

Foundations of Trustworthy AI

Jonathan Ullman
CS7180/7880 · Fall 2026

Class 10: Robustness in High Dimension

Oct 13, 2026

Roadmap

Send project
brainstorms
tomorrow night

Robustness Overview

Robust Mean Estimation in $\mathbb{R}$

- Means, trimmed means, medians

Robust Mean Estimation in $\mathbb{R}^d$

- Resilience and Tukey medians
- Efficient algorithms via filtering

Robust Optimization

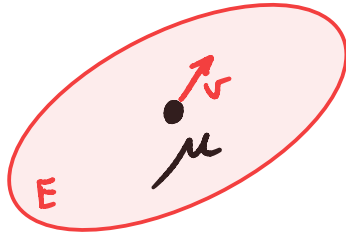
Distribution Shift

- Connection to calibration

Gaussians in $\mathbb{R}^d$

- Defined by a mean $\mu \in \mathbb{R}^d$, covariance $\Sigma \in \mathbb{R}_{\text{sym}}^{d \times d}$ $\Sigma \succcurlyeq 0$

$\downarrow$
 $\mathcal{N}(\mu, \Sigma)$



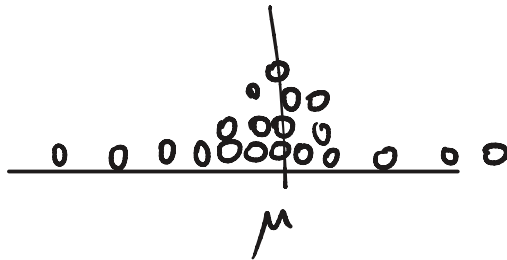
In any direction v , $\|v\|_2 = 1$
the variance is $v^T \Sigma v$

More specifically $X \sim \mathcal{N}(\mu, \Sigma)$

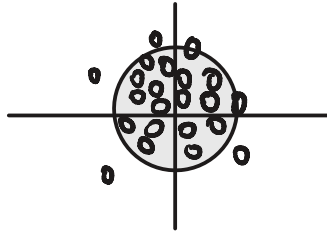
$$E = \{x : \|\Sigma^{-1/2}(x - \mu)\|_2 = 1\}$$

then $\langle v, X \rangle \sim \mathcal{N}(\langle v, \mu \rangle, v^T \Sigma v)$

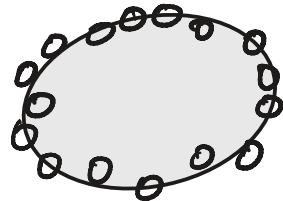
Univariate



Bivariate



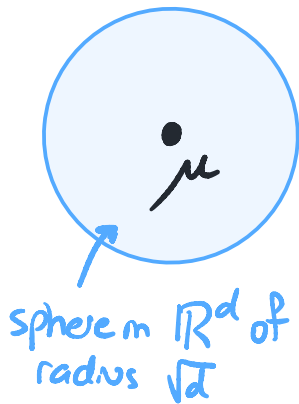
Multivariate $d \rightarrow \infty$



Gaussians in $\mathbb{R}^d$ (Isotropic, "Identity Covariance")

- Defined by a mean $\mu \in \mathbb{R}^d$, covariance $\mathbb{I}_d = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

$N(\mu, \mathbb{I})$

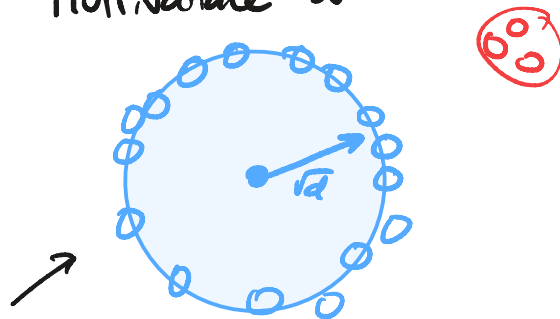


In any direction v
 $\langle X, v \rangle \sim N(\langle \mu, v \rangle, 1)$

$X \sim N(\mu, \mathbb{I}_d) \Leftrightarrow X_1, \dots, X_d$ independent
 $X_i \sim N(\mu_i, 1)$

$$\mathbb{E}[\|X - \mu\|_2^2] = \sum_{i=1}^d \mathbb{E}[(X_i - \mu_i)^2] = d \times 1$$

Multivariate $d \rightarrow \infty$



Gaussians in $\mathbb{R}^d$

- Defined by a mean $\mu \in \mathbb{R}^d$, covariance $\mathbb{I}_d = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

In any direction v
 $\langle X, v \rangle \sim N(\langle \mu, v \rangle, 1)$

$X \sim N(\mu, \mathbb{I}_d) \Leftrightarrow X_1, \dots, X_d$ independent
 $X_i \sim N(\mu_i, 1)$

- Sample $X = (X_1, \dots, X_n) \stackrel{\text{iid}}{\sim} N(\mu, \mathbb{I})$

- Sample mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \frac{1}{n} \mathbb{I}_d)$

- $\mathbb{E}[\|\hat{\mu} - \mu\|_2^2] = \frac{d}{n}$

- $\mathbb{P}\left[\|\hat{\mu} - \mu\|_2^2 \geq \frac{C(d + \log(1/\delta))}{n}\right] \leq \delta$

Robust Gaussian Mean in $\mathbb{R}^d$

Simple Algorithms:

- Coordinatewise mean estimation
- Trim any point further than $\epsilon\sqrt{d}$ away

Adversary: clean data is $\mathcal{N}(0, \mathbb{I}_d)$
 corrupted points are set to $(1, 1, \dots, 1)$

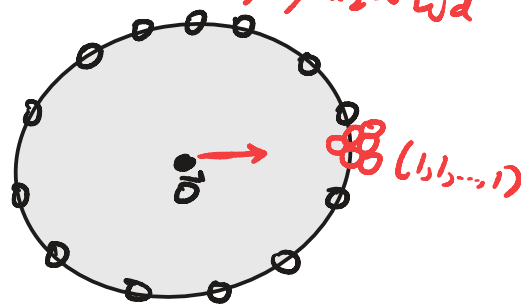
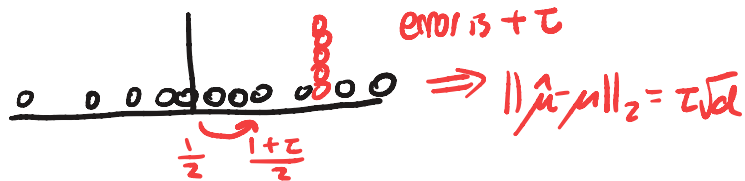
↓ coordinate

↓ trimming

$$\hat{\mu} \approx \frac{\tau n \cdot (1, 1, \dots, 1)}{n}$$

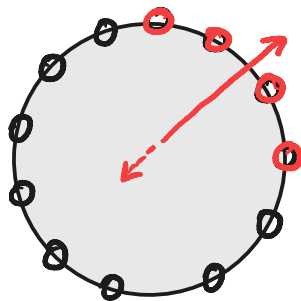
$$\|\hat{\mu} - \mu\|_2 \approx \tau\sqrt{d}$$

In each dimension $\mathcal{N}(0, 1)$



Resilience

Start by thinking about a deletion-only adversary



Any subset of $(1-\epsilon)n$ points
has a mean close to μ

"Resilience"

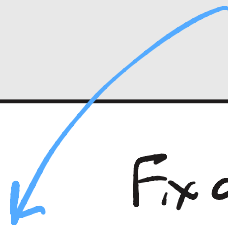
Definition (Resilience of a dataset): A dataset $X = (X_1, \dots, X_m)$ is (ρ, η) -resilient at μ if for every subset $T \subseteq [m]$ of size $\geq (1-\eta)m$

$$\bar{X}_T = \frac{1}{|T|} \sum_{i \in T} X_i \quad \|\bar{X}_T - \mu\| \leq \rho$$

Gaussians are Resilient

Lemma: If $X_1, \dots, X_m \stackrel{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ then with prob $\geq 1 - \delta$ the set $X_1, \dots, X_n$ is (ρ, η) -resilient around μ

$$\rho \approx \sqrt{\frac{d + \log(1/\delta)}{m}} + \underbrace{\eta \sqrt{\log(1/\eta)}}_{\text{independent of } d}$$



Immediately
say we can
estimate with
robustness to
deletions

Fix any $S \subseteq [m]$ $\frac{1}{|S|} \sum_{i \in S} X_i \sim \mathcal{N}(\mu, \frac{\Sigma}{|S|})$

Removing these pts shifts the mean by $\approx \eta \left(\frac{1}{|S|} \sum_{i \in S} X_i - \mu \right)$

There are $\approx m^{\eta n}$ possible choices

Resilience Implies Full Robustness

$A(\tilde{X}_1, \dots, \tilde{X}_n):$

If data is Gaussian, then clean data is such a set

- 1: Find a resilient set S' of size $\geq (1-\tau)n$ that is (ρ, η) -resilient around some $\hat{\mu}$
- 2: Output $\hat{\mu}$

set $\eta = \frac{\tau}{1-\tau}$

$\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_n)$

clean data is S , resilient set is S'

$T = S \cap S' \quad |T| \geq (1-2\tau)n$

resilience of S'

resilience of S

set $\eta = \frac{\tau}{1-\tau}$

$= (1-\eta)|S|$

$= (1-\eta)|S'|$

$\|\hat{\mu} - \mu\|_2 \leq \|\hat{\mu} - \bar{X}_T\|_2 + \|\bar{X}_T - \mu\|_2$

$\leq \rho + \rho = 2\rho$

