Lecture 11: Examples of testing multiple hypotheses

Lecturer: Yanjun Han

May 3, 2021

Announcements

- scribe required starting from the next lecture
- one student per lecture
- sign-up link: `https://bit.ly/31quUIb`
- choice of literature review paper due Sunday
- submit your choice via gradescope or email

Today's plan

Examples of testing multiple hypotheses

- key: how to construct different hypotheses for a given problem
- example I: nonparametric density estimation
- example II: learning theory
- example III: theory of aggregation
- example IV: stochastic optimization

Example I: nonparametric density estimation

An overview of different nonparametric estimation problems:

- estimating the density at a point (two-point method)
- estimating the quadratic functional (point vs. mixture)
- estimating a non-smooth functional (mixture vs. mixture)
- estimating the global density (multiple hypotheses testing)

Density estimation over Sobolev space

- model: $X_1, \cdots, X_n \sim f$ supported on $[0,1]$
- smoothness assumption: f belongs to a Sobolev ball $\mathcal{W}^{k,p}(L)$:

$$\mathcal{W}^{k,p}(L) = \{f \in C[0,1] : \|f\|_p + \|f^{(k)}\|_p \leq L\}$$

- target: estimate density f under L_q norm, i.e. $L(f, T) = \|f - T\|_q$

Claim

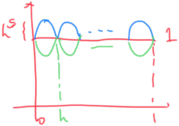
dense case

$$R^{\star}_{n,k,p,q} \asymp \begin{cases} n^{-k/(2k+1)} & \text{if } q < (1+2k)p, \\ (\log n/n)^{(k-1/p+1/q)/(2(k-1/p)+1\cancel{/q})} & \text{if } q \geq (1+2k)p. \end{cases}$$

sparse case

Dense case: $q < (1+2k)p$

- w.l.o.g assume that $q = 1$
- hypothesis construction: for $v \in \{\pm 1\}^{1/h}$, choose



$$f_v(x) = 1 + \sum_{i=1}^{h^{-1}} v_i \cdot h^s \underbrace{g}_{\int g = 0}\left(\frac{x-(i-1)h}{h}\right)$$

- smoothness requirement: $s = k$

$f_v^{(k)}(x) = \sum_i v_i \cdot h^{s-k} g^{(k)}\left(\frac{x-(i-1)h}{h}\right)$ $\quad$ $\|f^{(k)}\|_p \le L \Downarrow s \ge k$

- separation condition: $\Delta \asymp h^{s+1}$ in Assouad's lemma
- neighboring χ^2-divergence:

$\|f_v - f_{v'}\|_1 \propto d_H(v, v')$

$\Delta \asymp \int_0^h h^s |g(\frac{x}{h})| dx = h^{s+1}$

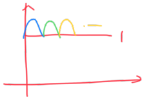
$$\max_{d_H(v,v')=1} \chi^2(f_v^{\otimes n}, f_{v'}^{\otimes n}) \lesssim n\|f_v - f_{v'}\|_2^2 \lesssim nh^{2s+1}$$

- application of Assouad's lemma:

$h = n^{-\frac{1}{2k+1}}$ $(s = k)$

$$R^\star_{n,k,p,q} \gtrsim \frac{1}{h} \cdot h^{s+1}(1 - O(\sqrt{nh^{2s+1}}))$$

Sparse case: $q \geq (1+2k)p$



- hypothesis construction: for $v \in [h^{-1}]$, choose

$$f_v(x) = 1 + h^s g\left(\frac{x-(v-1)h}{h}\right) \cdot 1(x \in [(v-1)h, vh))$$

- smoothness requirement: $s = k - 1/p$

$f_v^{(k)}(x) = h^{s-k} g^{(k)}\left(\frac{x-(v-1)h}{h}\right)$

$\|f_v^{(k)}\|_p \lesssim h^{s-k} \cdot h^{1/p}$

- separation condition: $\Delta \asymp h^{s+1/q}$ in Fano's method
- mutual information:

$\Delta \asymp h^s \|g(\frac{x}{h})\|_q \asymp h^s \cdot h^{1/q}$

$$I(V;X) \leq \max_{v \neq v'} D_{\mathrm{KL}}(f_v^{\otimes n} \| f_{v'}^{\otimes n}) \lesssim nh^{2s+1}$$

- application of Fano's inequality:

$h = \left(\frac{\log n}{n}\right)^{\frac{1}{2(k-1/p)+1}}$

$$R^\star_{n,k,p,q} \gtrsim h^{s+1/q}\left(1 - \frac{nh^{2s+1} + \log 2}{\log(1/h)}\right)$$

Example II: learning theory

- model: $(x_1, y_1), \cdots, (x_n, y_n) \sim P_{XY}$ with $\mathcal{Y} = \{-1, 1\}$
- assumption: a given function class $\mathcal{F}$ with VC dimension d
- target: find classifier $\widehat{f}$ with a small excess risk:

$$L_{\mathcal{F}}(P_{XY}, \widehat{f}) = P_{XY}(Y \neq \widehat{f}(X)) - \min_{f^\star \in \mathcal{F}} P_{XY}(Y \neq f^\star(X))$$

Definition (VC dimension)

The VC dimension of $\mathcal{F}$ is the largest integer d such that there are d points in $\mathcal{X}$ that can be shattered by $\mathcal{F}$. In other words, for all $v \in \{\pm 1\}^d$ there exists a function $f_v \in \mathcal{F}$ such that

$$f_v(x_i) = v_i, \qquad \forall i \in [d].$$

Optimistic vs. pessimistic case

- classical VC theory distinguishes into two cases depending on whether $Y = f^\star(X)$ for some $f^\star \in \mathcal{F}$
- well-specified (optimistic) case:

$$\mathcal{P}_{\text{opt}}(\mathcal{F}) = \{P_{XY} : \boxed{\min_{f \in \mathcal{F}} P_{XY}(Y \neq f^\star(X)) = 0}\}$$

- misspecified (pessimistic) case: no assumption on P_{XY}

Claim

$$R^\star_{\text{opt}} = \inf_{\widehat{f}} \sup_{P_{XY} \in \mathcal{P}_{\text{opt}}(\mathcal{F})} \mathbb{E}[L(P_{XY}, \widehat{f})] \asymp \min\{d/n, 1\}$$

$$R^\star_{\text{pes}} = \inf_{\widehat{f}} \sup_{P_{XY}} \mathbb{E}[L(P_{XY}, \widehat{f})] \asymp \min\left\{\sqrt{d/n}, 1\right\}$$

Optimistic case

- VC dimension: $\exists x_1, \cdots, x_d \in \mathcal{X}$ and $f_v \in \mathcal{F}$ such that $f_v(x_i) = v_i$ for all $v \in \{\pm 1\}^d$
- hypothesis construction: for $u \in \{\pm 1\}^{d-1}$, define

$$P_X(\{x_i\}) = p_0, \quad i \in [d-1], \qquad P_X(\{x_d\}) = \underbrace{1-(d-1)p_0}_{\geq 0}$$

  and $Y \overset{\text{a.s.}}{=} f_{(u,1)}(X)$ under hypothesis P_u

- separation condition: $\Delta = p_0$ in Assouad's lemma
- neighboring TV distance: $P_u(Y \neq \hat{f}(X)) + P_{u'}(Y \neq \hat{f}(X)) \geq p_0 \cdot d_H(u, u')$

$$\max_{d_{\mathrm{H}}(u,u')=1} \|P_u^{\otimes n} - P_{u'}^{\otimes n}\|_{\mathrm{TV}} = (1-p_0)^n$$

- application of Assouad's lemma:

$p_0 = \frac{1}{n} \wedge \frac{1}{d-1}$

$$R^\star_{\mathrm{opt}} \geq \frac{(d-1)p_0}{2}(1-(1-p_0)^n)$$

Pessimistic case

- VC dimension: $\exists x_1, \cdots, x_d \in \mathcal{X}$ and $f_v \in \mathcal{F}$ such that $f_v(x_i) = v_i$ for all $v \in \{\pm 1\}^d$
- hypothesis construction: for $v \in \{\pm 1\}^d$, let P_X be the uniform distribution on $\{x_1, \cdots, x_d\}$, and

$$P_v(Y = v_i \mid X = x_i) = \frac{1}{2} + \delta$$

- separation condition: $\Delta = 2\delta/d$ in Assouad's lemma

$\min_{f^* \in \mathcal{F}} P_{XY}(f^*(X) \neq Y) = \frac{1}{2} - \delta$

$P^v_{XY}(\hat{f}(X) \neq Y) + P^{v'}_{XY}(\hat{f}(X) \neq Y) \geq 2\left(\frac{1}{2} \cdot d_H(v,v') + (\frac{1}{2} - \delta)\right.$.
$(d - d_H(v,v'))$

- neighboring KL divergence:

$$\max_{d(v,v')=1} D_{\mathrm{KL}}(P_v^{\otimes n} \| P_{v'}^{\otimes n}) \lesssim \frac{n}{d} \cdot \delta^2$$

- application of Assouad's lemma:

$$R^\star_{\mathrm{pes}} \geq \delta \left(1 - O(\sqrt{n\delta^2/d})\right)$$

$\delta = \sqrt{\frac{d}{n}} \wedge \frac{1}{4}$

Generalization

- an intermediate regime: define

$$\mathcal{P}(\mathcal{F}, \varepsilon) = \left\{ P_{XY} : \inf_{f^\star \in \mathcal{F}} P_{XY}(Y \neq f^\star(X)) \leq \varepsilon \right\}.$$

- the minimax excess risk:

$$R^\star(\mathcal{F}, \varepsilon) = \inf_{\widehat{f}} \sup_{P_{XY} \in \mathcal{P}(\mathcal{F}, \varepsilon)} \mathbb{E}\left[P_{XY}(Y \neq \widehat{f}(X)) - \inf_{f^\star \in \mathcal{F}} P_{XY}(Y \neq f^\star(X)) \right]$$

Claim (HW3)

$$R^\star(\mathcal{F}, \varepsilon) \asymp \min\left\{ \sqrt{\frac{d}{n} \cdot \varepsilon} + \frac{d}{n}, 1 \right\}.$$

Example III: theory of aggregation

- model: $x_1, \cdots, x_n \sim P_X$, and $y_i \sim \mathcal{N}(f(x_i), 1)$ with $\|f\|_\infty \leq 1$
- assumption: a candidate set of functions $\mathcal{F} = \{f_1, \cdots, f_M\}$
- loss function:

$$L(f, \widehat{f}) = \|\widehat{f} - f\|_{L_2(P_X)}^2 - \inf_{\lambda \in \Theta} \|f_\lambda - f\|_{L_2(P_X)}^2$$

  with $f_\lambda = \sum_{i=1}^M \lambda_i f_i$
- target: characterize the minimax rate of aggregation

$$R_{n,M}^\star(\Theta) = \sup_{\mathcal{F}} \inf_{\widehat{f}} \sup_{\|f\|_\infty \leq 1} \mathbb{E}_f[L(f, \widehat{f})]$$

Different types of aggregation

- $\Theta = \mathbb{R}^M$: linear aggregation
- $\Theta = \{\lambda \in \mathbb{R}_+^M : \sum_{i=1}^M \lambda_i \leq 1\}$: convex aggregation
- $\Theta = \{e_1, \cdots, e_M\}$: model selection aggregation

Claim

$$R^\star_{n,M}(\mathrm{L}) \asymp \min\{M/n, 1\}$$

$$R^\star_{n,M}(\mathrm{C}) \asymp \min\left\{M/n, \sqrt{\log(M/\sqrt{n}+1)/n}, 1\right\}$$

$$R^\star_{n,M}(\mathrm{MS}) \asymp \min\{(\log M)/n, 1\}$$

Linear aggregation

- hypothesis construction: for $v \in \{\pm 1\}^M$, choose

$$f_v(x) = \gamma \cdot \sum_{i=1}^{M} v_i f_i(x) \qquad \|f_v\|_\infty = \gamma \le 1$$

  with $f_i(x) = 1(x \in \mathcal{X}_i)$, where $\mathcal{X}_i$ disjoint and has P_X-prob. $1/M$
- separation condition: $\Delta = 2\gamma^2/M$ in Assouad's lemma
- neighboring KL divergence:

$$\max_{d_H(v,v')=1} D_{KL}(f_v^{\otimes n} \| f_{v'}^{\otimes n}) \lesssim n\|f_v - f_{v'}\|_{L_2(P_X)}^2 \lesssim \frac{n\gamma^2}{M}$$

- application of Assouad's lemma:

$$\gamma = \sqrt{\frac{M}{n}} \wedge 1$$

$$R_{n,M}^\star(\mathrm{L}) \ge \gamma^2 \left(1 - O\left(\sqrt{n\gamma^2/M}\right)\right)$$

Model selection aggregation

- hypothesis construction: for $v \in [M]$, choose $f_v(x) = \gamma \cdot \phi_v(x)$, with orthonormal $\{\phi_v\}$ on $L_2(\mathcal{X})$ with $\|\phi_v\|_\infty = O(1)$
- separation condition: $\Delta \asymp \gamma^2$ in Fano's inequality

  $\|f_v - f_{v'}\|_2^2 \asymp \gamma^2$
- mutual information:

$$I(V;X) \le \max_{v \neq v'} D_{\mathrm{KL}}(f_v^{\otimes n} \| f_{v'}^{\otimes n}) \lesssim n\gamma^2$$

- application of Fano's inequality:

$$R^\star_{n,M}(\mathrm{MS}) \gtrsim \gamma^2 \left(1 - \frac{O(n\gamma^2) + \log 2}{\log M}\right)$$

$\gamma^2 = \frac{\log M}{n} \wedge 1$

Convex aggregation

- suffices to consider the case $M = \Omega(\sqrt{n})$
- hypothesis construction: for $v \in \{0, 1/m\}^M$ with m non-zero entries, choose

$$f_v(x) = \gamma \cdot \sum_{i=1}^{M} v_i f_i(x)$$

  with $f_i(x) = \phi_i(x)$ being orthonormal, and $\gamma \asymp 1$
- separation condition: choosing $\Delta \asymp 1/m$, we have

$$\log(1/p_\Delta) = \Omega(m \log(1 + M/m))$$

- mutual information:

$$I(V; X) \le \frac{n}{2} \cdot \mathbb{E}_v[\|f_v\|_{L_2(P_X)}^2] \lesssim \frac{n}{m}$$

- application of generalized Fano's inequality:

$$R_{n,M}^\star(\mathrm{C}) \gtrsim \frac{1}{m}\left(1 - \frac{O(n/m) + \log 2}{m \log(1 + M/m)}\right)$$

$m = \sqrt{\frac{n}{\log(1+\frac{M}{\sqrt{n}})}}$

Example IV: stochastic optimization

- model: at each time $t \in [T]$,
  - learner queries x_t with $\|x_t\|_p \le 1$
  - oracle returns y_t, z_t with $\mathbb{E}[y_t] = f(x_t), \mathbb{E}[z_t] = \nabla f(x_t)$, and $\|z_t\|_q \le 1$
  - q is the conjugate of p: $p^{-1} + q^{-1} = 1$
- function class $\mathcal{F}$: f convex, with $\|\nabla f(x)\|_q \le 1$ everywhere
- loss function:

$$L(f, \widehat{x}) = f(\widehat{x}) - \min_{\|x^\star\|_p \le 1} f(x^\star)$$

- minimax optimality gap of stochastic optimization:

$$R^\star_{T,d,p} = \inf_{\widehat{x}} \sup_{f \in \mathcal{F}} \sup_{\mathcal{O}_p} \mathbb{E}_{f,\mathcal{O}_p}[L(f, \widehat{x})]$$

Claim

$$R^\star_{T,d,p} \asymp \begin{cases} T^{-1/2} & \text{if } 1 \le p \le 2, \\ \min\{T^{-1/p}, d^{1/2-1/p} T^{-1/2}\} & \text{if } p > 2. \end{cases}$$

Indistinguishability condition

- idea: choose $f(x) = \mathbb{E}_\xi[F(x;\xi)]$, and $y_t = F(x_t;\xi_t), z_t = \nabla F(x_t;\xi_t)$
- hypothesis construction: for $v \in \{\pm 1\}^d$, choose $f_v(x) = \mathbb{E}_{P_v}[F(x;\xi)]$:

$$P_v(\xi = e_i) = \frac{1+\delta v_i}{2d}, \quad P_v(\xi = -e_i) = \frac{1-\delta v_i}{2d}, \qquad i \in [d].$$

- neighboring KL divergence:

$$\max_{d_{\mathrm{H}}(v,v')=1} D_{\mathrm{KL}}(P_v^{\otimes T} \| P_{v'}^{\otimes T}) \lesssim \frac{T\delta^2}{d}$$

Separation condition

- condition on the optimization distance:

$$\min_x f_v(x) + \min_x f_{v'}(x) - \min_x (f_v(x) + f_{v'}(x)) \geq \Delta \cdot d_{\mathrm{H}}(v, v')$$

- choice of F:

$$F(x;\xi) = |x_i - \lambda \xi_i|, \quad \text{if } \xi = \pm e_i$$

$$\Longrightarrow f_v(x) = \lambda - \frac{\delta}{d} \sum_{i=1}^{d} v_i x_i, \quad \text{if } \|x\|_\infty \leq \lambda.$$

- choice of λ: $\lambda = d^{-1/p}$, so that $\min_x f_v(x) = (1-\delta) d^{-1/p}$
- separation condition: $\Delta \asymp \delta\lambda/d$ in Assouad's lemma
- application of Assouad's lemma:

$\delta = \sqrt{\frac{d}{T}} \quad (d \leq T)$

$$R^\star_{T,d,p} \gtrsim \delta d^{-1/p} \left(1 - O(\delta\sqrt{T/d})\right)$$

References

- Arkadi Nemirovski. "Topics in non-parametric statistics." *Ecole d'Eté de Probabilités de Saint-Flour* 28 (2000): 85.
- Vlamimir Vapnik. "Statistical learning theory." Wiley, New York (1998): 156–160.
- Alexandre B. Tsybakov, "Optimal rates of aggregation." *Learning theory and kernel machines*. Springer, Berlin, Heidelberg, 2003. 303–313.
- Olivier Bousquet, Daniel Kane, and Shay Moran. "The optimal approximation factor in density estimation." In *Conference on Learning Theory*, 2019.
- Alekh Agarwal, Peter L. Bartlett, Pradeep K. Ravikumar, and Martin J. Wainwright. "Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization." *IEEE Transactions on Information Theory*, 5(58):3235–3249.

Next lecture: Global Fano's method