Lec 10 : Entropic upper bounds of density estimation

Yanjun Han

Last lecture: use covering/packing to prove statistical lower bound via Fano

This lecture: they can also prove upper bounds!

Density estimation: $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P$, where $P \in \mathcal{P}$ is an unknown distribution

target: for $D \in \{D_{KL}, TV, H^2\}$, construct an estimator $\hat{P} = \hat{P}(X^n)$ s.t.

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{X^n \sim P^{\otimes n}}[D(P, \hat{P})] \quad \text{is small.}$$

Overview of results.

- KL (Yang-Barron): $\exists \hat{P}$ s.t.

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[D_{KL}(P \| \hat{P})] \lesssim \inf_{\varepsilon > 0} \left(\varepsilon^2 + \frac{1}{n} \log N_{KL}(\mathcal{P}, \varepsilon)\right)$$

- TV (Yatracos): $\exists \hat{P}$ s.t.

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[TV^2(P, \hat{P})] \lesssim \inf_{\varepsilon > 0} \left(\varepsilon^2 + \frac{1}{n} \log N_{TV}(\mathcal{P}, \varepsilon)\right)$$

- H^2 (Le Cam-Birgé): $\exists \hat{P}$ s.t.

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[H^2(P, \hat{P})] \lesssim \inf_{\varepsilon > 0} \left(\varepsilon^2 + \frac{1}{n} \log N_H(\mathcal{P}, \varepsilon)\right).$$

Example 1. For finite-dimensional models $\mathcal{P}$ with d parameters, usually $N_D(\mathcal{P}, \varepsilon) \asymp d \log \frac{1}{\varepsilon}$. (volume bound)

In this case. $\inf_{\hat{P}} \sup_{P} \mathbb{E}_P[D(P, \hat{P})] \lesssim \inf_{\varepsilon > 0} \left(\varepsilon^2 + \frac{d}{n} \log \frac{1}{\varepsilon}\right) \lesssim \frac{d \log n}{n}$.

usually optimal up to $\log n$ factor.

Example 2. For nonparametric classes $\mathcal{P}$ with $N_D(\mathcal{P}, \varepsilon) \asymp \varepsilon^{-\alpha}$.

$$\inf_{\hat{P}} \sup_{P} \mathbb{E}_P[D(P, \hat{P})] \lesssim \inf_{\varepsilon > 0} \left(\varepsilon^2 + \frac{1}{n\varepsilon^\alpha}\right) \lesssim n^{-\frac{2}{2+\alpha}}.$$

Yang-Barron: progressive mixing / online-to-batch conversion

"Online" guarantee: similar to global Fano. Let $P_1, \cdots, P_N$ be an ε-covering of $\mathcal{P}$, i.e.

$$\sup_{P \in \mathcal{P}} \min_{i \in [N]} D_{KL}(P \| P_i) \le \varepsilon^2.$$

Let $Q_{X^{n+1}}$ be the average product distribution:

$$Q_{X^{n+1}} = \frac{1}{N} \sum_{i=1}^N P_i^{\otimes(n+1)}.$$

Lemma. $$\sup_{P \in \mathcal{P}} D_{KL}(P^{\otimes(n+1)} \| Q_{X^{n+1}}) \le (n+1)\varepsilon^2 + \log N.$$

Pf. Similar to global Fano: for any $P \in \mathcal{P}$,

$$\begin{aligned} D_{KL}(P^{\otimes(n+1)} \| Q_{X^{n+1}}) &= \mathbb{E}_{X^{n+1} \sim P^{\otimes(n+1)}}\left[\log \frac{P^{\otimes(n+1)}(X^{n+1})}{\frac{1}{N}\sum_{i=1}^N P_i^{\otimes(n+1)}(X^{n+1})}\right] \\ &\le \mathbb{E}_{X^{n+1} \sim P^{\otimes(n+1)}}\left[\min_{i \in [N]} \log \frac{P^{\otimes(n+1)}(X^{n+1})}{P_i^{\otimes(n+1)}(X^{n+1})} + \log N\right] \\ &\le \min_{i \in [N]} D_{KL}(P^{\otimes(n+1)} \| P_i^{\otimes(n+1)}) + \log N \\ &\le (n+1)\varepsilon^2 + \log N. \end{aligned}$$

□

This is called an "online" guarantee as it concerns the density estimation performance for joint distributions of $X_1, \cdots, X_{n+1} \overset{i.i.d.}{\sim} P$.

"Online-to-batch" conversion: given $Q_{X^{n+1}}$, we can define

$$\hat{P}(x) = \frac{1}{n+1} \sum_{t=0}^n Q_{X_{t+1}=x | X^t}.$$

Note that $\hat{P}$ is a well-defined estimator and depends on X^n. Expanding out the defn. of $Q_{X^{n+1}}$ gives

$$\hat{P}(x) = \frac{1}{n+1} \sum_{t=0}^n \frac{\frac{1}{N}\sum_{i=1}^N \left(\prod_{s \le t} P_i(X_s)\right) P_i(x)}{\frac{1}{N}\sum_{i=1}^N \prod_{s \le t} P_i(X_s)} \in \text{conv}(\mathcal{P})$$

"progressive mixing"

Yang-Barron result follows from the next result:

Lemma. $\mathbb{E}_P[D_{KL}(P \| \hat{P})] \le \frac{1}{n+1} D_{KL}(P^{\otimes(n+1)} \| Q_{X^{n+1}})$.

Pf.

$$\mathbb{E}_P[D_{KL}(P \| \hat{P})] = \mathbb{E}_P\left[D_{KL}\left(P \| \frac{1}{n+1}\sum_{t=0}^{n} Q_{X_{t+1}|X^t}\right)\right]$$

$$\le \frac{1}{n+1}\sum_{t=0}^{n} \mathbb{E}_P[D_{KL}(P \| Q_{X_{t+1}|X^t})] \quad \text{(convexity)}$$

$$= \frac{1}{n+1} D_{KL}(P^{\otimes(n+1)} \| Q_{X^{n+1}}) \quad \text{(chain rule)}$$

□

Remarks: ① This online-to-batch conversion provides a general paradigm for converting "redundancy" bound to prediction risk bound, even beyond i.i.d. data; see more next lecture.

② The Yang-Barron estimator is often improper (i.e $\hat{P} \in \text{conv}(\mathcal{P})$ but often $\hat{P} \notin \mathcal{P}$), and computationally hard to obtain.

Yatracos: minimum distance estimator

The TV density estimation result is a corollary of the following general result in the robust case.

Thm. Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} P$, and $Q_1, \cdots, Q_N$ be arbitrary candidate distributions. Then there exists an estimator $\hat{P}$ s.t.

$$TV(P, \hat{P}) \le 3 \min_{i \in [N]} TV(P, Q_i) + \varepsilon_n, \text{ with } \mathbb{E}[\varepsilon_n^2] = O\left(\frac{\log N}{n}\right).$$

Note that by choosing $\{Q_1, \cdots, Q_N\}$ be an ε-covering of $\mathcal{P}$ under TV, with $N = N_{TV}(\mathcal{P}, \varepsilon)$, this implies the density estimation result.

Next we prove the theorem using a minimum-distance estimator:

$$\hat{P} = \underset{Q \in \{Q_1, \dots, Q_N\}}{\operatorname{argmin}} \widetilde{TV}(P_n, Q),$$

where $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ is the empirical distribution, and $\widetilde{TV}$ is a psendo-distance.

(What if $\widetilde{TV} = TV$? If $Q_1, \dots, Q_N$ are all continuous distributions, since P_n is discrete. $TV(P_n, Q_i) = 1$ for all i, so it's not useful.)

Let's defer the choice of $\widetilde{TV}$ and proceed to the analysis. Let

$$Q^* = \underset{Q \in \{Q_1, \dots, Q_N\}}{\operatorname{argmin}} TV(P, Q).$$

Then

$$\begin{aligned}
TV(\hat{P}, P) &\le TV(\hat{P}, Q^*) + TV(Q^*, P) \\
&\overset{\text{hope}}{=} \widetilde{TV}(\hat{P}, Q^*) + TV(Q^*, P) \\
&\le \widetilde{TV}(\hat{P}, P_n) + \widetilde{TV}(P_n, Q^*) + TV(Q^*, P) \\
&\le 2\widetilde{TV}(P_n, Q^*) + TV(Q^*, P) \quad \text{(definition of } \hat{P}\text{)} \\
&\le 2\widetilde{TV}(P_n, P) + 2\widetilde{TV}(P, Q^*) + TV(P, Q^*) \\
&\overset{\text{hope}}{\le} 2\widetilde{TV}(P_n, P) + 3\,TV(P, Q^*).
\end{aligned}$$

To make the analysis go through, we need:

① $\widetilde{TV}(P, Q) \le TV(P, Q)$. $\quad \forall P, Q$

② $\widetilde{TV}(Q_i, Q_j) = TV(Q_i, Q_j)$. $\quad \forall i, j \in [n]$

③ $\mathbb{E}[\widetilde{TV}(P_n, P)^2]$ is small.

Motivated by ① + ②, we define

$$\widetilde{TV}(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|,$$

where $\mathcal{A} = \{A_{ij} : i, j \in [N]\}$ with

$$A_{ij} := \{x : Q_i(x) \ge Q_j(x)\}.$$

Then: ① is immediate as $TV(P, Q) = \sup_A |P(A) - Q(A)|$

② is also true as $TV(Q_i, Q_j) = |Q_i(A_{ij}) - Q_j(A_{ij})| \le \widetilde{TV}(Q_i, Q_j)$

③ notes that $|\mathcal{A}| \le \binom{N}{2}$, and for fixed A,

$$\mathbb{P}(|P(A) - P_n(A)| > \varepsilon) \le 2\exp(-2n\varepsilon^2) \quad \text{by Hoeffding.}$$

Therefore, a union bound over $\mathcal{A}$ gives

$$\mathbb{P}(\widetilde{TV}(P, P_n) > \varepsilon) \le 2N^2 \exp(-2n\varepsilon^2)$$

$$\Rightarrow \mathbb{E}[\widetilde{TV}^2(P, P_n)] = \int_0^\infty \mathbb{P}(\widetilde{TV}^2(P, P_n) \ge r)\, dr$$

$$\le \int_0^\infty \min\{1, 2N^2 e^{-2nr}\}\, dr$$

$$\le \frac{\log(2N^2)}{2n} + \int_{\frac{\log(2N^2)}{2n}}^\infty 2N^2 e^{-2nr}\, dr$$

$$= O\left(\frac{\log N}{n}\right).$$

Remark: ① The Yatracos estimator is proper, i.e. $\hat{P} \in \mathcal{P}$.

② The above proof also shows a high-probability guarantee on $TV(\hat{P}, P)$

③ It's known that the constant 3 is not improvable if the estimator is required to be proper.

④ There are some recent interests in the computationally efficient version of Yatracos.

Le Cam-Birgé: pairwise comparison

Composite hypothesis testing. H_0: $X_1, \dots, X_n \sim P$ with $P \in \mathcal{P}$

H_1: $X_1, \dots, X_n \sim Q$ with $Q \in \mathcal{Q}$

test: $T = T(X^n) \in \{0, 1\}$

type-I error: $\sup_{P \in \mathcal{P}} P^{\otimes n}(T = 1)$

type-II error: $\sup_{Q \in \mathcal{Q}} Q^{\otimes n}(T = 0)$

Lemma. $\inf_T \left(\sup_{P \in \mathcal{P}} P^{\otimes n}(T=1) + \sup_{Q \in \mathcal{Q}} Q^{\otimes n}(T=0) \right) \le e^{-\frac{n}{2} H^2(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q}))}$.

where $H^2(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q})) := \inf_{\substack{P \in \mathrm{conv}(\mathcal{P}) \\ Q \in \mathrm{conv}(\mathcal{Q})}} H^2(P, Q)$.

Pf. In Lec 8 we know that

$$\mathrm{LHS} = 1 - TV(\mathrm{conv}(\mathcal{P}^{\otimes n}), \mathrm{conv}(\mathcal{Q}^{\otimes n})) \quad \left(\mathcal{P}^{\otimes n} = \{P^{\otimes n} : P \in \mathcal{P}\}, \; TV(\mathcal{P}, \mathcal{Q}) = \inf_{\substack{P \in \mathcal{P} \\ Q \in \mathcal{Q}}} TV(P, Q) \right)$$

$$\le 1 - \frac{H^2}{2}(\mathrm{conv}(\mathcal{P}^{\otimes n}), \mathrm{conv}(\mathcal{Q}^{\otimes n})) \quad (TV \ge \tfrac{H^2}{2})$$

$$\le \left(1 - \frac{H^2}{2}(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q}))\right)^n \quad (\text{next lemma})$$

$$\le \exp\left(-\tfrac{n}{2} H^2(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q}))\right) \qquad \square$$

Lemma. $1 - \frac{H^2}{2}\left(\mathrm{conv}\left(\bigotimes_{i=1}^n P_i\right), \mathrm{conv}\left(\bigotimes_{i=1}^n Q_i\right)\right) \le \prod_{i=1}^n \left(1 - \frac{H^2}{2}(\mathrm{conv}(P_i), \mathrm{conv}(Q_i))\right)$.

Pf. Suffice to prove the case $n=2$. Note that

$$1 - \frac{H^2}{2}(P, Q) = 1 - \frac{1}{2}\int (\sqrt{P} - \sqrt{Q})^2 = \int \sqrt{PQ}.$$

and any $P_{XY} \in \mathrm{conv}(\mathcal{P}_1 \otimes \mathcal{P}_2)$ can be written as $P_{XY} = \mathbb{E}_Z[P_{X|Z} P_{Y|Z}]$ with $P_{X|Z} \in \mathcal{P}_1$, $P_{Y|Z} \in \mathcal{P}_2$. Then

$$1 - \frac{H^2}{2}(P_{XY}, Q_{XY}) = \int \sqrt{P_{XY} Q_{XY}}$$

$$= \int_x \sqrt{P_X Q_X} \int_y \sqrt{P_{Y|X} Q_{Y|X}}$$

$$\left(= \mathbb{E}_{Z|X}[P_{Y|Z}] \in \mathrm{conv}(\mathcal{P}_2) \right)$$

$$\le \int_x \sqrt{P_X Q_X} \cdot \left(1 - \frac{H^2}{2}(\mathrm{conv}(\mathcal{P}_2), \mathrm{conv}(\mathcal{Q}_2))\right)$$

$$\left(= \mathbb{E}_Z[P_{X|Z}] \in \mathrm{conv}(\mathcal{P}_1) \right)$$

$$\le \left(1 - \frac{H^2}{2}(\mathrm{conv}(\mathcal{P}_1), \mathrm{conv}(\mathcal{Q}_1))\right)\left(1 - \frac{H^2}{2}(\mathrm{conv}(\mathcal{P}_2), \mathrm{conv}(\mathcal{Q}_2))\right). \qquad \square$$

(Note: the same proof holds for all Rényi divergences $D_\alpha = \frac{1}{\alpha-1}\log\int P^\alpha Q^{1-\alpha}$.)

This lemma will be applied in the following setting:

$$H_0: \quad X_1, \cdots, X_n \sim P, \qquad P \in B_H(P_0, \varepsilon) = \{P: H^2(P, P_0) \le \varepsilon^2\}$$
$$H_1: \quad X_1, \cdots, X_n \sim Q, \qquad Q \in B_H(Q_0, \varepsilon).$$

Corollary. If $H(P_0, Q_0) \ge 4\varepsilon$, then

$$\inf_T \left(\sup_{P \in B_H(P_0,\varepsilon)} P^{\otimes n}(T=1) + \sup_{Q \in B_H(Q_0,\varepsilon)} Q^{\otimes n}(T=0) \right) \le e^{-\frac{n}{8}H^2(P_0, Q_0)}.$$

Pf. Since $(P, Q) \mapsto H^2(P, Q)$ is jointly convex (Lec 3), both balls $B_H(P_0, \varepsilon)$ and $B_H(Q_0, \varepsilon)$ are convex. Therefore, the result follows from

$$H(B_H(P_0, \varepsilon), B_H(Q_0, \varepsilon)) = \inf_{\substack{P \in B_H(P_0,\varepsilon) \\ Q \in B_H(Q_0,\varepsilon)}} H(P, Q)$$
$$\ge \inf_{\substack{P \in B_H(P_0,\varepsilon) \\ Q \in B_H(Q_0,\varepsilon)}} H(P_0, Q_0) - H(P, P_0) - H(Q, Q_0)$$
$$\ge H(P_0, Q_0) - 2\varepsilon \ge \frac{1}{2} H(P_0, Q_0) \qquad \square$$

Le Cam - Birgé pairwise comparison estimator.

Let $P_1, \cdots, P_N$ be a maximal ε-packing of $\mathcal{P}$ under H, i.e.

$$H(P_i, P_j) \ge \varepsilon, \quad \forall\, i \ne j.$$

Since a maximal ε-packing is also an ε-covering,

$$\sup_{P \in \mathcal{P}} \min_{i \in [N]} H(P, P_i) \le \varepsilon.$$

For $\delta = 4\varepsilon$ and $H(P_i, P_j) \ge \delta$, construct a test T_{ij} for

$$H_0: \quad P \in B_H(P_i, \varepsilon) \qquad \text{vs.} \qquad H_1: \quad P \in B_H(P_j, \varepsilon).$$

By the above corollary, $\exists\, T_{ij}$ (and $T_{ji} := 1 - T_{ij}$)

$$\sup_{P \in B_H(P_i, \varepsilon)} P(T_{ij} = 1) \le e^{-\frac{n}{8} H(P_i, P_j)^2}.$$

Now define the following estimator:

- For $i \in [N]$, let $\psi_i = \max\{ H(P_i, P_j) : T_{ij} = 1,\ H(P_i, P_j) \ge \delta \}$
  ($\psi_i = 0$ if no such j exists)
- $\hat{P} = P_{i^*}$, with $i^* = \operatorname{argmin}_{i \in [N]} \psi_i$.

Thm. If $n\varepsilon_n^2 \ge \max\{\log N_H(\mathcal{P}, \varepsilon_n), 1\}$, then the above estimator $\hat{P}$ with $\varepsilon = \varepsilon_n$ satisfies

$$\sup_{P \in \mathcal{P}} P(H(P, \hat{P}) > 4t\varepsilon_n) \le Ce^{-t^2}, \quad \forall t \ge 1.$$

Consequently, $\sup_{P \in \mathcal{P}} \mathbb{E}[H^2(P, \hat{P})] = O(\varepsilon_n^2)$.

Pf. Since $\{P_1, \cdots, P_N\}$ is an ε-covering, WLOG assume that $H(P, P_1) \le \varepsilon$.

For $\delta = 4\varepsilon$ and $t \ge 1$,

$$\begin{aligned}
\{ H(\hat{P}, P_1) \ge t\delta \} &= \{ H(P_{i^*}, P_1) \ge t\delta \} \\
&\subseteq \{ \max\{ \psi_{i^*}, \psi_1 \} \ge t\delta \} \quad (\text{one of } T_{i i^*} \text{ and } T_{i^* 1} \text{ must be one}) \\
&= \{ \psi_1 \ge t\delta \} \quad (\psi_{i^*} \le \psi_1) \\
&\subseteq \bigcup_{j : H(P_1, P_j) \ge t\delta} \{ T_{1j} = 1 \}.
\end{aligned}$$

By a union bound,

$$P(H(\hat{P}, P_1) \ge t\delta) \le N \cdot e^{-\frac{n}{8}(t\delta)^2} = N_H(\mathcal{P}, \varepsilon)\, e^{-2nt^2\varepsilon^2}.$$

Since $n\varepsilon^2 \ge \max\{1, \log N_H(\mathcal{P}, \varepsilon)\}$, this probability is at most $O(e^{-t^2})$.

Finally, $P(H(\hat{P}, P) \ge t\delta) \le P(H(\hat{P}, P_1) \ge t\delta - \varepsilon)$. ∎

Remark: ① $\hat{P}$ is proper, i.e. $\hat{P} \in \mathcal{P}$.

② A high-probability upper bound on $H(\hat{P}, P)$ is established above.

Refinement via local entropy

It turns out that the global entropy $\log N_H(\mathcal{P}, \varepsilon)$ can be improved to a local entropy $\log N_{loc}(\mathcal{P}, \varepsilon)$, with

$$N_{loc}(\mathcal{P}, \varepsilon) = \sup_{P \in \mathcal{P}} \sup_{\eta \geq \varepsilon} N_H\left(B_H(P, \eta) \cap \mathcal{P}, \frac{\eta}{2}\right).$$

(In other words, we are using balls of radius $\frac{\eta}{2}$ to cover balls of radius η)

Example. For many d-dimensional family $\mathcal{P}$, we usually have

$$\log N_H(\mathcal{P}, \varepsilon) \asymp d \log \frac{1}{\varepsilon}, \quad \log N_{loc}(\mathcal{P}, \varepsilon) \asymp d.$$

Therefore, using local entropy improves the Hellinger result from $O(\frac{d \log n}{n})$ to $O(\frac{d}{n})$.

Thm. The same guarantee holds for Le Cam-Birgé pairwise comparison estimator, with N_H replaced by N_{loc}.

Pf. Let $2^\ell \leq t < 2^{\ell+1}$. We decompose $\{j \in [N] : H(P_1, P_j) \geq t\delta\} \subseteq \bigcup_{k \geq \ell} A_k$, where

$$A_k = \{j \in [N] : 2^k \delta \leq H(P_1, P_j) < 2^{k+1}\delta\}.$$

By union bound,

$$\begin{aligned} \mathbb{P}(H(\hat{P}, P_1) \geq t\delta) &\leq \mathbb{P}(\psi_1 \geq t\delta) \\ &\leq \sum_{k \geq \ell} \mathbb{P}(2^k\delta \leq \psi_1 < 2^{k+1}\delta) \\ &\leq \sum_{k \geq \ell} |A_k| e^{-\frac{n}{8}(2^k\delta)^2}. \end{aligned}$$

To upper bound $|A_k|$, since $\{P_1, \cdots, P_N\}$ is an ε-packing,

$$\begin{aligned} |A_k| &\leq M(\{P \in \mathcal{P} : 2^k\delta \leq H(P_1, P) < 2^{k+1}\delta\}, \varepsilon) \\ &\leq M(B_H(P_1, 2^{k+1}\delta) \cap \mathcal{P}, \varepsilon) \\ &\leq N(B_H(P_1, 2^{k+1}\delta) \cap \mathcal{P}, \tfrac{\varepsilon}{2}) \\ &\leq N_{loc}(\mathcal{P}, \varepsilon)^{k+4} \qquad \text{(see lemma below)} \end{aligned}$$

$$\Rightarrow \quad \mathbb{P}(H(\hat{P}, P_1) \geq t\delta) \leq \sum_{k \geq \ell} e^{(k+4)\log N_{loc}(\mathcal{P}, \varepsilon) - 2n\varepsilon^2 4^k} \leq e^{-\Omega(4^\ell)} = e^{-\Omega(t^2)}.$$

($n\varepsilon^2 \geq \max\{1, \log N_{loc}(\mathcal{P}, \varepsilon)\}$.)

Lemma. For $\eta \geq \varepsilon$. $N_H(B_H(P, 2^k\eta) \cap \mathcal{P}, \frac{\eta}{2}) \leq N_{loc}(P, \varepsilon)^{k+1}$.

Pf. Induction on k. Base case $k=0$ is the definition of $N_{loc}(P, \varepsilon)$.

For the inductive step, first cover $B_H(P, 2^k\eta) \cap \mathcal{P}$ using balls of radius $2^{k-1}\eta$, then cover each ball using small balls of radius $\frac{\eta}{2}$.

$\Rightarrow \quad N_k \leq N_{k-1} \cdot N_0 \leq N_{loc}(P, \varepsilon)^k \cdot N_{loc}(P, \varepsilon) = N_{loc}(P, \varepsilon)^{k+1}$. ☐

Special Topic: Guest lecture by J. Qian on his recent work on high-probability density estimation under KL.