Lecture-30: Fano's method

1 Fano's method

Definition 1.1. Let (V,d) be a metric space and $\Theta \subseteq V$, then an ϵ-ball with centre $v \in V$ and radius $\epsilon \in \mathbb{R}_+$ is defined as the closed set $B(v,\epsilon) \triangleq \{u \in V : d(u,v) \leqslant \epsilon\}$.

(a) We say $\{v_1,\ldots,v_N\} \subseteq V$ is an *ϵ-covering* or *ϵ-net* of Θ if $\Theta \subseteq \cup_{i=1}^N B(v_i,\epsilon)$. Equivalently, for all $\theta \in \Theta$, there exists $i \in [N]$ such that $d(\theta,v_i) \leqslant \epsilon$.

(b) We say $\{\theta_1,\ldots,\theta_M\} \subseteq \Theta$ is an *ϵ-packing* of Θ if $\min_{i\neq j} \|\theta_i - \theta_j\| > \epsilon$; or equivalently, the set of balls $\{B(\theta_i,\epsilon/2) : j \in [M]\}$ are disjoint.

Definition 1.2. Consider a metric space (V,d) and a parameter space $\Theta \subseteq V$.

(a) The covering number is defined as $N(\Theta,d,\epsilon) \triangleq \min\{n \in \mathbb{N} : \text{ there exists an } \epsilon - \text{covering of } \Theta \text{ of size } n\}$.

(b) The packing number is defined as $M(\Theta,d,\epsilon) \triangleq \max\{m \in \mathbb{N} : \text{ there exists an } \epsilon - \text{packing of } \Theta \text{ of size } m\}$.

We discuss another method for proving minimax lower bound by reduction to multiple hypothesis testing. We call this program *Fano's method*, based on the Fano's inequality to show the impossibility result in one of the steps.

Step 1. We assume that the loss function $\ell : \Theta \times \Theta \to \mathbb{R}_+$ is a metric on parameter space Θ.

Step 2. Consider an ϵ-packing $\Theta' \triangleq \{\theta_i \in \Theta : i \in [M]\}$ of the parameter space Θ.

Step 3. Suppose we can show that given observation X one cannot reliably distinguish these hypotheses in Θ'. That is, $P_{\theta_i}\{\tilde{\theta}(X) \neq \theta_i\} > 0$.

Step 4. Then the best estimation error $\mathbb{E}\ell(\hat{\theta},\theta)$ is at least proportional to ϵ.

Step 5. The impossibility of testing is often shown by applying Fano's inequality in Corollary A.3, which bounds the probability of error of testing in terms of the mutual information.

Theorem 1.3 (Fano). *Let $\ell : \Theta \times \Theta \to \mathbb{R}_+$ be a metric on parameter space Θ. Fix an estimator $\hat{\theta}$. For any $T \subseteq \Theta$ and $\epsilon > 0$,*

$$P\left\{\ell(\theta,\hat{\theta}) \geqslant \frac{\epsilon}{2}\right\} \geqslant 1 - \frac{C(T) + \ln 2}{\ln M(T,\ell,\epsilon)}, \tag{1}$$

where $C(T) \triangleq \sup_{\pi\in\mathcal{M}(T)} I(\theta;X)$ is the capacity of the channel $\theta \to X$ with input space T. Consequently, for any $r \in \mathbb{N}$

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \ell(\theta,\hat{\theta})^r \geqslant \sup_{T\subset\Theta,\epsilon>0} \left(\frac{\epsilon}{2}\right)^r \left(1 - \frac{C(T) + \ln 2}{\ln M(T,\ell,\epsilon)}\right).$$

Proof. It suffices to show (1), since the second result follows from the first by applying Markov inequality to increasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$ defined as $f(x) \triangleq x^r$ for any $x \in \mathbb{R}_+$ and $r \geqslant 1$. Fix $T \subseteq \Theta$. Consider an ϵ-packing $T' \triangleq \{\theta_1,\ldots,\theta_M\} \subset T$ such that $\min_{i\neq j\in[M]} \ell(\theta_i,\theta_j) > \epsilon$. For each $i \in [M]$, we define $\frac{\epsilon}{2}$ balls $B(\theta_i,\frac{\epsilon}{2}) \triangleq \left\{\theta \in \Theta : \ell(\theta_i,\theta) \leqslant \frac{\epsilon}{2}\right\}$, and observe that $\left\{B(\theta_i,\frac{\epsilon}{2}) : i \in [M]\right\}$ is a set of disjoint balls. Let $\theta : \Omega \to T'$ be uniformly distributed and $X \sim P_\theta$ conditioned on parameter θ. Given any estimator $\hat{\theta} : \Omega \to \Theta^{\mathcal{X}}$, construct a test by rounding estimate $\hat{\theta}(X)$ to output $\tilde{\theta}(X) \triangleq \arg\min_{\theta\in T'} \ell(\theta,\hat{\theta}(X))$. From the triangle inequality for metric ℓ and definition of $\tilde{\theta}(X)$, we get $\ell(\theta,\tilde{\theta}(X)) \leqslant \ell(\theta,\hat{\theta}(X)) + \ell(\hat{\theta}(X),\tilde{\theta}(X)) \leqslant 2\ell(\theta,\hat{\theta}(X))$ for any $\theta \in T'$, and thus

$$P\left\{\theta \neq \tilde{\theta}(X)\right\} = \mathbb{E}\mathbb{1}_{\left\{\theta\neq\tilde{\theta}(X)\right\}} \leqslant \mathbb{E}\mathbb{1}_{\left\{\ell(\theta,\tilde{\theta}(X))>\epsilon\right\}} \leqslant P\left\{\ell(\theta,\hat{\theta}(X)) > \frac{\epsilon}{2}\right\}.$$

Recall that $|T'| = M$. The result follows from the application of Fano's inequality from Corollary A.3 to the Markov chain $\theta \to X \to \hat{\theta} \to \tilde{\theta}$ with uniformly distributed $\theta \in T'$ to lower bound $P\left\{\theta \neq \tilde{\theta}(X)\right\}$, using the definition of $C(T) \geqslant I(\theta;X)$, and taking supremum over all ϵ-packing T' of T which is the packing number $M(T,\ell,\epsilon)$. □

Remark 1. In applying Fano's method, since it is often difficult to evaluate the capacity $C(T)$, it is useful to recall from Theorem 5.9 that $C(T)$ coincides with the KL radius of the set of distributions $\{P_\theta : \theta \in T\}$, namely, $C(T) \triangleq \inf_Q \sup_{\theta\in T} D(P_\theta\|Q)$. As such, choosing any Q leads to an upper bound on the capacity. As an application, we revisit the d-dimensional GLM in Corollary 28.8 under the ℓ_q loss for $1 \leqslant q \leqslant \infty$, with the particular focus on the dependency on the dimension. For a different application in sparse setting see Exercise VI.12.

Example 1.4. Consider GLM with *i.i.d.* sample size m, where $P_\theta = \mathcal{N}(\theta, I_d)^{\otimes m}$. Taking natural logarithms here and below, we have

$$D(P_\theta\|P_{\theta'}) = \mathbb{E}_{X\sim P_\theta} \ln \frac{dP_\theta}{dP_{\theta'}}(X) = \frac{1}{2}\mathbb{E}_{X\sim P_\theta}(\|X-\theta'\|_2^2 - \|X-\theta\|_2^2) = \frac{m}{2}\|\theta-\theta'\|_2^2.$$

In other words, KL-neighborhoods are ℓ_2 balls. As such, we can apply Theorem 1.3 to $T = B_2(0,\rho)$ where $B_q(0,\rho) \triangleq \left\{x \in \mathbb{R}^d : \|x\|_q \leqslant \rho\right\}$ for some $q \geqslant 1$ and $\rho > 0$ to be specified. Then $C(T) \leqslant \sup_{\theta\in T} D(P_\theta\|P_0) = \frac{m}{2}\rho^2$. To bound the packing number from below, we applying the volume bound in Theorem 27.3,

$$M(B_2(0,\rho), \|\cdot\|_q, \epsilon) \geqslant \frac{\rho^d \operatorname{vol}(B_2(0,1))}{\epsilon^d \operatorname{vol}(B_q(0,1))} \geqslant \Big(\frac{c_q \rho d^{\frac{1}{q}}}{\epsilon\sqrt{d}}\Big)^d.$$

for some constant c_q, where the last step follows the volume formula (27.13) for ℓ_q balls. Choosing $\rho = \sqrt{\frac{d}{m}}$ and $\epsilon = \frac{c_q}{e^2}\rho d^{\frac{1}{q}-\frac{1}{2}}$, an application of Theorem 1.3 yields the minimax lower bound for risk $R_q \triangleq \inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^d} \mathbb{E}_{X\sim P_\theta}\|\hat\theta-\theta\|_q$, as

$$R_q \geqslant \sup_{\rho,\epsilon>0}\Big(\frac{\epsilon}{2}\Big)\Big(1 - \frac{C(T)+\ln 2}{\ln M(T,\ell,\epsilon)}\Big)^{1/q} \geqslant \frac{c_q d^{\frac{1}{q}}}{2\sqrt{m}e^2}\Big(\frac{3}{4} - \frac{1}{2d}\ln 2\Big)^{\frac{1}{q}} \geqslant C_q\frac{d^{\frac{1}{q}}}{\sqrt{m}} \tag{2}$$

for some constant C_q depending on q. This is the same lower bound as that in (30.9) obtained via the mutual information method plus the Shannon lower bound which is also volume-based.

Remark 2. For any $q \geqslant 1$, (2) is rate-optimal since we can apply the MLE $\hat\theta(X) = X$. Note that at $q = \infty$, the constant C_q is still finite since $\operatorname{vol}(B_\infty) = 2^d$. However, for the special case of $q = \infty$, (2) does not depend on the dimension at all, as opposed to the correct dependency $\sqrt{\ln d}$ shown in Corollary 28.8. In fact, previously in Example 31.4 the application of Assouad's lemma yields the same suboptimal result. So is it possible to fix this looseness with Fano's method? It turns out that the answer is yes and the suboptimality is due to the volume bound on the metric entropy, which, as we have seen in Section 27.3, can be ineffective if ϵ scales with dimension. Indeed, if we apply the tight bound of $M(B_2, \|\cdot\|_\infty, \epsilon)$ in (27.18)[1], with $\epsilon = \sqrt{\frac{c\ln d}{m}}$ and $\rho = \sqrt{\frac{c'\ln d}{m}}$ for some absolute constants c, c', we do get $R_\infty \gtrsim \sqrt{\frac{\ln d}{m}}$ as desired.

Remark 3. It is sometimes convenient to further bound the KL radius by the KL diameter, since $C(T) \leqslant \operatorname{diam}_{\mathrm{KL}}(T) \triangleq \sup_{\theta,\theta'\in T} D(P_{\theta'}\|P_\theta)$ (cf. Corollary 5.8). This suffices for Example 31.5.

Remark 4. In Theorem 1.3 we actually lower bound the global minimax risk by that restricted on a parameter subspace $T \subset \Theta$ for the purpose of controlling the mutual information, which is often difficult to compute. For the GLM considered in Example 31.5, the KL divergence is proportional to squared ℓ_2 distance and T is naturally chosen to be a Euclidean ball. For other models such as the covariance model (Exercise VI.16) wherein the KL divergence is more complicated, the KL neighborhood T needs to be chosen carefully. Later in Section 32.4 we will apply the same Fano's method to the infinite-dimensional problem of estimating smooth density.

[1] In fact, in this case we can also choose the explicit packing $\{\epsilon e_1, \ldots, \epsilon e_d\}$

A Fano's inequality

Definition A.1 (Binary entropy and KL divergence). Consider binary random variables $X, Y : \Omega \to \mathcal{X} \triangleq \{0,1\}$ with respective probability mass functions $(p, 1-p), (q, 1-q) \in \mathcal{M}(\mathcal{X})$ for any $p, q \in [0,1]$. Then binary entropy $h : [0,1] \to [0,1]$ is defined as $h(p) \triangleq H(X) = -p \ln p - (1-p)\ln(1-p)$ for all $p \in [0,1]$, and the binary KL divergence is defined as $d : [0,1] \times [0,1] \to \mathbb{R}_+$ as $d(p,q) \triangleq D(P_X \| P_Y) = p \ln \frac{p}{q} + (1-p) \ln \frac{(1-p)}{(1-q)}$ for all $p, q \in [0,1]$.

Theorem A.2 (Fano's inequality). *Let $|\mathcal{X}| = M < \infty$, $X \to Y \to \hat{X}$ be a Markov chain, and $P_e \triangleq P\{X \neq \hat{X}\}$. Then the following are true*
(a) $H(X \mid Y) \leqslant F_M(1 - P_e) \triangleq P_e \ln(M-1) + h(P_e)$.
(b) If $P_{\max} \triangleq \max_{x \in \mathcal{X}} P_X(x) > 0$, *then* $I(X;Y) \geqslant (1 - P_e) \ln \frac{1}{P_{\max}} - h(Pe)$ *regardless of* $|\mathcal{X}|$.

Proof. Consider two joint distributions $P_{X,Y,\hat{X}} = P_X P_{Y|X} P_{\hat{X}|Y}$ and $Q_{X,Y,\hat{X}} = Q_X P_Y P_{\hat{X}|Y}$, and the data processor (kernel) $(X, Y, \hat{X}) \mapsto \mathbb{1}_{\{X \neq \hat{X}\}}$. We note that X and Y are independent under Q, and the observation Y has identical marginal $P_Y = Q_Y$ under both P and Q. Further, the kernel $P_{\hat{X}|Y} = Q_{\hat{X}|Y}$, i.e. the estimator for X is same for both distributions based on the observation Y. We recall the KL divergence data processing inequality for Markov chain $X \to Y \to \hat{X}$, which implies

$$D(P_{X,Y,\hat{X}} \| Q_{X,Y,\hat{X}}) \geqslant D(P_{X,\hat{X}} \| Q_{X,\hat{X}}) \geqslant d(P\{X = \hat{X}\} \| Q\{X = \hat{X}\}).$$

From the definition of KL divergence, mutual information, and joint distributions P, Q, we get

$$D(P_{X,Y,\hat{X}} \| Q_{X,Y,\hat{X}}) = \mathbb{E}_P \ln \frac{dP_X dP_{Y|X}}{dQ_X dP_Y} = D(P_X \| Q_X) + I(X;Y) \geqslant d(P\{X = \hat{X}\} \| Q\{X = \hat{X}\}).$$

(a) Let $U_X \in \mathcal{M}(\mathcal{X})$ be a uniform distribution over $\mathcal{X}$. Then for $Q_X = U_X$, we obtain $D(P_X \| Q_X) = \ln M - H(X)$ and since $I(X;Y) = H(X) - H(X \mid Y)$, we get $D(P_X \| Q_X) + I(X;Y) = \ln M - H(X \mid Y)$. Further, since X and Y are independent under Q, we get $Q\{X = \hat{X}\} = \sum_{x \in \mathcal{X}} Q\{X = \hat{X} = x\} = \sum_{x \in \mathcal{X}} Q_X(x) Q_{\hat{X}}(x) = \frac{1}{M}$. It follows that

$$\ln M - H(X \mid Y) \geqslant d\Big(P_e \| 1 - \frac{1}{M}\Big) = -h(P_e) + (1 - P_e) \ln M + P_e \ln \frac{M}{M-1}.$$

(b) When $P_X = Q_X$, we get $D(P_X \| Q_X) = 0$ and $Q\{X = \hat{X}\} = \sum_{\mathcal{X}} P_X(x) Q_{\hat{X}}(x) \leqslant P_{\max}$. Therefore,

$$\begin{aligned} I(X;Y) &\geqslant d(1 - P_e \| Q\{X = \hat{X}\}) = -h(P_e) - (1 - P_e) \ln Q\{X = \hat{X}\} - P_e \ln Q\{X \neq \hat{X}\} \\ &\geqslant -h(P_e) + (1 - P_e) \ln \frac{1}{Q\{X = \hat{X}\}} \geqslant -h(P_e) - (1 - P_e) \ln P_{\max}. \end{aligned}$$

□

The following corollary of the previous result emphasizes its role in providing converses or impossibility results for statistics and data transmission.

Corollary A.3 (Lower bound on average probability of error). *Consider Markov chain $W \to X \to Y \to \hat{W}$, where W is uniform on $[M] \triangleq \{1, \ldots, M\}$. Then*

$$P_e \triangleq P\{W \neq \hat{W}\} \geqslant 1 - \frac{I(X;Y) + h(P_e)}{\ln M} \geqslant 1 - \frac{I(X;Y) + \ln 2}{\ln M}.$$

Proof. Apply Theorem A.2 and the data processing for mutual information $I(W;\hat{W}) \leqslant I(X;Y)$. □