Density Estimation:

R.V. X has density f on $\mathbb{R}^d$ when

$$P\{X \in A\} = \int_A f(x)\,dx \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^d), \text{ Borel sets in } \mathbb{R}^d$$

[ie, $P\{X \in A\} = f(x) \cdot \mathrm{Vol}(A)$ if $A = \{x' \in \mathbb{R}^d : d(x', x) \le r\}$ for small $r > 0$.

Purpose

Estimate unknown density f from an i.i.d. sample $X_1, X_2, \ldots, X_n$ drawn from f.

density estimate

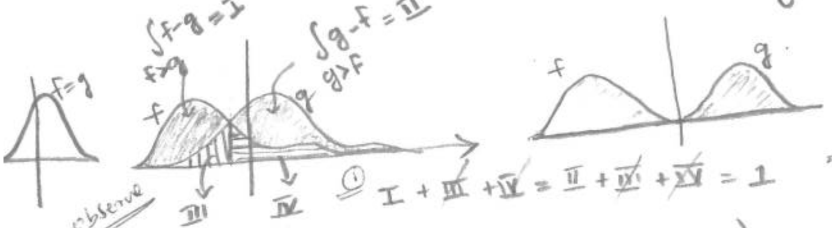
$$f_n(x) = f_n(x; X_1, \ldots, X_n) : (\mathbb{R}^d)^{n+1} \longrightarrow \mathbb{R}.$$

quality of f_n measured by T.V. distance

$$\text{error} = \sup_{B \in \mathcal{B}(\mathbb{R}^d)} \left| \int_B f_n - \int_B f \right|$$

if This is $< \varepsilon$ then all probabilities will be estimated with errors not exceeding ε

The distance between two densities f & g can be measured by their L_1 distance $\int |f-g|$



① $\mathrm{I} + \cancel{\mathrm{III}} + \cancel{\mathrm{IV}} = \mathrm{II} + \cancel{\mathrm{III}} + \cancel{\mathrm{IV}} = 1$

② & $\int |f-g| = \mathrm{I} + \mathrm{II}$

Thm 1 (Scheffé's Identity). Let f and g be two functions defined on $\mathbb{R}^d$ satisfying $\int f = \int g = 1$.

$$\sup_{B \in \mathcal{B}(\mathbb{R}^d)} \left| \int_B f - \int_B g \right| = \frac{1}{2} \int |f-g|$$

Proof

$$\sup_{B\in\mathcal{B}(\mathbb{R}^d)} \left| \int_B f - \int_B g \right| = \int_{f>g} (f-g) = \int_{g>f} (g-f) = \frac{1}{2}\int |f-g| \quad \blacksquare$$

why L_1 (TV) ① absolute scale ② Physical interpretation

$0 \le \frac{1}{2}\int |f-g| \le 1$, $\int |f-g| < 0.02$

$\Rightarrow$ differences in probabilities are at most $\frac{1}{2} \times 0.02 = 0.01$

③ scale invariance

$$\sup_{B} |P\{X\in B\} - P\{Y\in B\}| = \sup_{B} |P\{T(X)\in T(B)\} - P\{T(Y)\in T(B)\}|$$

if T is a bijection and $\{T(B) : B\in\mathcal{B}(\mathbb{R}^d)\} = \mathcal{B}(T(\mathbb{R}^d))$.

④ TV distance decreases on any Borel measurable mapping $T: \mathbb{R}^d \to S \subseteq \mathbb{R}^k$, i.e. for any R.V.s X and Y on $\mathcal{B}(\mathbb{R}^d)$

$$\sup_{A\in\mathcal{B}(\mathbb{R}^d)} |P\{X\in A\} - P\{Y\in A\}| \ge \sup_{A\in\mathcal{B}(\mathbb{R}^k)} |P\{T(X)\in A\} - P\{T(Y)\in A\}|$$

skip

proof:

$$\sup_{A\in\mathcal{B}(\mathbb{R}^k)} |P\{T(X)\in A\} - P\{T(Y)\in A\}|$$

$$= \sup_{A\in\mathcal{B}(\mathbb{R}^k)} |P\{X\in T^{-1}(A)\} - P\{Y\in T^{-1}(A)\}|$$

$$\le \sup_{A\in\mathcal{B}(\mathbb{R}^d)} |P\{X\in A\} - P\{Y\in A\}|$$

since $\{T^{-1}(A) : A\in\mathcal{B}(\mathbb{R}^k)\} \subseteq \mathcal{B}(\mathbb{R}^d)$ $\blacksquare$

Minimum Distance Estimate

setting: $X_1, \ldots, X_n \overset{iid}{\sim} f$

objective: non parametric density estimate f_n from $x_1, \ldots, x_n$ with universal performance guarantees ($\int |f_n - f|$ is small for any $f \in L_1$ as n gets large).

consider The simpler problem of choosing between two densities f_n and g_n, i.e construct ϕ_n such that,

$$\int |\phi_n - f| \approx \min\left(\int |f_n - f|, \int |g_n - f| \right)$$

Ideas use empirical measure $\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \in A]} = \frac{\#\text{ of data point in } A}{n}$

& scheffé set of ordered pair (f_n, g_n) is:

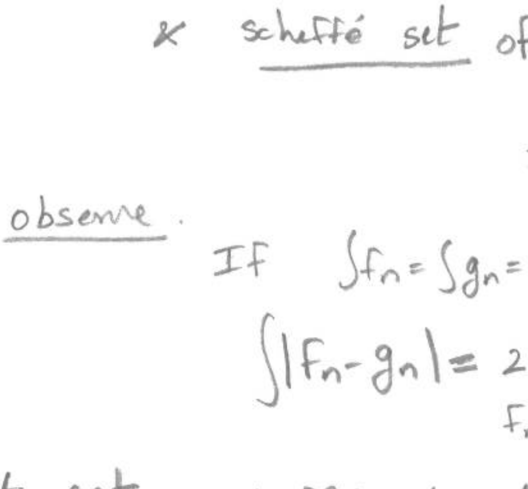
$$A = A(f_n, g_n) = \{ x : f_n(x) > g_n(x) \}$$

observe. If $\int f_n = \int g_n = 1$ then by Scheffé's identity,

$$\int |f_n - g_n| = 2 \int_{f_n > g_n} (f_n - g_n) = 2 \int_A f_n - 2 \int_A g_n$$

to get Scheffé Estimate

$$f_n^* = \begin{cases} f_n & \text{if } \left| \int_A f_n - \mu_n(A) \right| < \left| \int_A g_n - \mu_n(A) \right| \\ g_n & \text{otherwise} \end{cases}$$



eg1: given data ???...?

IF $\mu_n(A) = 8/10$, $\int_A f_n = 7/10$, $\int_A g_n = \frac{4}{10}$.

then $f_n^* = f_n$.

eg2:- different data $\times\times\ldots\times$ from another f

IF $\mu_n(A) = 5/10$.

then $f_n^* = g_n$

<u>Thm</u> (DL 6.1) Let f_n and g_n be two density estimates with $\int f_n = \int g_n = 1$. For the Scheffé estimate f_n^*, we have

$$\int |f_n^* - f| \le 3 \min\left(\int |f_n - f|, \int |g_n - f|\right) + 4 \max_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|,$$

where $\mathcal{A} = \{ \{f_n > g_n\}, \{g_n > f_n\} \}$

<u>Proof</u>. (let's prove a slightly more general Thm)

<u>Note</u>. f_n^* has an error that is within $E_n = 4 \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|$ of 3 times the best possible error among f_n and g_n.

<u>Now</u> consider the problem of <u>selecting from k densities</u>. (our main problem here)

f_{ni}, $1 \le i \le k$, $\int f_{ni} = 1$ for all i

Let A_{ij} denote the Scheffé set $A_{ij} = A(f_{ni}, f_{nj})$
$= \{x : f_{ni}(x) > f_{nj}(x)\}$.

Let the <u>Yatracos</u> <u>class</u> of such Scheffé sets be:

$$\mathcal{A} = \{ A_{ij}, A_{ji} : 1 \le i < j \le k \}$$

The Minimum Distance Estimate (MDE) ψ_n is the $f_{n,i}$ of smallest index that <u>minimizes</u>

$$\Delta_i = \sup_{A \in \mathcal{A}} \left| \int_A f_{ni} - \mu_n(A) \right|$$

$$\psi_n = f_{ni^*}, \quad i^* = \min\left\{ \operatorname*{argmin}_{1 \le i \le k} \Delta_i \right\}.$$

Thm 2 (Universal Performance Bound of MDE from a finite set of densities) [DL 6.3]

For the MDE ψ_n, we have (for each n).

$$\int|\psi_n - f| \leq 3 \min_i \int |f_{ni} - f| + 4\Delta, \quad \Delta = \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|$$

Proof: Let $\psi_n = f_{ni}$ and let $f_{nj} = \operatorname{argmin}_{1 \leq l \leq k} \int |f_{nl} - f|$.

Assume $j \neq i$, then

$$(*) \qquad \int|\psi_n - f| = \int |f_{ni} - f| \leq \int |f_{nj} - f| + \int |f_{ni} - f_{nj}| \quad \left(\Delta^{u} \neq \text{ of } L_1\right)$$

Now, assuming WLOG $i < j$:

$$\int |f_{ni} - f_{nj}| = 2 \sup_{A \in \{A_{ij}, A_{ji}\}} \left| \int_A f_{ni} - \int_A f_{nj} \right| \quad \left(\text{by Scheffé's identity } \underline{\text{Thm 1}}\right)$$

$$\leq 2 \sup_{A \in \mathcal{A}} \left| \int_A f_{ni} - \int_A f_{nj} \right| \quad \left(\{A_{ij}, A_{ji}\} \subseteq \mathcal{A}\right)$$

$$\leq 2 \sup_{A \in \mathcal{A}} \left| \int_A f_{ni} - \mu_n(A) \right| + 2 \sup_{A \in \mathcal{A}} \left| \int_A f_{nj} - \mu_n(A) \right|$$

$$= 2 \underbrace{\sup_{A \in \mathcal{A}} \left| \int_A \psi_n - \mu_n(A) \right|}_{= \Delta_i} + 2 \underbrace{\sup_{A \in \mathcal{A}} \left| \int_A f_{nj} - \mu_n(A) \right|}_{= \Delta_j} \quad \left(\psi_n = f_{ni} \text{ by assumption}\right)$$

$$\leq 4 \sup_{A \in \mathcal{A}} \left| \int_A f_{nj} - \mu_n(A) \right| \quad \left(\Delta_i \leq \Delta_j\right)$$

$$\leq 4 \sup_{A \in \mathcal{A}} \left| \int_A f_{nj} - \int_A f \right| + 4 \underbrace{\sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right|}_{= \Delta}$$

$$\leq 4 \sup_{B \in \mathcal{B}(\mathbb{R}^d)} \left| \int_B f_{nj} - \int_B f \right| + 4\Delta \quad \left(\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)\right)$$

$$(**) \qquad = 2 \int |f_{nj} - f| + 4\Delta$$

$(*)+(**)$ give: $\int |\psi_n - f| \leq 3 \int |f_{nj} - f| + 4\Delta = 3 \min_i \int |f_{ni} - f| + 4\Delta$ ∎

hus to obtain error bounds for MDE we need bounds on $\boxed{\Delta = \sup_{A \in \mathcal{A}} \left| \mu(A) - \mu_n(A) \right|}$

no matter what $f \in L_1$ is generating data $X_1, \dots, X_n$. $\mu(A) = \int_A f$

We are interested in Δ, The maximal deviation of μ_n from μ over $\mathcal{A}$:

$$g(X_1, \dots, X_n) = \Delta = \sup_{A \in \mathcal{A}} \left| \mu_n(A) - \mu(A) \right|$$

we will first show (using concentration $\neq$'s) that

$$P\{|g - Eg| \geq \varepsilon\} \leq 2e^{-2n\varepsilon^2}$$

and thus The maximal deviation is sharply concentrated around its mean and Then we will show (using uniform deviation $\neq$) that this mean $(Eg = E\Delta)$ can be bounded using combinatorial characteristics of $\mathcal{A}$.

Using concentration $\neq$s i.e., $\neq$s of the form:

$$P\{|X_n - EX_n| \geq \varepsilon\} \leq c_1 e^{-nc_2\varepsilon^2}$$

Recall Markov's $\neq$ For any non-negative RV X and any $t > 0$:

$$P\{X \geq t\} \leq \frac{EX}{t}$$

proof: $EX = \int_0^\infty x\,dF(x) = \int_0^t x\,dF(x) + \int_t^\infty x\,dF(x) \geq \int_t^\infty x\,dF(x) \geq t\int_t^\infty dF(x)$

$= t\,P\{X \geq t\}$ ∎

Let $s \in (0,\infty)$, Then for any RV X and any $t > 0$, by Markov's $\neq$, $P\{X \geq t\} = P\{e^{sX} \geq e^{st}\} \leq \frac{E\{e^{sX}\}}{e^{st}}$

In Chernov's bounding method we find an $s > 0$ that minimizes $E\{e^{sX}\}/e^{st}$, The upper bound.

IF $X_1, X_2, \dots, X_n \overset{iid}{\sim} X$, and $S_n = \sum_{i=1}^n X_i$, we get

$$P\{S_n - ES_n \geq t\} \leq e^{-st} E\left\{\exp\left(s\sum_{i=1}^n (X_i - EX_i)\right)\right\}$$

$$= e^{-st} \prod_{i=1}^n E\{e^{s(X_i - EX_i)}\} \quad \text{(by independence)}$$

So finding tight bounds $\Leftrightarrow$ good upper bounds for M.G.F. of $X_i - EX_i$.

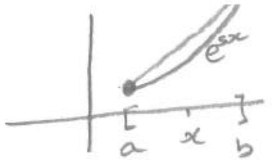
Lemma 4 (upper bounding the MGF of a bounded RV.)

Let X be RV with $EX = 0$, $a \leq X \leq b$.

Then for $s > 0$:

$$E\{e^{sX}\} \leq e^{s^2(b-a)^2/8}$$

proof: by convexity



$$e^{sx} \le \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$

for $a \le x \le b$.

Since $E(X)=0$,

$$E\{e^{sX}\} \le E\left\{\frac{X-a}{b-a} e^{sb} + \frac{b-X}{b-a} e^{sa}\right\} = \underbrace{\left(\frac{-a}{b-a}\right)}_{= p} e^{sb} + \underbrace{\left(\frac{b}{b-a}\right)}_{= 1-p} e^{sa}$$

$$= p e^{sb} + (1-p) e^{sa} = \left(\frac{p e^{sb}}{e^{sa}} + \frac{(1-p)e^{sa}}{e^{sa}}\right) e^{sa}$$

$$= \left(1-p+pe^{s(b-a)}\right) e^{sa} = \left(1-p+pe^{s(b-a)}\right) e^{-ps(b-a)}$$

$$\stackrel{\text{def}}{=} e^{\phi(u)}$$

where $u = s(b-a)$ and $\phi(u) = -pu + \log(1-p+pe^{u})$

now, $\phi'(u) = \frac{d}{du}\phi(u) = -p + \frac{pe^{u}}{(1-p+pe^{u})} = -p + \frac{p}{p+(1-p)e^{-u}}$

Thus, $\phi(0) = \phi'(0) = 0$

and $\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \le \frac{1}{4}$, since $0 < p < 1$ and $e^{-u} < 1$

Then, by Taylor's theorem, for some $\xi \in [0, u]$

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\xi) \le \frac{u^2}{8} = \frac{s^2(b-a)^2}{8} \quad \blacksquare$$

<u>Thm 5</u> (Hoeffding's $\neq$) Let $X_1, \dots, X_n$ be indep RVs such that $X_i \in [a_i, b_i]$ w.p 1. Then for any $t > 0$,

$$P\{S_n - ES_n \ge t\} \le e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2}$$

and

$$P\{S_n - ES_n \le -t\} \le e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2}$$

Proof:

From Chernov's bounding method, for any $s>0$, $t>0$

$$P\{S_n - ES_n \geq t\} \leq e^{-st}\prod_{i=1}^{n} E\{e^{s(X_i - EX_i)}\}$$

$$\leq e^{-st}\prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8}$$ by Lemma 4. since $a_i \leq X_i \leq b_i$ and $E(X_i - EX_i) = 0$

$$= e^{-st} e^{s^2\sum_{i=1}^{n}(b_i-a_i)^2/8}$$

$$= e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2}$$ by choosing $s = \underset{s}{\operatorname{argmin}}\left(-st + \frac{s^2\sum_{i=1}^{n}(b_i-a_i)^2}{8}\right) = 4t\Big/\left(\sum_{i=1}^{n}(b_i-a_i)^2\right)$

IIIly for $P\{S_n - ES_n \leq -t\} = P\{ES_n - S_n \geq t\}$ ∎

Example $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p) \Rightarrow P\{S_n/n - p \geq \varepsilon\} \leq e^{-2n\varepsilon^2}$

Lemma 6 (Expected Maximal Deviation ≠)

Let $\sigma > 0$, $n \geq 2$, and let $Y_1, \ldots, Y_n$ be real-valued RVs such that for all $s>0$ and $1 \leq i \leq n$, $E\{e^{sY_i}\} \leq e^{s^2\sigma^2/2}$. Then

$$E\{\max_{i\leq n} Y_i\} \leq \sigma\sqrt{2\ln n}$$

If, additionally, $E\{e^{s(-Y_i)}\} \leq e^{s^2\sigma^2/2}$ for every $s>0$ and $1 \leq i \leq n$, Then for any $n \geq 1$,

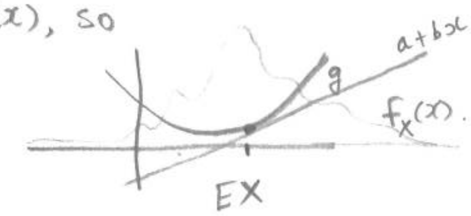
$$E\{\max_{i\leq n} |Y_i|\} \leq \sigma\sqrt{2\ln(2n)}$$

Proof: First let's recall Jensen's ≠ $\begin{cases} g \text{ is convex} \Rightarrow Eg(X) \geq g(EX) \\ g \text{ is concave} \Rightarrow Eg(X) \leq g(EX) \end{cases}$

Let $L(x) = a + bx$ be a line tangent to $g(x)$ at The point EX

Since g is convex it lies above $L(x)$, so

$$Eg(X) \geq EL(X) = E(a+bX) = a + bEX$$
$$= L(EX)$$
$$= g(EX)$$



Proof

By Jensen's $\neq$ and convexity of $\exp(\cdot)$, for all $s > 0$,

$$e^{sE\{\max_{i \leq n} Y_i\}} \leq E\{e^{s \max_{i\leq n} Y_i}\}$$
$$= E\{\max_{i \leq n} e^{sY_i}\}$$
$$\leq \sum_{i=1}^{n} E\{e^{sY_i}\}$$
$$\leq n e^{s^2\sigma^2/2}$$

Thus,
$$E\{\max_{i\leq n} Y_i\} \leq \frac{\ln n}{s} + \frac{s\sigma^2}{2}$$

for any $s > 0$, and taking
$$s = \operatorname{argmin}_s \frac{\ln n}{s} + \frac{s\sigma^2}{2} = \sqrt{2\ln n/\sigma^2}$$

$$\operatorname{argmin}_s \frac{\ln n}{s} + \frac{s\sigma^2}{2} = \sqrt{2 \ln n/\sigma^2}$$

$$\frac{d}{ds}\left((\ln n)s^{-1} + \frac{\sigma^2}{2}s\right) = 0$$

$$-\ln n\, s^{-2} + \frac{\sigma^2}{2} = 0$$

$$s^{-2} = \sigma^2/2\ln n$$

$$s = \sqrt{2\ln n/\sigma^2}$$

and $\frac{d^2}{ds^2} = \ln n\, s^{-3} > 0$

gives
$$E\{\max_{i\leq n} Y_i\} \leq \frac{\ln n}{\sqrt{2\ln n}/\sigma} + \frac{\frac{\sqrt{2\ln n}}{\sigma}\cdot \sigma^2}{2}$$
$$= \frac{2(\ln n)\sigma + (\sqrt{2\ln n})^2\sigma}{2\sqrt{2\ln n}} = \frac{\cancel{2}(2\ln n)\sigma}{\cancel{2}\sqrt{2\ln n}} = \sigma\sqrt{2\ln n}$$

$$E\{\max_{i\leq n} Y_i\} \leq \sigma\sqrt{2\ln n} \qquad (\star)$$

Finally, $\max_{i\leq n} |Y_i| = \max(Y_1, -Y_1, Y_2, -Y_2, \cdots, Y_n, -Y_n)$

and applying $(\star)$ to $2n$ terms we get

$$E\{\max_{i\leq n} |Y_i|\} \leq \sigma\sqrt{2\ln(2n)} \qquad \blacksquare$$

Lemma 7: [a conditional straightforward extension of Lemma 4]

Let V and Z be RVs such that $E\{V|Z\}=0$ with prob. 1, and for some function h and constant $c \geq 0$:

$$h(Z) \leq V \leq h(Z) + c.$$

Then for all $s > 0$:

$$E\{e^{sV}|Z\} \leq e^{s^2 c^2/8}$$

Now, we get to an extension of Hoeffding's $\neq$.

Thm 8 (Bounded Difference $\neq$)

Let A be some set and suppose the function $g: A^n \to \mathbb{R}$ satisfies the bounded difference assumption

$$\sup_{\substack{x_1,\dots,x_n,\\ x_i' \in A}} |g(x_1,\dots,x_n) - g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

(i.e., changing the i^{th} variable of g while fixing all others does not change the value of g by more than c_i)

and let $X_1,\dots,X_n$ be independent RVs. taking values in A.

Then, for all $t > 0$,

$$P\{g(X_1,\dots,X_n) - Eg(X_1,\dots,X_n) \geq t\} \leq e^{-2t^2/\sum_{i=1}^n c_i^2}$$

and

$$P\{Eg(X_1,\dots,X_n) - g(X_1,\dots,X_n) \geq t\} \leq e^{-2t^2/\sum_{i=1}^n c_i^2}$$

Proof [by martingale technique of Mc Diarmid (1989)] proof

Let $V = g(X_1, \ldots, X_n) - E g(X_1, \ldots, X_n) = g - Eg$

let $V_i = E\{g | X_1, \ldots, X_i\} - E\{g | X_1, \ldots, X_{i-1}\}$, $i = 1, \ldots, n$

Then $V = \sum_{i=1}^{n} V_i$.

Let $H_i(X_1, \ldots, X_i) = E\{g(X_1, \ldots, X_n) | X_1, \ldots, X_i\}$

and F_i be the distribution of X_i for $i = 1, \ldots, n$

Then,

$$V_i = H_i(X_1, \ldots, X_i) - \int H_i(X_1, \ldots, X_{i-1}, x) F_i(dx)$$

Let $W_i = \sup_u \left(H_i(X_1, \ldots, X_{i-1}, u) - \int H_i(X_1, \ldots, X_{i-1}, x) F_i(dx) \right)$

and $Z_i = \inf_v \left(H_i(X_1, \ldots, X_{i-1}, v) - \int H_i(X_1, \ldots, X_{i-1}, x) F_i(dx) \right)$

then $Z_i \le V_i \le W_i$ for $i = 1, \ldots, n$ with prob 1

and by bounded difference assumption:

$$W_i - Z_i = \sup_u \sup_v \left(H_i(X_1, \ldots, X_{i-1}, u) - H_i(X_1, \ldots, X_{i-1}, v) \right) \le c_i$$

Therefore by lemma 7, for $i = 1, \ldots, n$:

$$E\{e^{sV_i} | X_1, \ldots, X_{i-1}\} \le e^{s^2 c_i^2 / 8}$$

Using the fact that if X, Y are arbitrary bounded RVs

then $E\{XY\} = E\{E\{XY | Y\}\} = E\{Y E\{X | Y\}\}$

and Chernov's bounding method; for any $s > 0$:

$$P\{g - Eg > t\} = P\{V > t\} = P\{\sum_{i=1}^{n} V_i > t\}$$

$$\le \frac{E\{e^{s \sum_{i=1}^{n} V_i}\}}{e^{st}} = E\{e^{s \sum_{i=1}^{n-1} V_i} e^{sV_n}\}$$

$$= \frac{E\{e^{s \sum_{i=1}^{n-1} V_i} E\{e^{sV_n} | X_1, \ldots, X_{n-1}\}\}}{e^{st}}$$

13

$$\leq e^{s^2 c_n^2/8} \frac{E\{e^{s\sum_{i=1}^{n-1} V_i}\}}{e^{st}}$$

$\vdots$ (by repeating n times)

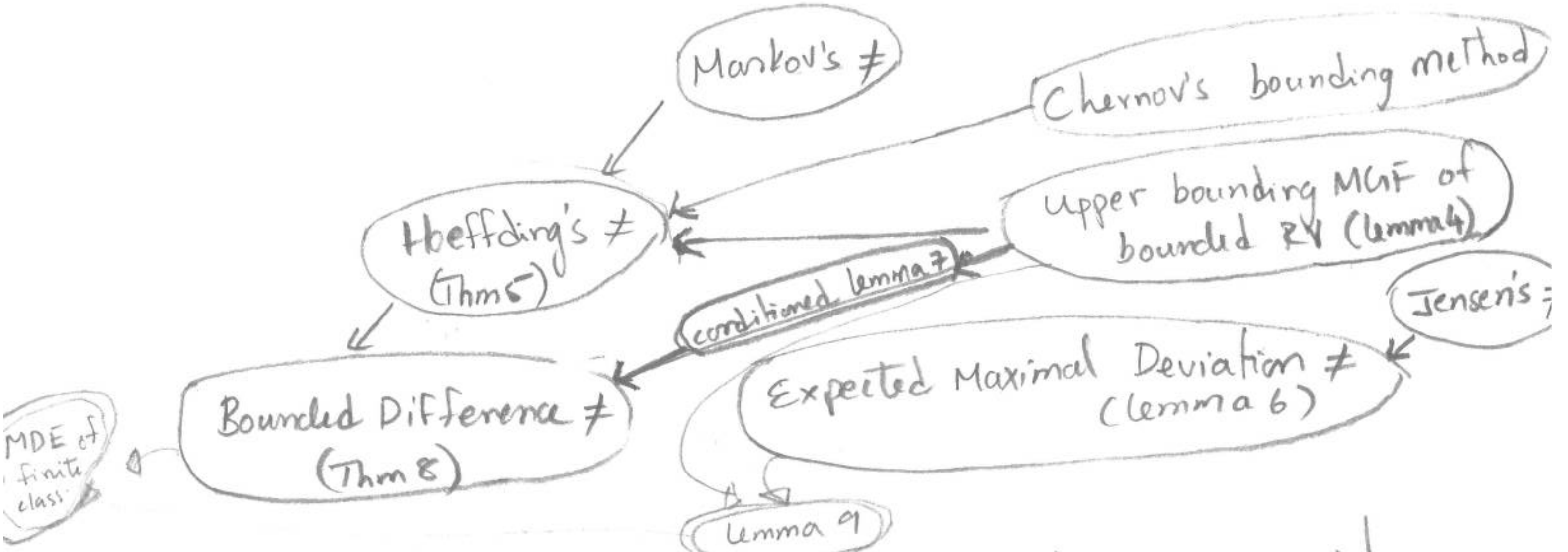
$$\leq e^{-st} e^{s^2 \sum_{i=1}^{n} c_i^2/8}$$

Now choosing the $s = 4t / \sum_{i=1}^{n} c_i^2$ gives

$$P\{g - Eg \geq t\} \leq e^{-\frac{4t^2}{\sum_{i=1}^n c_i^2} + \frac{\cancel{16}^{2} t^2}{(\sum_{i=1}^n c_i^2)^{\cancel{2}}} \frac{\cancel{\sum_{i=1}^n c_i^2}}{\cancel{8}}} = e^{-2t^2/\sum_{i=1}^n c_i^2}$$

$P\{Eg - g \geq t\} \leq e^{-2t^2/\sum_{i=1}^n c_i^2}$ is proved similarly ▨

Recall our main reason for <u>tour of concentration ≠s</u>



is to upper bound $\Delta = g(X_1, \dots, X_n) = \sup_{A \in \mathcal{A}} |\mu(A) - \mu_n(A)|$

where $X_1, \dots, X_n$ are iid RVs taking values in $\mathcal{X}$,

$\mathcal{A}$ is a collection of subsets of $\mathcal{X}$ with $\mu(A) = P\{X_1 \in A\}$

and $\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}$, for any $A \in \mathcal{A}$.

Regardless of the nature of $\mathcal{A}$, by changing one X_i,

g can change by at most $1/n$.

Therefore by bounded difference ≠ :

$$P\left\{\left|\sup_{A\in\mathcal{A}}|\mu(A)-\mu_n(A)| - E\left\{\sup_{A\in\mathcal{A}}|\mu(A)-\mu_n(A)|\right\}\right| > t\right\}$$

$$\leq 2e^{-2t^2/\sum_{i=1}^{n}(\frac{1}{n})^2} = 2e^{-2t^2/\frac{n}{n^2}} = 2e^{-2nt^2}.$$

for any n and $t>0$.

This shows that ~~The~~ for any class $\mathcal{A}$, the maximal deviation of empirical measure μ_n from true measure μ is sharply concentrated around its expected value. ($\Delta = g \to 0$ in probability if and only if $g \to 0$ almost surely).
Thus, we only worry about bounding the expected value of The maximal deviation, i.e

$$E\left\{\sup_{A\in\mathcal{A}}|\mu(A)-\mu_n(A)|\right\}.$$

To do this in generality we need <u>uniform deviation $\neq$s</u>

but when $|\mathcal{A}|<\infty$ as in our Thm 2's MDE setting of <u>selecting from k densities</u> f_{ni}, $1\leq i\leq k$, $\int f_{ni}=1$ for all i, with Yatracos class $\mathcal{A}=\{A_{ij}, A_{ji}: 1\leq i<j\leq k\}$ made up of Scheffé sets $A_{ij}=\{x: f_{ni}(x)>f_{nj}(x)\}$ then all we need are lemmas 4 and 6.

<u>Lemma 9</u> If $\mathcal{A}=\{\{x: f_{ni}(x)>f_{nj}(x)\};\ i,j\in\{1,\dots,k\}, i\neq j\}$ with $k<\infty$

Then $E\Delta = E\left\{\sup_{A\in\mathcal{A}}|\mu(A)-\mu_n(A)|\right\} \leq \sqrt{\frac{\ln(2k^2)}{2n}}$

<u>proof</u> Note that $\mathcal{A} = \{A_{ij}, A_{ji} : 1 \le i < j < k\}$ has at most k^2 sets.

Let $X_A = \mu(A) - \mu_n(A)$ for each $A \in \mathcal{A}$. For $s > 0$,

$$E\{e^{sX_A}\} = E\left\{e^{s\left(\mu(A) - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \in A\}}\right)}\right\}$$

$$= E\left\{e^{s\left(\frac{1}{n}\sum_{i=1}^n (\mu(A) - \mathbb{1}_{\{X_i \in A\}})\right)}\right\}$$

$$= \prod_{i=1}^n E\left\{e^{s(\mu(A) - \mathbb{1}_{\{X_i \in A\}})/n}\right\}$$ by independence of $X_1, \dots, X_n$.

$$\le \prod_{i=1}^n e^{s^2 (1/n)^2/8}$$

by lemma 4 since $E\{(\mu(A) - \mathbb{1}_{\{X_i \in A\}})/n\} = 0$ and $\mu(A) \le \mu(A) - \mathbb{1}_{\{X_i \in A\}} \le \mu(A) + \frac{1}{n}$

$$= e^{s^2/8 \sum_{i=1}^n 1/n^2}$$

$$= e^{s^2 \left(\frac{1}{2\sqrt{n}}\right)^2/2}$$

Note that we also have $E\{e^{s(-X_A)}\} \le e^{s^2\left(\frac{1}{2\sqrt{n}}\right)^2/2}$.

Thus for each $A \in \mathcal{A}$, $X_A = \mu(A) - \mu_n(A)$, $E\{e^{sX_A}\} \le e^{s^2\sigma^2/2}$, where $\sigma = \frac{1}{2\sqrt{n}}$. Therefore by lemma 6 (expected maximal deviation $\neq$)

$$E\left\{\max_{A \in \mathcal{A}} |\mu(A) - \mu_n(A)|\right\} = E\left\{\max_{A \in \mathcal{A}} X_A\right\}$$

$$\le \frac{1}{2\sqrt{n}} \sqrt{2 \ln(2|\mathcal{A}|)} \le \frac{1}{2\sqrt{n}} \sqrt{2 \ln(2k^2)}$$

$$= \sqrt{\frac{\ln(2k^2)}{2n}} \qquad \blacksquare$$

Finally from <u>Lemma 9</u> <u>and</u> taking Expectations in <u>Thm 2</u> we get our full asymptotic story for MDE for the problem of selecting from k densities.

$\boxed{\text{Corollary 10}}$

$$E\left(\int |\psi_n - f|\right) \le 3 \min_{i \le k} \int |f_{ni} - f| + \sqrt{\frac{8 \ln(2k^2)}{n}}$$

Corollary 10 has applications in situations where k is allowed to grow with sample size n, say as k_n, such that $\min_{1\le i\le k_n}\int|f_{ni}-f| \xrightarrow[n\to\infty]{} 0$ as $k_n\to\infty$

and $\sqrt{\dfrac{8\ln(2k_n^2)}{n}} \longrightarrow 0$, for e.g. $\ln(k_n)\le\sqrt{n} \Leftrightarrow k_n\le\exp(n^{10/20})$

concretely,

n	$k_n=\lfloor\exp(n^{7/20})\rfloor$	$\sqrt{\frac{8\ln(2k_n^2)}{n}}$	$k_n=\lfloor\exp(n^{5/20})\rfloor$	$\sqrt{\frac{8\ln(2k_n^2)}{n}}$
100	150	0.926	23	0.750
1,000	7,4621	0.430	276	0.310
10,000	$\approx 8.1\times10^{10}$	0.202	22,026	0.129
100,000	$\approx 2.6\times10^{24}$	0.095	$\approx 5.2\times10^{7}$	0.054
1000,000	$\approx 4.7\times10^{54}$	0.045	$\approx 5.4\times10^{13}$	0.023

How to control first error term $3\min_{1\le i\le k}\int f_{ni}-f$?

Situation 1

* If $f\in\{f_{ni}: 1\le i\le k\}$ Then $\min_{1\le i\le k}\int f_{ni}-f=0$

(this is useful in some simulation settings only)

Situation 2

* suppose $f\in\tilde{\mathcal{F}}$, a class of totally bounded densities i.e., $\forall\varepsilon>0, \exists\, \mathcal{G}_\varepsilon=\{g_1,g_2,\dots,g_{N_\varepsilon}\}\subseteq\tilde{\mathcal{F}}$, s.t.

$$\tilde{\mathcal{F}}\subseteq\bigcup_{i=1}^{N_\varepsilon}B_{g_i,\varepsilon}\quad,\quad B_{g,r}=\{f:\int|f-g|\le r\}$$

The smallest such N_ε is the complexity or covering number of $\tilde{\mathcal{F}}$, $\boxed{\log_2 N_\varepsilon}$ is the Kolmogorov entropy of $\tilde{\mathcal{F}}$ and $\mathcal{G}_\varepsilon$ is the skeleton of $\tilde{\mathcal{F}}$

Data compression # bits to store an integer between $1,\dots,N_\varepsilon$

Define The Yatracos Class

$$\mathcal{A}_\varepsilon = \{ \{x : g_i(x) > g_j(x)\} ; \ g_i, g_j \in G_\varepsilon \}$$

with $|\mathcal{A}_\varepsilon| \leq N_\varepsilon^2$

Now, the MDE of Yatracos (skeleton estimate) is:

$$\psi_n = \underset{g_i \in G_\varepsilon}{\operatorname{argmin}} \ \sup_{A \in \mathcal{A}_\varepsilon} \left| \int_A g_i - \mu_n(A) \right|$$

Thus, by corollary 10 and Dfn of N_ε we have:

$$E\left(\int |\psi_n - f| \right) \leq 3 \min_{g_i \in G_\varepsilon} \int |g_i - f| + \sqrt{\frac{8 \ln(2 N_\varepsilon^2)}{n}}$$

$$= 3\varepsilon + \sqrt{\frac{8 \ln(2N_\varepsilon^2)}{n}} \qquad \left[\varepsilon > 0 \text{ s.t. } N_\varepsilon \geq 2 \right]$$

Remark. Regardless of how quickly $N_\varepsilon \to \infty$ as $\varepsilon \to 0$, we can choose an $\varepsilon = \varepsilon_n = \inf\{\varepsilon : \log N_\varepsilon \leq \sqrt{n}\}$ so that the expected L_1 error of the MDE ψ_n converges to 0 uniformly, provided $f \in \tilde{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ is totally bounded.

Situation 3 (realistic). But if $f \notin \tilde{\mathcal{F}}$ and $f \in \mathcal{F}$, the set of all L_1 densities.

then by Corollary 10

$$E\left(\int |\psi_n - f| \right) \leq 3 \min_{g \in G_\varepsilon} \int |g - f| + \sqrt{\frac{8 \ln(2N_\varepsilon^2)}{n}}$$

$$\leq 3 \min_{g \in \tilde{\mathcal{F}}} \int |g - f| + 3\varepsilon + \sqrt{\frac{8 \ln(2N_\varepsilon^2)}{n}}$$

(by dfn. N_ε an $\Delta^k \neq$)

MDE is robust and accounts for the distance between true f and The best estimate in the class $\tilde{\mathcal{F}}$ we are looking in. If f is close to $\tilde{\mathcal{F}}$ then we are o[...]

<u>Situation 4</u> $f \in L_1$ and we want to construct from data $X_1, \dots, X_n$ a set of densities $\{f_{ni} : 1 \le i \le k_n\}$ s.t.

$$3 \ \min_{1 \le i \le k_n} \int |f_{ni} - f| \longrightarrow 0 \quad \text{and} \quad 4 \ \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_n(A) \right| \to 0$$

Want data-adaptive strategies for constructing $\{f_{ni} : 1 \le i \le k_n\}$.

Want better bounds for this that that based on $|\mathcal{A}|$.

<u>Thm 11</u> (Vapnik-Chervonenkis $\neq$)

$$E\left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right\} \le 2\sqrt{\frac{\ln(2 S_{\mathcal{A}}(n))}{n}}$$

where $S_{\mathcal{A}}(n) = \max_{x_1, \dots, x_n \in \mathbb{R}^d} \left| \{ \{x_1, \dots, x_n\} \cap A ; A \in \mathcal{A} \} \right|$.

$S_{\mathcal{A}}(n)$ is V-C shatter coefficient and gives the maximal number of different subsets of a set of n points that can be obtained by intersecting with elements of $\mathcal{A}$

<u>Proof</u> (Giné & Zinn (1984))

<u>Introduce</u> $X_1', \dots, X_n'$, as an independent copy of $X_1, \dots, X_n$

$\sigma_1, \dots, \sigma_n$, as n iid sign variables with $P\{\sigma = -1\} = P\{\sigma = +1\} = \frac{1}{2}$ that are also independent of $X_1, X_1', X_2, X_2', \dots, X_n, X_n'$.

Let $\mu_n'(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[X_i' \in A]}$

Then,

$$
\begin{aligned}
& E\left\{\sup_{A\in\mathcal{A}} |\mu_n(A)-\mu(A)|\right\} \\
&= E\left\{\sup_{A\in\mathcal{A}} \left| E\{\mu_n(A)-\mu_n'(A) \mid X_1,\dots,X_n\}\right|\right\} \\
&\le E\left\{\sup_{A\in\mathcal{A}} E\{|\mu_n(A)-\mu_n'(A)| \mid X_1,\dots,X_n\}\right\} && \text{(by Jensen's } \neq \text{ and convexity of } |\cdot| \\
&\le E\left\{\sup_{A\in\mathcal{A}} |\mu_n(A)-\mu_n'(A)|\right\} && \begin{pmatrix} \sup E(\cdot) \le E\sup(\cdot) \\ \& \ E(E(\cdot|\cdot)) = E(\cdot) \end{pmatrix} \\
&= \frac{1}{n} E\left\{\sup_{A\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i (\mathbb{1}_{[X_i\in A]} - \mathbb{1}_{[X_i'\in A]})\right|\right\} && \begin{pmatrix} \text{since} \\ X_1, X_1', \dots, X_n, X_n' \text{ are} \\ \text{iid.} \end{pmatrix} \\
&= \frac{1}{n} E\left\{E\left\{\sup_{A\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i (\mathbb{1}_{[X_i\in A]} - \mathbb{1}_{[X_i'\in A]})\right| \;\middle|\; X_1, X_1', \dots, X_n, X_n'\right\}\right\}
\end{aligned}
$$

Since σ_i's are independent of $X_1, X_1', \dots, X_n, X_n'$, let us fix $X_1 = x_1$, $X_1' = x_1'$, $\dots$, $X_n = x_n$, $X_n' = x_n'$ and investigate

$$
E\left\{\sup_{A\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i (\mathbb{1}_{[x_i\in A]} - \mathbb{1}_{[x_i'\in A]})\right|\right\}.
$$

Let $\hat{\mathcal{A}} \subseteq \mathcal{A}$ be a collection of sets s.t. any two sets in $\hat{\mathcal{A}}$ have different intersections with the set

$$\{x_1, x_1', \dots, x_n, x_n'\}$$

and every possible intersection is represented once

Thus, $|\hat{\mathcal{A}}| \le S_{\mathcal{A}}(2n)$, and

$$
\begin{aligned}
& E\left\{\sup_{A\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i (\mathbb{1}_{[x_i\in A]} - \mathbb{1}_{[x_i'\in A]})\right|\right\} \\
&= E\left\{\max_{A\in\hat{\mathcal{A}}} \left|\sum_{i=1}^n \sigma_i (\mathbb{1}_{[x_i\in A]} - \mathbb{1}_{[x_i'\in A]})\right|\right\}
\end{aligned}
$$

Now, since each $\sigma_i (1_{[x_i \in A]} - 1_{[x_i' \in A]})$ has mean zero and range $[-1, 1]$, by Lemma 4, we have (upperbounding MGF of a bounded RV). for any $s > 0$:

$$E\left\{ e^{s \sum_{i=1}^{n} \sigma_i (1_{[x_i \in A]} - 1_{[x_i' \in A]})} \right\}$$

$$= \prod_{i=1}^{n} E\left\{ e^{s\sigma_i (1_{[x_i \in A]} - 1_{[x_i' \in A]})} \right\} \leq \prod_{i=1}^{n} e^{s^2 2^2/8} = e^{ns^2/2}.$$

Since the distrn. of $\sigma_i (1_{[x_i \in A]} - 1_{[x_i \in A]})$ is symmetric, Lemma 6 (expected Maximal Deviation $\neq$) implies that

$$E\left\{ \max_{A \in \hat{\mathcal{A}}} \left| \sum_{i=1}^{n} \sigma_i (1_{[x_i \in A]} - 1_{[x_i' \in A]}) \right| \right\} \leq \sqrt{2n \log 2 S_{\mathcal{A}}(2n)}$$

$$\leq \sqrt{2n \log 2 S_{\mathcal{A}}(n)^2}$$

$$(\text{since } S_{\mathcal{A}}(2n) \leq S_{\mathcal{A}}(n)^2)$$

by dividing by $\frac{1}{n}$ on both sides we get V-C $\neq$:

$$E\left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \right\} \leq 2\sqrt{\frac{\log 2 S_{\mathcal{A}}(n)}{n}} \quad \blacksquare$$