GML 23 - Lecture 12 (Interactive Session): Minimax lower bound for semi-supervised learning

We will the following tight lower bound on the estimation error:

Theorem 1 (SSL Minimax Rate for Estimation Error)**.** *For any $0 < s \leq 1$ the following holds when $n_u \gtrsim (1/s)^2$, $n_l \gtrsim \frac{\log n_u}{s^2}$ and $d \geq 2$:*

$$\inf_{\mathcal{A}_{SSL}} \sup_{\|\theta^*\|=s} \mathbb{E}[\mathcal{R}_{estim}(\mathcal{A}_{SSL}(\mathcal{D}_l, \mathcal{D}_u), \theta^*)] \gtrsim \min\left\{s, \sqrt{\frac{d}{n_l + s^2 n_u}}\right\}.$$

We will prove Theorem 1 via Fano's method. The proof is divided into the following exercises:

Question 1: Fano's method for GMMs

Consider an arbitrary set of predictors $\mathcal{M} = \{\theta_i\}_{i=0}^M$. Prove the following:

$$\begin{aligned}&\inf_{\mathcal{A}_{\mathrm{SSL}}} \sup_{\|\theta^*\|=s} \mathbb{E}_{\mathcal{D}_l, \mathcal{D}_u}[\mathcal{R}_{\mathrm{estim}}(\mathcal{A}_{\mathrm{SSL}}(\mathcal{D}_l, \mathcal{D}_u), \theta^*)] \geq \\ &\frac{1}{2} \min_{i,j\in[M], i\neq j} \|\theta_i - \theta_j\| \left(1 - \frac{1 + n_l \max_{i\in[M]} D(P_{XY}^{\theta_i} \| P_{XY}^{\theta_0}) + n_u \max_{i\in[M]} D(P_X^{\theta_i} \| P_X^{\theta_0})}{\log M}\right),\end{aligned} \tag{1}$$

where $D(\cdot\|\cdot)$ denotes the KL divergence.

Hint: first, prove that given a collection of distributions $\{P_1, ..., P_M\}$ and their mixture distribution $\overline{Q} = \frac{1}{M}\sum_{i=1}^M P_i$, it holds that

$$\frac{1}{M}\sum_{i=1}^M D(P_i \| \overline{Q}) \leq \frac{1}{M}\sum_{i=1}^M D(P_i \| Q)$$

for any other distribution Q (Exercise 15.11 in MW).

Solution

We first prove the hint. Assuming existence of all densities, we write for any Q:

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^M D(P_i\|Q) &= \frac{1}{M}\sum_{i=1}^M \int p_i(x) \log\left(\frac{p_i(x)}{q(x)}\right) dx = \frac{1}{M}\sum_{i=1}^M \int p_i(x)\log\left(\frac{1}{q(x)}\right) dx + const \\ &= \int \overline{q}(x) \log\left(\frac{\overline{q}(x)}{q(x)}\right) dx + const = D(\overline{Q}\|Q) + const,\end{aligned}$$

where all *const* terms do not depend on the distribution Q. Thus, the original expression is minimized by the mixture distribution $Q = \overline{Q}$ and the statement follows.

To prove (1), we first note that our set $\mathcal{M}$ is a 2δ-packing with $\delta = \frac{1}{2}\min_{i,j\in[M], i\neq j} \|\theta_i - \theta_j\|_2$. Combining the estimation vs. testing lemma (MW Prop 15.1) and Fano's method, we obtain

$$\begin{aligned}&\inf_{\mathcal{A}_{\mathrm{SSL}}} \sup_{\|\theta^*\|=s} \mathbb{E}_{\mathcal{D}_l, \mathcal{D}_u}[\mathcal{R}_{\mathrm{estim}}(\mathcal{A}_{\mathrm{SSL}}(\mathcal{D}_l, \mathcal{D}_u), \theta^*)] \geq \\ &\frac{1}{2}\min_{i,j\in[M], i\neq j} \|\theta_i - \theta_j\| \left(1 - \frac{I(\mathcal{D}, J) + \log 2}{\log M}\right).\end{aligned}$$

For the mutual information, it holds that

$$I(\mathcal{D}, J) = \frac{1}{M} \sum_{i=1}^{M} D(P_{\theta_i} \| \overline{Q}),$$

see also MW Eq. 15.30. We thus have

$$I(\mathcal{D}, J) = \frac{1}{M} \sum_{i=1}^{M} D(P_{\theta_i} \| \overline{Q}) \leq \frac{1}{M} \sum_{i=1}^{M} D(P_{\theta_i} \| P_{\theta_0}),$$

where we have used the hint with the choice $Q = P_{\theta_0}$. We now recall that P_{θ_i} corresponds to the product distribution of n_l labeled and n_u unlabeled samples, i.e. $P_{\theta_i} = (P_{XY}^{\theta_i})^{n_l} \times (P_X^{\theta_i})^{n_u}$. Using the decoupling property of the KL divergence for product distributions, we thus obtain

$$\frac{1}{M} \sum_{i=1}^{M} D(P_{\theta_i} \| P_{\theta_0}) = \frac{1}{M} \sum_{i=1}^{M} (n_l D(P_{XY}^{\theta_i} \| P_{XY}^{\theta_0}) + n_u D(P_X^{\theta_i} \| P_X^{\theta_0})).$$

We now upper bound both averages by the maximum to obtain

$$I(\mathcal{D}, J) \leq \frac{1}{M} \sum_{i=1}^{M} D(P_{\theta_i} \| P_{\theta_0}) \leq n_l \max_{i \in [M]} D(P_{XY}^{\theta_i} \| P_{XY}^{\theta_0}) + n_u \max_{i \in [M]} D(P_X^{\theta_i} \| P_X^{\theta_0}).$$

Inserting this in Fano's bound and additionally bounding $\log 2 < 1$ yields the claim.

Question 2: Upper bounds on KL divergence for GMMs

Assume that you are given a packing $\{\theta_i\}_{i=0}^M$ which is constructed as follows: given positive absolute constants c_0 and C_0, we take a c_0-packing $\tilde{\mathcal{M}} = \{\psi_1, ..., \psi_M\}$ on the unit sphere S^{d-2} such that $|\tilde{\mathcal{M}}| \geq e^{C_0 d}$. For an absolute constant $\alpha \in [0, 1]$, we now construct the following packing:

$$\mathcal{M} = \left\{ \theta_i = s \begin{bmatrix} \sqrt{1-\alpha^2} \\ \alpha \psi_i \end{bmatrix}, \quad \psi_i \in \tilde{\mathcal{M}} \right\},$$

and define $\theta_0 = \begin{bmatrix} s, 0, ..., 0 \end{bmatrix}$.

1) Visualize the given packing and study its properties. Where are θ_0 and θ_i located? What is the distance between different elements of the packing? Is there an intuition for this particular choice? Discuss with your partner why this choice of a packing is better for use in (1) as compared to, for instance, a uniform packing on the sphere S^{d-1}.

2) Compute the KL divergence between two GMMs with identity covariance matrices, i.e. show that

$$D(P_{XY}^{\theta_i} \| P_{XY}^{\theta_0}) = \frac{1}{2} \|\theta_i - \theta_0\|_2^2 \leq \alpha^2 s^2, \text{ for all } i \in [M]. \tag{2}$$

Solution

1) The given packing maximizes the tradeoff between the max and the min in the lower bound. $(\theta_i)_i$ are points on a "circle" around $\theta_0 = (s, 0, ..., 0)$ which is located on the sphere S^{d-1}. Since we have chosen $(\theta_i)_i$ to be the largest possible (up to constants) packing of S^{d-2}, $(\theta_i)_i$ are the maximum amount of points with distance at least c_0 from each other while *simultaneously* being all relatively close to θ_0 due to the geometry of the sphere. In other words, this construction puts as many points as possible close to some point θ_0 (chosen here to be $(s, 0, ..., 0)$ for simplicity), while maintaining as large as possible distance between the points themselves (packing). This allows us to optimize the tradeoff between the term $\frac{1}{2} \min_{i,j \in [M], i \neq j} \|\theta_i - \theta_j\|_2$ (which we want to maximize) and $D(P_{\theta_i} \| P_{\theta_0})$ (which we want to minimize).

2) We compute

$$D(P_{XY}^{\theta_i} \| P_{XY}^{\theta_0}) = \int p_{XY}^{\theta_i}(x, y) \log \frac{p_{XY}^{\theta_i}(x, y)}{p_{XY}^{\theta_0}(x, y)} dx dy =$$
$$\frac{1}{2} \int p^{\theta_i}(x|Y=1) \log \frac{p^{\theta_i}(x|Y=1)}{p^{\theta_0}(x|Y=1)} dx + \frac{1}{2} \int p^{\theta_i}(x|Y=-1) \log \frac{p^{\theta_i}(x|Y=-1)}{p^{\theta_0}(x|Y=-1)} dx.$$

Recalling that $p^{\theta_i}(x|Y=1) = \frac{1}{(2\pi)^{d/2}}\exp(-\frac{1}{2}\|x-\theta_i\|_2^2)$ and $p^{\theta_i}(x|Y=-1) = \frac{1}{(2\pi)^{d/2}}\exp(-\frac{1}{2}\|x+\theta_i\|_2^2)$, we obtain

$$D(P_{XY}^{\theta_i}\|P_{XY}^{\theta_0}) = \frac{1}{2}D(p^{\theta_i}(x|Y=1)\|p^{\theta_0}(x|Y=1)) + \frac{1}{2}D(p^{\theta_i}(x|Y=-1)\|p^{\theta_0}(x|Y=-1)) = \frac{1}{2}\|\theta_i-\theta_0\|_2^2,$$

where we have used that the KL divergence between two isotropic Gaussians with means μ_1 and μ_2 is equal to $\frac{1}{2}\|\mu_1-\mu_2\|_2^2$. Now due to the construction of the packing it holds for any i that

$$\|\theta_i-\theta_0\|_2^2 = (s-s\sqrt{1-\alpha^2})^2 + s^2\alpha^2\|\psi_i\|_2^2 = 2s^2 - 2s^2\sqrt{1-\alpha^2} \le 2s^2\alpha^2,$$

where we have used that for $0 \le \alpha \le 1$ it holds that $1-\sqrt{1-\alpha^2} \le \alpha^2$. The claim then follows.

Question 3: Proof of Theorem 1

Assume that additionally to (2), we have proven the following upper bound for the KL divergence between marginal distributions:

$$D(P_X^{\theta_i}\|P_X^{\theta_0}) \le C\|\frac{1}{s}\theta_i - \frac{1}{s}\theta_0\|^2 \le 2C\alpha^2 s^4. \tag{3}$$

Utilizing these two results as well as Question 1, prove Theorem 1. *(You might need to optimize over one of the constants.)*

Solution

Given the expression

$$\frac{1}{2}\min_{i,j\in[M],i\neq j}\|\theta_i-\theta_j\|_2\left(1-\frac{1+n_l\max_{i\in[M]}D(P_{XY}^{\theta_i}\|P_{XY}^{\theta_0})+n_u\max_{i\in[M]}D(P_X^{\theta_i}\|P_X^{\theta_0})}{\log M}\right)$$

from Question 1, we have for our packing $\frac{1}{2}\min_{i,j\in[M],i\neq j}\|\theta_i-\theta_j\|_2 \ge \frac{1}{2}c_0 s\alpha$ (follows direcly from θ_i definition). Furthermore, we have $\log M \le C_0 d$ (as $|\mathcal{M}| \ge e^{C_0 d}$). Additionally, we insert both bounds for the KL divergence from Question 2 and 3 to obtain

$$\inf_{\mathcal{A}_{\mathrm{SSL}}}\sup_{\|\theta^*\|=s}\mathbb{E}_{\mathcal{D}_l,\mathcal{D}_u}[\mathcal{R}_{\mathrm{estim}}(\mathcal{A}_{\mathrm{SSL}}(\mathcal{D}_l,\mathcal{D}_u),\theta^*)] \ge \frac{1}{2}c_0 s\alpha\left(1-\frac{1+n_l\alpha^2 s^2+2Cn_u\alpha^2 s^4}{C_0 d}\right).$$

The RHS is a cubic polynomial in α which we now want to maximize w.r.t. α. We obtain the maximum $\sqrt{\frac{C_0 d-1}{3s^2 n_l+3C_1 s^4 n_u}}$. Since $0 \le \alpha \le 1$ and the maximizing value can be larger than 1, we set α to be $\alpha = \min\left\{1, \sqrt{\frac{C_0 d-1}{3s^2 n_l+3C_1 s^4 n_u}}\right\}$. Inserting both values of α in the bound and neglecting multiplicative constants, we obtain the final result

$$\inf_{\mathcal{A}_{\mathrm{SSL}}}\sup_{\|\theta^*\|=s}\mathbb{E}[\mathcal{R}_{\mathrm{estim}}(\mathcal{A}_{SSL}(\mathcal{D}_l,\mathcal{D}_u),\theta^*)] \gtrsim \min\left\{s, \sqrt{\frac{d}{n_l+s^2 n_u}}\right\}.$$