Robust Mean Estimation

Why Should We Care About Robustness

- model misspecification

- crowdsourcing data
  * Quality vs Quantity tradeoff.

- Data poisoning attacks

Contamination Models

① No Contamination

Samples $X_1, \ldots X_n$ are drawn from $N(\mu, I_d)$ independently & identically.

② Strong Contamination

Samples $X_1, \ldots X_n$ drawn as before, except adversary may remove εn points & add εn points as she wants!

Contamination Models

① No Contamination

Samples $X_1, \dots X_n$ are drawn from $N(\mu, I_d)$ independently & identically.

$\boxed{\text{Goal: output } \hat{\mu} \text{ close to true } \mu}$

② Strong Contamination

Samples $X_1, \dots X_n$ drawn as before, except adversary may **remove** εn points & **add** εn points as she wants!

No Contamination

Thm (Chernoff - Hoeffding)

Let $X_1, \ldots, X_n$ be (sub) Gaussian r.v.'s & independent

w/ mean $\mathbb{E} X_i = \mu$.

$$\Rightarrow \Pr\left[\frac{1}{n}\sum X_i - \mu \geq t\right] \leq O\left(\exp(-nt^2)\right)$$

ex Bernoulli

ex Gaussian

ex Bounded

No Contamination

Thm (Chernoff - Hoeffding)

Let $X_1, \ldots X_n$ be (sub) Gaussian r.v.'s & independent w/ mean $\mathbb{E}\, X_i = \mu$.

$$\Rightarrow \Pr\left[\frac{1}{n}\sum X_i - \mu \geq t\right] \leq O\left(\exp(-nt^2)\right)$$

ex Bernoulli

ex Gaussian

ex Bounded

- After $n \geq \Omega\left(1/\varepsilon^2\right)$ samples, estimate up to $\pm\,\varepsilon$ in 1-D
- After $n \geq \Omega\left(d/\varepsilon^2\right)$ samples, " " " " in d-D.

Is The Mean Robust?

No! A single corrupted point causes drift arbitrarily far from μ!

(take some point to ∞!)

but ...

Is the Mean Robust?

No! A single corrupted point causes drift arbitrarily far from μ!

(take some point to ∞!)

but ...

the median naturally ignores points far from μ

or any order statistic

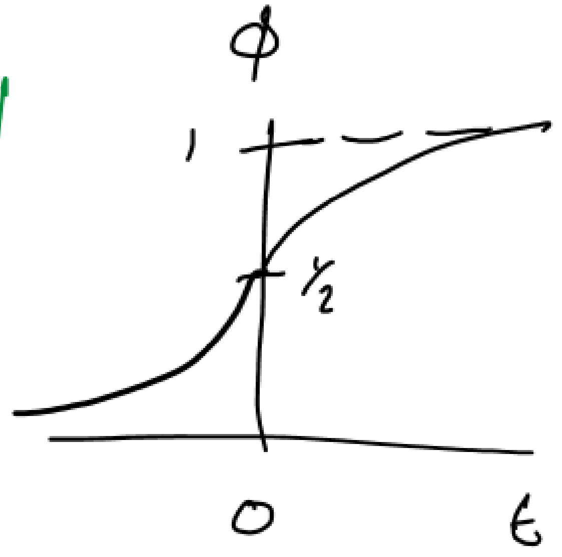
$\Rightarrow$ **up next:** the median achieves $O(\varepsilon)$ error in 1-D.

Analysis of the Median

Thm Let ϕ be the Gaussian CDF,

$$\mathbb{P}\left[|\mathrm{med}(S)-\mu| > t\sigma\right] \le 2\exp\left(-2n\left(\phi(t)-\frac{1}{2}-\varepsilon\right)^2\right)$$

$$\Rightarrow \mathbb{P}\left[|\mathrm{med}(S)-\mu| > (1+b)\sqrt{2\pi}\,\varepsilon\right] \le 2\exp\left(-\Omega(b n \varepsilon^2)\right)$$

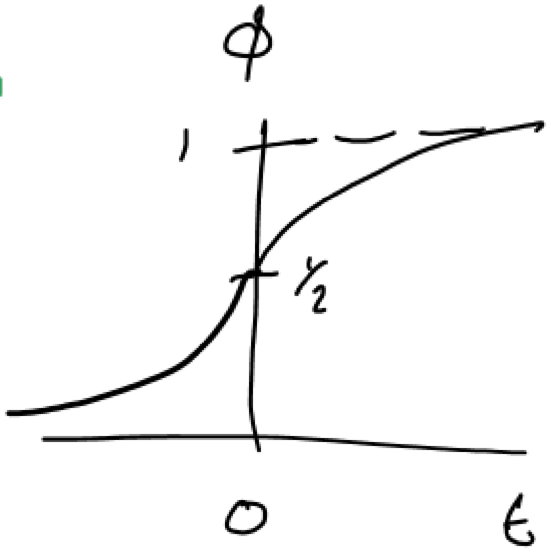


Analysis of the Median

Thm Let ϕ be the Gaussian CDF,

$$\mathbb{P}\left[|\mathrm{med}(S)-\mu| > t\sigma\right] \leq 2\exp\left(-2n\left(\phi(t)-\frac{1}{2}-\varepsilon\right)^2\right)$$

$$\Rightarrow \mathbb{P}\left[|\mathrm{med}(S)-\mu| > (1+b)\sqrt{2\pi}\,\varepsilon\right] \leq 2\exp\left(-\Omega(bn\varepsilon^2)\right)$$



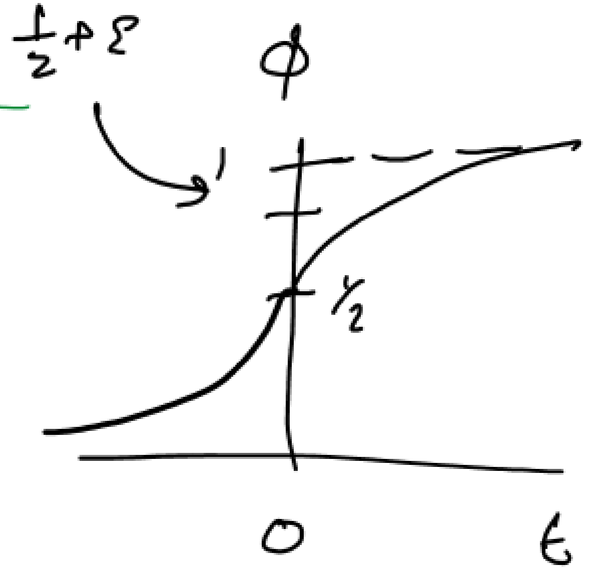
Thus, $|\mathrm{med}(S)-\mu| = \boxed{O(\varepsilon)}$ as long as $n = \Omega\left(1/\varepsilon^2\right)$.

optimal rate!!

Analysis of the Median

Thm Let ϕ be the Gaussian CDF,

$$\mathbb{P}\left[|\mathrm{med}(S) - \mu| > t\sigma\right] \leq 2\exp\left(-2n\left(\phi(t) - \tfrac{1}{2} - \varepsilon\right)^2\right)$$



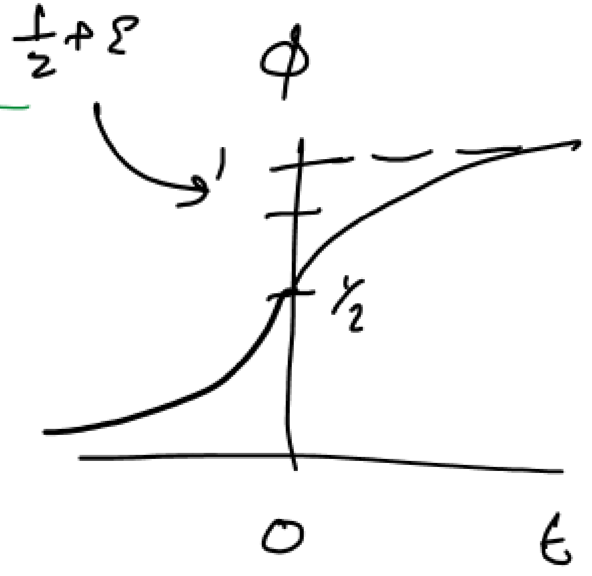
Pf After ε-corruption, the median can move only up to $1/2 + \varepsilon$ of the original CDF

$\Rightarrow |\mathrm{med}(S) - \mu| > t\sigma$ if & only if $1/2 - \varepsilon$ fraction of pts exceed $\mu + t\sigma$

Analysis of the Median

Thm Let ϕ be the Gaussian CDF,

$$\mathbb{P}\left[|\mathrm{med}(S) - \mu| > t\sigma\right] \leq 2\exp\left(-2n\left(\phi(t) - \tfrac{1}{2} - \varepsilon\right)^2\right)$$



Pf After ε-corruption, the median can move only up to $1/2 + \varepsilon$ of the original CDF

$\Rightarrow |\mathrm{med}(S) - \mu| > t\sigma$ if & only if $1/2 - \varepsilon$ fraction of pts exceed $\mu + t\sigma$

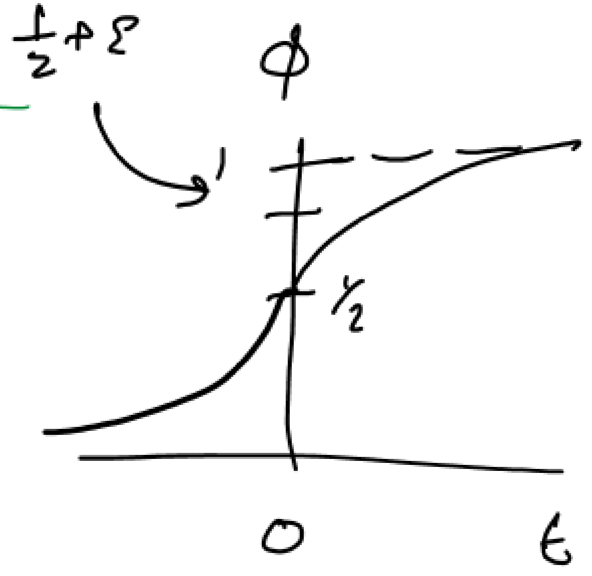
the probability of each event is $1 - \phi(t)$:

$$\mathbb{P}\left[\frac{1}{n}\sum Y_i \geq 1 - \phi(t) + s\right] \leq \exp(-2ns^2)$$

Analysis of the Median

Thm Let ϕ be the Gaussian CDF,

$$\mathbb{P}\left[|\mathrm{med}(S) - \mu| > t\sigma\right] \leq 2\exp\left(-2n\left(\phi(t) - \frac{1}{2} - \varepsilon\right)^2\right)$$



Pf After ε-corruption, the median can move only up to $1/2 + \varepsilon$ of the original CDF

$\Rightarrow |\mathrm{med}(S) - \mu| > t\sigma$ if & only if $1/2 - \varepsilon$ fraction of pts exceed $\mu + t\sigma$

the probability of each event is $1 - \phi(t)$:

$$\mathbb{P}\left[\frac{1}{n}\sum Y_i \geq 1 - \phi(t) + s\right] \leq \exp(-2ns^2)$$

$\Rightarrow$ taking $s = \phi(t) - \frac{1}{2} - \varepsilon$ gives the desired result!

How can we generalize to d-dimensions?

<u>Note</u>: the goal is to learn μ to min l_2 error

How can we generalize to d-dimensions?

Note: the goal is to learn μ to min ℓ_2 error

① Compute median in all directions . . .

. . . this gives $O(\varepsilon)$ in each, so $O(\varepsilon\sqrt{d})$ total.

How can we generalize to d-dimensions?

Note: the goal is to learn μ to w/in ℓ_2 error

① Compute median in all directions...

... this gives $O(\varepsilon)$ in each, so $O(\varepsilon\sqrt{d})$ total.

② "Tukey Median" gives $O(\varepsilon)$ total, but requires computing ∞-many medians*

How can we generalize to d-dimensions?

Note: the goal is to learn μ to w/in ℓ_2 error

① Compute median in all directions...

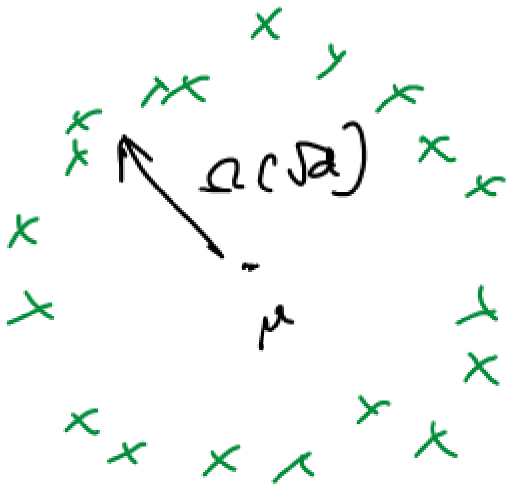
... this gives $O(\varepsilon)$ in each, so $O(\varepsilon\sqrt{d})$ total.

② "Tukey Median" gives $O(\varepsilon)$ total, but requires computing ∞-many medians*

③ Polytime alternatives that give $O(\varepsilon\sqrt{\log(1/\varepsilon)})$ total!!

High-Dimensional Intuitions

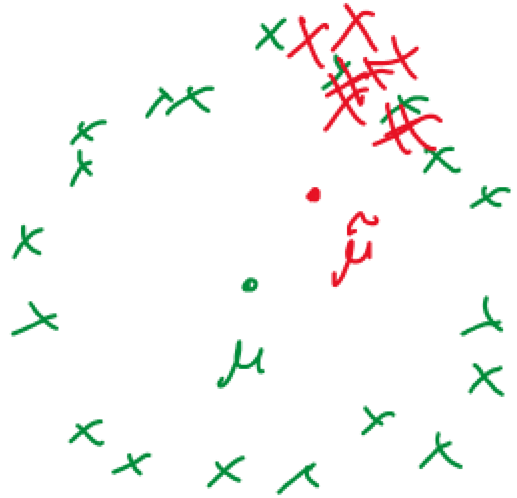
- What is the "adversarial" strategy in $\mathbb{R}^d$?



Fact: $X \sim N(\mu, I_d)$ concentrates around $\Theta(\sqrt{d})$ from μ, w/ error $O(1)$

High-Dimensional Intuitions

- What is the "adversarial" strategy in $\mathbb{R}^d$?

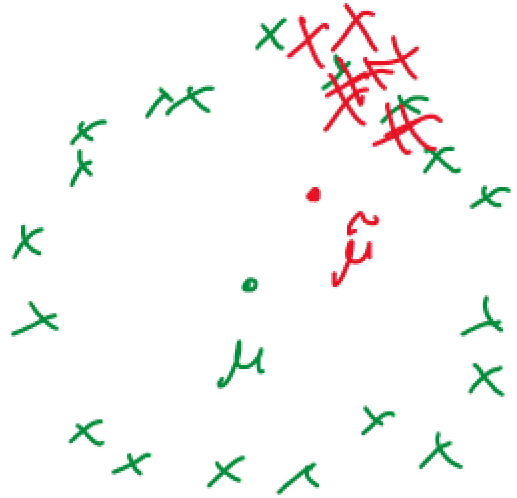


Fact: $X \sim N(\mu, I_d)$ concentrates around $\Theta(\sqrt{d})$ from μ, w/ error $O(1)$

Rather than putting points at ∞, the adversary has to "clump" points to shift the mean!

High-Dimensional Intuitions

- What is the "adversarial" strategy in $\mathbb{R}^d$?



$\boxed{\text{Fact: } X \sim N(\mu, I_d) \text{ concentrates around } \Theta(\sqrt{d}) \text{ from } \mu, \text{ w/ error } O(1)}$

Rather than putting points at ∞, the adversary has to "clump" points to shift the mean!

* If $\mu \to \tilde{\mu}$ by corruption, the "variance" in direction $\hat{\mu} - \mu$ must be large

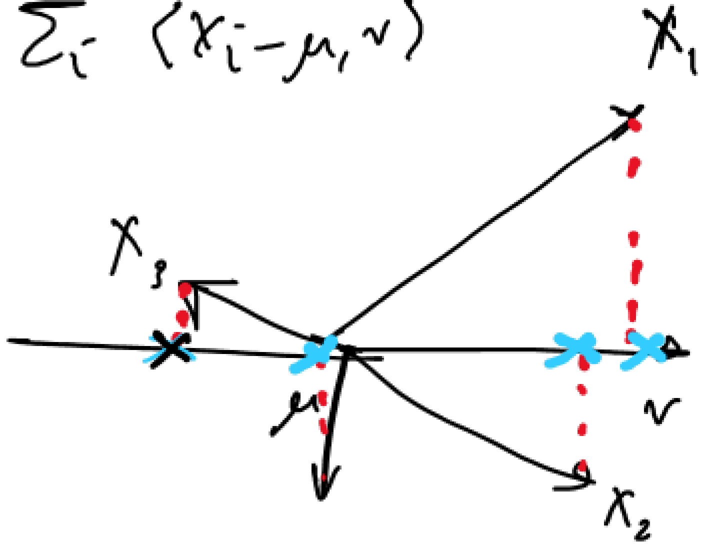
Covariance Function

Definition The covariance matrix of $X_1, \dots X_n \in \mathbb{R}^d$ is

$$\Sigma = \frac{1}{n} \sum_i (X_i - \mu)(X_i - \mu)^T ; \quad \mu = \frac{1}{n} \sum_{i=1} X_i$$

The covariance naturally encodes variance along <u>all</u> directions

$$v^T \Sigma v = \frac{1}{n} \sum_i v^T (X_i - \mu)(X_i - \mu)^T v = \frac{1}{n} \sum_i \langle X_i - \mu, v \rangle^2$$



Covariance Function

Definition The covariance matrix of $X_1, .. X_n \in \mathbb{R}^d$ is

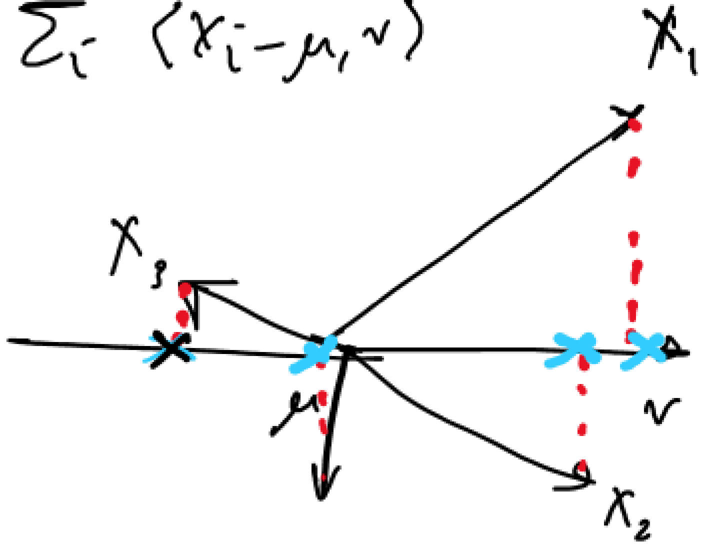
$$\Sigma = \frac{1}{n} \sum_i (X_i - \mu)(X_i - \mu)^T \; ; \quad \mu = \frac{1}{n} \sum_{i=1} X_i$$

The covariance naturally encodes variance along <u>all</u> directions

$$v^T \Sigma v = \frac{1}{n} \sum_i v^T (X_i - \mu)(X_i - \mu)^T v = \frac{1}{n} \sum_i \langle X_i - \mu, v \rangle^2$$

So, the "top principal component" is

$\underset{\|v\|=1}{\text{argmax}} \; v^T \Sigma v$, the direction of greatest variance.

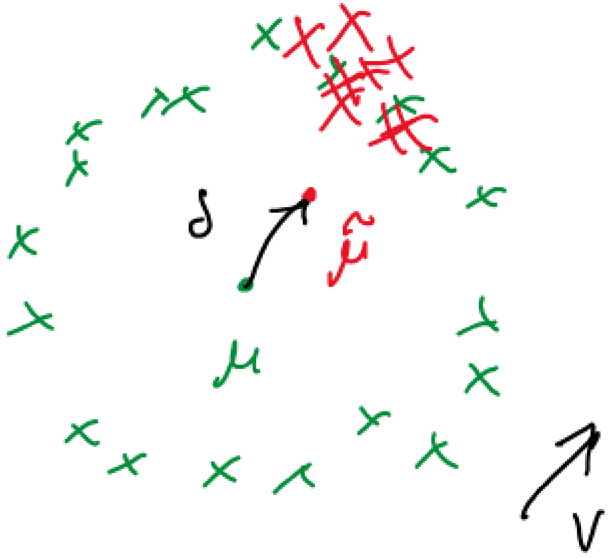


High-Dimensional Intuitions

If the corrupted points move μ by distance δ

$\Rightarrow$ corrupted points have projection onto v

at least δ/ε

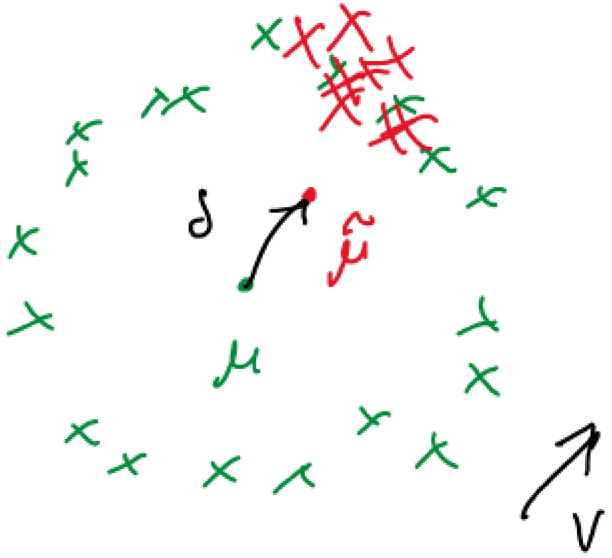


High-Dimensional Intuitions

If the corrupted points move μ by distance δ

$\Rightarrow$ corrupted points have projection onto v

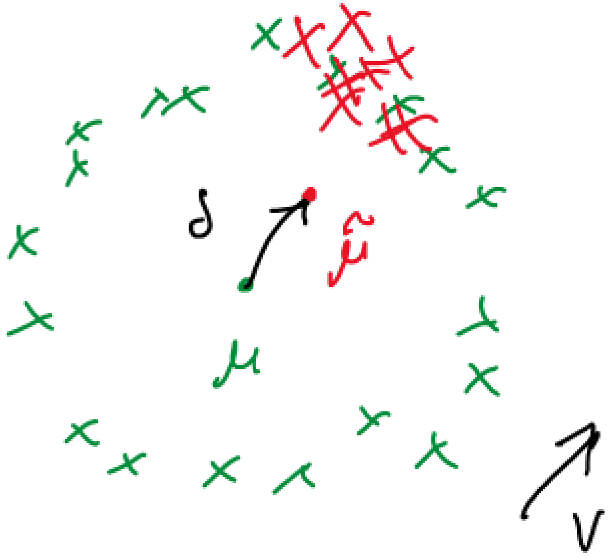
at least δ/ε.



$\Rightarrow$ corrupted points have $\varepsilon \cdot (\delta/\varepsilon)^2 = \delta^2/\varepsilon$ variance onto v.

High-Dimensional Intuitions

If the corrupted points move μ by distance δ



$\Rightarrow$ corrupted points have projection onto v

at least δ/ε.

$\Rightarrow$ corrupted points have $\varepsilon \cdot (\delta/\varepsilon)^2 = \delta^2/\varepsilon$ variance onto v.

__Conjecture:__ A "random" set satisfies that if the corrupted version has "**small**" operator norm

$\Rightarrow$ $\|\mu_T - \mu_X\|_2$ is "small".

Stability –

We formalize the notion of "random" as a stable set

Defn (ε, δ)-stable wrt X for $\varepsilon \in (0, \frac{1}{2})$, $\delta \geq \varepsilon$ if $\forall v \in S^{d-1}$, $S' \subseteq S$ of size at least $(1-\varepsilon)|S|$,

1. $\left| \frac{1}{|S'|} \sum_{x \in S'} \langle v, x - \mu_X \rangle \right| \leq \delta$

2. $\left| \frac{1}{|S'|} \sum_{x \in S'} \langle v, x - \mu_X \rangle^2 - 1 \right| \leq \delta^2/\varepsilon$

variance bound

mean bound

Intuition: Ability to "match" moments on a sample up to deletions

Stability

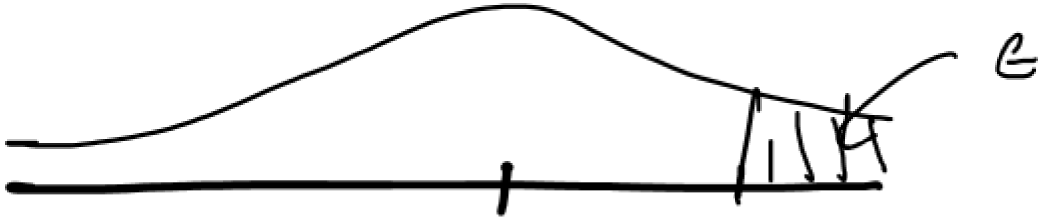
We formalize the notion of "random" as a stable set

Def^n (ε,δ)-stable wrt X for $\varepsilon \in (0, \tfrac{1}{2})$, $\delta \geq \varepsilon$ if $\forall v \in S^{d-1}$, $S' \subseteq S$ of size at least $(1-\varepsilon)|S|$,

1. $\left| \frac{1}{|S'|} \sum_{x \in S'} \langle v, x - \mu_X \rangle \right| \leq \delta$ $\rightsquigarrow \varepsilon\sqrt{\log(1/\varepsilon)}$

2. $\left| \frac{1}{|S'|} \sum_{x \in S'} \langle v, x - \mu_X \rangle^2 - 1 \right| \leq \delta^2/\varepsilon$ $\rightsquigarrow \varepsilon . \log(1/\varepsilon)$

Prop: A subgaussian distribution is $(\varepsilon, O(\varepsilon\sqrt{\log 1/\varepsilon}))$-stable



Thm (Conjecture)

Let S be (ϵ,δ) set wrt X, let T be ϵ-corrupted of S

if μ_T, Σ_T are mean & variance, $\|\Sigma_T\| \leq 1+\lambda$

$$\Rightarrow \|\mu_T - \mu_X\|_2 \leq O(\delta + \sqrt{\epsilon\lambda})$$

Cor If W is a distribution differing from S by TV of ϵ.

if $\|\Sigma_W\|_2 \leq 1+\lambda \Rightarrow \|\mu_W - \mu_X\|_2 \leq O(\delta + \sqrt{\epsilon\lambda})$

Intuition: W is a "weight" corresponding to each datapt of "confidence in S"

Def^n

$$C = \left\{ W : W \text{ supp on } T,\ W(x) \leq \frac{1}{|T|(1-\epsilon)} \right\}$$

Defⁿ $C = \left\{ W : W \text{ supp on } T,\ W(x) \le \dfrac{1}{|T|(1-\epsilon)} \right\}$

Prop If, for some $W \in C$, Σ_W has no large eigenvalue

$\Rightarrow$ μ_W is a good approx to μ_X

(by previous lemma)

Def^n $\mathcal{C} = \left\{ W : W \text{ supp on } T,\ W(x) \leq \frac{1}{|T|(1-\epsilon)} \right\}$

★ Take $w^* = \text{unif}(T \cap S)$

By property: $\|\Sigma_{w^*}\| \leq 1 + \delta^2/\epsilon$

gives ℓ_2 error $O(\delta)$ for

S being $(3\epsilon, \delta)$-stable!

Prop If, for some $W \in \mathcal{C}$, Σ_W has no large eigenvalue

$\Rightarrow \mu_W$ is a good approx to μ_X

(by previous lemma)

Summary

1 Consistency condition w/ removing samples

* (ε, δ)-stable sets

Summary

1 Consistency condition w/ removing samples

* (ε, δ)-stable sets

2 Consistency condition happens whp on subGaussian distr

* for n large, $(\epsilon, O(\epsilon\sqrt{\log 1/\epsilon}))$-stable

3 Corruption of consistency spectrally certifies closeness in ℓ_2

* Any TV-close corruption shows a covariance $\to \ell_2$ reduction

Summary

1 Consistency condition w/ removing samples

* (ε, δ)-stable sets

2 Consistency condition happens whp on subGaussian distr

* for n large, $(\epsilon, O(\epsilon\sqrt{\log 1/\epsilon}))$-stable

3 Corruption of consistency spectrally certifies closeness in ℓ_2

* Any TV-close corruption shows a covariance $\to \ell_2$ reduction

4 Produce a convex set w/ the optimum being the **set** of uncorrupted points.

Towards Robust Algorithms

Reduced mean estimation to a "convex" problem on the set $\mathcal{C}$ of weight distributions

strictly speaking, not true, but technique applies.

Towards Robust Algorithms

Reduced mean estimation to a "convex" problem on the set $\mathcal{C}$ of weight distributions

Use the "ellipsoid method" to approach W^*

Towards Robust Algorithms

Reduced mean estimation to a "convex" problem on the set $\mathcal{C}$ of weight distributions

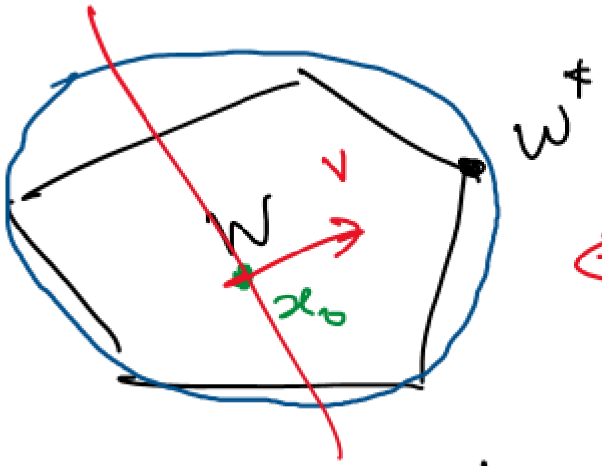
Use the "ellipsoid method" to approach w^*

feasible set &
optima

Towards Robust Algorithms

Reduced mean estimation to a "convex" problem on the set $\mathcal{C}$ of weight distributions.

Use the "ellipsoid method" to approach w^*



separation oracle gives a direction st $\langle v, x_0 \rangle < \langle v, w^* \rangle$

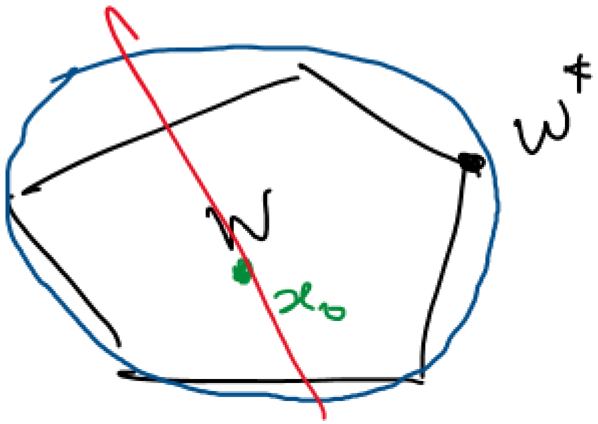
feasible set & optima

& center of ellipse

Towards Robust Algorithms

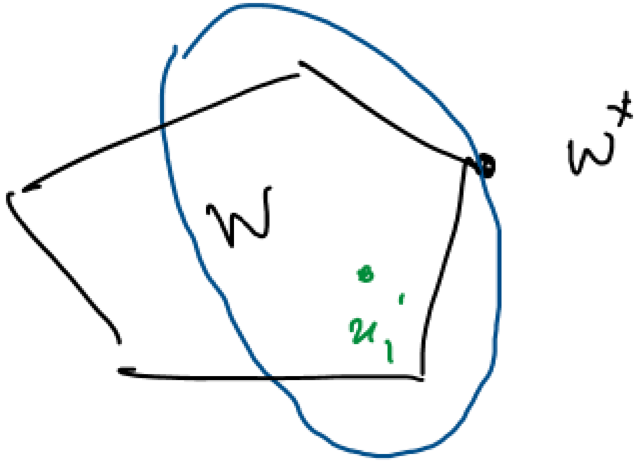
Reduced mean estimation to a "convex" problem on the set $\mathcal{C}$ of weight distributions.

Use the "ellipsoid method" to approach w^*



feasible set & optima
& center of ellipse

$\Longrightarrow$



new feasible set & ellipse.

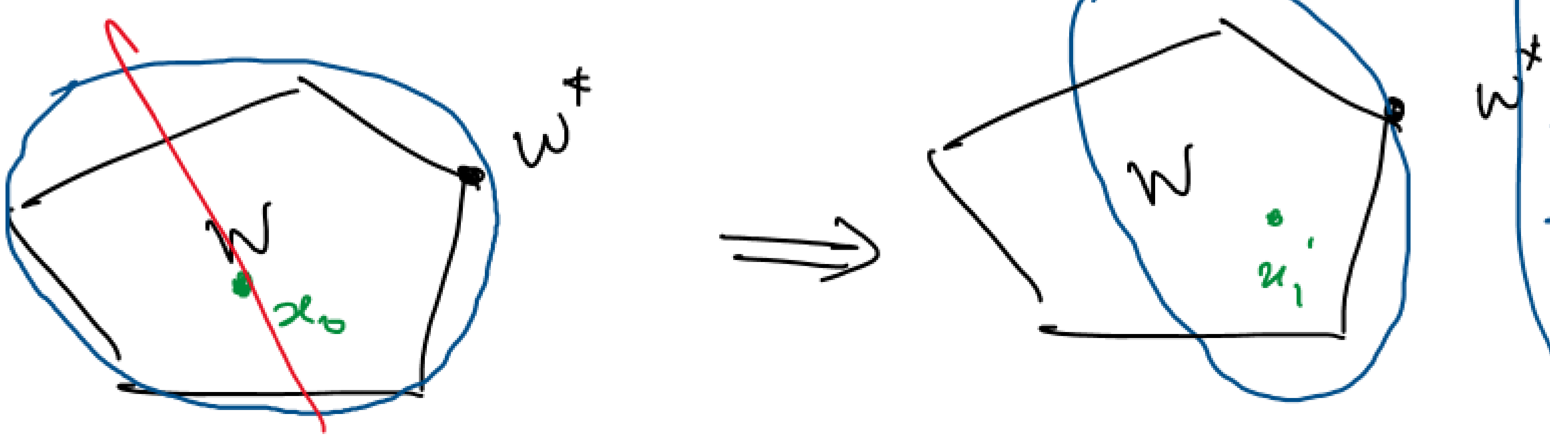
Towards Robust Algorithms

Reduced mean estimation to a "convex" problem on the set $\mathcal{C}$ of weight distributions

Use the "ellipsoid method" to approach w^*.

Without detail

$\exists$ a separation oracle for this nonconvex problem $\Rightarrow$ poly(d) iters suffice to approach w^*.



feasible set & optima & center of ellipse

new feasible set & ellipse.

Towards Robust Algorithms

An even <u>simpler</u> (but harder to analyze) approach is via <u>filtering</u>

Towards Robust Algorithms

An even simpler (but harder to analyze) approach is via filtering

Alg

1 If no directions have large variance, output empirical mean

By ~~pop~~ on stable sets

Towards Robust Algorithms

An even simpler (but harder to analyze) approach is via filtering

Alg

1. If no directions have large variance, output empirical mean

2. Throw out data along the high variance direction that is "too far" from the mean

"trimmed mean" analogue

Towards Robust Algorithms

An even simpler (but harder to analyze) approach is via filtering

Alg

1. If no directions have large variance, output empirical mean
2. Throw out data along the high variance direction that is "too far" from the mean

Analysis: Each time we throw out data, the density of good points increases!

Lemma 3.3 (informal). *Let* $S = G \cup E \backslash S_r$ be an ε-corrupted set of points from $\mathcal{N}(\mu, I)$ *of size at least* $\tilde{\Omega}(d/\varepsilon^2)$. *Then, with probability at least* 0.99 *and after a simple preprocessing step, the filter satisfies the following property: Given any* $S' \subseteq S$ *satisfying* $\Gamma(G, S') \leq 2\varepsilon$, *the filter either*

value of δ

(a) outputs $\hat{\mu}$ *so that* $\|\mu - \hat{\mu}\|_2 \leq \boxed{O(\varepsilon\sqrt{\log 1/\varepsilon})}$, *or*

(b) outputs T *so that* $\Gamma(G, T) \leq \Gamma(G, S') - \varepsilon/\alpha$, *where* $\alpha = d \log(d/\varepsilon) \log(d \log(d/\varepsilon))$.

iteration complexity

Thanks For Listening!