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

Lecture 3: January 21

Lecturer: Siva Balakrishnan *Scribe: Ian Waudby-Smith*

Intuitively, we might expect that if a set is sufficiently nice its metric entropy might be related to its volume, relative to the volume of the balls we are using to cover the set. We will first formalize this in the case when the set is a unit norm ball in some norm.

3.1 Metric entropy and Gaussian width continued

Consider some unit ball, B measured with some norm, $\|\cdot\|$, i.e.

$$B = \{x : \|x\| \le 1\}.$$

We'll cover this with δ-balls, $\delta B'$ in some other norm, $\|\cdot\|'$.

Proposition 3.1

$$\frac{vol(B)}{\delta^d vol(B')} \le N(\delta; B, \|\cdot\|') \le \frac{vol(B + \frac{\delta}{2}B')}{(\frac{\delta}{2})^d vol(B')}$$

Proof:(lower bound)

$$\begin{aligned}
\mathrm{vol}(B) &\le \mathrm{vol}\left(\bigcup_{i=1}^{N(\delta,B,\|\cdot\|')} \delta B_i'\right)\\
&\le N(\delta, B, \|\cdot\|')\mathrm{vol}(\delta B')\\
&= N(\delta; B, \|\cdot\|')\delta^d \mathrm{vol}(B')
\end{aligned}$$

Therefore, $N(\delta; B, \|\cdot\|') \ge \frac{\mathrm{vol}(B)}{\delta^d \mathrm{vol}(B')}$ ■

Proof:(upper bound)

Take a maximal packing $\{\theta_1, \ldots, \theta_M\}$ of B.

Now, $\bigcup_{i=1}^M B(\delta/2; \theta_i) \subseteq B + \frac{\delta}{2}B'$ where + here is the Minkowski sum.

Therefore,

$$M \cdot \mathrm{vol}\left(\frac{\delta}{2}B'\right) \le \mathrm{vol}\left(B + \frac{\delta}{2}B'\right)$$

And so we have

$$N(\delta; B, \|\cdot\|') \leq M(\delta; B, \|\cdot\|') \leq \frac{\text{vol}\left(B + \frac{\delta}{2}B'\right)}{\left(\frac{\delta}{2}\right)^d \text{vol}(B')}$$

as required.

■

3.1.1 Functions

Now, let's switch gears to discuss function spaces.

Say that we're interested in functions of the form

$$f = \sum_{j=1}^{\infty} \theta_j \varphi_j : [0,1] \to \mathbb{R}$$

such that $\sum_{j=1}^{\infty} \theta_j^2 < \infty$ and the φ_j form an orthonormal basis for the inner product space, $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ where for any $\psi_i, \psi_j \in \mathcal{F}$,

$$\langle \psi_i, \psi_j \rangle = \int_0^1 \psi_i(x)\psi_j(x)dx$$

In other words, for any φ_j, φ_k in our basis,

$$\int_0^1 \varphi_j(x)\varphi_k(x)dx = \mathbb{I}(j = k)$$

We'd like to impose some kind of structure on these functions. One natural way to do so is to assume that "most" of the mass of the sequence is in the early coefficients. Concretely, we may assume that,

$$\sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1,$$

for some sequence of coeffcients $\mu_1 \geq \mu_2 \geq \ldots \geq 0$.

In particular, consider **Sobolev ellipsoids**, functions for which we have $\mu_j = j^{-2\alpha}$ for some $\alpha > 1/2$. When $\alpha = 1$, these functions should be thought of as a generalization of Lipschitz functions.

Example 3.2

$$y = \theta^* + \epsilon$$

where θ^* is an infinite-dimensional vector.

If we just use y as an estimator for θ^*, we'll get bad rates. Instead, let's use this Sobolev structure.

If we use $t \in \mathbb{N}$ as a truncation point beyond which we simply don't estimate $\theta_{t+1}, \theta_{t+2}, \ldots$, then our bias and variance will (heuristically) look like

$$\text{Bias}^2 : \sum_{j=t+1}^{\infty} \theta_j^2, \qquad \text{Var: } \frac{t}{n}.$$

We can now bound this bias using our Sobolev ellipsoid assumption, i.e.

$$\sum_{j=t+1}^{\infty} \theta_j^2 = \sum_{j=t+1}^{\infty} \frac{\mu_j \theta_j^2}{\mu_j} \leq \mu_{t+1} \sum_{j=t+1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq \mu_{t+1} \leq t^{-2\alpha}.$$

Choosing $t \asymp n^{1/(1+2\alpha)}$, we see that the MSE should scale as $n^{-2\alpha/(2\alpha+1)}$ – which is indeed the rate at which we can estimate Sobolev functions in 1D.

Let us return to the task of bounding the metric entropy of the Sobolev ellipsoid. Consider the space of coefficients:

$$\mathcal{E} = \left\{ (\theta_1, \theta_2, \ldots) : \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

and

$$\widetilde{\mathcal{E}} = \{\theta \in \mathcal{E} : \theta_j = 0 \text{ for all } j > t\}$$

We claim that if we choose t to be the smallest integer such that $\mu_t \leq \delta^2$, then a δ-cover of $\widetilde{\mathcal{E}}$ is a $\sqrt{2}\delta$-cover of $\mathcal{E}$.

Proof: Let $\{\theta^1, \ldots, \theta^N\}$ be a cover for $\widetilde{\mathcal{E}}$. For any $\theta \in \mathcal{E}$,

$$\begin{aligned}
\|\theta - \theta^i\|_2^2 &= \sum_{j=1}^{t}(\theta_j - \theta_j^i)^2 + \sum_{j=1}^{\infty}\theta_j^2 \\
&\leq \delta^2 + \mu_t \sum_{j=t+1}^{\infty}\frac{\theta_j^2}{\mu_j} \\
&\leq 2\delta^2.
\end{aligned}$$

■

Therefore, we need only consider the covering of $\widetilde{\mathcal{E}}$. In class we did a quick argument that leads to a bound that is sub-optimal by a log-factor. We note that $\widetilde{\mathcal{E}}$ is contained within the t-dimensional unit ball, so

$$N(\widetilde{\mathcal{E}}; \delta, \|\cdot\|_2) \leq N(B; \delta, \|\cdot\|_2) \leq \left(1 + \frac{2}{\delta}\right)^t,$$

so selecting

$$t := \lceil (1/\delta)^{1/\alpha} \rceil,$$

we have that the metric entropy is upper bounded as:

$$\log N(\widetilde{\mathcal{E}}; \delta, \|\cdot\|_2) \lesssim \left(\frac{1}{\delta}\right)^{1/\alpha} \log(1/\delta).$$

Here is a sharper argument from the Wainwright book. As a subset of $\mathbb{R}^t$, $\widetilde{\mathcal{E}}$ contains the ℓ_2 ball, $B_2(\delta)$ and thus

$$\mathrm{vol}(\widetilde{\mathcal{E}} + B_2(\delta/2)) \leq \mathrm{vol}(2\widetilde{\mathcal{E}}).$$

And thus by proposition 3.1,

$$N \leq \frac{\mathrm{vol}(\widetilde{\mathcal{E}} + B_2(\delta/2))}{\left(\frac{\delta}{2}\right)^t \mathrm{vol}(B_2(1))} \leq \left(\frac{4}{\delta}\right)^t \frac{\mathrm{vol}(\widetilde{\mathcal{E}})}{\mathrm{vol}(B_2(1))}$$

By volume of ellipsoids,

$$\frac{\mathrm{vol}(\widetilde{\mathcal{E}})}{\mathrm{vol}(B_2(1))} = \prod_{j=1}^{t}\sqrt{\mu_j}$$

Therefore, substituting $\mu_j = j^{-2\alpha}$, we get

$$\log N \leq t \log\left(\frac{4}{\delta}\right) + \frac{1}{2}\sum_{j=1}^{t} \log \mu_j = t \log\left(\frac{4}{\delta}\right) - \alpha \sum_{j=1}^{t} \log j$$

Using the inequality, $\sum_{j=1}^{t} \log j \geq t \log t - t$, and selecting,

$$t := \lceil (1/\delta)^{1/\alpha} \rceil,$$

we have

$$\log N \leq \left\{ \left(\frac{1}{\delta}\right)^{\frac{1}{\alpha}} + 1 \right\} (\log 4 + \alpha)$$

And so,

$$\log N \lesssim \left(\frac{1}{\delta}\right)^{\frac{1}{\alpha}}.$$

3.1.2 Gaussian and Rademacher complexity

Definition 3.3 (Gaussian complexity of a set) *The Gaussian complexity of a set $\mathcal{K}$ is defined as*

$$\mathcal{G}(\mathcal{K}) = \mathbb{E} \sup_{\theta \in \mathcal{K}} \langle \theta, w \rangle \qquad w \sim N(0, I_d)$$

Definition 3.4 (Rademacher complexity of a set) *The Rademacher complexity of a set $\mathcal{K}$ is defined as*

$$\mathcal{R}(\mathcal{K}) = \mathbb{E} \sup_{\theta \in \mathcal{K}} \langle \theta, \epsilon \rangle \qquad \forall j,\ \epsilon_j \sim \mathrm{Unif}\{-1, 1\}.$$

Example 3.5 (Gaussian complexity of a Euclidean ball B_2^d)

$$\begin{aligned}
\mathcal{G}(B_2(1)) &= \mathbb{E} \sup_{\|\theta\|_2 \le 1} \langle \theta, w \rangle \\
&= \mathbb{E}\|w\|_2 \\
&\le \sqrt{\mathbb{E}\|w\|_2^2} \quad \text{by Jensen's inequality} \\
&\le \sqrt{d}.
\end{aligned}$$

Example 3.6 (Gaussian complexity of B_1^d)

$$\begin{aligned}
\mathcal{G}(B_1(1)) &= \mathbb{E} \sup_{\|\theta\|_1 \le 1} \langle \theta, w \rangle \\
&= \mathbb{E}\|w\|_\infty \\
&= \mathbb{E} \max_{j=1,\ldots,d} |w_j| \\
&\asymp \sqrt{2 \log d}.
\end{aligned}$$

Key takeaway: an ℓ_2 ball has much larger Gaussian complexity than an ℓ_1 ball.

Example 3.7 (Gaussian complexity of truncated B_0^d) *Let*

$$\mathbb{S}^d(s) = \left\{ \theta \in \mathcal{K} : \sum_{j=1}^{d} \mathbb{I}(\theta_j \neq 0) \le s \right\} \bigcap B_2(1)$$

$$\begin{aligned}
\mathcal{G}\left(\mathbb{S}^d(s) \right) &= \mathbb{E} \sup_{\|\theta\|_0 \le s, \|\theta\|_2 \le 1} \langle \theta,\, w \rangle \\
&= \mathbb{E} \max_{|S| \le s} \|w_s\|_2 \\
&\lesssim \sqrt{s \log\left(\frac{ed}{s} \right)}.
\end{aligned}$$