36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 2: January 16

Lecturer: Siva Balakrishnan *Scribe: Neil Xu*

2.1 Gaussian complexity

2.1.1 Motivating example

Let's try to derive an estimator for a parameter $\theta^* \in K \subset \mathbb{R}^d$. We have an observation of this θ^*, y, that is generated through the following process:

$$\epsilon \sim \mathcal{N}\left(0, \frac{\sigma^2 I_d}{n}\right)$$
$$y = \theta^* + \epsilon.$$

Note that semantically, we can view y as an already computed mean of several draws from a standard normal centered at θ^*. One estimator for θ^* we can consider is the MLE:

$$\begin{aligned}
\widehat{\theta} &= \operatorname*{argmax}_{\theta \in K} \mathcal{L}(y \mid \theta) \\
&= \operatorname*{argmax}_{\theta \in K} \sqrt{\frac{1}{(2\pi)^d \det(\sigma^2 I_d/n)}} \exp\left(-\frac{1}{2}\|y-\theta\|_2^2\right) \\
&= \operatorname*{argmin}_{\theta \in K} \|y-\theta\|_2^2.
\end{aligned}$$

We would like to consequently measure the distance between $\widehat{\theta}$ and θ^*. Define the vector between the two points as $\Delta = \widehat{\theta} - \theta^*$.

One property that follows from the definition of $\widehat{\theta}$, which we term the **basic inequality**, is:

$$\|y - \widehat{\theta}\|_2^2 \leq \|y - \theta^*\|_2^2$$

We use this basic inequality to get an upper bound on $\|\Delta\|_2^2$:

$$\begin{aligned}
\|\theta^* + \epsilon - \widehat{\theta}\|_2^2 &\leq \|\epsilon\|_2^2 \\
\|\Delta + \epsilon\|_2^2 &\leq \|\epsilon\|_2^2 \\
\|\Delta\|_2^2 + 2\langle \Delta, \epsilon \rangle + \|\epsilon\|_2^2 &\leq \|\epsilon\|_2^2 \\
\|\Delta\|_2^2 &\leq -2\langle \Delta, \epsilon \rangle
\end{aligned}$$

We could try to bound $-2\langle \Delta, \epsilon \rangle$ in a couple of ways:

- Cauchy-Schwarz:

$$\begin{aligned} \|\Delta\|_2^2 &\leq 2\|\epsilon\|_2\|\Delta\|_2 \\ \|\Delta\| &\leq 2\|\epsilon\| \\ \mathbb{E}[\|\Delta\|] &\leq 2\sigma\sqrt{\frac{d}{n}} \end{aligned}$$

  This is a result of the following inequality

$$\mathbb{E}[\|\epsilon\|] \leq \sqrt{\mathbb{E}\left[\|\epsilon\|_2^2\right]} = \sqrt{\sum_{i=1}^{d} \mathbb{E}[\epsilon_i^2]} = \sqrt{\frac{\sigma^2 d}{n}}.$$

- Other Hölder inequalities: an example of another bound that could be useful is if we knew a fact like $K = \{\, \theta |\ \|\theta\|_1 \leq t \,\}$. Then we could get the following bound:

$$\begin{aligned} \|\Delta\|_2^2 &\leq 2\|\epsilon\|_\infty\|\Delta\|_1 \\ &\leq 2t\|\epsilon\|_\infty \end{aligned}$$

  We could then use some kind of Gaussian tail bound on $\|\epsilon\|_\infty$. In particular, we can use the fact that

$$\mathbb{E}\|\epsilon\|_\infty \leq \sigma\sqrt{\frac{2\log(d)}{n}}.$$

2.1.2 Formulation

In the previous bounds, we either used no knowledge of K (in the Cauchy-Schwarz case) or imposed too much structure on K (having it being a specific kind of set in the Hölder case). We use K in a more general way in the following upper bound:

$$-2\langle \epsilon, \Delta \rangle \leq 2 \sup_{\widetilde{\theta} \in K - \theta^*} \left\langle -\epsilon, \widetilde{\theta} \right\rangle$$

Definition 2.1 *We define the* ***Gaussian complexity*** *of a set K as*

$$\mathcal{G}(K) = \mathbb{E}\left[\sup_{\theta \in K} \langle \epsilon, \theta \rangle\right]$$

where $\epsilon \sim \mathcal{N}(0, I_d)$.

We can view this as a measurement of the size of K in sense, and is similar to Rademacher complexity (where ϵ would be a vector of independent Rademacher variables) as a tool for characterizing the deviation of an empirical estimate from the true parameter.

Example 2.2 (Finite set) *When there are a finite set of vectors, $K = \{\theta_1, \ldots, \theta_N\}$ we can expect:*

$$\mathcal{G}(K) = \mathbb{E}\left[\max_{\theta \in K} \langle \epsilon, \theta \rangle\right] \asymp \sqrt{\log(N)}$$

Example 2.3 (L1 and L2 balls) *Let $K = \{\, \theta |\ \|\theta\|_2 \leq 1 \,\}$ be a the unit sphere in L2 distance. The Gaussian complexity is then:*

$$\mathcal{G}(K) = \mathbb{E}\left[\max_{\|\theta\|_2 \leq 1} \langle \epsilon, \theta \rangle\right] = \mathbb{E}\left[\|\epsilon\|_2\right] \leq \sqrt{d}.$$

We can view the L2 unit sphere case as when the Cauchy-Schwarz bound becomes equality. On the other hand, for a L1 unit ball we have the following Gaussian complexity (which we will prove in a later lecture):

$$\mathcal{G}(K) \approx \sqrt{\log(d)}$$

Note that Gaussian complexity is a lossy bound on $-2\langle \epsilon, \Delta \rangle$, since it considers the worst case (furthest) distance between elements of K. In future lectures, we will also consider a notion of localized Gaussian complexity that will utilize an additional inductive bias $\widehat{\theta}$ is close to θ^*.

2.2 Covering and Packing Numbers

Gaussian complexity maybe difficult to directly compute, so we consider a different notion of the size of a set that we can use to then bound Gaussian complexity.

Let $\mathbb{T}, \rho$ be a metric space i.e. ρ be a metric on $\mathbb{T}$.

Definition 2.4 *$\rho : \mathbb{T} \times \mathbb{T} \to \mathbb{R}$ is a **metric** on $\mathbb{T}$ iff if it satisfies the following properties on arbitrary $\theta_i, \theta_j, \theta_k \in \mathbb{T}$:*

1. *$\rho(\theta_i, \theta_j) \geq 0$*
2. *$\rho(\theta_i, \theta_j) = 0$ iff $i = j$*

3. $\rho(\theta_i, \theta_j) = \rho(\theta_i, \theta_j)$
4. $\rho(\theta_i, \theta_k) \leq \rho(\theta_i, \theta_j) + \rho(\theta_j, \theta_k)$

Example 2.5 (Metric spaces)

- $\mathbb{T} = \mathbb{R}^d, \rho(\theta_i, \theta_j) = \|\theta_i - \theta_j\|_p$ *for some* $p \geq 0$.
- $\mathbb{T} = \mathcal{F}[0,1]$ *where* $\|f\|_\infty = \sup_{x \in [0,1]} f(x) \leq 1$, $\rho(f_i, f_j) = \sup_{x \in [0,1]} \|f_i(x) - f_j(x)\|_p$ *for some* $p \geq 0$.

Definition 2.6 *A* ***covering*** *of set* $\mathbb{T}$ *under metric* ρ *with balls of radius* $\delta > 0$ *is a set* $\{\theta^1 \ldots, \theta^N\}$ *such that for all* $\theta \in \mathbb{T}$*, there exists* $i \in [N]$ *such that* $\rho(\theta^i, \theta) \leq \delta$.

Definition 2.7 *The* ***covering number*** *of set* $\mathbb{T}$ *with respect to metric* ρ *for balls of radius* $\delta > 0$ *is denoted as* $N(\delta; \mathbb{T}, \rho)$ *and defined as the cardinality of a minimum cover of* $\mathbb{T}$ *with radius* δ *and metric* ρ. ***Metric entropy*** *is simply defined as* $\log N(\delta; \mathbb{T}, \rho)$.

Definition 2.8 *A metric space* $\mathbb{T}, \rho$ *is considered* ***totally bounded*** *iff for all* $\delta > 0$, $N(\delta; \mathbb{T}, \rho)$ *is finite.*

Definition 2.9 *A* ***packing*** *of set* $\mathbb{T}$ *under metric* ρ *with balls of radius* $\delta > 0$ *is a set* $\{\theta^1 \ldots, \theta^M\} \subseteq \mathbb{T}$ *such that for all* $i, j \in [M]$ *and* $i \neq j$, $\rho(\theta^i, \theta^j) > \delta$.

Definition 2.10 *The* ***packing number*** *of set* $\mathbb{T}$ *with respect to metric* ρ *for balls of radius* $\delta > 0$ *is denoted as* $M(\delta; \mathbb{T}, \rho)$ *and defined as the cardinality of a maximum packing of* T *with radius* δ *and metric* ρ.

Proposition 2.11 (Relationship between packing and covering)

$$M(2\delta; \mathbb{T}, \rho) \leq N(\delta; T, \rho) \leq M(\delta; T, \rho).$$

Proof: For the upper bound, consider a maximum δ-packing. That means for all θ in $\mathbb{T}$, there exists some θ^i in the packing such that $\rho(\theta, \theta^i) \leq \epsilon$, else we would be able to add θ to our packing, and violate the definition of maximum. Consequently, this maximum packing is also covering. Thus, the cardinality of the minimum covering can only be smaller, and consequently the upper bound is shown.

For the lower bound, we assume for sake of contradiction that $M(2\delta; \mathbb{T}, \rho) \geq N(\delta; \mathbb{T}, \rho) + 1$. Consider an arbitrary maximum 2δ-packing and minimum δ-covering. By pigenohole principle, there is an element of the minimum covering, θ^k, that is the closest element in the covering to at least two elements of the packing, θ^i, θ^j. By definition of covering, $\rho(\theta^i, \theta^k) \leq \delta, \rho(\theta^j, \theta^k) \leq \delta$. By triangle inequality, this means that $\rho(\theta^i, \theta^j) \leq 2\delta$ which contradicts the definition of 2δ-packing. ∎

Here, packing numbers are useful because they provides us with both a lower and upper bound on the covering number.

Example 2.12 (Unit intervals) *Let $\mathbb{T} = [-1, 1], \rho(\theta^i, \theta^j) = |\theta^i - \theta^j|$ be a metric space. We can compute a δ-cover simply by taking numbers every 2δ, starting at -1 i.e. $\{-1, -1 + 2\delta, \ldots, -1 + 2\delta N\}$ where N is the smallest integer such that $1 \leq -1 + 2\delta N + \delta$. This gives us an upper bound of $\lceil 1/\delta \rceil$ on $N(\delta; \mathbb{T}, \rho)$. The cover we have constructed is also a 2δ-packing. By proposition 2.11, this packing is also a lower bound on $N(\delta; \mathbb{T}, \rho)$ which gives a equality of $N(\delta; \mathbb{T}, \rho) = \lceil 1/\delta \rceil$.*

We can generalize this to d dimensions i.e. $\mathbb{T} = [-1, 1]^d, \rho(\theta^i, \theta^j) = \|\theta^i - \theta^j\|_\infty$ and $N(\delta; \mathbb{T}, \rho) = (1/\delta)^d$

Example 2.13 (Lipschitz functions on the unit interval) *Let*

$$\mathbb{T} = \mathcal{F}^L[0, 1] = \{ f | \ f : [0, 1] \to \mathbb{R}, f(0) = 0, |f(x) - f(y)| \leq L |x - y| \}$$
$$\rho(f^i, f^j) = \|f^i - f^j\|_\infty$$

Define:

$$\beta \in \{-1, +1\}^M$$
$$x_i = (i - 1)h$$
$$\phi(z) = \mathbf{1}\{z \geq 0\}$$

where h is a bandwidth we choose. To achieve a lower bound, we construct a packing with the family of functions:

$$f_\beta(x) = \sum_{i=1}^{M} \beta_i L h \phi\left(\frac{y - x_i}{h}\right)$$

For each pair of functions f_{β^i}, f_{β^j}, there exists index k such that $\beta_k^i \neq \beta_k^j$ which means that $\|f_{\beta^i} - f_{\beta^j}\|_\infty \geq 2Lh$. If $2Lh = 2\delta$, this is a valid 2δ-packing. Consequently, $h = \delta/L$ which means there are at least $2^{L/\delta}$ elements of the packing. This gives an asymptotic lower bound of $\log N(\delta; \mathbb{T}, \rho) \gtrsim L/\delta$.

We can also prove upper bound by showing this packing is also a cover - a proof sketch of this is to divide the up the graph into grid squares of $\delta \times \delta$ size, and show that for the grid squares $f \in \mathcal{F}^L[0,1]$ passes through, there exists a f_β that passes through the same grid squares that is no more than δ away from f.

If we consider f over $[0,1]^d$ instead, an analogical argument shows that $\log N(\delta; \mathbb{T}, \rho) \asymp (L/\delta)^d$ which is exponential when compared to the metric entropy of unit intervals.

Metric entropy can be used in a heuristic calculation for rates of convergence for estimators in some set, as demonstrated in the Le Cam equation.

Definition 2.14 *The **Le Cam equation** is defined as:*

$$n\epsilon^2 \asymp \log N(\epsilon; \mathbb{T}, \rho)$$

where ϵ is the "error" of an "estimator", n is the number of samples used to construct the estimator, and $\mathbb{T}, \rho$ form a metric space. The equation is a heuristic for computing the convergence rate of some estimator for a parameter in $\mathbb{T}$.

Example 2.15 (Using Le Cam equation) *For a parametric estimator, we get a covering number of $d \log(1/\epsilon)$, which when plugged into the Le Cam equation, $n\epsilon^2 \asymp d \log(1/\epsilon)$, gets us an asymptotic error of $\epsilon \asymp \sqrt{(d/n)\log(1/\epsilon)}$.*

For a nonparametric estimator, we have a covering number of $(1/\epsilon)^{d/\alpha}$ where α is based on constraints of how smooth the distribution is. Consequently, this gets us $\epsilon \asymp n^{-\alpha/(2\alpha+d)}$ which is typically $n^{-1/3}$ or slower, making it slower than parametric rates of convergence.