On the Relation between Distributionally Robust Optimization and Data Curation (Student Abstract)

Agnieszka Słowik[1], **Léon Bottou**[2], **Sean B. Holden**[1], **Mateja Jamnik**[1]

[1]Department of Computer Science and Technology, University of Cambridge
15 JJ Thompson Avenue, Cambridge, CB3 0FD, United Kingdom
[2]Facebook AI Research
agnieszka.slowik@cl.cam.ac.uk

Abstract

Machine learning systems based on minimizing average error have been shown to perform inconsistently across important subsets of the data, and this defect is not exposed by a low average error for the entire dataset. In some social and economic applications, where data represent people, this can lead to discrimination of underrepresented gender, ethnic and other groups. Distributionally Robust Optimization (DRO) attempts to address 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 a weighted training dataset. A practical implication of our results is that neither DRO nor curation of the training set represent a complete solution for bias mitigation.

Introduction

Machine learning algorithms are increasingly used to support real-world decision-making. Optimizing for the loss averaged on the overall population can yield models that perform poorly on specific subpopulations, amplifying injustices in our society (Chouldechova 2017).

Distributionally Robust Optimization (DRO) (Ben-Tal, Ghaoui, and Nemirovski 2009) bridges two perspectives on this problem. DRO seems to offer a promising solution because it minimizes the worst loss observed on multiple distributions (which e.g. represent each subpopulation). However, it can be shown that, under weak conditions, DRO is closely related to minimizing average loss on some mixture of those distributions – that is, a training set in which the subpopulations have been weighted. Our contributions are:

1. We establish results that clarify the relation between DRO and the optimization of the same loss averaged on a correctly weighted training set.
2. We also show that neither DRO nor curation of the training set are a complete solution of our initial problem due to the implicit assumptions DRO makes on the data.
3. We use this mathematical understanding to provide a minimal set of practical recommendations with which to approach real-life bias mitigation. This is guided by our results which show DRO is not applicable if we are unable to obtain an acceptable result with systems optimized for each subpopulation alone.

Proofs and an extended discussion of the results and their implications are included in the supplemental file.

DRO and Data Curation

Traditionally, training a machine learning model seeks parameters that minimize a risk $C_P(w)$ that is the expectation of a loss function with respect to a *single distribution of training examples*. Alas, even when the training distribution is representative of the actual testing conditions, the trained system might perform very poorly on selected subsets of examples (Chouldechova 2017). In real life, this can be a source of major injustice. DRO *seemingly addresses this problem* by considering instead a collection $\mathcal{Q}$ of 'training distributions' and minimizing the expected risk observed on the *most adverse* distribution:

$$\min_w \max_{P \in \mathcal{Q}} C_P(w) \,. \tag{1}$$

We can introduce *calibration coefficients* r_P that control how we compare costs for different distributions:

$$\min_w \max_{P \in \mathcal{Q}} \left(C_P(w) - r_P\right) \,. \tag{2}$$

For convex cost functions we already know that finding a local minimum of the DRO problem (1) is equivalent to minimizing the usual expected risk with respect to a single, well-crafted, training distribution, because one can reformulate the DRO problem as a constrained optimization problem and rely on standard convex duality results (Bertsekas 2009). We show that similar results *hold for the local minima of the nonconvex costs* typical of modern deep learning systems, and also *hold when the family $\mathcal{Q}$ is infinite*.

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 $z \in \mathbb{R}^n$ are examples. The following theorem generalizes the result by Arjovsky et al. (Arjovsky et al. 2019) by eliminating the Karush-Kuhn-Tucker (KKT) conditions.

Theorem 1 (Finite case). *Let $\mathcal{Q} = \{P_1, \ldots, P_K\}$ be a finite set of probability distributions on $\mathbb{R}^n$ and let w^* be a local minimum of the DRO problem* (1) *or the calibrated DRO problem* (2)*. Let the costs $C_P(w) = \mathbb{E}_{z \sim P}[\ell(z, w)]$ 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 $\nabla C_{P_{\text{mix}}}(w^*) = 0$.*

When the collection $\mathcal{Q}$ is infinite (possibly uncountably) but satisfies a *tightness* condition (Billingsley 1999), we can still show that a DRO local minimum is a stationary point for a well-crafted training distribution. *Adversarial robustness* is an example of applying DRO on an infinite family of distributions.

Theorem 2 (Infinite case)**.** *Let $\mathcal{Q}$ be a tight family of probability distributions on $\mathbb{R}^n$. Let w^* be a local minimum of problem* (2)*. Let $\mathcal{Q}_{\rm mix}$ 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 $P \in \mathcal{Q}_{\rm mix}$ and such that $\|h(z,w) - h(z,w')\| \leq M\|w-w'\|$ for almost all $z \in \mathbb{R}^n$. Then $\mathcal{Q}_{\rm mix}$ contains a distribution $P_{\rm mix}$ such that $\nabla_w C_{P_{\rm mix}}(w^*) = 0$.*

Conversely, we consider a local minimum of the expectation of the loss with respect to an arbitrary mixture of distributions from $\mathcal{Q}$. Such a local minimum always is a local minimum of a *calibrated DRO* problem.

Theorem 3 (Converse)**.** *Let $P_{\rm mix} = \sum_k \lambda_k P_k$ be an arbitrary mixture of distributions $P_k \in \mathcal{Q}$. If w^* is a local minimum of $C_{P_{\rm mix}}$, then w^* is a local minimum of the calibrated DRO problem* (2) *with calibration coefficients $r_{P_k} = C_{P_k}(w^*)$.*

Proofs are given in the supplemental file.

Calibration Problems

The calibration constants r_P might be a better way than mixture coefficients λ_{P_k} to specify which performance discrepancies are deemed acceptable across subpopulations because there are useful reference points for choosing them. An intuitive approach is to use the calibration constants r^*_P representing the best performance we can reach with our machine learning model on each distribution P in isolation: $r^*_P = \min_w C_P(w)$. Solving the DRO problem for these calibration constants amounts to constructing a single machine learning system that performs nearly as well on each distribution P as a system trained for distribution P alone.

Note that based on Theorems 1 and 3, regardless of the chosen calibration constants, no DRO solution can achieve a performance better than r^*_P on any distribution P. If this were the case, it would mean that r^*_P was not correctly estimated, and the new performance would become the corrected r^*_P. This simple observation forms the basis for a minimal set of recommendations to machine learning engineers who face the difficult task of constructing and deploying bias-sensitive machine learning systems. These recommendations (summarized in Inset 1) represent intuitively sensible steps that are supported by our mathematical insights.

Conclusion

Whether fighting bias in machine learning systems is a data curation or an algorithmic problem has been the object of much discussion. Our results clarify the relation between a well-known algorithmic approach, DRO, and the optimization of the expected cost on a well-crafted data distribution. This analysis also clarifies that this well-crafted distribution is not universal but depends on often implicit details of the DRO problem setup such as calibration constants.

Fighting bias with DRO: recommendations

1. Identify subpopulations P_k at risk (based on 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 with calibration coefficients $r^*_{P_k}$ to construct a single machine learning system (these are the calibration coefficients).
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: Summary of practical recommendations.

Using DRO for fairness without a clear understanding of its algorithmic limitations can have a negative societal impact. Our recommendations aim to prevent misuses of DRO, such as lowering performances on the remaining subpopulations to match the error on the most difficult distribution. It follows from our results that it is also necessary to address the underlying problems in the most challenging distribution. We hope that our results and discussion will give more context to the debate on the sources of bias in machine learning and help with bias mitigation in real-life scenarios.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *arXiv CoRR*, abs/1907.02893.

Ben-Tal, A.; Ghaoui, L. E.; and Nemirovski, A. 2009. *Robust Optimization*, volume 28 of *Princeton Series in Applied Mathematics*. Princeton University Press. ISBN 978-1-4008-3105-0.

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

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

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.