Information and Coding Theory **Winter 2021**

Lecture 10: February 11, 2021

Lecturer: Madhur Tulsiani

1 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 $\ell(\widehat{\theta}, \theta) = \left\|\widehat{\theta} - \theta\right\|_2^2$ 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.

1.1 Lower bound

Let $\mathcal{V} = \{e_1, \ldots, e_d\}$ 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

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

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.

1.2 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 1.1. *Let $x_1, \ldots, x_n \in \mathbb{R}^d$. 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 $\bar{\mathbf{x}} = (x_1, \ldots, x_n)$, let the η 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 $\widehat{\mu}$ to only consist of the largest entry (in absolute value) of η, and set all other entries to zero i.e.,

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

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 η 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 1.2. *Check that for $\widehat{\mu}$ defined as above*

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

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 η as an intermediate object in the analysis. We need the following basic properties

Proposition 1.3. *Let $\bar{\mathbf{x}} \sim (N(\mu, I_d))^n$ 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 $N\left(0, \frac{1}{n} \cdot I_d\right)$.*

Proof: Since different coordinates are independent in each of $x_1, \ldots, x_n$, 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 $(x_1, \ldots, x_n)$, 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

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

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

Corollary 1.4. *Let* $\overline{\mathbf{x}} = (x_1, \ldots, x_n) \sim (N((\mu, I_d))^n$ *as above. Then,*

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

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

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

Using Proposition 1.3 for each coordinate $\eta_j - \mu_j$, and taking a union bound over all $j \in [d]$ gives the desired bound. ■

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

Claim 1.5. *For the estimator* $\widehat{\mu}$ *as above*

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

Proof: We will prove that

$$\|\mu - \widehat{\mu}\|_2 \geq t \quad \Rightarrow \quad \exists j \in [d] \;\left|\eta_j - \mu_j\right| \geq t/3 \,.$$

Using this, together with Corollary 1.4 will prove the claim. Recall that both μ and $\widehat{\mu}$ have at most one non-zero coordinate. If $\mu = 0$ and $\widehat{\mu}_j \neq 0$, then we must have $\left|\widehat{\mu}_j\right| = \left|\eta_j - \mu_j\right| \geq t$. The case when $\widehat{\mu} = 0$ 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 $\widehat{\mu}_1 \neq 0$, then we again have

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

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

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

Also, since η_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

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

This yields the following bound.

Claim 1.6. *For the estimator $\widehat{\mu}$ as above*

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

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

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

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

2 I-Projections and applications

We will now talk more about finding a distribution in a set Π that minimizes $D(P\|Q)$ for a fixed distribution Q. We encountered this when discussing Sanov's theorem and hypothesis testing, and will now discuss its properties in some detail. When Q is the uniform distribution on $\mathcal{X}$. Then we also have,

$$D(P||Q) = \log|\mathcal{X}| - H(P)$$

Hence, in this case P^* is a distribution that maximizes entropy. In general, when the given information does not uniquely determine a distribution, we choose P^* that maximizes entropy. This can be thought of as picking P^* in the set of distributions Π, subject to the least amount of additional assumptions. This is sometimes called the *Maximum Entropy Principle*. In this lecture, we will characterize the distributions obtained by minimizing Kl-divergence (or maximizing entropy).

For closed convex set Π, such a P is called the I-projection of Q onto Π.

Definition 2.1. *Let Π be a closed convex set of distributions over $\mathcal{X}$. In addition, assume that* $\mathrm{Supp}(Q) = \mathcal{X}$. *Then*

$$\mathsf{Proj}_{\Pi}(Q) \;:=\; \arg\min_{P\in\Pi} D(P\|Q) \;=\; P^*$$

Note that the assumption $\mathrm{Supp}(Q) = \mathcal{X}$ above is without loss of generality since $D(P\|Q) = \infty$ for any P such that $\mathrm{Supp}(P) \not\subseteq \mathrm{Supp}(Q)$. Use the (strict) convexity of KL-divergence to check the following.

Exercise 2.2. *For a closed, convex set Π, the projection $P^* = \mathsf{Proj}_{\Pi}(Q)$ exists and is unique.*

It is immediate from definition that if $P \in \Pi$, then $D(P\|Q) \geq D(P^*\|Q)$. In fact, P^* tells us more. It also tells us how "far" P is away from Q in KL-divergence measure.

Theorem 2.3. *Let $P^* = \mathsf{Proj}_{\Pi}(Q)$. Then, for all $P \in \Pi$,*

$$\begin{aligned} \mathrm{Supp}(P) &\subseteq \mathrm{Supp}(P^*) \\ D(P\|Q) &\geq D(P\|P^*) + D(P^*\|Q) \end{aligned}$$

Proof: Define $P_t = tP + (1-t)P^*$, where $t \in [0,1]$. By minimality of P^*, it is clear that $D(P_t||Q) - D(P^*||Q) \geq 0$. By the mean value theorem, we also have that

$$0 \;\leq\; \frac{1}{t}\cdot\left(D(P_t\|Q) - D(P^*\|Q)\right) \;\leq\; \frac{d}{dt}D(P_t\|Q)\bigg|_{t=t'\in[0,t]}$$

Since $t' \to 0$ as $t \to 0$, we get

$$\lim_{t\downarrow 0}\frac{d}{dt}D(P_t\|Q) \;\geq\; 0\,.$$

We now compute $\frac{d}{dt}D(P_t||Q)$.

$$\frac{d}{dt}D(P_t\|Q) = \sum_{x\in\mathcal{X}}\frac{d}{dt}p_t(x)\log\frac{p_t(x)}{q(x)} + \sum_{x\in\mathcal{X}}p_t(x)\frac{d}{dt}(\log p_t(x) - \log q(x))$$

Note that

$$\begin{aligned}\frac{d}{dt}p_t(x) &= p(x) - p^*(x)\\ \frac{d}{dt}\log p_t(x) &= \frac{1}{\ln 2}\frac{1}{p_t(x)}(p(x) - p^*(x))\end{aligned}$$

Using these facts, we have

$$\begin{aligned}\frac{d}{dt}D(P_t||Q) &= \sum_{x\in\mathcal{X}}(p(x) - p^*(x))\log\frac{p_t(x)}{q(x)} + \sum_{x\in\mathcal{X}}\frac{1}{\ln 2}(p(x) - p^*(x))\\ &= \sum_{x\in\mathcal{X}}(p(x) - p^*(x))\log\frac{p_t(x)}{q(x)}\end{aligned}$$

Here, note that if $(\exists x)$ such that $p(x) > 0$ and $p^*(x) = 0$, then $\lim_{t\downarrow 0}\frac{d}{dt}D(P_t\|Q) \to -\infty$, which contradicts the fact that $\frac{d}{dt}D(P_t\|Q) \geq 0$. Hence, if $p(x) > 0$, then $p^*(x) > 0$ and therefore, $\text{Supp}(P) \subseteq \text{Supp}(P^*)$. This proves the first part of the theorem. Now we evaluate $\frac{d}{dt}D(P_t\|Q)$ at $t = 0$.

$$\begin{aligned}\frac{d}{dt}D(P_t||Q)|_{t=0} &= \sum_{x\in\mathcal{X}}p(x)\log\frac{p^*(x)}{q(x)} - p^*(x)\log\frac{p^*(x)}{q(x)}\\ &= \sum_{x\in\mathcal{X}}p(x)\log\frac{p^*(x)}{q(x)}\frac{p(x)}{p(x)} - D(P^*||Q)\\ &= \sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)} - \sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{p^*(x)} - D(P^*||Q)\\ &= D(P||Q) - D(P||P^*) - D(P^*||Q) \geq 0\end{aligned}$$

Hence, $D(P||Q) \geq D(P||P^*) + D(P^*||Q)$. ■

Consider the following example, which shows that the inequality can in fact be strict.

Exercise 2.4. *Let $\mathcal{X} = \{0,1\}$ and $\Pi = \{P : p(1) \leq 1/2\}$. Let Q be defined as*

$$Q = \begin{cases}1 & \textit{with prob. } 3/4\\ 0 & \textit{with prob. } 1/4\end{cases}$$

1. *Show that*
$$P^* = \begin{cases} 1 & \textit{with prob. } 1/2 \\ 0 & \textit{with prob. } 1/2 \end{cases}$$

2. *Show that* $D(P||Q) > D(P||P^*) + D(P^*||Q)$ *for the above example.*

Next, we show how to compute and characterize I-projections for some special sets of distributions.

2.1 Linear families and I-projections

Definition 2.5. *For any given real-valued functions* $f_1, f_2, ..., f_k$ *on* $\mathcal{X}$ *and* $\alpha_1, \alpha_2, ..., \alpha_k \in \mathbb{R}$*, the set*

$$\mathcal{L} = \left\{ P \mid \sum_{x \in \mathcal{X}} p(x) \cdot f_i(x) \; = \; \mathop{\mathbb{E}}_{x \sim P}[f_i(x)] \; = \; \alpha_i, \; \forall i \in [k] \right\}$$

is called a linear family *of distributions.*

We show that for linear families, the inequality proved above, is in fact tight. Moreover, the projection P^* lies in the interior of the polytope defining $\mathcal{L}$.

Lemma 2.6. *Let* $\mathcal{L}$ *be a linear family given by*

$$\mathcal{L} = \left\{ P : \sum_{x \in \mathcal{X}} p(x) \cdot f_i(x) = \alpha_i, \; i \in [k] \right\}$$

and $\bigcup_{P \in \mathcal{L}} \mathsf{Supp}(P) = \mathcal{X}$. *Let* $P^* = \mathsf{Proj}_{\mathcal{L}}(Q)$. *Then, for all* $P \in \mathcal{L}$

1. *There exists* $\beta > 0$ *such that for* $t \in [-\beta, 0]$, $P_t = tP + (1-t)P^* \in \mathcal{L}$.

2. $D(P\|Q) \; = \; D(P\|P^*) + D(P^*\|Q)$

Then the I-Projection P^* *of* Q *onto* $\mathcal{L}$ *satisfies the Pythagorean identity*

$$D(P\|Q) \; = \; D(P\|P^*) + D(P^*\|Q)$$

Proof: Recall that $\mathsf{Supp}(P) \subseteq \mathsf{Supp}(P^*)$ and $p_t(x) = t \cdot p(x) + (1-t) \cdot p^*(x)$. Since the conditions defining $\mathcal{L}$ are linear, we have that for *all* $t \in \mathbb{R}$ and all $i \in [k]$

$$\sum_{x \in \mathcal{X}} p_t(x) \cdot f_i(x) \; = \; t \cdot \sum_{x \in \mathcal{X}} p(x) \cdot f_i(x) + (1-t) \cdot \sum_{x \in \mathcal{X}} p^*(x) \cdot f_i(x) \; = \; \alpha_i$$

However, we may not have $p_t(a) \geq 0$ for all $t < 0$. We find a $\beta > 0$ such that for $t \in [-\beta, 0]$

$$p_t(x) \ \geq \ 0 \quad \Leftrightarrow \quad t(p(x) - p^*(x)) \ \geq \ -p^*(x)$$

Note that above inequality clearly holds if $p(x) - p^*(x) < 0$. Now choose β such that

$$\beta = \min_{x:p(x)-p^*(x)>0} \left\{ \frac{p^*(x)}{p(x) - p^*(x)} \right\}$$

Notice that $\beta > 0$ since $\mathrm{Supp}(P^*) \supseteq \cup_{P \in \mathcal{L}} \mathrm{Supp}(P)$.
The above implies that $\frac{d}{dt} D(P_t||Q)|_{t=0} = 0$ by the minimality of P^*, which in turn implies the equality $D(P||Q) = D(P||P^*) + D(P^*||Q)$. ■

The above can also be used to show that the I-projection onto $\mathcal{L}$ is of a special form. To describe this, we define the following family of distributions.

Definition 2.7. *Let Q be a given distribution. For any given functions $g_1, g_2, ..., g_k$ on $\mathcal{X}$, the set*

$$\mathcal{E}_Q(g_1, \ldots, g_k) \ := \ \left\{ P \ \mid \ \exists \lambda_1, \ldots, \lambda_k \in \mathbb{R} \ \forall x \in \mathcal{X}, \quad p(x) = c \cdot q(x) \cdot \exp\left(\sum_{i=1}^{k} \lambda_i g_i(x) \right) \right\}$$

is called an exponential family *of distributions.*

We will show that $P^* = \mathsf{Proj}_{\mathcal{L}}(Q) \in \mathcal{E}_Q(f_1, ..., f_k)$. We prove this for a linear family defined by a single constraint. The proof for families with multiple constraints is identical. Let $f : \mathcal{X} \to \mathbb{R}$ and let $\mathcal{L}$ be defined as

$$\mathcal{L} = \left\{ P \ \mid \ \sum_{x \in \mathcal{X}} p(x) \cdot f(x) \ = \ \mathop{\mathbb{E}}_{x \sim P} [f(x)] \ = \ \alpha \right\}$$

The projection P^* is the optimal solution to the convex program

$$\begin{array}{rrcl} \text{minimize} & D(P \| Q) & & \\ \text{subject to} & \sum_{x \in \mathcal{X}} p(x) \cdot f(x) & = & \alpha \\ & \sum_{x \in \mathcal{X}} p(x) & = & 1 \\ & p(x) & \geq & 0 \quad \forall x \in \mathcal{X}. \end{array}$$

For $\lambda_0, \lambda_1 \in \mathbb{R}$, we write the Lagrangian as

$$\Lambda(P; \lambda_0, \lambda_1) \ = \ D(P \| Q) + \lambda_0 \cdot \left(\sum_x p(x) - 1 \right) + \lambda_1 \cdot \left(\sum_x p(x) \cdot f(x) - \alpha \right).$$

The problem above can be written in terms of the Lagrangian as

$$\inf_{P \geq 0} \sup_{\lambda_0, \lambda_1 \in \mathbb{R}} \Lambda(P; \lambda_0, \lambda_1) .$$

From Lemma 2.6, we know that P^* lies in the relative interior of the polytope defining $\mathcal{L}$. Then, strong duality holds for the above program and we can write

$$\inf_{P \geq 0} \sup_{\lambda_0, \lambda_1 \in \mathbb{R}} \Lambda(P; \lambda_0, \lambda_1) \; = \; \sup_{\lambda_0, \lambda_1 \in \mathbb{R}} \inf_{P \geq 0} \Lambda(P; \lambda_0, \lambda_1) .$$

We now characterize the form of the optimal solution by considering the second (dual) program. For a given value of λ_0, λ_1, we can find the optimal solution P^* by setting the derivative of $\Lambda(P; \lambda_0, \lambda_1)$ with respect to $p(x)$ to zero, for every $x \in \mathcal{X}$. This gives

$$\log \left(\frac{p^*(x)}{q(x)} \right) + \frac{1}{\ln 2} + \lambda_0 + \lambda_1 \cdot f(x) \; = \; 0$$

Thus, we have for all $a \in \mathcal{X}$

$$p^*(x) \; = \; q(x) \cdot 2^{-\lambda_0 - \lambda_1 \cdot f(x)} .$$

The proof for linear families defined by multiple constraints is identical. The above also shows that maximum entropy distributions subject to linear constraints, always belong to an exponential family. Exponential families have many interesting applications, and more material on these can be found in the survey by Jordan and Wainwright [WJ08]. A good reference for looking up the convex-duality based arguments above, is Chapter 5 of the excellent book by Boyd and Vandenberghe [BV04].

References

[BV04] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004. 9

[WJ08] Martin J Wainwright and Michael Irwin Jordan, *Graphical models, exponential families, and variational inference*, Now Publishers Inc, 2008. 9