

## Lecture 10: February 8, 2025

Lecturer: Madhur Tulsiani

## 1 Lower bounds for minimax rates via multiple hypotheses

We now consider a high-dimensional problem, where we can prove lower bounds using bounds for testing multiple hypotheses. Recall that for a random variable  $V$  uniformly distributed over a set of hypotheses  $\mathcal{V}$ , the probability of error for any classifier  $T(\bar{\mathbf{x}})$  with input  $\bar{\mathbf{x}}$  coming from  $P_v^n$  for a randomly chosen  $v \in \mathcal{V}$ , is lower bounded as

$$\mathbb{P}[T(\bar{\mathbf{x}}) \neq V] \geq 1 - \frac{n \cdot \mathbb{E}_{v_1, v_2 \in \mathcal{V}} [D(P_{v_1} \| P_{v_2})] + 1}{\log |\mathcal{V}|}.$$

As before, we will combine the above bound with our reduction to hypothesis testing, to prove the desired lower bound on the minimax rate using

$$\mathcal{M}_n(\Pi, \ell) = \inf_{\hat{\theta}} \sup_{P \in \Pi} \mathbb{E}_{\bar{\mathbf{x}} \sim P^n} [\ell(\hat{\theta}(\bar{\mathbf{x}}), \theta(P))] \geq \Phi(\delta) \cdot \inf_T \{\mathbb{P}[T(\bar{\mathbf{x}}) \neq V]\}$$

To use the above bounds, we need to come up with a set of distributions which are far in terms of the property  $\theta$  (so that the second bound is large), but close on average in terms of KL-divergence (so that the first bound is large). This is also known as the *local Fano* method since we derived the first bound using Fano's inequality, and are applying it by using (a local bound on) KL-divergence for every pair of distributions  $P_{v_1}, P_{v_2}$  (recall that we used convexity of KL-divergence to reduce to the local setting). You can find other variants of this method in the notes by Duchi [Duc16].

### 1.1 Gaussian mean estimation

While binary hypothesis testing was used show a bound for estimating the mean of Bernoulli random variables, the multiple hypotheses setting is often useful in considering high-dimensional problems. We take  $\Pi$  to be the set of  $d$ -dimensional Gaussian distributions as below

$$\Pi = \left\{ N(\mu, I_d) \mid \mu \in \mathbb{R}^d \right\}.$$

Let the property  $\theta$  be the mean as before, and let  $\ell(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2$ . We first check the expected loss for the empirical mean estimator.

**Proposition 1.1.** Let  $\hat{\theta}(x_1, \dots, x_n) = \frac{1}{n} \cdot \sum_{i \in [n]} x_i$ . Then, for any  $\mu \in \mathbb{R}^d$ , we have that

$$\mathbb{E}_{\bar{x} \sim (N(\mu, I_d))^n} \left[ \left\| \frac{1}{n} \sum_{i \in [n]} X_i - \mu \right\|_2^2 \right] = \frac{d}{n}.$$

**Proof:** The proof is similar to the case of Bernoulli random variables. By the (pairwise) independence of the  $n$  samples, we have that

$$\begin{aligned} \mathbb{E}_{\bar{x} \sim (N(\mu, I_d))^n} \left[ \left\| \frac{1}{n} \sum_{i \in [n]} X_i - \mu \right\|_2^2 \right] &= \frac{1}{n^2} \cdot \sum_{i \in [n]} \mathbb{E} \left[ \|X_i - \mu\|_2^2 \right] + \frac{1}{n^2} \cdot \sum_{i \neq j} \mathbb{E} [\langle X_i - \mu, X_j - \mu \rangle] \\ &= \frac{n}{n^2} \cdot \mathbb{E}_{X \sim N(\mu, I_d)} \left[ \|X - \mu\|_2^2 \right] \\ &= \frac{n}{n^2} \cdot d = \frac{d}{n}. \end{aligned}$$

■

We will apply the local Fano method to prove the optimality of the above bound in terms of both  $d$  and  $n$ . We first need the following expression for KL-divergence of two normal distributions.

**Exercise 1.2.** Prove (using the chain rule) that

$$D(N(\mu_1, I_d) \parallel N(\mu_2, I_d)) = \frac{1}{2 \ln 2} \cdot \|\mu_1 - \mu_2\|_2^2.$$

Thus, we need to find a large collection of distributions, equivalent to finding a large collection of means, such that for any two  $\mu_1 \neq \mu_2$ , we have that  $\|\mu_1 - \mu_2\|$  is somewhat large (to lower bound the loss), but still  $\|\mu_1 - \mu_2\|$  is small on average (to upper bound the average KL-divergence). This is the content of the following lemma.

**Lemma 1.3** (Packing lemma). *There exists a collection of vectors  $\mathcal{V} \subseteq \mathbb{R}^d$  such that  $|\mathcal{V}| \geq 2^d$  and for all  $v_1, v_2 \in \mathcal{V}, v_1 \neq v_2$ , we have*

$$\frac{1}{2} \leq \|v_1 - v_2\|_2 \leq 2.$$

It is actually quite easy to prove the packing lemma above (with slightly weaker parameters) but we will take a slightly longer route through covering and packing numbers to illustrate a general method. We first the lower bound, assuming the packing lemma. Let  $\mathcal{V}$  be a collection as in [Lemma 1.3](#). We consider the set of distributions

$$\{N(4\delta \cdot v, I_d) \mid v \in \mathcal{V}\}.$$

We have that for all  $P, P' \in \Pi$ ,  $\|\theta(P) - \theta(P')\| \geq 2\delta$ . Also, since  $\|v - v'\| \leq 2$  for any  $v, v' \in \mathcal{V}$ , we get that for any  $P, P' \in \Pi$ , the means are at distance at most  $8\delta$ . Hence,

$$D(P\|P') = \frac{1}{2\ln 2} \cdot \|\mu - \mu'\|_2^2 \leq \frac{1}{2\ln 2} \cdot (64\delta^2) = \frac{32\delta^2}{\ln 2}.$$

Applying the lower bound on minimax loss in terms of KL-divergences gives

$$\mathcal{M}_n(\Pi, \ell) \geq \delta^2 \cdot \left(1 - \frac{n \cdot (32\delta^2 / \ln 2) + 1}{\log |\mathcal{V}|}\right) \geq \delta^2 \cdot \left(1 - \frac{n \cdot (32\delta^2 / \ln 2) + 1}{d}\right).$$

## 1.2 Covering and packing numbers

**Definition 1.4.** Let  $S$  be a set of points with a metric  $\rho(\cdot, \cdot)$ . A collection of points  $\mathcal{C} \subseteq S$  is called a  $\delta$ -covering of  $S$  (with respect to the metric  $\rho$ ) if

$$\forall x \in S, \exists y \in \mathcal{C} \quad \rho(x, y) \leq \delta.$$

A set of points  $\mathcal{P}$  is called a  $\delta$ -packing if

$$\forall x, y \in \mathcal{P}, x \neq y \quad \rho(x, y) > \delta.$$

The size of the minimum  $\delta$ -covering, denoted as  $N(\delta, S, \rho)$ , is called the  $\delta$ -covering number of  $S$  and the size of the maximum  $\delta$ -packing is called the  $\delta$ -packing number. The quantity  $\log N(\delta, S, \rho)$  is also called the metric entropy of  $S$ .

We will take the required collection in [Lemma 1.3](#) to be a  $(1/2)$ -packing of the unit ball in  $\mathbb{R}^d$  (under the Euclidean distance). We will show a lower bound on the size of this collection (the packing number) by using a relationship between the packing and covering numbers.

**Exercise 1.5.** For any set  $S$ , metric  $\rho$  and  $\delta > 0$ , show that

$$M(2\delta, S, \rho) \leq N(\delta, S, \rho) \leq M(\delta, S, \rho).$$

**(Hint:** First prove that an optimal  $\delta$ -packing must also be a  $\delta$ -covering.)

Let  $B_d(x, r)$  denote the ball in  $\mathbb{R}^d$  of radius  $r$  (in the Euclidean distance) with its center at  $x$ . We know that  $\text{Vol}(B_d(x, r)) = c_d \cdot r^d$  for some constant  $c_d \geq 0$ . Note that if  $\mathcal{C} \subseteq B_d(0, 1)$  is a  $\delta$ -covering of  $B_d(0, 1)$ , then

$$B(0, 1) \subseteq \bigcup_{x \in \mathcal{C}} B_d(x, \delta).$$

Thus, we have

$$c_d = \text{Vol}(B_d(0,1)) \leq \sum_{x \in \mathcal{C}} \text{Vol}(B_d(x,\delta)) = N(\delta, B_d(0,1), \|\cdot\|_2) \cdot c_d \cdot \delta^d.$$

Combining with the previous relationship between covering and packing numbers, this gives

$$M(\delta, B_d(0,1), \|\cdot\|_2) \geq N(\delta, B_d(0,1), \|\cdot\|_2) \geq \frac{1}{\delta^d}.$$

Thus, there exists a  $(1/2)$ -packing of  $B_d(0,1)$  of size at least  $2^d$ . Note that for any  $v_1, v_2$  in the packing

$$\frac{1}{2} < \|v_1 - v_2\| \leq \|v_1\| + \|v_2\| \leq 2,$$

which proves the packing lemma.

### 1.3 Another proof of (a weaker) packing lemma

One can also sample points on the Boolean cube  $\{-1,1\}^d$  to prove a weaker version of the packing lemma, which also suffices for our application. For some applications, this additional structure in the set of points may be helpful.

**Lemma 1.6.** *There exists a collection of vectors  $\mathcal{V} \subseteq \frac{1}{\sqrt{d}} \cdot \{-1,1\}^d$  such that  $|\mathcal{V}| \geq 2^{d/20}$  and for all  $v_1, v_2 \in \mathcal{V}, v_1 \neq v_2$ , we have*

$$\frac{1}{2} \leq \|v_1 - v_2\|_2 \leq 2.$$

**Remark 1.7.** *The above bound is a crude one and is just proved as an illustration. A more sophisticated sampling argument can prove a lower bound of  $\exp(d/8)$  on the size of the set  $\mathcal{V}$ .*

**Proof:** Sample  $y_1, \dots, y_N \in \{-1,1\}^d$  uniformly and independently at random, and take  $\mathcal{V} = \left\{ \frac{1}{\sqrt{d}} \cdot y_1, \dots, \frac{1}{\sqrt{d}} \cdot y_N \right\}$ . For a fixed pair  $v_i, v_j$ , we have

$$\|v_i - v_j\| \leq \frac{1}{2} \Leftrightarrow \langle v_i, v_j \rangle \geq \frac{7}{8} \Leftrightarrow \langle y_i, y_j \rangle \geq \frac{7d}{8}$$

Since  $\langle y_i, y_j \rangle$  is a sum of  $d$  independent variables in  $\{-1,1\}$ , we have by Chernoff-Hoeffding bounds that

$$\mathbb{P} \left[ \langle y_i, y_j \rangle \geq \frac{7d}{8} \right] \leq 2^{-\frac{d}{6} \cdot \left(\frac{7}{8}\right)^2} \leq 2^{-d/8}.$$

By a union bound

$$\mathbb{P} \left[ \exists i, j \in [N] \langle y_i, y_j \rangle \geq \frac{7d}{8} \right] \leq N^2 \cdot 2^{-d/8},$$

The probability above is  $o(1)$  for  $N \ll 2^{d/16}$ , and thus, with high probability, we have for all  $i \neq j$ ,  $\|v_i - v_j\| \geq \frac{1}{2}$ . The upper bound on the distance also holds since all vectors lie on the unit sphere. ■

## 2 Sparse mean estimation

We will conclude our discussion of minimax rates, with this final example of estimating the mean, when we are given the additional condition that the mean is a *sparse* vector. Consider the set of normal distributions, where the mean has only *one* non-zero coordinate.

$$\Pi = \left\{ N(\mu, I_d) \mid \mu \in \mathbb{R}^d, \|\mu\|_0 \leq 1 \right\}.$$

Let  $\theta(P) = \mathbb{E}_{x \sim P} [x]$  be the mean, and let  $\ell(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2$  as before. From the previous examples, it seems like the empirical mean estimator is always the best one, and the role of information theory is primarily for proving lower bounds. However, it can also serve as a guide for the right bound to aim for. For this problem, it will be much easier to prove a lower bound. We will then show an estimator which matches this bound.

### 2.1 Lower bound

Let  $\mathcal{V} = \{e_1, \dots, e_d\}$  be the set of standard basis vectors in  $\mathbb{R}^d$ . Consider the set of distributions  $P_v = N(\sqrt{2\delta} \cdot v, I_d)$  for all  $v \in \mathcal{V}$ . Note that the means  $\mu_v = \sqrt{2\delta} \cdot v$  satisfy  $\|\mu_{v_1} - \mu_{v_2}\| = 2\delta$  for all  $v_1 \neq v_2$ . Using the bound from the previous lecture, we get

$$\begin{aligned} \mathcal{M}_n(\Pi, \ell) &\geq \delta^2 \cdot \left( 1 - \frac{n \cdot \mathbb{E}_{v_1, v_2 \in \mathcal{V}} [D(P_{v_1} \| P_{v_2})] + 1}{\log |\mathcal{V}|} \right) \\ &\geq \delta^2 \cdot \left( 1 - \frac{n \cdot (4\delta^2 / (2 \ln 2)) + 1}{\log d} \right) \\ &\geq c \cdot \frac{\log d}{n}, \end{aligned}$$

for an appropriate constant  $c > 0$ , using a choice of  $\delta^2 = c' \cdot \frac{\log d}{n}$ . We will now show that this lower bound is actually tight.

### 2.2 Upper bound

The optimal estimator for the above problem actually extends the definition of the mean as the minimizer of the total square distance (from the sample points). Recall the following.

**Exercise 2.1.** Let  $x_1, \dots, x_n \in \mathbb{R}^d$ . Then the empirical mean  $\eta = \frac{1}{n} \cdot \sum_{i=1}^n x_i$  satisfies

$$\sum_{i=1}^n \|x_i - \eta\|_2^2 = \inf_{v \in \mathbb{R}^d} \left\{ \sum_{i=1}^n \|x_i - v\|_2^2 \right\}.$$

Given a sequence of samples  $\bar{x} = (x_1, \dots, x_n)$ , let the  $\eta$  denote the empirical mean

$$\eta := \frac{1}{n} \cdot \sum_{i=1}^n x_i.$$

As we saw above, the empirical mean is the minimizer of the least square distance. However, it is not sparse. We take our estimator  $\hat{\mu}$  to only consist of the largest entry (in absolute value) of  $\eta$ , and set all other entries to zero i.e.,

$$\hat{\mu}_j := \begin{cases} \eta_j & \text{if } j = \operatorname{argmax}_{k \in [d]} |\eta_k| \\ 0 & \text{otherwise} \end{cases}.$$

Note that the above definition does not make sense if the the coordinate maximizing  $|\eta_k|$  is not unique. In such a case, we arbitrarily pick one of the maximizing coordinates. Check that this definition is a constrained version of the above definition for empirical mean. While the empirical mean  $\eta$  is the minimizer over all of  $\mathbb{R}^d$ , of the average squared distance from the sample points, the estimator above is the minimizer over all sparse vectors.

**Exercise 2.2.** Check that for  $\hat{\mu}$  defined as above

$$\sum_{i=1}^n \|x_i - \hat{\mu}\|_2^2 = \inf_{\|v\|_0 \leq 1} \left\{ \sum_{i=1}^n \|x_i - v\|_2^2 \right\}.$$

While we will use the above estimator, the operation of picking the largest coordinate does not combine well with analytic expressions such as expectation etc. For this reason, we will use the empirical mean  $\eta$  as an intermediate object in the analysis. We need the following basic properties

**Proposition 2.3.** Let  $\bar{x} \sim (N(\mu, I_d))^n$  be a sequence of  $n$  independent samples, and let  $\eta = \frac{1}{n} \cdot \sum_{i=1}^n x_i$  be the empirical mean. Then  $\eta - \mu$  is distributed according to the Gaussian distribution  $N(0, \frac{1}{n} \cdot I_d)$ .

**Proof:** Since different coordinates are independent in each of  $x_1, \dots, x_n$ , they are also independent in  $\delta - \mu$ . For any single coordinate  $j \in [d]$ , we have

$$(\eta - \mu)_j = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)_j.$$

By definition of  $(x_1, \dots, x_n)$ , each term  $(x_i - \mu)_j$  is independently distributed according to  $N(0, 1)$ . Since a linear combination of independent Gaussians is still a Gaussian, and variances add for independent variables, we get

$$\text{Var} [(\eta - \mu)_j] = \frac{1}{n^2} \sum_{i=1}^n \text{Var} [(x_i - \mu)_j] = \frac{1}{n^2} \cdot n = \frac{1}{n}.$$

Combined with  $\mathbb{E} [x_i - \mu] = 0$ , this completes the proof.  $\blacksquare$

**Corollary 2.4.** *Let  $\bar{x} = (x_1, \dots, x_n) \sim (N((\mu, I_d)))^n$  as above. Then,*

$$\mathbb{P} [\exists j \in [d] \quad |\mu_j - \eta_j| \geq t] \leq 2d \cdot \exp(-nt^2/2).$$

**Proof:** Using the standard Gaussian tail bound, we know that for  $y \sim N(0, \sigma^2)$ , we have

$$\mathbb{P} [|y| \geq t] \leq 2 \cdot \exp(-t^2/(2\sigma^2)).$$

Using [Proposition 2.3](#) for each coordinate  $\eta_j - \mu_j$ , and taking a union bound over all  $j \in [d]$  gives the desired bound.  $\blacksquare$

Recall that our goal is to bound the expected loss  $\mathbb{E}_{\bar{x} \sim (N(\mu, I_d))^n} [\|\mu - \hat{\mu}(\bar{x})\|_2^2]$ . Using the above, we can first prove a tail bound: the probability that the loss is too large, is small.

**Claim 2.5.** *For the estimator  $\hat{\mu}$  as above*

$$\mathbb{P} [\|\mu - \hat{\mu}\|_2 \geq t] \leq 2d \cdot \exp(-nt^2/18).$$

**Proof:** We will prove that

$$\|\mu - \hat{\mu}\|_2 \geq t \quad \Rightarrow \quad \exists j \in [d] \quad |\eta_j - \mu_j| \geq t/3.$$

Using this, together with [Corollary 2.4](#) will prove the claim. Recall that both  $\mu$  and  $\hat{\mu}$  have at most one non-zero coordinate. If  $\mu = 0$  and  $\hat{\mu}_j \neq 0$ , then we must have  $|\hat{\mu}_j| = |\eta_j - \mu_j| \geq t$ . The case when  $\hat{\mu} = 0$  can be handled similarly.

If  $\mu \neq 0$ , then let unique the non-zero coordinate be 1 (without loss of generality) i.e.,  $|\mu_1| > 0$ . If  $\hat{\mu}_1 \neq 0$ , then we again have

$$|\mu_1 - \eta_1| = |\mu_1 - \hat{\mu}_1| = \|\mu - \hat{\mu}\|_2 \geq t,$$

and we are done. So let's assume  $\hat{\mu}_1 = 0$  and  $\hat{\mu}_j \neq 0$  for some  $j > 1$ . Since we must have  $\hat{\mu}_j = \eta_j$  in this case, we have

$$|\mu_1| + |\eta_j| = |(\mu - \hat{\mu})_1| + |(\mu - \hat{\mu})_j| \geq \|\mu - \hat{\mu}\|_2 \geq t.$$

Also, since  $\eta_j$  must be the largest coordinate in absolute value, we have

$$|\eta_j| \geq |\eta_1| \geq |\mu_1| - |\mu_1 - \eta_1|.$$

Adding the above inequalities gives

$$|\mu_1 - \eta_1| + 2 \cdot |\eta_j| = |\mu_1 - \eta_1| + 2 \cdot |\mu_j - \eta_j| \geq t.$$

Hence, either  $|\mu_1 - \eta_1| \geq t/3$  or  $|\mu_j - \eta_j| \geq t/3$ , which is what we wanted to prove. ■

We can now finish the computation of the expected loss, using the above tail bound. Using  $s = t^2$  in the above bound, we can write it as

$$\mathbb{P} \left[ \|\mu - \hat{\mu}\|_2^2 \geq s \right] \leq 2d \cdot \exp(-ns/18).$$

This yields the following bound.

**Claim 2.6.** For the estimator  $\hat{\mu}$  as above

$$\mathbb{E}_{\bar{\mathbf{x}} \sim (N(\mu, I_d))^n} \left[ \|\mu - \hat{\mu}(\bar{\mathbf{x}})\|_2^2 \right] = O\left(\frac{\log d}{n}\right).$$

**Proof:** We use the fact that for a non-negative random variable  $Z$ ,  $\mathbb{E}[Z] = \int_s \mathbb{P}[Z \geq s]$ . Using this, we get

$$\begin{aligned} \mathbb{E}_{\bar{\mathbf{x}} \sim (N(\mu, I_d))^n} \left[ \|\mu - \hat{\mu}(\bar{\mathbf{x}})\|_2^2 \right] &= \int_0^\infty \mathbb{P} \left[ \|\mu - \hat{\mu}\|_2^2 \geq s \right] ds \\ &= \int_0^u \mathbb{P} \left[ \|\mu - \hat{\mu}\|_2^2 \geq s \right] ds + \int_u^\infty \mathbb{P} \left[ \|\mu - \hat{\mu}\|_2^2 \geq s \right] ds \\ &\leq \int_0^u 1 ds + \int_u^\infty 2d \cdot \exp(-ns/18) ds \\ &= u + \frac{36d}{n} \cdot \exp(-nu/18). \end{aligned}$$

Choosing  $u = c \cdot \frac{\log d}{n}$  for an appropriate constant  $c$ , then finishes the proof. ■

## References

[Duc16] John Duchi, *Lecture notes on Information Theory and Statistics*, 2016. 1