HDS36 - Le Cam's convex hull and Fano's method

Yangjianchen Xu

Department of Biostatistics
University of North Carolina at Chapel Hill

02/04/2022

Overview

- Le Cam's convex hull method
- Fano's method
    - Bounds based on local packings
    - Local packings with Gaussian entropy bounds
    - Yang–Barron version of Fano's method

Le Cam's convex hull method

Consider two subsets $\mathcal{P}_0$ and $\mathcal{P}_1$ of $\mathcal{P}$ that are 2δ-separated, in the sense that

$$\rho\left(\theta\left(\mathbb{P}_0\right), \theta\left(\mathbb{P}_1\right)\right) \geq 2\delta \quad \text{for all } \mathbb{P}_0 \in \mathcal{P}_0 \text{ and } \mathbb{P}_1 \in \mathcal{P}_1.$$

Lemma (15.9)

For any 2δ-separated classes of distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ contained within $\mathcal{P}$, any estimator $\widehat{\theta}$ has worst-case risk at least

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] \geq \frac{\delta}{2} \sup_{\substack{\mathbb{P}_0 \in \mathrm{conv}(\mathcal{P}_0) \\ \mathbb{P}_1 \in \mathrm{conv}(\mathcal{P}_1)}} \left\{1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}}\right\}. \tag{1}$$

Proof of Lemma 15.9

Proof. For any estimator $\widehat{\theta}$, let us define the random variables

$$V_j(\widehat{\theta}) = \frac{1}{2\delta} \inf_{\mathbb{P}_j \in \mathcal{P}_j} \rho(\widehat{\theta}, \theta(\mathbb{P}_j)), \qquad \text{for } j = 0, 1.$$

We then have

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] &\geq \frac{1}{2} \left\{ \mathbb{E}_{\mathbb{P}_0} \left[\rho\left(\widehat{\theta}, \theta(\mathbb{P}_0)\right) \right] + \mathbb{E}_{\mathbb{P}_1} \left[\rho\left(\widehat{\theta}, \theta(\mathbb{P}_1)\right) \right] \right\} \\ &\geq \delta \left\{ \mathbb{E}_{\mathbb{P}_0} \left[V_0(\widehat{\theta}) \right] + \mathbb{E}_{\mathbb{P}_1} \left[V_1(\widehat{\theta}) \right] \right\}. \end{aligned}$$

Since the right-hand side is linear in $\mathbb{P}_0$ and $\mathbb{P}_1$, we can take suprema over the convex hulls, and thus obtain the lower bound

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] \geq \delta \sup_{\substack{\mathbb{P}_0 \in \mathrm{conv}(\mathcal{P}_0) \\ \mathbb{P}_1 \in \mathrm{conv}(\mathcal{P}_1)}} \left\{ \mathbb{E}_{\mathbb{P}_0} \left[V_0(\widehat{\theta}) \right] + \mathbb{E}_{\mathbb{P}_1} \left[V_1(\widehat{\theta}) \right] \right\}.$$

Proof of Lemma 15.9, cont.

By the triangle inequality, we have

$$\rho\left(\widehat{\theta}, \theta\left(\mathbb{P}_0\right)\right) + \rho\left(\widehat{\theta}, \theta\left(\mathbb{P}_1\right)\right) \geq \rho\left(\theta\left(\mathbb{P}_0\right), \theta\left(\mathbb{P}_1\right)\right) \geq 2\delta.$$

Taking infima over $\mathbb{P}_j \in \mathcal{P}_j$ for each $j = 0, 1$, we obtain

$$\inf_{\mathbb{P}_0 \in \mathcal{P}_0} \rho\left(\widehat{\theta}, \theta\left(\mathbb{P}_0\right)\right) + \inf_{\mathbb{P}_1 \in \mathcal{P}_1} \rho\left(\widehat{\theta}, \theta\left(\mathbb{P}_1\right)\right) \geq 2\delta,$$

which is equivalent to $V_0(\widehat{\theta}) + V_1(\widehat{\theta}) \geq 1$. Since $V_j(\widehat{\theta}) \geq 0$ for $j = 0, 1$, the variational representation of the TV distance (see Exercise 15.1) implies that, for any $\mathbb{P}_j \in \operatorname{conv}\left(\mathcal{P}_j\right)$, we have

$$\mathbb{E}_{\mathbb{P}_0}\left[V_0(\widehat{\theta})\right] + \mathbb{E}_{\mathbb{P}_1}\left[V_1(\widehat{\theta})\right] \geq 1 - \left\|\mathbb{P}_1 - \mathbb{P}_0\right\|_{\mathrm{TV}},$$

which completes the proof.

Example 15.10: Sharpened bounds for Gaussian location family

- Setting $\theta = 2\delta$ as before, consider the two families $\mathcal{P}_0 = \{\mathbb{P}_0^n\}$ and $\mathcal{P}_1 = \left\{\mathbb{P}_\theta^n, \mathbb{P}_{-\theta}^n\right\}$.
- The mixture distribution $\overline{\mathbb{P}} := \frac{1}{2}\mathbb{P}_\theta^n + \frac{1}{2}\mathbb{P}_{-\theta}^n$ belongs to $\operatorname{conv}(\mathcal{P}_1)$.
- From the second-moment bound explored in Exercise 15.10(c), we have

$$\|\mathbb{P} - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left\{e^{\frac{1}{2}\left(\frac{\sqrt{n}\theta}{\sigma}\right)^4} - 1\right\} = \frac{1}{4}\left\{e^{\frac{1}{2}\left(\frac{2\sqrt{n}\delta}{\sigma}\right)^4} - 1\right\}.$$

- Setting $\delta = \frac{\sigma t}{2\sqrt{n}}$ for some parameter $t > 0$ to be chosen, the convex hull Le Cam bound (1) yields

$$\min_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\hat{\theta} - \theta|] \geq \frac{\sigma}{4\sqrt{n}} \sup_{t>0}\left\{t\left(1 - \frac{1}{2}\sqrt{e^{\frac{1}{2}t^4} - 1}\right)\right\} \geq \frac{3}{20}\frac{\sigma}{\sqrt{n}}.$$

Fano's method: recall basic set-up

- We are interested in lower bounding the probability of error in an M-ary hypothesis testing problem, based on a family of distributions $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^M}\}$.
- A sample Z is generated by choosing an index J uniformly at random from the index set $[M] := \{1, \ldots, M\}$, and then generating data according to $\mathbb{P}_{\theta^J}$.
- In this way, the observation follows the mixture distribution $\mathbb{Q}_Z = \overline{\mathbb{Q}} := \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}$.
- Goal: to identify the index J of the probability distribution from which a given sample has been drawn.

Kullback–Leibler divergence and mutual information

- Difficulty: the amount of dependence between the observation Z and the unknown random index J.
- Question: How to measure the amount of dependence between a pair of random variables?
- A natural way is by computing some type of divergence measure between the joint distribution and the product of marginals.
- The mutual information between the random variables (Z, J) is defined in exactly this way:

$$I(Z, J) := D\left(\mathbb{Q}_{Z,J} \| \mathbb{Q}_Z \mathbb{Q}_J\right),$$

which uses the Kullback-Leibler divergence as the underlying measure of distance

Kullback–Leibler divergence and mutual information

- Given our set-up and the definition of the KL divergence, the mutual information can be written as

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \overline{\mathbb{Q}}\right), \tag{2}$$

  corresponding to the mean KL divergence between component distribution $\mathbb{P}_{\theta^j}$ and the mixture distribution $\bar{\mathbb{Q}} = \mathbb{Q}_J$, averaged over the choice of index j.
- Consequently, the mutual information is small if the distributions $\mathbb{P}_{\theta^j}$ are hard to distinguish from the mixture distribution $\overline{\mathbb{Q}}$ on average.

Fano lower bound on minimax risk

The Fano method is based on the following lower bound:

$$\mathbb{P}[\psi(Z) \neq J] \geq 1 - \frac{I(Z; J) + \log 2}{\log M}.$$

When combined with the reduction from estimation to testing given in Proposition 15.1, we obtain the following lower bound on the minimax error:

Proposition (15.12)

Let $\{\theta^1, \ldots, \theta^M\}$ be a 2δ-separated set in the ρ semi-metric on $\Theta(\mathcal{P})$, and suppose that J is uniformly distributed over the index set $\{1, \ldots, M\}$, and $(Z \mid J = j) \sim \mathbb{P}_{\theta^j}$. Then for any increasing function $\Phi : [0, \infty) \to [0, \infty)$, the minimax risk is lower bounded as

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left\{ 1 - \frac{I(Z; J) + \log 2}{\log M} \right\}, \tag{3}$$

where $I(Z; J)$ is the mutual information between Z and J.

Fano lower bound on minimax risk

- As we shrink δ, then the 2δ-separation criterion becomes milder, so that the cardinality $M \equiv M(2\delta)$ in the denominator increases.
- At the same time, in a generic setting, the mutual information $I(Z; J)$ will decrease, since the random index $J \in [M(2\delta)]$ can take on a larger number of potential values.
- By decreasing δ sufficiently, we may thereby ensure that

$$\frac{I(Z;J) + \log 2}{\log M} \leq \frac{1}{2} \tag{4}$$

so that the lower bound (3) implies that $\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$.

Fano lower bound on minimax risk

In order to derive lower bounds in this way, there remain two technical and possibly challenging steps:

1. To specify 2δ-separated sets with large cardinality $M(2\delta)$.
2. To compute or upper bound the mutual information $I(Z; J)$.

Bounds based on local packings

- Using this convexity and the mixture representation (2), we find that

$$I(Z;J) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right). \tag{5}$$

- Suppose that we can construct a 2δ-separated set contained within Ω such that, for some quantity c, the Kullback-Leibler divergences satisfy the uniform upper bound

$$\sqrt{D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right)} \leq c\sqrt{n}\delta \quad \text{for all } j \neq k. \tag{6}$$

- The bound (5) then implies that $I(Z;J) \leq c^2 n\delta^2$, and hence the bound (4) will hold as long as

$$\log M(2\delta) \geq 2\left\{c^2 n\delta^2 + \log 2\right\}. \tag{7}$$

Example 15.14: Minimax risks for linear regression

- The standard linear regression model $y = \mathbf{X}\theta^* + w$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a fixed design matrix, and the vector $w \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_n\right)$ is observation noise.
- Goal: to obtain lower bounds on the minimax risk in the prediction (semi-)norm $\rho_{\mathbf{X}}(\widehat{\theta}, \theta^*) := \frac{\left\|\mathbf{X}(\widehat{\theta} - \theta^*)\right\|_2}{\sqrt{n}}$.
- For a tolerance $\delta > 0$ to be chosen, consider the set

$$\left\{\gamma \in \operatorname{range}(\mathbf{X}) \mid \|\gamma\|_2 \leq 4\delta\sqrt{n}\right\},$$

  and let $\left\{\gamma^1, \ldots, \gamma^M\right\}$ be a $2\delta\sqrt{n}$-packing in the ℓ_2-norm.
- Since this set sits in a space of dimension $r = \operatorname{rank}(\mathbf{X})$, Lemma 5.7 implies that we can find such a packing with $\log M \geq r \log 2$ elements.

Example 15.14: Minimax risks for linear regression, cont.

- We thus have a collection of vectors of the form $\gamma^j = \mathbf{X}\theta^j$ for some $\theta^j \in \mathbb{R}^d$, and such that

$$\frac{\|\mathbf{X}\theta^j\|_2}{\sqrt{n}} \leq 4\delta, \text{ for each } j \in [M],$$

$$2\delta \leq \frac{\|\mathbf{X}(\theta^j - \theta^k)\|_2}{\sqrt{n}} \leq 8\delta \text{ for each } j \neq k \in [M] \times [M].$$

- Under $\mathbb{P}_{\theta^j}$, the observed vector $y \in \mathbb{R}^n \sim \mathcal{N}\left(\mathbf{X}\theta^j, \sigma^2 \mathbf{I}_n\right)$. By Exercise 15.13,

$$D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right) = \frac{1}{2\sigma^2}\|\mathbf{X}(\theta^j - \theta^k)\|_2^2 \leq \frac{32n\delta^2}{\sigma^2}.$$

- Consequently, for r sufficiently large, the lower bound (7) can be satisfied by setting $\delta^2 = \frac{\sigma^2}{64}\frac{r}{n}$, and we conclude that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\frac{1}{n}\|\mathbf{X}(\widehat{\theta} - \theta)\|_2^2\right] \geq \frac{\sigma^2}{128}\frac{\text{rank}(\mathbf{X})}{n}.$$

Example 15.16: Minimax risks for sparse linear regression

- The high-dimensional linear regression model $y = \mathbf{X}\theta^* + w$, where the regression vector θ^* is known *a* priori to be sparse, say with at most $s < d$ non-zero coefficients.
- It is then natural to consider the minimax risk over the set

$$\mathbb{S}^d(s) := \mathbb{B}_0^d(s) \cap \mathbb{B}_2(1) = \left\{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\right\}$$

  of s-sparse vectors within the Euclidean unit ball.
- From our earlier results in Chapter 5, there exists a $1/2$-packing of this set with log cardinality at least $\log M \geq \frac{s}{2} \log \frac{d-s}{s}$.
- We follow the same rescaling procedure as in Example 15.14 to form a δ-packing such that $\left\|\theta^j - \theta^k\right\|_2 \leq 4\delta$ for all pairs of vectors in our packing set.

Example 15.16: Minimax risks for sparse linear regression, cont.

- Since the vector $\theta^j - \theta^k$ is at most $2s$-sparse, we have

$$\sqrt{D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right)} = \frac{1}{\sqrt{2}\sigma} \|\mathbf{X}(\theta^j - \theta^k)\|_2 \leq \frac{\gamma_{2s}}{\sqrt{2}\sigma} 4\delta\sqrt{n}$$

where $\gamma_{2s} := \max_{|T|=2s} \sigma_{\max}\left(\mathbf{X}_T\right) / \sqrt{n}$.

- Putting together the pieces, we see that the minimax risk is lower bounded by any $\delta > 0$ for which

$$\frac{s}{2} \log \frac{d-s}{s} \geq 128 \frac{\gamma_{2s}^2}{\sigma^2} n\delta^2 + 2\log 2.$$

- As long as $s \leq d/2$ and $s \geq 10$, the choice $\delta^2 = \frac{\sigma^2}{400\gamma_{2s}^2} s \log \frac{d-s}{s}$ suffices. We conclude that in the range $10 \leq s \leq d/2$, the minimax risk is lower bounded as

$$\mathfrak{M}\left(\mathbb{S}^d(s); \|\cdot\|_2\right) \gtrsim \frac{\sigma^2}{\gamma_{2s}^2} \frac{s \log \frac{ed}{s}}{n}.$$

Local packings with Gaussian entropy bounds

Lemma (15.17)

Suppose J is uniformly distributed over $[M] = \{1, \ldots, M\}$ and that Z conditioned on $J = j$ has a Gaussian distribution with covariance Σ^j. Then the mutual information is upper bounded as

$$I(Z; J) \leq \frac{1}{2} \left\{ \log \det \operatorname{cov}(Z) - \frac{1}{M} \sum_{j=1}^{M} \log \det \left(\boldsymbol{\Sigma}^j \right) \right\}. \tag{8}$$

In the special case when $\boldsymbol{\Sigma}^j = \boldsymbol{\Sigma}$ for all $j \in [M]$, it takes on the simpler form

$$I(Z; J) \leq \frac{1}{2} \log \left(\frac{\det \operatorname{cov}(Z)}{\det(\Sigma)} \right). \tag{9}$$

Example 15.18: Variable selection in sparse linear regression

- Return to the model of sparse linear regression from Example 15.16.
- Goal: to lower bound the minimax risk for the problem of determining the support set $S = \{j \in \{1, 2, \ldots, d\} \mid \theta_j^* \neq 0\}$.
- The problem of interest is itself a multiway hypothesis test-namely, that of choosing from all $\binom{d}{s}$ possible subsets.
- We show that, in order to achieve a probability of error below $1/2$, any method requires a sample size of at least

$$n > \max\left\{8\frac{\log(d+s-1)}{\log(1+\frac{\theta_{\min}^2}{\sigma^2})}, 8\frac{\log\binom{d}{s}}{\log(1+s\frac{\theta_{\min}^2}{\sigma^2})}\right\}, \tag{10}$$

as long as $\min\left\{\log(d+s-1), \log\binom{d}{s}\right\} \geq 4\log 2$. $\theta_{\min} = \min_{j \in S} |\theta_j^*|$.

Example 15.18: Variable selection in sparse linear regression, cont.

- We derive lower bounds by first conditioning on a particular instantiation $\mathbf{X} = \{x_i\}_{i=1}^n$ of the design matrix, and using a form of Fano's inequality that involves the mutual information $I_{\mathbf{X}}(y; J)$.
- In particular, we have

$$\mathbb{P}\left[\psi(y, \mathbf{X}) \neq J \mid \mathbf{X} = \{x_i\}_{i=1}^n\right] \geq 1 - \frac{I_{\mathbf{X}}(y; J) + \log 2}{\log M}$$

so that by taking averages over $\mathbf{X}$, we can obtain lower bounds on $\mathbb{P}[\psi(y, \mathbf{X}) \neq J]$ that involve the quantity $\mathbb{E}_{\mathbf{X}}\left[I_{\mathbf{X}}(y; J)\right]$.

Example 15.18: Ensemble A

- Consider the class $M = \binom{d}{s}$ of all possible subsets of cardinality s.
- For the ℓ th subset S^ℓ, let $\theta^\ell \in \mathbb{R}^d$ have values $\theta_{\min}$ for all indices $j \in S^\ell$, and zeros in all other positions.
- For a fixed covariate vector $x_i \in \mathbb{R}^d$, an observed response $y_i \in \mathbb{R}$ then follows the mixture distribution $\frac{1}{M} \sum_{\ell=1}^{M} \mathbb{P}_{\theta^\ell}$, where $\mathbb{P}_{\theta'}$ is the distribution of a $\mathcal{N}\left(\left\langle x_i, \theta^\ell \right\rangle, \sigma^2\right)$ random variable.

Example 15.18: Ensemble A, cont.

- By the definition of mutual information, we have

$$
\begin{aligned}
I_{\mathbf{X}}(y; J) &= H_{\mathbf{X}}(y) - H_{\mathbf{X}}(y \mid J) \\
&\overset{\text{(i)}}{\leq} \left[\sum_{i=1}^{n} H_{\mathbf{X}}\left(y_i\right)\right] - H_{\mathbf{X}}(y \mid J) \\
&\overset{\text{(ii)}}{=} \sum_{i=1}^{n} \left\{H_{\mathbf{X}}\left(y_1\right) - H_{\mathbf{X}}\left(y_1 \mid J\right)\right\} \\
&= \sum_{i=1}^{n} I_{\mathbf{X}}\left(y_i; J\right)
\end{aligned}
$$

where step (i) follows since independent random vectors have larger entropy than dependent ones (see Exercise 15.4), and step (ii) follows since $(y_1, \ldots, y_n)$ are independent conditioned on J.

Example 15.18: Ensemble A, cont.

- Next, applying Lemma 15.17 repeatedly for each $i \in [n]$ with $Z = y_i$, conditionally on the matrix $\mathbf{X}$ of covariates, yields

$$I_{\mathbf{X}}(y; J) \leq \frac{1}{2} \sum_{i=1}^{n} \log \frac{\operatorname{var}(y_i \mid x_i)}{\sigma^2}.$$

- Now taking averages over $\mathbf{X}$ and using the fact that the pairs (y_i, x_i) are jointly i.i.d., we find that

$$\mathbb{E}_{\mathbf{X}}\left[I_{\mathbf{X}}(y; J)\right] \leq \frac{n}{2} \mathbb{E}\left[\log \frac{\operatorname{var}(y_1 \mid x_1)}{\sigma^2}\right] \leq \frac{n}{2} \log \frac{\mathbb{E}_{x_1}\left[\operatorname{var}(y_1 \mid x_1)\right]}{\sigma^2},$$

where the last inequality follows Jensen's inequality, and concavity of the logarithm.

Example 15.18: Ensemble A, cont.

- Since the random vector y_1 follows a mixture distribution with M components, we have

$$\begin{aligned}\mathbb{E}_{x_1}\left[\operatorname{var}\left(y_1 \mid x_1\right)\right] &\leq \mathbb{E}_{x_1}\left[\mathbb{E}\left[y_1^2 \mid x_1\right]\right] \\ &= \mathbb{E}_{x_1}[x_1^{\mathrm{T}}\{\frac{1}{M}\sum_{j=1}^{M}\theta^j \otimes \theta^j\}x_1 + \sigma^2] \\ &= \operatorname{trace}(\frac{1}{M}\sum_{j=1}^{M}(\theta^j \otimes \theta^j)) + \sigma^2.\end{aligned}$$

- Now each index $j \in \{1, 2, \ldots, d\}$ appears in $\binom{d-1}{s-1}$ of the total number of subsets $M = \binom{d}{s}$, so that

$$\operatorname{trace}(\frac{1}{M}\sum_{j=1}^{M}\theta^j \otimes \theta^j) = d\frac{\binom{d-1}{s-1}}{\binom{d}{s}}\theta_{\min}^2 = s\theta_{\min}^2.$$

Example 15.18: Ensemble A, cont.

- Putting together the pieces, we conclude that

$$\mathbb{E}_{\mathbf{X}}\left[I_{\mathbf{X}}(y;J)\right] \leq \frac{n}{2}\log\left(1+\frac{s\theta_{\min}^2}{\sigma^2}\right),$$

- The Fano lower bound implies that

$$\mathbb{P}[\psi(y,\mathbf{X}) \neq J] \geq 1 - \frac{\frac{n}{2}\log\left(1+\frac{s\theta_{\min}^2}{\sigma^2}\right)+\log 2}{\log\binom{d}{s}},$$

from which the first lower bound in equation (10) follows as long as $\log\binom{d}{s} \geq 4\log 2$, as assumed.

Example 15.18: Ensemble B

- Let $\bar{\theta} \in \mathbb{R}^d$ be a vector with $\theta_{\min}$ in its first $s-1$ coordinates, and zero in all remaining $d-s+1$ coordinates.
- Define $\theta^j := \bar{\theta} + \theta_{\min} e_j$ for $j = s, \ldots, d$.
- By a straightforward calculation, we have $\mathbb{E}[Y \mid x] = \langle x, \gamma \rangle$, where $\gamma := \bar{\theta} + \frac{1}{M}\theta_{\min} e_{s \to d}$, and the vector $e_{s \to d} \in \mathbb{R}^d$ has ones in positions s through d, and zeros elsewhere.
- By the same argument as for ensemble A, it suffices to upper bound the quantity $\mathbb{E}_{x_1}[\operatorname{var}(y_1 \mid x_1)]$. Using the definition of our ensemble, we have

$$\mathbb{E}_{x_1}[\operatorname{var}(y_1 \mid x_1)] = \sigma^2 + \operatorname{trace}\left\{ \frac{1}{M} \sum_{j=1}^{M} \left(\theta^j \otimes \theta^j - \gamma \otimes \gamma\right) \right\} \leq \sigma^2 + \theta_{\min}^2.$$

Yang–Barron version of Fano's method

Lemma (15.21 (Yang-Barron method))

Let $N_{\mathrm{KL}}(\epsilon; \mathcal{P})$ denote the ϵ-covering number of $\mathcal{P}$ in the square-root KL divergence. Then the mutual information is upper bounded as

$$I(Z;J) \leq \inf_{\epsilon>0} \left\{\epsilon^2 + \log N_{\mathrm{KL}}(\epsilon; \mathcal{P})\right\}. \tag{11}$$

Proof. We observe that for any distribution $\mathbb{Q}$, the mutual information is upper bounded by

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \overline{\mathbb{Q}}\right) \overset{\text{(i)}}{\leq} \frac{1}{M}\sum_{j=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \mathbb{Q}\right) \leq \max_{j=1,\ldots,M} D\left(\mathbb{P}_{\theta^j} \| \mathbb{Q}\right), \tag{12}$$

where inequality (i) uses the fact that the mixture distribution $\bar{\mathbb{Q}} := \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_{\theta^j}$ minimizes the average Kullback-Leibler divergence over the family $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^m}\}$ (Exercise 15.11).

Proof of Lemma 15.21

Since the upper bound (12) holds for any distribution $\mathbb{Q}$, we are free to choose it: in particular, we let $\{\gamma^1, \ldots, \gamma^N\}$ be an ϵ-covering of Ω in the square-root KL pseudo-distance, and then set $\mathbb{Q} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{P}_{\gamma^k}$. By construction, for each θ^j with $j \in [M]$, we can find some γ^k such that $D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\gamma^k}\right) \leq \epsilon^2$. Therefore, we have

$$
\begin{aligned}
D\left(\mathbb{P}_{\theta^j} \| \mathbb{Q}\right) &= \mathbb{E}_{\theta^j}\left[\log \frac{d\mathbb{P}_{\theta^j}}{\frac{1}{N}\sum_{\ell=1}^{N} d\mathbb{P}_{\gamma^k}}\right] \\
&\leq \mathbb{E}_{\theta^j}\left[\log \frac{d\mathbb{P}_{\theta_j}}{\frac{1}{N} d\mathbb{P}_{\gamma^k}}\right] \\
&= D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\gamma^k}\right) + \log N \\
&\leq \epsilon^2 + \log N.
\end{aligned}
$$

Since this bound holds for any choice of $j \in [M]$ and any choice of $\epsilon > 0$, the claim (11) follows.

Lemma 15.21 allows us to prove a minimax lower bound of the order δ as long as the pair $(\delta, \epsilon) \in \mathbb{R}_+^2$ are chosen such that

$$\log M(\delta; \rho, \Omega) \geq 2 \left\{\epsilon^2 + \log N_{\mathrm{KL}}(\epsilon; \mathcal{P}) + \log 2\right\}.$$

Finding such a pair can be accomplished via a two-step procedure:

(A) First, choose $\epsilon_n > 0$ such that

$$\epsilon_n^2 \geq \log N_{\mathrm{KL}}\left(\epsilon_n; \mathcal{P}\right). \tag{13}$$

(B) Second, choose the largest $\delta_n > 0$ that satisfies the lower bound

$$\log M\left(\delta_n; \rho, \Omega\right) \geq 4\epsilon_n^2 + 2\log 2. \tag{14}$$

Example 15.23: Minimax risks for generalized Sobolev families

- Recall that the standard regression model is based on i.i.d. observations of the form

$$y_i = f^*(x_i) + \sigma w_i, \qquad \text{for } i = 1, 2, \ldots, n,$$

where $w_i \sim \mathcal{N}(0, 1)$.

- Assuming that the design points $\{x_i\}_{i=1}^n$ are drawn in an i.i.d. fashion from some distribution $\mathbb{P}$, let us derive lower bounds in the $L^2(\mathbb{P})$-norm:

$$\|\widehat{f} - f^*\|_2^2 = \int_X [\widehat{f}(x) - f^*(x)]^2 \mathbb{P}(dx).$$

Example 15.23: Minimax risks for generalized Sobolev families, cont.

- For a smoothness parameter $\alpha > 1/2$, consider the ellipsoid $\ell^2(\mathbb{N})$ given by $\mathcal{E}_\alpha = \{(\theta_j)_{j=1}^{\infty} \mid \sum_{j=1}^{\infty} j^{2\alpha}\theta_j^2 \leq 1\}$.
- Given an orthonormal sequence $(\phi_j)_{j=1}^{\infty}$ in $L^2(\mathbb{P})$, we can then define the function class $\mathscr{F}_\alpha := \{f = \sum_{j=1}^{\infty} \theta_j \phi_j \mid (\theta_j)_{j=1}^{\infty} \in \mathcal{E}_\alpha\}$.
- For any such function class, we claim that the minimax risk in squared $L^2(\mathbb{P})$-norm is lower bounded as

$$\inf_{\widehat{f}} \sup_{f \in \mathscr{F}_a} \mathbb{E}\left[\|\widehat{f} - f\|_2^2\right] \gtrsim \min\left\{1, \left(\frac{\sigma^2}{n}\right)^{\frac{2a}{2a+1}}\right\}.$$

Example 15.23: Minimax risks for generalized Sobolev families, cont.

- Consider a function of the form $f = \sum_{j=1}^{\infty} \theta_j \phi_j$ for some $\theta \in \ell^2(\mathbb{N})$, and observe that by the orthonormality of $(\phi_j)_{j=1}^{\infty}$, Parseval's theorem implies that $\|f\|_2^2 = \sum_{j=1}^{\infty} \theta_j^2$.
- Consequently, the metric entropy of $\mathscr{F}_\alpha$ scales as $\log N(\delta; \mathscr{F}_\alpha, \|\cdot\|_2) \asymp (1/\delta)^{1/\alpha}$ (Example 5.12).
- Accordingly, we can find a δ-packing $\{f^1, \ldots, f^M\}$ of $\mathscr{F}_\alpha$ in the $\|\cdot\|_2$-norm with $\log M \gtrsim (1/\delta)^{1/\alpha}$ elements.

Example 15.23: Step A

- For each j, let $\mathbb{P}_{f^j}$ denote the distribution of y given $\{x_i\}_{i=1}^n$ when the true regression function is f^j, and let $\mathbb{Q}$ denote the n-fold product distribution over the covariates $\{x_i\}_{i=1}^n$.
- For any distinct pair of indices $j \neq k$, we have

$$
\begin{aligned}
& D\left(\mathbb{P}_{f^j} \times \mathbb{Q} \| \mathbb{P}_{f^k} \times \mathbb{Q}\right) = \mathbb{E}_x\left[D\left(\mathbb{P}_{f^j} \| \mathbb{P}_{f^k}\right)\right] \\
& = \mathbb{E}_x[\frac{1}{2\sigma^2} \sum_{i=1}^n (f^j(x_i) - f^k(x_i))^2] \\
& = \frac{n}{2\sigma^2} \|f^j - f^k\|_2^2
\end{aligned}
$$

- Consequently, we find that

$$
\log N_{\mathrm{KL}}(\epsilon) = \log N(\frac{\sigma\sqrt{2}}{\sqrt{n}}\epsilon; \mathscr{F}_\alpha, \|\cdot\|_2) \precsim (\frac{\sqrt{n}}{\sigma\epsilon})^{1/\alpha}.
$$

- Inequality (13) in step A can be satisfied by setting $\epsilon_n^2 \asymp \left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha+1}}$.

Example 15.23: Step B

- It remains to choose $\delta > 0$ to satisfy the inequality (14) in step B. Given our choice of ϵ_n and the scaling of the packing entropy, we require

$$(1/\delta)^{1/\alpha} \geq c \left\{ \left(\frac{n}{\sigma^2} \right)^{\frac{1}{2\alpha+1}} + 2 \log 2 \right\}$$

- As long as n/σ^2 is larger than some universal constant, the choice $\delta_n^2 \asymp (\frac{\sigma^2}{n})^{\frac{2\alpha}{2\alpha+1}}$ satisfies this condition.