Robust Distribution Learning with Local and Global Adversarial Corruptions

Sloan Nietert NIETERT@CS.CORNELL.EDU
Ziv Goldfeld GOLDFELD@CORNELL.EDU
Soroosh Shafiee SHAFIEE@CORNELL.EDU
Cornell University

Editors: Shipra Agrawal and Aaron Roth

Abstract

As the scope of data-driven decision-making grows, so does the risk posed by data poisoning attacks. To address this, researchers have developed learning algorithms with strong guarantees despite adversarial corruptions. Existing models for data corruption are primarily *global*, allowing an adversary to arbitrarily modify a bounded fraction of samples, or *local*, allowing for all samples to be slightly perturbed. For example, Huber's ε-contamination model in classical robust statistics (Huber, 1964) and the total variation (TV) ε-contamination model (Donoho and Liu, 1988) give the adversary an ε fraction of data to arbitrarily and globally corrupt. Popularized recently in the setting of adversarial training (Sinha et al., 2018), Wasserstein corruption models permit all of the data to be locally perturbed, bounding the average perturbation size by some radius $\rho \geq 0$. Recall that the p-Wasserstein distance is defined between distributions P, Q by

$$\mathsf{W}_p(P, Q) \coloneqq \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi} \left[\|X - Y\|_2^p\right]^{\frac{1}{p}},$$

where $\Pi(P, Q)$ is the set of their couplings. This metric naturally lifts the geometry of $\mathbb{R}^d$ to the space of distributions $\mathcal{P}(\mathbb{R}^d)$ with finite p-th absolute moments.

This work[1] examines non-parametric distribution learning under a unified framework supporting both local and global corruptions. Specifically, we permit an ε-fraction of clean samples to be arbitrarily modified and the remaining perturbations have average magnitude bounded by ρ (i.e., a TV perturbation of size ε and a W_1 perturbation of size ρ). This combined model presents new computational and statistical challenges which cannot be properly addressed via naive combinations of existing methods. To overcome this, we build upon recent techniques in algorithmic robust statistics and take a perspective rooted in optimal transport.

Formally, given access to n such corrupted samples, we seek a computationally efficient estimator $\hat{P}_n$ that minimizes the Wasserstein distance $\mathsf{W}_1(\hat{P}_n, P)$. In fact, we attack the fine-grained task of minimizing $\mathsf{W}_1(\Pi_\sharp \hat{P}_n, \Pi_\sharp P)$ for all orthogonal projections $\Pi \in \mathbb{R}^{d\times d}$, with performance scaling with $\mathrm{rank}(\Pi) = k$. This allows us to account simultaneously for mean estimation ($k = 1$), distribution estimation ($k = d$), as well as the settings interpolating between these two extremes. We characterize the optimal population-limit risk for this task and then develop an efficient finite-sample algorithm with error bounded by $\sqrt{\varepsilon k} + \rho + \tilde{O}(k\sqrt{d}n^{-1/k})$ when P has bounded covariance. Our procedure relies on a novel trace norm approximation of an ideal yet intractable 2-Wasserstein projection estimator.

We apply this algorithm to robust stochastic optimization, and, in the process, uncover a new method for overcoming the curse of dimensionality in Wasserstein distributionally robust optimization (WDRO). We show that, when applied to learning problems with k-dimensional affine structure, standard WDRO adapts to this structure and is effective with a significantly smaller ambiguity set than that suggested by classic theory. Specifically, our risk bounds scale proportionally to the error bound above, whereas a naive approach would require taking $k = d$.

Keywords: robust statistics, optimal transport, distributionally robust optimization

1. Extended abstract. Full version appears as [http://arxiv.org/abs/2406.06509, v1].

References

David L. Donoho and Richard C. Liu. The "Automatic" Robustness of Minimum Distance Functionals. *The Annals of Statistics*, 16(2):552 – 586, 1988.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.