Density Estimation:

Random variable \( X \) has density \( f \) on \( \mathbb{R}^d \) when

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

[i.e., \( P\{x \in A\} = f(x) \cdot \text{Vol}(A) \) if \( A = \{x' \in \mathbb{R}^d : d(x', x) \leq r\} \) 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

\[
f_n(x) = f_n(x; x_1, \ldots, x_n) : (\mathbb{R}^d)^{n+1} \to \mathbb{R}.
\]

Quality of \( f_n \) measured by T.V. distance

\[
\text{error} = \sup_{B \in \mathcal{B}(\mathbb{R}^d)} \int_B |f_n - f| \, d\lambda
\]

If this is \( \leq \varepsilon \) then all probabilities will be estimated with errors not exceeding \( \varepsilon \).

The distance between two densities \( f \) and \( g \) can be measured by their \( L_1 \) distance \( \|f - g\|_1 \).

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

\[
\sup_{B \in \mathcal{B}(\mathbb{R}^d)} \int_B |f - g| \, d\lambda = \frac{1}{2} \int |f - g| \, d\lambda
\]
Proof
\[ \sup_{B \in \mathcal{B}(\mathbb{R}^d)} |f - g| = \int_{\mathbb{R}^d} (f - g) = \int_{\mathbb{R}^d} (g - f) = \frac{1}{2} \int |f - g| \]

\[ \Rightarrow 0 \leq \frac{1}{2} \int |f - g| \leq 1, \quad \int |f - g| < 0.02 \]

\[ \Rightarrow \text{differences in probabilities are at most } \frac{1}{2} \times 0.02 = 0.01. \]

3) Scale invariance
\[ \sup_{B} |P\{X \in B\} - P\{Y \in B\}| = \sup_{B} \left| P\{T(X) \in T(B)\} - P\{T(Y) \in T(B)\} \right| \]
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 \( T : \mathbb{R}^d \rightarrow \mathbb{S} \subseteq \mathbb{R}^k \), i.e., for any R.V.s \( X \) and \( Y \) on \( \mathcal{B}(\mathbb{R}^d) \)
\[ \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left| P\{X \in A\} - P\{Y \in A\} \right| \leq \sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left| P\{T(X) \in A\} - P\{T(Y) \in A\} \right| \]

Proof:
\[ \sup_{A \in \mathcal{B}(\mathbb{R}^k)} \left| P\{T(X) \in A\} - P\{T(Y) \in A\} \right| \]
\[ = \sup_{A \in \mathcal{B}(\mathbb{R}^k)} \left| P\{X \in T^{-1}(A)\} - P\{Y \in T^{-1}(A)\} \right| \]
\[ \leq \sup_{A \in \mathcal{B}(\mathbb{R}^k)} \left| P\{X \in A\} - P\{Y \in A\} \right| \]

Since \( \{T^{-1}(A) : A \in \mathcal{B}(\mathbb{R}^k)\} \subseteq \mathcal{B}(\mathbb{R}^d) \). \( \square \)
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 (\( S[\hat{f}_n - f] \) is small for any \( f \in \mathbb{L}_1 \) as \( n \) gets large).

Consider the simpler problem of choosing between two densities \( f_n \) and \( g_n \), i.e., construct \( \hat{f}_n \) such that,

\[
\int |\hat{f}_n - f| \leq \min \left( S[f_n - f], S[g_n - f] \right)
\]

**Ideas:**

- Use empirical measure \( M_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \in A]} = \frac{\# \text{ of data points in } A}{n} \)
- A Schéffe set of ordered pair \( (f_n, g_n) \) is:
  \[
  A = A(f_n, g_n) = \{ x : f_n(x) > g_n(x) \}
  \]

**Observe:**

If \( S[f_n - g_n] = 1 \) then by Schéffe's identity,

\[
\int |f_n - g_n| = 2 \sum_{A}(f_n - g_n) = 2 S[f_n - g_n]_{A}
\]

**to get Schéffe Estimate**

\[
\hat{f}_n = \begin{cases} 
  f_n & \text{if } |S[A]_{f_n - M_n(A)}| < |S[A]_{g_n - M_n(A)}| \\
  g_n & \text{otherwise}
\end{cases}
\]

**Eq. 1:** Given data: \( x_1, x_2, \ldots, x_n \), \( \mu_n(A) = 8/10 \), \( S[f_n] = 7/10 \), \( S[g_n] = 4/10 \).

**Eq. 2:** different data \( x_1, x_2, \ldots, x_n \) from \( f \)

If \( \mu_n(A) = 5/10 \), then \( f_n^* = g_n \).
Thm (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

\[
\int |f_n^* - f| \leq 3 \min (\int |f_n - f|, \int |g_n - f|) + 4 \max_{A \in \mathcal{A}} \left( \int_{\mathcal{A}} |f - \mathcal{M}_n(A)| \right),
\]

where \( \mathcal{A} = \{ \{f_n > g_n\}, \{g_n > f_n\} \} \).

Proof: (Let's prove a slightly more general Thm.)

Note: \( f_n^* \) has an error that is within \( E_n = 4 \sup_{A \in \mathcal{A}} \left( \int_{\mathcal{A}} |f - \mathcal{M}_n(A)| \right) \) of 3 times the best possible error among \( f_n \) and \( g_n \).

Now consider the problem of selecting from \( k \) densities. (Our main problem here)

\( f_{n_i}, 1 \leq i \leq k, \) \( \int f_{n_i} = 1 \) for all \( i \)

Let \( A_{ij} \) denote the Scheffé set \( A_{ij} = A(f_{n_i}, f_{n_j}) \)

\( \{ x : f_{n_i}(x) > f_{n_j}(x) \} \).

Let the Yatsaras class of such Scheffé sets be:

\( \mathcal{A} = \{ A_{ij}, A_{ji} : 1 \leq i < j \leq k \} \)

The Minimum Distance Estimate (MDE) \( \psi_n \) is the \( f_{n_i} \) of smallest index that minimizes

\[
\Delta_i = \sup_{A \in \mathcal{A}} \left( \int_{\mathcal{A}} |f_{n_i} - \mathcal{M}_n(A)| \right)
\]

\( \psi_n = f_{n_i^*}, \quad i^* = \min \{ \arg \min_{1 \leq i \leq k} \Delta_i \} \).
Thm 2 (Universal Performance Bound of MDE) [DL 6.3]

For the MDE \( \psi_n \), we have (for each \( n \))

\[
\int |\psi_n - f| \leq 3 \min_i \int |f_{n_i} - f| + 4 \Delta, \quad \Delta = \sup_{A \in \mathcal{F}} |S_f - M_A|
\]

**Proof:**

Let \( \psi_n = f_{n_i} \) and let \( f_{n_j} = \arg \min_{1 \leq k \leq n} \int |f_{n_k} - f| \).

Assume \( j \neq i \), then

\[
(\#) \quad \int |\psi_n - f| = \int |f_{n_i} - f| \leq \int |f_{n_j} - f| + \int |f_{n_i} - f_{n_j}| \quad (\text{of L.1})
\]

Now, assuming WLOG \( i < j \):

\[
\int |f_{n_i} - f_{n_j}| = 2 \sup_{A \in \mathcal{A}} \left| \int f_{n_i} - f_{n_j} \right| \quad (\text{by Sheffe's identity})
\]

\[
\leq 2 \sup_{A \in \mathcal{A}} \left| \int f_{n_i} - f_{n_j} \right| \quad (\text{by Thm 1})
\]

\[
\leq 2 \sup_{A \in \mathcal{A}} \left| \int f_{n_i} - M_A \right| + 2 \sup_{A \in \mathcal{A}} \left| \int f_{n_j} - M_A \right| \quad (\psi_n = f_{n_i} \text{ by assumption})
\]

\[
\leq 4 \sup_{A \in \mathcal{A}} \left| \int f_{n_j} - M_A \right| + 4 \sup_{A \in \mathcal{A}} \left| \int f_{n_i} - M_A \right| \quad (\Delta \leq \Delta_i)
\]

\[
\leq 4 \sup_{B \in \mathcal{B}(R^n)} \left| \int_{B} f_{n_j} - f_{B} \right| + 4 \Delta \quad (\Delta \leq \Delta_i)
\]

\[
(\ast \ast) \quad \int |f_{n_j} - f| \leq 4 \Delta
\]

\[
\int |\psi_n - f| = 3 \int |f_{n_i} - f| + 4 \Delta = 3 \min_i \int |f_{n_i} - f| + 4 \Delta
\]
Thus to obtain error bounds for MDE we need bounds on $\Delta = \sup_{A \in \mathcal{A}} |\mu(A) - \mu_n(A)|$ no matter what $f \in L_1$ is generating data $x_1, \ldots, x_n$.

We are interested in $\Delta$, the maximal deviation of $\mu_n$ from $\mu$ over $\mathcal{A}$:

$$g(x_1, \ldots, x_n) = \Delta = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$$

We will first show (using concentration $\neq \delta$) 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 $\neq \delta$) that this mean ($Eg = E\Delta$) can be bounded using combinatorial characteristics of $\mathcal{A}$. 


Using concentration \( \leq \delta \), i.e., \( \leq \delta \) of the form:

\[
P\{ |X_n - EX_n| \geq \delta \} \leq \frac{\delta^{nc}}{c}
\]

Recall Markov's \( \leq \) for any non-negative RV \( X \) and any \( t > 0 \):

\[
P\{ X \geq t \} \leq \frac{EX}{t}
\]

**Proof:**

\[
EX = \int_0^\infty x dF(x) = \int_0^t x dF(x) + \int_t^\infty x dF(x) \geq \int_t^\infty x dF(x) = t \int_t^\infty df(x) = t P\{ X > t \}
\]

Let \( S \in (0, \infty) \), then for any RV \( X \) and any \( t > 0 \), by Markov's \( \leq \),

\[
P\{ X \geq t \} = P\{ e^{sx} \geq e^{st} \} \leq E\{ e^{sx} \}
\]

In Chernov's bounding method we find \( \leq \) a \( s > 0 \) that minimizes \( E\{ e^{sx} \} / e^{st} \), the upper bound.

If \( X_1, X_2, \ldots, X_n \overset{iid}{\rightleftharpoons} X \), and \( S_n = \sum_{i=1}^n X_i \), we get

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

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

So finding tight bounds \( \leq \) good upper bounds for MGF 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 \( s > 0 \):

\[
E\{ e^{sx} \} \leq e^{s^2(b-a)^2 / 8s}
\]
proof: by convexity

\[ e^{x} \leq \frac{x-a}{b-a} e^{b} + \frac{b-x}{b-a} e^{a} \]

for \( a \leq x \leq b \).

Since \( E(x) = 0 \),

\[ E \{ e^{X} \} \leq E \{ \frac{X-a}{b-a} e^{b} + \frac{b-x}{b-a} e^{a} \} = \left( \frac{a}{b-a} \right) e^{b} + \left( \frac{b}{b-a} \right) e^{a} \]

\[ = p e^{b} + (1-p) e^{a} = \left( \frac{p e^{b} + (1-p) e^{a}}{e^{a}} \right) e^{a} \]

\[ = (1-p + p e^{(b-a)}) e^{a} = (1-p + p e^{(b-a)}) e^{a} \]

\[ \frac{d}{du} e^{u} = \phi(u) \]

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

now,

\[ \phi'(u) = \frac{1}{du} \phi(u) = -p + \frac{b e^{u}}{1 - p + p e^{u}} = -p + \frac{p}{p + (1-p) e^{u}} \]

Thus,

\[ \phi'(0) = \phi'(0) = 0 \]

and

\[ \phi''(u) = \frac{p (1-p) e^{-u}}{(1 - p + p e^{u})^2} \leq \frac{1}{4} \]

since \( 0 < p < 1 \) and \( e^{u} < 1 \).

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

\[ \phi(u) = \phi(0) + u \phi'(0) + \frac{u^2}{2} \phi''(\xi) \leq \frac{u^2}{2} \leq \frac{s^2(b-a)^2}{2} \]

Thm 5 (Hoeffding's #: ) Let \( X_1, \ldots, X_n \) be indep. RVs such that \( X_i \in [a_i, b_i] \) w.p. 1. Then for any \( t > 0 \),

\[ P \{ S_n - ES_n \geq t \} \leq e^{-2t^2/\sum_{i=1}^{n} (b_i - a_i)^2} \]

and

\[ P \{ S_n - ES_n \leq -t \} \leq e^{-2t^2/\sum_{i=1}^{n} (b_i - a_i)^2} \]
From Chernov's bounding method, for any $s > 0$, $t > 0$

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

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

by Lemma 4.

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

by choosing

$$s = \arg \min_{s} \left( -st + \frac{s^2 \sum_{i=1}^{n} (b_i - a_i)^2}{8} \right)$$

$$= 4t \left( \sum_{i=1}^{n} (b_i - a_i)^2 \right)$$

Similarly for

$$\Pr \{ S_n - E S_n \leq -t \} = \Pr \{ E S_n - S_n \geq t \}$$

Example

$$X_1, \ldots, X_n \text{ i.i.d. Bernoulli (p)} \Rightarrow \Pr \{ S_n/n - p \geq \varepsilon \} \leq e^{-2n \varepsilon^2}$$

Lemma 5 (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 $1 \leq i \leq n$, $E \{ e^{s Y_i} \} \leq e^{s \sigma^2/2}$

Then

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

If, additionally, $E \{ e^{s (-Y_i)} \} \leq e^{s \sigma^2/2}$ 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

$$\int g \text{ is convex } \Rightarrow E g(X) \geq g(E X)$$

$$g \text{ is concave } \Rightarrow E g(X) \leq g(E X)$$
Let \( L(x) = ax + bx \) be a line tangent to \( g(x) \) at the point \( EX \).

Since \( g \) is convex it lies above \( L(x) \), so
\[
E g(x) \geq E L(x) = E(a + bX) = a + bEX = \underbrace{L(EX)}_{=g(EX)}
\]

**Proof**

By Jensen's inequality and convexity of \( \exp(x) \), for all \( s > 0 \),
\[
E \{ \max_{i \leq n} Y_i \} \leq E \{ e^{s \max_{i \leq n} Y_i} \} \\
= E \{ \max_{i \leq n} e^{sY_i} \} \\
\leq \sum_{i=1}^{n} E \{ e^{sY_i} \} \\
\leq n \cdot e^{s \cdot \frac{\sigma^2}{2}}
\]

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

for any \( s > 0 \), and taking
\[
s = \text{argmin}_s \frac{\ln n}{s} + \frac{s \cdot \sigma^2}{2} = \sqrt{2 \ln n / \sigma^2}
\]
gives
\[
E \{ \max_{i \leq n} Y_i \} \leq \frac{\ln n}{\sqrt{2 \ln n / \sigma^2}} + \frac{2 \ln n \cdot \sigma^2}{2 \cdot 2 \ln n} = \frac{2 \ln n \sigma^2 + (\ln n)^2 \sigma^2}{2 \cdot 2 \ln n} = \frac{2 \ln n \sigma^2}{2 \cdot 2 \ln n} = \sigma \sqrt{2 \ln n} \quad (\star)
\]

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

and applying (\star) to \( 2n \) terms we get
\[
E \{ \max_{i \leq n} |Y_i| \} \leq \sigma \sqrt{2 \ln (2n)} \quad (\star\star)
\]
Lemma 7: [Extension of lemma 4]

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

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

Then for all $s > 0$:

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

Now, we get to an extension of tboofding'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

$$\sup_{x_i, \ldots, x_n, x'_i \in A} \left| g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n) \right| \leq c_i,$$

$$1 \leq i \leq n.$$ 

(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 $X_1, \ldots, X_n$ be independent RVs, taking values in $A$.

Then, for all $t > 0$,

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

and

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

Proof [by martingale technique of McDiarmid (1989)]
Proof

Let $V = g(X, \ldots, X_n) - E g(X, \ldots, X_n) = y - E g$

Let $V_i = E \{ g \mid X, \ldots, X_i \} - E \{ g \mid X, \ldots, X_{i-1} \}$, $i = 1, \ldots, n$

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

Let $H_i(x, \ldots, X_i) = E \{ g(x, \ldots, X_n) \mid x, \ldots, X_i \}$

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

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

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

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

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

and by bounded difference assumption:

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

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

$E \{ e^{SV_i} \mid X, \ldots, X_{i-1} \} \leq e^{S^2 c_i^2 / 8}$

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 $s > 0$:

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

$\leq E \{ e^{S \sum_{i=1}^{n} V_i} \} = E \{ e^{S \sum_{i=1}^{n} V_i / E[V]} \}$

$= E \{ e^{S \sum_{i=1}^{n} V_i} \} E \{ e^{S V_n \mid X, \ldots, X_n} \}$

$\leq e^{S^2 c_i^2 / 8}$
\[
\begin{align*}
\leq \ e^{-\frac{s^2 c^2}{18} E \left\{ e^{s \sum_{i=1}^{n} \xi_i^2} \right\} / e^{st}}.
\end{align*}
\]

(by repeating \(n\) times)

\[
\leq \ e^{-st \ e^{s \sum_{i=1}^{n} \xi_i^2 / 8}}
\]

Now choosing \(s = 4t / \sum_{i=1}^{n} \xi_i^2\) gives

\[
P \{ g - \mathbb{E} g \geq t \} \leq e^{-\frac{4t^2}{\sum_{i=1}^{n} \xi_i^2} + \left( \frac{2 \sum_{i=1}^{n} \xi_i^2}{\left( \sum_{i=1}^{n} \xi_i^2 \right)^2} \right) \sum_{i=1}^{n} \xi_i^2 \} = e^{-2t^2 / \sum_{i=1}^{n} \xi_i^2}
\]

\[
P \{ \mathbb{E} g - g \geq t \} \leq e^{-2t^2 / \sum_{i=1}^{n} \xi_i^2} \]

is proved similarly.

Recall our main reason for tour of concentration is

\[\text{Markov's} \quad \rightarrow \quad \text{Chernov's bounding method} \quad \rightarrow \quad \text{Hoeffding's} \quad \rightarrow \quad \text{Upper bounding MGF of bounded RV (Lemma 4)} \quad \rightarrow \quad \text{combined lemma?} \quad \rightarrow \quad \text{Expected Maximal Deviation} \quad \rightarrow \quad \text{Jensen's?}\]

is to upper bound \( \Delta = g(X_1, \ldots, X_n) = \sup_{A \in \mathcal{A}} |\mu(A) - \mathbb{E}g(A)| \)

where \(X_1, \ldots, X_n\) are iid RVs taking values in \(X\),
\(\mathcal{A}\) is a collection of subsets of \(X\) with \(\mu(A) = \mathbb{P}\{X_i \in A\}\) and \(\mathbb{E}g(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}\) for any \(A \in \mathcal{A}\).

Regardless of the nature of \(\mathcal{A}\), by changing one \(X_i\), \(g\) can change by at most \(\frac{1}{n}\).

Therefore the bounded difference is:
\[
P\{ \sup_{A \in \mathcal{A}} |M_n(A) - M_n'(A)| - E\{ \sup_{A \in \mathcal{A}} |M_n(A) - M_n'(A)| \} > t \} \\
\leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n} \sigma_i^2}} = 2e^{-\frac{2t^2}{\sigma^2}} = 2e^{-\frac{2nt^2}}
\]
for any \( n \) and \( t > 0 \).

This shows that for any class \( \mathcal{A} \), the maximal deviation of empirical measure \( M_n \) from true measure \( \mu \) 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.

\[
E\{ \sup_{A \in \mathcal{A}} |M_n(A) - M_n'(A)| \}^2
\]

To do this in generality, we need uniform deviation \( \Delta \)

but when \( |\mathcal{A}| < \infty \) as in our Thm 2's MDE setting of selecting from \( k \) densities \( f_{ni}, 1 \leq i \leq k, \sum_{i=1}^{k} f_{ni} = 1 \)

for all \( i \), \( \mathcal{A} = \{ A_{ij}, A_{ji}: 1 \leq i < j \leq k \} \) made up of Scheffe sets \( A_{ij} = \{ x: f_{ni}(x) > f_{nj}(x) \} \). Then all we need are lemmas 4 and 6.

**Lemma 9**: If \( \mathcal{A} = \{ A_{ij} \}; i, j \in \{1, \ldots, k\}, \) with \( k < \infty \)

\[ \text{Then } E \Delta = E \left\{ \sup_{A \in \mathcal{A}} |M_n(A) - M_n'(A)| \right\} \leq \sqrt{\frac{\ln(2k^2)}{2n}} \]
proof. Note that \( A = \{ A_{ij}, A_{ji} : 1 \leq i \leq j \leq k \} \) has at most \( k^2 \) sets.

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

\[
E \{ e^{s X_A} \} = \frac{E \{ e^{s (M(A) - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{X_i \in A})} \}}{\sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{X_i \in A}}
\]

\[
= \frac{1}{n} E \{ e^{s (M(A) - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{X_i \in A})} \} \text{ by independence of } X_1, \ldots, X_n
\]

\[
\leq \frac{1}{n} e^{s^2 \left( \frac{1}{2 n^2} \right) / 2}
\]

by lemma 4, since

\[
E \{ \left( M(A) - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{X_i \in A} \right)^2 / n \} = 0
\]

and

\[
M_n(A) \leq M(A) - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \in A} \leq M(A) + \frac{1}{n}
\]

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

Thus for each \( A \in \mathcal{A} \), \( X_A = M(A) - M_n(A) \), \( E \{ e^{s X_A} \} \leq e^{s^2 \left( \frac{1}{2 n^2} \right) / 2} \).

where \( \sigma = \frac{1}{2 n} \). Therefore by lemma 6 (expected maximal deviation ^ ?)

\[
E \{ \max_{A \in \mathcal{A}} | M(A) - M_n(A) | \} = \frac{1}{2 \sqrt{n}} \sqrt{ \ln(\phi(A)) } \leq \frac{1}{2 \sqrt{n}} \sqrt{ \ln(2k) } \]

\[
= \sqrt{ \frac{\ln(2k^2)}{2n} }
\]

Finally from lemma 9 and taking Expectations in Thm 2, we get our full asymptotic story for MDE for the problem of selecting from k densities.

Corollary 10

\[
E( \| \hat{f} - f \| ) \leq 2 \min \left\{ \left\| \hat{f}_n - f \right\| + \sqrt{ \frac{\ln(2k^2)}{n} } \right\}
\]
Corollary 10 has applications in situations where \( k \) is allowed to grow with sample size \( n \), say as \( k_n \), such that \[ \min_{1 \leq i \leq k_n} \int |f_{n i} - f| \to 0 \quad \text{as} \quad k_n \to \infty \]

and \[ \sqrt{\frac{\xi \ln(2k_n^2)}{n}} \to 0, \quad \text{for e.g.} \quad \ln(k_n) \leq \sqrt{n} \quad \implies \quad k_n = \exp(n^{1/2}) \]

Concretely,

<table>
<thead>
<tr>
<th>( n )</th>
<th>( k_n = \exp(n^{1/2}) )</th>
<th>( \frac{\xi \ln(2k_n^2)}{n} )</th>
<th>( k_n = \exp(n^{5/20}) )</th>
<th>( \frac{\xi \ln(2k_n^2)}{n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>150</td>
<td>0.926</td>
<td>2.3</td>
<td>0.750</td>
</tr>
<tr>
<td>1,000</td>
<td>7,4621</td>
<td>0.430</td>
<td>2.76</td>
<td>0.310</td>
</tr>
<tr>
<td>10,000</td>
<td>( \approx 8.1 \times 10^6 )</td>
<td>0.202</td>
<td>22.026</td>
<td>0.129</td>
</tr>
<tr>
<td>100,000</td>
<td>( \approx 2.6 \times 10^{24} )</td>
<td>0.095</td>
<td>( \approx 5.2 \times 10^4 )</td>
<td>0.054</td>
</tr>
<tr>
<td>1,000,000</td>
<td>( \approx 4.7 \times 10^{54} )</td>
<td>0.045</td>
<td>( \approx 5.4 \times 10^{13} )</td>
<td>0.023</td>
</tr>
</tbody>
</table>

How to control first error term \( \min_{1 \leq i \leq k} \int |f_{ni} - f| \) ?

**Situation 1**

If \( f \in \{ f_{ni} : 1 \leq i \leq k \} \) Then \( \min_{1 \leq i \leq k} \int |f_{ni} - f| = 0 \) (this is useful in some simulation settings only)

**Situation 2**

Suppose \( f \in \tilde{f} \), a class of totally bounded densities i.e., \( \forall \epsilon > 0, \exists \mathcal{G}_\epsilon = \{ g_1, g_2, \ldots, g_{N_{\epsilon}} \} \subseteq \tilde{f} \), s.t.

\[
\tilde{f} \subseteq \bigcup_{i=1}^{N_{\epsilon}} B_{g_i, \epsilon}, \quad B_{g, \epsilon} = \{ f : \| f - g \|_1 \leq \epsilon \}
\]

The smallest such \( N_{\epsilon} \) is the complexity or covering number of \( \tilde{f} \) and \( \log_2 N_{\epsilon} \) is the Kolmogorov entropy of \( \tilde{f} \)

and \( \mathcal{G}_\epsilon \) is the skeleton of \( \tilde{f} \).
Define the Yatracos class:

$A^*_e = \{ \{ x : g_i(x) > g_\delta(x) \} ; g_i, g_\delta \in G^2_e \}

with $|A^*_e| \leq N_e^2$.

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

$\psi_n = \arg\min_{A \in A^*_e} \sup_{g_i \in G^2_e} |g_i - M_n(A)|$

Thus, by corollary 10 and dfn of $N_e$ we have:

$E(\psi_n - f) \leq 3 \min_{g \in G^2_e} \int |g - f| + \sqrt{\frac{8 \log(2N_e^2)}{n}}$

$= 3 \varepsilon + \sqrt{\frac{8 \log(2N_e^2)}{n}}$  \hspace{1cm} \varepsilon > 0 \hspace{1cm} N_e \geq 2$

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

Situation 3 (realistic): But if $f \not\in \tilde{F}$ and $f \not\in \tilde{f}$, the set of all $L_1$ densities, then by corollary 10

$E(\psi_n - f) \leq 3 \min_{g \in G^2_e} \int |g - f| + \sqrt{\frac{8 \log(2N_e^2)}{n}}$

$\leq 3 \min_{g \in G^2_e} \int |g - f| + 3\varepsilon + \sqrt{\frac{8 \log(2N_e^2)}{n}}$

(by dfn. $N_e$ an $\Delta_e$)

MDE is robust and accounts for the distance between true $f$ and the best estimate in the class $\psi_n$. If $f$ is close to one of the $g_i$, $\psi_n$ will give the best estimate.
Situation 4. \( f \in L_1 \) and we want to construct \( \frac{\text{from data}}{X_1, \ldots, X_n} \) a set of densities \( \{ F_{n_i} : 1 \leq i \leq k_n \} \) s.t.

\[
3 \min_{1 \leq i \leq k_n} \int |F_{n_i} - f| \to 0 \quad \text{and} \quad 4 \sup_{A \in \mathcal{A}} |\text{Pr}_f(M_{n_i}(A)) - \text{Pr}_A(A)| \to 0
\]

Want data-adaptive strategies for constructing \( \{ F_{n_i} : 1 \leq i \leq k_n \} \).

Want better bounds for this that are based on \( |\mathcal{A}| \).

Theorem (Vapnik-Chervonenkis ≤)

\[
E \left\{ \sup_{A \in \mathcal{A}} |M_{n}(A) - \mu(A)| \right\} \leq 2 \sqrt{\frac{\ln(2S_A(n))}{n}}
\]

where \( S_A(n) = \max_{x_1, \ldots, x_n \in \mathbb{R}^d} \left| \{ x_1, \ldots, x_n \cap A : A \in \mathcal{A} \} \right| \).

\( S_A(n) \) 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 \( \mathcal{A} \).

Proof. (Giné & Zinn (1984))

Introduce \( X'_1, \ldots, X'_n \), as an independent copy of \( X_1, \ldots, X_n \)

\( \sigma_1, \ldots, \sigma_n \), as \( n \) i.i.d. sign variables with

\[
P\{ \sigma_1 = -1 \} = P\{ \sigma_1 = +1 \} = \frac{1}{2}
\]

that are also independent of \( X_1, X'_1, X_2, X'_2, \ldots, X_n, X'_n \).

Let \( \mu'_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[X'_i \in A]} \)
Then,
\[ E \left\{ \sup_{A \in \mathcal{A}} |M_n(A) - M(A)| \right\} \]
\[ = E \left\{ \sup_{A \in \mathcal{A}} \left[ E \left\{ |M_n(A) - M_n'(A)| \mid X_1, \ldots, X_n \right\} \right] \right\} \]
\[ \leq E \left\{ \sup_{A \in \mathcal{A}} E \left\{ |M_n(A) - M_n'(A)| \mid X_1, \ldots, X_n \right\} \right\} \quad \text{(by Jensen's inequality and convexity of } 1-t) \]
\[ = \frac{1}{n} E \left\{ \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i \left( 1_{[X_i \in A]} - 1_{[X_i' \in A]} \right) \right| \right\} \]
\[ = \frac{1}{n} E \left\{ \left| \sum_{i=1}^{n} \sigma_i \left( 1_{[X_i \in A]} - 1_{[X_i' \in A]} \right) \right| \mid X_1, X_1', \ldots, X_n, X_n' \right\} \]

Since \( \sigma_i \)'s are independent of \( X_1, X_1', \ldots, X_n, X_n' \), let us fix \( X_1 = x_1, X_1' = x_1', \ldots, X_n = x_n, X_n' = x_n' \) and investigate
\[ E \left\{ \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i \left( 1_{[X_i \in A]} - 1_{[X_i' \in A]} \right) \right| \right\} \]

Let \( \mathcal{A} \subseteq \mathcal{A} \) be a collection of sets such that any two sets in \( \mathcal{A} \) have different intersections with the set \( \{x_1, x_1', \ldots, x_n, x_n'\} \) and every possible intersection is represented once. Thus, \( |\mathcal{A}| \leq S_{\mathcal{A}}(2n) \), and
\[ E \left\{ \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i \left( 1_{[X_i \in A]} - 1_{[X_i' \in A]} \right) \right| \right\} \]
\[ = E \left\{ \max_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i \left( 1_{[X_i \in A]} - 1_{[X_i' \in A]} \right) \right| \right\} \]
Now, since each \( \bar{\sigma}_i (1 \mathbb{1}_{x_i \in A} - 1 \mathbb{1}_{x_i \in \hat{A}}) \) has mean zero and range \([-1, 1]\), by \textbf{Lemma 4}, we have (upperbounding MGF of a bounded RV).

\[
E \left\{ e^{s \sum_{i=1}^{n} \bar{\sigma}_i (1 \mathbb{1}_{x_i \in A} - 1 \mathbb{1}_{x_i \in \hat{A}})} \right\} \leq \prod_{i=1}^{n} E \left\{ e^{s^2 \mathbb{1}_{x_i \in A}} e^{s^2 \mathbb{1}_{x_i \in \hat{A}}} \right\} \leq \prod_{i=1}^{n} e^{s^2 2^{-1/2}} = e^{ns^{2/2}}.
\]

Since the distribution of \( \bar{\sigma}_i (1 \mathbb{1}_{x_i \in A} - 1 \mathbb{1}_{x_i \in \hat{A}}) \) is symmetric, \textbf{Lemma 6} (expected maximal deviation \#) implies that

\[
E \left\{ \max_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \bar{\sigma}_i (1 \mathbb{1}_{x_i \in A} - 1 \mathbb{1}_{x_i \in \hat{A}}) \right| \right\} \leq \sqrt{2n \log 2 S_2^2(2n)}
\]

\[
\leq \sqrt{2n \log 2 S_A(n)^2}
\]

(by dividing by \( n \) on both sides we get \( V-C \neq \))

\[
E \left\{ \sup_{A \in \mathcal{A}} \left| Mn(A)-M(\hat{A}) \right| \right\} \leq 2 \sqrt{\frac{\log 2 S_A(n)}{n}}
\]