Outlier-Robust Moment Estimation

Average-case

SoS-approach to avg case algo design

Last time: clustering mixtures of Gaussians

method to check purported solutions.
$\hookrightarrow$ "certifiability":

Completeness: all true sol^{n}s pass.

Soundness: if you pass then close to a true sol^n

In addition we ask:

1. checks are low deg poly inequalities.
2. Soundness claim's conclusion should also be a poly inequality
3. poly inequalities in checks $\Rightarrow$ poly inequality in soundness

low-deg SoS proof

$\rightarrow$ can efficiently find a approx. sol^n

Robust-Mean Estimation

Focus on $d=1$

D : unknown dist^n on $\mathbb{R}^d$., $\underline{Var(D) \leq 1}$., $\boxed{\mu}$ true mean.

Input: $X = \{x_1, \ldots, x_n\} \overset{iid}{\sim} D$.

approx. the $\mu(D) = \mathbb{E}_{x \sim D} x$.

Tim's algo : $\underbrace{\frac{1}{n} \Sigma_i x_i}$: empirical mean

Claim: $\mathbb{E}_X \left(\frac{1}{n} \Sigma_i x_i - \mu\right)^2 \leq \frac{1}{n}$.

Input: $Y = \{y_1, \ldots, y_n\}$. (ε-corrupted sample)

$$|Y \cap X| \geq (1-\varepsilon) \cdot n$$

$\boxed{\underline{\varepsilon} < 0.1}$

Goal: to approx. $\mu(D)$.

What happens to $\underbrace{\frac{1}{n} \Sigma_i y_i}$? Ans. arbitrarily bad!

take the median : under additional assumptions of $D \longrightarrow$ decent estimate.

$\downarrow$

don't generalize easily to even 2-D.

Robust Statistics : non-eff in dim algo in 60s
2016- eff algo for several related. problems.

$\underline{X} = \{ x_1, \dots, x_n \} \overset{iid}{\sim} D$ on $\mathbb{R}^d$. $(d=1)$

<u>assume</u>: $\frac{1}{n} \sum_i x_i = \mu$ ← our goal to estimate

$\forall v: v^T \frac{1}{n} \sum_i (x_i-\mu)(x_i-\mu)^T v \leq \|v\|_2^2$ ✓

$\equiv \frac{1}{n} \sum_i \langle x_i-\mu, v \rangle^2 \leq \|v\|_2^2$

$\frac{1}{n} \sum_i (x_i-\mu)^2 \leq 1.$ empirical variance.

<u>Input</u>: Y : $|Y \cap X| \geq (1-\varepsilon)n$ $\quad \underline{\varepsilon < 0.1}$

<u>Goal</u>: $\hat{\mu}$ so that $|\hat{\mu}-\mu| \leq ??$

Focus on <u>checking</u> procedure [Certifiability of sol^n].

<u>Given</u>. Y

<u>Oracle</u>: μ' ← "purported sol^n", $\underline{X}' = \{x_1', \dots, x_n'\}$

how to Check if μ' is a good sol^n ?

What checks to make?

① $|\{ \dots \}| \geq (1-\varepsilon)\cdot n.$

① $|\{i \mid X'_i = Y_i\}| \geq (1-2\varepsilon)n$

② $\mu' = \frac{1}{n}\sum_i X'_i$.

③ $\frac{1}{n}\sum_i (X'_i - \mu')^2 \leq 1$.

$\forall v:$
$v^T \frac{1}{n}\sum_i (X'_i - \mu')(X'_i - \mu')^T v \leq \|v\|_2^2$
$\equiv \frac{1}{n}\sum_i \langle X'_i - \mu', v\rangle^2 \leq \|v\|_2^2$

<u>Completeness</u>: (X, μ) satisfy ①, ②, ③.

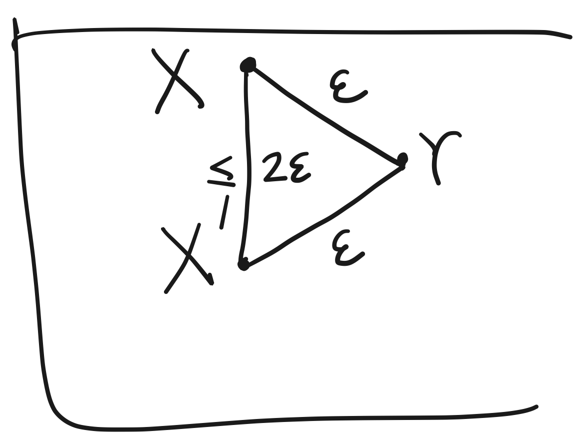
$\boxed{\text{Soundness:}}$ Suppose X', μ' satisfies ①, ②, ③.

<u>Then</u>, $|\mu' - \mu|^2 \leq \left(\frac{12\varepsilon}{1-6\varepsilon}\right)$ if $\varepsilon < \frac{1}{12}$ $\leq 24\varepsilon$

$\|\mu - \mu'\|_2^2 \leq \frac{12\varepsilon}{1-6\varepsilon}$ $\Leftarrow$ $\forall v:$ $\langle \mu' - \mu, v\rangle^2 \leq \left(\frac{12\varepsilon}{1-6\varepsilon}\right)\cdot\|v\|_2^2$

<u>Claim</u>: we get an ∞-time algo to estimate μ up to <u>$O(\sqrt{\varepsilon})$</u> error.

Proof of Soundness:



$|\mu - \mu'|^2$: $\langle \mu - \mu', v\rangle^2$

$= \left|\frac{1}{n}\sum_i X'_i - \mu\right|^2$

$$= \left| \frac{1}{n}\sum_i x_i' - \frac{1}{n}\sum_i x_i \right|^2$$

2 samples X & X' with bdd var (≤ 1) & $(1-2\varepsilon)n$ intersection must have $\leq O(\sqrt{\varepsilon})$-diff means.

$$= \left| \frac{1}{n}\sum_i (x_i - x_i') \right|^2$$

$\rightarrow \frac{1}{n}\sum_i \langle x_i - x_i', v\rangle^2$

$$= \left| \frac{1}{n}\sum_i \mathbb{1}(x_i \neq x_i') \cdot (x_i - x_i') \right|^2$$

0-1 indicator of $x_i \neq x_i'$.

$\boxed{\text{Cauchy-Schwarz}}$

$$\leq \underbrace{\left(\frac{1}{n}\sum_i \mathbb{1}(x_i \neq x_i')^2\right)}_{\leq 2\varepsilon} \underbrace{\left(\frac{1}{n}\sum_i (x_i - x_i')^2\right)}_{\leq 6 + 3(\mu-\mu')^2}$$

$\rightarrow \frac{1}{n}\sum_i \langle x_i - x_i', v\rangle^2$

$$\frac{1}{n}\sum_i (x_i - x_i')^2 = \frac{1}{n}\sum_i \left[(x_i - \mu) + (\mu - \mu') + (\mu' - x_i')\right]^2$$

$$\leq 3\left[\underbrace{\frac{1}{n}\sum_i (x_i - \mu)^2}_{} + \underbrace{\frac{1}{n}\sum_i (x_i' - \mu')^2}_{} + (\mu - \mu')^2\right]$$

Almost triangle inequality

$$(a+b)^2 \leq 2\cdot(a^2+b^2)$$

$$(a+b+c)^2 \leq 3\cdot(a^2+b^2+c^2)$$

$\leq 6 + 3(\mu-\mu')^2$

Thus: $|\mu-\mu'|^2 \leq 12\varepsilon + 6\varepsilon \cdot (\mu-\mu')^2$

$(1-6\varepsilon)(\mu-\mu')^2 \leq 12\varepsilon$

$$(\mu-\mu')^2 \leq \left[\frac{12\varepsilon}{1-6\varepsilon}\right] \qquad \square$$

Claim: There exist X & X' s.t.

① $\mathrm{Var}(X), \mathrm{Var}(X') \leq 1$

② $\frac{1}{n}\sum_i x_i = 0$, $\left|\left(\frac{1}{n}\sum_i x_i'\right)\right| = \sqrt{\varepsilon}$

③ $|X \cap X'| = (1-\varepsilon)n$.

Proof: $X: \{0, 0, \ldots, 0\}$

$\longrightarrow X': \{\underbrace{0 \ldots 0}_{(1-\varepsilon)n}, \underbrace{1/\sqrt{\varepsilon}, \ldots, \frac{1}{\sqrt{\varepsilon}}}_{\varepsilon n}\}$

$\mathrm{Var}(X') \leq$

$\frac{1}{n}\sum_i x_i'^2$

$$= \varepsilon \cdot \frac{1}{\varepsilon} = 1$$

$$\mu(X') = \frac{1}{n} \sum_i X_i' = \varepsilon \cdot \frac{1}{\sqrt{\varepsilon}} = \sqrt{\varepsilon} \quad \square$$

SoSization : proof of soundness.

Variables: X', μ'

Input: Y

① Intersection Constraints | Checks :

① $w_i^2 = w_i$, $1 \le i \le n$

② $\sum_i w_i = (1-\varepsilon) \cdot n$.

③ $\forall i$: $w_i \cdot (Y_i - X_i') = 0$.

"whenever $w_i = 1$, $Y_i = X_i'$"

② $\mu' = \frac{1}{n} \sum_i X_i'$ ④

③ $\forall v$: $\frac{1}{n} \sum_i \langle X_i' - \mu', v \rangle^2 \le \|v\|_2^2$ ⑤

$\longrightarrow$ degree 2 poly inequality.

$\boxed{\text{Quadratic system}}$: $X', \mu', w_1 — w_n$

Conclusion of soundness. $\forall v$: $\langle \mu-\mu', v\rangle^2 \leq \left(\frac{12\cdot\varepsilon}{1-6\varepsilon}\right)\cdot \|v\|_2^2$

"quadratic inequality"

Soundness [SoSized]:

$\mathcal{A} :\overset{def}{=}$ ①, ②, ③, ④, ⑤

$$\mathcal{A} \vdash^{4}_{x',\mu',w} \left\{ \langle \mu-\mu', v\rangle^2 \leq \frac{12\varepsilon}{1-6\varepsilon}\|v\|_2^2 \;\middle|\; v\in\mathbb{R}^d \right\}$$

Proof:

$$\langle \mu-\mu', v\rangle^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n}\langle x_i - x_i', v\rangle\right)^2$$

"$\underbrace{\mathbb{1}(x_i'=y_i)}$".

Obs^n 1: $w_i' = \underbrace{\mathbb{1}(x_i=y_i)}\cdot w_i$

$$\mathcal{A} \vdash^{2} \{w_i'\cdot(x_i-x_i')=0\}.$$

Pf: $\mathcal{A} \vdash^{2} \{w_i'\cdot(x_i-x_i') = w_i'\cdot(x_i-y_i)$

$$+ w_i' \cdot (y_i - x_i').$$
$$= 0. \}$$

$$= \frac{1}{n} \sum_{i=1}^{n} [w_i' + (1-w_i')] \cdot \langle x_i - x_i', v \rangle^2$$

$$\overset{\text{obs}^n 1}{=} \left(\frac{1}{n} \sum_{i=1}^{n} (1-w_i') \cdot \langle x_i - x_i', v \rangle \right)^2$$

Cauchy-Schwarz.

$$\leq \frac{1}{n} \sum_{i=1}^{n} (1-w_i')^2 \cdot \frac{1}{n} \sum_{i=1}^{n} \langle x_i - x_i', v \rangle^2$$

Lemma [SoS Cauchy-Schwarz].

$a_1, a_2, \ldots, a_n$ indeterminates

$b_1, b_2, \ldots, b_n$

$$\vdash^{4}_{\substack{a_1 - a_n \\ b_1 - b_n}} \left\{ \left(\sum_i a_i b_i\right)^2 \leq \left(\sum_i a_i^2\right) \cdot \left(\sum_i b_i^2\right) \right\}$$

Pf: $\left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right) - \left(\sum_i a_i b_i\right)^2 \leftarrow$

$$= \sum_{i,j} (a_i b_j - a_j b_i)$$ Lagrange's identity

$$\frac{1}{n} \sum_i (1-\omega_i')^2 \overset{?}{\leq} 2\varepsilon$$

$$\underset{0,1}{(1-\omega_i')^2} \leq \underbrace{(1 - \mathbb{1}(x_i = y_i))} + \underbrace{(1 - w_i \quad)}$$

$$w_i' = \underbrace{\mathbb{1}(x_i = y_i) \cdot \mathbb{1}(x_i' = y_i)}.$$

$$\frac{1}{n} \sum_i (1-\omega_i)^2 \leq \frac{1}{n} \sum_i (1 - \mathbb{1}(x_i = y_i)) + \frac{1}{n} \sum_i (1 - w_i)$$

$$\leq \varepsilon + \varepsilon = 2\varepsilon. \quad \checkmark$$

$$A \vdash^2 \{ w_i'^2 = \underbrace{\mathbb{1}(x_i = y_i)^2} \cdot w_i^2 = \mathbb{1}(x_i = y_i) \cdot w_i = w_i' \}$$

$$\Downarrow$$

$$A \vdash^2 \{ (1-\omega_i')^2 \leq (1 - \mathbb{1}(x_i = y_i)) + (1 - w_i) \}$$

$\checkmark$

fact in 1-variable: w_i

$\Rightarrow$ ~~many~~ proof $\Rightarrow$ deg 2 proof!

Almost Δ Inequality

$\vdash^2 \{ 2ab \le a^2+b^2 \}$

$\vdash^2 \{ (a+b+c)^2 = a^2+b^2+c^2+2ab+2bc+2ca$

$\overset{\text{sos}}{\le} 3a^2+3b^2+3c^2 \}$

$\boxed{\text{Algorithm}}$

① Find a $\underline{\zeta}$ p.d. of deg 4 in X', μ', w satisfying $\mathcal{A}$

$\underline{\text{Then}}$, $\underline{\forall v}$: $\underbrace{\tilde{\mathbb{E}}_{\zeta} \langle \mu' - \mu, v \rangle^2 \le \left(\frac{12\varepsilon}{1-6\varepsilon} \right) \cdot \|v\|_2^2}$

② return $\hat{\mu} = \tilde{\mathbb{E}}_{\underline{\zeta}} \mu'$

<u>Analysis:</u> $\langle \hat{\mu} - \mu, v \rangle^2$

$$= \langle \tilde{\mathbb{E}} \mu' - \mu, v \rangle^2$$

<u>Claim:</u> $\langle \tilde{\mathbb{E}}_{\zeta} \mu' - \mu, v \rangle^2 \leq \tilde{\mathbb{E}}_{\zeta} \langle \mu' - \mu, v \rangle^2$

$$\leq \left(\frac{12\varepsilon}{1 - 6\varepsilon} \right) \cdot \|v\|_2^2$$

<u>Jensen:</u> $\forall f: \; (\tilde{\mathbb{E}}_{\zeta} f)^2 \leq \tilde{\mathbb{E}}_{\zeta} f^2 \longleftarrow$

$$f = \langle \mu' - \mu, v \rangle$$

<u>Lemma [Pseudo-expectation C.S]</u>

$\forall$ $\underline{\deg 2k \;\; \zeta}$, $\underline{f}, \underline{g}$: poly$^{\deg} \leq k$

$$\tilde{\mathbb{E}}\, f \cdot g \leq \underline{\sqrt{\tilde{\mathbb{E}} f^2} \cdot \sqrt{\tilde{\mathbb{E}}\, g^2}} \rightarrow 1$$

$f = \underline{\langle \mu' - \mu, v \rangle}$, $\underline{g = 1}$

<u>Proof:</u> Suppose: $\tilde{\mathbb{E}} f^2 = 0$ $\quad \forall \underline{\delta > 0}$.

$$\tilde{\mathbb{E}} (f - \delta g)^2 \geq 0$$

We know: $\mathbb{E}(f+\delta\cdot g)$

$$\delta^2\cdot\tilde{\mathbb{E}}g^2 \geq -2\delta\tilde{\mathbb{E}}f\cdot g$$

$$\tilde{\mathbb{E}}(f-\delta g)^2 \geq 0$$

$$\delta^2\cdot\tilde{\mathbb{E}}g^2 \geq 2\delta\cdot\tilde{\mathbb{E}}\cdot f\cdot g$$

$$\lim_{\delta\to 0}\left|\hat{\mathbb{E}}\cdot fg\right| \leq \frac{\delta}{2}\cdot\tilde{\mathbb{E}}\cdot g^2$$

$$= 0$$

Suppose: $\tilde{\mathbb{E}}f^2, \tilde{\mathbb{E}}g^2 > 0$.

$$\bar{f} = \frac{f}{\sqrt{\tilde{\mathbb{E}}f^2}}, \quad \bar{g} = \frac{g}{\sqrt{\tilde{\mathbb{E}}g^2}}$$

Then, $\tilde{\mathbb{E}}\cdot(\bar{f}+\bar{g})^2 \geq 0$

$$\hat{\mathbb{E}}(\bar{f}-\bar{g})^2 \geq 0$$

$$2\left|\tilde{\mathbb{E}}\cdot\bar{f}\cdot\bar{g}\right| \leq \hat{\mathbb{E}}\cdot\bar{f}^2 + \hat{\mathbb{E}}\cdot\bar{g}^2$$

$$= 2$$

$$\Rightarrow \left|\hat{\mathbb{E}}\, \bar{f}\cdot\bar{g}\right| \leq 1$$

$$\Rightarrow \left|\hat{\mathbb{E}}\, f\cdot g\right| \leq \sqrt{\hat{\mathbb{E}} f^2}\sqrt{\hat{\mathbb{E}} g^2} \qquad \square$$

Succinct representations of constraints

$$\frac{1}{n}\sum_i \langle x_i' - \mu', v\rangle^2 \leq \|v\|_2^2$$

Can add because:

$$\frac{1}{n}\sum_i \langle x_i - \mu, v\rangle^2 \leq \|v\|_2^2$$

$\bar{x}_i = x_i - \mu$ $\quad \forall v:\ v^T\left(\frac{1}{n}\sum_i \bar{x}_i\cdot\bar{x}_i^T\right)v \leq \|v\|_2^2$ — $(*)$

$$\Leftrightarrow \quad \mathbb{I} - \frac{1}{n}\sum_i \bar{x}_i\cdot\bar{x}_i^T \succeq 0$$

"quantifiers elimination over v"

$$\Leftrightarrow \quad \mathbb{I} - \frac{1}{n}\sum_i \bar{x}_i\cdot\bar{x}_i^T = V\cdot V^T$$

for some matrix V.

Add constraint: $\mathbb{I} - \frac{1}{n} \cdot \sum_i (x_i' - \mu')(x_i' - \mu')^T$

$= V \cdot V^T$

V: $d \times d$ matrix-valued indeterminate

$\boxed{\text{Our final system:}}$ x', μ', w, V

Last time: $\forall v: \frac{k}{n} \cdot \sum_i w_i \langle x_i - \frac{k}{n} \sum_i w_i x_i, v \rangle^{2t}$

$\leq \underbrace{1 \cdot 1}_{} \cdot \underbrace{(2t-1)!!}_{\sim t^t} \cdot \underbrace{\|v\|_2^{2t}}_{}$

Certifiable-t-Subgaussian dist^n

$z_1, \ldots, z_n$ are t-certifiably

Subgaussian if

$\vdash_v^{2t} \left\{ \underbrace{\frac{1}{n} \cdot \sum_i \langle z_i - \frac{1}{n} \sum_i z_i, v \rangle^{2t}}_{} \leq \underbrace{(t)}_{}^t \underbrace{\left(\frac{1}{n} \sum_i \langle z_i - \frac{1}{n} \sum_i z_i, v \rangle^2 \right)^t}_{} \right\}$

E.g. Gaussians are certifiably subgaussian with $C=1 \ \forall t$.

$$(1,v)^{\otimes t} \cdot \underbrace{P \cdot P^T}_{} (1,v)^{\otimes t}$$

$$\frac{1}{n}\sum_i \left\langle z_i - \frac{1}{n}\sum_i z_i, v\right\rangle^{2t} + \underbrace{SOS(v)}_{} = (Ct)^t \cdot \left(\frac{1}{n}\sum_i \left\langle z_i - \frac{1}{n}\sum_i z_i, v\right\rangle^2\right)^t$$

$$\frac{1}{n}\sum_i \left(z_i - \frac{1}{n}\sum_i z_i\right)^{\otimes 2t} + \underline{P \cdot P^T} = (Ct)^t \cdot \left[\frac{1}{n}\sum_i \left(z_i - \frac{1}{n}\sum_i z_i\right)^{\otimes 2}\right]^{\otimes t}$$

Idea:

$$\text{Inequality} \xrightarrow[\text{proof}]{\text{SOS}} \text{Equality}$$

$\downarrow$ SOS polynomials

$\equiv$ quadratic form of PSD matrix

$=$ PSDness

Improved algo (Sketch)

Let $X = \{x_1 \text{ --- } x_n\}$, $\mu = \frac{1}{n}\sum_i x_i$

$$\frac{1}{n}\sum_i \langle x_i - \mu, v\rangle^2 \le 1$$

& $\frac{1}{n}\sum_i \langle x_i-\mu, v\rangle^t$ ✓

$$\le C\cdot\left(\frac{1}{n}\sum_i\langle x_i-\mu, v\rangle^2\right)^2$$

$\downarrow$

$$\frac{1}{n}\sum_i\langle x_i-\mu, v\rangle^{2t} \le (Ct)^t\left(\frac{1}{n}\sum_i\langle x_i-\mu, v\rangle^2\right)^t$$ ✓

Thm: X : t - Subgaussian in addition.

then $\deg(4t^2)$ - SoS finds an estimate $\hat{\mu}$ s.t. $|\langle\hat{\mu}-\mu, v\rangle| \le O\left(t\varepsilon^{1-\frac{1}{2t}}\right)$.

$\downarrow$ dependence tight

Test: 4

$$\frac{1}{n}\sum_i\langle x_i'-\mu', v\rangle^{2t} \le (Ct)^t\cdot\left(\frac{1}{n}\sum_i\langle x_i'-\mu', v\rangle^2\right)^t$$

Soundness proof: $\rightarrow$ key step replace Cauchy-Schwarz

$$\langle\mu'-\mu, v\rangle^{2t} = \left(\frac{1}{n}\sum_i\langle x_i-x_i', v\rangle\right)^{2t}$$

$$= \left(\frac{1}{n}\cdot\sum_i\cdot\mathbb{1}(x_i\neq x_i')^{2t-1}\cdot\langle x_i-x_i', v\rangle\right)^{2t}$$

$\left(\frac{}{n} \ldots\right)$

$$\underset{\text{inequality}}{\overset{\text{Hölder's}}{\leq}} \left(\frac{1}{n}\sum_i \mathbb{1}(x_i \neq x_i')\right)^{\frac{2t-1}{2t}} \cdot \left(\frac{1}{n}\sum_i \langle x_i - x_i', v\rangle^{2t}\right)$$

$\underline{\forall t}$: $\sum_i a_i b_i \leq \underbrace{\left(\sum_i a_i^{\frac{2t}{2t-1}}\right)^{\frac{2t-1}{2t}}}\cdot \underbrace{\left(\sum_i b_i^{2t}\right)^{1/2t}}$

$$4t^2 \left(\sum_i a_i^{2t-1} \cdot b_i\right)^{2t} \leq \left(\sum_i a_i^{2t}\right)^{2t-1} \cdot \left(\sum_i b_i^{2t}\right)$$

deg inequality

$\rightarrow$ SOS Hölder's Inequality

$$\left\{ \mathbb{I} - \frac{1}{n}\sum_i (x_i' - \mu')(x_i' - \mu')^T = VV^T \right\} \subseteq \mathcal{A}$$

$\underline{\text{then}}$:

$$\mathcal{A} \vdash_v^2 \left\{ \underline{\|v\|_2^2 \geq \frac{1}{n}\sum_i \langle x_i' - \mu', v\rangle^2} \right\}$$

$$\mathcal{A} \vdash_v^2 \left\{ \|v\|_2^2 - \frac{1}{n}\sum_i \langle x_i' - \mu', v\rangle^2 = \underbrace{\|Vv\|_2^2} \right\}$$

Chapter 4.3 of "Semialgebraic Proofs

Chapter

& Efficient Algo Design".