Lecture 18: Adaptation lower bounds

Lecturer: Yanjun Han

May 26, 2021

Today's plan

penalty of adaptation:

$$\inf_T \max_{m \geq 1} \frac{\sup_{\theta \in \Theta_m} \mathbb{E}_\theta[L(\theta, T)]}{\inf_{T_m} \sup_{\theta \in \Theta_m} \mathbb{E}_\theta[L(\theta, T_m)]} \gg 1,$$

$$\inf_T \sup_{L \in \mathcal{L}} \frac{\sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, T)]}{\inf_{T_L} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, T_L)]} \gg 1.$$

- constrained risk inequality
- penalty of adaptation to nested parameter sets
- penalty of adaptation to loss functions

Constrained risk inequality

Theorem (Brown and Low, 1996)

If $L(\theta, a) = (\theta - a)^2$ and $\mathbb{E}_{\theta_0}[L(\theta_0, T)] \leq R$, then for every $\theta_1 \in \Theta$,

$$\mathbb{E}_{\theta_1}[L(\theta_1, T)] \geq \left(|\theta_1 - \theta_0| - \sqrt{R(1 + \chi^2(P_{\theta_1}, P_{\theta_0}))} \right)_+^2 \cdot \quad (x)_+ = \max(x, 0)$$

Proof:

$$\begin{aligned} \mathbb{E}_{\theta_1}[|T - \theta_1|] &= \mathbb{E}_{\theta_0}\left[|T - \theta_0| \frac{dP_{\theta_1}}{dP_{\theta_0}}\right] \\ &\leq \mathbb{E}_{\theta_0}[|T - \theta_0|^2]^{1/2} \cdot \mathbb{E}_{\theta_0}\left[\frac{dP_{\theta_1}^2}{dP_{\theta_0}^2}\right]^{1/2} \\ &\leq \sqrt{R} \cdot \sqrt{1 + \chi^2} \end{aligned}$$

$$\Rightarrow \mathbb{E}_{\theta_1}[|T - \theta_1|] \geq |\theta_1 - \theta_0| - \mathbb{E}_{\theta_1}[|T - \theta_0|] \geq |\theta_1 - \theta_0| - \sqrt{R(1 + \chi^2)}$$

$$\Rightarrow \mathbb{E}_{\theta_1}[|T - \theta_1|^2] \geq (\mathbb{E}_{\theta_1}[|T - \theta_1|])^2$$

- idea: if a small risk is achieved at one point, then a large risk will be achieved at another point
- asymmetric version of two-point method
- works for generalized two-point method as well

Generalized constrained risk inequality

Theorem (Generalization of Duchi and Ruan, 2021)

If $\min_{a \in \mathcal{A}}[\sqrt{L(\theta_0, a)} + \sqrt{L(\theta_1, a)}] \geq \Delta$ and $\mathbb{E}_{\theta_0}[L(\theta_0, T)] \leq R$, then

$$\mathbb{E}_{\theta_1}[L(\theta_1, T)] \geq \left(\Delta - \sqrt{R(1 + \chi^2(P_{\theta_1}, P_{\theta_0}))}\right)_+^2.$$

$\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$

$\mathbb{E}_{\theta_1}[\sqrt{L(\theta_0, T)}] \leq \sqrt{R(1+\chi^2)}$ (C-S)

$\mathbb{E}_{\theta_1}[\sqrt{L(\theta_1, T)}] \geq \Delta - \sqrt{R(1+\chi^2)}$

$\mathbb{E}_{\theta_1}[L(\theta_1, T)] \geq \left(\mathbb{E}_{\theta_1}[\sqrt{L(\theta_1, T)}]\right)^2$

Duchi and Ruan (2021): $L(\theta, a) = \ell(\|\theta - a\|_2)$ with ℓ increasing or convex

Example: Hodges' estimator

- Gaussian location model: $X_1, \cdots, X_n \sim \mathcal{N}(\theta, 1)$, with $\theta \in \mathbb{R}$
- Hodges' estimator:

$$\widehat{\theta}_n = \begin{cases} \bar{X} & \text{if } |\bar{X}| \geq n^{-1/4}, \\ 0 & \text{if } |\bar{X}| < n^{-1/4}. \end{cases}$$

- asymptotic normality:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \begin{cases} 0 & \text{if } \theta = 0, \\ \mathcal{N}(0,1) & \text{if } \theta \neq 0. \end{cases}$$

Consequence of superefficiency

For any estimator $\widehat{\theta}_n$, if $\mathbb{P}_0(\sqrt{n}|\widehat{\theta}_n| \geq t) \leq \delta_n$ with $\delta_n \to 0$, then for every $\theta \in [2t/\sqrt{n}, \sqrt{c \log(1/\delta_n)/n}]$ with $c \in (0,1)$, it holds that

$$\mathbb{P}_\theta(\sqrt{n}|\widehat{\theta}_n - \theta| \geq t) \to 1.$$

Applying the constrained risk inequality

- loss function: $L(\theta, a) = 1(\sqrt{n}|\theta - a| \geq t)$
- separation parameter: if $\theta \geq 2t/\sqrt{n}$, then

$$\min_{a \in \mathbb{R}} \left[\sqrt{L(0,a)} + \sqrt{L(\theta,a)} \right] \geq 1$$

- χ^2-divergence:

$$\chi^2(\mathcal{N}(\theta,1)^{\otimes n}, \mathcal{N}(0,1)^{\otimes n}) + 1 = \exp(n\theta^2)$$

$|\theta| \leq \sqrt{\frac{c \log(1/\delta_n)}{n}}$

$\Rightarrow \chi^2 + 1 \leq \delta_n^{-c}$

- constrained risk inequality:

$$\mathbb{E}_\theta[L(\theta, \widehat{\theta}_n)] \geq \left(1 - \sqrt{\delta_n \cdot \exp(n\theta^2)}\right)_+^2 = (1 - \delta_n^{(1-c)/2})^2$$

Adaptation to a nested family of parameter sets

- nested family of parameter sets: $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$
- optimal rate of adaptation:

$$R^\star_{\text{ada}}(\{\Theta_m\}_{m\geq 1}, L) = \inf_T \max_{m\geq 1} \frac{\sup_{\theta\in\Theta_m} \mathbb{E}_\theta[L(\theta, T)]}{\inf_{T_m} \sup_{\theta\in\Theta_m} \mathbb{E}_\theta[L(\theta, T_m)]}$$

- constrained risk inequality is a prominent tool to lower bound $R^\star_{\text{ada}}$
- a rich literature to upper bound $R^\star_{\text{ada}}$ as well

Example I: nonparametric estimation at a point

- $X_1, \cdots, X_n \sim f$ supported on $[0,1]$ with Hölder smoothness s
- target: estimate $f(1/2)$ under loss $L(f, T) = |T - f(1/2)|$
- adaptive estimation: $s \in [0, s_{\max}]$ is unknown

Claim (Brown and Low, 1996)

$$R^\star_{\text{ada}}(n, s_{\max}) \asymp (\log n)^{\frac{s_{\max}}{2s_{\max}+1}}$$

- same two-point construction: $f_0 \equiv 1, f_1(x) = 1 + c_0 \cdot h^s g((x-1/2)/h)$
- constrained risk inequality:

$$R^\star_{\text{ada}} \cdot n^{-\frac{s}{2s+1}} \geq \left(c_1 \sqrt{h^s} - \sqrt{R^\star_{\text{ada}} \cdot n^{-\frac{s_{\max}}{2s_{\max}+1}} \cdot (1 + c_2 h^{2s+1})^n} \right)_+^2$$

- choose $h \asymp (\log n / n)^{1/(2s+1)}$ and $s \approx s_{\max}$

Example II: quadratic functional estimation

non-adaptive: $R^*_{n,s} \asymp n^{-\frac{1}{2}} + n^{-\frac{4s}{4s+1}}$

- $X_1, \cdots, X_n \sim f$ with Hölder smoothness $s > 0$
- target: estimate $Q(f) = \int_0^1 f(x)^2 dx$, with $L(f, T) = |T - Q(f)|$
- adaptive estimation: $s \in [0, s_{\max}]$ is unknown, with $s_{\max} \leq 1/4$

Claim (Efromovich and Low, 1996)

$$R^\star_{\text{ada}}(n, s_{\max}) \asymp (\log n)^{\frac{2s_{\max}}{4s_{\max}+1}}$$

- same two-point construction with χ^2-method giving that

$$1 + \chi^2(\mathbb{E}[f_v^{\otimes n}], f_0^{\otimes n}) \leq \exp(cn^2 h^{4s+1})$$

- constrained risk inequality:

$$R^\star_{\text{ada}} \cdot n^{-\frac{4s}{4s+1}} \geq \left(c_1 \sqrt{h^{2s}} - \sqrt{R^\star_{\text{ada}} \cdot n^{-\frac{4s_{\max}}{4s_{\max}+1}} \cdot \exp(c_2 n^2 h^{4s+1})} \right)_+^2$$

$\Rightarrow h \asymp \left(\frac{\log n}{n^2}\right)^{\frac{1}{4s+1}}$

Example III: adaptive testing with L^2-alternatives

- $X_1, \cdots, X_n \sim f$ with Hölder smoothness $s > 0$
- target: test between

$$H_0 : f \equiv 1 \quad \text{v.s.} \quad H_1 : \|f - 1\|_2 \geq \rho_{n,s}$$

- non-adaptive minimax separation rate:

$$\rho_{n,s} \asymp n^{-\frac{2s}{4s+1}}$$

- adaptive testing: unknown $s \in [0, s_{\max}]$

Claim (Spokoiny, 1998)

Adaptive testing is impossible if

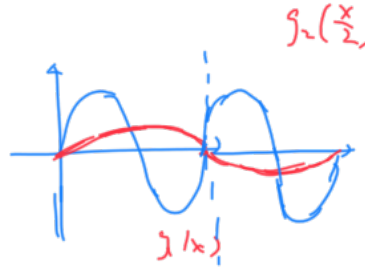
$$\rho_{n,s} = o(\rho^{\star}_{n,s}) = o\left(n^{-\frac{2s}{4s+1}} \cdot (\log\log n)^{\frac{s}{4s+1}}\right)$$

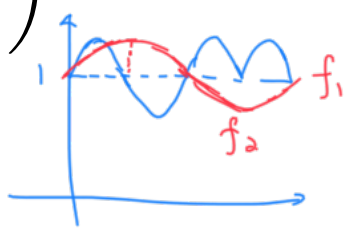
Proof of the claim

- high-level idea: additional layer of mixture over s
- detailed construction: find smooth functions $\{g_j\}_{j\in[m]}$ such that $\{1, g_1(x), g_2(2^{-1}(x+k_2)), \cdots, g_m(2^{-m}(x+k_m))\}$ are orthogonal, for all $j \in [m]$ and $k_j \in \{0, 1, \cdots, 2^j - 1\}$
- the final mixture: with prob. $1/m$, choose the mixture

$$f_{v,j}(x) = 1 + c_0 \cdot \sum_{i=1}^{1/h_j} v_i h_j^{s_j} g_j\left(\frac{x-(i-1)h_j}{h_j}\right)$$

with $h_j = 2^{j-1}h \asymp (\rho^\star_{n,s_j})^{1/s_j}$, and $m \asymp \log n$





- the key observation: $\int_0^1 [f_{v,j}(x) f_{v',j'}(x)/f_0(x)]dx = 1$ if $j \neq j'$
- χ^2-method:

$$1 + \chi^2(\mathbb{E}_{v,j}[f_{v,j}^{\otimes n}], f_0^{\otimes n}) \lesssim \frac{1}{m^2}\sum_{j=1}^{m} \exp(cn^2 h_j^{4s_j+1})$$

doesn't depend on j.

Adaptation to different loss functions

- an important special case:

$$R^\star(\Theta, \mathcal{L}) = \inf_T \sup_{\theta \in \Theta} \sup_{L \in \mathcal{L}} \mathbb{E}_\theta[L(\theta, T)]$$

- a typical target:

$$\inf_T \sup_{\theta \in \Theta} \sup_{L \in \mathcal{L}} \mathbb{E}_\theta[L(\theta, T)] \gg \sup_{L \in \mathcal{L}} \inf_T \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, T)]$$

- modification of the testing idea: find $\theta_1, \cdots, \theta_M \in \Theta$ and $L_1, \cdots, L_M \in \mathcal{L}$ with the same indistinguishability condition and a new separation condition: for all $i \neq j$,

$$\inf_a [L_i(\theta_i, a) + L_j(\theta_j, a)] \geq \Delta.$$

A toy example

- model: $X_1, \cdots, X_n \sim \text{Bern}(p)$ with unknown $p \in [0, 1]$
- loss: $L_1(p, T) = |T - p|, L_2(p, T) = |T - (1 - p)|$
- claim:

$$R_n^\star(\{L_1, L_2\}) = \frac{1}{2}$$

- proof: choosing $p_1 = p_2 = 0$, then TV $= 0$, while

$$\inf_a [L_1(p_1, a) + L_2(p_2, a)] = 1.$$

A non-toy example

- model: $X_1, \cdots, X_n \sim p = (p_1, \cdots, p_k)$
- family of loss functions $\mathcal{L}_{\text{Lip}}$: given a 1-Lipschitz function f,

$$L_f(p, \widehat{p}) = \left| \sum_{i=1}^{k} f(\widehat{p}_i) - \sum_{i=1}^{k} f(p_i) \right| \in \mathcal{L}_{\text{Lip}}$$

Claim (Han, 2021)

$$R_{n,k}(\mathcal{L}_{\text{Lip}}) \asymp \begin{cases} \sqrt{\frac{k}{n \log n}} & \text{if } n^{1/3} \ll k \lesssim n \log n, \\ \sqrt{\frac{k}{n}} & \text{if } k \ll n^{1/3}. \end{cases}$$

- background: unified approaches for functional estimation
- strict penalty of unified approaches compared with ad-hoc approaches:

$$\sup_{L \in \mathcal{L}_{\text{Lip}}} \inf_{\widehat{p}} \sup_{p} \mathbb{E}_p[L(p, \widehat{p})] \asymp \sqrt{\frac{k}{n \log n}}, \quad \log n \lesssim k \lesssim n \log n$$

References

- Lawrence D Brown and Mark G Low. "A constrained risk inequality with applications to nonparametric functional estimation." *The Annals of Statistics*, 24(6):2524–2535, 1996.
- John C Duchi and Feng Ruan. "A constrained risk inequality for general losses." *AISTATS*, 2021.
- Sam Efromovich and Mark Low. "On optimal adaptive estimation of a quadratic functional." *The Annals of Statistics*, 24(3):1106–1125, 1996.
- Vladimir Spokoiny. "Adaptive and spatially adaptive testing of a nonparametric hypothesis." *Math. Methods Statist.*, 7:245–273, 1998.
- Yanjun Han. "On the High Accuracy Limitation of Adaptive Property Estimation." *AISTATS*, 2021.

Recap of the course

- understand why a given problem is difficult - assumption, target, information structure, etc.
- rigorously argue the difficulty based on the intuition
- statistical formulation: distribution class, parameter set, loss
- idea of reduction: f-divergence, deficiency, equivalence
- idea of testing: two points (Le Cam, χ^2-method, moment matching), multiple points (Fano, Assouad, global Fano)
- other ideas: compression (of models and algorithms), experimental design (SDPI, quantized Fisher information, direct modeling), geometry (information theory, high-dimensional probability), online or sequential experiment (order of the game), adaptation (constrained risk inequality, and others), etc.

Next week: project presentations!