Robust Estimation for Non-Parametric Families

An Approach based on Generative Adversarial Networks

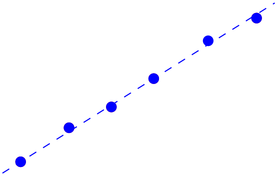
Banghua Zhu, with Jiantao Jiao and Michael I. Jordan

Department of Electrical Engineering and Computer Sciences,
University of California, Berkeley

June 28th, 2022

Motivation

Statistical inference with corrupted data:

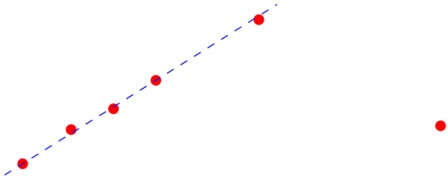


Clean data

Motivation

Statistical inference with corrupted data:

- Data Poisoning Attack / Backdoor Attack
- Byzantine clients in Distributed Learning
- Hamming's Adversarial Channel



Corruption under Total Variation (TV) distance

$$\mathsf{TV}(p, q) = \sup_A |p(A) - q(A)| \leq \epsilon$$

Setting

population distribution

$$p^* \in \mathcal{G}$$ distributional assumption

$$\downarrow$$

samples

$$\hat{p}_n : X_1, \ldots, X_n \in \mathbb{R}^d$$

$$\downarrow$$

estimated parameters cost

$$\hat{\theta}(X_1, \ldots, X_n) \longrightarrow L(p^*, \hat{\theta})$$

Setting

true distribution corrupted distribution

$$\begin{array}{ccc} p^* & \xrightarrow{\mathrm{TV}(p^*, p) \le \epsilon} & p \\ \Cap & & \big\downarrow \\ \mathcal{G} & & \end{array}$$

distributional assumption samples

$$\hat{p}_n : X_1, \ldots, X_n \in \mathbb{R}^d$$

$$\big\downarrow$$

estimated parameters cost

$$\hat{\theta}(X_1, \ldots, X_n) \longrightarrow L(p^*, \hat{\theta})$$

- Fundamental limit: $\inf_{\hat{\theta}(\hat{p}_n)} \sup_{\mathrm{TV}(p^\star, p) \le \epsilon, p^\star \in \mathcal{G}} L(p^\star, \hat{\theta}(\hat{p}_n))$
- In this talk: $L(p^\star, \theta) = \|\mu_{p^\star} - \theta\|$

Goal

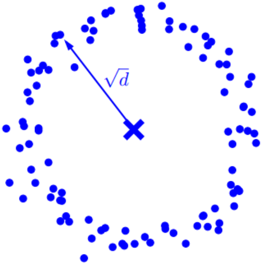
Fundamental limit: $\inf_{\hat{\theta}(\hat{p}_n)} \sup_{\mathsf{TV}(p^\star, p) \le \epsilon, p^\star \in \mathcal{G}} \|\mu_{p^\star} - \hat{\theta}(\hat{p}_n)\|$

- Population: infinite samples, settling the ϵ-dependence in the fundamental limit
- Generalization: finite samples, design algorithms with near-optimal sample complexity (ϵ, n, d-dependence)
- Computation: computationally efficient algorithm

Failure of Naïve algorithm

Suppose clean data is Gaussian:

$$x_i \sim \underbrace{\mathcal{N}(\mu, I)}_{\substack{\text{Gaussian mean } \mu \\ \text{variance 1 each coord.}}}$$

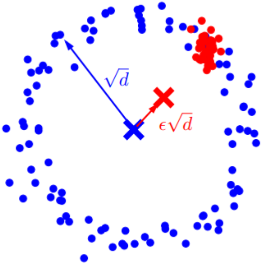


$$\|x_i - \mu\|_2 \approx \sqrt{1^2 + \cdots + 1^2} = \sqrt{d}$$

Failure of Naïve algorithm

Suppose clean data is Gaussian:

$$x_i \sim \underbrace{\mathcal{N}(\mu, I)}_{\substack{\text{Gaussian mean } \mu \\ \text{variance 1 each coord.}}}$$



$$\|x_i - \mu\|_2 \approx \sqrt{1^2 + \cdots + 1^2} = \sqrt{d}$$

- Failed to utilize distributional assumption.

Projection: minimum distance functionals

Population Algorithm: Minimum Distance Functionals / Projection

Return $\mu_q = \mathbb{E}_q[X]$, where

$$q = \mathrm{argmin}_{q \in \mathcal{G}} \mathsf{TV}(q, p)$$

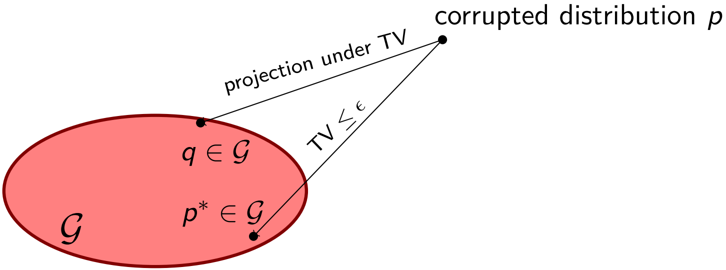
- Performance upper bounded by modulus of continuity [Donoho and R. C. Liu, 1988]:

$$\|\mu_{p^\star} - \mu_q\| \leq \sup_{p, p' \in \mathcal{G}, \mathsf{TV}(p, p') \leq 2\epsilon} \|\mu_p - \mu_{p'}\|$$

Proof of modulus as an upper bound

$$\|\mu_{p^\star} - \mu_q\| \leq \sup_{p,p' \in \mathcal{G}, \mathrm{TV}(p,p') \leq 2\epsilon} \|\mu_p - \mu_{p'}\|$$

$$q = \mathrm{argmin}_{q \in \mathcal{G}} \mathrm{TV}(q, p)$$



Mean estimation

Theorem (Zhu, Jiao, and Steinhardt, 2019)

Let $\mathcal{G}_\psi$ be the family of Orlicz-norm bounded distributions, i.e. $\mathcal{G}_\psi = \{\sup_{v\in\mathbb{R}^d, \|v\|_*\leq 1} \mathbb{E}_p[\psi(\frac{|v^\top(X-\mu_p)|}{\sigma})] \leq 1\}$ for Orlicz function ψ, we have modulus $\Theta(\sigma\epsilon\psi^{-1}(1/\epsilon))$.

Mean estimation

Theorem (Zhu, Jiao, and Steinhardt, 2019)

Let $\mathcal{G}_\psi$ be the family of Orlicz-norm bounded distributions, i.e. $\mathcal{G}_\psi = \{\sup_{v\in\mathbb{R}^d, \|v\|_*\le 1} \mathbb{E}_p[\psi(\frac{|v^\top(X-\mu_p)|}{\sigma})] \le 1\}$ for Orlicz function ψ, we have modulus $\Theta(\sigma\epsilon\psi^{-1}(1/\epsilon))$.

Example:

- $\psi(x) = exp(x^2) - 1$, $\mathcal{G}_\psi =$ sub-Gaussian, modulus $\Theta(\sigma\epsilon\sqrt{\log(1/\epsilon)})$
- $\psi(x) = x^k$, $\mathcal{G}_\psi =$ bounded k-th moment, modulus $\Theta(\sigma\epsilon^{1-1/k})$.
- $\psi(x) = x^2$, $\mathcal{G}_\psi = \{\|\Sigma_p\| \le \sigma^2\}$ bounded covariance, modulus $\Theta(\sigma\sqrt{\epsilon})$.

Finite-sample Algorithms

- Finite sample: only observe corrupted empirical distribution $\hat{p}_n$ instead of corrupted population distribution p

Finite-sample Algorithms

- Finite sample: only observe corrupted empirical distribution $\hat{p}_n$ instead of corrupted population distribution p
- One attempt: Do projection anyway! Take $\mathrm{argmin}_{q \in \mathcal{G}} \mathrm{TV}(q, \hat{p}_n)$.

Finite-sample Algorithms

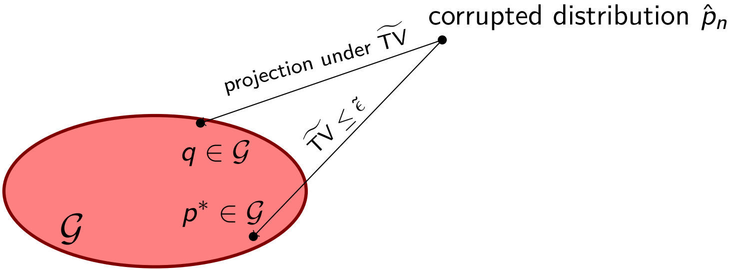
- Finite sample: only observe corrupted empirical distribution $\hat{p}_n$ instead of corrupted population distribution p
- One attempt: Do projection anyway! Take $\mathrm{argmin}_{q \in \mathcal{G}} \mathsf{TV}(q, \hat{p}_n)$.
- Problem: if $\mathcal{G}$ contains only continuous distributions, $\mathsf{TV}(q, \hat{p}_n) = 1, \forall q \in \mathcal{G}$

Solution I: Weaken the distance

- Although $\mathrm{TV}(\hat{p}_n, p) = 1$ for continuous p, find $\widetilde{\mathrm{TV}} \leq \mathrm{TV}$ such that $\widetilde{\mathrm{TV}}(\hat{p}_n, p)$ small.

Solution I: Weaken the distance

- Although $\mathrm{TV}(\hat{p}_n, p) = 1$ for continuous p, find $\widetilde{\mathrm{TV}} \leq \mathrm{TV}$ such that $\widetilde{\mathrm{TV}}(\hat{p}_n, p)$ small.



$\widetilde{\mathrm{TV}}$ for Mean Estimation

- Design $\widetilde{\mathrm{TV}}$ for mean estimation [Donoho and R. C. Liu, 1988]:

$$\widetilde{\mathrm{TV}}(p, q) = \sup_{t \in \mathbb{R}, v \in \mathbb{R}^d} |p(v^\top X \geq t) - q(v^\top X \geq t)|.$$

- Then

$$\widetilde{\mathrm{TV}} \leq \mathrm{TV}$$

and w.p. $1 - \delta$, [Vapnik and Chervonenkis, 1971; Dudley, 1978]

$$\widetilde{\mathrm{TV}}(\hat{p}_n, p) \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right)$$

Computation: Weakening the Distance

Weaken the distance: $\mathrm{argmin}_{q\in\mathcal{G}} \widetilde{\mathsf{TV}}(q, \hat{p}_n)$.

$$\begin{aligned}\widetilde{\mathsf{TV}}(p,q) &= \sup_{t\in\mathbb{R}, v\in\mathbb{R}^d} |p(v^\top X \geq t) - q(v^\top X \geq t)| \\ &= \sup_{t\in\mathbb{R}, v\in\mathbb{R}^d} |\mathbb{E}_p[1(v^\top X \geq t)] - \mathbb{E}_q[1(v^\top X \geq t)]|\end{aligned}$$

Computation: Weakening the Distance

Weaken the distance: $\mathrm{argmin}_{q\in\mathcal{G}}\,\widetilde{\mathrm{TV}}(q,\hat{p}_n)$.

$$\begin{aligned}\widetilde{\mathrm{TV}}(p,q) &= \sup_{t\in\mathbb{R},v\in\mathbb{R}^d} |p(v^\top X\geq t)-q(v^\top X\geq t)| \\ &= \sup_{t\in\mathbb{R},v\in\mathbb{R}^d} |\mathbb{E}_p[1(v^\top X\geq t)]-\mathbb{E}_q[1(v^\top X\geq t)]|\end{aligned}$$

- Indicator loss is hard to optimize. Can we find surrogates?

$$A(p,q) = \sup_{t\in\mathbb{R},v\in\mathbb{R}^d} |\mathbb{E}_p[T(v^\top X+t)]-\mathbb{E}_q[T(v^\top X+t)]|$$

Computation: Weakening the Distance

Algorithm: Weaken the Distance

Return μ_q, where $q = \operatorname{argmin}_{q \in \mathcal{G}_\psi} A(q, \hat{p}_n)$,

$$A(p, q) = \sup_{t \in \mathbb{R}, v \in \mathbb{R}^d} |\mathbb{E}_p[T(v^\top X + t)] - \mathbb{E}_q[T(v^\top X + t)]|$$

Theorem (Zhu, Jiao, and Jordan, 2022)

Let $T(\cdot) = \mathrm{sigmoid}(\cdot)$*,* $\tilde{\epsilon} = 2\epsilon + C\sqrt{(d + \log(1/\delta))/n}$*. Then the above algorithm guarantees a mean estimation error of* $O(\tilde{\epsilon}\psi^{-1}(1/\tilde{\epsilon}) + \tilde{\epsilon}\log(1/\tilde{\epsilon}))$*.*

- Minimizing A can be approximately solved via GANs.
- $T(\cdot, \cdot)$ can be replaced with any neural network with multiple layers + sigmoid activation.

Interpretation: Weaken the Distance

Mean estimation error: $O(\tilde{\epsilon}\psi^{-1}(1/\tilde{\epsilon}) + \tilde{\epsilon}\log(1/\tilde{\epsilon}))$,
$\tilde{\epsilon} = 2\epsilon + C\sqrt{(d + \log(1/\delta))/n}$.

- Lower bound: $\Omega(\epsilon\psi^{-1}(1/\epsilon) + \sqrt{d/n})$
- Subexponential, error $\widetilde{\Theta}(\tilde{\epsilon}\log(1/\tilde{\epsilon}))$
- Bounded covariance, error $O(\sqrt{\tilde{\epsilon}}) = O(\sqrt{\epsilon + \sqrt{d/n}})$

Solution II: Expand the set

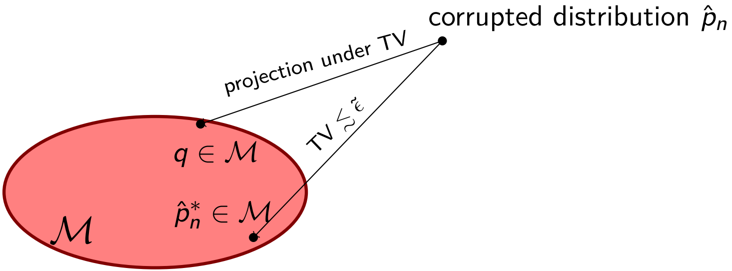
Solution II: expand the set $\mathcal{G}$ in $\mathrm{argmin}_{q\in\mathcal{G}}\, \mathsf{TV}(q, \hat{p}_n)$.

- Instead of changing TV, try to find a larger set $\mathcal{M} \supset \mathcal{G}$ such that $\hat{p}_n^\star \in \mathcal{M}$.

Solution II: Expand the set

Solution II: expand the set $\mathcal{G}$ in $\mathrm{argmin}_{q\in\mathcal{G}}\,\mathsf{TV}(q,\hat{p}_n)$.

- Instead of changing TV, try to find a larger set $\mathcal{M} \supset \mathcal{G}$ such that $\hat{p}_n^\star \in \mathcal{M}$.



Solution II: Expand the set

Algorithm: Expand the Set

Find $q \in \mathcal{M}$ such that q is an ϵ-deletion of $\hat{p}_n$: $q_i \leq \frac{1}{(1-\epsilon)n}$.

Theorem (Zhu, Jiao, and Steinhardt, 2019; Zhu, Jiao, and Steinhardt, 2020)

Assume $\|\Sigma_{p^\star}\|_2 \leq \sigma^2$. *Take* $\mathcal{M} = \{p \mid \|\Sigma_p\|_2 \leq C(1 + \frac{d\log(d/\delta)}{n\epsilon})\sigma^2\}$. *Projection under* TV *guarantees error* $O(\sqrt{\epsilon + \frac{d\log(d/\delta)}{n}})$ *w.p.* $1-\delta$ *when* $\epsilon < 1/2$.

- Optimal breakdown point and near optimal rate.

Weakening v.s. Expanding

- Weakening the distance:
  - find $p^* \in \mathcal{G}$ via projection onto $\mathcal{G}$ under $\widetilde{\mathrm{TV}}$
  - statistics literature [Maronna, 1976; Huber, 1973; Donoho, 1982; Donoho and R. C. Liu, 1988; Adrover and Yohai, 2002; Chen et al., 2018; Gao, J. Liu, et al., 2018; Gao, Yao, et al., 2019]
- Expanding the set:
  - find $\hat{p}_n^* \in \mathcal{M}$ via projection onto $\mathcal{M}$ under TV
  - theoretical computer science literature [Diakonikolas et al., 2016; Diakonikolas et al., 2017; Prasad et al., 2018; Klivans et al., 2018; Cheng et al., 2019; Steinhardt et al., 2018; Steinhardt, 2018]

Application: Byzantine-Robust Federated Learning

Federated Learning: m worker machines, each with n $i.i.d.$ samples, send local gradient to master machine. ϵ fraction of the workers may be Byzantine.

Theorem (Zhu, Wang, et al., 2022)

No-regret algorithm, when applied to gradient aggregation in federated learning, incurs statistical error $\widetilde{\Theta}(\sqrt{\frac{\epsilon}{n} + \frac{d}{mn}})$ *for strongly convex and smooth loss functions.*

- Yin et al., 2018 applies coordinate-wise median or coordinate-wise trimmed mean and achieves a rate of $\widetilde{O}(\sqrt{\frac{\epsilon d}{n} + \frac{d^2}{mn}})$.

Future Work

- Achieving sub-Gaussian rate $O(\sqrt{\epsilon + \frac{d+\log(1/\delta)}{n}})$ and high breakdown point simultaneously under bounded covariance assumption.
- Application to decentralized systems / fraud detection.

References I

Adrover, Jorge and Vıctor Yohai (2002). "Projection estimates of multivariate location." In: *The Annals of Statistics* 30.6, pp. 1760–1781.

Chen, Mengjie, Chao Gao, Zhao Ren, et al. (2018). "Robust covariance and scatter matrix estimation under Huber's contamination model." In: *The Annals of Statistics* 46.5, pp. 1932–1960.

Cheng, Yu, Ilias Diakonikolas, Rong Ge, and David Woodruff (2019). "Faster Algorithms for High-Dimensional Robust Covariance Estimation." In: *arXiv preprint arXiv:1906.04661*.

Diakonikolas, Ilias, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart (2016). "Robust estimators in high dimensions without the computational intractability." In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 655–664.

References II

Diakonikolas, Ilias, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart (2017). "Being robust (in high dimensions) can be practical." In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 999–1008.

Donoho, David L (1982). *Breakdown properties of multivariate location estimators*. Tech. rep. Technical report, Harvard University, Boston.

Donoho, David L and Richard C Liu (1988). "The "automatic" robustness of minimum distance functionals." In: *The Annals of Statistics* 16.2, pp. 552–586.

Dudley, Richard M (1978). "Central limit theorems for empirical measures." In: *The Annals of Probability*, pp. 899–929.

Gao, Chao, Jiyi Liu, Yuan Yao, and Weizhi Zhu (2018). "Robust Estimation and Generative Adversarial Nets." In: *arXiv preprint arXiv:1810.02030*.

References III

- Gao, Chao, Yuan Yao, and Weizhi Zhu (2019). "Generative Adversarial Nets for Robust Scatter Estimation: A Proper Scoring Rule Perspective." In: *arXiv preprint arXiv:1903.01944*.
- Huber, Peter J (1973). "Robust regression: asymptotics, conjectures and Monte Carlo." In: *The Annals of Statistics* 1.5, pp. 799–821.
- Klivans, Adam, Pravesh K Kothari, and Raghu Meka (2018). "Efficient algorithms for outlier-robust regression." In: *arXiv preprint arXiv:1803.03241*.
- Maronna, Ricardo Antonio (1976). "Robust M-estimators of multivariate location and scatter." In: *The annals of statistics*, pp. 51–67.
- Prasad, Adarsh, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar (2018). "Robust estimation via robust gradient estimation." In: *arXiv preprint arXiv:1802.06485*.

References IV

Steinhardt, Jacob (2018). "Robust Learning: Information Theory and Algorithms." Doctoral dissertation. Stanford University.

Steinhardt, Jacob, Moses Charikar, and Gregory Valiant (2018). "Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers." In: *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Vol. 94. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, p. 45.

Vapnik, Vladimir N and A Ya Chervonenkis (1971). "On the uniform convergence of relative frequencies of events to their probabilities." In: *Measures of complexity*. Springer, pp. 11–30.

Yin, Dong, Yudong Chen, Kannan Ramchandran, and Peter Bartlett (2018). "Byzantine-robust distributed learning: Towards optimal statistical rates." In: *arXiv preprint arXiv:1803.01498*.

References V

Zhu, Banghua, Jiantao Jiao, and Michael I Jordan (2022). "Robust Estimation for Nonparametric Families via Generative Adversarial Networks." In: *arXiv preprint arXiv:2202.01269*.

Zhu, Banghua, Jiantao Jiao, and Jacob Steinhardt (2019). "Generalized Resilience and Robust Statistics." In: *arXiv preprint arXiv:1909.08755*.

— (2020). "Robust estimation via generalized quasi-gradients." In: *arXiv preprint arXiv:2005.14073*.

Zhu, Banghua, Lun Wang, Qi Pang, Shuai Wang, Jiantao Jiao, Dawn Song, and Michael I Jordan (2022). "Byzantine-Robust Federated Learning with Optimal Statistical Rates and Privacy Guarantees." In: *arXiv preprint arXiv:2205.11765*.