L5: Hardness of Estimation: Mean Estimation

Review Learning Theory

VC-Dimension: $\nu \approx$ number of parameters in model class F
"dimension for model F"

Labeled data (X, y) size n

Sample Error (training error)

- separable: $n \approx (\nu/\varepsilon)\log(\nu/\varepsilon)$
  $error(X) \leq \varepsilon \approx (\nu/n)\log(\nu/n)$
- non-separable: $n \approx \nu/\varepsilon^2$
  $error(X) \leq \varepsilon \approx \nu/\sqrt{n}$

increases with ν increasing

Model Error

- $\gamma_F(X) = \min_{f\in F} error(f, X, y)$

decreases with ν increasing

Total Error (test error) = Sample Error + Model Error

Parameter Estimation

(Bayesian-y view)

Assume data $X \sim g(\alpha)$ for some model g with parameters $\alpha \in \mathbb{R}^d$.

Distributional so g provides probability distribution
e.g., each $x \in X$ from "perfect" $h(\alpha) + Noise$, where $Noise$ is random, independent of h.

The simplest case is

- $h(\alpha) = \alpha$
- $Noise = \mathcal{G}_d(0, I)$ (d-dimensional Gaussian/Normal noise)
- Goal: from $X \sim g$, recover α

Alternatively, if

- $h(\alpha) = 0$
- $Noise = \mathcal{G}_d(\alpha, I)$
- Goal: from $X \sim g$, recover α
  the same problem, but now clear we are aiming to recover the **mean**
  which is $E_{X\sim Noise}[X] = \alpha$

Chebyshev Inequality (Law of Large Numbers)
For n iid RVs $X_1, X_2, \dots, X_n$ with $Var[X_i] = \sigma^2$

$$Pr[|\bar{x} - E[X_j]| \geq \eta] \leq \frac{\sigma^2}{n\eta^2}$$

Note that simple Chebyshev we have
$Pr[|X_j - E[X_j]| \geq \eta] \leq \frac{\sigma^2}{\eta^2}$,
but for $Var[\bar{x}] = Var[X_j]/n = \sigma^2/n$ so
$Pr[|X - E[X]| \geq \eta] \leq \frac{Var[\bar{x}]}{\eta^2} = \frac{\sigma^2}{n\eta^2}$

Chernoff-Hoeffding Inequality (simplified as in Azuma)
For n iid RVs $X_1, X_2, \dots X_n$ with $X_i \in [0, \Delta]$

$$Pr[|X - E[X]| \geq \eta] \leq 2\exp(-\frac{2\eta^2 n}{\Delta^2})$$

Thus as n increases our bound on the error η from an expected value decreases with $1/\sqrt{n}$.
Fix either $Pr[...] = \delta$, and solve for η as a function of n.

Trouble with High-Dimensional Mean Estimation

For each $x \in X \sim G_d(\alpha, I)$, then

$$E[\|x - \alpha\|^2] = \sum_{j=1}^{d} E[(x_j - \alpha_j)^2] = \sum_{j=1}^{d} E[(x_j - E[x_j])^2] = \sum_{j=1}^{d} Var[x_j] = d$$

Then for $\bar{x} = \frac{1}{n}\sum_{j=1}^{d} x_i$

$$E[\|\bar{x} - \alpha\|^2] = \sum_{j=1}^{d} Var[\bar{x}_j] = d/n$$

We can also analyze the convergence

$$Pr[\|\bar{x} - \alpha\| > \eta] \leq d/(n\eta^2)$$

To show this we will use the **Union Bound** that if there are k events $E_1, \dots, E_k$, then the probability all events are true $Pr[E_1 \& ... \& E_k] \leq 1 - \sum_{j=1}^{k} Pr[E_j = FALSE]$.

$Pr[(\bar{x}_j - \alpha_j)^2 > (\eta')^2] \leq \frac{Var[X_j]}{n(\eta')^2} = \frac{1}{n(\eta')^2} = \delta'$
So applying the union bound on d coordinates, with $\delta = \delta' \cdot d$,
Setting $\eta = \eta' \cdot \sqrt{d}$ so $\eta^2 = (\eta')^2 d$, we have
$Pr[\|\bar{x} - \alpha\|^2 > \eta^2] \leq \delta$
Solving for $\eta = \eta' \cdot \sqrt{d}$ and $\eta' = \frac{1}{\sqrt{n\delta}}$, so $\eta = \frac{\sqrt{d}}{\sqrt{n\delta}}$.
Or $n = d/(\eta^2 \delta)$

Two Mean Example:
Consider two mean estimations in $\mathbb{R}^d$
$X_1 \sim \mathcal{G}_d(\alpha, I)$ and $X_2 \sim \mathcal{G}_d(\alpha', I)$
where we are promised that $|\alpha_1 - \alpha'_1| = 2$ and $\alpha_j = \alpha'_j$ for $j > 1$.

As n increases, we can get estimates of α_1 and α'_1 to concentrate to values η much less than 2.
But each $x \in X_1$ has $E[\|x - \alpha\|^2] = d$. and $E[\|\bar{x} - \alpha\|^2] = d/n$
Moreover $Pr[\|\bar{x} - \alpha\| \geq \sqrt{d/n\delta}] \leq \delta$.

Let $\bar{x}_1 = \frac{1}{|X_1|}\sum_{x \in X_1} x$ and similar for $\bar{x}_2$.
$E[\|\bar{x}_1 - \alpha\|^2] = d/|X_1|$.
To get $\|\bar{x}_1 - \alpha\| \leq \eta$, we need approximately d/η^2 samples.

Outliers:

One outlier can significantly affect sample mean $\bar{x}$
In $d = 1$, the median is a good estimate for α and resistant to outliers.
What is analog in high dimensions?
- coordinate-wise median: $v = (v_1, ..., v_d)$ has v_j as median of j th coordinates.
- L1 median (geometric median): v minimize sum of distances to $x \in X$ - centerpoint: v so no halfspace containing v contains more than $|X|/(d+1)$ points - Turkey median: $v = \arg\max_{v \in \mathbb{R}^d} \min_{h \in H, v \in h} \frac{|X \cap h|}{|X|}$

Turkey median works well (uses d/η^2 samples, even with outliers), but best algorithms take about $|X|^{d-1}$ time to compute.

Robust High-Dimensional Mean Estimation

With $n = d/\eta^2$ samples, one can use new approaches - that allow for η-fraction of outliers, rest from $\mathcal{G}_d(0, I)$ - in $\mathsf{poly}(nd/\eta)$ time - find a point $\hat{v} \in \mathbb{R}^d$ so $\|v - \hat{v}\| \leq \tilde{O}(\eta)$

Careful Pruning

Start with Sample Mean $\bar{x}$, and center data.
Compute top principal vector u, project data along $u : X_u = \{x_u = \langle x, u \rangle \mid x \in X\}$.
Compare CDF to that of 1-d normal. Prune extreme points that are too far off. [*]
Repeat until no sign of outliers.

[*] Vershynin (2011): $\Sigma \in \mathbb{R}^{n \times d}$ so each entry iid $\Sigma_{i,j} \sim \mathcal{N}(0, 1)$.
With probability at least $1 - 2\exp(-t^2/2)$

$$\sqrt{n} - \sqrt{d} - t \leq s_{\min}(\Sigma) \leq \|\Sigma\|_2 \leq \sqrt{n} + \sqrt{d} + t$$

Median of Means

Decompose X into k components randomly (e.g., for $k = 3, 5$, or 7)
Compute mean of each $\bar{x}_1, \dots, \bar{x}_k$.
Return coordinate-wise median of set $\{\bar{x}_1, \dots, \bar{x}_k\}$.

Works quite well if $|X|$ is large enough to split.