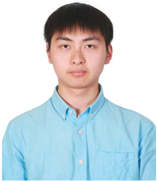
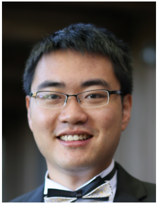
Generalized Resilience and Robust Statistics

Jacob Steinhardt
with Banghua Zhu and Jiantao Jiao



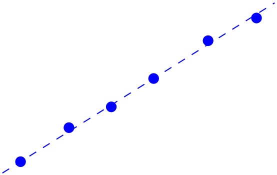


UC Berkeley

August 8, 2019

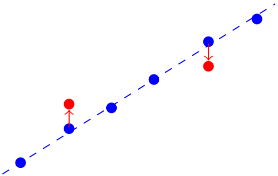
Motivation

Would like to design robust estimators:



Motivation

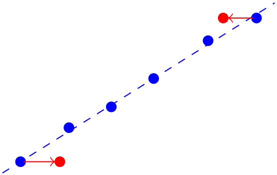
Would like to design robust estimators:



- Process error

Motivation

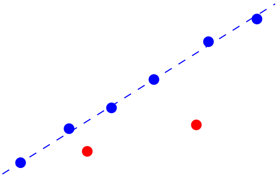
Would like to design robust estimators:



- Process error
- Measurement error

Motivation

Would like to design robust estimators:



- Process error
- Measurement error
- Outliers

The Difficulty

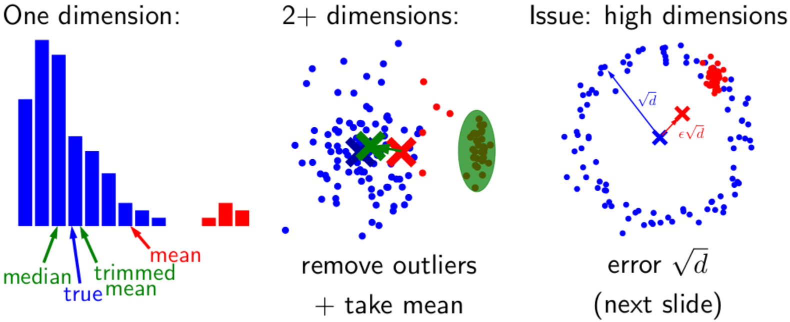
Simple example: mean estimation.

- Estimate mean of distribution in $\mathbb{R}^d$ with ε fraction of outliers.

The Difficulty

Simple example: mean estimation.

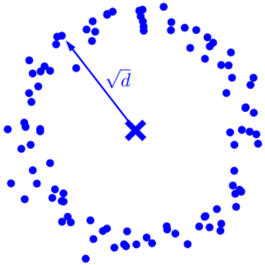
- Estimate mean of distribution in $\mathbb{R}^d$ with ε fraction of outliers.



The Difficulty

Suppose clean data is Gaussian:

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

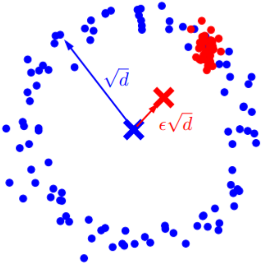


$$\|x_i - \mu\|_2 \approx \sqrt{1^2 + \cdots + 1^2} = \sqrt{d}$$

The Difficulty

Suppose clean data is Gaussian:

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



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

Context and Overview

Recent work designs **outlier-robust** estimators in many settings:

- mean estimation [DKKLMS16/17, LRV16, CSV17, SCV18, ...]
- regression [KK18, PSBR18, DKKLSS18]
- classification [KLS09, ABL14, DKS17], etc.

Will generalize and extend the insights:

- general treatment of population limit in presence of outliers
- new finite-sample analysis based on generalized KS distance
- robustness to **Wasserstein corruptions** based on “friendly perturbations”

Setting

population distribution

$$p$$

$$\downarrow$$

samples

$$X_1, \dots, X_n$$

$$\downarrow$$

estimated parameters　　　cost

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

Setting

true distribution $\quad$ corrupted distribution

$$p^* \xrightarrow{D(p^*, p) \le \varepsilon} \tilde{p}$$

$$\downarrow$$

samples

$$X_1, \ldots, X_n$$

$$\downarrow$$

estimated parameters $\quad$ cost

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

Setting

true distribution — corrupted distribution

$$p^* \xrightarrow{D(p^*, p) \leq \varepsilon} \tilde{p}$$

$$\downarrow$$

samples

$$X_1, \dots, X_n$$

$$\downarrow$$

estimated parameters — cost

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

Example $D = W_c$: cost $c(x, y)$ to move x to y, average cost $\leq \varepsilon$.

- $c(x, y) = \mathbb{I}[x \neq y]$: *TV* distance (outliers)
- $c(x, y) = \|x - y\|_2$: earthmover distance (measurement error)
- $c(x, y) = \|x - y\|_0$: corrupted entries

Setting

true distribution — corrupted distribution

$$p^* \xrightarrow{D(p^*, p) \le \varepsilon} \tilde{p}$$

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

distributional assumptions

$\tilde{p} \downarrow$ samples

$$X_1, \ldots, X_n$$

$\downarrow$ estimated parameters — cost

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

Example $D = W_c$: cost $c(x, y)$ to move x to y, average cost $\le \varepsilon$.

- $c(x, y) = \mathbb{I}[x \ne y]$: *TV* distance (outliers)
- $c(x, y) = \|x - y\|_2$: earthmover distance (measurement error)
- $c(x, y) = \|x - y\|_0$: corrupted entries

Setting

true distribution — corrupted distribution

$$p^* \xrightarrow{D(p^*, p) \le \epsilon} \tilde{p}$$

$p^* \in \mathcal{G}$ (distributional assumptions)

$\tilde{p} \longrightarrow$ samples $X_1, \ldots, X_n$

$\longrightarrow$ estimated parameters, cost

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

Three goals:

- large $\mathcal{G}$
- population limit
- statistical rate

Example $D = W_c$: cost $c(x, y)$ to move x to y, average cost $\le \epsilon$.

- $c(x, y) = \mathbb{I}[x \neq y]$: *TV* distance (outliers)
- $c(x, y) = \|x - y\|_2$: earthmover distance (measurement error)
- $c(x, y) = \|x - y\|_0$: corrupted entries

Warm-up: TV, mean estimation

Warm-up problem: $D = TV$, $L(p, \theta) = \|\mu(p) - \theta\|$, where $\mu(p) = \mathbb{E}_{x \sim p}[x]$.

- Mean estimation with outliers.

Warm-up: TV, mean estimation

Warm-up problem: $D = TV$, $L(p, \theta) = \|\mu(p) - \theta\|$, where $\mu(p) = \mathbb{E}_{x\sim p}[x]$.

- Mean estimation with outliers.

Key lemma: projection estimator. First observed by Donoho and Liu (1988).

Lemma

Suppose $p^* \in \mathcal{G}$, and define $\hat{\theta}(p) = \mu(q)$, where $q = \operatorname{argmin}_{q\in\mathcal{G}} TV(p, q)$. Then $L(p^*, \hat{\theta}(\tilde{p}))$ is upper-bounded by $\mathrm{modu}(\mathcal{G}, 2\varepsilon)$, where

$$\mathrm{modu}(\mathcal{G}, \varepsilon) := \sup_{p,p'\in\mathcal{G}, TV(p,p')\le\varepsilon} \|\mu(p) - \mu(p')\|.$$

Warm-up: TV, mean estimation

Warm-up problem: $D = TV$, $L(p, \theta) = \|\mu(p) - \theta\|$, where $\mu(p) = \mathbb{E}_{x \sim p}[x]$.

- Mean estimation with outliers.

Key lemma: projection estimator. First observed by Donoho and Liu (1988).

Lemma

Suppose $p^* \in \mathcal{G}$, and define $\hat{\theta}(p) = \mu(q)$, where $q = \mathrm{argmin}_{q \in \mathcal{G}} TV(p, q)$. Then $L(p^*, \hat{\theta}(\tilde{p}))$ is upper-bounded by $\mathrm{modu}(\mathcal{G}, 2\varepsilon)$, where

$$\mathrm{modu}(\mathcal{G}, \varepsilon) := \sup_{p, p' \in \mathcal{G}, TV(p, p') \leq \varepsilon} \|\mu(p) - \mu(p')\|.$$

Proof: $TV(p^*, q) \leq 2\varepsilon$, and p^*, q both lie in $\mathcal{G}$.

Modulus: Examples

$$\mathrm{modu}(\mathcal{G}, \varepsilon) := \sup_{p, p' \in \mathcal{G}, TV(p, p') \leq \varepsilon} \|\mu(p) - \mu(p')\|.$$

Modulus: Examples

$$\mathrm{modu}(\mathcal{G}, \varepsilon) := \sup_{p,p' \in \mathcal{G}, TV(p,p') \leq \varepsilon} \|\mu(p) - \mu(p')\|.$$

Example: Gaussians. $\mathcal{G} = \{\mathcal{N}(\mu, I) \mid \mu \in \mathbb{R}^d\}$.

- $\mathsf{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I)) \approx \|\mu - \mu'\|_2$.
- Hence $\mathrm{modu}(G, \varepsilon) \approx \varepsilon$.

Modulus: Examples

$$\mathrm{modu}(\mathcal{G}, \varepsilon) := \sup_{p,p' \in \mathcal{G}, TV(p,p') \leq \varepsilon} \|\mu(p) - \mu(p')\|.$$

Example: Gaussians. $\mathcal{G} = \{\mathcal{N}(\mu, I) \mid \mu \in \mathbb{R}^d\}$.

- $\mathrm{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I)) \approx \|\mu - \mu'\|_2$.
- Hence $\mathrm{modu}(G, \varepsilon) \approx \varepsilon$.

Generalization: $\mathcal{G}$ = sub-Gaussians (parameter σ).

- Can show that $\mathrm{modu}(G, \varepsilon) = \mathcal{O}(\sigma\varepsilon\sqrt{\log(1/\varepsilon)})$.
- Key lemma: thin tails $\implies$ ε-perturbation can't change mean much.

Modulus: Examples

$$\mathrm{modu}(\mathcal{G}, \varepsilon) := \sup_{p,p' \in \mathcal{G}, TV(p,p') \leq \varepsilon} \|\mu(p) - \mu(p')\|.$$

Example: Gaussians. $\mathcal{G} = \{\mathcal{N}(\mu, I) \mid \mu \in \mathbb{R}^d\}$.

- $\mathrm{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I)) \approx \|\mu - \mu'\|_2$.
- Hence $\mathrm{modu}(G, \varepsilon) \approx \varepsilon$.

Generalization: $\mathcal{G}$ = sub-Gaussians (parameter σ).

- Can show that $\mathrm{modu}(G, \varepsilon) = \mathcal{O}(\sigma\varepsilon\sqrt{\log(1/\varepsilon)})$.
- Key lemma: thin tails $\implies$ ε-perturbation can't change mean much.

General property: *resilience*.

Resilience

Definition (Resilience)

A distribution p is (ρ, ε)-resilient if $\|\mu(p) - \mu(r)\| \leq \rho$ whenever $r \leq \frac{p}{1-\varepsilon}$.

(The condition $r \leq \frac{p}{1-\varepsilon}$ means that r is an ε-deletion of p.)

Resilience

Definition (Resilience)

A distribution p is (ρ, ε)-resilient if $\|\mu(p) - \mu(r)\| \leq \rho$ whenever $r \leq \frac{p}{1-\varepsilon}$.

(The condition $r \leq \frac{p}{1-\varepsilon}$ means that r is an ε-deletion of p.)

Lemma (Resilience $\implies$ bounded modulus)

Let $\mathcal{G}(\rho, \varepsilon) = \{p \mid p \text{ is } (\rho, \varepsilon)\text{-resilient}\}$. *Then* $\mathrm{modu}(\mathcal{G}(\rho, \varepsilon), \varepsilon) \leq 2\rho$.

Resilience

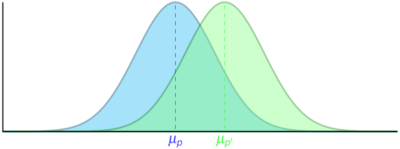
Definition (Resilience)

A distribution p is (ρ, ε)-resilient if $\|\mu(p) - \mu(r)\| \leq \rho$ whenever $r \leq \frac{p}{1-\varepsilon}$.

(The condition $r \leq \frac{p}{1-\varepsilon}$ means that r is an ε-deletion of p.)

Lemma (Resilience $\implies$ bounded modulus)

Let $\mathcal{G}(\rho, \varepsilon) = \{p \mid p \text{ is } (\rho, \varepsilon)\text{-resilient}\}$. *Then* $\mathrm{modu}(\mathcal{G}(\rho, \varepsilon), \varepsilon) \leq 2\rho$.



Proof: Let $p, p' \in \mathcal{G}(\rho, \varepsilon)$.

Resilience

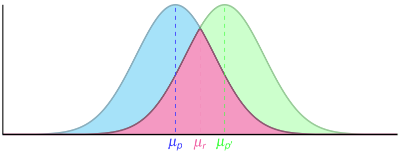
Definition (Resilience)

A distribution p is (ρ, ε)-resilient if $\|\mu(p) - \mu(r)\| \le \rho$ whenever $r \le \frac{p}{1-\varepsilon}$.

(The condition $r \le \frac{p}{1-\varepsilon}$ means that r is an ε-deletion of p.)

Lemma (Resilience $\implies$ bounded modulus)

Let $\mathcal{G}(\rho, \varepsilon) = \{p \mid p \text{ is } (\rho, \varepsilon)\text{-resilient}\}$*. Then* $\mathrm{modu}(\mathcal{G}(\rho, \varepsilon), \varepsilon) \le 2\rho$*.*



Proof: Let $p, p' \in \mathcal{G}(\rho, \varepsilon)$. Define midpoint $r = \frac{\min(p,p')}{1-\mathrm{TV}(p,p')}$. Then $r \le \frac{p}{1-\varepsilon}, \frac{p'}{1-\varepsilon}$.

Resilience

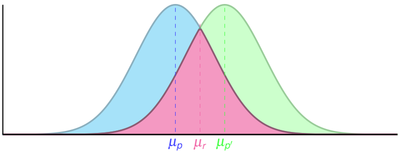
Definition (Resilience)

A distribution p is (ρ,ε)-resilient if $\|\mu(p)-\mu(r)\| \leq \rho$ whenever $r \leq \frac{p}{1-\varepsilon}$.

(The condition $r \leq \frac{p}{1-\varepsilon}$ means that r is an ε-deletion of p.)

Lemma (Resilience $\implies$ bounded modulus)

Let $\mathcal{G}(\rho,\varepsilon) = \{p \mid p \text{ is } (\rho,\varepsilon)\text{-resilient}\}$. *Then* $\mathrm{modu}(\mathcal{G}(\rho,\varepsilon),\varepsilon) \leq 2\rho$.



Proof: Let $p, p' \in \mathcal{G}(\rho,\varepsilon)$. Define midpoint $r = \frac{\min(p,p')}{1-\mathrm{TV}(p,p')}$. Then $r \leq \frac{p}{1-\varepsilon}, \frac{p'}{1-\varepsilon}$.
Thus $\|\mu(p)-\mu(p')\| \leq \|\mu(p)-\mu(r)\| + \|\mu(p')-\mu(r)\| \leq 2\rho$. □

Resilience

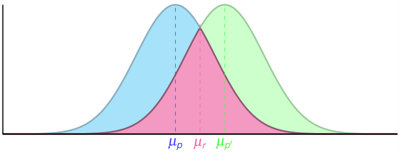
Definition (Resilience)

A distribution p is (ρ,ε)-resilient if $\|\mu(p)-\mu(r)\| \leq \rho$ whenever $r \leq \frac{p}{1-\varepsilon}$.

(The condition $r \leq \frac{p}{1-\varepsilon}$ means that r is an ε-deletion of p.)

Lemma (Resilience $\implies$ bounded modulus)

Let $\mathcal{G}(\rho,\varepsilon) = \{p \mid p \text{ is } (\rho,\varepsilon)\text{-resilient}\}$. *Then* $\mathrm{modu}(\mathcal{G}(\rho,\varepsilon),\varepsilon) \leq 2\rho$.



Modulus lemma yields optimal bound in most known cases!

- Sub-Gaussian: $\rho = \mathcal{O}(\varepsilon\sqrt{\log(1/\varepsilon)})$
- Bounded kth moments: $\rho = \mathcal{O}(\varepsilon^{1-1/k})$

Proof: Let $p, p' \in \mathcal{G}(\rho,\varepsilon)$. Define midpoint $r = \frac{\min(p,p')}{1-\mathrm{TV}(p,p')}$. Then $r \leq \frac{p}{1-\varepsilon}, \frac{p'}{1-\varepsilon}$.
Thus $\|\mu(p)-\mu(p')\| \leq \|\mu(p)-\mu(r)\| + \|\mu(p')-\mu(r)\| \leq 2\rho$. $\square$

Finite-sample estimation

Resilience characterizes error when $n = \infty$, what about finite samples?

Projection algorithm: take $\text{argmin}_{q \in \mathcal{G}} \text{TV}(\tilde{p}, q)$.

- Problem: if $\tilde{p}$ is discrete and q is continuous, $\text{TV}(\tilde{p}, q) = 1$!

Finite-sample estimation

Resilience characterizes error when $n = \infty$, what about finite samples?

Projection algorithm: take $\operatorname{argmin}_{q \in \mathcal{G}} \mathrm{TV}(\tilde{p}, q)$.

- Problem: if $\tilde{p}$ is discrete and q is continuous, $\mathrm{TV}(\tilde{p}, q) = 1$!

Solution: relax the distance!

$$\widetilde{\mathrm{TV}}_{\mathcal{H}}(p, q) = \sup_{t \in \mathbb{R}, h \in \mathcal{H}} |p(h(X) \geq t) - q(h(X) \geq t)|.$$

Finite-sample estimation

Resilience characterizes error when $n = \infty$, what about finite samples?

Projection algorithm: take $\operatorname{argmin}_{q \in \mathcal{G}} \mathsf{TV}(\tilde{p}, q)$.

- Problem: if $\tilde{p}$ is discrete and q is continuous, $\mathsf{TV}(\tilde{p}, q) = 1$!

Solution: relax the distance!

$$\widetilde{\mathsf{TV}}_{\mathcal{H}}(p, q) = \sup_{t \in \mathbb{R}, h \in \mathcal{H}} |p(h(X) \geq t) - q(h(X) \geq t)|.$$

Lemmas:

- Modulus is still bounded if we replace TV with $\widetilde{\mathsf{TV}}_{\mathcal{H}}$, where $\mathcal{H} = \{x \mapsto \langle v, x \rangle \mid v \in \mathbb{R}^d\}$.
- $\widetilde{\mathsf{TV}}_{\mathcal{H}}(p, \hat{p}_n) = \mathcal{O}(\sqrt{\mathrm{vc}(\mathcal{H})/n})$ [Devroye and Lugosi]

Upshot: projection still works, but use $\widetilde{\mathsf{TV}}_{\mathcal{H}}$ instead of TV.

General TV case

Focused so far on mean estimation. Now generalize to arbitrary loss.

- Stick with $D = \mathrm{TV}$, but replace $\|\mu(p) - \theta\|$ with arbitrary $L(p, \theta)$.

General TV case

Focused so far on mean estimation. Now generalize to arbitrary loss.

- Stick with $D = \mathrm{TV}$, but replace $\|\mu(p) - \theta\|$ with arbitrary $L(p, \theta)$.

Modulus still gives bound:

Lemma

Suppose $p^* \in \mathcal{G}$, and define $\hat{\theta}(p) = \theta^*(q)$, where $q = \mathrm{argmin}_{q \in \mathcal{G}} TV(p, q)$. Then $L(p^*, \hat{\theta}(\tilde{p}))$ is upper-bounded by $\mathrm{modu}(\mathcal{G}, 2\varepsilon)$, where

$$\mathrm{modu}(\mathcal{G}, \varepsilon) := \sup_{p, p' \in \mathcal{G}, TV(p, p') \leq \varepsilon} L(p, \theta^*(p')).$$

Can we generalize resilience to this setting?

Resilience: Arbitrary loss

Recall before: p is resilient if $\|\mu(p) - \mu(r)\|$ small whenever $r \le \frac{p}{1-\varepsilon}$.

Resilience: Arbitrary loss

Recall before: p is resilient if $\|\mu(p) - \mu(r)\|$ small whenever $r \le \frac{p}{1-\varepsilon}$.

Now two conditions: $\mathcal{G}_{\downarrow}$, $\mathcal{G}_{\uparrow}$.

$$\mathcal{G}_{\downarrow}(\rho_1, \varepsilon) = \{p \mid L(r, \theta^*(p)) \le \rho_1 \text{ whenever } r \le \tfrac{p}{1-\varepsilon}\},$$
$$\mathcal{G}_{\uparrow}(\rho_1, \rho_2, \varepsilon) = \{p \mid L(p, \theta) \le \rho_2 \text{ whenever } L(r, \theta) \le \rho_1 \text{ and } r \le \tfrac{p}{1-\varepsilon}\}.$$

Resilience: Arbitrary loss

Recall before: p is resilient if $\|\mu(p) - \mu(r)\|$ small whenever $r \le \frac{p}{1-\varepsilon}$.

Now two conditions: $\mathcal{G}_\downarrow$, $\mathcal{G}_\uparrow$.

$$\mathcal{G}_\downarrow(\rho_1, \varepsilon) = \{p \mid L(r, \theta^*(p)) \le \rho_1 \text{ whenever } r \le \tfrac{p}{1-\varepsilon}\},$$
$$\mathcal{G}_\uparrow(\rho_1, \rho_2, \varepsilon) = \{p \mid L(p, \theta) \le \rho_2 \text{ whenever } L(r, \theta) \le \rho_1 \text{ and } r \le \tfrac{p}{1-\varepsilon}\}.$$

Lemma (Resilience $\Longrightarrow$ small modulus)

Let $\mathcal{G} = \mathcal{G}_\downarrow(\rho_1, \varepsilon) \cap \mathcal{G}_\uparrow(\rho_1, \rho_2, \varepsilon)$*. Then* $\mathrm{modu}(\mathcal{G}, \varepsilon) \le \rho_2$*.*

Proof:

$$p \xrightarrow{D(p,p') \le \varepsilon} p'$$

$$p \xrightarrow{r \le \frac{p}{1-\varepsilon}} r, \qquad p' \xrightarrow{r \le \frac{p'}{1-\varepsilon}} r$$

$$r = \frac{\min(p, p')}{1 - \mathrm{TV}(p, p')}$$

$$p' \in \mathcal{G}_\downarrow \Rightarrow B(r, \theta^*(p')) \le \rho_1 \overset{p \in \mathcal{G}_\uparrow}{\Longrightarrow} L(p, \theta^*(p')) \le \rho_2$$

Example: Linear regression

Linear regression: $L(p,\theta) = \mathbb{E}_{(x,y)\sim p}[(y-\theta^\top x)^2] - \mathbb{E}_{(x,y)\sim p}[(y-(\theta^*)^\top x)^2]$.

Proposition (Sufficient conditions for linear regression)

Let $Z = Y - (\theta^*)^\top X$ be the regression error under the true parameters θ^*. Suppose that

$$\mathbb{E}[Z^{2k}] \le 1 \text{ and } \mathbb{E}[(v^\top X)^{2k}] \le \tau^{2k}\mathbb{E}[(v^\top X)^2]^k \ \forall v \in \mathbb{R}^d.$$

Then p^* is resilient with $\rho_2 = \mathcal{O}(\tau^2 \varepsilon^{2-2/k})$.

Example: Linear regression

Linear regression: $L(p, \theta) = \mathbb{E}_{(x,y)\sim p}[(y - \theta^\top x)^2] - \mathbb{E}_{(x,y)\sim p}[(y - (\theta^*)^\top x)^2]$.

Proposition (Sufficient conditions for linear regression)

Let $Z = Y - (\theta^*)^\top X$ be the regression error under the true parameters θ^*. Suppose that

$$\mathbb{E}[Z^{2k}] \leq 1 \text{ and } \mathbb{E}[(v^\top X)^{2k}] \leq \tau^{2k}\mathbb{E}[(v^\top X)^2]^k \ \forall v \in \mathbb{R}^d.$$

Then p^* is resilient with $\rho_2 = \mathcal{O}(\tau^2 \varepsilon^{2-2/k})$.

Comparisons:

- Delete points to minimize regression error (Klivans-Kothari-Mekha 2018): suboptimal error $\varepsilon^{1-1/k}$
- Diakonikolas-Kong-Stewart (2019) delete points to enforce moment condition: requires isotropy + 4th moments similar to Gaussian

Example: Covariance estimation

Given distribution with mean μ_p and covariance Σ_p.
Goal: output μ, Σ such that

$$\|I - \Sigma_p^{-1/2} \Sigma \Sigma_p^{-1/2}\|_2 \text{ and } \|\Sigma_p^{-1/2}(\mu_p - \mu)\|_2$$

are both small.

Example: Covariance estimation

Given distribution with mean μ_p and covariance Σ_p.
Goal: output μ, Σ such that

$$\|I - \Sigma_p^{-1/2} \Sigma \Sigma_p^{-1/2}\|_2 \text{ and } \|\Sigma_p^{-1/2}(\mu_p - \mu)\|_2$$

are both small.

Proposition (Sufficient condition for covariance estimation)

Suppose that $\mathbb{E}[(v^\top \Sigma_p^{-1/2}(X - \mu_p))^{2k}] \le \sigma^{2k} \|v\|_2^{2k}\ \forall v \in \mathbb{R}^d$. Then we can output Σ, μ such that

$$\|I - \Sigma_p^{-1/2} \Sigma \Sigma_p^{-1/2}\|_2 \le \mathcal{O}(\sigma \varepsilon^{1-1/k}) \text{ and } \tag{1}$$

$$\|\Sigma_p^{-1/2}(\mu_p - \mu)\|_2 \le \mathcal{O}(\sigma \varepsilon^{1-1/2k}). \tag{2}$$

Extension to other perturbations (W_c)

Recap: modulus determines robustness, resilience is sufficient condition for robustness in TV case.

Extension to other perturbations (W_c)

Recap: modulus determines robustness, resilience is sufficient condition for robustness in TV case.

Next extend results from TV to other W_c (transportation) distances.

- Recall $W_c(p, q)$ is cost to "move" p to q if moving $x \to y$ costs $c(x, y)$.
- Formally: $W_c(p, q) = \min_\pi\{\mathbb{E}_\pi[c(x, y)] \mid \pi(x) = p(x), \pi(y) = q(y)\}$.

Extension to other perturbations (W_c)

Recap: modulus determines robustness, resilience is sufficient condition for robustness in TV case.

Next extend results from TV to other W_c (transportation) distances.

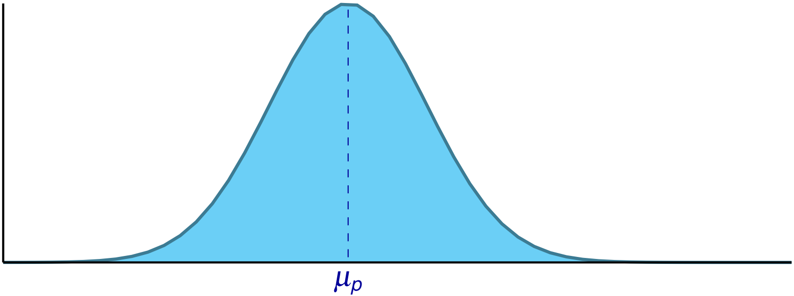
- Recall $W_c(p, q)$ is cost to "move" p to q if moving $x \to y$ costs $c(x, y)$.
- Formally: $W_c(p, q) = \min_\pi \{\mathbb{E}_\pi[c(x, y)] \mid \pi(x) = p(x), \pi(y) = q(y)\}$.

Key **midpoint** property of resilience: if $\mathrm{TV}(p, q) \le \varepsilon$, there exists midpoint r such that $r \le \frac{p}{1-\varepsilon}$ and $r \le \frac{q}{1-\varepsilon}$.

- How to generalize to W_c?

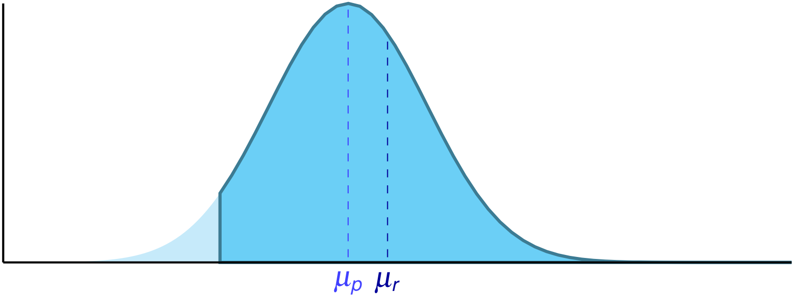
Friendly perturbations

Consider one-dimensional case:



Friendly perturbations

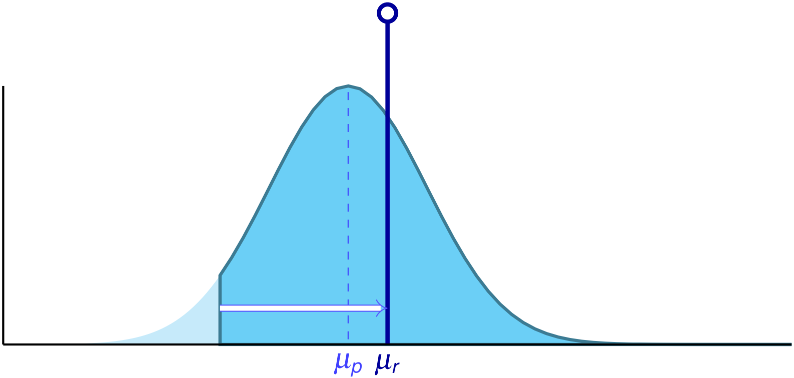
Consider one-dimensional case:



Delete ε-mass: $\mu_p \to \mu_r$.

Friendly perturbations

Consider one-dimensional case:

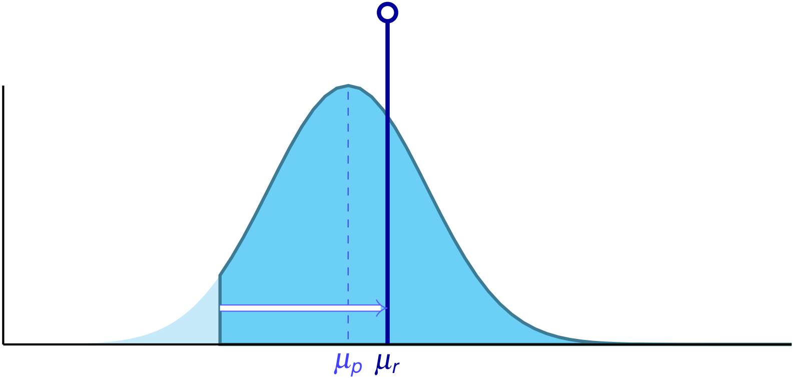


Delete ε-mass: $\mu_p \to \mu_r$.

- Alternative: **move** ε-mass **towards** μ_r.

Friendly perturbations

Consider one-dimensional case:



Delete ε-mass: $\mu_p \to \mu_r$.

- Alternative: **move** ε-mass **towards** μ_r.

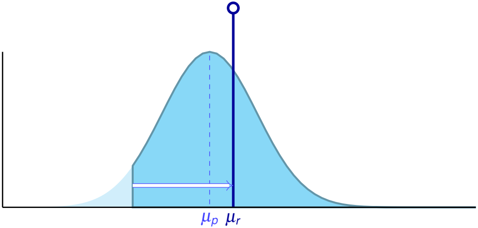
Doesn't reference deletion, defined for any W_c!

Friendly perturbation: formal definition

Definition (Friendly perturbation)

For a distribution p over X, fix a function $f : X \to \mathbb{R}$. A distribution r is an ε-*friendly perturbation* of p if there is a coupling π between p and r such that:

- The cost $\mathbb{E}_\pi[c(x, y)]$ is at most ε.
- All points move towards the mean of r: $f(y)$ is between $f(x)$ and $\mathbb{E}_r[f(x)]$ almost surely.

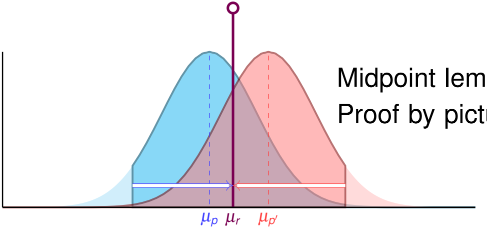


Friendly perturbation: formal definition

Definition (Friendly perturbation)

For a distribution p over X, fix a function $f : X \to \mathbb{R}$. A distribution r is an ε-*friendly perturbation* of p if there is a coupling π between p and r such that:

- The cost $\mathbb{E}_{\pi}[c(x, y)]$ is at most ε.
- All points move towards the mean of r: $f(y)$ is between $f(x)$ and $\mathbb{E}_r[f(x)]$ almost surely.



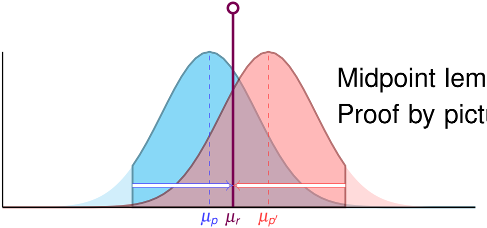
Midpoint lemma:
Proof by picture

Friendly perturbation: formal definition

Definition (Friendly perturbation)

For a distribution p over X, fix a function $f : X \to \mathbb{R}$. A distribution r is an ε-*friendly perturbation* of p if there is a coupling π between p and r such that:

- The cost $\mathbb{E}_\pi[c(x, y)]$ is at most ε.
- All points move towards the mean of r: $f(y)$ is between $f(x)$ and $\mathbb{E}_r[f(x)]$ almost surely.



Midpoint lemma:
Proof by picture

Lemma: if X has "nice topology", any p and p' with $W_c(p, p') \leq \varepsilon$ have an ε-friendly midpoint.

Resilience for W_c

Definition (Resilience for fixed f)

For any distribution p, we say that p is (ρ, ε, f)-resilient if every ε-friendly perturbation r of p has $|\mathbb{E}_r[f] - \mathbb{E}_p[f]| \leq \rho$.

Resilience for W_c

Definition (Resilience for fixed f)

For any distribution p, we say that p is (ρ, ε, f)-resilient if every ε-friendly perturbation r of p has $|\mathbb{E}_r[f] - \mathbb{E}_p[f]| \leq \rho$.

How to extend from one-dimensional f to arbitrary loss $L(p, \theta)$?

Resilience for W_c

Definition (Resilience for fixed f)

For any distribution p, we say that p is (ρ, ε, f)-resilient if every ε-friendly perturbation r of p has $|\mathbb{E}_r[f] - \mathbb{E}_p[f]| \le \rho$.

How to extend from one-dimensional f to arbitrary loss $L(p, \theta)$?

Answer: if $L(p, \theta)$ is convex in p, use Fenchel-Moreau theorem:

$$L(p, \theta) = \sup_{f \in \mathcal{F}_\theta} \mathbb{E}_p[f] - L^*(f, \theta)$$

Then apply to each f in Fenchel-Moreau representation.

Example: Linear regression

Under roughly similar assumptions to TV case, get $\varepsilon^{1-1/k}$ error assuming bounded $2(k+1)$ moments.

Example: Linear regression

Under roughly similar assumptions to TV case, get $\varepsilon^{1-1/k}$ error assuming bounded $2(k+1)$ moments.

- Error $\varepsilon^{1-1/k}$ likely suboptimal (should be $\varepsilon^{2-2/k}$).
- $k+1$ vs k in moment condition is typical behavior for W_1 vs TV

Example: Linear regression

Under roughly similar assumptions to TV case, get $\varepsilon^{1-1/k}$ error assuming bounded $2(k+1)$ moments.

- Error $\varepsilon^{1-1/k}$ likely suboptimal (should be $\varepsilon^{2-2/k}$).
- $k+1$ vs k in moment condition is typical behavior for W_1 vs TV

Finite-sample analysis:

- Can construct $\widetilde{W}_{\mathcal{H}}$ analogous to $\widetilde{\mathrm{TV}}_{\mathcal{H}}$.
- However, construction more complex and doesn't always work.
- Can at least show $\widetilde{W}_{\mathcal{H}}(p, \hat{p}_n) = \mathcal{O}((d/n)^{1/2} + (1/n)^{1/3})$ when p has bounded 3rd moments.

Summary

- Resilience criterion bounds population limit for TV perturbations.
- $\widetilde{\mathrm{TV}}_{\mathcal{H}}$ gives finite-sample analysis for projection algorithm.
- Friendly perturbations allow us to generalize resilience to W_c-perturbations.
- Many open questions for W_c case!
  - Better finite-sample analysis.
  - Efficient algorithms.
  - Beyond W_c?