ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 15: Sudakov, Maurey, and duality of metric entropy

Lecturer: Yihong Wu *Scribe: Aolin Xu, Mar 17, 2016 [Ed. Mar 24]*

In this lecture we study the upper and lower bounds on $N(B_1, \|\cdot\|_2, \epsilon)$.

From the last lecture, we know that for any $\Theta \in \mathbb{R}^d$ and any $\epsilon > 0$,

$$\frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(\epsilon B)} \leq M(\Theta, \|\cdot\|, 2\epsilon) \leq N(\Theta, \|\cdot\|, \epsilon) \leq M(\Theta, \|\cdot\|, \epsilon) \leq \frac{\mathrm{vol}(\Theta + \frac{\epsilon}{2}B)}{\mathrm{vol}(\frac{\epsilon}{2}B)}.$$

where B is the ball of radius 1 measured by $\|\cdot\|$. Therefore,

$$M(B_1, \|\cdot\|_2, \epsilon) \leq \frac{\mathrm{vol}(B_1 + \frac{\epsilon}{2}B_2)}{\mathrm{vol}(\frac{\epsilon}{2}B_2)} \leq \frac{\mathrm{vol}((1+\frac{\epsilon\sqrt{d}}{2})B_1)}{\mathrm{vol}(\frac{\epsilon}{2}B_2)} = \left(\frac{1+\frac{\epsilon\sqrt{d}}{2}}{\frac{\epsilon}{2}}\right)^d \left(\frac{c_1}{\sqrt{d}}\right)^d \leq \left(1 + \frac{c_2}{\epsilon\sqrt{d}}\right)^d,$$

where we have used the fact that $B_2 \subset \sqrt{d}B_1$ by Cauchy-Schwarz inequality, $\mathrm{vol}(B_1)^{1/d} \asymp \frac{1}{d}$, and $\mathrm{vol}(B_2)^{1/d} \asymp \frac{1}{\sqrt{d}}$. On the other hand,

$$M(B_1, \|\cdot\|_2, \epsilon) \geq \frac{\mathrm{vol}(B_1)}{\mathrm{vol}(\epsilon B_2)} = \left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(B_1)}{\mathrm{vol}(B_2)} = \left(\frac{c}{\epsilon\sqrt{d}}\right)^d.$$

Note that the lower bound derived above is useful only when $\epsilon \lesssim \frac{1}{\sqrt{d}}$.

15.1 Upper bound via Sudakov minorization

Recall that the Gaussian width of $\Theta \subset \mathbb{R}^d$ is defined as[1]

$$w(\Theta) = \mathbb{E}\sup_{\theta\in\Theta}\langle\theta, Z\rangle, \qquad \text{where } Z \sim N(0, \mathbf{I}_d).$$

Theorem 15.1 (Sudakov minorization)**.** *For any $\Theta \in \mathbb{R}^d$ and any $\epsilon > 0$,*

$$w(\Theta) \gtrsim \epsilon\sqrt{\log M(\Theta, \|\cdot\|_2, \epsilon)}.$$

The proof of Theorem 15.1 relies on Slepian's Gaussian comparison lemma:

Lemma 15.1 (Slepian's lemma)**.** *Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be Gaussian random vectors. If $\mathbb{E}(Y_i - Y_j)^2 \leq \mathbb{E}(X_i - X_j)^2$ for all i, j, then $\mathbb{E}\max Y_i \leq \mathbb{E}\max X_i$.*

For a self-contained proof see [Cha05].[2] See also [Pis99, Lemma 5.7, p. 70] for a simpler proof of a weaker version $\mathbb{E}\max X_i \leq 2\mathbb{E}\max Y_i$, which suffices for our purposes though.

[1] To avoid measurability difficulty, $w(\Theta)$ should be understood as $\sup_{T\subset\Theta, |T|<\infty}\mathbb{E}\max_{\theta\in T}\langle\theta, Z\rangle$.

[2] If you took ECE 534 last fall, you should revisit Problem 5 of `http://maxim.ece.illinois.edu/teaching/fall15a/homework/hw4.pdf` which follows [Cha05].

Proof of Theorem 15.1 assuming Slepian. Let $\{\theta_1, \dots, \theta_M\}$ be an he optimal ϵ-packing of Θ. Let $X_i = \langle \theta_i, Z\rangle$ for $i \in [M]$, where $Z \sim \mathcal{N}(0, I_d)$. Let $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \epsilon^2/2)$. Then for any pair i, j, X_i and X_j are jointly Gaussian, and

$$\mathbb{E}(X_i - X_j)^2 = (\theta_i - \theta_j)'\mathbb{E}[ZZ'](\theta_i - \theta_j) = \|\theta_i - \theta_j\|_2^2 \geq \epsilon^2 = \mathbb{E}(Y_i - Y_j)^2.$$

It follows from Lemma 15.1 that

$$\mathbb{E} \max_{1 \leq i \leq M} X_i \geq \mathbb{E} \max_{1 \leq i \leq M} Y_i \asymp \epsilon\sqrt{\log M}.$$

This completes the proof because $\mathbb{E}\sup_{\theta\in\Theta}\langle \theta, Z\rangle \geq \mathbb{E}\max_{1\leq i\leq M} X_i$. □

We can apply this theorem to $\Theta = B_1$. In this case, by the definition of the dual norm,

$$w(B_1) = \mathbb{E} \sup_{x \in \mathbb{R}^d \colon \|x\|_1 \leq 1} \langle x, Z\rangle = \mathbb{E}\|Z\|_\infty \asymp \sqrt{\log d}.$$

The theorem then implies that

$$\log M(B_1, \|\cdot\|_2, \epsilon) \lesssim \frac{\log d}{\epsilon^2}. \tag{15.1}$$

This bound is almost optimal: When $\epsilon \gg 1/\sqrt{d}$, this upper bound is (in fact optimal and) much better than what we get from the volume argument, which is

$$\log M(B_1, \|\cdot\|_2, \epsilon) \lesssim d \log\left(1 + \frac{c}{\epsilon\sqrt{d}}\right).$$

However, (15.1) is not always sharp. For example, when $\epsilon \asymp 1/\sqrt{d}$, it gives $d\log d$ and we know (even from volume bound) that the correct behavior is d. This suggests we need a more refined bound that interpolates between volume and Sudakov.

15.2 Upper bound via Maurey's empirical method

We can construct a covering of B_1 using the probabilistic method. Let $\{e_i,\ i = 1, \dots, d\}$ be the standard basis of $\mathbb{R}^d$. For an arbitrary $x \in B_1$, define a d dimensional random vector Z as

$$Z = \begin{cases} \operatorname{sgn}(x_i)e_i & \text{w.p. } |x_i| \\ 0 & \text{w.p. } 1 - \|x\|_1 \end{cases}$$

Z has the property that $\mathbb{E}Z_i = x_i$ for $i = 1, \dots, d$, hence $\mathbb{E}Z = x$, and $\operatorname{Var}[Z_i] = \mathbb{E}(Z_i - x_i)^2$ for $i = 1, \dots, d$. Let $Z_{(1)}, \dots, Z_{(k)}$ be i.i.d. copies of Z, and let $\bar{Z} = \frac{1}{k}\sum_{j=1}^k Z_{(j)}$. Then

$$\mathbb{E}\|\bar{Z} - x\|_2^2 = \sum_{i=1}^d \mathbb{E}(\bar{Z}_i - x_i)^2 = \sum_{i=1}^d \operatorname{Var}[\bar{Z}_i] = \frac{1}{k}\sum_{i=1}^d \operatorname{Var}[Z_i] = \frac{1}{k}\mathbb{E}\|Z - x\|_2^2 \leq \frac{1}{k}\mathbb{E}\|Z - x\|_1^2 \leq \frac{1}{k},$$

where we have used the facts that $\operatorname{Var}[\bar{Z}_i] = \frac{1}{k}\operatorname{Var}[Z_i]$ and $\|Z - x\|_2 \leq \|Z - x\|_1 \leq 1$. If we choose $k = 1/\epsilon^2$, then $\mathbb{E}\|\bar{Z} - x\|_2 \leq \sqrt{\mathbb{E}\|\bar{Z} - x\|_2^2} \leq \epsilon$. So there is a realization $\bar{z}$ of $\bar{Z}$ such that

$$\|\bar{z} - x\|_2 \leq \epsilon.$$

Now we examine how many different values $\bar{Z}$ can take regardless of x. Note that

$$\bar{Z} = \frac{1}{k}\sum_{j=1}^{k} Z_{(j)} = \frac{1}{k}(K_1, \ldots, K_d),$$

where

$$\sum_{i=1}^{d} K_i \le k, \qquad \text{with } K_i \in \mathbb{Z}, \text{ and } 0 \le |K_i| \le k \text{ for } i = 1, \ldots, d. \tag{15.2}$$

For any $(K_1, \ldots, K_d)$ satisfying inequality (15.2), we get a solution for the following inequality

$$\sum_{i=1}^{d} K_i^+ + K_i^- \le k, \qquad \text{with } K_i^+, K_i^- \in \mathbb{Z}, \text{ and } 0 \le K_i^+, K_i^- \le k \text{ for } i = 1, \ldots, d, \tag{15.3}$$

by setting $K_i^+ = K_i$ and $K_i^- = 0$ if $K_i \ge 0$, and setting $K_i^+ = 0$ and $K_i^- = -K_i$ if $K_i < 0$. Therefore, the number of values $\bar{Z}$ can take is upper bounded by the number of solutions of inequality (15.3). Note that there are $\binom{k+2d-1}{2d-1}$ solutions for

$$\sum_{i=1}^{d} K_i^+ + K_i^- = k, \qquad \text{with } K_i^+, K_i^- \in \mathbb{Z}, \text{ and } 0 \le K_i^+, K_i^- \le k \text{ for } i = 1, \ldots, d,$$

because the solutions are all possible types of the sequences of length k with alphabet size $2d$. It follows that the number of solutions of inequality (15.3) is

$$\binom{0+2d-1}{2d-1} + \binom{1+2d-1}{2d-1} + \ldots + \binom{k+2d-1}{2d-1} = \binom{k+2d}{2d} = \binom{k+2d}{k},$$

which is an upper bound on the number of $\bar{Z}$'s regardless of x. We thus have shown the existence of an ϵ-covering of B_1 in $\|\cdot\|_2$ with cardinality upper bounded by

$$\binom{\frac{1}{\epsilon^2} + 2d}{2d} = \binom{\frac{1}{\epsilon^2} + 2d}{\frac{1}{\epsilon^2}}.$$

Therefore,

$$\log N(B_1, \|\cdot\|_2, \epsilon) \le 2d \log\left(1 + \frac{1}{2\epsilon^2 d}\right) \wedge \frac{1}{\epsilon^2} \log\left(1 + 2d\epsilon^2\right).$$

We can see that the first upper bound recovers the result from the volume argument, while the second upper bound is even stronger than the result obtained from Sudakov's minorization.

15.3 Lower bound via packing Hamming spheres

Let $S_k = \{x \in \{0,1\}^d : w_{\mathrm{H}}(x) = k\}$ be the Hamming sphere of radius k. For the 2ρ-packing of S_k in Hamming distance $\|\cdot\|_{\mathrm{H}}$, we have

$$\log M(S_k, \|\cdot\|_{\mathrm{H}}, 2\rho) \ge \frac{|S_k|}{|B_{\mathrm{H}}(\rho)|} = \frac{\binom{d}{k}}{\sum_{i=0}^{\rho} \binom{d}{i}}.$$

This leads to the following lemma.

Lemma 15.2 (Gilbert-Varshamov). *There exist constants c_1 and c_2 such that for all $d \in \mathbb{N}$ and any $k \in [d]$,*

$$\log M(S_k, \|\cdot\|_{\mathrm{H}}, c_1 k) \geq c_2 k \log \frac{ed}{k}.$$

Now we construct a packing of B_1 based on a packing of S_k. Let $\{x_1, \ldots, x_M\}$ be a $c_1 k$-packing of S_k. Let $\theta_i = x_i/k$. Then $\theta_i \in B_1$ for $i = 1, \ldots, M$, and

$$\|\theta_i - \theta_j\|_2^2 = \frac{1}{k^2}\|x_i - x_j\|_{\mathrm{H}} \geq \frac{c_1}{k}.$$

Therefore, $\{\theta_1, \ldots, \theta_M\}$ is a $\sqrt{c_1/k}$-packing of B_1 in $\|\cdot\|_2$. Choosing $k = 1/\epsilon^2$, it follows from Lemma 15.2 that

$$\log M(B_1, \|\cdot\|_2, \sqrt{c_1}\epsilon) \geq \frac{c_2}{\epsilon^2} \log\left(ed\epsilon^2\right)$$

for some constants c_1 and c_2.

To summarize, combining the upper and lower bounds, we have

$$\log N(B_1, \|\cdot\|_2, \epsilon) \asymp \begin{cases} \frac{1}{\epsilon^2}\log\left(\epsilon^2 d\right) & \epsilon \gtrsim \frac{1}{\sqrt{d}} \\ d & \epsilon \asymp \frac{1}{\sqrt{d}} \\ d \log \frac{1}{\epsilon^2 d} & \epsilon \lesssim \frac{1}{\sqrt{d}} \end{cases}. \tag{15.4}$$

15.4 Duality

First we define a more general notoin of covering number. For $K, T \subset \mathbb{R}^d$, define the covering number of K using translates of T as

$$N(K, T) = \min\{N : \exists x_1, \ldots, X_N \in \mathbb{R}^d \text{ such that } K = \cup_{i=1}^N T + x_i\}.$$

An amazing theorem of Artstein-Milman-Szarek [AMS04] establishes the following duality result for metric entropy: There exist constants α and β such that for any symmetric convex body K,

$$\frac{1}{\beta}\log N\left(B_2, \frac{\epsilon}{\alpha}K^\circ\right) \leq \log N(K, \epsilon B_2) \leq \log N(B_2, \alpha\epsilon K^\circ),$$

where B_2 is the usual unit ℓ_2-ball,

$$K^\circ = \left\{y : \sup_{x \in K}\langle x, y\rangle \leq 1\right\}$$

is the polar body of K. For example, $B_p^\circ = B_q$ whenever $\frac{1}{p} + \frac{1}{q} = 1$. Therefore by duality, (15.4) also applies to $\log N(B_2, \|\cdot\|_\infty, \epsilon)$, which is what is needed for application to minimax risk.

15.5 Example

Finally, we use the results in this lecture to derive the minimax lower bound for the p-dimension, n-sample Gaussian location model with respect to the distortion function $\|\theta - \hat{\theta}\|_\infty^2$.

We can construct an ϵ-packing of $B_2(\delta)$ in $\|\cdot\|_\infty$. From the Fano's method,

$$
\begin{aligned}
R^* &\gtrsim \epsilon^2 \left(1 - \frac{\mathrm{diam}_{\mathrm{KL}}(\{N(\theta, \frac{1}{n}\mathbf{I}_p), \theta \in B_2(\delta)\}) + \log 2}{\log M(B_2(\delta), \|\cdot\|_\infty, \epsilon)}\right) \\
&= \epsilon^2 \left(1 - \frac{n\delta^2 + \log 2}{\log M(B_2(\delta), \|\cdot\|_\infty, \epsilon)}\right) \\
&= \epsilon^2 \left(1 - \frac{n\delta^2 + \log 2}{\log M(B_1, \|\cdot\|_2, \epsilon/\delta)}\right) \\
&\gtrsim \epsilon^2 \left(1 - \frac{n\delta^2 + \log 2}{\frac{\delta^2}{\epsilon^2} \log\left(1 + \frac{p\epsilon^2}{\delta^2}\right)}\right)
\end{aligned}
$$

where we have used that fact that $\mathrm{diam}_{\mathrm{KL}}(\{N(\theta, \frac{1}{n}\mathbf{I}_p), \theta \in B_2(\delta)\}) = n\mathrm{diam}^2_{\|\cdot\|_2}(B_2(\delta))$, the duality theorem, and the upper bound on $\log M(B_1, \|\cdot\|_2, \epsilon/\delta)$. Choosing $\epsilon = c_1\sqrt{\frac{\log p}{n}}$ and $\delta = c_2\epsilon$ with appropriate c_1 and c_2 such that the parenthesis in the lower bound is a positive constant, we obtain

$$
R^* \gtrsim \frac{\log p}{n}.
$$

An alternative proof of this result is by choosing the packing set as $\tau\{e_1, \ldots, e_p\}$ for some $\tau > 0$ to be determined later. This set is a τ-packing of $\mathbb{R}^d$ in $\|\cdot\|_\infty$, because $\|\tau(e_i - e_j)\|_\infty = \tau$ for all pairs $\{i, j\}$. We also have $\|\tau(e_i - e_j)\|_2^2 = 2\tau^2$. Then by Fano's method,

$$
R^* \geq \tau^2 \left(1 - \frac{2n\tau^2 + \log 2}{\log p}\right).
$$

Choosing $\tau = c\sqrt{\frac{\log p}{n}}$ with some appropriate constant c such that the parenthesis in the above bound is a positive constant, we obtain

$$
R^* \gtrsim \frac{\log p}{n}.
$$

References

[AMS04] Shiri Artstein, Vitali Milman, and Stanisław J Szarek. Duality of metric entropy. *Annals of mathematics*, pages 1313–1328, 2004.

[Cha05] Sourav Chatterjee. An error bound in the Sudakov-Fernique inequality. *arXiv preprint arXiv:0510424*, 2005.

[Pis99] G. Pisier. *The volume of convex bodies and Banach space geometry.* Cambridge University Press, 1999.