Infotheory for Statistics and Learning

Lecture 5

- Repeated iid experiments [PW:28.4–5]
- The Gaussian location model [PW:28.2]
- The mutual information method [PW:30]
- Fano's method [PW:6.3,31.4],[CT:2.10]
- Capacity and information radius [PW:5.3,30.1]

Product of Experiments

Consider the model $P_\theta = P_{\theta_1} \otimes \cdots \otimes P_{\theta_p}$ for $\theta_i \in \Theta_i$ with observation

$$X = (X_1, \ldots, X_p) \sim P_\theta$$

and loss

$$\ell(\theta, \hat{\theta}) = \sum_{i=1}^{p} \ell_i(\theta_i, \hat{\theta}_i)$$

For $R^* =$ minimax risk of product, $R_i^* =$ minimax risk of individual and $S_i^* = \sup_{\pi_i} R_{\pi_i}^* =$ worst-case Bayes of individual, we have

$$\sum_{i=1}^{p} S_i^* \le R^* \le \sum_{i=1}^{p} R_i^*$$

Thus if the minimax theorem holds for each i, we get $R^* = \sum_i R_i^*$

Repeated iid Experiments

Consider instead n repeated independent and identically distributed (iid) experiments:

$$X = (X_1, \ldots, X_n), \quad X_i \sim P_\theta \text{ and independent}$$

The resulting minimax risk R_n^* is non-increasing, and usually $\to 0$ as $n \to \infty$

Sample complexity

$$n^*(\varepsilon) = \min\{n : R_n^* \leq \varepsilon\}$$

Example: Gaussian location model (GLM),

$$X_i \sim \mathcal{N}(\theta, \sigma^2 I_p), \;\; i = 1, \ldots, n$$

iid in i, and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$. First, let $n = 1$ and $X = X_1$:

For any π, $\theta \sim \pi$, $R_\pi(\hat{\theta}) = E_\pi[E_\theta\{E[\|\theta - \hat{\theta}\|^2 | X]\}]$

Let $g(x) = E[\theta | X = x]$, then for each $X = x$

$$E[\|\theta - \hat{\theta}\|^2 | X = x] = E[\|\theta - g(x)\|^2 | X = x] + E[\|g(x) - \hat{\theta}(x)\|^2 | X = x]$$

Thus $\hat{\theta}^*(x) = g(x)$

We also know that

$$|E[(\theta - g(X))(\theta - g(X))^T]| \geq \frac{1}{(2\pi e)^p} 2^{2h(\theta|X)}$$

where the RHS is maximized for $P_{\theta|X}$ Gaussian and the LHS = RHS for θ and X jointly Gaussian

Because

$$h(\theta|X) \le \sum_{k=1}^{p} h(\theta_k|X_k)$$

with $=$ if (θ_k, X_k) are independent in k, we can take $\theta_k \sim \mathcal{N}(0, \gamma)$ and independent in k, to get

$$h(\theta|X) = \frac{p}{2}\log 2\pi e \frac{\sigma^2\gamma}{\sigma^2+\gamma} \Rightarrow \frac{1}{(2\pi e)^p} 2^{2h(\theta|X)} = \left(\frac{\sigma^2\gamma}{\sigma^2+\gamma}\right)^p$$

and since $E[(\theta - g(X))(\theta - g(X))^T] = E[(\theta_i - g_i(X))^2] I_p$ we get

$$\sup_\pi R_\pi^* = \lim_{\gamma\to\infty} p\frac{\sigma^2\gamma}{\sigma^2+\gamma} = p\sigma^2$$

and since $\sup_\pi R_\pi^* \le R^*$ and $R(\hat{\theta}(x)) = p\sigma^2$ is achieved by $\hat{\theta}(x) = x$ we have also $R^* = p\sigma^2$ (for $n = 1$)

For $n > 1$, let $\bar{x}_n = n^{-1}\sum_i x_i$, then

$$\begin{aligned} f(x|\theta) &= \frac{1}{(2\pi\sigma^2)^{(pn)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\|x_i - \theta\|^2\right) \\ &= \frac{1}{n^{p/2}(2\pi\sigma^2)^{(n-1)p/2}} f(\bar{x}_n|\theta)\, e^{-\frac{1}{2\sigma^2}(\sum_i \|x_i\|^2 - n\|\bar{x}_n\|^2)} \end{aligned}$$

Thus $\bar{X}_n$ is a sufficient statistic of X for θ and observing $\{X_i\}$ is equivalent to seeing $\bar{X}_n \sim \mathcal{N}(\theta, (\sigma^2/n)I_p)$, and consequently

$$R_n^* = p\frac{\sigma^2}{n} \quad \text{and} \quad n^*(\varepsilon) = \left\lceil p\frac{\sigma^2}{\varepsilon} \right\rceil$$

$\Rightarrow$ fundamental trade-off between p and n

Information Bounds

For a given P_θ, $\theta \sim \pi$ and $P_{\hat{\theta}|X}$ such that $E[\ell(\theta, \hat{\theta})] \leq D$, we have

$$R(D) = \inf_{P_{\hat{\theta}|\theta}: E[\ell(\theta,\hat{\theta})] \leq D} I(\theta; \hat{\theta}) \leq I(\theta; \hat{\theta}) \leq I(\theta; X) \leq \sup_\pi I(\theta; X)$$

Assume $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^r$ (rth power of a norm over $\mathbb{R}^p$),

$$\begin{aligned} R(D) &= \inf_{P_{\hat{\theta}|\theta}: E[\|\theta-\hat{\theta}\|^r] \leq D} \{h(\theta) - h(\theta - \hat{\theta}|\hat{\theta})\} \\ &\geq h(\theta) - \sup_{P_{\hat{\theta}|\theta}: E[\|\theta-\hat{\theta})\|^r] \leq D} h(\theta - \hat{\theta}) \\ &\geq h(\theta) - \log\left(V_p \left(\frac{Dre}{p}\right)^{p/r} \Gamma\left(1 + \frac{p}{r}\right)\right) \end{aligned}$$

where

$$V_p = \int_{\|x\| \leq 1} dx$$

The RHS of the bound = the Shannon lower bound on $R(D)$

The bound is tight as $D \to 0$

For $p = 1$, $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, we get $V_1 = 1$, $\Gamma(3/2) = \sqrt{\pi}/2$ and

$$R(D) \geq h(\theta) - \frac{1}{2}\log(2\pi e D) = \frac{1}{2}\log\frac{\sigma^2}{D} - D(\pi\|g)$$

with $g = \mathcal{N}(0, \sigma^2)$ and $\sigma^2 = E[\theta^2]$, recovering our previous bound

For the GLM $\bar{X}_n \sim \mathcal{N}(\theta, (1/n)I_p)$ with $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^r$, we get

$$\frac{p}{2}\log\left(1+n\gamma\right) \geq I(\theta; \bar{X}_n) \geq I(\theta, \hat{\theta}) \geq R(R_\pi^*)$$
$$\geq h(\theta) - \log\left(V_p\left(\frac{R_\pi^* re}{p}\right)^{p/r}\Gamma\left(1+\frac{p}{r}\right)\right)$$

for any π s.t. $E\|\theta\|^2 = p\gamma < \infty$. Thus

$$R_\pi^* \geq \frac{p}{re}\left(V_p\,\Gamma\left(1+\frac{p}{r}\right)\right)^{-r/p} 2^{(r/p)(h(\theta)-(p/2)\log(1+n\gamma))}$$

Maximizing over π s.t. $E\|\theta\|^2 = p\gamma$ and then letting $\gamma \to \infty$ we thus get

$$R_n^* \geq \frac{p}{re}\left(V_p\,\Gamma\left(1+\frac{p}{r}\right)\right)^{-r/p}\left(\frac{2\pi e}{n}\right)^{r/2}$$

Sanity check, $p = 1$ and $r = 2 \Rightarrow$ RHS $= 1/n = R_n^*$

Fano Bounds

Consider a discrete and finite Θ, i.e. $\theta \in \{\theta_1, \ldots, \theta_M\}$

For π uniform on Θ and $\theta \to X \to \hat{\theta}$ use

$$\ell(\theta, \hat{\theta}) = \mathbb{1}(\{\theta \neq \hat{\theta}\}) \Rightarrow E[\ell(\theta, \hat{\theta})] = \Pr(\hat{\theta} \neq \theta) = P_e$$

Recall that for $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = P_{Y|X} \circ Q_X$ (two input distributions through the same kernel), $D(P_Y\|Q_Y) \leq D(P_X\|Q_X)$

With $P_X \to P_{\theta,\hat{\theta}}$, $Q_X \to \pi \otimes P_{\hat{\theta}}$, and $P_{Y|X} \to P_{Z|\theta,\hat{\theta}}$ where $Z \in \{0, 1\}$ and $P_{Z|\theta,\hat{\theta}}(\{Z = 1\}|\theta, \hat{\theta}) = \ell(\theta, \hat{\theta})$ we get

$$P_Z = P_{Z|\theta,\hat{\theta}} \circ P_{\theta,\hat{\theta}}, \quad Q_Z = P_{Z|\theta,\hat{\theta}} \circ (\pi \otimes P_{\hat{\theta}})$$

that is, $P_Z(\{Z = 1\}) = P_e$ and $Q_Z(\{Z = 1\}) = 1 - 1/M$. Thus

$$I(\theta; \hat{\theta}) \geq D(P_Z\|Q_Z) = \log M - P_e \log(M-1) - H(P_Z)$$

Hence since $H(P_Z) \leq 1$, we arrive at Fano's inequality

$$P_e \geq 1 - \frac{I(\theta; \hat{\theta}) + 1}{\log M} \geq 1 - \frac{I(\theta; X) + 1}{\log M}$$

For $\Theta = \mathbb{R}^p$ and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ (for $\|\cdot\|$ a norm on $\mathbb{R}^p$)

Pick a discrete subset $\tilde{\Theta} = \{\theta_1, \ldots, \theta_M\}$ such that $\|\theta_i - \theta_j\| \geq \varepsilon$

Consider $\theta \to X \to \hat{\theta} \to f(\hat{\theta})$, $f(\hat{\theta}) = \theta_i$ if $\|\hat{\theta} - \theta_i\| \leq \|\hat{\theta} - \theta_j\|$

Assume the true θ is $\theta_k \in \tilde{\Theta}$, then

$$\Pr(f(\hat{\theta}) \neq \theta_k) \leq \Pr\left(\|\hat{\theta} - \theta_k\| \geq \frac{\varepsilon}{2}\right) \leq \frac{E\|\hat{\theta} - \theta_k\|}{\varepsilon/2}$$

Define π on $(\mathbb{R}^p, \mathcal{B}^p)$ by picking θ uniformly from $\tilde{\Theta}$, then

$$\begin{aligned} R^* &\geq \sup_\pi R^*_\pi \geq \frac{1}{M}\sum_{i=1}^{M} E\|\hat{\theta} - \theta_i\| \\ &\geq \frac{\varepsilon}{2M}\sum_{i=1}^{M} \Pr(f(\hat{\theta}) \neq \theta_i) \geq \frac{\varepsilon}{2}\left(1 - \frac{\sup_\pi I(\theta; X) + 1}{\log M}\right) \end{aligned}$$

Hence R^* is limited from below by the capacity

$$C = \sup_\pi I(\theta; X)$$

of the link $\theta \to X$ in $\theta \to X \to \hat{\theta}$

Information Radius

For any function $f(x,y)$, $x \in A$, $y \in B$, set $g(x) = \inf_y f(x,y)$

Thus $\sup_x g(x) \leq \sup_x f(x,y)$ for all $y \in B$, in particular $\sup_x g(x) \leq \inf_y(\sup_x f(x,y))$, and thus

$$\sup_x \inf_y f(x,y) \leq \inf_y \sup_x f(x,y)$$

For some set Ω and $\ell : \Omega^2 \to [0,\infty)$, the radius of $A \subset \Omega$ is

$$r(A) = \inf_{y \in \Omega} \sup_{x \in A} \ell(x,y)$$

and the diameter of A is

$$d(A) = \sup_{(x,y) \in A^2} \ell(x,y)$$

Note that $r(A) \leq d(A)$

For $\theta \to X \to \hat{\theta}$, let $\mathcal{P} = \{P_\theta\}$ and remember that

$$I(X;Y) = \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X)$$

Thus for the capacity

$$C = \sup_\pi I(\theta;X) = \sup_\pi \inf_{Q_X} D(P_{X|\theta} \| Q_X | \pi) \leq \inf_{Q_X} \sup_\pi D(P_{X|\theta} \| Q_X | \pi)$$
$$= \inf_Q \sup_\theta D(P_\theta \| Q) = r(\mathcal{P}) \leq d(\mathcal{P}) = \sup_{\theta \neq \theta'} D(P_\theta \| P_{\theta'})$$

with radius and diameter in the sense of $\ell(P,Q) = D(P\|Q)$

So, e.g. for the Fano bound

$$R^* \geq \frac{\varepsilon}{2}\left(1 - \frac{r(\mathcal{P}) + 1}{\log M}\right) \geq \frac{\varepsilon}{2}\left(1 - \frac{\sup_{\theta \neq \theta'} D(P_\theta \| P_{\theta'}) + 1}{\log M}\right)$$