On Distributionally Robust Optimization and Data Rebalancing

Agnieszka Słowik
Dep. of Computer Science and Technology
University of Cambridge
Cambridge, UK

Léon Bottou
Facebook AI Research, New York, NY, USA,
and New York University, New York, NY, USA

Abstract

Machine learning systems based on minimizing average error have been shown to perform inconsistently across notable subsets of the data, which is not exposed by a low average error for the entire dataset. Distributionally Robust Optimization (DRO) seemingly addresses this problem by minimizing the worst expected risk across subpopulations. We establish theoretical results that clarify the relation between DRO and the optimization of the same loss averaged on an adequately weighted training dataset. The results cover finite and infinite number of training distributions, as well as convex and non-convex loss functions. An implication of our results is that for each DRO problem there exists a data distribution such that learning this distribution is equivalent to solving the DRO problem. Yet, important problems that DRO seeks to address (for instance, adversarial robustness and fighting bias) cannot be reduced to finding the one 'unbiased' dataset. Our discussion section addresses this important discrepancy.

1 INTRODUCTION

Distributionally Robust Optimization (DRO) (Ben-Tal et al., 2009; Levy et al., 2020; Sagawa et al., 2020a; Liu et al., 2021; Zhen et al., 2021) aims to make machine learning systems more robust by replacing the optimization of a single expected error criterion by the simultaneous optimization of the expected errors with respect to a predefined family $\mathcal{Q}$ of distributions.

In this work, we explore the relation between solving a DRO problem and optimizing the expected error for a single distribution constructed by mixing distributions from the family $\mathcal{Q}$, with a focus on the nonconvex deep learning systems that are increasingly used for important real tasks. We also discuss the consequences of these findings for two application domains of DRO, namely fighting bias in machine learning systems and achieving robustness against adversarial examples. We find that DRO does not really solve the problem at hand but merely displaces important aspects into the precise formulation of the DRO problem, such as the choice of *calibration coefficients*.

We first present our theoretical results (Section 2), and discuss their loose ends (Section 3). We then discuss the consequences of these results in practical applications of DRO, that is when using DRO to address representation disparity (Section 4.1) and in adversarial robustness (Section 4.2). In the context of mitigating the majority bias, we provide simple recommendations based on our theoretical findings. We give a brief overview of related work (Section 5) on DRO and mitigating bias.

2 THEORETICAL RESULTS

2.1 Setup

Let $\ell(z, w)$ be the loss of a machine learning model where $w \in \mathbb{R}^d$ represent the parameters of the model and $z \in \mathbb{R}^n$ belongs to the space of examples. For instance, the examples z may be pairs (x, y) and the loss may be the squared loss $\ell(z, w) = \frac{1}{2}\|y - f_w(x)\|^2$.

Distributionally Robust Optimization Instead of assuming the existence of a probability distribution $P(z)$ over the examples z and formulating an Expected Risk Minimization (ERM) problem:

$$\min_w \quad \left\{ \; C_P(w) \overset{\Delta}{=} E_{z\sim P}[\ell(z, w)] \; \right\}, \qquad (1)$$

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

DRO considers a family $\mathcal{Q}$ of distributions and seeks to minimize

$$\min_w \left\{ C_{\mathcal{Q}}(w) \triangleq \max_{P\in\mathcal{Q}} C_P(w) \right\}. \qquad (2)$$

Many authors define $\mathcal{Q}$ with the purpose of constructing a learning algorithm with additional robustness properties. For instance, $\mathcal{Q}$ may be the set of all distributions located within a certain neighborhood of the training distribution (Bagnell, 2005; Namkoong and Duchi, 2016; Blanchet et al., 2019; Staib and Jegelka, 2019). Different ways to define this neighborhood lead to different and sometimes surprising solutions (e.g., Hu et al., 2018). Interesting theoretical possibilities appear when $\mathcal{Q}$ also contains the discrete distributions that represent finite training sets. Besides these theoretically justified choices of $\mathcal{Q}$, many practical concerns can be viewed through the prism of DRO on ad-hoc families $\mathcal{Q}$ of distributions.

Example 1 (Fighting bias)**.** Let the example distributions P_1 to P_K represent identified subpopulations for which we want to ensure consistent error rates. This can be achieved by minimizing the worst error, that is, formulating a DRO problem with $\mathcal{Q} = \{P_1 \dots P_K\}$ (see Section 4.1 for a discussion). Although this formulation is far too simple to address all forms of biases, it illustrates how DRO provides means to move away from focusing on a single optimization objective.

Example 2 (Fighting adversarial attacks)**.** Szegedy et al. (2014) have shown that one can almost arbitrarily change the output of a deep learning vision system by modifying the patterns in nearly invisible ways. Let Φ be the set of all measurable functions φ that map an example pattern z to another pattern $\varphi(z)$ that is assumed visually indistinguishable from z according to a certain psycho-visual criterion. Let P_φ represent the distribution followed by $\varphi(z)$ when z follows the distribution P. Robust solutions against the class of adversarial perturbation Φ can be found with DRO with the distribution family $\mathcal{Q} = \{P_\varphi : \varphi \in \Phi\}$ (see Section 4.2 for a discussion).

Calibration coefficients The simple DRO formulation makes sense when we know that all distributions define problems of comparable difficulty. It is however easy to imagine that a particular distribution emphasises harder examples. We can introduce calibration terms r_P in the DRO formulation to prevent any single distribution P to dominate the maximum

$$\min_w \quad \left\{ \quad C_{\mathcal{Q},r}(w) \;\triangleq\; \max_{P\in\mathcal{Q}}\{C_P(w) - r_P\} \quad \right\}. \quad (3)$$

Correctly setting the calibration terms is both difficult and application-specific. A simple but costly approach consists in letting r_P be equal to the optimum cost for that distribution alone, $r_P = \min_w C_P(w)$. Calibrated DRO (3) then controls the loss of performance incurred by seeking a solution that works for all distributions as opposed to solutions that are specific to each distribution. Another approach (Meinshausen and Bühlmann, 2015) relies instead on the variance of the predicted quantity.

Calibration terms can also be used to counter the effect of finite training data. For instance, when we only have n examples for a certain distribution $P \in \mathcal{Q}$, the expected risk $C_P(w)$ can be replaced by its empirical estimate $C_{P_n}(w)$ augmented with a calibration constant that decreases when the number n of training examples increases (Sagawa et al., 2020a).

2.2 A local minimum of a DRO problem is a stationary point of an expected loss mixture

Finite case We first address the case where $\mathcal{Q}$ is a finite set of distributions $P_1 \dots P_K$. The following result simplifies Proposition 2 of Arjovsky et al. (2019) by eliminating the KKT constraint qualification requirement. In the rest of this work, we always assume that the mixture coefficients λ_k are nonnegative and sum to one. We also assume that the parameters w are real numbers, hence, the set of possible parameters is convex.

Theorem 1. *Let $\mathcal{Q} = \{P_1, \dots, P_K\}$ be a finite set of probability distributions on $\mathbb{R}^n$ and let w^* be a local minimum of the DRO problem* (2) *or the calibrated DRO problem* (3)*. Let the cost $C_P(w)$ for any distribution P on $\mathbb{R}^n$ be defined as $C_P(w) = \mathbb{E}_{z\sim P}[\ell(z,w)]$. Let $C_P(w)$ be differentiable in w for all $P \in \mathcal{Q}$. Then there exists a mixture distribution $P_{\text{mix}} = \sum_k \lambda_k P_k$ such that $\nabla C_{P_{\text{mix}}}(w^*) = 0$.*

The proof relies on a simple hyperplane separation lemma closely related to Farkas' lemma (Boyd and Vandenberghe, 2014, Sec.2.5 and Ex.2.20).

Lemma 1. *A nonempty closed convex subset A of $\mathbb{R}^n$ either contains the origin or is strictly separated from the origin by a certain hyperplane, that is, there exists a vector $u \in \mathbb{R}^n$ and a scalar $c > 0$ such that, for all $x \in A$, $\langle u, x\rangle \geq c$.*

Proof. Assume $0 \notin A$. Let $u \in A$ be the projection of the origin onto the closed convex set A. For all $x \in A$ and all $t \in [0,1]$, the point $u + t(x-u)$ also belongs to the convex set A. Since u is the point of A closest to the origin, for all $t \in [0,1]$,

$$\begin{aligned} r(t) &= \|u + t(x-u)\|^2 \\ &= \|u\|^2 + 2t\,\langle u, x-u\rangle + t^2\|x-u\|^2 \geq \|u\|^2. \end{aligned}$$

Therefore $r'(0) = 2\langle u, x-u\rangle \geq 0$, and, as a consequence, $\langle u,x\rangle \geq \langle u,u\rangle > 0$. □

Proof of Theorem 1. Let $A \subset \mathbb{R}^n$ be the convex hull of the $g_k{=}\nabla C_{P_k}(w^*)$ for $k = 1\dots K$. A is closed and convex. If A does not contain the origin, according to the lemma, there exist u and c such that $\forall\ x \in A$, $\langle u,x\rangle \geq c > 0$. Therefore, for all $t > 0$, moving from w^* to $w^* - tu$ reduces all costs C_{P_k} by at least $tc + o(t)$. As a consequence, $\max_k C_{P_k}$ is also reduced by at least $tc + o(t)$, contradicting the assumption that w^* is a local minimum. Hence A contains the origin, i.e. there are positive mixture coefficients λ_k summing to one such that $\sum_k \lambda_k \nabla C_{P_k}(w) = \nabla_w C_{P_{\mathrm{mix}}}(w) = 0$. □

All local and global solutions of the DRO problem (2) or (3) are therefore stationary points of the expected risk (1) associated with a mixture of the distributions of $\mathcal{Q}$. The exact mixture coefficients depend on the loss functions, the distributions included in $\mathcal{Q}$ and, in the case of the calibrated version of DRO, on the calibration constants r_P.

This result raises several important questions. Is this result valid when $\mathcal{Q}$ is not finite? Are these stationary points always local minima? Is the converse true? What is the relation between the mixture coefficients λ_k and the calibration constants r_P? How far can such results go without assuming convex losses? These questions will be discussed in the rest of this paper.

Infinite case The infinite case differs because the convex hull of an infinite set of vectors is not necessarily closed, even when the original set is closed. Therefore, we cannot directly apply Lemma 1 to the convex hull A of the gradients $g_P = \nabla C_P(w^*)$ for all $P \in \mathcal{Q}$. Applying it instead to the closure $\bar{A}$ of A yields a substantially weaker result: if w^* is a local DRO minimum, then for each $\varepsilon > 0$, there is a mixture $P^{(\varepsilon)}_{\mathrm{mix}}$ of distributions from $\mathcal{Q}$ such that $\|\nabla C_{P^{(\varepsilon)}_{\mathrm{mix}}}\| \leq \varepsilon$.

There is no guarantee that $P^{(\varepsilon)}_{\mathrm{mix}}$ converges to an actual distribution when ε converges to zero.[1] Therefore this weaker result does not help relating the solution of a DRO problem with the solutions of an ERM problem for a suitable training distribution. However, this stronger result can be obtained at the price of a *tightness* assumption (Billingsley, 1999).

Definition 1. A family of distributions $\mathcal{Q}$ on a Polish space[2] Ω is *tight* when, for any $\epsilon > 0$, there is a compact subset $K \subset \Omega$ such that $\forall P \in \mathcal{Q}$, $P(K) \geq 1 - \epsilon$.

Tightness is obvious when all the examples belong to a bounded domain. Even when this is not the case, it is known that any finite set of probability distributions on a Polish space is tight (Billingsley, 1999). This often provides the means to prove the tightness of an infinite family $\mathcal{Q}$ of distributions that are "close" enough to a single distribution such as the training data distribution. For instance, in the case of adversarial examples (Example 2), tightness is doubly obvious, first because all images belong to a bounded domain, second because the visual similarity criterion ensures that the distance between z and $\varphi(z)$ is bounded.

Theorem 2. *Let $\mathcal{Q}$ be a tight family of probability distributions on $\mathbb{R}^n$. Let w^* be a local minimum of problem (3). Let $\mathcal{Q}_{\mathrm{mix}}$ be the weak convergence closure of the convex hull of $\mathcal{Q}$. Let there be a bounded continuous function $h(z,w)$ defined on a neighborhood $\mathcal{V}$ of w^* such that $\nabla C_P(w) = \mathbb{E}_{z\sim P}[h(z,w)]$ for all $P \in \mathcal{Q}_{\mathrm{mix}}$ and such that $\|h(z,w) - h(z,w')\| \leq M\|w - w'\|$ for almost all $z \in \mathbb{R}^n$. Then $\mathcal{Q}_{\mathrm{mix}}$ contains a distribution P_{mix} such that $\nabla_w C_{P_{\mathrm{mix}}}(w^*) = 0$.*

This theorem does not require the loss $\ell(z,w)$ to be differentiable everywhere as long as the purported derivative $h(z,w)$ has the correct expectation (Bottou et al., 2018). For our purposes, it must also be bounded, continuous on $\mathcal{V}$, and satisfy a Lipschitz continuity requirement.

Proof. Let $\bar{A}$ be the closure of the convex hull of the $g_P = \nabla C_P(w^*)$ for all $P \in \mathcal{Q}$. According to Lemma 1, if $\bar{A}$ does not contain the origin, then there are u and $c > 0$ such that $\forall x \in A, \langle u,x\rangle > c$. In particular, for all $P \in \mathcal{Q}$, we have $\langle u, \nabla C_P(w^*)\rangle > c > 0$. Thanks to the Lipschitz continuity of $h(z,w)$, we have $C_P(w^* - tu) < C_P(w^*) - tc + Mt^2$ for all $P \in \mathcal{Q}$. Therefore for any $0 < t < c/2M$ and any $P \in \mathcal{Q}$, we have $C_P(w^* - tu) < C_P(w^*) - tc/2$ contradicting the assumption that w^* is a local DRO miminum. Therefore $\bar{A}$ contains the origin. This means that for any $t > 0$, there exists a mixture $P^{(1/t)}_{\mathrm{mix}}$ of distributions from $\mathcal{Q}$ such that $\|\nabla C_{P^{(\varepsilon)}_{\mathrm{mix}}}(w^*)\| < 1/t$. Note that if $\mathcal{Q}$ is tight, the convex hull of $\mathcal{Q}$ is also tight. Therefore the sequence $P^{(1/t)}_{\mathrm{mix}}$ is also tight, and, by Prokhorov's theorem, contains a weakly convergent subsequence whose limit P_{mix} belongs to the closure $\mathcal{Q}_{\mathrm{mix}}$ of the convex hull of $\mathcal{Q}$. Because $h(z,w^*)$ is continuous and bounded, the map $P \mapsto \nabla C_P(w^*)$ is continuous for the weak topology. Therefore $\nabla C_{P_{\mathrm{mix}}}(w^*) = 0$. □

[1] Suppose for instance that $\mathcal{Q}$ contains all Gaussians with unit variance with arbitrary means in $\mathbb{R}$. For any $t > 0$, let $P^{(1/t)}_{\mathrm{mix}}$ be the equal mixture of t^2 equally spaced Gaussians in interval $[-t,+t]$. Neither this sequence not any of its subsequences converge to a distribution because there is no such thing as a uniform distribution on all of $\mathbb{R}$.

[2] For our purposes, it is sufficient to know that $\mathbb{R}^n$ is a Polish space!

2.3 A local minimum of an expected loss mixture is a local minimum of a calibrated DRO problem

The following elementary result states that if w^* is a local minimum of an expected cost mixture C_{P_mix}, then it also is a local minimum of the calibrated DRO problem (3) with calibration constants r_P equal to the costs $C_P(w^*)$.

Theorem 3 (Converse)**.** *Let $P_\text{mix} = \sum_k \lambda_k P_k$ be an arbitrary mixture of distributions $P_k \in \mathcal{Q}$. If w^* is a local minimum of C_{P_mix}, then w^* is a local minimum of the calibrated DRO problem* (3) *with calibration coefficients $r_P = C_P(w^*)$.*

Proof. By contradiction, assume that w^* is not a local minimum of (3), that is, for all $\epsilon > 0$ there exists u such that $\|u\| < \epsilon$ and $\max_{P\in\mathcal{Q}} \{C_P(w^*+u) - r_P\} < \max_{P\in\mathcal{Q}} \{C_P(w^*) - r_P\}$. Recalling our choice of r_P yields $\max_{P\in\mathcal{Q}} \{C_P(w^*+u) - C_P(w^*)\} < 0$. Since $C_P(w * +u) < C_P(w^*)$ for all $P \in \mathcal{Q}$, $C_{P_\text{mix}}(w^* + u) < C_{P_\text{mix}}(w^*)$, and w^* cannot be a local minimum of C_{P_mix}. □

3 DISCUSSION

3.1 Convex loss

Note the slight discrepancy between the statements of Theorem 3 and Theorems 1–2. The former requires a local minimum of the expected loss mixture, whereas the latter only provides a stationary point.

This distinction is of course moot when the loss functions $\ell(z, w)$ is convex in w because convexity makes all stationary points global minima as well.[3] Theorems 1 and 3 then provide an exact equivalence between finding a minimum of the calibrated DRO problem (3) and finding a minimum of a well-chosen expected loss mixture.

When $\mathcal{Q}$ is finite, this equivalence is a natural consequence of convex duality theory (Bertsekas, 2009) because we can restate the DRO problem as a convex optimization problem using a slack variable L,

$$\min_{w,L} L \quad \text{s.t.} \quad \forall P \in \mathcal{Q}. \;\; C_P(w) - r_P - L \leq 0 \; . \tag{4}$$

Theorem 2 shows that this equivalence still holds when $\mathcal{Q}$ is infinite and satisfies a tightness assumption.

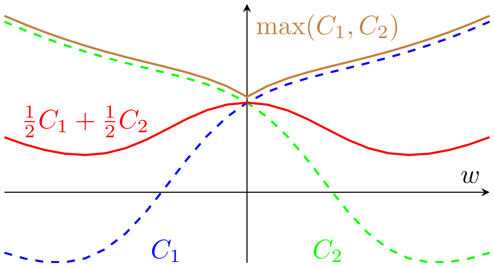

Figure 1: The minimum of $\max\{C_1(w), C_2(w)\}$ (which is $w^* = 0$) is a stationary point of $C_\text{mix}(w) = \frac{1}{2}C_1(w)+\frac{1}{2}C_2(w)$. However, this stationary point is not a local minimum but a local maximum of the mixture cost.

3.2 Nonconvex loss

The nonconvex case is more challenging because the stationary points identified by Theorem 1 need not be local minima. Consider for instance the two real functions

$$C_1(w) = \tanh(1 + w) + \epsilon w^2$$
$$C_2(w) = \tanh(1 - w) + \epsilon w^2$$

where the term ϵw^2 with $0 < \epsilon \ll 1$ is only present to ensure that each of these functions has a well defined optimum. As shown in Figure 1, their maximum $\max\{C_1(w), C_2(w)\}$ has a a minimum in $w^* = 0$. As predicted by Theorem 1, this solution is a stationary point of the the mixture $C_\text{mix} = \frac{1}{2}C_1(w) + \frac{1}{2}C_2(w)$. However, this stationary point is not a local minimum but a local *maximum*.

In this counter-example, the solution $w^* = 0$ of the DRO problem $\min_w \max\{C_1(w), C_2(w)\}$ falls in negative curvature regions of the functions C_1 and C_2. As a result any mixture of these two costs also has negative curvature in w^*. Therefore, w^* cannot be a local minimum.

Conversely, when a local optimum of the DRO problem is achieved in a point where the individual cost functions have positive curvature, all mixtures must also have positive curvature, and the stationary point must be a local minima. This remark is important in practice because learning algorithms for deep learning problems tend to follow trajectories where the Hessian is very flat apart from a few positive eigenvalues (Sagun et al., 2018). Although more evidence and a formal framework are needed to make a definitive

[3] Convexity also provides easy means to weaken the differentiability assumption because of the existence of subgradients. One could similarly weaken the differentiability assumptions of Theorems 1-2 by assuming instead the existence of local sub– and super–gradients.

```
Algorithm 1: A Lagrangian DRO algorithm
  Input: Equal size training sets $D_k$. $k = 1 \dots K$
  Input: Calibration coeffs $r_k$. Initial $w_0$.
  Input: Temperature $\beta$. Stopping threshold $\epsilon$.
  Output: A sequence of weights $w_t$.
  $t \leftarrow K$
  $\lambda_k \leftarrow 1/t \ \ \forall k$
  repeat
      $w_{t+1} \leftarrow \texttt{Descend}\big(w_t, \{D_1 \star \lambda_1 \dots D_K \star \lambda_K\}\big)$
      $c_k \leftarrow \texttt{Cost}\big(w_{t+1}, \{D_k\}\big) \ \ \forall k$
      $\delta_k \leftarrow \frac{1}{Z} \exp(\beta(c_k - r_k)) \ \ \forall k$
          — with $Z$ such that $\sum_k \delta_k = 1$
      $\lambda_k \leftarrow \frac{1}{t+1}(t\lambda_k + \delta_k) \ \ \forall k$
      $t \leftarrow t + 1$
  until $\max_k |\lambda_k - \delta_k| < t\epsilon$
  return $w_t$
```

statement, this fact suggests that the issue presented in Figure 1 is often cured by overparametrization.

3.3 Lagrangian algorithms for DRO

When the loss function is convex, duality theory suggests to write the Lagrangian of problem (4) and solve instead a dual problem (Boyd and Vandenberghe, 2014; Bertsekas, 2009). Algorithm 1 is a typical example of this strategy.

Although such Lagrangian DRO algorithms are justified by convexity considerations, they are also widely used with deep learning system with nonconvex objectives (Sagawa et al., 2020a; Augustin et al., 2020). Our theoretical results provide partial support to this practice. In Appendix, we discuss failure modes that prevent the algorithm from finding DRO solutions that are not associated with a local minimum of an adequate loss mixture, including the scenario most relevant to overparameterized models (Appendix A).

3.4 Implications for overparameterized models

Overparametrization, on the other hand, dilutes the practical meaning of DRO or dataset rebalancing. When the optimization achieves near zero loss on all training examples, the expected losses for all subpopulations are near zero regardless of DRO or rebalancing techniques. What matters is now the implicit or explicit regularization that selects which of the many possible solutions achieve near-zero loss for all examples.

For instance, Byrd and Lipton (2019) find that importance sampling does not improve the average test error. On the other hand, Sagawa et al. (2020b) finds that the worst group error can be worse in overparametrized networks. Sagawa et al. (2020a) stress the impact of regularization when using DRO in overparametrized network, and Lopez-Paz (2021) finds that simple rebalancing techniques in overparametrized networks can improve the worst group as much as Sagawa's regularized group DRO.

Our theorems provide a satisfactory explanation of these facts when, instead of viewing the regularization as additional cost penalties, one views regularization as data augmentation, that is, replacing each training example by a local distribution centered on the training example (Leen, 1995). Since the effect of the regularization is then expressed by the expected losses $C_{P_k}(w)$, both DRO and rebalancing become meaningful and practically equivalent objectives.

4 PRACTICAL IMPLICATIONS

Up to nonconvex effects that we believe disappear with overparametrization (as discussed in Section 3.4), the theoretical results suggest that DRO is practically equivalent to training on a well chosen example distribution. Does this mean that it would be enough to acquire the true unbiased training data? Wasn't DRO supposed to provide a template to move away from optimizing a single averaged criterion? The theoretical results make it clear that this well chosen example distribution is far from universal, but depends on often overlooked assumptions hidden in the DRO problem statement, such as calibration coefficients. We now illustrate this assertion on two practical problems: addressing the majority bias (Section 4.1) and DRO in adversarial robustness (Section 4.2).

4.1 Fighting bias

Bias issues in machine learning take many forms that cannot be reduced to mere differences in error rates. However, these issue often arise because the training algorithm optimizes a single performance criterion (Barocas et al., 2019, p218). Minimizing the worst error measured on various subpopulations therefore can be used as a template to understand how DRO can or cannot be used to address bias. Following Example 1, we formulate the minimization of the worst error as the following DRO instance,

$$\min_w \left\{ \max_{P_k \in \mathcal{Q}} C_{P_k}(w) - r_{P_k} \right\} ,$$

where the P_k represent the distributions associated with K subpopulations of interest and the r_{P_k} are calibration coefficients. Although they are rarely made

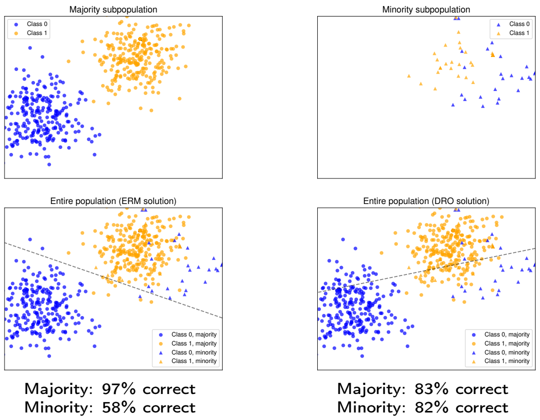

Figure 2: Illustration of the majority bias when using Expected Risk Minimization (ERM) in a linear binary classification problem, where the dataset can be partitioned to a *majority* and a *minority* subpopulation. The DRO solution was obtained using the Lagrangian algorithm (Algorithm 1), both solutions use the same linear SVM model.

explicit, there are legitimate reasons to introduce such calibration coefficients in bias figthing problems.

- Calibration coefficients can be used to favor certain subpopulations in order to counter *representation disparity* (which leads to the majority bias when optimizing for average performance, e.g. Figure 2) or *disparity amplification*. Representation disparity refers to the pheonomenon of achieving a high overall accuracy but low minority accuracy (Buolamwini and Gebru, 2018; Khan and Fu, 2021). For instance, ubiquitous speech recognition systems, such as voice assistants, struggle with accents and dialects (Behravan et al., 2016; Yang et al., 2018; Najafian and Russell, 2020). A minority user becomes discouraged by the poor performance of such system, which leads to disparity amplification over time due to the increasing gap between quantity of data provided by active users (majority groups, favored by the system from the beginning) and groups that experienced poor performance due to the initial representation disparity.

- Calibration coefficients can also be used to account for justifiable differences in difficulty across distributions. For instance, it might be known that one of the training distribution represents examples collected with a deficient method, such as, bad cameras, bad conditions, etc. Because the task is more difficult due to data limitations, the cost $C_P(w)$ for such a distribution will systematically be higher than for other distributions. The simple DRO formulation (Equation (2)) then amounts to optimizing only for this distribution. As a consequence, small gains for the deficient distribution will be obtained at the expense of a massive performance degradation for all other distributions, essentially making it as bad as the performance for the deficient distribution.

The results presented in Section 2 and the discussion presented in Section 3 suggest that, for all practical purposes, this is equivalent to minimizing the expected risk for a suitable mixture of the P_k distributions. However, this mixture is not universal but depends critically on the calibration coefficients r_{P_k}. In fact, specifying a set of calibration constants amounts to describing what we consider to be an *acceptable* outcome (acceptable subpopulation performance) for the original bias fighting problem. What is acceptable or not is obviously problem-dependent and can be the object of difficult controversies.

Consequently, DRO should not be seen as a complete solution to the bias fighting problem, but rather as a way to produce a single system that works almost as well on all subpopulations as the best system we can get on each subpopulation in isolation, which by themselves may or may not represent an acceptable combination of results. This is the motivation for our recommendations for addressing the majority bias in practice using DRO (Inset 1), which are discussed in more details in Appendix B.

Societal impact Using DRO for fairness or adversarial robustness without a clear understanding of its algorithmic limitations can have a negative societal impact. Recommendations in Inset 1 and Appendix Section B aim to prevent misuses of DRO, such as lowering performances on the remaining subpopulations to match the error on the most difficult distribution. However, our results show that it is also necessary to address the underlying problems in the most challenging distribution. On one hand, failure to address the issues in the minority subpopulation leaves it susceptible to discrimination, both in the application at hand and in the future applications, where the unresolved issues might persist. On the other hand, reducing the performance of the majority populations can lead to an unacceptable average performance, and as a result, the system is not going to be used — which might lead to a loss of interest in designing broadly accessible systems for this purpose (i.e., voice assistants robust to minority accents). We hope that our results and discussion will give more context to the debate on the sources of bias in machine learning (Hooker, 2021), as well as help in bias mitigation in real-life scenarios.

Fighting the majority bias with DRO: a minimal set of practical recommendations

1. Identify subpopulations P_k at risk in the available data.
2. For *each subpopulation, and in isolation*, determine the best performance $r^*_{P_k}$ that can be achieved with the machine learning model of choice.
3. Decide whether the $r^*_{P_k}$ represent an acceptable set of performances. *There is no point using DRO if this is not the case.* Instead, investigate why the model performs so poorly on the adverse distributions (insufficient data, inadequate model, etc.) until obtaining an acceptable set of $r^*_{P_k}$.
4. Use DRO to construct a single machine learning system whose performance on each subpopulation is not much worse than $r^*_{P_k}$. This can be achieved by using the $r^*_{P_k}$ as calibration coefficients in a Lagrangian algorithm.
5. Deploy the system on an experimental basis in order to collect more data. Sample the examples with the lowest accuracy in order to determine whether we missed a subpopulation at risk. If one is found, add the vulnerable subpopulation to the initial data and repeat all the steps.

Inset 1: A minimal set of practical recommendations. We elaborate on each step in the Appendix (Section B).

4.2 Adversarial examples

DRO is often presented as a good way to construct systems robust to adversarial examples (Szegedy et al., 2014; Madry et al., 2017). Following Example 2, this can be formalized by considering a set Φ of all measurable functions φ that map an example pattern z to another pattern $\varphi(z)$ assumed *visually indistinguishable* from z according to a predefined criterion. For instance, it is common to consider the set of all transformations φ such that $\|z-\varphi(z)\|_p \leq \kappa$, that is, transformations that can only modify an input pattern while remaining in a given L_p ball.

Let P_φ represent the distribution followed by $\varphi(z)$ when z follows the distribution P. Robust solutions against the class of adversarial perturbation Φ can be expressed as the DRO problem

$$\min_w \left\{ \quad \max_{\varphi\in\Phi} C_{P_\varphi}(w) \quad = \quad \max_{P_\varphi\in\mathcal{Q}} C_{P_\varphi}(w) \quad \right\} .$$

The distribution family $\mathcal{Q} = \{P_\varphi : \varphi\in\Phi\}$ is typically much larger than the ones considered in the bias fighting scenario. Instead of representing a finite number of subpopulations, the family $\mathcal{Q}$ is usually infinite and uncountable.

Theorem 2 handles infinite distribution classes using an additional *tightness* assumption. It is relatively easy to construct a sequence of mixtures distributions for which the expected gradient w^* tends to zero. The tightness assumption tells us that there exists a distribution that achieves that limit, that is, there exists a distribution for which the solution w^* is a stationary point of the expected loss.

Since the tightness assumption is trivially satisfied when the examples belong to a bounded domain (as is the case for images), this result suggests that there exists a distribution of images for which the ordinary training procedure yields a solution robust to adversarial examples. Is it true that we would not have adversarial example issues if only we had the right examples to start with?

More precisely, the theorem states that a DRO local minimum is a stationary point of the expected risk for an example distribution that depends on all the details of the DRO problem and in particular on the definition of the set Φ of adversarial perturbations, which itself encodes which images are considered visually indistinguishable from a reference image. On one hand, we could use DRO with a class of adversarial perturbations Φ whose effect is conservatively below the visual distinguishability threshold. For instance, the perturbation might be limited to changing pixel values by no more than a small threshold. Alas, the solution might be fooled by adversarial examples that do not satisfy this strict condition but nevertheless are still visually indistinguishable from the original image. On the other hand, we could use DRO with much broader class of perturbation, potentially including some that would be clear to a human observer. For instance, dithering patterns might occasionally introduce enough noise to be perceptually meaningful. Because such perturbations can dominate the DRO problem, it becomes necessary to introduce calibration constants in order to account for the variation in performance that can be justifiably expected with such perturbations.

Because DRO is fundamentally related to minimizing the expected cost for a well crafted example distribution, DRO does not really solve the original problem but merely displaces it into the specification of the class of adversarial perturbations or the selection of the associated cost calibration constants. However, the adversarial example scenario is substantially more challenging than the bias fighting scenario: because the number of potential perturbations is much larger than the number of potentially vulnerable subpopulations, we cannot work around the problem by first working on each of them in isolation as suggested in the Inset 1. We find concerning that using DRO for adversarial robustness without a reliable perceptual distance might be fundamentally flawed (Sharif et al., 2018).

5 RELATED WORK

This work is motivated by Proposition 2 of Arjovsky et al. (2019) which restates DRO for a finite class of distributions as a constrained optimization problem and shows under the usual KKT conditions that a solution of the DRO problem must be a stationary point of a certain mixture of the original distributions. Our analysis substantially broadens this result by showing that it still holds in the common setup where the class of distributions is infinite (Bagnell, 2005; Namkoong and Duchi, 2016; Blanchet et al., 2019; Staib and Jegelka, 2019) and by providing a sufficiency result. These extended results amount to a practical equivalence, with substantial consequences for important applications of DRO such as fighting bias in machine learning or constructing systems that resist adversarial examples.

As we discuss in later sections, the efficacy of DRO largely depends on setting the values of calibration coefficients (2.1). An existing approach, MaxiMin, (Meinshausen et al., 2015), sets calibration coefficients to the variance of a corresponding distribution, $r_P = \mathrm{Var}[Y_P]$, in order to maximize the minimum explained variance across distributions. Min-max regret is a related approach (Guillaume and Dubois, 2020).

Although biases in machine learning are more complex than differences in error rates (Barocas et al., 2019; Blodgett et al., 2020; Mehrabi et al., 2021), they often arise because the training algorithm optimizes a single performance criterion (Barocas et al., 2019, p218). Using DRO to minimize the worst error is therefore a useful template to understand how it can or cannot be used to address bias. For instance, DRO has been advocated to address biases in text autocompletion tasks (Hashimoto et al., 2018), to achieve robustness in the presence of noisy minority subpopulations (Wang et al., 2020), to predict recidivism (Duchi et al., 2020), or to protect sensitive attributes (Taskesen et al., 2020). Recent work (Zhou et al., 2021) analyses the failings of DRO in the presence of imperfect partitions.

DRO is often advocated to achieve robustness against adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2014a; Madry et al., 2017; Ren and Majumdar, 2021). Finding the appropriate choice for adversarial risk and making it match some notion of perceptual similarity is a topic of ongoing research. Various schemes have been proposed in the past, using L_0 norms (Papernot et al., 2015), L_2 norms (Szegedy et al., 2014), L_∞ norms (Goodfellow et al., 2014a; Madry et al., 2017), Wasserstein balls (Sinha et al., 2018), and perceptual criteria such as SSIM (Wang et al., 2004). Yet, all these criteria still get fooled by simple adversarial examples (Sharif et al., 2018). In this paper, we discuss how the practice of using DRO in adversarial robustness is problematic in the absence of a reliable perceptual distance.

6 CONCLUSION

We establish a series of theoretical results that clarify the relation between a well known algorithmic approach, DRO, and optimization of an expected error defined on a suitable combination of the original distributions. Contrary to the usual convex duality results, these results hold for nonconvex costs and for infinite families of distributions. These results also provide some support for the common practice of leveraging this quasi-equivalence to design efficient DRO algorithms. But it also becomes clear that running such an imperfect DRO algorithms is equivalent to optimizing the expected risk for a well crafted distribution.

We then discuss the consequences of these results for two practical problems, namely majority bias in machine learning and adversarial examples. Whether fighting bias in machine learning systems is a data curation problem or an algorithmic problem has been the object of much discussion. We extend this discussion by adding results which show an existence of an equivalent training distribution. However, this equivalent training distribution depends on minute details of the DRO formulation such as calibration coefficients or the exact definition of the family of distributions under consideration. Practically, this means that DRO is not a complete solution of the practical problem because it displaces the difficulty into setting the specifics of its formulation, e.g. the choice of calibration constants.

In the case of fighting bias against a discrete number of subpopulations, it makes sense to see DRO as a way to construct a single system that works almost as well as systems optimized for each subpopulation in isolation. This forms the basis for our minimal set of practical recommendations for addressing the majority bias (Inset 1) that can lead to a significant improvement in terms of understanding DRO relative to the current practice.

We then argue that a similar approach cannot be applied in the case of adversarial examples because potential attacks form an infinite and uncountable family. Although DRO has been shown to help, it may not be able to provide a complete solution without a precise understanding of what constitutes visually indistinguishable images.

Acknowledgements

Many thanks to the soon-to-be Dr Andrej Ivašković for the feedback on this work.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv CoRR*, abs/1907.02893, 2019.

K Arrow, L. Hurwicz, and H. Uzawa. *Studies in Nonlinear Programming.* Stanford Univ. Press, 1958.

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pages 228–245. Springer, 2020.

J. Andrew Bagnell. Robust Supervised Learning. In Manuela M. Veloso and Subbarao Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference*, pages 714–719. AAAI Press / The MIT Press, 2005.

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

Hamid Behravan, Ville Hautamäki, Sabato Marco Siniscalchi, Tomi Kinnunen, and Chin-Hui Lee. i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 24(1):29–41, 2016.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28 of *Princeton Series in Applied Mathematics.* Princeton University Press, 2009.

D.P. Bertsekas. *Convex Optimization Theory.* Athena Scientific optimization and computation series. Athena Scientific, 2009.

Patrick Billingsley. *Convergence of probability measures.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, 1999. ISBN 0-471-19745-9.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *Siam Reviews*, 60(2):223–311, 2018.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2014.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, 2018.

Jonathon Byrd and Zachary C. Lipton. What is the effect of importance weighting in deep learning? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 872–881. PMLR, 09–15 Jun 2019.

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014a.

Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014b.

Romain Guillaume and Didier Dubois. A min-max regret approach to maximum likelihood inference under incomplete data. *International Journal of Approximate Reasoning*, 121:135–149, 2020. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2020.03.003. URL `https://www.sciencedirect.com/science/article/pii/S0888613X19301215`.

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

Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100241. URL `https://www.sciencedirect.com/science/article/pii/S2666389921000611`.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does Distributionally Robust Supervised Learning Give Robust Classifiers? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2029–2037, 2018.

Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 587–597, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445920. URL `https://doi.org/10.1145/3442188.3445920`.

Brendan Klare, Mark James Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face Recognition Performance: Role of Demographic Information. *IEEE Trans. Inf. Forensics Secur.*, 7(6): 1789–1801, 2012.

Todd K. Leen. From data distributions to regularization in invariant learning. *Neural Computation*, 7 (5):974–981, 1995.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/64986d86a17424eeac96b08a6d519059-Abstract.html`.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 2021.

David Lopez-Paz. Simple baselines for worst group accuracy. Technical report, 2021. personal communication.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL `https://doi.org/10.1145/3457607`.

Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Nicolai Meinshausen, Peter Bühlmann, et al. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Maryam Najafian and Martin J. Russell. Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*, 122:44–55, 2020.

Hongseok Namkoong and John C. Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 2208–2216, 2016.

Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.*, 22 (10):1345–1359, 2010.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings. *CoRR*, abs/1511.07528, 2015.

Allen Z Ren and Anirudha Majumdar. Distributionally robust policy learning via adversarial environment generation. *arXiv preprint arXiv:2107.06353*, 2021.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A Survey of Deep Active Learning, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations, ICLR 2020*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 8346–8356. PMLR, 13–18 Jul 2020b.

Levent Sagun, Utku Evci, Veli Uğur Güney, Yann Dauphin, and Léon Bottou. Empirical Analysis of the Hessian of Over-Parametrized Neural Networks.

In *Sixth International Conference on Learning Representations (ICLR), Workshop paper*, 2018.

Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. On the Suitability of Lp-Norms for Creating and Preventing Adversarial Examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Matthew Staib and Stefanie Jegelka. Distributionally Robust Optimization and Generalization in Kernel Methods. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 9131–9141, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations, ICLR 2014*, 2014.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. *CoRR*, abs/1808.01974, 2018.

Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya R. Gupta, and Michael I. Jordan. Robust Optimization for Fairness with Noisy Protected Groups. *CoRR*, abs/2002.09343, 2020.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson. Joint Modeling of Accents and Acoustics for Multi-Accent Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, pages 5989–5993. IEEE, 2018.

Jianzhe Zhen, Daniel Kuhn, and Wolfram Wiesemann. Mathematical foundations of robust and distributionally robust optimization. *arXiv preprint arXiv:2105.00760*, 2021.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12857–12867. PMLR, 2021. URL `http://proceedings.mlr.press/v139/zhou21g.html`.

On Distribution Robust Optimization and Data Rebalancing Supplementary Material

A LAGRANGIAN ALGORITHMS FOR DRO

The calibrated DRO problem (3) is easily rewritten as a constrained optimization problem by introducing a slack variable L:

$$\min_{w,L} L \quad \text{s.t.} \quad \forall P \in \mathcal{Q}. \quad C_P(w) - r_P - L \leq 0 \ .$$

With convex loss function, finite $\mathcal{Q}$, and under adequate qualification conditions (Boyd and Vandenberghe, 2014; Bertsekas, 2009), convex duality theory suggests to write the Lagrangian

$$L(w, M, \lambda_1 \ldots \lambda_K) = M + \sum_k \lambda_k \big(C_{P_k}(w) - r_P - M \big) \ ,$$

and solve instead the dual problem,

$$\max_{\lambda_k \geq 0} \ \left\{ D(\lambda_1 \ldots \lambda_K) \triangleq \min_{w,M} L(w, M, \lambda_1 \ldots \lambda_K) \right\}$$

The solution must satisfy $\sum_k \lambda_k = 1$ because the dual $D(\lambda_1 \ldots \lambda_k)$ is $-\infty$ when this is not the case. With this knowledge, the dual problem becomes

$$\max_{\substack{\lambda_k \geq 0 \\ \sum_k \lambda_k = 1}} \ \left\{ D(\lambda_1 \ldots \lambda_K) = \left(\min_w \sum_k \lambda_k C_{P_k}(w) \right) - \left(\sum_k \lambda_k r_{P_k} \right) \right\} .$$

The inner optimization problem is precisely the minimization of the expected risk with respect to the mixture $\sum_k \lambda_k P_k$ and therefore lends itself to many popular gradient descent methods.

Convex duality also clarifies the relation between the mixture coefficients λ_k and the calibration constants r_{P_k}. Increasing the weight of a distribution in the mixture is equivalent to reducing the corresponding calibration coefficient. This observation then leads to a plethora of saddle-point seeking algorithms such as Uzawa iterations (Arrow et al., 1958).

This stategy is illustrated in Algorithm 1. Although this particular instance uses a temperature parameter β to smooth the mixture coefficient update rule, it is also common to focus on a single term with $\beta = +\infty$. When this is the case, each outer iteration of Algorithm 1 merely amounts to augmenting the training set with an extra copy of the examples associated with most adverse subpopulation.

Because of their simplicity and effectiveness, such Lagrangian DRO algorithms are widely used with deep learning system (Sagawa et al., 2020a; Augustin et al., 2020). Our theoretical results provide a measure support for the practice of applying such algorithm to nonconvex losses.

A crucial assumption for this algorithm is the idea that increasing the weight of a distribution in the mixture amounts to finding a local DRO minimum with a lower calibration coefficient for that distribution. This is true in the convex case. This requires a more precise discussion in the nonconvex case. Suppose for instance that one modifies the mixture coefficients by slightly increasing λ_1 by a small $\delta > 0$ and re-normalizing:

$$\lambda'_1 = \tfrac{1}{Z}(\lambda_1 + \delta) \qquad \lambda'_k = \tfrac{1}{Z}\lambda_k \ \ \forall k > 1$$

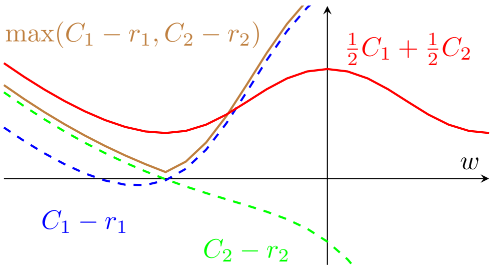

Figure 3: Both minima of $C_{\text{mix}}(w) = \frac{1}{2}C_1 + \frac{1}{2}C_2$ are solutions of a DRO problem, albeit one with different calibration constants r_1 and r_2. Here $r_2 > r_1$.

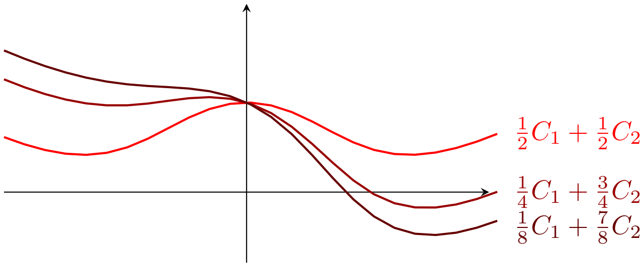

Figure 4: Increasing the weight of the second distribution beyond a certain threshold erases the first minimum and causes Algorithm 1 to jump to the other minimum which is a calibrated DRO minimum for $r_1 > r_2$.

Such a change can yield two outcomes. Either w^* remains a local minimum of the new expected cost mixture, or we can follow a descent trajectory and reach a new local minimum w':

$$Z \sum_k \lambda'_k C_{P_k}(w') \; < \; Z \sum_k \lambda'_k C_{P_k}(w^*) \; . \tag{5}$$

i) Let us first assume that the old cost function increases when one moves from its local minimum w^* to the local minimum w' of the new cost function

$$\sum_k \lambda_k C_{P_k}(w') \; \geq \; \sum_k \lambda_k C_{P_k}(w^*) \tag{6}$$

Subtracting (6) from (5) yields

$$\delta C_{P_1}(w') < \delta C_{P_1}(w^*) \; ,$$

which, according to Theorem 3, means that the new local minimum w' is a local minimum of a DRO problem with a reduced calibration coefficient for distribution P_1, just as for convex losses.

ii) However, it is also conceivable that (6) does not hold. This means that the new minimum w' achieves a lower cost than w^* for both the old and new mixture costs. In other words, tweaking the mixture allowed us to escape the attraction basin of the local minimum w^*. From the perspective of algorithm 1, this disrupts the determination of the mixture coefficient, but this is nevertheless progress because both the old and new mixture costs are lower. In theory, this can only happen a finite number of times in a neural network because there is only a finite number of attraction basins. In practice, this never happens: stochastic gradient descent in neural networks usually follows a path with slowly decreasing cost without hopping from one attraction basin to another one (Goodfellow et al., 2014b; Sagun et al., 2018).

As mentioned earlier, it is also conceivable that w^* remains a local minimum with the new mixture cost. Algorithm 1 then keeps increasing the weight of distribution P_1 as longs as the cost $C_{P_1}(w^*) = C_{P_1}(w')$ remains too high with respect to the desired calibration coefficients. This last case covers two distinct scenarios.

iii) The Lagrangian algorithm could keep increasing the weight of the first distribution without moving away from the local minimum w^*. The inner loop eventually minimizes the empirical risk for the first distribution only, yet without achieving progress. This suggests that we have reached a disappointing bound on the best performance achievable with our model using training data sampled from this first distribution.

iv) Alternatively, the old mixture local minimum w^* could stop being a local minimum of the new mixture once the first distribution weight reaches a certain threshold. Consider for instance the problem of Figure 1. Even though the DRO minimum corresponds to a local *maximum* of the mixture cost $C_{\text{mix}}(w) = \frac{1}{2}C_1 + \frac{1}{2}C_2$, Theorem 3 tells us that both minima of this mixture cost are also local DRO minima, albeit for different calibration constants r_i. Figure 3 shows the case where $r_2 > r_1$. Figure 4 shows that increasing the weight of the second cost function beyond a certain threshold eventually erases the left minimum and causes Algorithm 1 to jump to the condition $r_1 > r_2$. In other words, our algorithm is not able to simultaneously keep both cost functions as low as they could separately be. This either suggests that these two goals are incompatible, or that the model does not have enough capacity to simultaneously achieve them together. As usual with neural networks, the remedy is overparametrization…

One can derive two conclusions from this brief analysis. First, as long as we use a Lagrangian descent algorithm to solve the DRO problem, there is little point being concerned about stationary points of the mixture cost that are not local minima because (a) the algorithm is not going to find them anyway, and (b) overparametrizing the network is likely to make them disappear anyway (scenario *iv* above). Second, the most concerning scenario is the case where a single distribution or subpopulation dominates the DRO problem because our model is unable to achieve a satisfactory performance even when it is trained to minimize the expected cost for that distribution only. When this is the case, DRO cannot help.

B Practical recommendations

In this section, we provide a minimal set of practical recommendations to machine learning engineers who face the difficult task of constructing and deploying bias-sensitive machine learning systems. We do not pretend that these recommendations are sufficient to address the bias problem, but merely represent intuitively sensible steps that are supported by our mathematical insights and should not be avoided. We summarize these recommendations in Inset 1.

We also motivate and elaborate on each step below.

The *identification of the subpopulations* of concern frames the problem because it also defines the success criterion, that is, bias mitigation with respect to meaningful subpopulations. Key factors to consider are future users of the system, information on which groups have previously suffered from discrimination in similar scenarios, and the quantity and quality of the available data at the training time. In particular, we must at least have enough data to evaluate the subpopulation performances reliably. For instance, in a face recognition system, subpopulations might contain images of people representing distinct ethnicities (Klare et al., 2012).

Working *on each subpopulation in isolation* attempts to determine the best achievable performance on each subpopulation if this subpopulation were the only target. Data available for minority subpopulations might be limited. In such case, data from remaining subpopulations can be used as a regularizer to improve performance on the subpopulation P of interest. For instance, we can train on a mixture of data coming from both the subpopulation P (with weight 1) and the remaining subpopulations (with weights α_P). We then treat α_P as a hyperparameter that we tune to achieve the best validation performance on data from the subpopulation P. Our estimate of r_P^* is then the performance of the resulting system, either measured on the validation set, or on held out data if such data is available in sufficient quantity. This is why it is important to have sufficient data to reliably validate a model performance on each subpopulation. Techniques proposed to tackle noisy datasets and scenarios with limited labelled examples (active learning (Ren et al., 2020), transfer learning (Pan and Yang, 2010; Tan et al., 2018)) can be used to increase the performance.

We can then judge whether the r_P^* represent an *acceptable set of performances* for a final system. No DRO solution can perform better on a subpopulation P than a model trained for this subpopulation P only. If the set of performances obtained in the previous steps is not acceptable, we must identify the *root cause* of this problem. For instance, if poor performance stems from insufficient data quality for the subpopulation, this problem will persist at the step of finding a consistent system using DRO. We need to then focus on improving data quality for vulnerable subpopulations. We recommend investigating the root cause of insufficient performance for each of the vulnerable subpopulations in isolation.

If minimum cost that can be achieved per each subpopulation is acceptable, we can then build a system that works consistently well across the subpopulations using DRO. In the simplest case, calibration coefficients r_P per each subpopulation are going to be equal to the optimum expected risk for that subpopulation alone, $r_P = \min_w C_P(w)$. We can also adjust the calibration coefficients to prevent overfitting to individual subpopulations (Sagawa et al., 2020a). For n examples in a certain subpopulation P, the expected risk $C_P(w)$ can be replaced by its empirical estimate $C_{P_n}(w)+$ augmented with a calibration constant that decreases when the number n of training examples increases. Moreover, the model size often needs to be larger than the model size that achieves the best performance on each individual subpopulations. Intuitively, this is needed because handling all subpopulations at once might be more demanding than handling only one. In the main paper, we also argue that overparametrization improves the issues associated with DRO local minima that are stationary points of an expected loss mixture but are not local minima of this mixture. As a result, overparametrization helps practical Lagragian DRO algorithms to find a good solution.

Finally, we must remain aware that the final system critically depends on the initial selection of the subpopulations of interest. Therefore, it remains essential to cautiously deploy such a system and to *monitor* its performance during the ramp up. In particular, the worst performing cases should be examined for consistent patterns that might indicate that a vulnerable subpopulation was not considered in the problem specification. When this is the case, the correct solution is to include the initially omitted subpopulation and start again.