

Foundations of Trustworthy AI

Jonathan Ullman
CS7180/7880 · Fall 2026

Class 9: Robust Mean Estimation

Oct 9, 2026

Roadmap

Robust Statistics

Distribution Shift

Targeted Attacks

Next 2-3 Classes

- Robust Univariate Mean
- Origins of Robustness
- Robust Multivariate Mean
- Robust Optimization

Robust
Mean Estimation

(Non-Robust) Mean Estimation

Distribution D on $\mathbb{R}$ Mean $\mu = \mathbb{E}_D[X]$ Loss $\ell(\hat{\mu}; D) = (\hat{\mu} - \mu)^2$

Sample $X = (X_1, \dots, X_n) \stackrel{\text{iid}}{\sim} D$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

Sample Mean

Weak Condition: Bounded Variance

$$\sigma^2 = \mathbb{E}_D[(X - \mu)^2]$$

Strong Condition: D is Gaussian $N(\mu, \sigma^2)$

$$\mu = \mathbb{E}_D[X] \quad \sigma^2 = \mathbb{E}_D[(X - \mu)^2]$$

$\uparrow$
 σ is the "standard deviation"

Bounded Variance

$$\Rightarrow \mathbb{E}_D[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$$

Gaussian

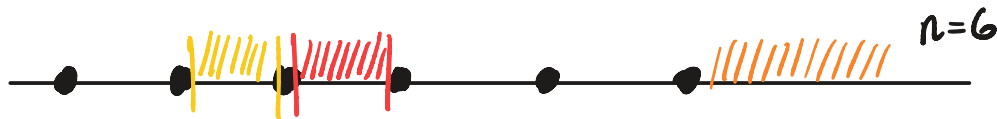
$$\Rightarrow \mathbb{P}_D[(\hat{\mu} - \mu)^2 \geq \frac{C \sigma^2 \log(1/\delta)}{n}] \leq \delta$$

Robust Mean Estimation

Classic robust estimator for $\mathbb{R}$ is the median

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

$$\text{med}(X) = \begin{cases} X_{(n/2)} & \text{if } n \text{ is even} \\ X_{((n+1)/2)} & \text{if } n \text{ is odd} \end{cases}$$



$$\text{med}(X) = \text{a point } z \text{ such that } \begin{aligned} z &\geq X_{(n/2)} \\ z &\leq X_{(n/2+1)} \end{aligned}$$

Robustness of Median

Robustness: Adversary

maps $X \mapsto \tilde{X}$ by
changing at most τn points

$$\text{med}(\tilde{X}) \approx \mathbb{E}_{\mathcal{D}}[X]$$

Classic robust estimator for IR is the median

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

$$\text{med}(X) = \begin{cases} X_{(n/2)} & \text{if } n \text{ is even} \\ X_{(n+1)/2} & \text{if } n \text{ is odd} \end{cases}$$

Median is not in general robust

$$\text{med}(0, 0, B) = 0$$

$$\text{med}(0, B, B) = B$$

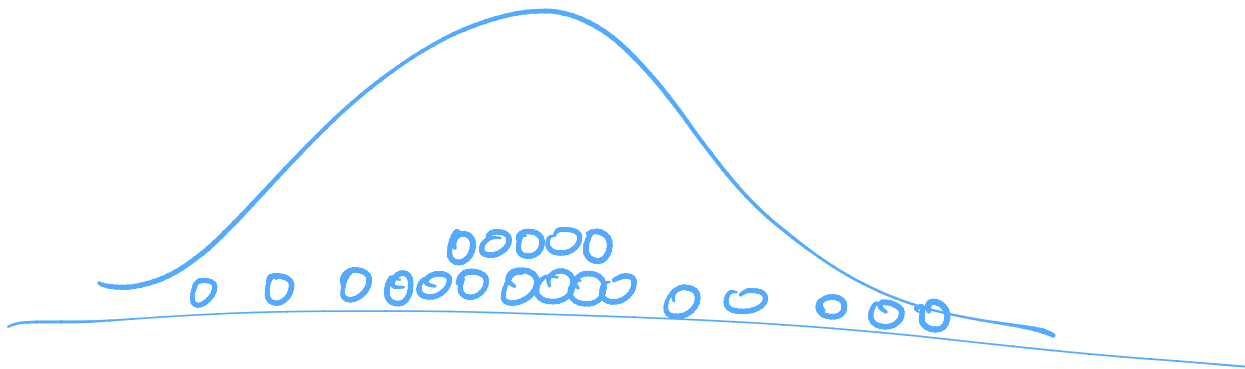
Fact: "Robustness is impossible for arbitrary distributions"

$$\mathcal{D} = \delta_0$$

Always
takes value 0

$$\tilde{\mathcal{D}} = (1-\tau)\delta_0 + \tau\delta_B$$

$$\left| \mathbb{E}_{\mathcal{D}}[X] - \mathbb{E}_{\tilde{\mathcal{D}}}[X] \right| = \tau B$$

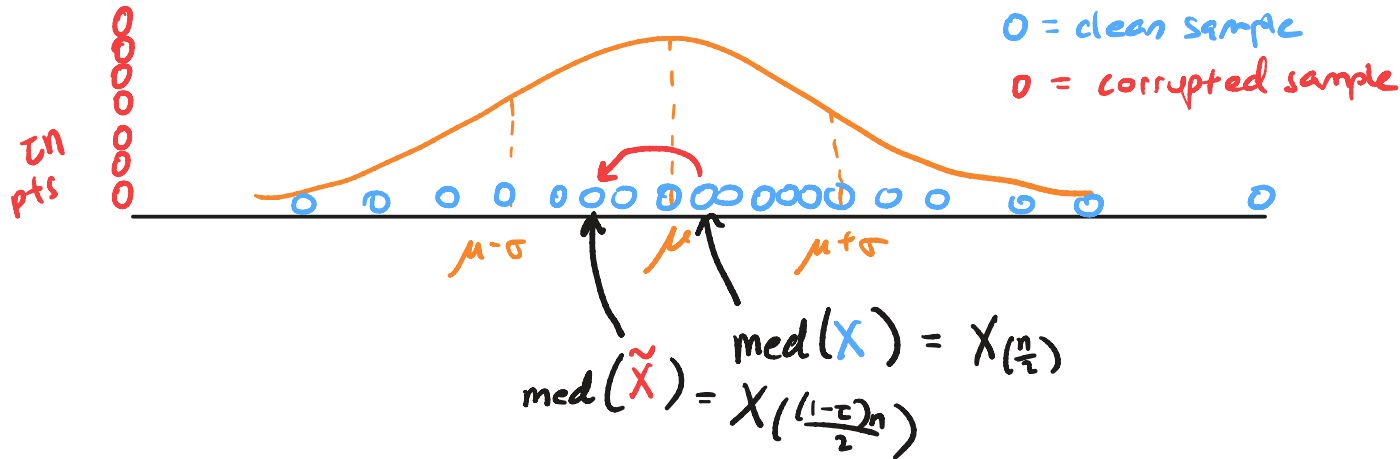


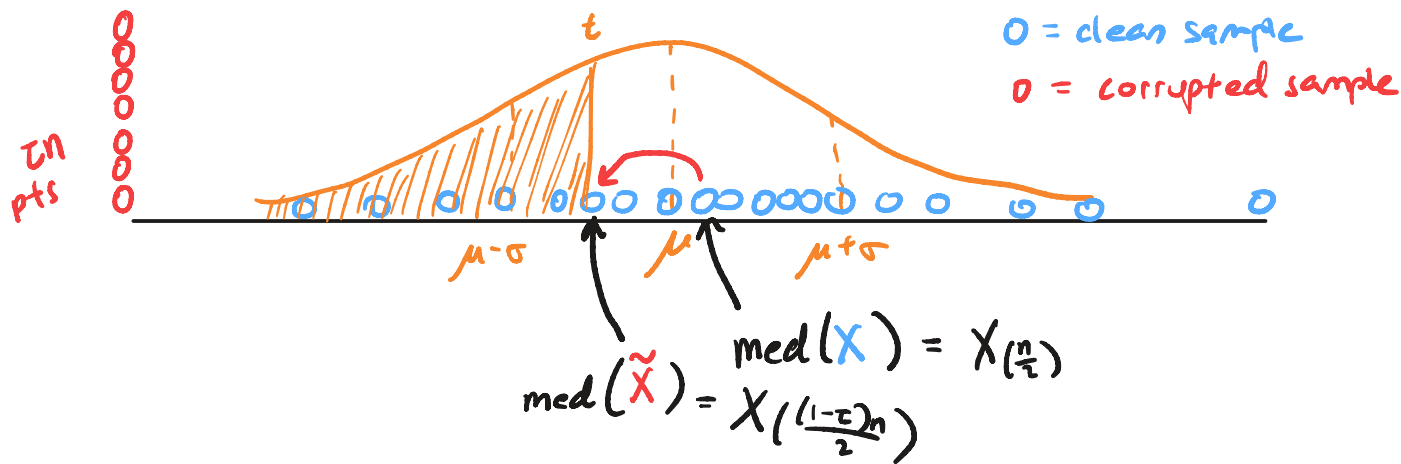
000

Robustness of Medians (Gaussian)

Theorem: If $D = \overset{\text{Gaussian}}{\mathcal{N}(\mu, \sigma^2)}$ and the adversary has corruption rate τ then with high probability $1 - \delta$

$$(\text{med}(\tilde{X}) - \mu)^2 = O\left(\sigma^2 \tau^2 + \frac{\sigma^2 \log(1/\delta)}{n}\right)$$

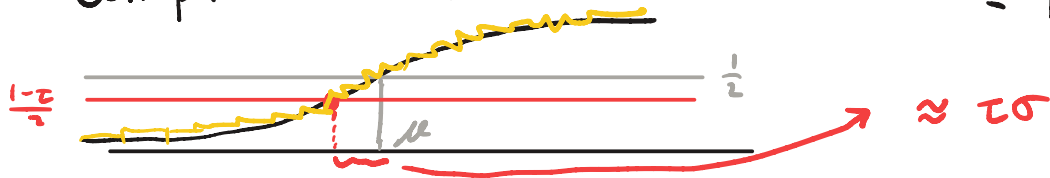




For intuition, look at the quantiles of D instead of X

CDF $\Phi_D(t) = P[X \leq t]$ Inverse CDF $\Phi_D^{-1}(q) = t \text{ s.t. } P[X \leq t] = q$

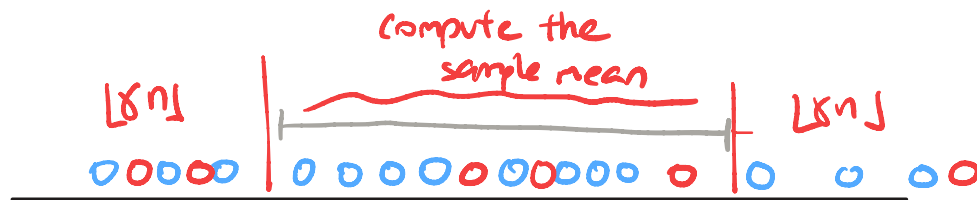
Corruption of the median introduces error $\leq \left| \Phi^{-1}\left(\frac{1}{2}\right) - \Phi^{-1}\left(\frac{1-\epsilon}{2}\right) \right|$



Trimmed Means

Trimmed means:

- ① Make sense for asymmetric data
- ② Nice "template" of removing potential outliers



$$\delta\text{-trimmed mean } \mu_{\text{trim}}(X) = \frac{1}{n - 2LxnJ} \sum_{i=LxnJ+1}^{n-LxnJ} X_{(i)}$$

For Gaussians the $\alpha\tau$ -trimmed mean give

$$(\hat{\mu} - \mu)^2 = \sigma^2 \tau^2 + \frac{\sigma^2}{n}$$

For Bounded variance the $\alpha\tau$ -trimmed mean gives

$$(\hat{\mu} - \mu)^2 = \sigma^2 \tau^2 + \frac{\sigma^2}{n}$$