STA355 Lecture 11

Final exam. Dec 19. 2-5pm.

- Similar format to midterm
- No electronics
- Bring your ID.
- 4 pages (double-sided) cheat sheet.

Before the exam.

- Office hours (TA/instructor).
- Practice question (released next week).

Nonparametric estimation.

$X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathbb{P}$, learn something about $\mathbb{P}$.

- Nonparametric model: $\mathbb{P}$ is not indexed by a finite-dimensional parameter.

e.g. cdf estimation.

Today, we'll discuss

(i) density estimation

(ii) nonparametric regression.

(i) $X_1, X_2, \cdots X_n \overset{iid}{\sim} p^*$ ($\in \mathcal{P}$, class of densities)

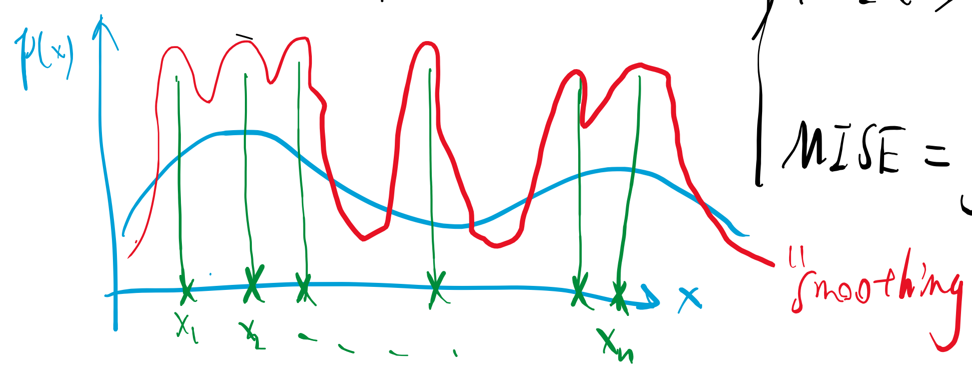
(p^* is the pdf)

Goal: recover p^*, more concretely, we want to minimize

Empirical estimator won't work!

$$MSE(x_0) = \mathbb{E}\left[|\hat{p}(x_0) - p^*(x_0)|^2\right]$$

$$MISE = \int MSE(x)\, dx.$$



involving bias-variance trade-offs.

(ii) Nonparametric regression.

$(X_i, Y_i)_{i=1}^n$

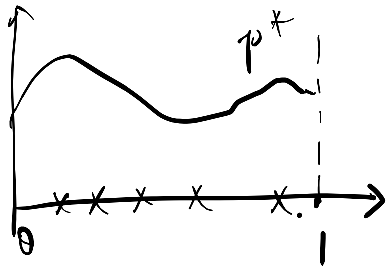
Fixed design, X_i's are deterministic
$Y_i = f^*(x_i) + \varepsilon_i$

Random design, $(X_i, Y_i)_{i=1}^n$ follows a joint distribution.

- Density estimation.

$$X_1, X_2, \dots X_n \overset{iid}{\sim} p^* \text{ on } [0,1].$$

"good density".



Assumption on p^*.

$$|p^*(x) - p^*(y)| \le L|x-y| \quad \forall x,y \in [0,1].$$

"a bad density"

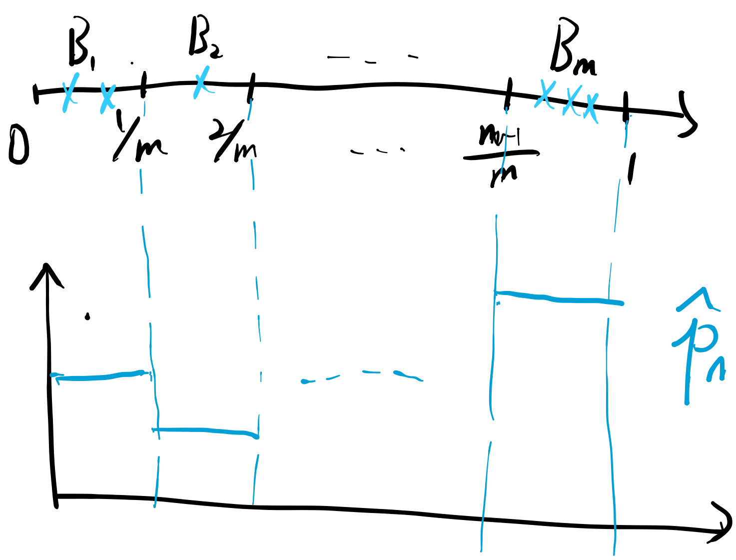
arbitrary fluctuation.

(L is a constant characterizing modulus of continuity).

(assume L is fixed and known).

"Histogram estimator".

m bins



$$\frac{1}{nh}\sum_{i=1}^{n} 1\{X_i \in B_j\} = \hat{p}_n(x) = \frac{|B_j \cap \{x_1, \dots, x_n\}|}{n/m}$$

for $x \in B_j$.

$\hat{p}_n(x)$: piecewise constant function.

Idea: $\hat{p}_n(x) = \dfrac{\frac{1}{n}\cdot(\#\text{points in } B_j)}{\text{width of } B_j}.$

$$p^*(x) = \lim_{h\to 0} \frac{\mathbb{P}(X \in [x, x+h])}{h}.$$

Remark: we can use this regardless of Lipschitz assumption. But we expect it to work well only with certain regularities on p^* (recall the bad example).

- Choice of m
  - Larger m, width $1/m$ smaller. This gives larger variance and smaller bias.

Theoretical analysis. (Under Lipschitz assumption).

Fact. $\mathbb{E}[\hat{p}_n(x)] = \dfrac{\mathbb{P}(X \in B_j)}{h}$. $(h := 1/m)$

(for $x \in B_j$)

$$\mathrm{Var}(\hat{p}_n(x)) = \frac{\mathbb{P}(X \in B_j)\cdot(1 - \mathbb{P}(X \in B_j))}{n h^2}.$$

Upper bounds for bias & variance.

- $$b(x) := \mathbb{E}\left[\hat{p}_n(x)\right] - p^*(x).$$

$$= \frac{\mathbb{P}(X \in B_j)}{h} - p^*(x)$$

$$= \frac{1}{h}\int_{B_j} p^*(u)\,du - p^*(x).$$

$$= \frac{1}{h}\int_{B_j} \left(p^*(u) - p^*(x)\right) du.$$

$$|b(x)| \leq \frac{1}{h}\int_{B_j} |p^*(u) - p^*(x)|\,du$$

$$\leq \frac{L}{h}\int_{B_j} |x-u|\,du \leq \frac{L}{h}\int_{B_j} h\,du$$

$$\leq L \cdot h.$$

- $$\sigma^2(x) = var\left(\hat{p}_n(x)\right) = \frac{\mathbb{P}(X \in B_j)\cdot\left(1-\mathbb{P}(X \in B_j)\right)}{nh^2}.$$

$$\leq \frac{\mathbb{P}(X \in B_j)}{nh^2} \leq \frac{p_{max}}{nh}$$

(Assuming $p^*(x) \leq p_{max}$ $\forall$ $x \in [0,1]$)

(Actually this is implied by Lipschitz assumption).

$$\mathbb{P}(X \in B_j) = \int_{B_j} p^*(x)\,dx \le \int_{B_j} p_{max}\,dx = p_{max} h.$$

- Bias-variance tradeoff

(Larger $m \Rightarrow$ smaller h

smaller bias

larger variance.

$$\mathbb{E}\left[|\hat{p}_n(x_0) - p^*(x_0)|^2\right]$$
$$= b(x_0)^2 + \sigma^2(x_0).$$
$$\le L^2 h^2 + \frac{p_{max}}{nh}$$

Minimize the sum. optimal $h_n = \left(\frac{p_{max}}{2L^2 n}\right)^{1/3}$.

This gives the error bound

$$MSE(x_0) \le 2 \cdot \left(\frac{p_{max} L^2}{n}\right)^{2/3}.$$

- Unlike parametric estimators, the convergence rate is of order $n^{-2/3}$ in MSE.

- Here we mainly focus on the rate w/o caring too much about the constant.
So we can choose h w/o knowing (L, p_{max}).

By choosing $h_n = C_0 n^{-1/3}$, we obtain

$$MSE(x_0) \le C' \cdot n^{-2/3}$$

where C' is a constant depending on (C_0, p_{max}, L).

- Integrating the bound w.r.t. x.

$$MISE = \int_0^1 MSE(x)\,dx \le C' \cdot n^{-2/3}.$$

Both MSE and MISE bounds are information-theoretically (rate) optimal for Lipschitz densities.

---

p^* may be more regular than just Lipschitz.

eg $\left|\frac{d^2 p^*}{dx^2}(x)\right| \le L \quad (\forall x \in [0,1])$.

(this is more than Lipschitz).

Better choice: kernel density estimator.

Given a kernel function $K: \mathbb{R} \to \mathbb{R}$.

satisfying

(i) $\int K(x)\,dx = 1.$

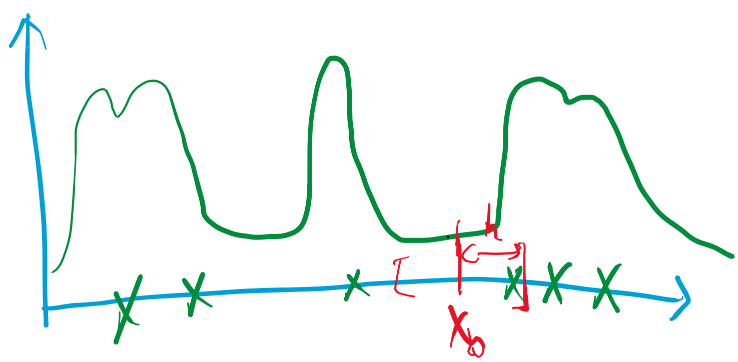
(ii) $\int x K(x)\,dx = 0$ (satisfied by even function K).

We estimate $p^*(x)$ by

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

$\frac{1}{nh}$ normalization ensures $\int \hat{p}_n(x)\,dx = 1.$

h: bandwidth parameter.



Choice of K:

- For Lipschitz / second-order smooth p^*,

  choice of K does not matter too much.

  You could simply use box-shaped K

  $$K(x) = \frac{1}{2} \cdot \mathbb{1}\{x \in [-1, 1]\}$$

  bandwidth choice is more important.

(when taking box-shaped K, KDE becomes local averaging).

- Choice of K matters for (≥ 3)-rd order smoothness p^*, beyond the scope.

Analysis of KDE.

- Bias
$$b(x) = \frac{1}{h}\int K\left(\frac{u-x}{h}\right)\cdot p^*(u)\,du - p^*(x)$$
$$= \frac{1}{h}\int K\left(\frac{u-x}{h}\right)\cdot \left(p^*(u) - p^*(x)\right)du.$$

- Variance
$$\sigma^2(x) = \frac{1}{nh^2}\cdot \mathrm{var}\left(K\left(\frac{X-x}{h}\right)\right)$$

(similar to histogram estimator).

$$\leq \frac{1}{nh^2}\cdot\int K^2\left(\frac{u-x}{h}\right)\cdot p^*(u)\,du.$$

(change of var! $y = \frac{u-x}{h}$).

$$= \frac{1}{nh}\int K^2(y)\cdot p^*(x+yh)\,dy.$$
$$\leq \frac{p_{max}}{nh}\cdot \boxed{\int K^2(y)\,dy}$$

Constant.

---

closer look at bias.

$$b(x) = \frac{1}{h}\int K\left(\frac{u-x}{h}\right)\left(p^*(u) - p^*(x)\right)du$$

- If we assume p^* is L-Lipschitz,
$$|b(x)| \leq \frac{1}{h}\int K\left(\frac{u-x}{h}\right)\cdot |u-x|\,du.$$

(Assuming that K is supported on $[-1,1]$

i.e. $K(x)=0$ when $|x|>1$).

$$= \frac{L}{h}\int_{x-h}^{x+h} \left|K\left(\frac{u-x}{h}\right)\right| |u-x|\, du.$$

(Assuming $|K(x)| \le K_{max}$)

$$\le \frac{LK_{max}}{h}\int_{x-h}^{x+h} |u-x|\, du = LK_{max}\, h.$$

Same as histogram method.

Assuming $|(p^*)''(x)| \le L \quad \forall x \in [0,1]$.

KDE can do better.

$$b(x) = \int_{-1}^{1} K(y)\cdot \left(p(x+yh)-p(x)\right) dy.$$

$(u-x = yh)$.

$$p(x+yh)-p(x) = yh\cdot p'(x) + \frac{y^2h^2}{2} p''(x+\tau_y yh)$$

where $\tau_y \in [0,1]$ depends on y.

Cancels out after integration.

$$|b(x)| \le \int_{-1}^{1} |K(y)| \frac{y^2 h^2}{2} \cdot |p''(x+\tau_y y h)| \, dy. \le L \cdot h^2 K_{max}$$

This is because

$$\int_{-1}^{1} K(y) \cdot y \cdot h p'(x) \, dy$$

$$= \left(\int_{-1}^{1} K(y) \cdot y \, dy \right) \cdot h p'(x) = 0.$$

Bias-var tradeoff.

$$\mathbb{E}\left[|\hat{p}_n(x_0) - p^*(x_0)|^2 \right] = b(x_0)^2 + \sigma^2(x_0)$$

$$\le (L h^2 K_{max})^2 + \frac{p_{max}}{nh} \cdot \int K(x)^2 \, dx.$$

$$= C_1 \cdot h^4 + C_2 \cdot \frac{1}{nh}.$$

Choose $h = C_0 \cdot n^{-1/5}$, we get

$$\mathbb{E}\left[|\hat{p}_n(x_0) - p^*(x_0)|^2 \right] \le C' \cdot n^{-4/5}.$$

Remark: faster rate available assuming more derivatives

But this requires clever choice of

(i) K (ii) h.

Brief intro to nonparametric regression.

Nadaraya-Watson estimator

Idea: "local averaging".

$$Y_i = f^*(x_i) + \varepsilon_i \qquad (i = 1, 2, \cdots n).$$

f^* is 1st/2nd order smooth.

$$\hat{f}_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}.$$

(e.g. when K is box-shaped, $\hat{f}_n(x)$ is local averaging).

Achieves similar convergence rate.

— $n^{-2/3}$ for 1st order differentiable f^*

— $n^{-4/5}$ for 2nd order differentiable f^*.

Final review.

- cdf / empirical estimation / bootstrap.
  - Concepts. How these are implemented.
  - Theory. (no ε-δ proofs, though ε-δ ideas are useful for understanding).

  e.g. When bootstrap works.
  (what it means for bootstrap to work).

- M-estimation, MLE, finite-dimensional asymptotics.
  - Consistency. $\hat{\theta}_n \xrightarrow{P} \theta^*$
  - Asymptotic normality.

  "big theorems"

  Assumptions and how to verify them.

  (— Error bounds).

- Testing.
  - Concepts of type I/II.
  - Wald / χ^2 tests / permutation tests.

  (— Testing radius).

- Decision theory / Bayes methods.
  - Concepts of minimax / Bayes estimators.
  - For calculations about Bayes estimator. Calculation won't be the burden. (We'll provide formula for conjugate prior).

- Regression / nonparametrics.
  (You'll get partial credit for solving 1-D).
  - Calculate concretely. (e.g. estimation error, inference etc.)
  - Nonparametric density estimation (need to be comfortable w/ convergence rate analysis).