Mitigating Source Bias for Fairer Weak Supervision

Anonymous Author(s)
Affiliation
Address
email

Abstract

Weak supervision enables efficient development of training sets by reducing the need for ground truth labels. However, the techniques that make weak supervision attractive—such as integrating any source of signal to estimate unknown labels—also entail the danger that the produced pseudolabels are highly biased. Surprisingly, given everyday use and the potential for increased bias, weak supervision has not been studied from the point of view of fairness. We begin such a study, starting with the observation that even when a fair model can be built from a dataset with access to ground-truth labels, the corresponding dataset labeled via weak supervision can be arbitrarily unfair. To address this, we propose and empirically validate a model for source unfairness in weak supervision, then introduce a simple counterfactual fairness-based technique that can mitigate these biases. Theoretically, we show that it is possible for our approach to simultaneously improve both accuracy and fairness—in contrast to standard fairness approaches that suffer from tradeoffs. Empirically, we show that our technique improves accuracy on weak supervision baselines by as much as 32% while reducing demographic parity gap by 82.5%. A simple extension of our method aimed at maximizing performance produces state-of-the-art performance in five out of ten datasets in the WRENCH benchmark.

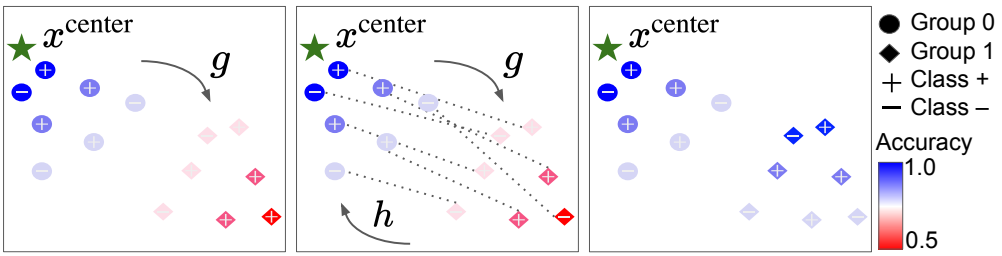
1 Introduction

Weak supervision (WS) is a powerful set of techniques aimed at overcoming the labeled data bottleneck [RSW$^+$16, FCS$^+$20, SLV$^+$22]. Instead of manually annotating points, users assemble noisy label estimates obtained from multiple sources, model them by learning source accuracies, and combine them into a high-quality pseudolabel to be used for downstream training. All of this is done without any ground truth labels. Simple, flexible, yet powerful, weak supervision is now a standard component in machine learning workflows in industry, academia, and beyond [BRL$^+$19]. Most excitingly, WS has been used to build models deployed to billions of devices.

Real-life deployment of models, however, raises crucial questions of fairness and bias. Such questions are tackled in the burgeoning field of fair machine learning [DHP$^+$12, HPS16]. However, weak supervision **has not been studied from this point of view**. This is not a minor oversight. The properties that make weak supervision effective (i.e., omnivorously ingesting any source of signal for labels) are precisely those that make it likely to suffer from harmful biases. This motivates the need to understand and mitigate the potentially disparate outcomes that result from using weak supervision.

The starting point for this work is a simple result. Even when perfectly fair classifiers are possible when trained on ground-truth labels, weak supervision-based techniques can nevertheless produce arbitrarily unfair outcomes. Because of this, simply applying existing techniques for producing fair outcomes to the datasets produced via WS is insufficient—delivering highly suboptimal datasets. Instead, a new approach, specific to weak supervision, must be developed.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.



(a) Data (b) Transformations (c) Recovered labels

Figure 1: Intuitive illustration for our setting and approach. (a): circles and diamonds are datapoints from group 0 and 1, respectively. Labeling function vote accuracy is colored-coded, with blue being perfect (1.0) and red random (0.5). Note that accuracy degrades as data points get farther from center x^{center} (star). (b) We can think of group 1 as having been moved far from the center via a transformation g, producing lower-quality estimates and violating fairness downstream. (c) Our technique uses counterfactual fairness to undo this transformation, obtaining higher quality estimates. and improved fairness.

We introduce a simple technique for improving the fairness properties of weak supervision-based models. Intuitively, a major cause of bias in WS is that particular sources are targeted at certain groups, and so produce far more accurate label estimates for these groups—and far more noise for others. We counterfactually ask what outgroup points would most be like if they were part of the 'privileged' group (with respect to each source), enabling us to borrow from the more powerful signal in the sources applied to this group. Thus the problem is reduced to finding a transformation between groups that satisfies this counterfactual. Most excitingly, while in standard fairness approaches there is a typical tradeoff between fairness and accuracy, with our approach, both the fairness and performance of WS-based techniques can be (sometimes dramatically) improved.

Theoretically, in certain settings, we provide finite-sample rates for recovering the counterfactual transformation. Empirically, we propose several ways to craft an efficiently-computed transformation building on optimal transport and some simple variations. We validate our claims on a diverse set of experiments. These include standard real-world fairness datasets, where we observe that our method can improve both fairness and accuracy by as much as 82.5% and 32.5%, respectively, versus weak supervision baselines. Our method can also be combined with other fair ML methods developed for fully supervised settings, further improving fairness. Finally, our approach has implications for WS beyond bias: we combined it with slice discovery techniques [EVS+22] to improve latent underperforming groups. This enabled us to **improve on state-of-the-art on the weak supervision benchmark** WRENCH [ZYL+21].

The contributions of this work include,

- The first study of fairness in weak supervision,
- A new empirically-validated model for weak supervision that captures labeling function bias,
- A simple counterfactual fairness-based correction to mitigate such bias, compatible with any existing weak supervision pipeline, as well as with downstream fairness techniques,
- Theoretical results showing that (1) even with a fair dataset, a weakly-supervised counterpart can be arbitrarily biased and (2) a finite-sample recovery result for the proposed algorithm,
- Experiments validating our claims, including on weakly-supervised forms of popular fairness evaluation datasets, showing gains in fairness metrics—and often simultaneously improvements in accuracy.

2 Background and Related Work

We present some high-level background on weak supervision and fairness in machine learning. Afterward, we provide setup and present the problem statement.

Weak Supervision Weak supervision frameworks build labeled training sets *with no access to ground truth labels*. Instead, they exploit multiple sources that provide noisy estimates of the label. These sources include heuristic rules, knowledge base lookups, pretrained models, and more [KOS11, MBSJ09, GM14, DZS+17, RBE+18]. Because these sources may have different—and

unknown—accuracies and dependencies, their outputs must be modeled in order to produce a combination that can be used as a high-quality pseudolabel.

Concretely, there is a dataset $\{(x_1,y_1),...,(x_n,y_n)\}$ with unobserved true label $y_i \in \{-1,+1\}$. We can access the outputs of m sources (labeling functions) $\lambda^1,\lambda^2,...,\lambda^m \in \{-1,+1\}$ outputting noisy estimates of the labels. These outputs are modeled via a generative model called the *label model*, $p_\theta(\lambda^1,...,\lambda^m,y)$. The goal is to estimate the parameters θ of this model, without accessing the latent y, and to produce a pseudolabel estimate $p_{\hat{\theta}}(y|\lambda^1,...,\lambda^m)$. For more background, see [ZHY+22].

Machine Learning and Fairness Fairness in machine learning is a large and active field that seeks to understand and mitigate biases. We briefly introduce high-level notions that will be useful in the weak supervision setting, such as the notion of fairness metrics. Two popular choices are demographic parity [DHP+12] and equal opportunity [HPS16]. Demographic parity is based on the notion that individuals of different groups should have equal treatment, i.e., if A is the group attribute, $P(\hat{Y}=1|A=1)=P(\hat{Y}=1|A=0)$. The equal opportunity principle requires that predictive error should be equal across groups, i.e., $P(\hat{Y}=1|Y=1,A=1)=P(\hat{Y}=1|Y=1,A=0)$.

A large number of works study, measure, and seek to improve fairness in different machine learning settings based on these metrics. Typically, the assumption is that the underlying dataset differs within groups in such a way that a trained model will violate, for example, the equal opportunity principle. In contrast, in this work, we focus on additional violations of fairness that are induced by weak supervision pipelines—which can create substantial unfairness even when the true dataset is perfectly fair. In the same spirit is [WH22], which considers fairness in positive-and-unlabeled (PU) settings, where true labels are available, but only for one class, while other points are unlabeled. Another related line of research is fairness under noisy labels [WLL21, KL22, WZN+23, ZZL+23]. These works consider the noise rate of labels in fair learning, enhancing the robustness of fair learning methods. *A crucial difference between such works and ours: in weak supervision, we have multiple sources of noisy labels—and we can exploit these to directly improve dataset fairness.*

Counterfactual Fairness Most closely related to the notion of fairness we use in this work is *counterfactual fairness*. [KLRS17] introduced such a counterfactual fairness notion, which implies that changing the sensitive attribute A, while keeping other variables causally not dependent on A, should not affect the outcome. While this notion presumes the causal structure behind the ML task, it is related to our work in the sense that our proposed method tries to remove the causal effect by A with a particular transformations. A more recent line of work has proposed bypassing the need for causal structures and directly tackling counterfactual fairness through optimal transport [GDBFL19, BYF20, SRT+20, SMBN21, BDB22]. The idea is to detect or mitigate unfairness by mapping one group to another group via such techniques. In this paper, we build on these tools to help improve fairness while avoiding the accuracy-fairness tradeoff common to most settings.

3 Mitigating Labeling Function-Induced Unfairness

We are ready to explain our approach to mitigating unfairness in weak supervision sources. First, we provide a flexible model that captures such behavior, along with empirical evidence supporting it. Next, we propose a simple solution to correct unfair source behavior via optimal transport.

Modeling Group Bias in Weak Supervision Weak supervision models the accuracies and correlations in labeling functions. The standard model, used in [RHD+19a, FCS+20] and others is $P(\lambda^1,...,\lambda^m,y)=\frac{1}{Z}\exp(\theta_y y+\sum_{j=1}^m \theta_j \lambda^j y)$, with $\theta_j \geq 0$. We leave out the correlations for simplicity; all of our discussion below holds when considering correlations as well. Here, Z is the normalizing partition function. The θ are *canonical parameters* for the model. θ_y sets the class balance. The θ_i's capture how accurate LF i is: if $\theta_i=0$, the LF produces random guesses. If θ_i is relatively large, the LF is highly accurate. A weakness of this model is that it *ignores the feature vector* x. It implies that LFs are uniformly accurate over the feature space—a highly unrealistic assumption. A more general model was presented in [CFA+22], where there is a model for each feature vector x, i.e.,

$$P_x(\lambda^1,...,\lambda^m,y)=\frac{1}{Z}\exp(\theta_y y+\sum_{j=1}^m \theta_{j,x}\lambda^j(x)y). \tag{1}$$

However, as we see only one sample for each x, it is impossible to recover the parameters θ_x. Instead, the authors assume a notion of *smoothness*. This means that the $\theta_{j,x}$'s do not vary in small neighborhoods, so that the feature space can be partitioned and a single model learned per part. Thus *model* (1) *from [CFA$^+$22] is more general, but still requires a strong smoothness assumption*. It also does not encode any notion of bias. Instead, we propose a model that encodes both smoothness and bias.

Concretely, let us assume that the data is drawn from some distribution on $\mathcal{Z} \times \mathcal{Y}$, where $\mathcal{Z}$ is a latent space. We do not observe samples from $\mathcal{Z}$. Instead, there are l transformation functions $g_1, \ldots, g_l$, where $g_k : \mathcal{Z} \to \mathcal{X}$. For each point z_i, there is an assigned group k and we observe $x_i = g_k(z_i)$. Then, our model is the following:

$$P(\lambda^1(z), \ldots, \lambda^m(z), y) = \frac{1}{Z} \exp\left(\theta_y y + \sum_{j=1}^{m} \frac{\theta_j}{1 + d(x^{\text{center}_j}, g_k(z))} \lambda^j(g_k(z)) y\right). \tag{2}$$

We explain this model as follows. We can think of it as a particular version of (1). However, instead of arbitrary $\theta_{j,x}$ parameters for each x, we explicitly model these parameters as two components: a feature-independent accuracy parameter θ_j and a term that modulates the accuracy based on the distance between feature vector x and some fixed center x^{center_j}. The center represents, for each LF, a *most accurate point*, where accuracy is maximized at a level set by θ_j. As the feature vector $x = g_k(z)$ moves away from this center, the denominator $1 + d(x^{\text{center}_j}, g_k(z))$ increases, and the LF votes increasingly poorly. This is an explicit form of smoothness that we validate empirically below.

For simplicity, we assume there are two groups, indexed by $0, 1$, that $\mathcal{X} = \mathcal{Z}$, and that $g_0(z) = z$. In other words, the transformation for group 0 is the identity, while this may not be the case for group 1. Simple extensions of our approach can handle cases where none of these assumptions are met.

Labeling Function Bias The model (2) explains how and when labeling functions might be biased. Suppose that g_k takes points z far from x^{center_j}. Then, the denominator term in (2) grows—and so the penalty for $\lambda(x)$ to disagree with y is reduced, making the labeling function less accurate. This is common in practice. For example, consider a situation where a bank uses features that include credit scores for loan review. Suppose the group variable is the applicant's nationality. Immigrants typically have a shorter period to build credit; this is reflected in a transformed distribution $g_1(z)$. A labeling function using a credit score threshold may be accurate for non-immigrants, but may end up being highly inaccurate when applied to immigrants.We validate this notion empirically. We used the Adult dataset [K$^+$96], commonly used for fairness studies, with a set of custom-built labeling functions.

In Figure 2, we track the accuracies of these LFs as a function of distance from an empirically-discovered center x^{center_j}. On the left is the high-accuracy group; as expected in our model, as we increase the distance, the accuracy decreases. On the right-hand side, we see the lower-accuracy group, whose labeling functions are voting $x_i = g_1(z_i)$. This transformation has sent these points further away from the center (note the larger distances). As a result, the overall accuracies have also decreased. Note, for example, how LF 5, in purple, varies between 0.9 and 1.0 accuracy in one group and is much worse—between 0.6 and 0.7—in the other.

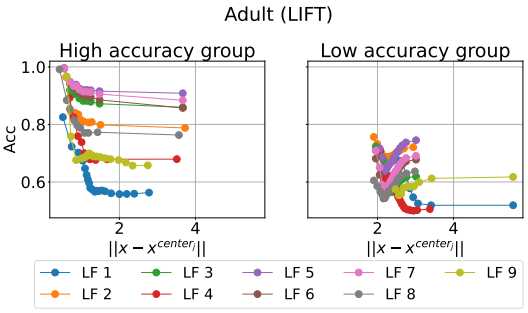


Figure 2: Average accuracy (y-axis) depending on the distance to the center point (x-axis). The center is obtained by evaluating the accuracy of their neighborhood data points.

3.1 Correcting Unfair LFs

Given the model (2), how can we reduce the bias induced by the g_k functions? A simple idea is to *reverse* the effect of the g_k's. If we could invert these functions, violations of fairness would be mitigated, since the accuracies of labeling functions would be uniformized over the groups.

Concretely, suppose that g_k is invertible and that h_k is this inverse. If we knew h_k, then we could ask the labeling functions to vote on $h_k(x) = h_k(g_k(z)) = z$, rather than on $x = g_k(z)$, and we could do so for any group, yielding equal-accuracy estimates for all groups. The technical challenge is how to estimate the inverses of the g_k's, without any parametric form for these functions. To do so, we

Algorithm 1: SOURCE BIAS MITIGATION (SBM)

Parameters: Features X_0, X_1 and LF outputs $\Lambda_0 = [\lambda_0^1,...,\lambda_0^m]$, $\Lambda_1 = [\lambda_1^1,...,\lambda_1^m]$ for groups 0, 1, transport threshold ε
Returns: Modified weak labels $\Lambda = [\lambda^1,...,\lambda^m]$
Estimate accuracy of λ^j in each group, $\hat{a}_0^j, \hat{a}_1^j$ from Λ_0, Λ_1 with Algorithm 2
for $j \in \{1,2,...,m\}$ **do**
    if $\hat{a}_1^j \geq \hat{a}_0^j + \varepsilon$ **then** update λ_0^j by transporting X_0 to X_1 (Algorithm 3)
    else if $\hat{a}_0^j \geq \hat{a}_1^j + \varepsilon$ **then** update λ_1^j by transporting X_1 to X_0 (Algorithm 3)
end for
return $\Lambda = [\lambda^1,...,\lambda^m]$

deploy optimal transport (OT) [PC$^+$19]. OT transports a probability distribution to another probability distribution by finding a minimal cost coupling. We use OT to recover the reverse map $h_k : \mathcal{X} \to \mathcal{Z}$ by $\hat{h}_k = \text{arginf}_{T(\nu)=\omega}\{\int_{x\in\mathcal{X}} c(x,T(x))d\nu(x)\}$, where c is a cost functon, ν is a probability measure in $\mathcal{X}$ and ω is a probability measure in $\mathcal{Z}$.

Our proposed approach, building on the use of OT, is called *source bias mitigation* (SBM). It seeks to reverse the group transformation g_k via OT. The core routine is described in Algorithm 1. The first step of the algorithm is to estimate the accuracies of each group so that we can identify which group is privileged, i.e., which of the transformations g_0, g_1 is the identity map. To do this, we use Algorithm 2 [FCS$^+$20] by applying it to each group separately.

After identifying the high-accuracy group, we transport data points from the low-accuracy group to it. Since not every transported point perfectly matches an existing high-accuracy group point, we find a nearest neighbor and borrow its label. We do this only when there is a sufficient inter-group accuracy gap, since the error in transport might otherwise offset the benefit. In practice, if the transformation is sufficiently weak, it is possible to skip optimal transport and simply use nearest neighbors. Doing this turned out to be effective in some experiments (Section 5.1). Finally, after running SBM, modified weak labels are used in a standard weak supervision pipeline, which is described in Appendix C.

4 Theoretical Results

We provide two types of theoretical results. First, we show that labeling function bias can be arbitrarily bad—resulting in substantial unfairness—regardless of whether the underlying dataset is fair. Next, we show that in certain settings, we can consistently recover the fair labeling function performance when using Algorithm 1, and provide a finite-sample error guarantee. Finally, we comment on extensions. All proofs are located in Appendix D.

Setting and Assumptions We assume that the distributions $P_0(x)$ and $P_1(x')$ are subgaussian with means μ_0 and μ_1 and positive-definite covariance matrices Σ_0 and Σ_1, respectively. Note that by assumption, $P_0(x) = P(z)$ and $P_1(x')$ is the pushforward of $P_0(x)$ under g_1. Let $\mathbf{r}(\Sigma)$ denote the effective rank of Σ [Ver18]. We observe n_0 and n_1 i.i.d. samples from groups 0 and 1, respectively. We use Euclidean distance as the distance $d(x,y) = ||x-y||$ in model (2). For the unobserved ground truth labels, y_i is drawn from some distribution $P(y|z)$. Finally, the labeling functions voting on our points are drawn via the model (2).

4.1 Labeling Functions can be Arbitrarily Unfair

We show that, as a result of the transformation g_1, the predictions of labeling functions can be arbitrarily unfair even if the dataset is fair. The idea is simple: the average group 0 accuracy, $\mathbb{E}_{z\in\mathcal{Z}}[P(\lambda(I(z)) = y)]$, is independent of g_1, so it suffices to show that $\mathbb{E}_{x'\in g_1(\mathcal{Z})}[P(\lambda(x') = y)]$ can deteriorate when g_1 moves data points far from the center x^{center_0}. As such, we consider the change in $\mathbb{E}_{x'\in g_1(\mathcal{Z})}[P(\lambda(x') = y)]$ as the group 1 points are transformed increasingly far from x^{center_0} in expectation.

Theorem 4.1. *Let $g_1^{(k)}$ be an arbitrary sequence of functions such that $\lim_{k\to\infty}\mathbb{E}_{x'\in g_1^{(k)}(\mathcal{Z})}[||x' - x^{center_0}||] \to \infty$. Suppose our assumptions above our met; in particular, that the label y is independent of the observed features $x = I(z)$ or $x' = g_1^{(k)}(z), \forall k$, conditioned on the latent features z. Then,*

$$\lim_{k\to\infty} \mathbb{E}_{x'\in g_1^{(k)}(\mathcal{Z})}[P(\lambda(x') = y)] = \frac{1}{2},$$

Table 1: Tabular dataset results

	Adult				Bank Marketing			
	Acc (↑)	F1 (↑)	Δ_{DP} (↓)	Δ_{EO} (↓)	Acc (↑)	F1 (↑)	Δ_{DP} (↓)	Δ_{EO} (↓)
FS	0.824	0.564	0.216	0.331	0.912	0.518	0.128	0.117
WS (Baseline)	0.717	0.587	0.475	0.325	0.674	0.258	0.543	0.450
SBM (w/o OT)	0.720	**0.592**	0.439	0.273	0.876	**0.550**	0.106	**0.064**
SBM (OT-L)	0.560	0.472	0.893	0.980	**0.892**	0.304	0.095	0.124
SBM (OT-S)	0.723	0.590	0.429	0.261	0.847	0.515	0.122	0.080
SBM (w/o OT) + LIFT	0.704	0.366	0.032	0.192	0.698	0.255	**0.088**	0.137
SBM (OT-L) + LIFT	0.700	0.520	0.015	**0.138**	**0.892**	0.305	0.104	0.121
SBM (OT-S) + LIFT	**0.782**	0.448	**0.000**	0.178	0.698	0.080	0.109	0.072

which corresponds to random guessing.

It is easy to construct such a sequence of functions $g_1^{(k)}$, for instance by letting $g_1^{(k)}(z) = z + ku$, where u is a d-dimensional vector of ones. When the distribution of group 1 points lies far from x^{center_0} while the distribution of group 0 points lies near to x^{center_0}, the accuracy parity of λ suffers. With adequately large expected $d(x^{\text{center}_0}, g_1^{(k)}(z))$, the performance of λ on group 1 points approaches random guessing.

4.2 Finite-Sample Bound for Mitigating Unfairness

Next, we provide a result bounding the difference in LF accuracy between group 0 points, $\mathbb{E}_{x \in \mathcal{Z}}[P(\lambda(x) = y)]$, and group 1 points transformed using our method, $\mathbb{E}_{x' \in \mathcal{X}}[P(\lambda(\hat{h}(x')) = y)]$. A tighter bound on this difference corresponds to better accuracy intra-group parity.

Theorem 4.2. *Set τ to be $\max\big(\mathbf{r}(\Sigma_0)/n_0, \mathbf{r}(\Sigma_1)/n_1, t/\min(n_0,n_1), t^2/\max(n_0,n_1)^2\big)$, and let C be a constant. Under the assumptions described above, when using Algorithm 1, for any $t > 0$, we have that with probability $1 - e^{-t} - 1/n_1$,*

$$|\mathbb{E}_{x \in \mathcal{Z}}[P(\lambda(x) = y)] - \mathbb{E}_{x' \in \mathcal{X}}[P(\lambda(\hat{h}(x')) = y)]| \leq 4\theta_0 C \sqrt{\tau \mathbf{r}(\Sigma_1)},$$

Next we interpret Theorem 4.2. LF accuracy recovery scales with $\max\big(1/\sqrt{n_1}, 1/\sqrt{n_2}\big)$. This does not present any additional difficulties compared to vanilla weak supervision—it is the same rate we need to learn LF accuracies. In other words, there is no sample complexity penalty for using our approach. Furthermore, LF accuracy recovery scales inversely to $\max\Big(\sqrt{\mathbf{r}(\Sigma_0)\mathbf{r}(\Sigma_1)}, \mathbf{r}(\Sigma_1)\Big)$. That is, when the distributions $P_0(x)$ or $P_1(x')$ have greater spread, it is more difficult to restore fair behavior.

Finally, we briefly comment on extensions. It is not hard to extend these results to a setting with less strict assumptions. For example, we can take P to be a mixture of Gaussians. In this case, it is possible to combine algorithms for learning mixtures [CGT18] with the approach we presented.

5 Experiments

The primary objective of our experiments is to validate that SBM improves fairness while often enhancing model performance as well. In real data experiments, we confirm that our methods work well with real-world fairness datasets (Section 5.1). In the synthetic experiments, we validate our theory claims in a fully controllable setting—showing that our method can achieve perfect fairness and performance recovery (Section 5.2). In addition, we show that our method **is compatible with other fair ML techniques** developed for fully supervised learning (Section 5.3). Finally, we demonstrate that our method can improve weak supervision performance beyond fairness by applying techniques to discover underperforming data slices (Section 5.4). This enables us to outperform state-of-the-art on a popular weak supervision benchmark.

Table 2: NLP dataset results

Dataset	Methods	Acc	F1	Δ_{DP}	Δ_{EO}
Civil	FS	0.893	0.251	0.083	0.091
	WS (Baseline)	0.854	**0.223**	0.560	0.546
	SBM (w/o OT)	0.879	0.068	0.048	0.047
	SBM (OT-L)	0.880	0.070	0.042	0.039
	SBM (OT-S)	**0.882**	0.047	**0.028**	**0.026**
Hate	FS	0.698	0.755	0.238	0.121
	WS (Baseline)	0.584	0.590	0.170	0.133
	SBM (w/o OT)	0.592	0.637	0.159	0.138
	SBM (OT-L)	**0.670**	0.606	0.120	0.101
	SBM (OT-S)	0.612	**0.687**	**0.072**	**0.037**

Table 3: Vision dataset results

Dataset	Methods	Acc	F1	Δ_{DP}	Δ_{EO}
CelebA	FS	0.897	0.913	0.307	0.125
	WS (Baseline)	0.866	0.879	0.308	0.193
	SBM (w/o OT)	0.870	0.883	0.309	0.192
	SBM (OT-L)	0.870	0.883	**0.306**	0.185
	SBM (OT-S)	**0.872**	**0.885**	**0.306**	**0.184**
UTKF	FS	0.810	0.801	0.133	0.056
	WS (Baseline)	0.791	0.791	0.172	0.073
	SBM (w/o OT)	0.797	0.790	0.164	0.077
	SBM (OT-L)	0.800	0.793	0.135	0.043
	SBM (OT-S)	**0.804**	**0.798**	**0.130**	**0.041**

5.1 Real data experiments

Claims Investigated In real data settings, we hypothesize that our methods can reduce the bias of LFs, leading to better fairness and improved performance of the weak supervision end model.

Setup and Procedure We used 6 datasets in three different domains: tabular (Adult and Bank Marketing), NLP (CivilComments and HateXplain), and vision (CelebA and UTKFace). Their task and group variables are summarized in Appendix E, Table 7. LFs are either heuristics or pretrained models. More details are included in Appendix E.3.

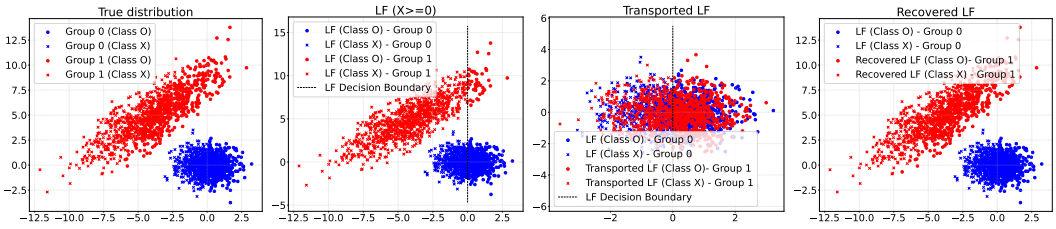
For the weak supervision pipeline, we followed a standard procedure. First, we generate weak labels from labeling functions in the training set. Secondly, we train the label model on weak labels. In this experiment, we used Snorkel [BRL$^+$19] as the label model in weak supervision settings. Afterwards, we generate pseudolabels from the label model, train the end model on these, and evaluate it on the test set. We used logistic regression as the end model. The only difference between our method and the original weak supervision pipeline is a procedure to fix weak labels from each labeling function. As a sanity check, a fully supervised learning result (FS), which is the model performance trained on the true labels, is also provided. Crucially, however, *in weak supervision, we do not have such labels*, and therefore fully supervised learning is simply an upper bound to performance—and not a baseline.

We ran three variants of our method. *SBM (w/o OT)* is a 1-nearest neighbor mapping to another group without any transformation. *SBM (OT-L)* is a 1-nearest neighbor mapping with a linear map learned via optimal transport. *SBM (OT-S)* is a 1-nearest neighbor mapping with a Monge mapping learned via the Sinkhorn algorithm. To see if our method can improve both fairness and performance, we measured the demographic parity gap (Δ_{DP}) and the equal opportunity gap (Δ_{EO}) as fairness metrics, and computed accuracy and F1 score as performance metrics as well.

Results The tabular dataset result is reported in Table 1. As expected, our method improves accuracy while reducing demographic parity gap and equal opportunity gap. However, we observed *SBM (OT-L)* critically fails at Adult dataset, contrary to what we anticipated. We suspected this originates in one-hot coded features, which might distort computing distances in the nearest neighbor search. To work around one-hot coded values in nearest neighbor search, we deployed LIFT [DZZ$^+$22], which encodes the input as natural language (e.g. "She/he is <race attribute>. She/he works for <working hour attribute> per week ...") and embeds them with language models (LMs). We provide heuristic rules to convert feature columns into languages in Appendix E.2, and used BERT as the language model. The result is given in Table 1 under the dashed lines. While it sacrifices a small amount of accuracy, it substantially reduces the unfairness as expected.

The results for NLP datasets are provided in Table 2. In the CivilComments and HateXplain datasets, we observed our methods mitigate bias consistently, as we hoped. While our methods improve performance as well in the HateXplain dataset, enhancing other metrics in CivilComments results in drops in the F1 score. We believe that a highly unbalanced class setting ($P(Y=1) \approx 0.1$) is the cause of this result.

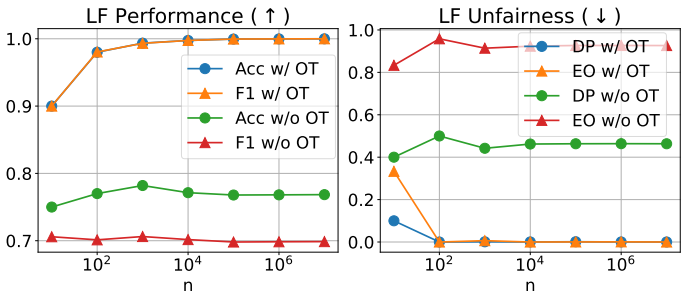
The results for vision datasets are given in Table 3. Though not as dramatic as other datasets since here the LFs are pretrained models, none of which are heavily biased, our methods can still improve accuracy and fairness. In particular, our approach shows clear improvement over the baseline, which yields performance closer to the fully supervised learning setting while offering less bias.



(a) True distribution (b) LF (c) Transported LF (d) Recovered LF

Figure 3: Synthetic datasets. In (a), seemingly different data distributions from the two groups actually have perfect achievable fairness. However, the labeling function in (b) only works well in group 0, which leads to unfairness. Via OT (c), the input distribution can be matched and the LF applied to similar groups—original and recovered. As a result, LFs on group 1 works as well as on 0 (d).

5.2 Synthetic experiment



(a) LF Performance w/ OT (b) LF Unfairness w/ OT

Figure 4: Synthetic experiment result. As OT recovers the transformation induced by the group attribute, the performance improves and the unfairness drops.

Claim Investigated We hypothesized that our method can recover both fairness and accuracy (as a function of the number of samples available) by transporting the distribution of one group to another group when our theoretical assumptions are almost exactly satisfied. To show this, we generate unfair synthetic data and LFs and see if our method can remedy LF fairness and improve LF performance.

Setup and Procedure We generated a synthetic dataset that has perfect fairness as follows. First, n input features in $\mathbb{R}^2$ are sampled as $X_0 \sim \mathcal{N}(\mathbf{0}, I)$ for group 0, and labels Y_0 are set by $Y_0 = \mathbb{1}(X_0[0] \geq 0.5)$, i.e. 1 if the first dimension is positive or equal. Afterwards, n input features in $\mathbb{R}^2$ are sampled as $\tilde{X}_1 \sim \mathcal{N}(0, I)$ for group 1, and the labels are also set by $Y_1 = \mathbb{1}(\tilde{X}_1[0] \geq 0.5)$. Then, a linear transformation is applied to the input distribution: $X_1 = \Sigma\tilde{X}_1 + \mu$ where $\mu = \begin{bmatrix} -4 \\ 5 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, which is the distribution of group 1. Clearly, we can see that $X_1 = \Sigma X_0 + \mu \sim \mathcal{N}(\mu, \Sigma)$. Here we applied the same labeling function $\lambda(x) = \mathbb{1}(x[0] \geq 0)$, which is the same as the true label distribution in group 0.

We apply our method (SBM OT-L) since our data model fits its basic assumption. Again, we evaluated the results by measuring accuracy, F1 score, Δ_{DP}, and Δ_{EO}. The setup and procedure are illustrated in Figure 3. We varied the number of samples n from 10^2 to 10^7

Results The result is reported in Figure 4. As we expected, we saw the accuracy and F1 score are consistently improved as the linear Monge map is recovered when the number of samples n increases. Most importantly, we observed that perfect fairness is achieved after only a small number of samples (10^2) are obtained.

5.3 Compatibility with other fair ML methods

Claim Investigated Our method corrects labeling function bias at the individual LF—and not model—level. We expect our methods can work cooperatively, in a constructive way, with other fair ML methods developed for fully supervised learning settings.

Table 4: Compatibility with other fair ML methods (HateXplain dataset)

	Acc	F1	Δ_{DP}	Δ_{EO}
FS	0.698	0.755	0.238	0.121
WS (Baseline)	0.584	0.590	0.171	0.133
SBM (OT-S)	**0.612**	0.687	0.072	0.037
WS (Baseline) + OTh-DP	0.539	0.515	0.005	0.047
SBM (OT-S) + OTh-DP	0.607	**0.694**	**0.002**	**0.031**

Table 5: Slice discovery with SBM results in WRENCH. Evaluation metric is accuracy for iMDb, F1 for the rest.

Methods	Basket ball	Census	iMDb	Mush room	Tennis
FS	0.855	0.634	0.780	0.982	0.858
WS (HyperLM)	0.259	0.551	0.753	0.866	0.812
SBM (w/o OT)	**0.261**	**0.568**	0.751	0.790	**0.819**
SBM (OT-L)	0.242	0.547	**0.756**	0.903	0.575
SBM (OT-S)	0.260	0.552	**0.756**	**0.935**	0.663

Setup and Procedure We used the same real datasets, procedures, and metrics as before. We combined the optimal threshold method [HPS16] with WS (baseline) and our approach, SBM (Sinkhorn). We denote the optimal threshold with demographic parity criteria as OTh-DP.

Results The results are shown in Table 4. As we expected, we saw the effect of optimal threshold method, which produces an accuracy-fairness (DP) tradeoff. This has the same effect upon our method. Thus, when optimal threshold is applied to both, our method has better performance and fairness aligned with the result without optimal threshold. More experimental results with other real datasets and additional fair ML methods are reported in Appendix E.5.

5.4 Beyond fairness: maximizing performance with slice discovery

Claim Investigated We postulated that even outside the context of improving fairness, our techniques can be used to boost the performance of weak supervision approaches. In these scenarios, there are no pre-specified groups. Instead, underperforming latent groups (slices) must first be discovered. Our approach then uses transport to improve labeling function performance on these groups.

Setup and Procedure We used Basketball, Census, iMDb, Mushroom, and Tennis dataset from the WRENCH benchmark [ZYL+21], which is a well-known weak supervision benchmark but does not include any group information. We generated group annotations by slice discovery [KGZ19, SNS+21, ddWLB22, EVS+22], which is an approach to discover data slices that share a common characteristic. To find groups with a large accuracy gap, we used Domino [EVS+22]. It discovers regions of the embedding space based on the accuracy of model. Since the WS setting does not allow access to true labels, we replaced true labels with pseudolabels obtained from the label model and model scores with label model probabilities. In order to show we can increase performance *even for state-of-the-art weak supervision*, we used the recently-proposed state-of-the-art Hyper Label Model [WCZC] as the label model. We used the group information generated by the two discovered slices to apply our methods. We used logistic regression as the end model, and used the same weak supervision pipeline and metrics as in the other experiments, excluding fairness.

Results The results can be seen in Table 5. As expected, even without known group divisions, we still observed improvements in accuracy and F1 score. We see the most significant improvements on the Mushroom dataset, where we substantially close the gap to fully-supervised. These gains suggest that it is possible to generically combine our approach with other principled methods for subpopulation discovery to substantially improve weak supervision in general settings.

6 Conclusion

Weak supervision has been successful in overcoming manual labeling bottlenecks, but its impact on fairness has not been adequately studied. Our work has found that WS can easily induce additional bias due to unfair LFs. In order to address this issue, we have proposed a novel approach towards mitigating bias in LFs and further improving model performance. We have demonstrated the effectiveness of our approach using both synthetic and real datasets and have shown that it is compatible with traditional fair ML methods. We believe that our proposed technique can make weak supervision safer to apply in important societal settings and so encourages its wider adoption.

References

[ABD+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[BDB22] Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. In *Advances in Neural Information Processing Systems*, 2022.

[BDE+20] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

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

[BRL+19] Stephen H Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, et al. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, pages 362–375, 2019.

[BYF20] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

[CFA+22] Mayee F Chen, Daniel Y Fu, Dyah Adila, Michael Zhang, Frederic Sala, Kayvon Fatahalian, and Christopher Ré. Shoring up the foundations: Fusing model embeddings and weak supervision. In *Uncertainty in Artificial Intelligence*, pages 357–367. PMLR, 2022.

[CGT18] Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Optimal transport for gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018.

[Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[ddWLB22] Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981, 2022.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[DS79] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

[DZS+17] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conferenceon Research and Development in Information Retrieval*, 2017.

[DZZ+22] Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. In *Advances in Neural Information Processing Systems*, 2022.

[EVS+22] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

[FCG+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[FCS+20] Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2020.

[FLF19] Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*, 2019.

[GDBFL19] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR, 2019.

[GM14] Sonal Gupta and Christopher D Manning. Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108, 2014.

[GNRV19] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

[GZS+22] Ozgur Guldogan, Yuchen Zeng, Jy-yong Sohn, Ramtin Pedarsani, and Kangwook Lee. Equal improvability: A new fairness notion considering the long-term impact. *arXiv preprint arXiv:2210.06732*, 2022.

[HNG19] Hoda Heidari, Vedant Nanda, and Krishna P Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *arXiv preprint arXiv:1903.01209*, 2019.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[HU20] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[HWZW20] Wen Huan, Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, pages 743–751, 2020.

[K+96] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[KGZ19] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[KL22] Nikola Konstantinov and Christoph H Lampert. Fairness-aware pac learning from corrupted data. *The Journal of Machine Learning Research*, 23(1):7173–7232, 2022.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[KOS11] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.

[KS84] Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984.

[KSM+21] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[LVS] Hunter Lang, Aravindan Vijayaraghavan, and David Sontag. Training subset selection for weak supervision. In *Advances in Neural Information Processing Systems*.

[MBSJ09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[MCR14] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[MSY+21] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

[MW18] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[RBE+18] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the 44th International Conference on Very Large Data Bases (VLDB)*, Rio de Janeiro, Brazil, 2018.

[RHD+19a] A. J. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019.

[RHD+19b] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771, 2019.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[RSW+16] A. J. Ratner, Christopher M. De Sa, Sen Wu, Daniel Selsam, and C. Ré. Data programming: Creating large training sets, quickly. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, 2016.

[SLV+22] Changho Shin, Winfred Li, Harit Vishwakarma, Nicholas Carl Roberts, and Frederic Sala. Universalizing weak supervision. In *International Conference on Learning Representations (ICLR)*, 2022.

[SMBN21] Nian Si, Karthyek Murthy, Jose Blanchet, and Viet Anh Nguyen. Testing group fairness via optimal transport projections. In *International Conference on Machine Learning*, pages 9649–9659. PMLR, 2021.

[SNS+21] Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12853–12862, 2021.

[SRT+20] Chiappa Silvia, Jiang Ray, Stepleton Tom, Pacchiano Aldo, Jiang Heinrich, and Aslanides John. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3633–3640, 2020.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[VS22] Harit Vishwakarma and Frederic Sala. Lifting weak supervision to structured prediction. In *Advances in Neural Information Processing Systems*, 2022.

[WCZC] Renzhi Wu, Shen-En Chen, Jieyu Zhang, and Xu Chu. Learning hyper label model for programmatic weak supervision. In *The Eleventh International Conference on Learning Representations*.

[WH22] Ziwei Wu and Jingrui He. Fairness-aware model-agnostic positive and unlabeled learning. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2022.

[WLL21] 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, 2021.

[WZN+23] Jiaheng Wei, Zhaowei Zhu, Gang Niu, Tongliang Liu, Sijia Liu, Masashi Sugiyama, and Yang Liu. Fairness improves learning from noisily labeled long-tailed data. *arXiv preprint arXiv:2303.12291*, 2023.

[ZDC22] Xianli Zeng, Edgar Dobriban, and Guang Cheng. Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*, 2022.

[ZHY+22] Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*, 2022.

[ZSQ17] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[ZYL+21] Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*, 2021.

[ZZL+23] Yixuan Zhang, Feng Zhou, Zhidong Li, Yang Wang, and Fang Chen. Fair representation learning with unreliable labels. In *International Conference on Artificial Intelligence and Statistics*, pages 4655–4667. PMLR, 2023.