Lecture-28: Le Cam's method: applications

1 Applications of Le Cam's method

Corollary 1.1. *Consider a simple statistical decision theory setting with $\Theta = \hat{\Theta}$, and loss function $\ell : \Theta \times \Theta \to \mathbb{R}_+$ that is an α-metric on parameter space Θ. Then, the minimax risk $R^*(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\theta, \hat{\theta})$ satisfies*

$$R^*(\Theta) \geqslant \sup_{\theta_0,\theta_1 \in \Theta} \frac{\ell(\theta_0,\theta_1)}{2\alpha}(1 - \mathrm{LC}(P_{\theta_0}, P_{\theta_1})) \geqslant \sup_{\theta_0,\theta_1 \in \Theta} \frac{\ell(\theta_0,\theta_1)}{2\alpha}(1 - H^2(P_{\theta_0}, P_{\theta_1})). \tag{1}$$

Proof. For $x > 0$, we have $(1 - \sqrt{x})^2 \geqslant 0$ and hence $2(1 + x) \geqslant (1 + \sqrt{x})^2$. It follows that $(1 - \sqrt{x})^2 \geqslant \frac{(1-x)^2}{2(1+x)}$. From the definition of squared Hellinger distance and Le Cam distance and monotonicity of expectation, we observe that $H^2(P,Q) \geqslant \mathrm{LC}(P,Q)$. □

Example 1.2 (One-dimensional GLM). Consider *i.i.d.* observation sample $X : \Omega \to \mathcal{X}^m$ with common distribution $\mathcal{N}(\theta, 1)$ for $\theta \in \Theta \triangleq \mathbb{R}$. Considering the sufficient statistic $\bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i$, the model is simply $\left\{\mathcal{N}(\theta, \frac{1}{m}) : \theta \in \mathbb{R}\right\}$. We observe that $\sqrt{m}(\bar{X} - \theta_0) \sim \mathcal{N}(\sqrt{m}(\theta - \theta_0), 1)$. From the shift and scale invariance of the total variation distance from Lemma A.1, we have

$$\mathrm{TV}(\mathcal{N}(\theta_0, \frac{1}{m}), \mathcal{N}(\theta_1, \frac{1}{m})) = \mathrm{TV}(P_{\bar{X}|\theta_0}, P_{\bar{X}|\theta_1}) = \mathrm{TV}(P_{\sqrt{m}(\bar{X}-\theta_0)|\theta_0}, P_{\sqrt{m}(\bar{X}-\theta_0)|\theta_1}) = \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(s,1)),$$

where $s \triangleq \sqrt{m}(\theta_1 - \theta_0)$. Applying Le Cam's Theorem to $\Theta' \triangleq \{\theta_0, \theta_1\} \subset \Theta$ and quadratic loss with $\alpha = 2$, we obtain

$$R^* \geqslant \sup_{\theta_0,\theta_1 \in \mathbb{R}} \frac{1}{4}\left|\theta_0 - \theta_1\right|^2 (1 - \mathrm{TV}(\mathcal{N}(\theta_0, \frac{1}{m}), \mathcal{N}(\theta_1, \frac{1}{m})) = \frac{1}{4m} \sup_{s>0} s^2 (1 - \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(s,1))).$$

We can compute the total variation distance between two unit variance Gaussians with means 0 and $s > 0$, as

$$\begin{aligned}\mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(s,1)) &= \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\frac{s}{2}} (e^{-\frac{1}{2}x^2} - e^{-\frac{1}{2}(x-s)^2}) dx + \frac{1}{2\sqrt{2\pi}} \int_{\frac{s}{2}}^{\infty} (e^{-\frac{1}{2}(x-s)^2} - e^{-\frac{1}{2}x^2}) dx \\ &= \Big(1 - 2Q\big(\frac{s}{2}\big)\Big).\end{aligned}$$

It follows that $\frac{s^2}{4m}(1 - \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(s,1))) = \frac{1}{2m}s^2 Q(\frac{s}{2})$ and $\sup_{s>0} \frac{1}{2}s^2 Q(\frac{s}{2}) = c$ for some absolute constant $c \approx 0.083$. It follows that $R^* \geqslant \frac{c}{m}$. On the other hand, we know that the minimax risk equals $\frac{1}{m}$, so the two-point method is rate-optimal in this case.

Remark 1. In the above example, for two points separated by $\Theta(\frac{1}{\sqrt{m}})$, the corresponding hypothesis cannot be tested with vanishing probability of error so that the resulting estimation risk (say in squared error) cannot be smaller than $\frac{1}{m}$. This convergence rate is commonly known as the *parametric rate* for smooth parametric families focusing on the Fisher information as the sharp constant. More generally, the $\frac{1}{m}$ rate is not improvable for models with locally quadratic behavior

$$H^2(P_{\theta_0}, P_{\theta_0+t}) \asymp t^2, \text{ for } t \to 0. \tag{2}$$

We have studied the sufficient conditions for this local behavior of f-divergences. Indeed, picking $\theta_0 \in \Theta^o$ and setting $\theta_1 \triangleq \theta_0 + \frac{1}{\sqrt{m}}$, so that $H^2(P_{\theta_0},P_{\theta_1}) = \Theta(\frac{1}{m})$ from (2). By Theorem A.2, we have $\mathrm{TV}(P_{\theta_0}^{\otimes m},P_{\theta_1}^{\otimes m}) \leqslant 1-c$ for some constant c and hence Le Cam's Theorem yields the lower bound $\Omega(\frac{1}{m})$ for the squared error.

Example 1.3 (Uniform family). Consider the parameter space $\Theta \triangleq \mathbb{R}_+$ and the parametric family of distributions $\mathcal{P}(\Theta) \triangleq (U_\theta : \theta \in \mathbb{R}_+)$ where $U_\theta : \Omega \to (0,\theta)$ is a uniform random variable, and quadratic loss function $\ell : \Theta \times \hat{\Theta} \to \mathbb{R}_+$ defined as $\ell(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$ for each pair $(\theta,\hat{\theta}) \in \Theta \times \hat{\Theta}$. Consider $\Theta' \triangleq \{\theta_0,\theta_1\}$. We can find the squared Hellinger distance between distributions P_{θ_0} and P_{θ_1} as

$$H^2(U(0,\theta_0),U(0,\theta_1)) = 2 - 2\int_0^{\theta_0\wedge\theta_1} \frac{1}{\sqrt{\theta_0\theta_1}}du = 2(1 - \sqrt{\frac{\theta_0}{\theta_1}} \wedge \sqrt{\frac{\theta_1}{\theta_0}}).$$

Since this distance only depends on the ratio, we can take $\theta_0 = 1$ without any loss of generality and $\theta_1 > \theta_0 = 1 + t$ for some $t > 0$. Note that as opposed to the quadratic behavior in (2), we have $H^2(U(0,\theta_0),U(0,\theta_1)) = 2\Big(1 - \sqrt{\frac{\theta_0}{\theta_1}}\Big) \asymp t$. For an m size *i.i.d.* sample $X : \Omega \to \mathcal{X}^m$, we have

$$H^2(U(0,\theta_0)^{\otimes m},U(0,\theta_1)^{\otimes m}) = 2 - 2\int_0^{\theta_0}\cdots\int_0^{\theta_0} \frac{du_1\ldots du_m}{(\theta_0\theta_1)^{\frac{m}{2}}} = 2\Big(1 - (\frac{\theta_0}{\theta_1})^{\frac{m}{2}}\Big) \asymp mt.$$

Recall that quadratic loss function ℓ is a 2-metric on $\mathbb{R}_+$ and $\ell(\theta_0,\theta_1) = t^2$. Applying Le Cam's theorem to Θ' and quadratic loss function, we obtain

$$R^* \geqslant \frac{1}{4}\sup_{t>0} t^2(1-mt) = \frac{1}{27m^2}.$$

This maximum is obtained for $t^* = 2/(3m)$. This rate is not achieved by the empirical mean estimator which only achieves $1/m$ rate, but by the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}(X) \triangleq \max\{X_1,\ldots,X_m\}$. To observe the rate of empirical mean estimator $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$, we note that

$$R_\theta = \mathbb{E}_{X\sim P_{\theta^{\otimes m}}}(\theta - 2\bar{X})^2 = \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}_{X_i\sim P_\theta}(2X_i - \theta)^2 = \frac{\theta^2}{m}\int_0^1 (2x-1)^2 dx = \frac{\theta^2}{3m}.$$

To derive the ML estimator, we observe that

$$dP_{X|\theta} = \prod_{i=1}^m dP_\theta(X_i) = \frac{1}{\theta^m}\prod_{i=1}^m \mathbb{1}_{\{X_i\leqslant\theta\}} = \frac{1}{\theta^m}\mathbb{1}_{\{\max_{i\in[m]} X_i\leqslant\theta\}}.$$

The right hand side is maximized for maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}(X) = \max_{i\in[m]} X_i$. Conditioned on the true parameter θ, the distribution of $\hat{\theta}_{\mathrm{ML}}(X)$ is

$$P_{\theta^{\otimes m}}\left\{\hat{\theta}_{\mathrm{ML}}(X) \leqslant x\right\} = P_{\theta^{\otimes m}} \cap_{i=1}^m \{X_i \leqslant x\} = \prod_{i=1}^m P_\theta\{X_i \leqslant x\} = \Big(\frac{x\wedge\theta}{\theta}\Big)^m \mathbb{1}_{\{x\geqslant 0\}}.$$

Recall that $\hat{\Theta} \triangleq \Theta \triangleq \mathbb{R}_+$ and the quadratic loss function ℓ is a 2-metric on Θ. Thus, the quadratic risk for ML estimator is

$$R_\theta = \mathbb{E}_{X\sim P_{\theta^{\otimes m}}}(\theta - \hat{\theta}_{\mathrm{ML}}(X))^2 = m\theta^2\int_0^1 (1-x)^2 x^{m-1}dx = m\theta^2\frac{(m-1)!2!}{(m+2)!} = \frac{2\theta^2}{(m+2)(m+1)}.$$

Other types of behavior in t, and hence the rates of convergence, can occur even in compactly supported location families.

Remark 2. The limitation of Le Cam's two-point method is that it does not capture the correct dependency on the dimensionality. To see this, let us revisit Example 1.2 for d dimensions.

Example 1.4 (d-dimensional GLM). Consider *i.i.d.* observation sample $X : \Omega \to \mathcal{X}^m$ with common distribution $\mathcal{N}(\theta, I_d)$ for $\theta \in \Theta \triangleq \mathbb{R}^d$. For the sufficient statistic $\bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i$, the model is simply $\left\{\mathcal{N}(\theta, \frac{1}{m}I_d) : \theta \in \mathbb{R}^d\right\}$. For quadratic loss $\ell(\theta,\hat{\theta}) \triangleq \left\|\theta - \hat{\theta}\right\|_2^2$ defined for all $\theta,\hat{\theta} \in \Theta \subseteq \mathbb{R}^d$, the exact minimax risk is known to be $R^* = \frac{d}{m}$ for any dimension d and sample size m. Let us compare this with the best two-point lower bound. From the shift and scale invariance of the total variation distance from Lemma A.1 and defining $\theta \triangleq \sqrt{m}(\theta_1 - \theta_0)$, we have

$$\mathrm{TV}(\mathcal{N}(\theta_0, \frac{1}{m}I_d), \mathcal{N}(\theta_1, \frac{1}{m}I_d)) = \mathrm{TV}(P_{\sqrt{m}(\bar{X}-\theta_0)|\theta_0}, P_{\sqrt{m}(\bar{X}-\theta_0)|\theta_1}) = \mathrm{TV}(\mathcal{N}(0, I_d), \mathcal{N}(\theta, I_d)).$$

Applying Le Cam's Theorem to $\Theta' \triangleq \{\theta_0, \theta_1\} \subset \Theta$ for quadratic loss which a 2-metric on Θ, we get

$$R^* \geqslant \sup_{\theta_0,\theta_1 \in \mathbb{R}^d} \frac{1}{4}\left\|\theta_0 - \theta_1\right\|_2^2 (1 - \mathrm{TV}(\mathcal{N}(\theta_0, \frac{1}{m}I_d), \mathcal{N}(\theta_1, \frac{1}{m}I_d))) = \sup_{\theta \in \mathbb{R}^d} \frac{1}{4m}\left\|\theta\right\|_2^2 (1 - \mathrm{TV}(\mathcal{N}(0, I_d), \mathcal{N}(\theta, I_d))).$$

From rotational invariance of isotropic Gaussians, we can rotate the vector θ to align with a coordinate vector $e_1 \triangleq (1,0,\ldots,0)$, which reduces the problem to one dimension, namely, $\mathrm{TV}(\mathcal{N}(0, I_d), \mathcal{N}(\theta, I_d)) = \mathrm{TV}(\mathcal{N}(0, I_d), \mathcal{N}(\left\|\theta\right\|_2 e_1, I_d) = \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(\left\|\theta\right\|_2, 1))$. Thus, we obtain

$$R^* \geqslant \frac{1}{4m}\sup_{s>0} s^2(1 - \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(s,1))).$$

Comparing the above display with (31.3), we see that the best Le Cam two-point lower bound in d dimensions coincide with that in one dimension.

Remark 3. Let us mention in passing that although Le Cam's two-point method is typically suboptimal for estimating a high-dimensional parameter θ, for functional estimation in high dimensions e.g. estimating a scalar functional $T(\theta)$, Le Cam's method is much more effective and sometimes even optimal. The subtlety is that as opposed to testing a pair of simple hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, we need to test $H_0 : T(\theta) = t_0$ versus $H_1 : T(\theta) = t_1$, both of which are composite hypotheses and require a sagacious choice of priors.

A Properties of total variation distance

Lemma A.1 (Shift and scale invariance of total variation). *Consider $\mathcal{X} \triangleq \mathbb{R}^d$. Consider a random vector $X : \Omega \to \mathcal{X}^{\{0,1\}}$ with marginals $P_{X_0}, P_{X_1} \in \mathcal{M}(\mathcal{X})$. Let $P_{X_0}, P_{X_1} \ll \mu \in \mathcal{M}(\mathcal{X})$, such that relative densities are $p_i \triangleq \frac{dP_{X_i}}{d\mu}$ for $i \in \{0,1\}$. We define shifted and scaled version of X as a random vector $Y : \Omega \to \mathcal{Y}^{\{0,1\}}$ where $Y_i \triangleq aX_i + b$ for $i \in \{0,1\}$ for some $a \in \mathbb{R}$ and $b \in \mathbb{R}^d$. Then, $\mathrm{TV}(P_{Y_0}, P_{Y_1}) = TV(P_{X_0}, P_{X_1})$.*

Proof. Recall that $\mathrm{TV}(P_X, P_Y) = \sup_{E \in \mathcal{B}(\mathcal{X})}(P\{X \in E\} - P\{Y \in E\})$. Therefore, we can write

$$\mathrm{TV}(P_{Y_0}, P_{Y_1}) = \sup_{E \in \mathcal{B}(\mathcal{X})}\left(P\left\{X_0 \in \frac{1}{a}(E-b)\right\} - P\left\{X_1 \in \frac{1}{a}(E-b)\right\}\right) = \mathrm{TV}(P_{X_0}, P_{X_1}).$$

□

Theorem A.2. *For any sequence of distributions $P, Q \in \mathcal{M}(\mathcal{X})^{\mathbb{N}}$, we have following equivalences as $m \to \infty$,*

$$\mathrm{TV}(P_m^{\otimes m}, Q_m^{\otimes m}) \to 0 \iff H^2(P_m, Q_m) = o\Big(\frac{1}{m}\Big), \quad \mathrm{TV}(P_m^{\otimes m}, Q_m^{\otimes m}) \to 1 \iff H^2(P_m, Q_m) = \omega\Big(\frac{1}{m}\Big),$$

Proof. For convenience, we assume that observation $X : \Omega \to \mathcal{X}^m$ is *i.i.d.* with common distribution $Q_m \in \mathcal{M}(\mathcal{X})$. Then,

$$H^2(P_m^{\otimes m}, Q_m^{\otimes m}) = 2 - 2\mathbb{E}\sqrt{\prod_{i=1}^m \frac{dP_m}{dQ_m}(X_i)} = 2 - 2\prod_{i=1}^m \mathbb{E}\sqrt{\frac{dP_m}{dQ_m}(X_i)} = 2 - 2\Big(\mathbb{E}\sqrt{\frac{dP_m}{dQ_m}(X_i)}\Big)^m.$$

Recall that $\mathrm{TV}(P_m^{\otimes m}, Q_m^{\otimes m}) \to 0$ if and only if $H^2(P_m^{\otimes m}, Q_m^{\otimes m}) \to 0$, which happens precisely when $H^2(P_m, Q_m) = o(1)$. Similarly, $\mathrm{TV}(P_m^{\otimes m}, Q_m^{\otimes m}) \to 1$ if and only if $H^2(P_m^{\otimes m}, Q_m^{\otimes m}) \to 2$, which is further equivalent to $H^2(P_m, Q_m) = \omega(\frac{1}{m})$. □