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

$$TV(p, q) = \sup_A |p(A) - q(A)| \leq \epsilon$$
Setting

population distribution

\[ p^* \in \mathcal{G} \quad \text{distributional assumption} \]

\[ \downarrow \]

samples

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

\[ \downarrow \]

estimated parameters

\[ \hat{\theta}(X_1, \ldots, X_n) \rightarrow L(p^*, \hat{\theta}) \]
Setting

true distribution $p^*$  corrupted distribution $p$

$\sup_{TV(p^*, p) \leq \epsilon} \inf_{\hat{\theta}(\hat{p}_n)} L(p^*, \hat{\theta}(\hat{p}_n))$

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

estimated parameters cost

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

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

Fundamental limit: \( \inf_{\hat{\theta}(\hat{\rho}_n)} \sup_{TV(p^*, p) \leq \epsilon, p^* \in G} \| \mu p^* - \hat{\theta}(\hat{\rho}_n) \| \)

- Population: infinite samples, settling the \( \epsilon \)-dependence in the fundamental limit

- Generalization: finite samples, design algorithms with near-optimal sample complexity (\( \epsilon, n, d \)-dependence)

- Computation: computationally efficient algorithm
Suppose clean data is Gaussian:

\[ x_i \sim \mathcal{N}(\mu, I) \]

Gaussian mean \( \mu \)
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:

\[ x_i \sim \mathcal{N}(\mu, I) \]

- Gaussian mean \( \mu \)
- Variance 1 each coord.

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

- Failed to utilize distributional assumption.
Population Algorithm: Minimum Distance Functionals / Projection

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

$$q = \arg\min_{q \in G} \text{TV}(q, p)$$

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

$$\|\mu_{p^*} - \mu_q\| \leq \sup_{p, p' \in G, \text{TV}(p, p') \leq 2\epsilon} \|\mu_p - \mu_{p'}\|$$
Proof of modulus as an upper bound

\[ \| \mu_{p^*} - \mu_q \| \leq \sup_{p, p' \in \mathcal{G}, \text{TV}(p, p') \leq 2\epsilon} \| \mu_p - \mu_{p'} \| \]

\[ q = \arg\min_{q \in \mathcal{G}} \text{TV}(q, p) \]
Mean estimation

Theorem (Zhu, Jiao, and Steinhardt, 2019)

Let $\mathcal{G}_\psi$ be the family of Orlicz-norm bounded distributions, i.e.

$$\mathcal{G}_\psi = \{ \sup_{v \in \mathbb{R}^d, \|v\|_* \leq 1} \mathbb{E}_p [\psi \left( \frac{|v^T (X - \mu_p)|}{\sigma} \right)] \leq 1 \}$$

for Orlicz function $\psi$, 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 $\Theta(\sigma \epsilon^{1 - 1/k})$.
  - $\psi(x) = x^2$, $\mathcal{G}_\psi = \{ \|\Sigma_p\| \leq \sigma^2 \}$ bounded covariance, modulus $\Theta(\sigma \epsilon \sqrt{\epsilon})$. 

Banghua Zhu (UC Berkeley) 
Robust Estimation via GANs 
June 28th, 2022 8 / 19
Mean estimation

Theorem (Zhu, Jiao, and Steinhardt, 2019)

Let $\mathcal{G}_\psi$ be the family of Orlicz-norm bounded distributions, i.e.

$$
\mathcal{G}_\psi = \{ \sup_{v \in \mathbb{R}^d, \|v\|_* \leq 1} \mathbb{E}_p[\psi(\frac{|v^T(X-\mu_p)|}{\sigma})] \leq 1 \} \text{ for Orlicz function } \psi, \text{ we have modulus } \Theta(\sigma \epsilon \psi^{-1}(1/\epsilon)).
$$

Example:

- $\psi(x) = \exp(x^2) - 1$, $\mathcal{G}_\psi = \text{sub-Gaussian}$, modulus $\Theta(\sigma \epsilon \sqrt{\log(1/\epsilon)})$
- $\psi(x) = x^k$, $\mathcal{G}_\psi = \text{bounded } k\text{-th moment}$, modulus $\Theta(\sigma \epsilon^{1-1/k})$.
- $\psi(x) = x^2$, $\mathcal{G}_\psi = \{ \|\Sigma_p\| \leq \sigma^2 \}$ 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 \( \arg\min_{q \in G} \text{TV}(q, \hat{p}_n) \).
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 $\text{argmin}_{q \in G} \text{TV}(q, \hat{p}_n)$.

- Problem: if $G$ contains only continuous distributions, $\text{TV}(q, \hat{p}_n) = 1, \forall q \in G$
Solution I: Weaken the distance

- Although $TV(\hat{p}_n, p) = 1$ for continuous $p$, find $\tilde{TV} \leq TV$ such that $\tilde{TV}(\hat{p}_n, p)$ small.
Solution I: Weaken the distance

- Although $TV(\hat{p}_n, p) = 1$ for continuous $p$, find $\tilde{TV} \leq TV$ such that $\tilde{TV}(\hat{p}_n, p)$ small.
**TV for Mean Estimation**

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

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

- Then $\tilde{\text{TV}} \leq \text{TV}$ and w.p. $1 - \delta$, [Vapnik and Chervonenkis, 1971; Dudley, 1978]

$$\tilde{\text{TV}}(\hat{p}_n, p) \leq O\left(\sqrt{\frac{d + \log(1/\delta)}{n}}\right).$$
Weaken the distance: \( \arg\min_{q \in G} \tilde{\text{TV}}(q, \hat{p}_n) \).

\[
\tilde{\text{TV}}(p, q) = \sup_{t \in \mathbb{R}, \nu \in \mathbb{R}^d} \left| p(\nu^\top X \geq t) - q(\nu^\top X \geq t) \right|
\]

\[
= \sup_{t \in \mathbb{R}, \nu \in \mathbb{R}^d} \left| \mathbb{E}_p[1(\nu^\top X \geq t)] - \mathbb{E}_q[1(\nu^\top X \geq t)] \right|
\]
Weaken the distance: \( \arg\min_{q \in \mathcal{G}} \tilde{\text{TV}}(q, \hat{p}_n) \).

\[
\tilde{\text{TV}}(p, q) = \sup_{t \in \mathbb{R}, v \in \mathbb{R}^d} |p(v^TX \geq t) - q(v^TX \geq t)|
\]

\[
= \sup_{t \in \mathbb{R}, v \in \mathbb{R}^d} |\mathbb{E}_p[1(v^TX \geq t)] - \mathbb{E}_q[1(v^TX \geq t)]|
\]

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

\[
A(p, q) = \sup_{t \in \mathbb{R}, v \in \mathbb{R}^d} |\mathbb{E}_p[T(v^TX + t)] - \mathbb{E}_q[T(v^TX + t)]|
\]
Computation: Weakening the Distance

Algorithm: Weaken the Distance

Return $\mu_q$, where $q = \arg\min_{q \in g} A(q, \hat{p}_n)$,

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

Theorem (Zhu, Jiao, and Jordan, 2022)

Let $T(\cdot) = \text{sigmoid}(\cdot)$, $\tilde{\epsilon} = 2\epsilon + C \sqrt{(d + \log(1/\delta))/n}$. 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: \( O(\hat{\epsilon} \psi^{-1}(1/\hat{\epsilon}) + \hat{\epsilon} \log(1/\hat{\epsilon})) \),
\[ \hat{\epsilon} = 2\epsilon + C \sqrt{d + \log(1/\delta)}/n. \]

- Lower bound: \( \Omega(\epsilon \psi^{-1}(1/\epsilon) + \sqrt{d/n}) \)
- Subexponential, error \( \tilde{\Theta}(\tilde{\epsilon}\log(1/\tilde{\epsilon})) \)
- Bounded covariance, error \( O(\sqrt{\hat{\epsilon}}) = O(\sqrt{\epsilon + \sqrt{d/n}}) \)
Solution II: Expand the set

Solution II: expand the set $\mathcal{G}$ in $\arg\min_{q \in \mathcal{G}} \text{TV}(q, \hat{p}_n)$.

- Instead of changing TV, try to find a larger set $\mathcal{M} \supset \mathcal{G}$ such that $\hat{p}_n^* \in \mathcal{M}$. 
Solution II: Expand the set $\mathcal{G}$ in $\arg\min_{q \in \mathcal{G}} TV(q, \hat{p}_n)$.

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

The diagram illustrates the projection under $TV$ of the corrupted distribution $\hat{p}_n$ onto the set $\mathcal{M}$, where $q \in \mathcal{M}$ and $\hat{p}_n^* \in \mathcal{M}$, with $TV \lesssim \varepsilon$. 

Banghua Zhu (UC Berkeley)
Solution II: Expand the set

Algorithm: Expand the Set

Find $q \in M$ such that $q$ is an $\epsilon$-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 $\|\Sigma_p^*\|_2 \leq \sigma^2$. Take $M = \{p \mid \|\Sigma_p\|_2 \leq C(1 + \frac{d \log(d/\delta)}{n \epsilon})\sigma^2\}$. 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 $\tilde{TV}$
  • 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]
Federated Learning: $m$ worker machines, each with $n$ $i.i.d.$ samples, send local gradient to master machine. $\epsilon$ 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 $\tilde{\Theta}(\sqrt{\frac{\epsilon}{n} + \frac{d}{mn}})$ 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 $\tilde{O}(\sqrt{\frac{\epsilon d}{n} + \frac{d^2}{mn}})$. 
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 III


References V


