

Foundations of Trustworthy AI

Jonathan Ullman
CS7180/7880 · Fall 2026

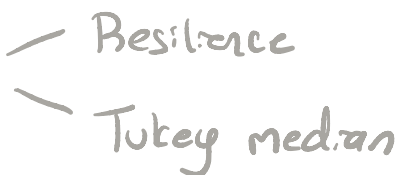
Class 11: Efficient Robust Means

Oct 16, 2026

Robust Estimation

One dimensional mean 

- Median
- Trimmed means

High dimensional mean
(inefficient) 

- Resilience
- Tukey median

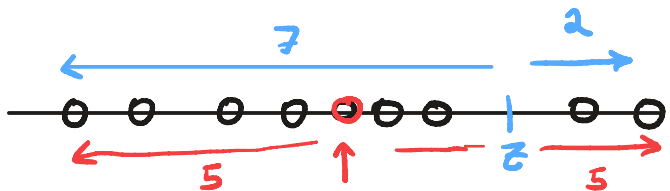
High dimensional mean
(efficient) — Spectral filtering

Robust optimization

Tukey Medians

Median in $\mathbb{R}$

Median is any point of maximum depth



$$\text{depth}(z) = 2/9$$

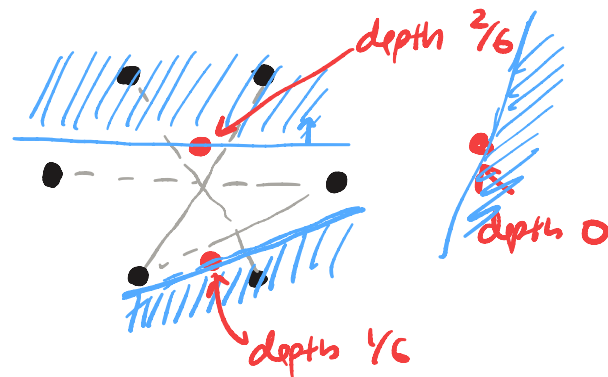
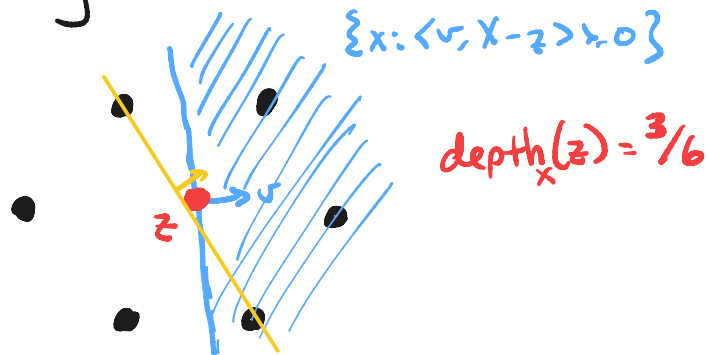
$$\text{depth}(z) = \frac{1}{n} \min \left\{ \begin{array}{l} \# \text{ of pts above } z, \\ \# \text{ of pts below } z \end{array} \right\}$$

$$\min \left\{ \frac{1}{n} \sum \mathbb{1}\{x_i - z \geq 0\}, \frac{1}{n} \sum \mathbb{1}\{-(x_i - z) \geq 0\} \right\}$$

minimum over all
halfspaces thru z of # of
pts in that halfspace

Tukey Medians

Tukey Median in $\mathbb{R}^2$



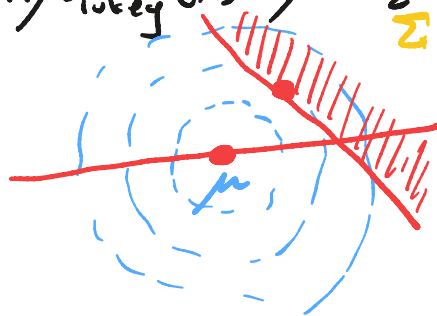
$$\text{depth}_x(z) = \min_{v \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\langle v, x_i - z \rangle > 0\}$$

$$\mu_{\text{Tukey}}(X) \in \arg\max_{z \in \mathbb{R}^d} \text{depth}_x(z)$$

Tukey Median for Gaussians

Lemma: If $X = (X_1, \dots, X_n) \sim \mathcal{N}(\mu, \Sigma)$ then with prob $\geq 1 - \delta$

$$\|\mu_{\text{Tukey}}(X) - \mu\|_{\Sigma}^2 = O\left(\frac{d + \log(1/\delta)}{n}\right)$$



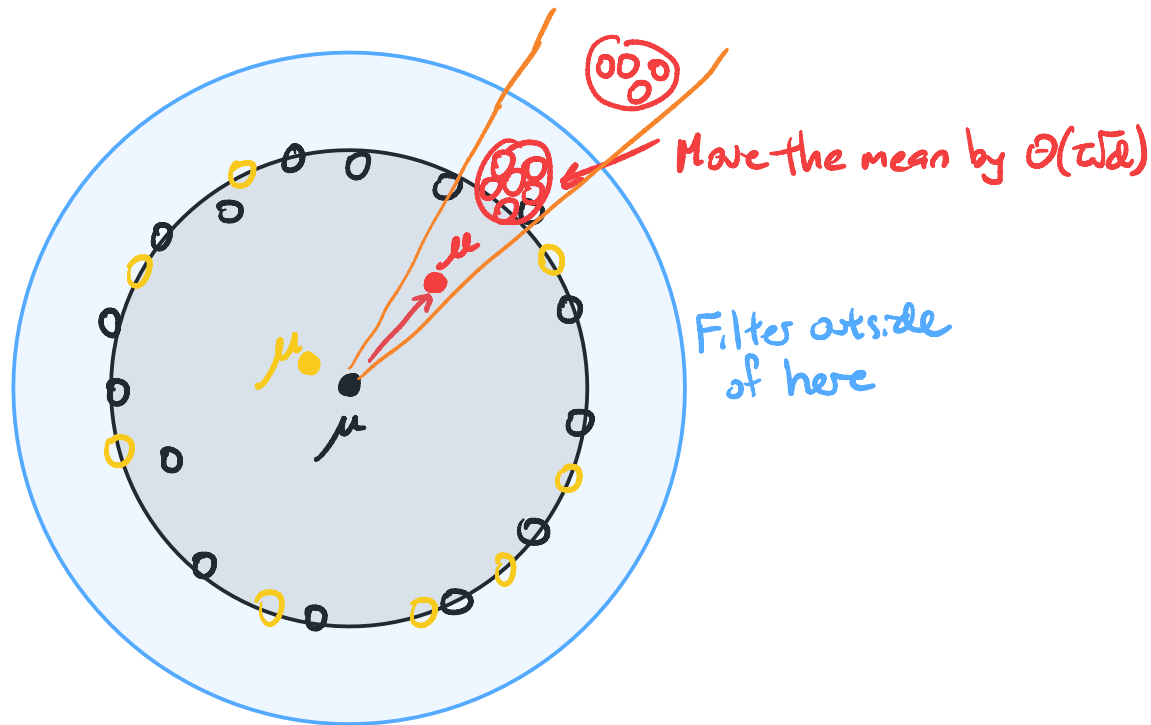
$$\|v\|_{\Sigma} = \|\Sigma^{-1/2}v\|_2$$

"VC Theorem": If we draw n pts from any dist $\mathcal{D}$ w.p. $1 - \delta$

then $\max_{\substack{v \in \mathbb{R}^d \\ z \in \mathbb{R}}} \left| \frac{1}{n} \sum \mathbb{1}\{\langle v, x - z \rangle \geq 0\} - \mathbb{P}[\langle v, x - z \rangle \geq 0] \right| = O\left(\frac{d + \log(1/\delta)}{n}\right)$

Efficient Gaussian Mean Estimation

$$X = (x_1, \dots, x_n) \sim \mathcal{N}(\mu, \Sigma)$$



Detecting Bad Corruptions

$$\mathcal{N}(0, \Sigma) \quad d=100$$

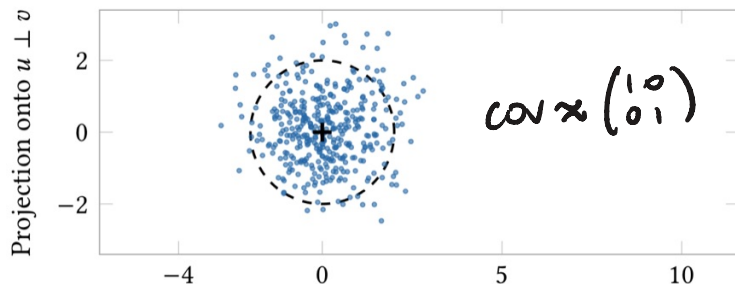
$$\hat{\Sigma} \approx \Sigma$$

$$\frac{v^T \hat{\Sigma} v}{\|v\|^2} \approx 1$$

Clean Gaussian data

$$\lambda_v = 1, \quad \lambda_u = 1$$

$$\mathcal{N}(0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix})$$



Projection onto $v = a/\|a\|_2$

$\tau = .1$ and corrupted pts are $(1, 1, \dots, 1)$

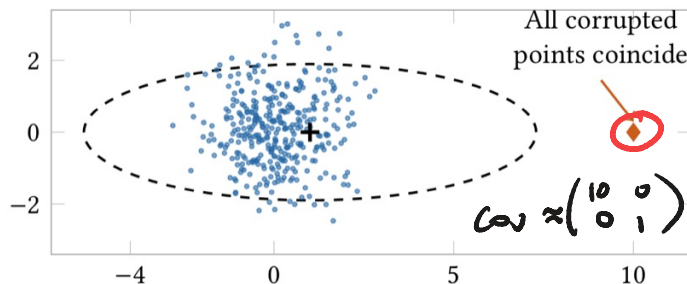
$$\frac{100 \text{ d.m.}}{\| \cdot \|_2 = 10} = \vec{a}$$

$\tilde{\Sigma}$ is sample cov

$$\frac{a^T \tilde{\Sigma} a}{\|a\|^2} \text{ is too big } (\approx 100)$$

10% corrupted at a

$$\lambda_v = 9.9, \quad \lambda_u = 0.9$$



Projection onto $v = a/\|a\|_2$

• Clean points ♦ Corrupted points + Mean - - - Covariance ellipse

Detecting Bad Corruptions

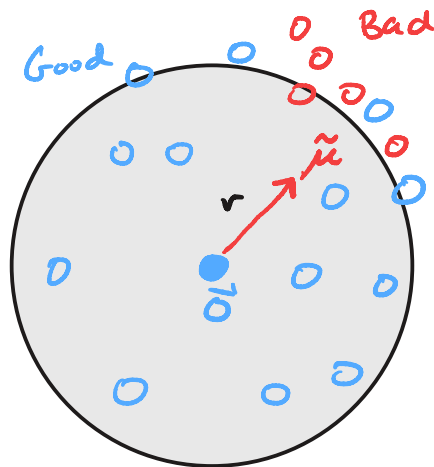
For the union of all the pts

$$\tilde{\mu} = \tau \mu_B$$

$$\tilde{\Sigma} = (1-\tau)\Pi + \tau\Sigma_B + \frac{1-\tau}{\tau} \tilde{\mu} \tilde{\mu}^T$$

$$\frac{\tilde{\mu}^T \tilde{\Sigma} \tilde{\mu}}{\|\tilde{\mu}\|^2} \geq 1-\tau + \frac{1-\tau}{\tau} \|\tilde{\mu}\|_2^2$$

Good data
has covariance $= \Pi$



Corruptions have
mean: μ_B
cov: Σ_B

Lemma: If the corruption moves the mean from $\hat{\mu}$ to $\tilde{\mu}$ then

$$\exists v, \|v\|_2=1 \text{ st } v^T \tilde{\Sigma} v \geq 1-\tau + \frac{1-\tau}{\tau} \|\tilde{\mu} - \hat{\mu}\|^2$$

Efficient Spectral Algorithm

Input: corrupted sample S

- 1: compute mean $\bar{\mu}_S$, covariance $\bar{\Sigma}_S$
- 2: compute $\lambda =$ top eigenvalue of $\bar{\Sigma}_S$
- 3: if $\lambda \approx 1$, return $\bar{\mu}_S$ and halt
- 4: let $v =$ top eigenvector of $\bar{\Sigma}_S$
- 5: filter out any pt x_i such that $\langle v, x_i - \bar{X}_S \rangle^2 \gg \text{threshold}$
- 6: repeat