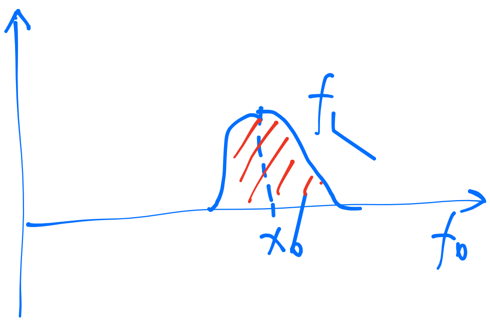
STA 3000 F   Lecture 12

Question: minimax risk for estimating f under integrated MSE

$$\inf_{\hat f}\ \sup_{f^*\in\mathcal{F}} \mathbb{E}\left[\int_0^1 |\hat f(x)-f^*(x)|^2\,dx\right].$$

Recall construction from last time.



$\int_0^1 |f_1(x)-f_0(x)|^2\,dx$ small.

Under Gaussian noise.

$x_i = i/n$,

$$D_{KL}\left(P_1^{\otimes n}\,\|\,P_0^{\otimes n}\right) = \frac{1}{2}\sum_{i=1}^n \left(f_1(i/n)-f_0(i/n)\right)^2$$

$$\approx \frac{n}{2}\int_0^1 \left(f_1(x)-f_0(x)\right)^2 dx.$$

If we want $D_{KL} \leq \frac{1}{2}$

$$\|f_1-f_0\|_{L^2[0,1]} \lesssim \frac{1}{\sqrt{n}}.$$

Upper bound $\|\hat f_n - f^*\|_{L^2[0,1]} \lesssim n^{-\frac{\beta}{2\beta+1}}$.

Introduction to introduction to Info theory.

(C. Shannon).

"entropy" $H(X)$

- Compress information
- Send it over noiseless channel.

$I(X;Y)$ "Mutual information".

X is discrete r.v.

$$H(X) \left(:= H(P_X) \right) := \sum_{x \in \mathcal{X}} \left(\log_2 \frac{1}{p(x)} \right) \cdot p(x).$$

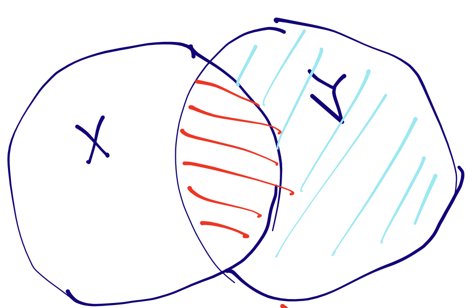
(Diff ent. $\int p(x) \log_2 \frac{1}{p(x)} \, dx$.)

On average, need $n \cdot H(X)$ bits to encode n iid copies of X.

Joint entropy $H(X,Y)$.

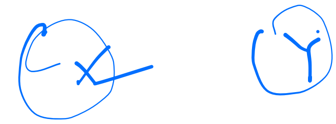
Conditional entropy $H(X|Y) := \sum_{y \in \mathcal{Y}} H(X|Y=y) \cdot P(Y=y)$. $(\leq H(X))$.

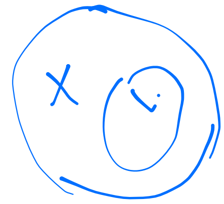
Mutual information: $I(X;Y) := H(X) - H(X|Y)$
$= H(Y) - H(Y|X)$.



$H(Y|X)$

$I(X;Y)$.

Independent: 

$Y = f(X)$: 

Shannon's theorem:

Channel capacity $= \sup_X I(X;Y)$.

where $Y = \text{channel}(X)$.

In Statistics

$\theta \sim \pi$. $X \mid \theta \sim P_\theta$.

$I(X;\theta)$ characterises possibility of recovering θ from X.

Back to minimax lower bounds.

Discrete decision problem:

$H_1: X \sim P_1$ vs. $H_2: X \sim P_2$ ---- $H_M: X \sim P_M$.

Decision $T: X \to [M]$. $X \sim P_J$, $J \sim \mathrm{Unif}([M])$.

Quantity of interest.

$$\mathbb{P}\left(T(X) \neq J \right).$$

Thm (Fano).

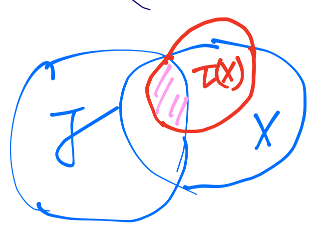
$$\mathbb{P}\left(T(X) \neq J \right) \geq 1 - \frac{I(X;J) + \log 2}{\log M}.$$

Proof of Fano.

· Step I. data processing inequality

$$J \to X \to T(X).$$

Then $$I(X;J) \geq I\left(T(X);J\right).$$



$$I(X;J) = I\left(X, T(X); J\right) = I\left(T(X);J\right) + I\left(X;J \mid T(X)\right)$$

· Step II. $$f(t,j) = \mathbb{1}\{t \neq j\}.$$

Detour

$$I(X;Y) = D_{KL}\left(\mathbb{P}_{X,Y} \,\|\, \mathbb{P}_X \times \mathbb{P}_Y \right)$$

$$I(T(X); J) = D_{KL}\left(\mathbb{P}_{T(X),J} \,\middle\|\, P_{T(X)} \times \text{Uniform}([M]) \right).$$

$$\boxed{D_{KL}(P_X \| P_Y) \geq D_{KL}(P_{f(X)} \| P_{f(Y)})}.$$

$$\geq D_{KL}\left(\mathbb{P}_{f(T(X),J)} \,\middle\|\, \mathbb{P}_{f(T(X), J')} \right)$$

$$(J' \perp X).$$

$$= D_{KL}\left(\text{Ber}(P_e) \,\middle\|\, \text{Ber}\left(1 - \frac{1}{M}\right) \right).$$

$$\left(P_e = \mathbb{P}(T(X) \neq J) \right).$$

$$D_{KL}\left(\text{Ber}(P_e) \,\middle\|\, \text{Ber}\left(1 - \frac{1}{M}\right) \right)$$

$$= P_e \log \frac{P_e}{1 - \frac{1}{M}} + (1 - P_e) \log \frac{1 - P_e}{1/M}.$$

$$\geq -\log 2 + \log M - P_e \log M.$$

$$\Longrightarrow \quad P_e \geq 1 - \frac{I(X; J) + \log 2}{\log M}.$$

Corollary. If $f_1, f_2, \cdots, f_M \in \mathcal{F}$:

$$J \sim \text{Unif}([M])$$

If $\|f_i - f_j\| \geq 2\delta$ for each i,j pair

Then Bayes risk lower bound

$$\inf_{\hat{f}} \underbrace{\mathbb{E}\left[\|\hat{f} - f_J\|^2\right]} \geq \delta^2 \cdot \left\{1 - \frac{I(X;J) + \log 2}{\log M}\right\}$$

Proof: reduce to testing.

Suppose we have estimator $\hat{f}$,

Construct test $T = \underset{j \in [M]}{\operatorname{argmin}} \|\hat{f} - f_j\|$

$$\mathbb{P}(T \neq J) \leq \frac{1}{\delta^2} \underbrace{\mathbb{E}\left[\|\hat{f} - f_J\|^2\right]}.$$

Construct $f_1, f_2, \dots f_M \in \mathcal{F}$. and make sure.

- M large
- $I(X;J)$ bounded
- f_i, f_j separated from each other.

make $\frac{1}{M} \sum_{j=1}^{M} D_{KL}(P_j \| Q)$ bounded.

Bound $I(X;J) = D_{KL}\left(\mathbb{P}_{X,J} \,\|\, \mathbb{P}_X \times \mathbb{P}_J\right)$ $J \sim \mathrm{Unif}([M])$.

$$= \frac{1}{M} \cdot \sum_{j=1}^{M} D_{KL}(P_j \| \bar{P})$$

where $P_j = \mathbb{P}(X \mid J=j)$.

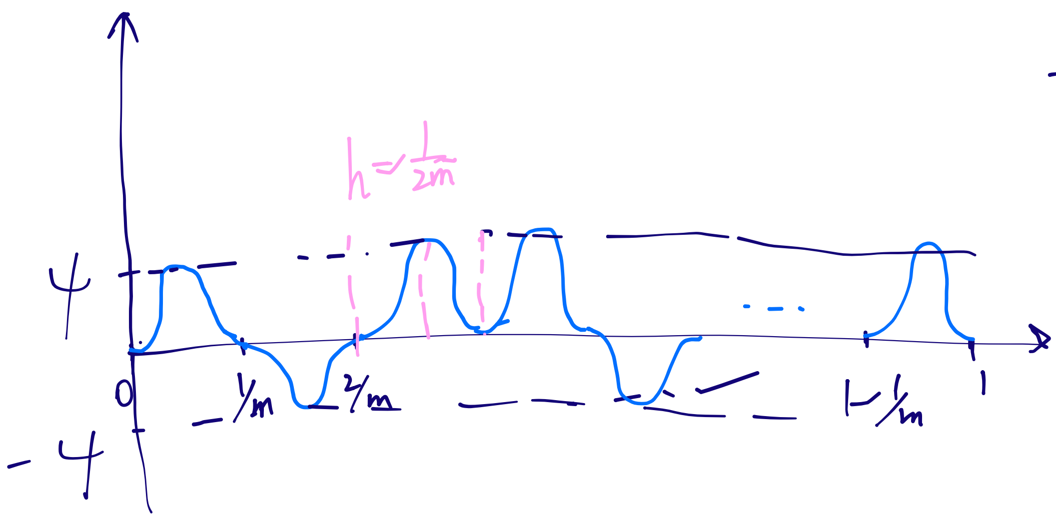
$$\bar{P} = \frac{1}{M} \sum_{j=1}^{M} P_j .$$

$$\frac{1}{M} \sum_{j=1}^{M} D_{KL}(P_j \| \bar{P}) \leq \frac{1}{M} \sum_{j=1}^{M} D_{KL}(P_j \| Q)$$

So we have

$$I(X;J) \leq \frac{1}{M} \sum_{j=1}^{M} D_{KL}(\mathbb{P}_j \| Q).$$

$f_0 = 0.$



For sign vector $z \in \{-1, 1\}^m$

Define $f_z(x) = 4 K\left(\frac{x - \frac{2i-1}{2m}}{h}\right) z_i$

for $x \in \left[\frac{i-1}{m}, \frac{i}{m}\right]$.

$$D_{KL}(P_z \| P_0) = \frac{n}{2} \| f_z \|_n^2 \leq n \cdot 4^2$$

For each z, z' pair

$$\| f_z - f_{z'} \|^2_{L^2[0,1]} = c \cdot \psi^2 \cdot h \cdot \sum_{i=1}^{m} \mathbb{1}_{z_i \neq z_i'}$$

Sample $z^{(1)}, z^{(2)}, \ldots z^{(M)} \overset{iid}{\sim} \mathrm{Unif}(\{\pm 1\}^m)$.

From Hw 1.

$$\mathbb{P}\left(\exists\, i, j,\ \| z^{(i)} - z^{(j)} \|_1 \leq \frac{m}{8} \right) \leq \frac{1}{2}$$

For $M = \exp(c' m)$. $(c' > 0)$.

Only for construction. Seen as deterministic for the rest of proof.

Use them as our prior distribution.

$$\forall\, i, j \quad \| f_{z^{(i)}} - f_{z^{(j)}} \|^2_{L^2} \geq c \cdot \psi^2 h \cdot \frac{m}{8}$$

$$= \frac{c}{16} \psi^2.$$

$$\log M = c' m.$$

$$\frac{1}{M} \sum_{j=1}^{M} D_{KL}(P_j \| Q) \leq n \cdot \psi^2.$$

From last lecture. for $f_z \in \Sigma(\beta, L)$.

$$\psi \leq c_1 h^\beta$$

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}\left[\|\hat{f} - f^*\|_{L^2}^2\right]$$

$$\geq C \cdot \psi^2 \left\{ 1 - \frac{\frac{1}{M} \sum_{j=1}^{M} D_{KL}(P_j \| Q) + \log 2}{m} \right\}$$

$$\geq C \cdot \psi^2 \left\{ 1 - \underbrace{\frac{n\psi^2 + \log 2}{m}}_{\frac{1}{2}} \right\}$$

$$\psi \leq h^\beta = m^{-\beta}$$

Choose : $\psi = n^{-\frac{\beta}{2\beta+1}}$ (up to const factors).

$m = n^{\frac{1}{2\beta+1}}$

So minimax rate $\geq n^{-\frac{2\beta}{2\beta+1}}$.

Adaptivity.

Recall in nonpara regression $f^* \in \Sigma(\beta, L)$

then $MSE(x_0) \lesssim n^{-\frac{2\beta}{2\beta+1}}$

Require knowledge about β.

Unknown β. hope to get optimal rate for underlying β.

$\forall \beta \in [\beta_{min}, \beta_{max}]$ $(\beta_{min} > 0)$.

Use a single estimator $\hat{f}$

s.t. $\forall f^* \in \Sigma(\beta, 1)$,

$$\mathbb{E}\left[|\hat{f}(x_0) - f^*(x_0)|^2 \right] \leq C \cdot n^{-\frac{2\beta}{2\beta+1}} ?$$

Surprisingly. this is not possible

(but we're not ~~too~~ far off).

Setup/ $Y_i = f^*(i/n) + \varepsilon_i$ $\quad \varepsilon_i \overset{iid}{\sim} N(0,1)$

$f^* \in \Sigma(\beta, 1)$ with $\beta \in (0, 1]$.

Thm. $0 < \beta_1 < a < \beta_2 \leq 1$, $\quad r_1^2(n) = \left(\frac{\log n}{n}\right)^{\frac{2\beta_1}{2\beta_1+1}}$

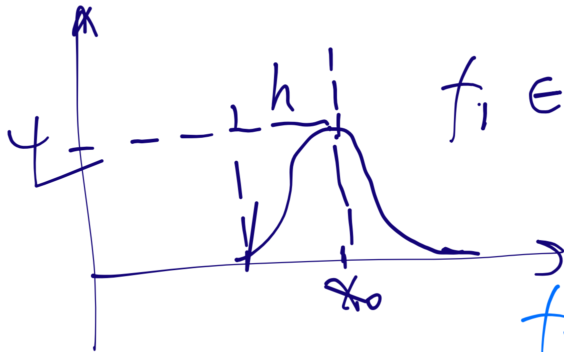
$$r_2^2(n) = n^{-\frac{2a}{2a+1}}$$

$$\inf_{\hat{f}} \ \sup_{i \in \{1,2\}} \ \sup_{f \in \Sigma(\beta_i)} \mathbb{E}_f \left[\frac{1}{r_i^2(n)} \left| \hat{f}_n(x_0) - f(x_0) \right|^2 \right] \geq C.$$

If we want idealized adaptivity,

need $\underline{r_i^2(n)} = n^{-\frac{2\beta_i}{2\beta_i+1}}$ for $i \in \{1,2\}$

Proof sketch.



$f_1 \in \Sigma(\beta_1)$

$\psi \leq h^{\beta_1}$.

$f_0 = 0 \in \Sigma(\beta_2)$

"Asymmetric version of two-pt method"

$$q_n =: \frac{r_1^2(n)}{r_2^2(n)} \quad (= \mathrm{poly}(n)).$$

$$R_{\text{adaptive-minimax}} \geq \frac{\psi_n^2}{c \cdot r_1(n)^2} \left(\mathbb{P}_1(\hat{T} = 0) + q_n \cdot \mathbb{P}_0(\hat{T} = 1) \right).$$

$$\mathbb{P}_1(\hat{T}=0) + q_n \mathbb{P}_0(\hat{T}=1)$$

$$\geq \int \min\left(P_1(x), \; q_n P_0(x) \right) \cancel{dx}$$

$$\geq 1 - \frac{\chi^2(P_1 \| P_0) + 1}{q_n}.$$

$$1 + \chi^2(P_1 \| P_0) \leq \exp\left(2 4_n^2 h_n \cdot n\right).$$

Allow $4_n^2 \cdot h_n \cdot n \leq \log n.$

Thm. $\exists$ an estimator $\hat{f}_{\text{Lepski}}$ s.t.

$$\sup_{\beta_{min} \leq \beta \leq \beta_{max}} \sup_{f \in \Sigma(\beta, L)} \frac{\mathbb{E}\left|\hat{f}_{\text{Lepski}}(x_0) - f(x_0)\right|^2}{\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}}} \leq C.$$

If β^* is the ground truth,

for $\beta < \beta^*$ (over-smooth)

$$\left| \hat{f}_{h_\beta}(x_0) - \hat{f}_{h_{\beta^*}}(x_0) \right| \leq C \cdot h_\beta^\beta \quad \text{for any } \beta < \beta^*.$$

$$\left(h_\beta = \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}}, \quad f_h \text{ local avg estimator w/ bandwidth } h \right).$$

Lepski's method.

Step 1. $B = \{ \beta_{min} = \beta_1 < \beta_2 < \cdots < \beta_N = \beta_{max} \}$

where $\beta_j - \beta_{j-1} = \frac{1}{\log n}$

$N = O(\log n)$ many grid pts

Rate $\quad \beta \in [\beta_j, \beta_{j+1}]$.

$$c \cdot n^{\frac{-2\beta_j}{2\beta_j+1}} \leq n^{\frac{-2\beta}{2\beta+1}} \leq n^{-\frac{2\beta_j}{2\beta_j+1}}$$

Step 2.

$$\hat{\beta} = \max \left\{ \beta \in B : \left| \hat{f}_{h_\beta}(x_0) - \hat{f}_{h_{\beta'}}(x_0) \right| \leq C \cdot h_{\beta'}^{\beta'} \quad \forall \beta' \leq \beta, \ \beta' \in B \right\}$$

$$\hat{f}_{Lepski}(x_0) = \hat{f}_{h_{\hat{\beta}}}(x_0).$$

Proof Sketch.

$$\mathcal{E}_j = \{ \hat{\beta} = \beta_j \}$$

Ground truth β_{j^*} $\qquad f \in \Sigma(\beta_{j^*}, 1).$

$$\mathbb{E}\left[|\hat{f}(x_0) - f(x_0)|^2 \right] = \sum_{j=1}^{N} \mathbb{E}\left[|\hat{f}(x_0) - f(x_0)|^2 \mathbb{1}_{\mathcal{E}_j} \right].$$

Case I: $j \geq j^*$. Event $\mathcal{E}_j$ makes it safe to use β_j

Case II: $j < j^*$. β^* does not pass the test happens with small prob ($\log n$ appears here).