Fairness Without Demographics in Repeated Loss Minimization

Tatsunori B. Hashimoto [1 2] **Megha Srivastava** [1] **Hongseok Namkoong** [3] **Percy Liang** [1]

Abstract

Machine learning models (e.g., speech recognizers) trained on average loss suffer from representation disparity—minority groups (e.g., non-native speakers) carry less weight in the training objective, and thus tend to suffer higher loss. Worse, as model accuracy affects user retention, a minority group can shrink over time. In this paper, we first show that the status quo of empirical risk minimization (ERM) amplifies representation disparity over time, which can even turn initially fair models unfair. To mitigate this, we develop an approach based on distributionally robust optimization (DRO), which minimizes the worst case risk over all distributions close to the empirical distribution. We prove that this approach controls the risk of the minority group at each time step, in the spirit of Rawlsian distributive justice, while remaining oblivious to the identity of the groups. We demonstrate that DRO prevents disparity amplification on examples where ERM fails, and show improvements in minority group user satisfaction in a real-world text autocomplete task.

1. Introduction

Consider a speech recognizer that is deployed to millions of users. State-of-the art speech recognizers achieve high overall accuracy, yet it is well known that such systems have systematically high errors on minority accents (Amodei et al., 2016). We refer to this phenomenon of high overall accuracy but low minority accuracy as a *representation disparity*, which is the result of optimizing for average loss. This representation disparity forms our definition of unfairness, and has been observed in face recognition (Grother et al., 2011), language identification (Blodgett et al., 2016; Jurgens et al., 2017), dependency parsing (Blodgett et al., 2016), part-of-speech tagging (Hovy & Sgaard, 2015), academic recommender systems (Sapiezynski et al., 2017), and automatic video captioning (Tatman, 2017).

Moreover, a minority user suffering from a higher error rate will become discouraged and more likely to stop using the system, thus no longer providing data to the system. As a result, the minority group will shrink and might suffer even higher error rates from a retrained model in a future time step. Machine learning driven feedback loops have been observed in predictive policing (Fuster et al., 2017) and credit markets (Fuster et al., 2017), and this problem of *disparity amplification* is a possibility in any deployed machine learning system that is retrained on user data.

In this paper, we aim to mitigate the representation disparity problem and its amplification through time. We focus on the following setting: at each time step, each user interacts with the current model and incurs some loss, based on which she decides to keep or quit using the service. A model is trained on the resulting user data which is used at the next time step. We assume that each user comes from one of K groups, and our goal is to minimize the worst case risk of any group across time. However, *the group membership and number of groups K are both unknown*, as full demographic information is likely missing in real online services.

We first show that empirical risk minimization (ERM) does not control the worst-case risk over the disparate K groups and show examples where ERM turns initially fair models unfair (Section 3). To remedy this issue, we propose the use of distributionally robust optimization (DRO) (Section 4). Given a lower bound on the smallest group proportion, we show that optimizing the worst-case risk over an appropriate chi-square divergence ball bounds the worst-case risk over groups. Our approach is computationally efficient, and can be applied as a small modification to a wide class machine learning models trained by stochastic-gradient descent methods. We show that DRO succeeds on the examples where ERM becomes unfair, and demonstrate higher average minority user satisfaction and lower disparity amplification on a Amazon Mechanical Turk based autocomplete task.

1.1. Fairness in Machine Learning

Recently, there has been a surge of interest in fairness in machine learning (Barocas & Selbst, 2016). Our work can

[1]Department of Computer Science, Stanford, USA [2]Department of Statistics, Stanford, USA [3]Management Science & Engineering, Stanford, USA. Correspondence to: Tatsunori Hashimoto <thashim@stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

be seen as a direct instantiation of John Rawls' theory on distributive justice and stability, where we view predictive accuracy as a resource to be allocated. Rawls argues that the *difference principle*, defined as maximizing the welfare of the worst-off group, is fair and stable over time since it ensures that minorities consent to and attempt to maintain the status quo (Rawls, 2001, p155).

In this work, we assume the task is general loss minimization, and demographic data is unavailable. This differs from the substantial body of existing research into fairness for classification problems involving protected labels such as the use of race in recidivism protection (Chouldechova, 2017). There has been extensive work (Barocas & Selbst, 2016) on guaranteeing fairness for classification over a protected label through constraints such as equalized odds (Woodworth et al., 2017; Hardt et al., 2016), disparate impact (Feldman et al., 2015) and calibration (Kleinberg et al., 2017). However, these approaches require the use of demographic labels, and are designed for classification tasks. This makes it difficult to apply such approaches to mitigate representation disparity in tasks such as speech recognition or natural language generation where full demographic information is often unavailable.

A number of authors have also studied individual notions of fairness, either through a fixed similarity function (Dwork et al., 2012) or subgroups of a set of protected labels (Kearns et al., 2018; Hébert-Johnson et al., 2017). Dwork et al. (2012) provides fairness guarantees without explicit groups, but requires a fixed distance function which is difficult to define for real-world tasks. Kearns et al. (2018); Hébert-Johnson et al. (2017) consider subgroups of a set of protected features, but defining non-trivial protected features which cover the latent demographics in our setting is difficult. Although these works generalize the demographic group structure, similarity and subgroup structure are both ill-defined for many real-world tasks.

In the online setting, works on fairness in bandit learning (Joseph et al., 2016; Jabbari et al., 2017) propose algorithms compatible with Rawls' principle on equality of opportunity—an action is preferred over another only if the true quality of the arm is better. Our work differs in considering Rawlsian fairness for distributive justice (Rawls, 2009). Simultaneous with our work, Liu et al. (2018) analyzed fairness over time in the context of constraint based fairness criteria, and show that enforcing static fairness constraints do not ensure fairness over time. In this paper, we consider latent demographic groups and study a loss-based approach to fairness and stability.

2. Problem setup

We begin by outlining the two parts of our motivation: *representation disparity* and *disparity amplification*.

Representation disparity: Consider the standard loss-minimization setting where a user makes a query $Z \sim P$, a model $\theta \in \Theta$ makes a prediction, and the user incurs loss $\ell(\theta; Z)$. We denote the expected loss as the risk $\mathcal{R}(\theta) = \mathbb{E}_{Z\sim P}[\ell(\theta; Z)]$. The observations Z are assumed to arise from one of K latent groups such that $Z \sim P := \sum_{k\in[K]} \alpha_k P_k$. We assume that neither the population proportions $\{\alpha_k\}$ nor the group distributions $\{P_k\}$ are known. The goal is to control the worst case risk over all K groups:

$$\mathcal{R}_{\max}(\theta) = \max_k \{\mathcal{R}_k(\theta) := \mathbb{E}_{P_k}[\ell(\theta; Z)] : k \in [K]\}. \quad (1)$$

Disparity amplification: To understand the amplification of representation disparity over time, we will make several assumptions on the behavior of users in response to observed losses. These assumptions are primarily for clarity of exposition—we will indicate whenever the assumptions can be relaxed leave generalizations to the supplement. Roughly speaking, minimizing the worst-case risk $\mathcal{R}_{\max}(\theta)$ should mitigate disparity amplification as long as lower losses lead to higher user retention. We now give assumptions that make this intuition precise.

In the sequential setting, loss minimization proceeds over $t = 1, 2, \ldots T$ rounds, where the group proportion $\alpha_k^{(t)}$ depends on t and varies according to past losses. At each round $\lambda_k^{(t+1)}$ is the expected number of users from group k, which is determined by $\nu(\mathcal{R}_k(\theta))$, the fraction of users retained, and b_k, the number of new users (see Definition 1). Here, ν is a differentiable, strictly decreasing retention function which maps a risk level $\mathcal{R}$ to the fraction of users who continue to use the system. Modeling user retention as a decreasing function of the risk implies that each user makes an independent decision of whether to interact with the system at time $t+1$ based on their expected loss at time t. For example, selecting $\nu(x) = 1 - x$ and $\mathcal{R}_k$ equal to the expected zero-one loss implies that users leave proportional to the misclassification rates of their queries.

At each round we learn parameters $\theta^{(t+1)}$ based on $n^{(t+1)} \sim \text{Pois}(\sum_k \lambda_k^{(t+1)})$ users (data points). While we define the sample size as a Poisson process for concreteness, our main results hold for any distribution fulfilling the strong law of large numbers, as we perform all stability analyses in the population limit.

Definition 1 (Dynamics)**.** *Given a sequence $\theta^{(t)}$, for each $t = 1 \ldots T$, the expected number of users λ and samples $Z_i^{(t)}$ starting at $\lambda_k^{(0)} = b_k$ is governed by,*

$$\lambda_k^{(t+1)} := \lambda_k^{(t)} \nu(\mathcal{R}_k(\theta^{(t)})) + b_k$$

$$\alpha_k^{(t+1)} := \frac{\lambda_k^{(t+1)}}{\sum_{k'\in[K]} \lambda_{k'}^{(t+1)}}$$

$$n^{(t+1)} := Pois(\sum_k \lambda_k^{(t+1)})$$
$$Z_1^{(t+1)} \dots Z_{n^{(t+1)}}^{(t+1)} \overset{\text{i.i.d.}}{\sim} P^{(t+1)} := \sum_{k\in[K]} \alpha_k^{(t+1)} P_k.$$

For example, if we use ERM at each time step the parameter sequence is defined as $\theta^{(t)} = \arg\min_{\theta\in\Theta} \sum_i \ell(\theta; Z_i^{(t)})$.

Our goal is to control the group-wise risk $\mathcal{R}_k(\theta^{(t)})$ over all $k = 1, \dots, K$ groups and time periods $t = 1, \dots, T$:

$$\mathcal{R}^T_{\max}(\theta^{(0)}, \cdots, \theta^{(T)}) = \max_{k,t} \left\{ \mathcal{R}_k(\theta^{(t)}) \right\} \tag{2}$$

given only the losses $\{\ell(\theta^{(t)}, Z_i)\}$ and samples Z_i.

Without knowledge of group membership labels, population proportions $\alpha_k^{(t)}$, new user rate b_k, and retention rate ν, minimizing $\mathcal{R}^T_{\max}$ gives rise to two major challenges. First, without group membership labels there is no way to directly measure the worst-case risk $\mathcal{R}^T_{\max}$, let alone minimize it. Second, we must ensure that the group proportions $\alpha_k^{(t)}$ are stable, since if $\alpha_k^{(t)} \to 0$ as $t \to \infty$ for some group $k \in [K]$, then no algorithm can control $\mathcal{R}^T_{\max}$ when a group has near zero probability of appearing in our samples.

We begin by illustrating how models that are initially fair with low representation disparity may become unfair over time if we use ERM (Section 3). We study real-world problems with representation disparity further in our experimental section (Section 5).

3. Disparity amplification

The standard approach to fitting a sequence of models $\theta^{(t)}$ is to minimize an empirical approximation to the population risk at each time period. In this section, we show that even minimizing the population risk fails to control minority risk over time, since expected loss (average-case) leads to disparity amplification. The decrease in user retention for the minority group exacerbates over time since once a group grows sufficiently small, it receives higher losses relative to others, leading to even fewer samples from the group.

3.1. Motivating example

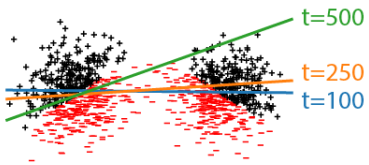


Figure 1. An example online classification problem which begins fair, but becomes unfair over time.

Consider the two-class classification problem in Figure 1a where the two groups are situated on the left/right and the true classification boundary is given along $x_2 = 0$. Assume that the sampling distribution evolves according to definition 1 with $\nu(x) = 1.0 - x$, ℓ equal to the zero one loss, and $b_0 = b_1 = n_0^{(0)} = n_1^{(0)} = 1000$. Initially, ERM has similar and high accuracy on both groups with the boundary $x_2 > 0$, but over time random fluctuations in accuracy result in slightly fewer samples from the cluster on the right. This leads to disparity amplification since ERM will further improve the loss on the *left* cluster at the expense of the *right* cluster. After 500 rounds, there are nearly no samples from the *right* cluster, and as a result, the *right* cluster ends up suffering high loss.

3.2. Conditions for disparity amplification

The example above demonstrated that disparity amplification can occur easily even from a completely fair and stable looking starting point, and result in the minority group to completely drop out. In general if we view the expected user counts $\lambda^{(t)}$ as a dynamical system, the long-term fairness properties are controlled by two factors - whether λ has a fair fixed point (such as $x_2 > 0$ in our motivating example) and whether this fixed point is stable.

Fixed points of risk minimization (average-case) are determined by a combination of user retention function ν and the models $\theta^{(t)}$, and without knowledge of ν it is hard to ensure that a model has a fair fixed point. Even if a fixed point is fair and we start at this fair fixed point, minimizing the average loss (e.g. ERM) may deviate from this fair fixed point over time.

To show this result, we study the dynamical system Φ, which is defined by dynamics in Definition 1 with θ derived from minimizing the population, rather than empirical risk.

Definition 2. *Let Φ be the update for the expected population size*

$$\lambda_k^{(t+1)} := \Phi(\lambda_k^{(t)}) = \lambda_k^{(t)} \nu(\mathcal{R}_k(\theta(\lambda_k^{(t)}))) + b_k,$$
$$\theta(\lambda_k^{(t)}) = \arg\min_\theta \mathbb{E}_{\sum_k \alpha_k^{(t)} P_k}[\ell(\theta; Z)].$$

The arrival intensity λ^* is called a fixed point if $\lambda^* = \Phi(\lambda^*)$. This fixed point is *stable* whenever the maximum modulus of the eigenvalues of the Jacobian of Φ is less than one and *unstable* whenever it is greater than one (Luo, 2012, Theorem 2.1). If a fair fixed point is unstable, then any perturbation results in disparity amplification where the population size deviates from λ^* over time.

Proposition 1 gives a precise statement of this phenomenon. We prove the result in Section A.1, and further show a generalization to general dynamics $\Phi(\lambda_k) = h(\lambda_k, \mathcal{R}_k)$ where h is differentiable and monotone in the second argument.

We denote by $\rho_{\max}(A)$ the maximum modulus of the eigenvalues of A.

Proposition 1. *Let $\lambda^* = \Phi(\lambda^*)$ be a fixed point, and $\theta^* = \arg\min_\theta \mathbb{E}_{\sum_k \alpha_k^* P_k}[\ell(\theta; Z)]$ be the minimizer at λ^*.*

Define $H_{\mathcal{R}}(\alpha^*)$ as the positive definite Hessian of the expected risk at θ^*, λ^* and define ∇L as the per-group parameter gradients at θ^*,

$$\nabla L = \begin{bmatrix} \nabla_\theta \mathbb{E}_{P_1}[\ell(\theta^*; Z)] \\ \vdots \\ \nabla_\theta \mathbb{E}_{P_k}[\ell(\theta^*; Z)] \end{bmatrix}.$$

The arrival intensity λ^* is unstable whenever

$$\rho_{\max}\Bigg(\operatorname{diag}(\nu(\mathcal{R}(\theta(\lambda^*)))) - \operatorname{diag}(\lambda^* \nu'(\mathcal{R}(\theta(\lambda^*)) \\ \nabla L H_{\mathcal{R}}(\alpha^*)^{-1} \nabla L^\top \left(\frac{I}{\sum_k \lambda_k^*} - \frac{\mathbf{1}\lambda^{*\top}}{(\sum_k \lambda_k^*)^2} \right) \Bigg) > 1.$$

We see that the major quantities which control risk are the retention rate ν and its derivative, as well as a $K \times K$ square matrix $\nabla L H_{\mathcal{R}}(\alpha^*)^{-1} \nabla L^\top$ which roughly encodes the changes in one group's risk as a function of another.

We can specialize the stability condition to obtain an intuitive and negative result for the stability of risk minimization (average-case). Even if we start at a fair fixed point with $\lambda_1^* = \lambda_2^* \ldots = \lambda_k^*$ and $\mathcal{R}_1 = \mathcal{R}_2 \ldots = \mathcal{R}_k$, if decreasing the risk for one group increases the risk for others sufficiently, the fixed point is unstable and the model will eventually converge to a different, possibly unfair, fixed point.

Corollary 1 (Counterexample under symmetry)**.** *Let $\lambda_1^* = \ldots \lambda_k^*$ be a fixed point with $\mathcal{R}_1 = \ldots \mathcal{R}_k$, then for any strongly convex loss,*

$$\rho_{\max}\left(\nabla L H_{\mathcal{R}}(\alpha^*)^{-1} \nabla L^\top \right) > \frac{1 - \nu(\mathcal{R}_1)}{-\nu'(\mathcal{R}_1)/k}. \quad (3)$$

is a sufficient condition for instability.

See Section A.2 for proof and generalizations.

The bound (3) has a straightforward interpretation. The left hand side is the stability of the model, where maximal eigenvalue of the matrix $\nabla L H_{\mathcal{R}}(\alpha^*)^{-1} \nabla L^\top$ represents the maximum excess risk that can be incurred due to a small perturbation in the mixture weights α. The right hand side represents the underlying stability of the dynamics and measures the sensitivity of λ with respect to risk.

Mean and median estimation: Consider a simple mean estimation example where each user belongs to one of two groups, -1 or 1 and incurs loss $(\theta - Z)^2$. $\theta = 0$ is clearly a fair fixed point, since it equalizes losses to both groups, with $H_{risk}(\alpha^*) = 1/2$ and $\nabla L = [2, -2]$ making $\rho_{\max}\left(\nabla L H_{\mathcal{R}}(\alpha^*)^{-1} \nabla L^\top \right) = 4$. If we select $\nu(x) = \exp(-x)$, the right hand side becomes $2(1 - e^{-1})e \approx 3.4$, and thus any perturbation will eventually result in $\lambda_1 \neq \lambda_2$. In this case the only other fixed points are the unfair solutions of returning the mean of either one of the groups.

The situation is even worse for models which are not strongly convex, such as median estimation. Replacing the squared loss above with the absolute value results in a loss which has a non-unique minimizer at 0 when $\lambda_1 = \lambda_2$ but immediately becomes -1 whenever $\lambda_1 > \lambda_2$. In this case, no conditions on the retention function ν can induce stability. This fundamental degeneracy motivates us to search for loss minimization schemes with better stability properties than ERM (average-case).

4. Distributionally robust optimization (DRO)

Recall that our goal is to control the worst-case risk (2) over all groups and over all time steps t. We will proceed in two steps. First, we show that performing distributionally robust optimization controls the worst-case risk $\mathcal{R}_{\max}(\theta^{(t)})$ for a single time step. Then, we show that this results in a lower bound on group proportions $\{\alpha_k^{(t)}\}_{k=1}^K$, and thus ensures control over the worst-case risk for all time steps. As a result of the two steps, we show in Section 4.4 that our procedure mitigates disparity amplification over *all time steps*. For notational clarity, we omit the superscript t in Sections 4.1-4.3.

4.1. Bounding the risk over unknown groups

The fundamental difficulty in controlling the worst-case group risk over a single time-step $\mathcal{R}_{\max}(\theta^{(t)})$ comes from not observing the group memberships from which the data was sampled. For many machine learning systems such as speech recognition or machine translation, such situations are common since we either do not ask for sensitive demographic information, or it is unclear apriori which demographics should be protected. To achieve reasonable performance across different groups, we postulate a formulation that protects against *all* directions around the data generating distribution. We build on the distributionally robust formulation of Duchi et al. (2016) which will allow controlling worst-case group risk $\mathcal{R}_{\max}(\theta^{(t)})$.

To formally describe our approach, let $D_{\chi^2}(P\|Q)$ be the χ^2-divergence between probability distributions P and Q given by $D_{\chi^2}(P\|Q) := \int \left(\frac{dP}{dQ} - 1\right)^2 dQ$. If P is not absolutely continuous with respect to Q, we define $D_{\chi^2}(P\|Q) := \infty$. Let $\mathcal{B}(P, r)$ be the chi-squared ball around a probability distribution P of radius r so that

$\mathcal{B}(P, r) := \{Q \ll P : D_{\chi^2}(Q\|P) \leq r\}$. We consider the worst-case loss over all r-perturbations around P,

$$\mathcal{R}_{\text{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]. \quad (4)$$

Intuitively, the distributionally robust risk $\mathcal{R}_{\text{dro}}(\theta; r)$ upweights examples Z with high loss $\ell(\theta; Z)$. If there is a group suffering high loss, the corresponding mixture component will be over-represented (relative to the original mixture weights) in the distributionally robust risk $\mathcal{R}_{\text{dro}}(\theta; r)$. We show in the following proposition that $\mathcal{R}_{\text{dro}}(\theta; r)$ bounds the risk of each group $\mathcal{R}_k(\theta)$, and hence the group-wise worst-case risk (1), for an appropriate choice of the robustness radius r.

Proposition 2. *For $P := \sum_{k\in[K]} \alpha_k P_k$, we have $\mathcal{R}_k(\theta) \leq \mathcal{R}_{\text{dro}}(\theta; r_k)$ for all $\theta \in \Theta$ where $r_k := (1/\alpha_k - 1)^2$ is the robustness radius.*

We prove the result in Section A.4. Roughly speaking, the above bound becomes tighter if the variation in the loss $\ell(\theta; Z)$ is substantially higher between groups than within each group. In particular, this would be the case if the loss distribution for each group have distinct support with relatively well-concentrated components within each group.

As a consequence of Proposition 2, if we have a lower bound on the group proportions $\alpha_{\text{min}} \leq \min_{k\in[K]} \alpha_k$, then we can control the worst-case group risk $\mathcal{R}_{\text{max}}(\theta)$ by minimizing the upper bound $\theta \mapsto \mathcal{R}_{\text{dro}}(\theta; r_{\text{max}})$ where $r_{\text{max}} := (1/\alpha_{\text{min}} - 1)^2$.

Similar formulations for robustness around the empirical distribution with radius shrinking as r/n had been considered in (Ben-Tal et al., 2013; Lam & Zhou, 2015; Duchi & Namkoong, 2016). While there are many possible robustness balls $\mathcal{B}$ which could provide upper bounds on group risk, we opt to use the Chi-squared ball since it is straightforward to optimize (Ben-Tal et al., 2013; Namkoong & Duchi, 2016; 2017) and found it empirically outperformed other f-divergence balls.

In the sequel, we provide intuition for (4) and show that minimization can be performed efficiently.

4.2. Interpreting the dual

The dual of the maximization problem (4) provides additional intuition on the behavior of the robust risk. The following proposition was first proved by Ben-Tal et al. (2013) for finitely supported distributions.

Proposition 3 ((Duchi & Namkoong, 2018)). *If $\ell(\theta; \cdot)$ is upper semi-continuous for any θ, then for $r_{\text{max}} \geq 0$ and any θ, $\mathcal{R}_{\text{dro}}(\theta; r_{\text{max}})$ is equal to the following expression*

$$\inf_{\eta\in\mathbb{R}} \left\{ F(\theta; \eta) := C \left(\mathbb{E}_P \left[[\ell(\theta, Z) - \eta]_+^2 \right] \right)^{\frac{1}{2}} + \eta \right\} \quad (5)$$

where $C = \left(2(1/\alpha_{min} - 1)^2 + 1\right)^{1/2}$

Denoting by $\eta^\star$ the optimal dual variable (5), we see from the proposition that all examples suffering less than $\eta^\star$-levels of loss are completely ignored, and large losses above $\eta^\star$ are upweighted due to the squared term.

The squared term in the dual can be seen as a form of regularization of the original objective. However, unlike standard parameter regularization techniques, which encourage θ to be close to some point, our objective biases the model to have fewer high loss examples which matches our goal of mitigating representation disparity.

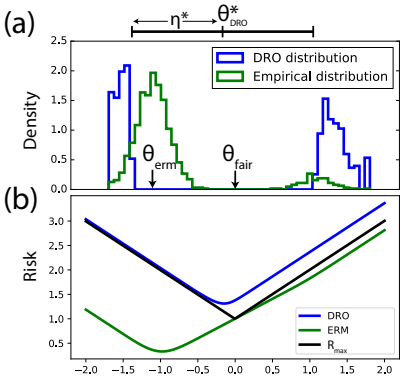


Figure 2. Chi-square distributionally robust optimization (DRO) regularizes the losses (top panel) such that the minimum loss estimate is fair to both groups (bottom panel).

Median Estimation: Recall the median estimation problem over two groups mentioned in Section 3.2 where the loss is $\ell(\theta; Z) = \|\theta - Z\|_1$. Figure 2 shows the behavior of both ERM and DRO on this median estimation task with unbalanced ($\alpha_{\text{min}} = 0.1$) groups. The parameter estimate which minimizes $\mathcal{R}_{\text{max}}$ for this problem is $\theta_{\text{fair}} = 0$ since this is equidistant from both groups. ERM on the other hand focuses entirely on the majority and returns $\theta_{\text{ERM}} \approx -1.0$.

DRO returns θ^*_{DRO} which is close to θ_{fair}. Analyzing the risk, we find that the single-step worst-case group risk $\mathcal{R}_{\text{max}}(\theta)$ in (1) is an upper bound on ERM, and DRO forms a tight upper bound this quantity (Figure 2b). We can also understand the behavior of DRO through the worst-case distribution Q in Equation 4. Figure 2a shows the worst-case distribution Q at the minimizer θ^*_{DRO} which completely removes points within distance η^*. Additionally, points far from θ^*_{DRO} are upweighted, resulting in a large contribution to the loss from the minority group.

Intuitively, we expect the DRO bound to be tight when losses are tightly clustered within a group regardless of θ – in this case thresholding by η^* roughly corresponds to essentially selecting some subset of groups and minimizing $\mathcal{R}_{\text{max}}(\theta)$, the worst-case group risk (1) directly.

On the other hand, the worst case for our approach is if α_{min}

is small, and a group with low expected loss has a high loss tail with population size $\alpha_{\min}$ regardless of θ. In this case DRO is a loose upper bound and optimizes the losses of the group with already low expected loss.

This is closely related to recent observations that the DRO bound can be loose for classification losses such as the zero-one loss due to the worst-case distribution consisting purely of misclassified examples (Hu et al., 2018). Even in this case, the estimated loss is still a valid upper bound on the worst case group risk, and as figure 2 shows, there are examples where the DRO estimate is nearly tight.

4.3. Optimization

We now show that minimizing $\theta \mapsto \mathcal{R}_{\mathrm{dro}}(\theta; r_{\max})$ can be done efficiently for a large class of problems. For models such as deep neural networks that rely on stochastic gradient descent, the dual objective $F(\theta; \eta)$ in (5) can be used directly since it only involves an expectation over the data generating distribution P.

Formally, the following procedure optimizes (4): for a given value of η, compute the approximate minimizer $\widehat{\theta}_\eta$

$$\underset{\theta \in \Theta}{\text{minimize}}\, \mathbb{E}_P \left[\ell(\theta; Z) - \eta\right]_+^2. \tag{6}$$

From Propositions 2 and 3, we have

$$\mathcal{R}_{\max}(\widehat{\theta}_\eta) \leq \mathcal{R}_{\mathrm{dro}}(\widehat{\theta}_\eta; r_{\max}) \leq F(\widehat{\theta}_\eta, \eta)$$

which implies that we can treat η as a hyperparameter and optimize over η using binary search. For convex losses $\theta \mapsto \ell(\theta; Z)$, the function $\eta \mapsto F(\widehat{\theta}_\eta, \eta)$ is convex and the binary search over η converges in linear time.

Alternatively, for models where there is a fast method for computing the minimizer $\theta^*(Q) \in \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_Q[\ell(\theta; Z)]$, we can use existing efficient primal solvers that computes the worst-case probability distribution $Q^*(\theta) \in \operatorname{argmax}_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]$ for a given θ based on projected gradient ascent on Q (Namkoong & Duchi, 2016). By alternating between optimization on θ and Q, we can efficiently find the saddle point (θ^*, Q^*) that satisfies $\theta^* = \theta^*(Q^*)$ and $Q^* = Q^*(\theta^*)$.

4.4. Stability of minority loss minimization

We have thus far demonstrated that for a single time step, the worst-case risk over all groups $\mathcal{R}_{\max}(\theta) = \max_k \mathcal{R}_k(\theta)$ can be controlled by the distributionally robust risk $\mathcal{R}_{\mathrm{dro}}(\theta; r_{\max})$ where $r_{\max} := (1/\alpha_{\min} - 1)^2$ and $\alpha_{\min}$ is the minority group proportion. Now, we study how the individual group risk $\mathcal{R}_k(\theta)$ affects user retention and hence *future* risk. By virtue of providing an upper bound to $\mathcal{R}_{\max}(\theta)$, optimizing $\mathcal{R}_{\mathrm{dro}}(\theta; r_{\max})$ at each time step can thus control the *future* group risk $\mathcal{R}_{\max}(\theta)$.

We show that if the initial group proportions satisfy $\alpha_k^{(0)} \geq \alpha_{\min}$ and the worst-case risk $\mathcal{R}_{\max}(\theta^{(t)})$ is sufficiently small at each time t, then we can ensure $\alpha_k^{(t+1)} > \alpha_{\min}$. Thus, to control $\mathcal{R}_{\max}^T$, the worst-case group risk over *all time steps*, it suffices to control $\mathcal{R}_{\mathrm{dro}}(\theta^{(t)}; r_{\max})$ using the procedure in Section 4.3.

Proposition 4. *Assume the retention model in Definition 1. Let* $\alpha_k^{(t)} > \alpha_{min}$, $\frac{b_k}{\sum_k b_k} > \alpha_{min}$, $n^{(t)} \leq \frac{\sum_k b_k}{1-\nu_{\max}}$, *and* $\nu(\mathcal{R}_k(\theta^{(t)})) < \nu_{\max}$. *Then, whenever we have*

$$\mathcal{R}_k(\theta^{(t)}) \leq \nu^{-1}\left(1 - \frac{(1-\nu_{max})b_k}{\alpha_{min}\sum_k b_k}\right),$$

$$\alpha_k^{(t+1)} = \frac{\lambda^{(t)}\alpha_k^{(t)}\nu(\mathcal{R}_k(\theta^{(t)})) + b_k}{\sum_l \lambda^{(t)}\alpha_l^{(t)}\nu(\mathcal{R}_l(\theta^{(t)})) + b_l} > \alpha_{min}.$$

We conclude that as long as we can guarantee

$$\mathcal{R}_{\mathrm{dro}}(\theta^{(t)}; r_{\max}) \leq \nu^{-1}\left(1 - \frac{(1-\nu_{max})b_k}{\alpha_{\min}\sum_k b_k}\right), \tag{7}$$

we can control $\mathcal{R}_{\max}^T(\theta^{(0)}, \ldots, \theta^{(T)})$, the unknown worst-case group risk over *all time steps* by optimizing $\mathcal{R}_{\mathrm{dro}}(\theta^{(t)}; r_{\max})$ at each step t. While the condition (7) is hard to verify in practice, we observe empirically in Section 5 that optimizing the distributionally robust risk $\mathcal{R}_{\mathrm{dro}}(\theta^{(t)}; r_{\max})$ at time step t indeed significantly reduces disparity amplification in comparison to using ERM.

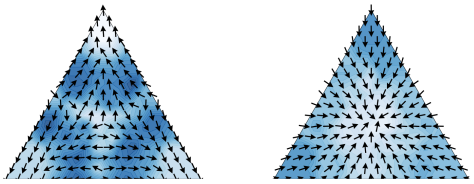
Proposition 4 gives stronger fairness guarantees than the stability conditions for ERM in Proposition 1. In ERM the best one can do is to add strong convexity to the model to stabilize to a possibly unfair fixed point. In contrast, Proposition 4 gives conditions for controlling $\mathcal{R}_{\max}$ over time without assumptions on the structure of fixed points.

Stability of median estimation: Returning to our running example of geometric median estimation, we can show that under the same dynamics, ERM is highly unstable while DRO is stable. Consider a three Gaussian mixture on the corners of the simplex, with L_2 loss, retention function $\nu(r) = \exp(-r)$, and $b_1 = b_2 = 50$, $n^{(t)} = 1000$. By construction, $(1/3, 1/3, 1/3)$ is the fair parameter estimate.

Figure 3 shows that ERM is highly unstable, with the only stable fixed points being the corners, where a single group dominates all others. The fair parameter estimate is an unstable fixed point for ERM, and any perturbation eventually results in a completely unfair parameter estimate. On the other hand, DRO has the reverse behavior, with the fair parameter estimate being the unique stable fixed point.

5. Experiments

We demonstrate the effectiveness of DRO on our motivating example (Figure 1) and human evaluation of a text autocom-



(a) ERM (b) DRO ($\alpha_k = 0.4$)

Figure 3. Dynamics of repeated median estimation - shading indicates velocity at each point. ERM results in unfair parameter estimates that favor one group. DRO is strongly stable, with an equal proportion groups being the unique stable equilibrium.

plete system on Amazon Mechanical Turk. In both cases DRO controls the worst-case risk $\mathcal{R}^T_{\max}$ over time steps and improves minority retention.

5.1. Simulated task

Recall the motivating example in Figure 1 which shows that logistic regression applied to a two-class classification problem is unstable, and becomes pathologically unfair.

The data is constructed by drawing from a mixture of two Gaussians centered at $(-1.5, 0)$ and $(0, 1.5)$. Each group is labeled according to linear decision boundary $(-3/2, \sqrt{3^2 - 1}/3)$ and $(3/2, \sqrt{3^2 - 1}/3)$ such that classifying with $x_2 > 0$ is accurate, but the optimal linear classifier on one group achieves 50% accuracy on the other.

At each round we fit a logistic regression classifier using ERM or DRO, fixing the margin distance to 1. Our dynamics follow definition 1 with $\nu(x) = 1 - x$, $\mathcal{R}$ the zero-one loss, and $b_k = 1000$. The DRO model is trained using the dual objective with logistic loss, and $\eta = 0.95$ which was the optimal dual solution to $\alpha_{\min} = 0.2$. The results do not qualitatively change for choices of $\alpha_{\min} < 0.5$, and we show that we obtain control even for group sizes substantially smaller than 0.2 (Figure 6).

Figure 5 shows that ERM is unstable and the minority group rapidly loses accuracy beyond $t = 300$, with most runs resulting in substantially lower accuracy for the minority group by iteration 500. On the other hand, DRO is completely stable, and maintains 0.8 accuracy.

This stability is due to the fact that the regularized loss for DRO prevents small losses in the minority fraction from amplifying, as we discuss in Proposition 4. Even when the minority fraction becomes as low as 1%, the DRO loss ensures that the accuracy of this minority fraction is high, with at least 75% accuracy (Figure 6).

5.2. Autocomplete task

We now present a real-world, human evaluation of user retention and satisfaction on a text autocomplete task. The task consists of the prediction of next words in a corpus of tweets built from two estimated demographic groups, African Americans and White Americans, from Blodgett et al. (2016). There are several distinguishing linguistic patterns between tweets from these groups, whose language dialects we henceforth refer to as African-American English (AAE) and Standard-American English (SAE), respectively following the nomenclature in Blodgett et al. (2016). Our overall experimental design is to measure the retention rate ν and risk $\mathcal{R}$ for various choices of demographic proportions $(\alpha_{\text{AAE}}, \alpha_{\text{SAE}})$ and simulate the implied dynamics, as running a fully on-line experiment would be prohibitively expensive.

For both ERM and DRO, we train a set of five maximum likelihood bigram language models on a corpus with 366,361 tweets total and a $(0.1, 0.4, 0.5, 0.6, 0.9)$ fraction of the tweets labeled as AAE tweets. This results in 10 possible autocomplete systems a given Mechanical Turk user can be assigned during the task.

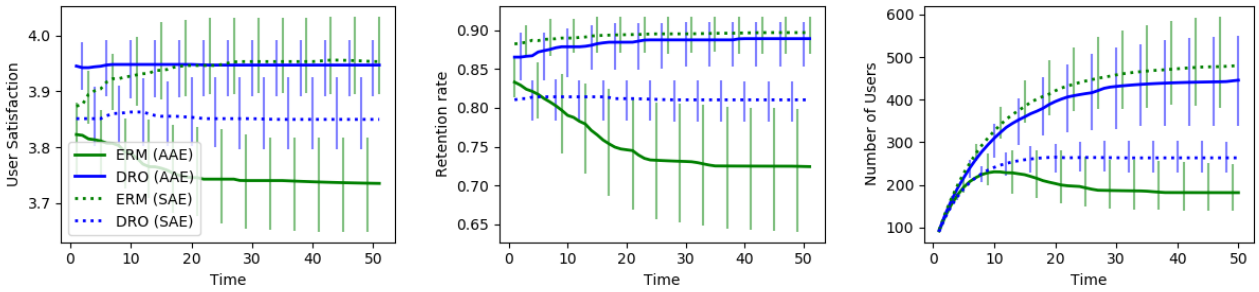
To evaluate the retention and loss for AAE and SAE separately, a user is assigned 10 tweets from either the held out AAE tweets or SAE tweets, which they must replicate using a web-based keyboard augmented by the autocomplete system. We induce a user's demographic by assigning them to one of these two held out set types. Details of the autocomplete task are included in the supplement.

After completing the task, users were asked to fill out a survey which included a rank from 1 to 5 on their satisfaction with the task, and a yes/no question asking whether they would continue to use such a system. We assign 50 users to one of these two held out set types, and one of the 10 autocomplete models, resulting in 1,000 users' feedback across autocomplete models and induced demographics.

The response to whether a user would continue to use the autocomplete system provides samples $\nu(\mathcal{R}_K(\alpha))$ with $n = 366361$ and the five values of demographic proportions α. The user satisfaction survey provides a surrogate for $\mathcal{R}_K(\alpha)$ at these same points. We interpolate ν and $\mathcal{R}_K$ to $\alpha \in [0, 1]$ via isotone regression which then allows us to simulate the user dynamics and satisfaction over time using definition 1. We estimate variability in these estimates via bootstrap replicates on the survey responses.

Our results in Figure 4 show an improvement in both minority satisfaction and retention rate: we improve the median user satisfaction from 3.7 to 4.0 and retention from 0.7 to 0.85, while only slightly decreasing the majority satisfaction and retention. Implied user counts follow the same trend with larger differences between groups due to compounding.

Counterintuitively, the minority group has higher satisfaction and retention under DRO. Analysis of long-form comments from Turkers suggest this is likely due to users valuing the model's ability to complete slang more highly than com-



(a) User satisfaction (b) User retention (c) User count

Figure 4. Inferred dynamics from a Mechanical Turk based evaluation of Autocomplete systems. DRO increases minority user satisfaction (panel a) and retention (panel b) leading to a corresponding increase in user count (panel c). Error bars indicates bootstrap quartiles.

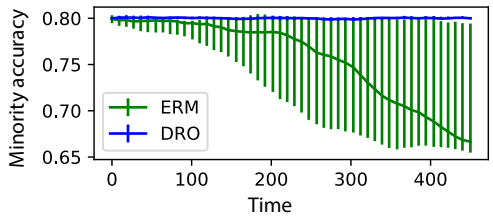


Figure 5. Disparity amplification in Figure 1 is corrected by DRO. Error bars indicate quartiles over 10 replicates.

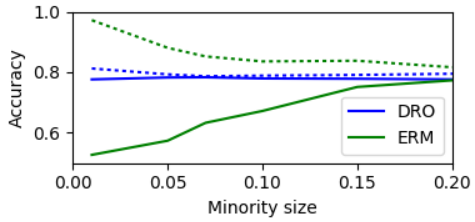


Figure 6. Classifier accuracy as a function of group imbalance. Dotted lines show accuracy on majority group.

pletion of common words and indicates a slight mismatch between our training loss and human satisfaction with an autocomplete system. This gap suggests that well-calibrated losses are critical for real-world applications of fairness through loss minimization.

6. Discussion

In this work we argued for a view of loss minimization as a distributive justice problem, and showed that ERM often results in disparity amplification and unfairness. We demonstrate that DRO provides a upper bound on the risk incurred by minority groups, where this bound is tight on simulations and empirically effective on an Autocomplete task. Our proposed algorithm is straightforward to implement, and induces distributional robustness, which can be viewed as a benefit in and of itself.

Our arguments against ERM and in favor of minority risk minimization closely mirror Rawls' arguments against utilitarianism, and thus inherit the critiques of Rawlsian distributive justice. Examples of such critiques are the focus on an abstract worst-off group rather than demographic groups or individuals (Altham, 1973), extreme risk-aversion (Mueller et al., 1974), and utilitarianism with diminishing returns as an alternative (Harsanyi, 1975). In this work, we do not discuss the normative question of whether Rawlsian justice is correct as many of these critiques have been discussed in earlier work (Rawls, 2001). It is an open question whether there are better frameworks for distributive justice in the context of machine learning.

There are two large open questions from our work. First, as fairness is fundamentally a causal question, observational approaches such as DRO can only hope to control limited aspects of fairness. The generality with which our algorithm can be applied also limits its ability to enforce fairness as a constraint, and thus our approach here is unsuitable for high-stakes fairness applications such as classifiers for loans, criminality, or admissions. In such problems the implied minorities from DRO may differ from well-specified demographic groups who are known to suffer from historical and societal biases. This gap arises due to looseness in the DRO bound (Hu et al., 2018), and could be tightened using smoothness assumptions (Dwork et al., 2012).

Second, distributional robustness proposed here runs counter to classical robust estimation for rejecting outlier samples, as high loss groups created by an adversary can easily resemble a minority group. Adversarial or high-noise settings loosen the DRO upper bound substantially, and it is an open question whether it is possible to design algorithms which are both fair to unknown latent groups and robust.

Reproducibility: Code to generate results available at `https://bit.ly/2sFkDpE`.

Acknowledgements: This work was funded by the Open Philanthropy Project Award.

References

Altham, J. J. Rawls' difference principle. *Philosophy*, 48: 75–78, 1973.

Amodei, D. et al. Deep speech 2 end to end speech recognition in English and mandarin. In *International Conference on Machine Learning (ICML)*, pp. 173–182, 2016.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *104 California Law Review*, 3:671–732, 2016.

Ben-Tal, A., den Hertog, D., Waegenaere, A. D., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of African-American English. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1119–1130, 2016.

Chouldechova, A. A study of bias in recidivism prediciton instruments. *Big Data*, pp. 153–163, 2017.

Duchi, J. C. and Namkoong, H. Variance-based regularization with convex objectives. *arXiv:1610.02581 [stat.ML]*, 2016.

Duchi, J. C. and Namkoong, H. Distributionally robust stochastic optimization: Minimax rates and asymptotics. Working Paper, 2018.

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv:1610.03425 [stat.ML]*, 2016. URL `https://arxiv.org/abs/1610.03425`.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pp. 214–226, 2012.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 259–268, 2015.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. Technical report, CEPR Discussion Papers, 2017.

Grother, P. J., Quinn, G. W., and Phillips, P. J. Report on the evaluation of 2d still-image face recognition algorithms. Technical report, NIST, 2011.

Hardt, M., Price, E., and Srebo, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3315–3323, 2016.

Harsanyi, J. C. Can the maximin principle serve as a basis for morality? a critique of john rawls's theory. *The American Political Science Review*, 69:594–606, 1975.

Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

Hovy, D. and Sgaard, A. Tagging performance correlates with age. In *Association for Computational Linguistics (ACL)*, pp. 483–488, 2015.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, 2018.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 1617–1626, 2017.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. Rawlsian fairness for machine learning. In *FATML*, 2016.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. Incorporating dialectal variability for socially equitable language identification. In *Association for Computational Linguistics (ACL)*, pp. 51–57, 2017.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2018.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS)*, 2017.

Lam, H. and Zhou, E. Quantifying input uncertainty in stochastic optimization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

Luo, A. C. *Regularity and complexity in dynamical systems*. Springer, 2012.

Mueller, D. C., Tollison, R. D., and Willet, T. D. The utilitarian contract: A generalization of rawls' theory of justice. *Theory and Decision*, 4:345–367, 1974.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.

Namkoong, H. and Duchi, J. C. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, 2017.

Rawls, J. *Justice as fairness: a restatement*. Harvard University Press, 2001.

Rawls, J. *A theory of justice: Revised edition*. Harvard University Press, 2009.

Sapiezynski, P., Kassarnig, V., Wilson, C., Lehmann, S., and Mislove, A. Academic performance prediction in a gender-imbalanced environment. In *FATREC*, volume 1, pp. 48–51, 2017.

Tatman, R. Gender and dialect bias in youtubes automatic captions. In *Workshop on Ethics in Natural Langauge Processing*, volume 1, pp. 53–59, 2017.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, pp. 1920–1953, 2017.