8 and 9, we show the performance and fairness of our method versus the base losses in the hardest setting: when using the estimated transition matrix (via Confident Learning [54]), and when using the noisy Brier score as the selection metric. We find that even in this challenging setting where

Table 5: Comparison of AUROC and Suf under synthetic Group Uniform (GU) and Group Asymmetric (GA) noise. Each entry is mean (std) across seven base losses.

		AUROC				Suf			
		Clean Brier	Default	Noisy Brier	Theory	Clean Brier	Default	Noisy Brier	Theory
income	GU	Base	0.744 (0.054)	0.642 (0.025)	0.639 (0.023)	0.645 (0.096)	0.044 (0.005)	0.047 (0.013)	0.047 (0.009)
		Base + Suf-REG	0.709 (0.058)	0.666 (0.019)	0.672 (0.061)	0.568 (0.092)	0.032 (0.012)	0.025 (0.006)	0.030 (0.009)
		Base + CMI-REG	0.729 (0.060)	0.704 (0.064)	0.694 (0.036)	0.597 (0.119)	0.035 (0.011)	0.024 (0.014)	0.019 (0.008)
	GA	Base	0.730 (0.074)	0.677 (0.057)	0.681 (0.059)	0.643 (0.101)	0.040 (0.007)	0.058 (0.020)	0.058 (0.012)
		Base + Suf-REG	0.738 (0.076)	0.689 (0.062)	0.711 (0.045)	0.558 (0.069)	0.039 (0.012)	0.059 (0.025)	0.056 (0.013)
		Base + CMI-REG	0.731 (0.074)	0.686 (0.061)	0.678 (0.059)	0.606 (0.107)	0.035 (0.011)	0.054 (0.023)	0.058 (0.014)
adult	GU	Base	0.529 (0.093)	0.619 (0.032)	0.650 (0.048)	0.640 (0.128)	0.100 (0.010)	0.109 (0.013)	0.107 (0.008)
		Base + Suf-REG	0.700 (0.108)	0.656 (0.103)	0.530 (0.126)	0.667 (0.162)	0.065 (0.028)	0.089 (0.026)	0.095 (0.008)
		Base + CMI-REG	0.836 (0.063)	0.640 (0.075)	0.684 (0.073)	0.706 (0.205)	0.053 (0.032)	0.095 (0.028)	0.091 (0.012)
	GA	Base	0.893 (0.006)	0.869 (0.032)	0.893 (0.008)	0.876 (0.030)	0.026 (0.004)	0.022 (0.004)	0.026 (0.004)
		Base + Suf-REG	0.886 (0.009)	0.876 (0.033)	0.886 (0.009)	0.802 (0.171)	0.028 (0.004)	0.025 (0.006)	0.027 (0.005)
		Base + CMI-REG	0.892 (0.005)	0.867 (0.033)	0.891 (0.011)	0.818 (0.181)	0.023 (0.006)	0.021 (0.005)	0.025 (0.004)
crime	GU	Base	0.581 (0.072)	0.608 (0.062)	0.592 (0.053)	0.585 (0.072)	0.086 (0.012)	0.107 (0.016)	0.086 (0.011)
		Base + Suf-REG	0.608 (0.074)	0.609 (0.049)	0.594 (0.086)	0.572 (0.092)	0.084 (0.015)	0.107 (0.018)	0.082 (0.019)
		Base + CMI-REG	0.572 (0.054)	0.589 (0.062)	0.552 (0.091)	0.552 (0.094)	0.102 (0.016)	0.105 (0.016)	0.101 (0.009)
	GA	Base	0.676 (0.125)	0.614 (0.084)	0.673 (0.129)	0.612 (0.126)	0.077 (0.017)	0.096 (0.021)	0.081 (0.014)
		Base + Suf-REG	0.714 (0.140)	0.636 (0.124)	0.685 (0.102)	0.618 (0.119)	0.066 (0.017)	0.082 (0.027)	0.082 (0.017)
		Base + CMI-REG	0.687 (0.135)	0.624 (0.102)	0.677 (0.110)	0.634 (0.134)	0.070 (0.018)	0.092 (0.023)	0.074 (0.021)
lsac	GU	Base	0.482 (0.123)	0.538 (0.038)	0.487 (0.119)	0.442 (0.113)	0.094 (0.022)	0.109 (0.007)	0.095 (0.024)
		Base + Suf-REG	0.509 (0.120)	0.518 (0.054)	0.466 (0.099)	0.539 (0.174)	0.097 (0.014)	0.100 (0.011)	0.094 (0.022)
		Base + CMI-REG	0.559 (0.139)	0.529 (0.043)	0.425 (0.174)	0.534 (0.167)	0.094 (0.019)	0.102 (0.008)	0.079 (0.030)
	GA	Base	0.478 (0.116)	0.542 (0.034)	0.512 (0.088)	0.533 (0.042)	0.095 (0.021)	0.110 (0.006)	0.102 (0.010)
		Base + Suf-REG	0.594 (0.107)	0.536 (0.070)	0.514 (0.095)	0.523 (0.054)	0.095 (0.008)	0.108 (0.007)	0.101 (0.010)
		Base + CMI-REG	0.564 (0.165)	0.538 (0.054)	0.439 (0.164)	0.442 (0.139)	0.085 (0.028)	0.107 (0.007)	0.084 (0.033)

Table 6: Clean test-set AUROC and sufficiency gap (Suf) of our regularizers when added to off-the-shelf loss functions, for tabular datasets with synthetic **group-uniform** noise at $\eta = 0.6$ and real vs. **estimated** transition matrices. Models are selected using the *clean Brier score*. The best method within each base loss is **bolded**; the overall best non-Oracle model for each metric/dataset is additionally **underlined**.

	Real T_g				Estimated T_g			
	income		adult		income		adult	
	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$
CE	0.794 (0.000)	0.050 (0.003)	0.760 (0.001)	0.093 (0.003)	0.794 (0.000)	0.050 (0.004)	0.760 (0.001)	0.093 (0.004)
CE + Suf-Reg	0.792 (0.001)	0.054 (0.001)	0.782 (0.020)	0.020 (0.000)	0.781 (0.006)	0.054 (0.002)	0.779 (0.012)	0.043 (0.016)
CE + CMI-Reg	0.802 (0.002)	0.043 (0.003)	0.878 (0.006)	0.022 (0.001)	0.782 (0.007)	0.025 (0.003)	0.782 (0.038)	0.061 (0.005)
Forward	0.797 (0.001)	0.041 (0.001)	0.799 (0.008)	0.080 (0.002)	0.786 (0.006)	0.040 (0.002)	0.686 (0.029)	0.098 (0.008)
Forward + Suf-Reg	0.782 (0.022)	0.041 (0.001)	0.854 (0.009)	0.045 (0.018)	0.768 (0.006)	0.048 (0.003)	0.558 (0.266)	0.067 (0.001)
Forward + CMI-Reg	0.779 (0.007)	0.031 (0.003)	0.866 (0.009)	0.040 (0.015)	0.785 (0.011)	0.038 (0.005)	0.833 (0.042)	0.053 (0.037)
Backward	0.788 (0.002)	0.048 (0.001)	0.795 (0.009)	0.082 (0.003)	0.769 (0.002)	0.045 (0.001)	0.687 (0.028)	0.098 (0.008)
Backward + Suf-Reg	0.790 (0.002)	0.048 (0.000)	0.818 (0.002)	0.073 (0.004)	0.763 (0.001)	0.068 (0.000)	0.810 (0.012)	0.013 (0.001)
Backward + CMI-Reg	0.786 (0.002)	0.043 (0.001)	0.838 (0.005)	0.061 (0.004)	0.770 (0.008)	0.039 (0.000)	0.832 (0.012)	0.060 (0.005)
DMI	0.674 (0.122)	0.038 (0.001)	0.784 (0.004)	0.098 (0.002)	0.674 (0.149)	0.038 (0.001)	0.784 (0.005)	0.098 (0.002)
DMI + Suf-Reg	0.624 (0.015)	0.037 (0.012)	0.791 (0.000)	0.091 (0.005)	0.625 (0.018)	0.038 (0.011)	0.783 (0.000)	0.097 (0.005)
DMI + CMI-Reg	0.666 (0.164)	0.037 (0.003)	0.791 (0.003)	0.095 (0.003)	0.628 (0.017)	0.038 (0.009)	0.784 (0.001)	0.097 (0.005)
ELR	0.811 (0.003)	0.045 (0.002)	0.778 (0.000)	0.092 (0.005)	0.811 (0.004)	0.045 (0.002)	0.778 (0.000)	0.092 (0.006)
ELR + Suf-Reg	0.784 (0.001)	0.042 (0.001)	0.861 (0.002)	0.033 (0.009)	0.775 (0.002)	0.043 (0.001)	0.804 (0.025)	0.014 (0.006)
ELR + CMI-Reg	0.785 (0.000)	0.036 (0.001)	0.815 (0.010)	0.036 (0.000)	0.799 (0.007)	0.039 (0.001)	0.793 (0.015)	0.046 (0.014)
GCE	0.798 (0.001)	0.039 (0.000)	0.802 (0.001)	0.087 (0.002)	0.798 (0.002)	0.039 (0.000)	0.802 (0.001)	0.087 (0.003)
GCE + Suf-Reg	0.792 (0.001)	0.052 (0.000)	0.778 (0.054)	0.032 (0.008)	0.779 (0.007)	0.052 (0.001)	0.797 (0.045)	0.024 (0.001)
GCE + CMI-Reg	0.821 (0.001)	0.020 (0.004)	0.852 (0.016)	0.053 (0.006)	0.792 (0.005)	0.029 (0.005)	0.864 (0.000)	0.043 (0.002)
Peer	0.792 (0.002)	0.043 (0.001)	0.756 (0.009)	0.082 (0.008)	0.792 (0.003)	0.043 (0.001)	0.756 (0.011)	0.082 (0.010)
Peer + Suf-Reg	0.754 (0.045)	0.031 (0.002)	0.816 (0.027)	0.028 (0.007)	0.737 (0.017)	0.018 (0.008)	0.803 (0.007)	0.066 (0.008)
Peer + CMI-Reg	0.712 (0.015)	0.019 (0.000)	0.882 (0.005)	0.026 (0.000)	0.767 (0.016)	0.036 (0.011)	0.813 (0.005)	0.076 (0.008)

Table 7: Clean test-set AUROC and sufficiency gap (Suf) of our regularizers when added to off-the-shelf loss functions, for tabular datasets with synthetic **group-uniform** noise at $\eta = 0.8$ and real vs. **estimated** transition matrices. Models are selected using the *clean Brier score*. The best method within each base loss is **bolded**; the overall best non-Oracle model for each metric/dataset is additionally underlined.

	Real T_g				Estimated T_g			
	income		adult		income		adult	
	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$
CE	0.755 (0.001)	0.043 (0.000)	0.476 (0.067)	0.090 (0.011)	0.755 (0.001)	0.043 (0.000)	0.476 (0.082)	0.090 (0.014)
CE + Suf-Reg	0.681 (0.021)	0.026 (0.011)	0.759 (0.060)	0.051 (0.008)	0.678 (0.020)	0.021 (0.007)	0.596 (0.151)	0.086 (0.005)
CE + CMI-Reg	0.767 (0.008)	0.025 (0.008)	0.831 (0.063)	0.047 (0.015)	0.669 (0.037)	0.038 (0.027)	0.622 (0.112)	0.101 (0.015)
Forward	0.753 (0.002)	0.043 (0.002)	0.473 (0.077)	0.090 (0.012)	0.756 (0.000)	0.043 (0.001)	0.469 (0.091)	0.089 (0.016)
Forward + Suf-Reg	0.710 (0.060)	0.015 (0.001)	0.845 (0.001)	0.040 (0.014)	0.639 (0.068)	0.049 (0.009)	0.361 (0.033)	0.037 (0.018)
Forward + CMI-Reg	0.759 (0.012)	0.032 (0.006)	0.889 (0.003)	0.021 (0.005)	0.598 (0.047)	0.015 (0.014)	0.500 (0.005)	0.093 (0.004)
Backward	0.755 (0.000)	0.041 (0.002)	0.472 (0.077)	0.090 (0.013)	0.684 (0.028)	0.031 (0.002)	0.469 (0.091)	0.089 (0.016)
Backward + Suf-Reg	0.765 (0.003)	0.040 (0.001)	0.680 (0.008)	0.049 (0.002)	0.750 (0.014)	0.029 (0.001)	0.407 (0.026)	0.063 (0.003)
Backward + CMI-Reg	0.756 (0.001)	0.040 (0.001)	0.791 (0.030)	0.081 (0.006)	0.653 (0.019)	0.038 (0.002)	0.545 (0.092)	0.087 (0.005)
DMI	0.649 (0.071)	0.040 (0.002)	0.411 (0.088)	0.099 (0.008)	0.649 (0.087)	0.040 (0.003)	0.411 (0.108)	0.099 (0.010)
DMI + Suf-Reg	0.585 (0.011)	0.052 (0.003)	0.631 (0.012)	0.115 (0.006)	0.589 (0.009)	0.051 (0.008)	0.423 (0.095)	0.099 (0.008)
DMI + CMI-Reg	0.649 (0.107)	0.041 (0.002)	0.411 (0.103)	0.099 (0.010)	0.567 (0.015)	0.047 (0.006)	0.443 (0.059)	0.096 (0.006)
ELR	0.755 (0.001)	0.043 (0.001)	0.476 (0.102)	0.088 (0.018)	0.755 (0.001)	0.043 (0.001)	0.476 (0.124)	0.088 (0.023)
ELR + Suf-Reg	0.757 (0.002)	0.045 (0.001)	0.658 (0.182)	0.079 (0.025)	0.662 (0.040)	0.043 (0.029)	0.597 (0.178)	0.056 (0.031)
ELR + CMI-Reg	0.759 (0.004)	0.041 (0.001)	0.726 (0.003)	0.096 (0.006)	0.671 (0.048)	0.046 (0.028)	0.545 (0.106)	0.084 (0.008)
GCE	0.753 (0.001)	0.043 (0.001)	0.472 (0.078)	0.090 (0.013)	0.753 (0.001)	0.043 (0.001)	0.472 (0.096)	0.090 (0.016)
GCE + Suf-Reg	0.756 (0.029)	0.038 (0.006)	0.759 (0.006)	0.061 (0.013)	0.640 (0.010)	0.048 (0.006)	0.394 (0.084)	0.050 (0.040)
GCE + CMI-Reg	0.759 (0.007)	0.037 (0.004)	0.892 (0.000)	0.021 (0.002)	0.661 (0.015)	0.066 (0.002)	0.569 (0.097)	0.082 (0.017)
Peer	0.755 (0.004)	0.042 (0.002)	0.720 (0.004)	0.108 (0.002)	0.755 (0.005)	0.042 (0.003)	0.500 (0.000)	0.093 (0.004)
Peer + Suf-Reg	0.695 (0.036)	0.023 (0.004)	0.771 (0.043)	0.035 (0.010)	0.642 (0.068)	0.017 (0.011)	0.500 (0.000)	0.093 (0.004)
Peer + CMI-Reg	0.699 (0.003)	0.017 (0.001)	0.858 (0.005)	0.041 (0.008)	0.640 (0.002)	0.043 (0.005)	0.500 (0.000)	0.093 (0.004)

no clearly labeled data is available, our regularizers consistently achieve lower sufficiency gaps than the base losses, and the selected models with the lowest sufficiency gaps across losses correspond our regularizers in six of eight combinations of datasets and noise types.

E Broader Impacts

In this work, we explored unfairness induced by label bias, and introduced regularizers so that models stay calibrated across groups with respect to the true labels. Though we evaluate on real-world data, we do not advocate for the blind deployment of ML models in any real-world safety critical setting (such as health, credit, or hiring). There are a myriad of other factors that should be taken into account before their deployment (e.g. privacy, regulation, interpretability, and socio-technical considerations). Misuse of such models could lead to real harm.

Table 8: Clean test-set AUROC and sufficiency gap (Suf) of our regularizers when added to off-the-shelf loss functions, for tabular datasets with synthetic **group-asymmetric** noise and **estimated** transition matrices. Models are selected using the *noisy Brier score*. The best method within each base loss is **bolded**; the best non-Oracle model overall for each metric/dataset is additionally underlined.

	income		adult		lsac		crime	
	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$
CE	0.682 (0.001)	0.065 (0.002)	0.889 (0.001)	0.026 (0.004)	0.561 (0.047)	0.109 (0.006)	0.740 (0.023)	0.072 (0.001)
CE + SUF-REG	0.679 (0.002)	0.076 (0.006)	0.884 (0.003)	0.028 (0.001)	0.477 (0.109)	0.092 (0.007)	0.722 (0.042)	0.074 (0.021)
CE + CMI-REG	0.663 (0.003)	0.079 (0.003)	0.892 (0.000)	0.023 (0.002)	0.479 (0.021)	0.102 (0.010)	0.731 (0.013)	0.078 (0.010)
Forward	0.633 (0.003)	0.109 (0.004)	0.886 (0.001)	0.029 (0.000)	0.446 (0.065)	0.095 (0.001)	0.689 (0.057)	0.057 (0.003)
Forward + SUF-REG	0.682 (0.030)	0.043 (0.018)	0.884 (0.003)	0.028 (0.001)	0.283 (0.059)	0.060 (0.028)	0.596 (0.025)	0.072 (0.009)
Forward + CMI-REG	0.609 (0.009)	0.103 (0.007)	0.890 (0.001)	0.026 (0.002)	0.554 (0.169)	0.092 (0.019)	0.637 (0.114)	0.084 (0.002)
Backward	0.607 (0.004)	0.092 (0.006)	0.886 (0.002)	0.029 (0.001)	0.445 (0.061)	0.094 (0.002)	0.564 (0.039)	0.098 (0.013)
Backward + SUF-REG	0.612 (0.000)	0.091 (0.003)	0.886 (0.002)	0.027 (0.001)	0.506 (0.098)	0.090 (0.001)	0.526 (0.001)	0.111 (0.022)
Backward + CMI-REG	0.601 (0.005)	0.106 (0.000)	0.887 (0.002)	0.027 (0.000)	0.476 (0.093)	0.094 (0.001)	0.563 (0.073)	0.099 (0.009)
DMI	0.590 (0.094)	0.047 (0.013)	0.888 (0.001)	0.024 (0.004)	0.477 (0.138)	0.090 (0.001)	0.475 (0.027)	0.108 (0.020)
DMI + SUF-REG	0.590 (0.113)	0.046 (0.016)	0.856 (0.003)	0.017 (0.002)	0.503 (0.130)	0.090 (0.007)	0.459 (0.029)	0.101 (0.037)
DMI + CMI-REG	0.587 (0.114)	0.046 (0.017)	0.890 (0.003)	0.024 (0.006)	0.480 (0.170)	0.089 (0.001)	0.453 (0.035)	0.097 (0.034)
ELR	0.656 (0.001)	0.077 (0.003)	0.887 (0.001)	0.026 (0.001)	0.520 (0.043)	0.101 (0.011)	0.656 (0.072)	0.088 (0.002)
ELR + SUF-REG	0.677 (0.001)	0.066 (0.002)	0.886 (0.002)	0.026 (0.000)	0.349 (0.065)	0.088 (0.011)	0.505 (0.051)	0.092 (0.038)
ELR + CMI-REG	0.650 (0.003)	0.080 (0.003)	0.883 (0.001)	0.020 (0.002)	0.375 (0.051)	0.095 (0.006)	0.650 (0.094)	0.098 (0.007)
GCE	0.675 (0.005)	0.070 (0.006)	0.884 (0.001)	0.029 (0.001)	0.491 (0.051)	0.100 (0.001)	0.743 (0.025)	0.083 (0.020)
GCE + SUF-REG	0.677 (0.003)	0.078 (0.007)	0.884 (0.003)	0.027 (0.001)	0.467 (0.118)	0.089 (0.010)	0.719 (0.049)	0.075 (0.027)
GCE + CMI-REG	0.663 (0.008)	0.082 (0.007)	0.891 (0.001)	0.024 (0.004)	0.623 (0.066)	0.098 (0.000)	0.701 (0.018)	0.084 (0.004)
Peer	0.730 (0.018)	0.030 (0.000)	0.888 (0.000)	0.025 (0.001)	0.566 (0.033)	0.112 (0.005)	0.696 (0.041)	0.080 (0.001)
Peer + SUF-REG	0.644 (0.046)	0.060 (0.037)	0.883 (0.001)	0.018 (0.007)	0.280 (0.030)	0.053 (0.007)	0.498 (0.064)	0.085 (0.026)
Peer + CMI-REG	0.589 (0.002)	0.054 (0.006)	0.892 (0.001)	0.021 (0.001)	0.481 (0.097)	0.101 (0.000)	0.663 (0.061)	0.100 (0.011)
GroupPeer	—	—	0.896 (0.002)	0.023 (0.003)	0.659 (0.001)	0.094 (0.018)	—	—
Oracle	0.854 (0.001)	0.003 (0.000)	0.919 (0.002)	0.013 (0.001)	0.838 (0.011)	0.018 (0.009)	0.825 (0.013)	0.046 (0.010)

Table 9: Clean test-set AUROC and sufficiency gap (Suf) of our regularizers when added to off-the-shelf loss functions, for tabular datasets with synthetic **group-uniform** noise and **estimated** transition matrices. Models are selected using the *noisy Brier score*. The best method within each base loss is **bolded**; the overall best non-Oracle model for each metric/dataset is additionally underlined.

	income		adult		lsac		crime	
	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$	AUROC $\uparrow$	Suf $\downarrow$
CE	0.630 (0.004)	0.053 (0.006)	0.634 (0.019)	0.108 (0.001)	0.565 (0.050)	0.111 (0.006)	0.609 (0.015)	0.090 (0.003)
CE + SUF-REG	0.651 (0.027)	0.042 (0.022)	0.573 (0.026)	0.109 (0.029)	0.464 (0.083)	0.093 (0.008)	0.647 (0.005)	0.090 (0.020)
CE + CMI-REG	0.681 (0.003)	0.034 (0.008)	0.593 (0.000)	0.112 (0.019)	0.494 (0.036)	0.101 (0.012)	0.650 (0.057)	0.092 (0.005)
Forward	0.764 (0.003)	0.047 (0.001)	0.656 (0.009)	0.107 (0.004)	0.456 (0.080)	0.095 (0.000)	0.609 (0.014)	0.103 (0.009)
Forward + SUF-REG	0.692 (0.005)	0.027 (0.007)	0.552 (0.352)	0.013 (0.004)	0.447 (0.131)	0.089 (0.020)	0.435 (0.037)	0.102 (0.003)
Forward + CMI-REG	0.649 (0.019)	0.034 (0.007)	0.592 (0.010)	0.095 (0.011)	0.453 (0.140)	0.087 (0.014)	0.655 (0.068)	0.084 (0.013)
Backward	0.747 (0.003)	0.048 (0.001)	0.669 (0.008)	0.106 (0.004)	0.454 (0.074)	0.096 (0.001)	0.531 (0.084)	0.094 (0.009)
Backward + SUF-REG	0.761 (0.006)	0.047 (0.002)	0.705 (0.014)	0.054 (0.013)	0.471 (0.140)	0.085 (0.018)	0.460 (0.003)	0.091 (0.009)
Backward + CMI-REG	0.659 (0.008)	0.041 (0.007)	0.530 (0.002)	0.096 (0.004)	0.485 (0.100)	0.095 (0.002)	0.538 (0.094)	0.089 (0.001)
DMI	0.660 (0.062)	0.034 (0.009)	0.641 (0.065)	0.077 (0.020)	0.478 (0.144)	0.088 (0.000)	0.489 (0.043)	0.109 (0.016)
DMI + SUF-REG	0.643 (0.090)	0.051 (0.003)	0.573 (0.026)	0.078 (0.020)	0.472 (0.171)	0.089 (0.002)	0.634 (0.087)	0.081 (0.023)
DMI + CMI-REG	0.642 (0.042)	0.025 (0.007)	0.604 (0.020)	0.107 (0.004)	0.482 (0.177)	0.088 (0.001)	0.488 (0.054)	0.110 (0.017)
ELR	0.630 (0.001)	0.056 (0.002)	0.632 (0.010)	0.112 (0.005)	0.523 (0.042)	0.101 (0.011)	0.608 (0.027)	0.088 (0.005)
ELR + SUF-REG	0.642 (0.012)	0.043 (0.013)	0.290 (0.026)	0.029 (0.015)	0.330 (0.077)	0.083 (0.013)	0.459 (0.049)	0.108 (0.007)
ELR + CMI-REG	0.694 (0.001)	0.029 (0.008)	0.405 (0.002)	0.066 (0.002)	0.375 (0.052)	0.094 (0.007)	0.645 (0.065)	0.096 (0.004)
GCE	0.646 (0.004)	0.052 (0.005)	0.629 (0.004)	0.112 (0.007)	0.488 (0.062)	0.097 (0.000)	0.606 (0.016)	0.090 (0.007)
GCE + SUF-REG	0.683 (0.046)	0.026 (0.020)	0.541 (0.028)	0.124 (0.019)	0.459 (0.086)	0.092 (0.008)	0.647 (0.006)	0.086 (0.001)
GCE + CMI-REG	0.637 (0.026)	0.043 (0.026)	0.607 (0.000)	0.101 (0.009)	0.614 (0.029)	0.096 (0.002)	0.651 (0.068)	0.093 (0.004)
Peer	0.767 (0.000)	0.045 (0.003)	0.640 (0.020)	0.107 (0.001)	0.573 (0.040)	0.110 (0.007)	0.618 (0.028)	0.085 (0.012)
Peer + SUF-REG	0.581 (0.028)	0.042 (0.014)	0.559 (0.282)	0.034 (0.025)	0.389 (0.100)	0.095 (0.026)	0.501 (0.074)	0.108 (0.010)
Peer + CMI-REG	0.654 (0.007)	0.042 (0.009)	0.513 (0.011)	0.092 (0.003)	0.487 (0.097)	0.099 (0.003)	0.657 (0.064)	0.096 (0.008)
GroupPeer	—	—	0.287 (0.095)	0.088 (0.003)	0.646 (0.007)	0.098 (0.024)	—	—
Oracle	0.854 (0.001)	0.003 (0.000)	0.919 (0.002)	0.013 (0.001)	0.838 (0.011)	0.018 (0.009)	0.825 (0.013)	0.046 (0.010)

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We have stated our contributions in the introduction, and we these contributions have been adequately justified by the theoretical analyses and experimental results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations in the Discussion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All assumptions are stated prior to the theoretical results, and all proofs are provided in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our experimental setup in the Experiments section, and provide all additional details needed to reproduce our experimental results in Appendix B. Our code is available at https://github.com/MLforHealth/sufficiency_label_bias.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided the code, including instructions to download the data, at https://github.com/MLforHealth/sufficiency_label_bias.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We describe our experimental setup in the Experiments section, and provide all additional details needed to reproduce our experimental results in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars as one standard deviation over 3 random seeds, which affects the dataset split, model initialization, and random noise for synthetic noise. See the Experiments section for details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See Appendix B.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research follows the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Appendix E.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks as we use publicly available datasets, and only train small classification models which do not have the potential for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper which produced the datasets have been cited in the main paper, and licenses and copyright information are provided in Appendix B.1.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our code is released on Github under the MIT license. There are no assets introduced otherwise.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not require IRB approval as we do not work with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our proposed method does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.