Consistency and rates of convergence for maximum likelihood estimators via empirical process theory

Jon A. Wellner

University of Washington

- Talk at University of Washington,
  Department of Statistics, October 24, 2005
- *Email: jaw@stat.washington.edu*
  http: //www.stat.washington.edu/jaw/jaw.research.html

Outline

- Introduction I: maximum likelihood estimation
- Introduction II: empirical process theory
- Consistency via Glivenko-Cantelli theorems
- Consistency: examples
- Introduction II, part 2: more empirical process theory
- Rates of convergence via empirical process theory
- Rates for MLE: examples
- Problems and challenges

1. Introduction I: maximum likelihood estimation

- Setting: dominated families; i.i.d. sampling.
- $X_1, \ldots, X_n$ are i.i.d. with density p_{θ_0} with respect to some dominating measure μ where $p_{\theta_0} \in \mathcal{P} = \{p_\theta : \ \theta \in \Theta\}$ for Θ a parameter space.
- The likelihood is

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i)\,.$$

- **Definition:** A Maximum Likelihood Estimator (or MLE) of θ_0 is any value $\hat{\theta} \in \Theta$ satisfying

$$L_n(\hat{\theta}) = \sup_{\theta \in \Theta} L_n(\theta)\,.$$

- Equivalently, the MLE $\hat{\theta}$ maximizes the log-likelihood

$$\frac{1}{n}\log L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log p_\theta(X_i) = \mathbb{P}_n \log p_\theta(X)$$

where $\mathbb{P}_n$ is the empirical measure,

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}.$$

- **Example 1.** Exponential (θ). $X_1, \ldots, X_n$ are i.i.d. p_{θ_0} where

$$p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x).$$

- Then the likelihood is

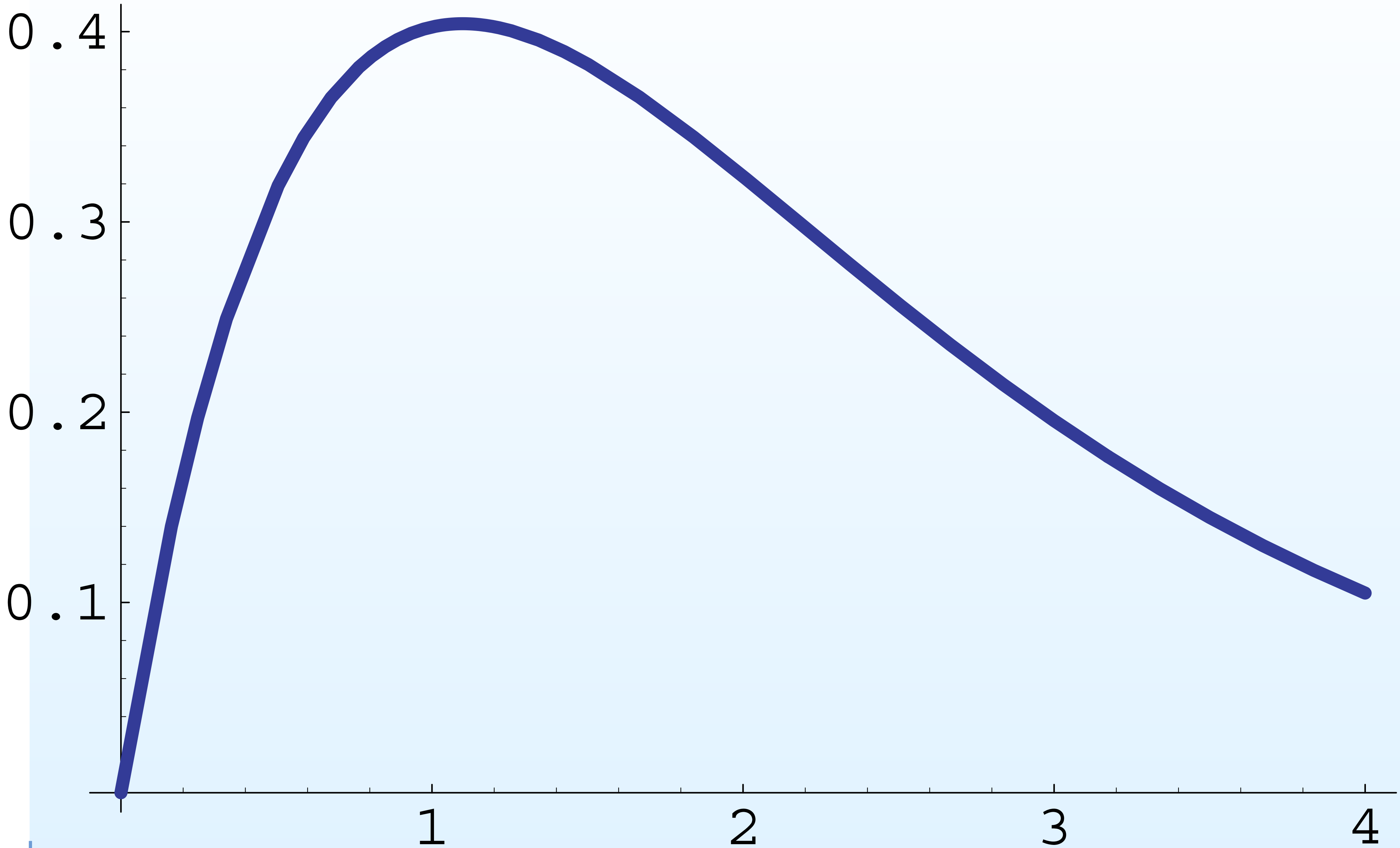
$$L_n(\theta) = \theta^n \exp(-\theta \sum_1^n X_i),$$

- so the log-likelihood is

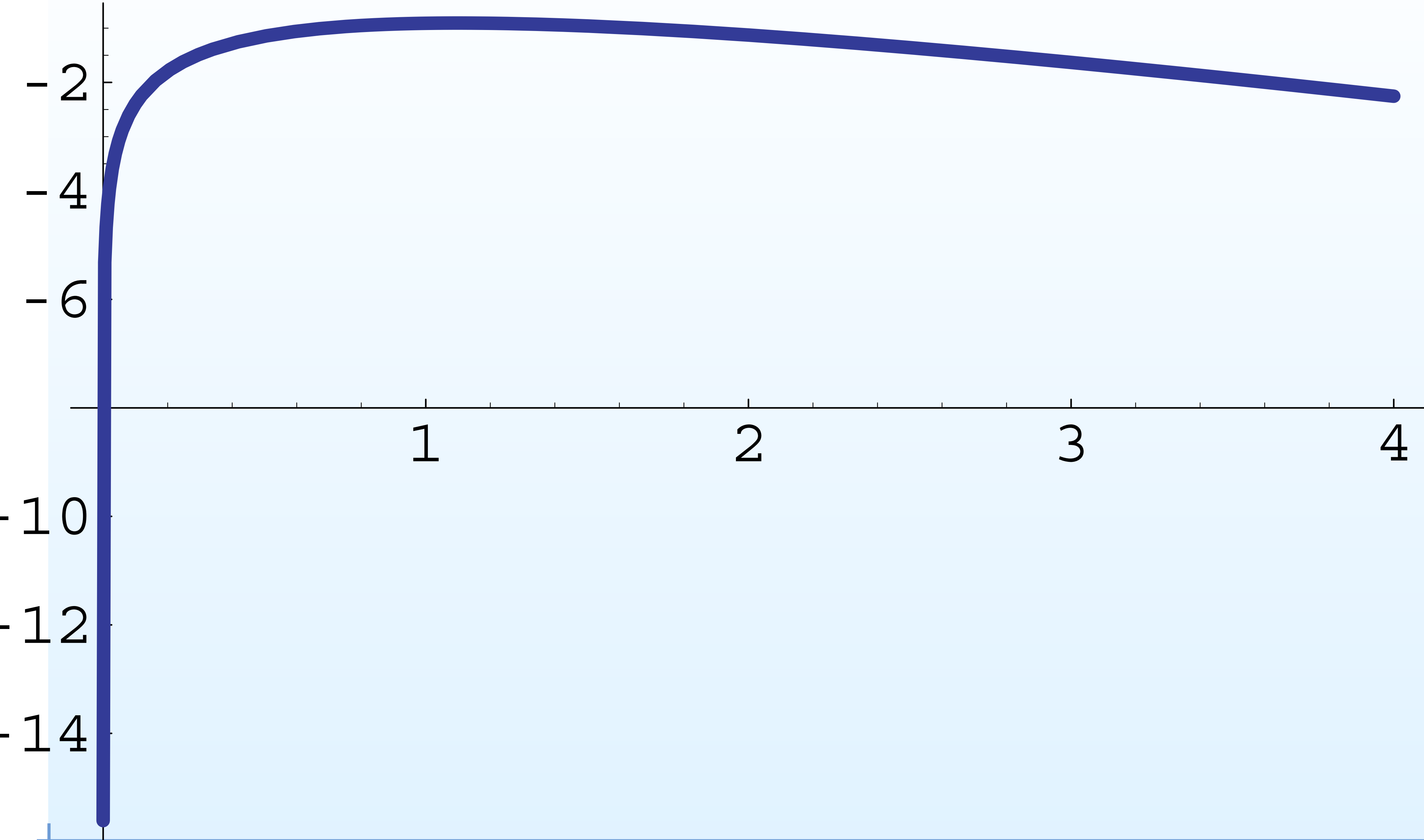
$$\log L_n(\theta) = n \log(\theta) - \theta \sum_1^n X_i$$

- and $\hat{\theta}_n = 1/\overline{X}_n$.

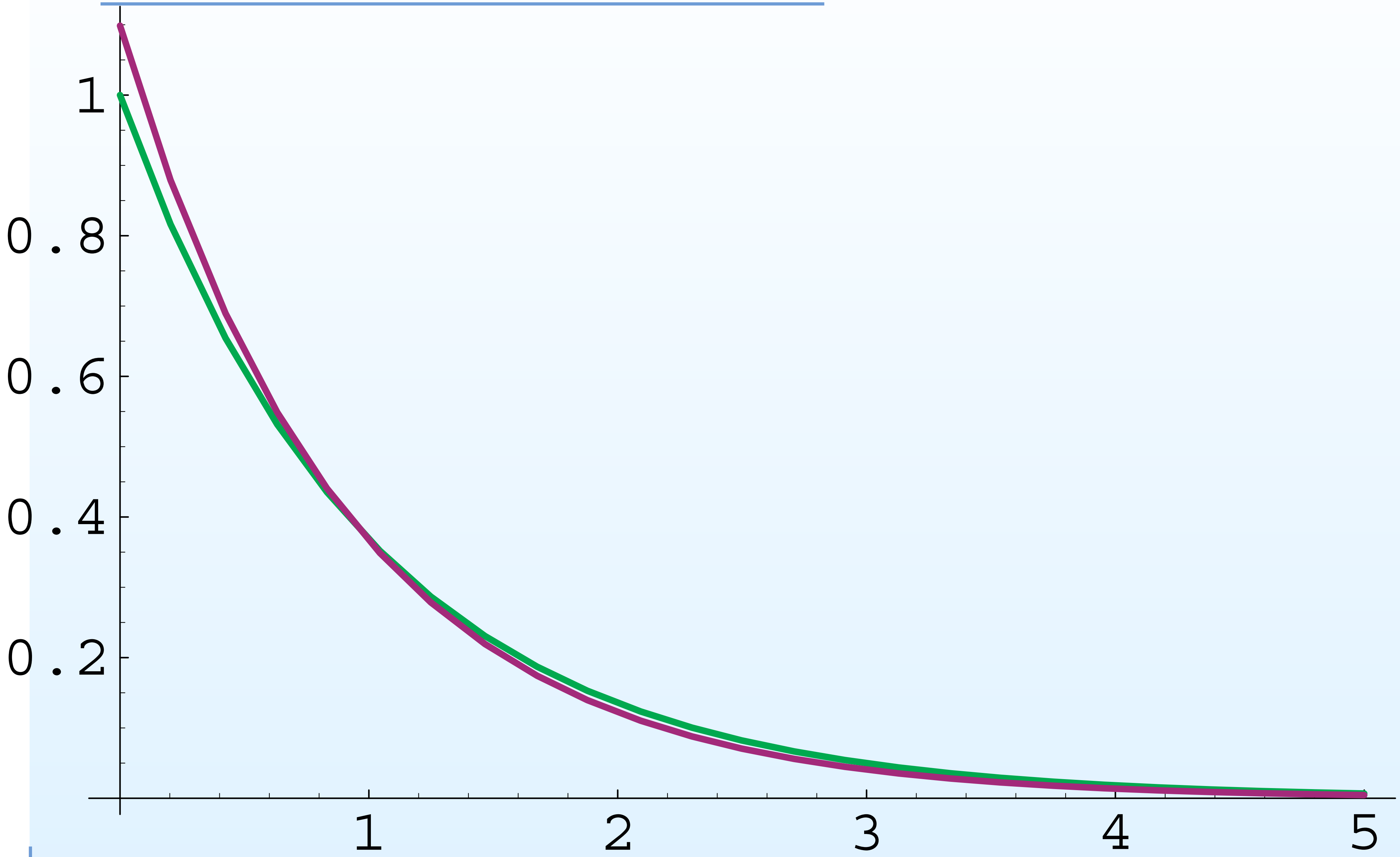
1/n power of likelihood, $n = 50$



1/n times log-likelihood, $n = 50$



MLE $p_{\hat{\theta}}(x)$ and true density $p_{\theta_0}(x)$



- **Example 2.** Monotone decreasing densities on $(0, \infty)$. $X_1, \ldots, X_n$ are i.i.d. $p_0 \in \mathcal{P}$ where

$$\mathcal{P} = \text{ all nonincreasing densities on } (0, \infty).$$

- **Example 2.** Monotone decreasing densities on $(0, \infty)$. $X_1, \ldots, X_n$ are i.i.d. $p_0 \in \mathcal{P}$ where

$$\mathcal{P} = \text{ all nonincreasing densities on } (0, \infty).$$

- Then the likelihood is $L_n(p) = \prod_{i=1}^n p(X_i)$;

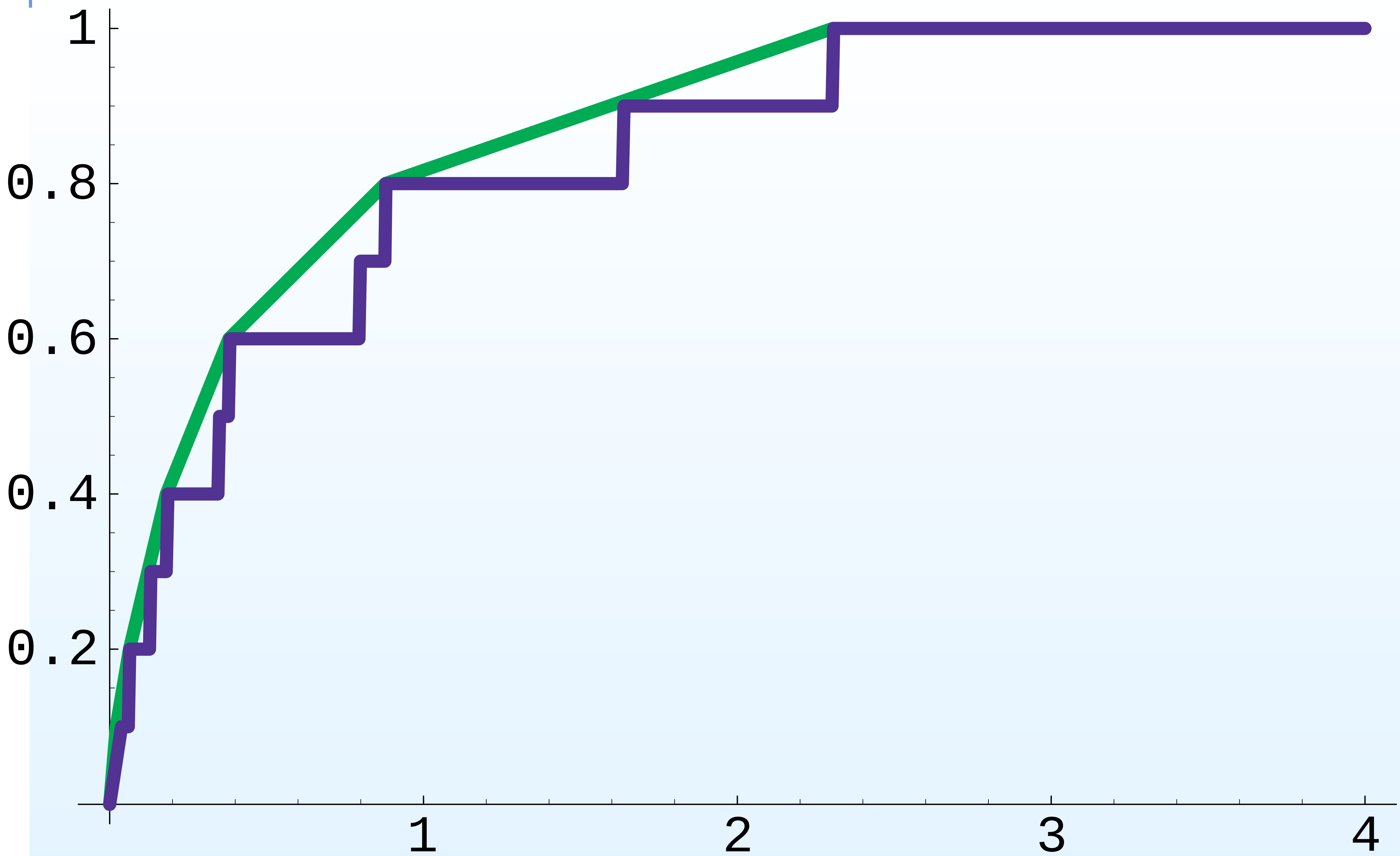
- **Example 2.** Monotone decreasing densities on $(0, \infty)$. $X_1, \ldots, X_n$ are i.i.d. $p_0 \in \mathcal{P}$ where

$$\mathcal{P} = \text{ all nonincreasing densities on } (0, \infty).$$

- Then the likelihood is $L_n(p) = \prod_{i=1}^n p(X_i)$;
- $L_n(p)$ is maximized by the Grenander estimator:

$$\hat{p}_n(x) \quad = \quad \text{left derivative at x of the Least Concave Majorant } \mathbb{C}_n \text{ of } \mathbb{F}_n$$

where $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^n 1\{X_i \leq x\}$


1
0.8
0.6
0.4
0.2
1
2
3
4

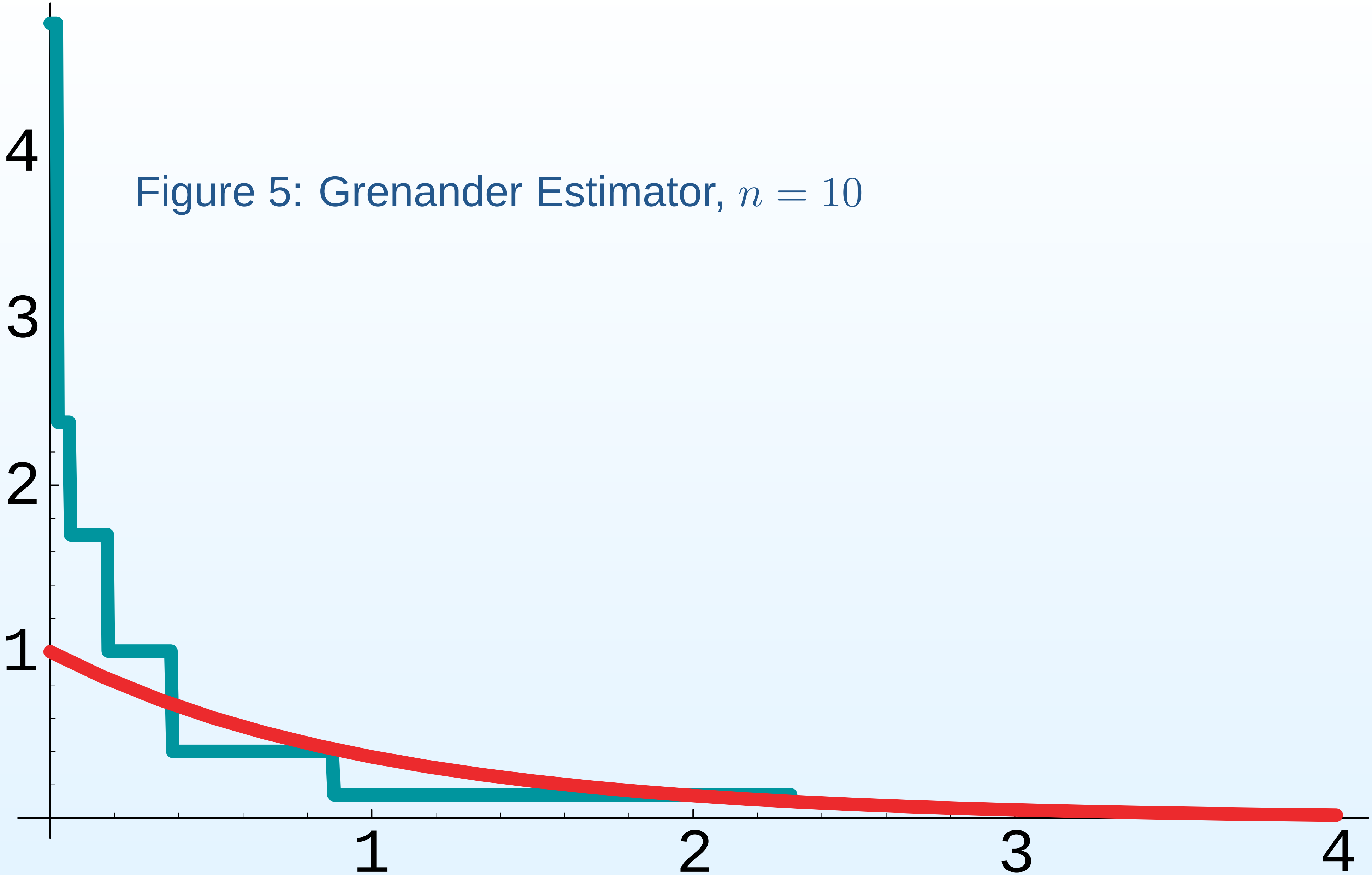
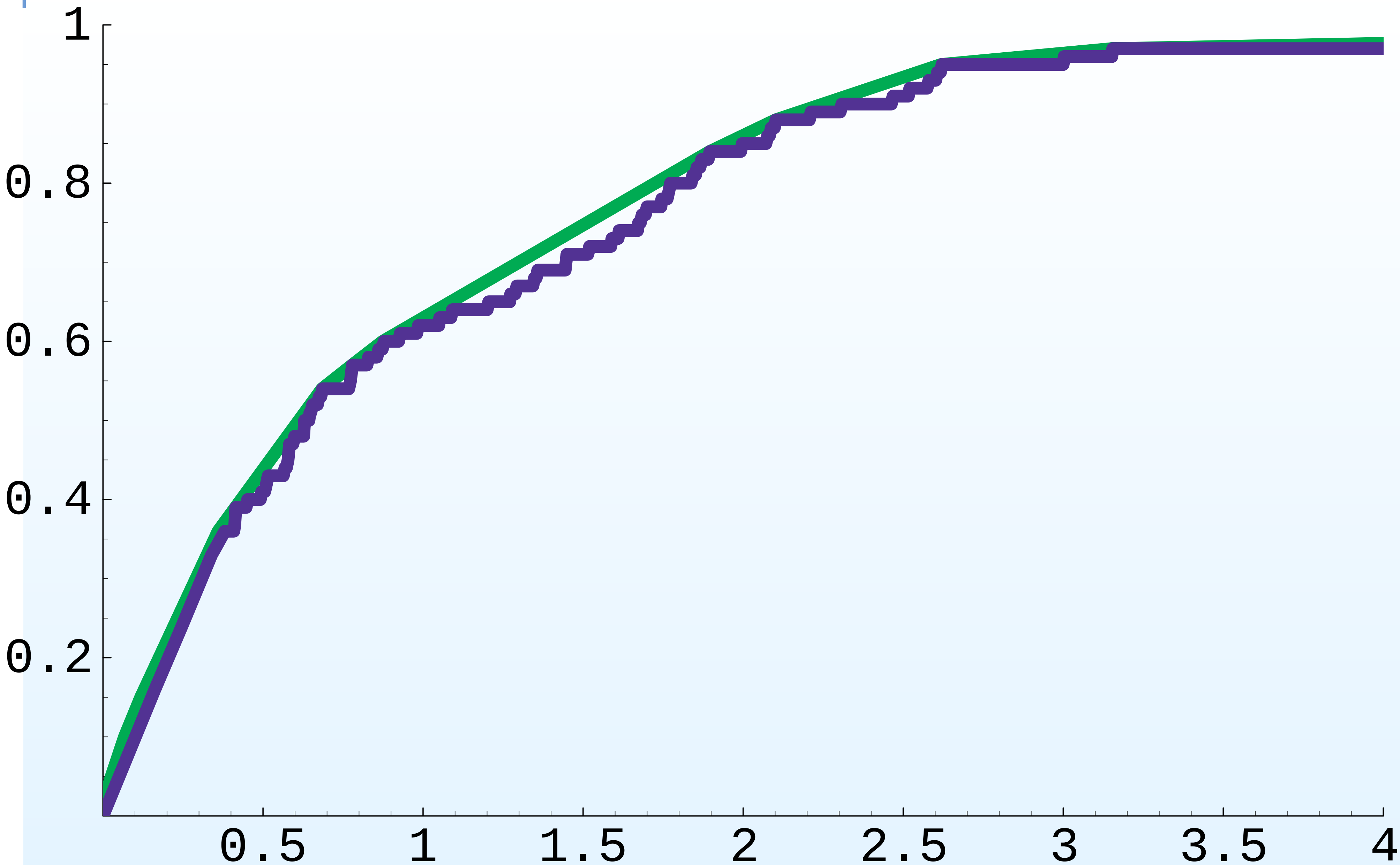


Figure 5: Grenander Estimator, $n = 10$



1
0.8
0.6
0.4
0.2
0.5
1
1.5
2
2.5
3
3.5
4

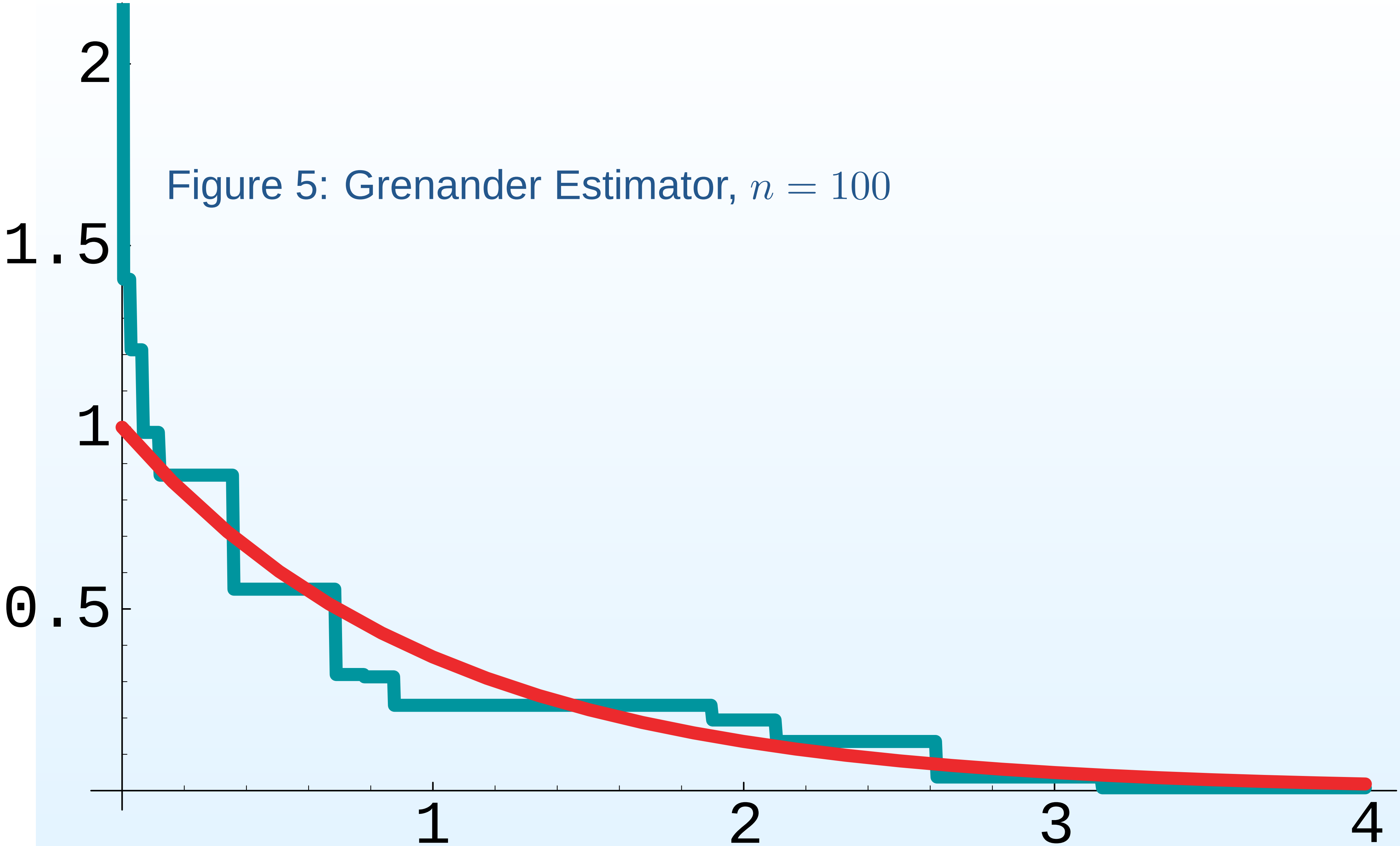


Figure 5: Grenander Estimator, $n = 100$
2
1.5
1
0.5
1
2
3
4

2. Introduction II: empirical process theory

- $X_1, \ldots, X_n$ are i.i.d. P on $(\mathcal{X}, \mathcal{A})$
- The empirical measure of the sample is

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

where

$$\delta_x(A) = 1_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

- Thus for a set $A \in \mathcal{A}$

$$\begin{aligned} \mathbb{P}_n(A) &= \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(A) = \frac{1}{n}\sum_{i=1}^n 1_A(X_i) \\ &= \frac{\#\{1 \le i \le n : \quad X_i \in A\}}{n}. \end{aligned}$$

- For a (measurable) function $f : \mathcal{X} \to \mathbb{R}$

$$\mathbb{P}_n(f) = \int f d\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n f(X_i).$$

- If $f \in L_1(P)$, so $P|f| = \int |f| dP < \infty$, then

$$\mathbb{P}_n(f) \rightarrow_{a.s.} P(f) = Ef(X) \tag{1}$$

  by the SLLN.

- Suppose that $\mathcal{F}$ is a collection of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$. If the convergence in (1) holds uniformly over $f \in \mathcal{F}$,

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n(f) - P(f)| \rightarrow_{a.s.} 0, \tag{2}$$

  then call $\mathcal{F}$ a Glivenko-Cantelli class for P.

- Bracketing numbers: for functions l, $u : \mathcal{X} \to \mathbb{R}$ with $l \le u$, the bracket $[l, u]$ is defined by

$$[l, u] \equiv \{f \in \mathcal{F} : \; l(x) \le f(x) \le u(x) \; \text{ for all } x \in \mathcal{X}\}. \tag{3}$$

$[l, u]$ is an $\epsilon-$bracket for $L_r(P)$ if $\|u - l\|_{L_r(P)} < \epsilon$.

- Bracketing numbers: for functions l, $u : \mathcal{X} \to \mathbb{R}$ with $l \leq u$, the bracket $[l, u]$ is defined by

$$[l, u] \equiv \{f \in \mathcal{F} : \; l(x) \leq f(x) \leq u(x) \;\text{ for all } x \in \mathcal{X}\}. \tag{4}$$

  $[l, u]$ is an $\epsilon-$bracket for $L_r(P)$ if $\|u - l\|_{L_r(P)} < \epsilon$.

- $N_{[\,]}(\epsilon, \mathcal{F}, L_r(P)) =$ minimal number of $\epsilon -$ brackets needed to cover $\mathcal{F}$

- **Covering numbers:**
  $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls of radius ϵ with respect to $\|\cdot\|$ needed to cover $\mathcal{F}$. If

$$B(f_j, \epsilon) = \{f \in \mathcal{F} : \ \|f - f_j\| < \epsilon\}$$

$$\begin{aligned} N(\epsilon, \mathcal{F}, \|\cdot\|) &= \min\{J : \ \mathcal{F} \subset \cup_{j=1}^{J} B(f_j, \epsilon) \\ &\qquad \text{for some } \ f_1, \ldots, f_J\} \end{aligned}$$

- **Covering numbers:**
  $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls of radius ϵ with respect to $\|\cdot\|$ needed to cover $\mathcal{F}$. If

$$B(f_j, \epsilon) = \{f \in \mathcal{F} : \ \|f - f_j\| < \epsilon\}$$

$$\begin{aligned} N(\epsilon, \mathcal{F}, \|\cdot\|) &= \min\{J : \ \mathcal{F} \subset \cup_{j=1}^{J} B(f_j, \epsilon) \\ &\qquad \text{for some} \ \ f_1, \ldots, f_J\} \end{aligned}$$

- **Envelope function F of a class $\mathcal{F}$:**

$$|f(x)| \leq F(x) \qquad \text{for all} \ \ x \in \mathcal{X}, \ \ \text{all} \ f \in \mathcal{F}.$$

- **Theorem:** (Bracketing Glivenko-Cantelli theorem) If $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$, (so that also $\mathcal{F}$ has envelope function F with $PF < \infty$), then $\mathcal{F}$ is $P-$Glivenko-Cantelli.

- **Theorem:** (Bracketing Glivenko-Cantelli theorem)
  If $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$, (so that also $\mathcal{F}$ has envelope function F with $PF < \infty$), then $\mathcal{F}$ is $P-$Glivenko-Cantelli.

- **Theorem:** (VC-Steele-Pollard-Giné-Zinn)
  If $\mathcal{F}$ has envelope function F with $PF < \infty$ and $\mathcal{F}_M \equiv \{f 1\{F \le M\} : \ f \in \mathcal{F}\}$ satisfies

$$n^{-1} E \log N(\epsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) \to 0,$$

  for all $\epsilon > 0$ and $M > 0$, then $\mathcal{F}$ is $P-$Glivenko-Cantelli (and conversely).

- Example: classical Glivenko-Cantelli theorem on $\mathbb{R}$

- Example: classical Glivenko-Cantelli theorem on $\mathbb{R}$
    - $\mathcal{X} = \mathbb{R}$, $\quad X \sim P$ on $(\mathbb{R}, \mathcal{B})$

- Example: classical Glivenko-Cantelli theorem on $\mathbb{R}$
  - $\mathcal{X} = \mathbb{R}, \qquad X \sim P$ on $(\mathbb{R}, \mathcal{B})$
  - $\mathcal{F} = \{x \mapsto 1_{(-\infty,t]}(x) : \ t \in \mathbb{R}\} \equiv \{f_t\}$

- Example: classical Glivenko-Cantelli theorem on $\mathbb{R}$
  - $\mathcal{X} = \mathbb{R}, \qquad X \sim P$ on $(\mathbb{R}, \mathcal{B})$
  - $\mathcal{F} = \{x \mapsto 1_{(-\infty,t]}(x) : \ t \in \mathbb{R}\} \equiv \{f_t\}$
  - $P(f_t) = P(X \leq t) \equiv F_X(t)$

- Example: classical Glivenko-Cantelli theorem on $\mathbb{R}$
  - $\mathcal{X} = \mathbb{R}$, $\quad X \sim P$ on $(\mathbb{R}, \mathcal{B})$
  - $\mathcal{F} = \{x \mapsto 1_{(-\infty,t]}(x) : \ t \in \mathbb{R}\} \equiv \{f_t\}$
  - $P(f_t) = P(X \le t) \equiv F_X(t)$
  - $\mathbb{P}(f_t) = \mathbb{P}_n(X \le t) \equiv \mathbb{F}_n(t)$

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F_X(t)| \to_{a.s.} 0.$$

- Example: classical Glivenko-Cantelli theorem on $\mathbb{R}$
    - $\mathcal{X} = \mathbb{R}, \qquad X \sim P$ on $(\mathbb{R}, \mathcal{B})$
    - $\mathcal{F} = \{x \mapsto 1_{(-\infty,t]}(x) : \ t \in \mathbb{R}\} \equiv \{f_t\}$
    - $P(f_t) = P(X \le t) \equiv F_X(t)$
    - $\mathbb{P}(f_t) = \mathbb{P}_n(X \le t) \equiv \mathbb{F}_n(t)$

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F_X(t)| \to_{a.s.} 0.$$

    - Here $F = 1$, and by VC - theory, for every $r \ge 1$, Q on $\mathbb{R}$

$$N(\epsilon, \mathcal{F}, L_r(Q)) \le K \left(\frac{M}{\epsilon}\right)^{r(V(\mathcal{F})-1)}$$

where $V(\mathcal{F}) = 2$

3. Consistency via Glivenko-Cantelli theorems

- **Inequality 1.** (van de Geer, 1993). Suppose that

$$\widehat{p}_n = \text{argmax}\{\mathbb{P}_n \log(p) : \ p \in \mathcal{P}\}.$$

Then with $h^2(p,q) = (1/2)\int[\sqrt{p} - \sqrt{q}]^2 d\mu = 1 - \int \sqrt{pq}\, d\mu$

$$\begin{aligned} h^2(\widehat{p}_n, p_0) &\le (\mathbb{P}_n - P_0)\left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) 1\{p_0 > 0\} \\ &\le \sup\{(\mathbb{P}_n - P_0)\left(\sqrt{\frac{p}{p_0}} - 1\right) 1\{p_0 > 0\} : \ p \in \mathcal{P}\} \\ &\to_{a.s.} 0 \end{aligned}$$

if $\mathcal{F} \equiv \left\{\left(\sqrt{\frac{p}{p_0}} - 1\right) 1\{p_0 > 0\} : \ p \in \mathcal{P}\right\}$ is P_0-Glivenko-Cantelli.

Proof of inequality 1.

- Since $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$,

$$
\begin{aligned}
0 &\le \frac{1}{2}\int_{[p_0>0]} \log\left(\frac{\widehat{p}_n}{p_0}\right) d\mathbb{P}_n \\
&\le \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d\mathbb{P}_n \quad \text{since} \quad \log(1+x) \le x \\
&= \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d(\mathbb{P}_n - P_0) \\
&\qquad + P_0\left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) 1\{p_0 > 0\} \\
&= \int_{[p_0>0]} \left(\sqrt{\widehat{p}_n/p_0} - 1\right) d(\mathbb{P}_n - P_0) - h^2(\widehat{p}_n, p_0)
\end{aligned}
$$

- **Inequality 2.** (Birgé and Massart, 1994).
  If $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$ over $\mathcal{P}$, then

$$
\begin{aligned}
h^2(\widehat{p}_n, p_0) &\le 12(\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1\{p_0 > 0\}\right) \\
&\le 12\sup\{(\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{p + p_0}{2p_0}\right)1\{p_0 > 0\}\right) \\
&\qquad : \ p \in \mathcal{P}\} \\
&\to_{a.s.} 0
\end{aligned}
$$

  if $\quad \mathcal{F} \equiv \left\{\left(\frac{1}{2}\log\left(\frac{p+p_0}{2p_0}\right)1\{p_0 > 0\}\right) : \ p \in \mathcal{P}\right\}$
  is P_0-Glivenko-Cantelli.

Proof of inequality 2.

- By concavity of $\log$,

$$\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right) 1\{p_0 > 0\} \geq \frac{1}{2}\log\left(\frac{\widehat{p}_n}{p_0}\right) 1\{p_0 > 0\}.$$

- Fact 1. $K(P, Q) \geq 2h^2(P, Q) \geq 0$.
- Fact 2. $h^2(P, Q) \leq 12h^2(P, (P + Q)/2)$.

Proof of inequality 2, cont'd

- Since $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$,

$$
\begin{aligned}
0 &\le \mathbb{P}_n \left(\frac{1}{4} \log \left(\frac{\widehat{p}_n}{p_0} \right) 1\{p_0 > 0\} \right) \\
&\le \mathbb{P}_n \left(\frac{1}{2} \log \left(\frac{\widehat{p}_n + p_0}{2p_0} \right) 1\{p_0 > 0\} \right) \\
&= (\mathbb{P}_n - P_0) \left(\frac{1}{2} \log \left(\frac{\widehat{p}_n + p_0}{2p_0} \right) 1\{p_0 > 0\} \right) \\
&\qquad + P_0 \left(\frac{1}{2} \log \left(\frac{\widehat{p}_n + p_0}{2p_0} \right) 1\{p_0 > 0\} \right) \\
&= (\mathbb{P}_n - P_0) \left(\frac{1}{2} \log \left(\frac{\widehat{p}_n + p_0}{2p_0} \right) 1\{p_0 > 0\} \right) \\
&\qquad - \frac{1}{2} K(P_0, (\widehat{P}_n + P_0)/2)
\end{aligned}
$$

Proof of inequality 2, cont'd

- $$\begin{aligned} &\leq (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1\{p_0 > 0\}\right) \\ &\quad - h^2(P_0, (\widehat{P}_n + P_0)/2) \\ &\leq (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1\{p_0 > 0\}\right) \\ &\quad - \frac{1}{12}h^2(P_0, (\widehat{P}_n + P_0)/2) \end{aligned}$$

  since

  $$h^2(P, Q) \leq 12h^2(P, (P + Q)/2).$$

- **Inequality 3.** (Pfanzagl,1988)
  If $\mathcal{P}$ is convex and $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$ over $\mathcal{P}$, then

$$
\begin{aligned}
h^2(\widehat{p}_n, p_0) &\leq (\mathbb{P}_n - P_0)\left(\frac{2\widehat{p}_n}{\widehat{p}_n + p_0}\right) \\
&\leq \sup\left\{(\mathbb{P}_n - P_0)\left(\frac{2p}{p + p_0}\right) : \ p \in \mathcal{P}\right\} \\
&\rightarrow_{a.s.} 0
\end{aligned}
$$

if $\quad \mathcal{F} \equiv \left\{\frac{2p}{p+p_0} : \ p \in \mathcal{P}\right\}$ is P_0-Glivenko-Cantelli.

4. Consistency: examples

- $\mathcal{P} = \{p_\theta(x) = \theta e^{-\theta x} 1\{x \geq 0\} : \ \theta > 0\}$ – Inequality 2

4. Consistency: examples

- $\mathcal{P} = \{p_\theta(x) = \theta e^{-\theta x} 1\{x \geq 0\} : \ \theta > 0\}$ – Inequality 2
- $\mathcal{P} = \{$all monotone decreasing densities on $\mathbb{R}^+\}$ – Inequality 3

4. Consistency: examples

- $\mathcal{P} = \{p_\theta(x) = \theta e^{-\theta x} 1\{x \geq 0\} : \ \theta > 0\}$ – Inequality 2
- $\mathcal{P} = \{$all monotone decreasing densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$all k-monotone densities on $\mathbb{R}^+\}$ – Inequality 3

4. Consistency: examples

- $\mathcal{P} = \{p_\theta(x) = \theta e^{-\theta x} 1\{x \geq 0\} : \ \theta > 0\}$ – Inequality 2
- $\mathcal{P} = \{$all monotone decreasing densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$all k-monotone densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$one-jump counting process with panel count observation scheme$\}$ (Schick and Yu (2000); van der Vaart and Wellner (2000))

4. Consistency: examples

- $\mathcal{P} = \{p_\theta(x) = \theta e^{-\theta x} 1\{x \ge 0\} : \ \theta > 0\}$ – Inequality 2
- $\mathcal{P} = \{$all monotone decreasing densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$all k-monotone densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$one-jump counting process with panel count observation scheme$\}$ (Schick and Yu (2000); van der Vaart and Wellner (2000))
- $\mathcal{P} = \{$competing risks model with current status censoring$\}$ (Maathuis)

4. Consistency: examples

- $\mathcal{P} = \{p_\theta(x) = \theta e^{-\theta x} 1\{x \ge 0\} : \ \theta > 0\}$ – Inequality 2
- $\mathcal{P} = \{$all monotone decreasing densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$all k-monotone densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$one-jump counting process with panel count observation scheme$\}$ (Schick and Yu (2000); van der Vaart and Wellner (2000))
- $\mathcal{P} = \{$competing risks model with current status censoring$\}$ (Maathuis)
- $\cdots$ (see Torgnon-Cortona-Delft notes) at

4. Consistency: examples

- $\mathcal{P} = \{p_\theta(x) = \theta e^{-\theta x} 1\{x \geq 0\} : \ \theta > 0\}$ – Inequality 2
- $\mathcal{P} = \{$all monotone decreasing densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$all k-monotone densities on $\mathbb{R}^+\}$ – Inequality 3
- $\mathcal{P} = \{$one-jump counting process with panel count observation scheme$\}$ (Schick and Yu (2000); van der Vaart and Wellner (2000))
- $\mathcal{P} = \{$competing risks model with current status censoring$\}$ (Maathuis)
- $\cdots$ (see Torgnon-Cortona-Delft notes) at http://www.stat.washington.edu/jaw/RESEARCH/

5. Introduction II, part 2: more empirical process theory

- If $f \in L_2(P)$, so $P|f|^2 = \int f^2 dP < \infty$, then

$$\sqrt{n}(\mathbb{P}_n(f) - P(f)) \rightarrow_d N(0, Var_P(f(X))) \tag{5}$$

  by the classical Central Limit Theorem.

- Suppose that $\mathcal{F}$ is a collection of real-valued functions $f : \mathcal{X} \rightarrow \mathbb{R}$. If the convergence in (5) holds uniformly over $f \in \mathcal{F}$,

$$\sqrt{n}(\mathbb{P}_n - P)(f) \Rightarrow \mathbb{G}_P(f) \qquad \text{in } \ \ell^\infty(\mathcal{F}) \tag{6}$$

  where $\mathbb{G}_P$ is a $P-$Brownian bridge process then call $\mathcal{F}$ a Donsker class for P.

Two Donsker theorems:

- **Theorem:** (Ossiander, 1987) If

$$\int_0^1 \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty$$

(so that $\mathcal{F}$ has an envelope F with $PF^2 < \infty$) then $\mathcal{F}$ is $P-$Donsker.

Two Donsker theorems:

- **Theorem:** (Ossiander, 1987) If

$$\int_0^1 \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty$$

(so that $\mathcal{F}$ has an envelope F with $PF^2 < \infty$) then $\mathcal{F}$ is $P-$Donsker.

- **Theorem:** (Pollard,1982; Koltchinskii, 1981)
If $\mathcal{F}$ has envelope function F with $PF^2 < \infty$ and

$$\int_0^1 \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon < \infty$$

then $\mathcal{F}$ is $P-$Donsker.

Two Donsker theorems:

- **Theorem:** (Ossiander, 1987) If

$$\int_0^1 \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty$$

  (so that $\mathcal{F}$ has an envelope F with $PF^2 < \infty$) then $\mathcal{F}$ is $P-$Donsker.

- **Theorem:** (Pollard,1982; Koltchinskii, 1981)
  If $\mathcal{F}$ has envelope function F with $PF^2 < \infty$ and

$$\int_0^1 \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon < \infty$$

  then $\mathcal{F}$ is $P-$Donsker.

- $\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P)) \le K\epsilon^{-r}$ with $r < 2$ suffices.
  $\sup_Q \log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \le K\epsilon^{-r}$ with $r < 2$ suffices.

6. Rates of convergence via empirical process theory

- Suppose Θ is a metric space with a metric d.
- Consider estimation of $\theta \in \Theta$ by maximizing

$$\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta(X), \qquad \theta \in \Theta$$

  for some collection of real-valued functions $m_\theta(X)$ from $\Theta \times \mathcal{X}$ to $\mathbb{R}$.

- Possibilities for m_θ:
  - $\diamond$ $m_\theta(X) = \log p_\theta(X)$
  - $\diamond$ $m_\theta(X) = \log q_\theta(X)$ for $q_\theta \in \mathcal{P}_0 \subset \mathcal{P}$, or
  - $\diamond$ $m_\theta(X) = -\int p_\theta^2 d\mu + 2p_\theta(X)$.
- Population version of criterion function:

$$\mathbb{M}(\theta) = P_0 m_\theta(X), \qquad \theta \in \Theta.$$

Rates via empirical process theory, contd.

- Now assume that θ_0 is a point maximizing $\mathbb{M}(\theta)$.
- When $\mathbb{M}$ is sufficiently smooth, the first derivative of $\mathbb{M}$ vanishes at θ_0 and the second derivative is typically negative definite. Hence it is very natural to assume that

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0) \tag{7}$$

for θ in a neighborhood of θ_0.

Rates via empirical process theory, contd.

Basic rate theorem: Suppose that:

- ⋄ (7) holds for θ in a neighborhood of θ_0;
- ⋄ $\mathbb{M}_n - \mathbb{M}$ satisfies

$$E^* \sup_{d(\theta,\theta_0)<\delta} |(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}},$$

  where ϕ_n are functions satisfying $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ (not depending on n).
- ⋄ $\hat{\theta}_n$ maximizes $\mathbb{M}_n(\theta)$
- ⋄ $\hat{\theta}_n \rightarrow_{p^*} \theta_0$

Rates via empirical process theory, contd.

◇ Then

$$r_n d(\widehat{\theta}_n, \theta_0) = O_p^*(1)$$

for r_n satisfying

$$r_n{}^2 \phi_n \left(\frac{1}{r_n} \right) \leq \sqrt{n} \qquad \text{for every } \; n\,.$$

Rates via empirical process theory, contd.

- If $\phi_n(\delta) = \delta^\beta$, then $r_n = n^{\frac{1}{2(2-\beta)}} \equiv n^s$.

β	s	name / situation
1	1/2	classical smoothness
1/2	1/3	bounded monotone on $\mathbb{R}$
3/4	2/5	convex on $\mathbb{R}$
3/4	2/5	bounded second derivative on $[0,1]$
$1 - d/4$	$2/(d+4)$	convex in $\mathbb{R}^d$

Rates via empirical process theory, contd.

- If $\phi_n(\delta) = \delta^\beta$, then $r_n = n^{\frac{1}{2(2-\beta)}} \equiv n^s$.

β	s	name / situation
1	1/2	classical smoothness
1/2	1/3	bounded monotone on $\mathbb{R}$
3/4	2/5	convex on $\mathbb{R}$
3/4	2/5	bounded second derivative on $[0,1]$
$1-d/4$	$2/(d+4)$	convex in $\mathbb{R}^d$

- How do we get $\phi_n(\delta)$? Empirical process theory ... !

Rates via empirical process theory, contd.

- When $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$ and $\mathbb{M}(\theta) = P_0 m_\theta$, then with $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$

$$\sqrt{n} \sup_{d(\theta,\theta_0)<\delta} |(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)| = \|\mathbb{G}_n\|_{\mathcal{M}_\delta(\theta_0)}$$

where

$$\mathcal{M}_\delta(\theta_0) = \{m_\theta - m_{\theta_0} : \; d(\theta, \theta_0) < \delta\}.$$

- Then the key oscillation condition of the theorem becomes:

$$E^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta(\theta_0)} \lesssim \phi_n(\delta)$$

Rates via empirical process theory, contd.

- Uniform entropy bounds and bracketing bounds yield

$$E\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J(1, \mathcal{M}_\delta)(PM_\delta^2)^{1/2}\,,$$
$$E\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim J_{[\,]}(1, \mathcal{M}_\delta, L_2(P))(PM_\delta^2)^{1/2}\,,$$

where M_δ is an envelope function for the class $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) \le \delta\}$,

$$J_{[\,]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[\,]}(\epsilon\|F\|, \mathcal{F}, \|\cdot\|)}\, d\epsilon$$
$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\epsilon\,.$$

Rates via empirical process theory, contd.

- Let $m_p = \log\left(\frac{p+p_0}{2p_0}\right)$, $\mathbb{M}_n(p) = \mathbb{P}_n m_p$.
- **Fact:** The MLE $\widehat{p}_n$ satisfies $\mathbb{M}_n(\widehat{p}_n) \geq \mathbb{M}_n(p_0)$
- **Theorem.** (Birgé and Massart). Suppose $p_0 \in \mathcal{P}$. Then

$$\mathbb{M}(p) - \mathbb{M}(p_0) = P_0(m_p - m_{p_0}) \lesssim -h^2(p, p_0)\,.$$

Furthermore, with $\mathcal{M}_\delta = \{m_p - m_{p_0} :\ h(p, p_0) \leq \delta\}$,

$$E^*_{P_0}\|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \tilde{J}_{[\,]}(\delta, \mathcal{P}, h)\left(1 + \frac{\tilde{J}_{[\,]}(\delta, \mathcal{P}, h)}{\delta^2\sqrt{n}}\right) \equiv \phi_n(\delta)$$

where

$$\tilde{J}_{[\,]}(\delta, \mathcal{P}, h) \equiv \int_{c\delta^2}^{\delta} \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{P}, h)}\, d\epsilon.$$

Rates via empirical process theory, contd.

- Thus the rate of convergence of the MLE $r_n = r_n^{mle}$ is determined by the solution of

$$\sqrt{n} r_n^{-2} = \int_{c r_n^{-2}}^{r_n^{-1}} \sqrt{\log N_{[\,]}(\epsilon, \mathcal{P}, h)}\, d\epsilon.$$

- If

$$\log N_{[\,]}(\epsilon, \mathcal{P}, h) \asymp \frac{K}{\epsilon^{1/\gamma}} \tag{8}$$

then r_n is given by

$$r_n = \begin{cases} n^{\gamma/(2\gamma+1)} & \text{if } \gamma > 1/2 \ \ \text{(upper limit dominant)} \\ n^{\gamma/2} & \text{if } \gamma < 1/2 \ \ \text{(lower limit dominant)}\,. \end{cases}$$

Rates via empirical process theory, contd.

- Le Cam (1973); Birgé (1983):
  optimal rate of convergence $r_n = r_n^{opt}$ determined by

$$nr_n^{-2} = \log N_{[\,]}(1/r_n, \mathcal{P}, h) \tag{9}$$

- If

$$\log N_{[\,]}(\epsilon, \mathcal{P}) \asymp \frac{K}{\epsilon^{1/\gamma}} \tag{10}$$

  (9) leads to the optimal rate of convergence

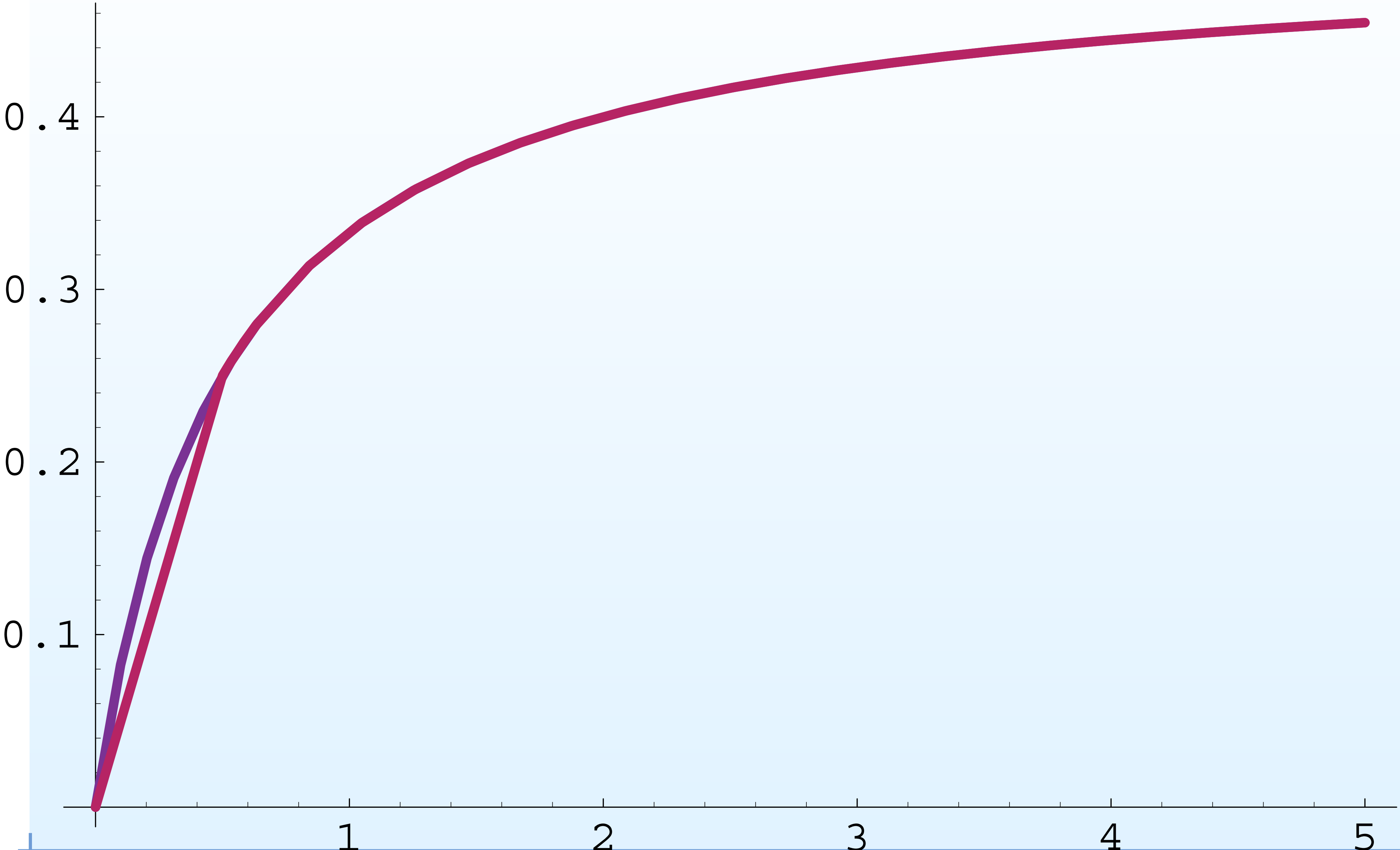
$$r_n^{opt} = n^{\gamma/(2\gamma+1)}\,.$$

- Conclusion: the MLE is (possibly) rate sub-optimal if $\gamma \le 1/2$.

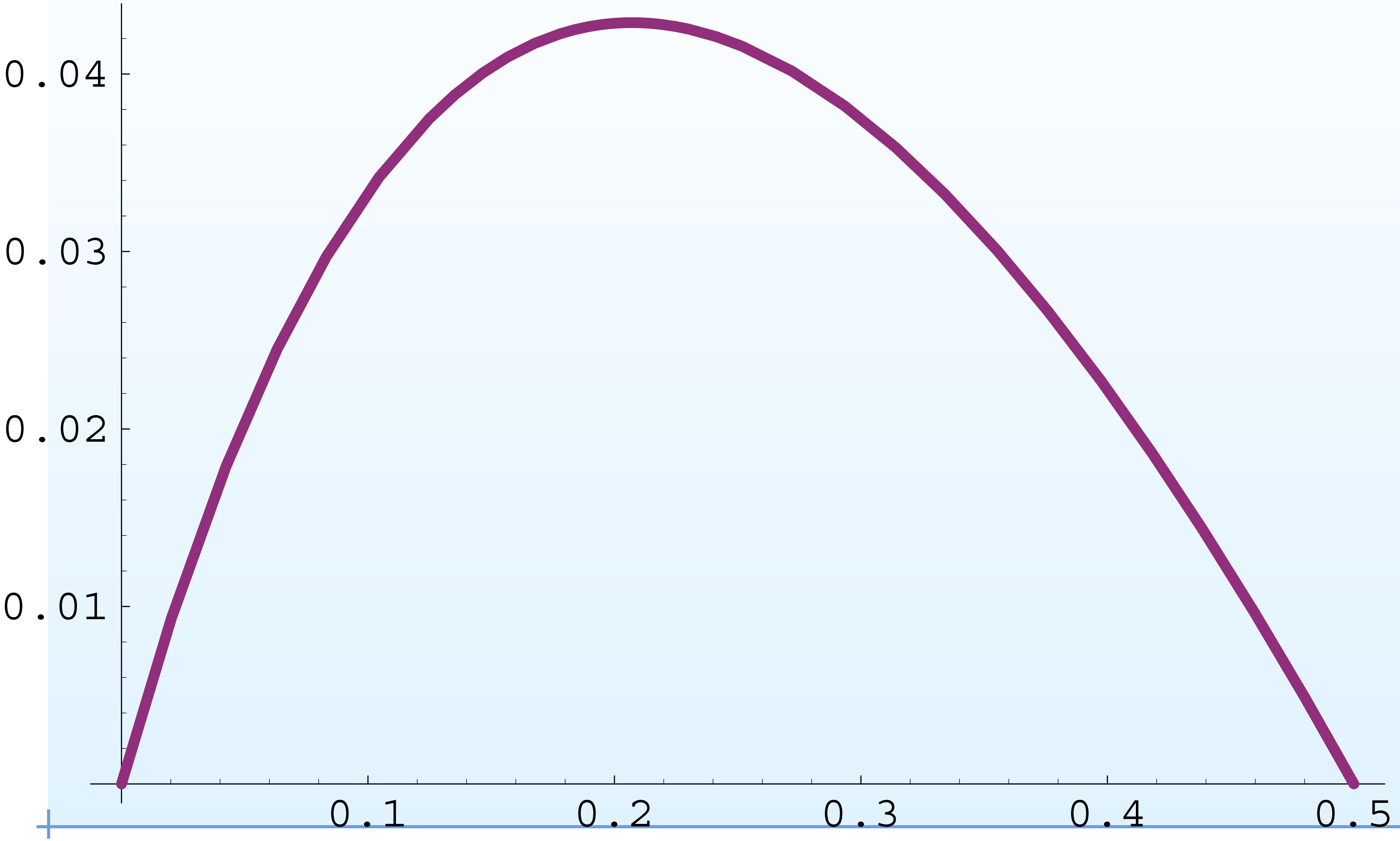
Rates via empirical process theory, contd.

- Typically
$$\frac{1}{\gamma} = \frac{d}{\alpha}$$
where d is the dimension of the underlying sample space and α is a measure of the “smoothness” (or number of derivatives) of the functions in $\mathcal{P}$.
- Hence
$$\alpha \leq \frac{d}{2}$$
leads to $\gamma \leq 1/2$.
- But there are many examples/problem with $\gamma > 1/2$!

Optimal rate and MLE global rate as a function of γ



Difference of rates $\gamma/(2\gamma+1) - \gamma/2$



7. Rates for MLEs: examples

- (Birgé-Massart, 1993): $\alpha-$Hölderian densities on $[0,1]$ with $\alpha < 1/2$. $r_n^{mle} = n^{\alpha/2}$, $r_n^{opt} = n^{\alpha/(2\alpha+1)}$.
- (Birgé, 1987, 1989): monotone density on $\mathbb{R}^+$. $1/\gamma = 1/1 = 1$. $r_n^{mle} = n^{1/3} = r_n^{opt}$.
- (Biau-Devroye, 2003): monotone decreasing densities in $\mathbb{R}^{+d}$. $1/\gamma = d/\alpha = d$. $r_n^{opt} = n^{1/(2+d)}$, $r_n^{mle} = n^{1/(2d)}$? (Entropies still unknown; rate of convergence of MLE unknown).
- (van de Geer, 1996, 2000): Interval censoring in $\mathbb{R}$. $1/\gamma = 1/1 = 1$ $r_n^{mle} = n^{1/3}$ (up to $\log$ terms); $r_n^{opt} =?$
- (Maathuis, 2004): competing risks with current status data. $1/\gamma = 1/1$, $r_n^{mle} = n^{1/3} = r_n^{opt}$.

7. Rates for MLEs: examples, cont'd.

- (Ghosal and van der Vaart, 2001): normal location mixtures on $\mathbb{R}$. $\log N_{[\,]}(\epsilon, \mathcal{P}, h) \le (\log(1/\epsilon))^{2r+1}$.
  $r_n^{mle} = r_n^{npbayes} = n^{1/2}/(\log n)^{1/2+r\vee 1/2}$.
- $k-$monotone densities on $\mathbb{R}^+$: $r_n^{opt} = n^{k/(2k+1)}$?
  $r_n^{mle} = n^{k/(2k+1)}$?

8. Problems and challenges

- Characterization of consistency of MLE's (dominated case)?
- Characterization of rate of convergence of MLE's?
- Is the MLE always rate-supoptimal when $\gamma \le 1/2$?
- Exact bounds for $N_{[\,]}(\epsilon, \mathcal{P}_{monotone,d}, h)$?
  Exact rates for MLE over $\mathcal{P}_{monotone,d}$?
- More entropy results for $N_{[\,]}(\epsilon, \mathcal{P}, h)$?
- Beyond consistency and global rates:
  - Tools for local rates? (No unifying method yet!)
  - Algorithms for computation?
    (Iterative Convex Minorant; Support reduction; ... ?)
  - Non-dominated case: characterization of consistency?
  - Methods for global rates when model assumptions fail?
    (Kleijn and van der Vaart (2005) treat nonparametric Bayes estimators)

Selected References

- Biau, G. and Devroye, L. (2003). On the risk of estimates for block decreasing densities. *J. Mult. Anal.* **86**, 143 - 165.
- Bickel, P. J. and Ritov, Y. (2003). Nonparametric estimators which can be “plugged-in”. *Ann. Statist.* **31**, 1033 - 1053.
- Birgé, L. (1987). On the risk of histograms for estimating decreasing densities. *Ann. Statist.* **15**, 1013 - 1022.

- Birgé, L. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17**, 1532-1549.
- Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields* **97**, 113 - 150.
- Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233 - 1263.

- Le Cam, L. (1973). Convergence rates under dimensionality restrictions. *Ann. Statist.* **1**, 38 - 53.
- Le Cam, L. (1990). Maximum likelihood: an introduction. *Internat. Statist. Rev.* **58**, 153 - 171.
- van de Geer, S. A. (1993) Helliinger consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14 - 44.
- van de Geer, S. A. (1996) Rates of convergence for the maximum likelihood estimator in mixture models. *J. Nonparametric Statist.* **6**, 293 - 310.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.
- Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios convergence rates for sieve MLEs. *Ann. Statist.* **23**, 339 - 362.