Maximum likelihood:

counterexamples, examples, and open problems

Jon A. Wellner

University of Washington

- Talk at University of Idaho ,
  Department of Mathematics, September 15, 2005
- *Email: jaw@stat.washington.edu*
  http: //www.stat.washington.edu/jaw/jaw.research.html

Outline

- Introduction: maximum likelihood estimation

Outline

- Introduction: maximum likelihood estimation
- Counterexamples

Outline

- Introduction: maximum likelihood estimation
- Counterexamples
- Beyond consistency: rates and distributions

Outline

- Introduction: maximum likelihood estimation
- Counterexamples
- Beyond consistency: rates and distributions
- Positive examples

Outline

- Introduction: maximum likelihood estimation
- Counterexamples
- Beyond consistency: rates and distributions
- Positive examples
- Problems and challenges

1. Introduction: maximum likelihood estimation

- Setting 1: dominated families

1. Introduction: maximum likelihood estimation

- Setting 1: dominated families
- Suppose that $X_1, \ldots, X_n$ 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 $\Theta \subset \mathbb{R}^d$.

1. Introduction: maximum likelihood estimation

- Setting 1: dominated families
- Suppose that $X_1, \ldots, X_n$ 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 $\Theta \subset \mathbb{R}^d$.
- The likelihood is

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

1. Introduction: maximum likelihood estimation

- Setting 1: dominated families
- Suppose that $X_1, \ldots, X_n$ 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 $\Theta \subset \mathbb{R}^d$.
- 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

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

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

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

- **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

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

- **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

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

- **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

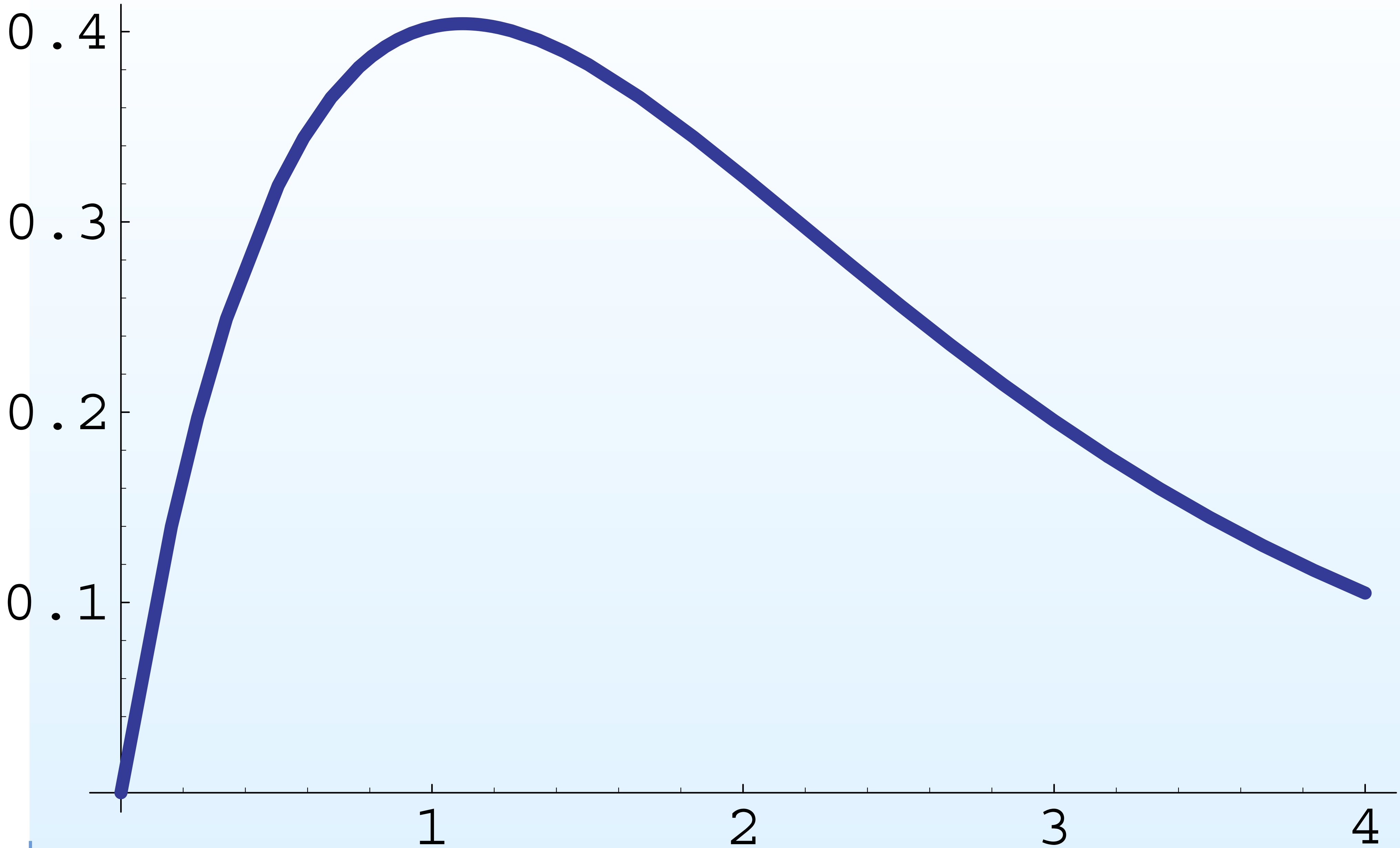
$$L_n(\theta) = \theta^n \exp(-\theta \sum_1^n X_i),$$

- so the log-likelihood is

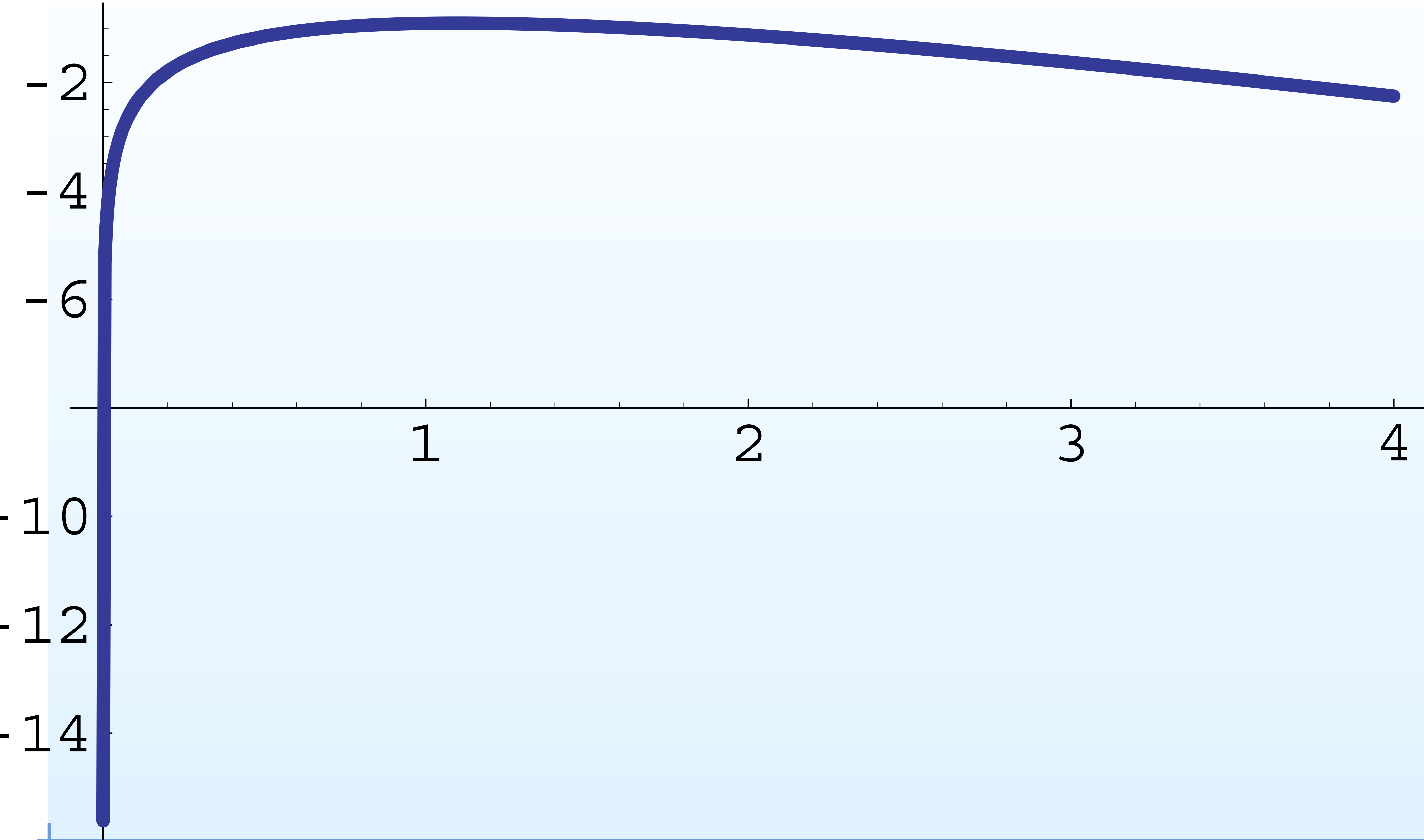
$$\log L_n(\theta) = n \log(\theta) - \theta \sum_1^n X_i$$

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

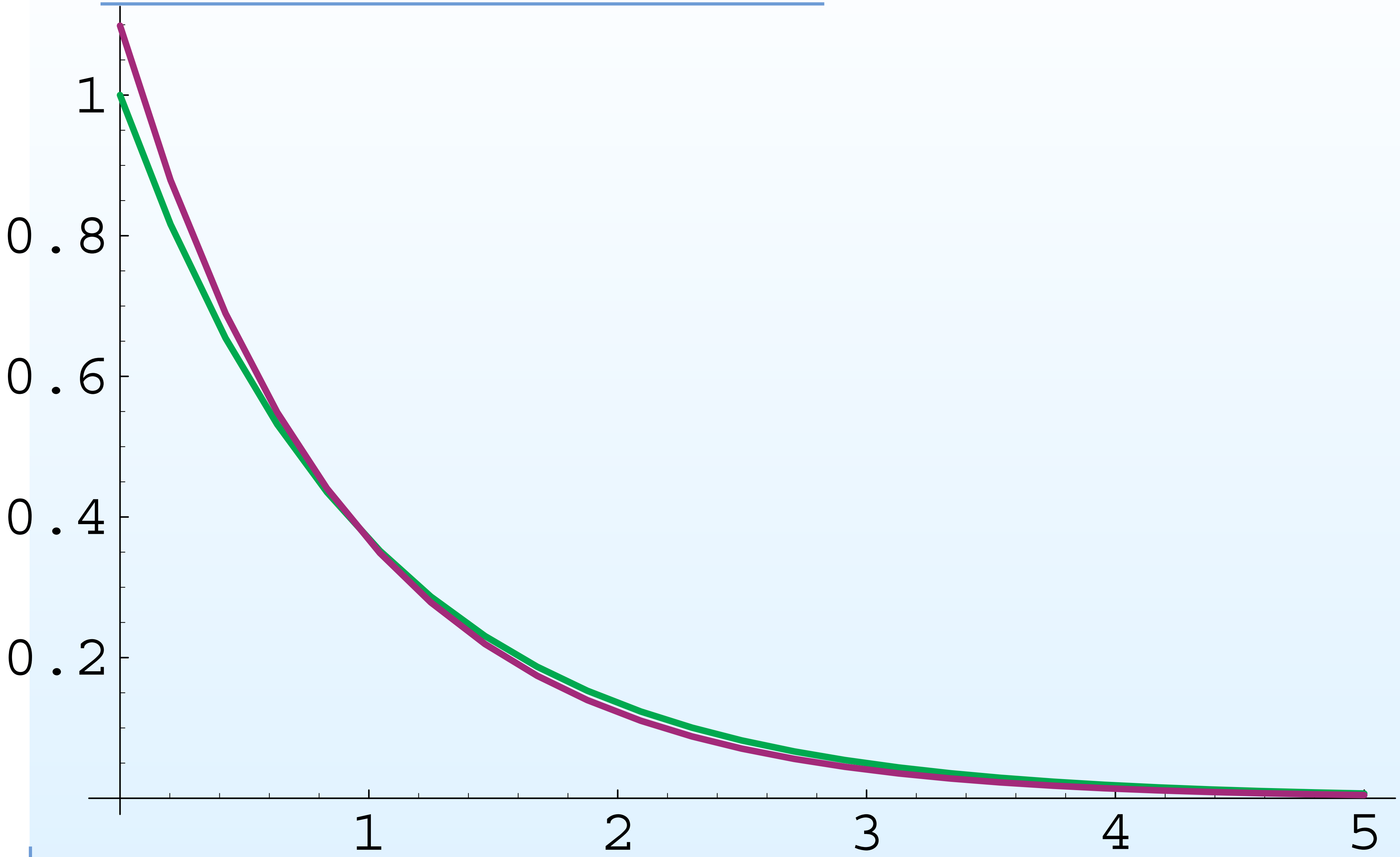
1/n power of likelihood, $n = 50$



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 $(0, \infty)$. $X_1, \dots, X_n$ 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 $L_n(p) = \prod_{i=1}^n p(X_i)$;

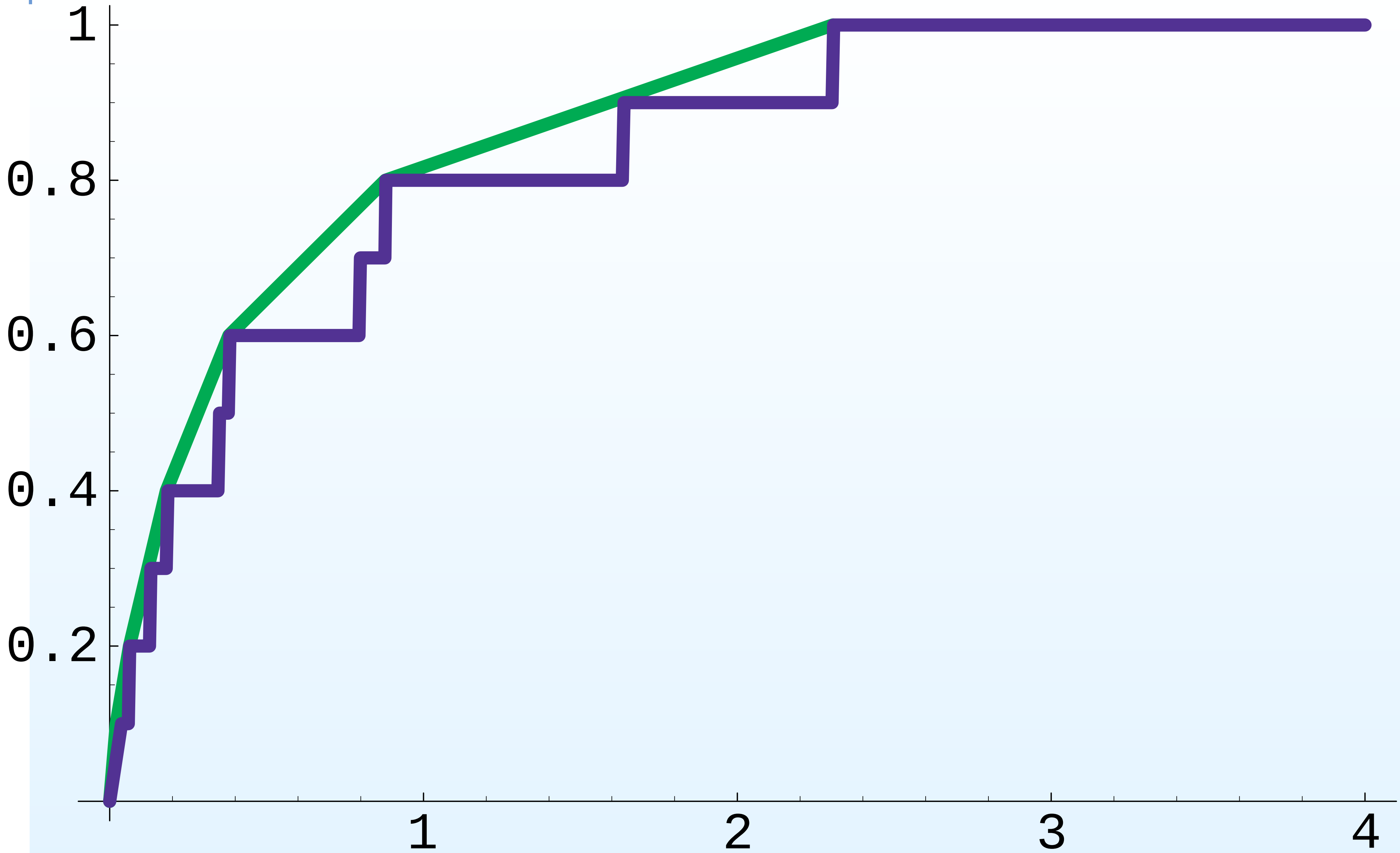
- **Example 2.** Monotone decreasing densities on $(0, \infty)$. $X_1, \ldots, X_n$ are i.i.d. $p_0 \in \mathcal{P}$ where

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

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

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

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


1
0.8
0.6
0.4
0.2
1
2
3
4

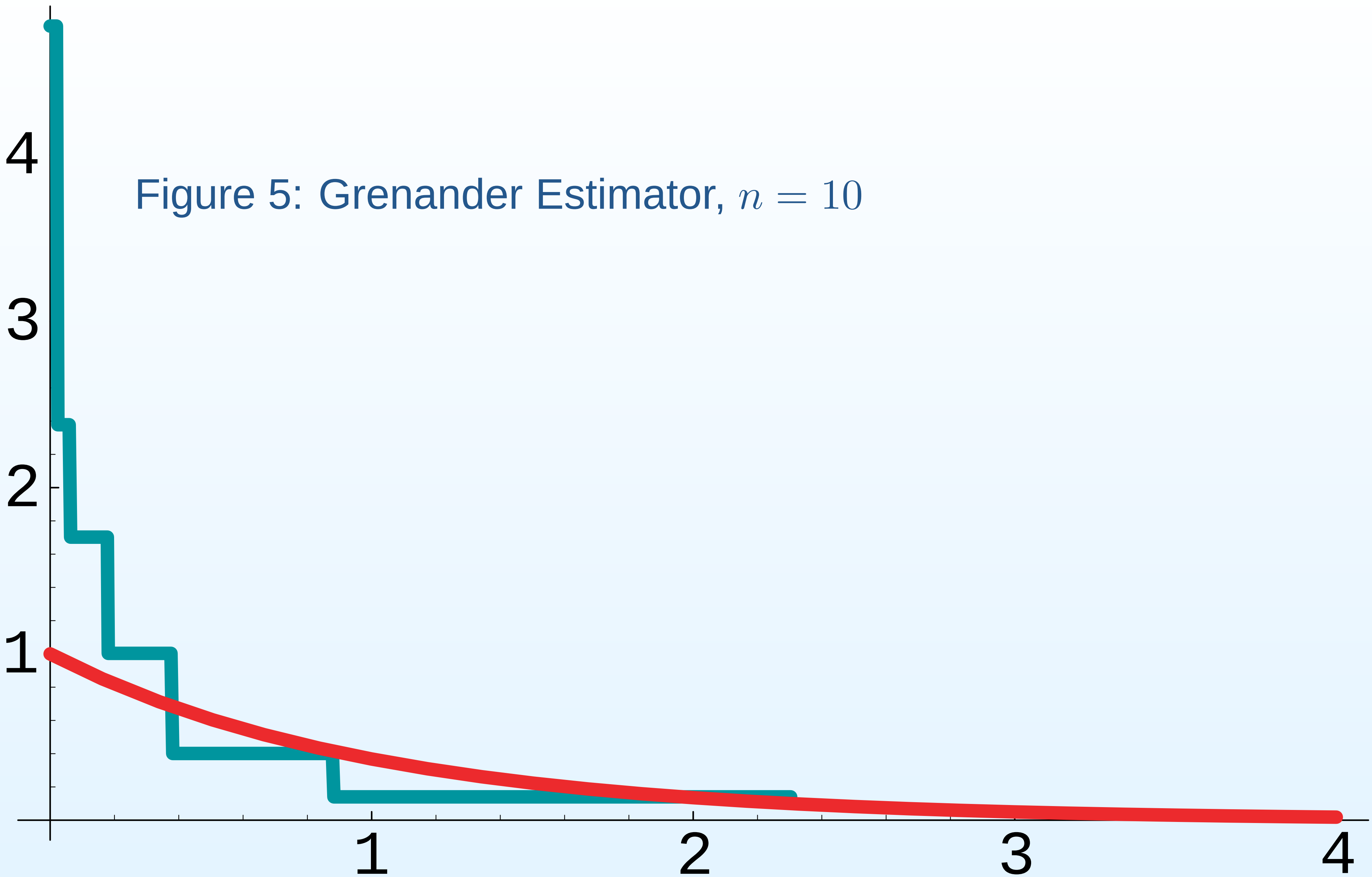
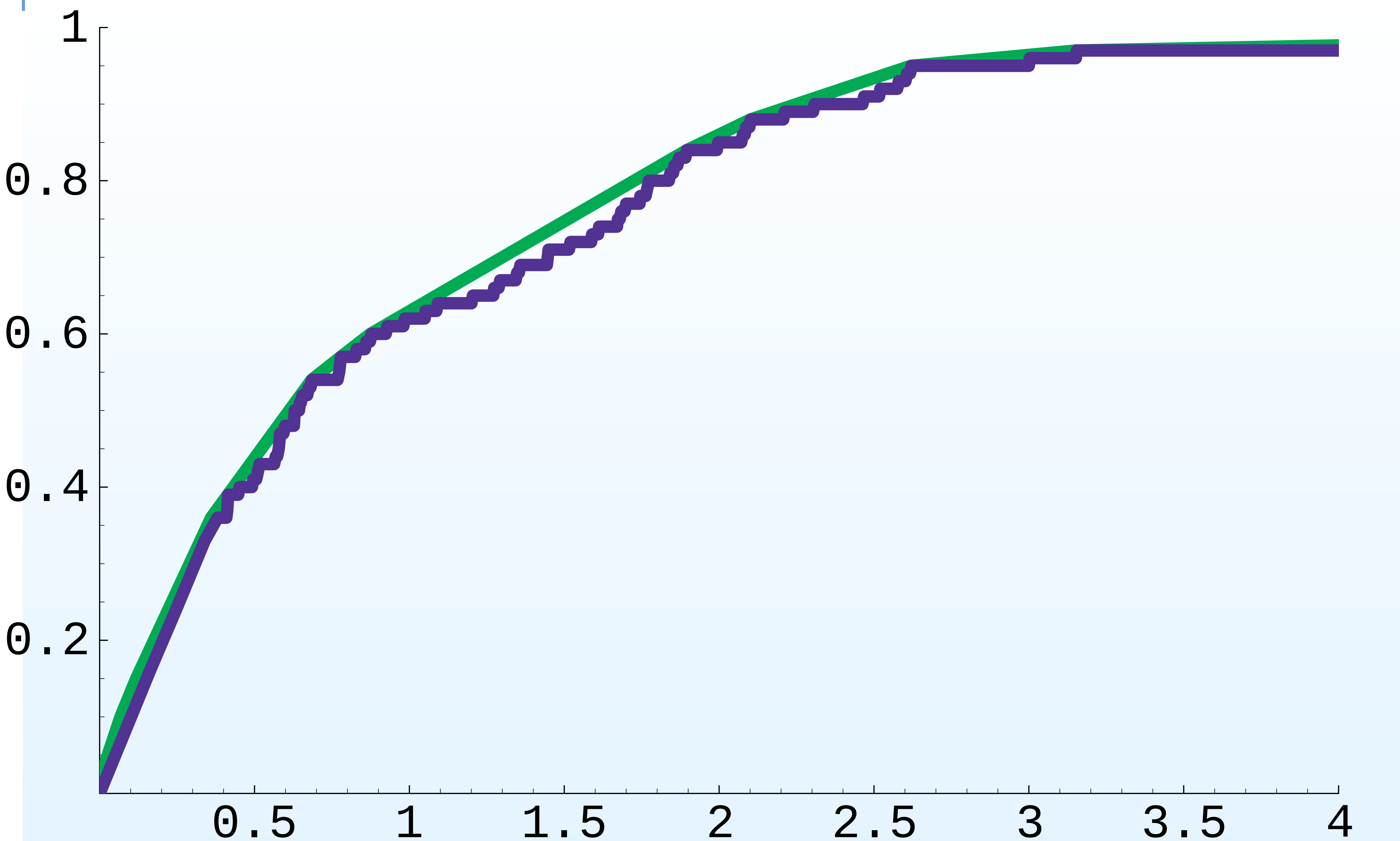


Figure 5: Grenander Estimator, $n = 10$


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

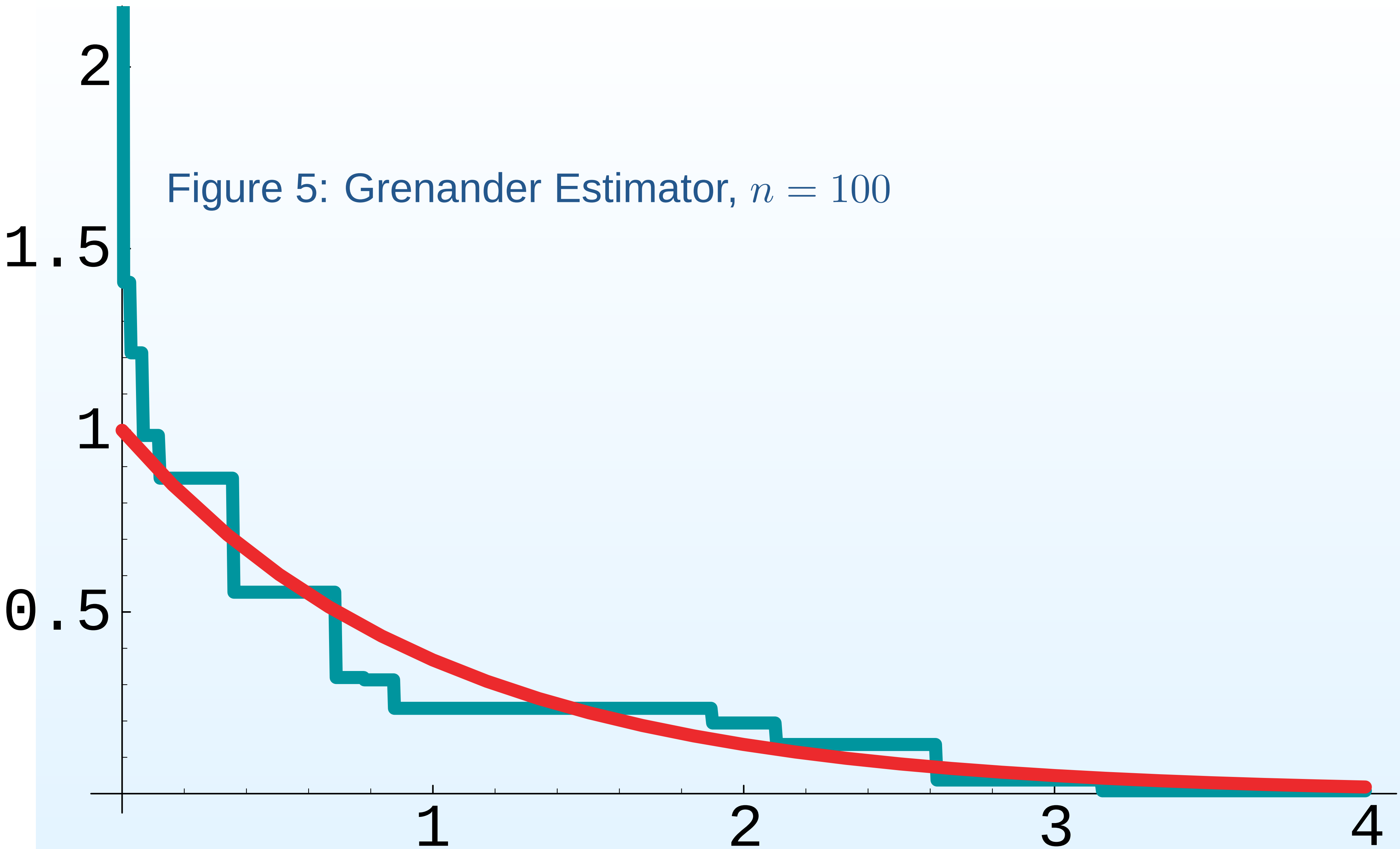


Figure 5: Grenander Estimator, $n = 100$

- **Setting 2: non-dominated families**

- **Setting 2: non-dominated families**
- Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_0 \in \mathcal{P}$ where $\mathcal{P}$ is some collection of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$.

- **Setting 2: non-dominated families**
- Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_0 \in \mathcal{P}$ where $\mathcal{P}$ is some collection of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$.
- If $P\{x\}$ denotes the measure under P of the one-point set $\{x\}$, the likelihood of $X_1, \ldots, X_n$ is defined to be

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

- **Setting 2: non-dominated families**
- Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_0 \in \mathcal{P}$ where $\mathcal{P}$ is some collection of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$.
- If $P\{x\}$ denotes the measure under P of the one-point set $\{x\}$, the likelihood of $X_1, \ldots, X_n$ is defined to be

$$L_n(P) = \prod_{i=1}^{n} P\{X_i\} .$$

- Then a Maximum Likelihood Estimator (or MLE) of P_0 can be defined as a measure $\hat{P}_n \in \mathcal{P}$ that maximizes $L_n(P)$; thus

$$L_n(\hat{P}) = \sup_{P \in \mathcal{P}} L_n(P)$$

- **Example 3.** (Empirical measure)

- **Example 3.** (Empirical measure)
- If $\mathcal{P}$ =all probability measures on $(\mathcal{X}, \mathcal{A})$, then

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

where $\delta_x(A) = 1_A(x)$.

- **Example 3.** (Empirical measure)
- If $\mathcal{P}$ =all probability measures on $(\mathcal{X}, \mathcal{A})$, then

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

where $\delta_x(A) = 1_A(x)$.

- Thus

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

Consistency of the MLE:

- Wald (1949)

Consistency of the MLE:

- Wald (1949)
- Kiefer and Wolfowitz (1956)

Consistency of the MLE:

- Wald (1949)
- Kiefer and Wolfowitz (1956)
- Huber (1967)

Consistency of the MLE:

- Wald (1949)
- Kiefer and Wolfowitz (1956)
- Huber (1967)
- Perlman (1972)

Consistency of the MLE:

- Wald (1949)
- Kiefer and Wolfowitz (1956)
- Huber (1967)
- Perlman (1972)
- Wang (1985)

Consistency of the MLE:

- Wald (1949)
- Kiefer and Wolfowitz (1956)
- Huber (1967)
- Perlman (1972)
- Wang (1985)
- van de Geer (1993)

Counterexamples:

- Neyman and Scott (1948)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)
- LeCam (1975), (1990)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)
- LeCam (1975), (1990)
- Barlow, Bartholomew, Bremner, and Brunk (4B's) (1972)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)
- LeCam (1975), (1990)
- Barlow, Bartholomew, Bremner, and Brunk (4B's) (1972)
- Boyles, Marshall, and Proschan (1985)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)
- LeCam (1975), (1990)
- Barlow, Bartholomew, Bremner, and Brunk (4B's) (1972)
- Boyles, Marshall, and Proschan (1985)
- bivariate right censoring
  Tsai, van der Laan, Pruitt

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)
- LeCam (1975), (1990)
- Barlow, Bartholomew, Bremner, and Brunk (4B's) (1972)
- Boyles, Marshall, and Proschan (1985)
- bivariate right censoring
  Tsai, van der Laan, Pruitt
- left truncation and interval censoring
  Chappell and Pan (1999)

Counterexamples:

- Neyman and Scott (1948)
- Bahadur (1958)
- Ferguson (1982)
- LeCam (1975), (1990)
- Barlow, Bartholomew, Bremner, and Brunk (4B's) (1972)
- Boyles, Marshall, and Proschan (1985)
- bivariate right censoring
  Tsai, van der Laan, Pruitt
- left truncation and interval censoring
  Chappell and Pan (1999)
- Maathuis and Wellner (2005)

2. Counterexamples: MLE's are not always consistent

- **Counterexample 1.** (Ferguson, 1982).
  Suppose that $X_1, \ldots, X_n$ are i.i.d. with density f_{θ_0} where

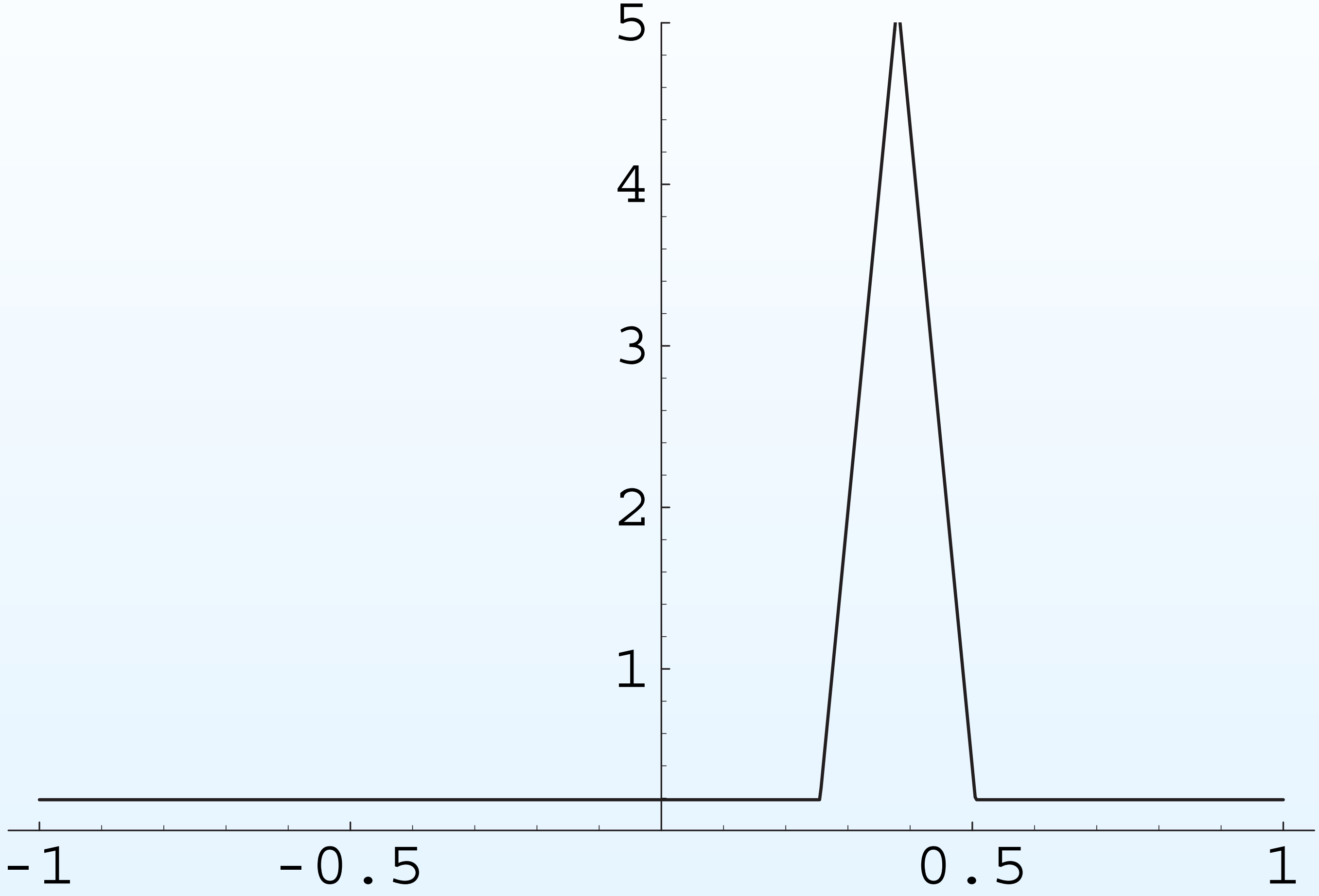
$$f_\theta(x) = (1-\theta)\frac{1}{\delta(\theta)} f_0\left(\frac{x-\theta}{\delta(\theta)}\right) + \theta f_1(x)$$

for $\theta \in [0,1]$ where

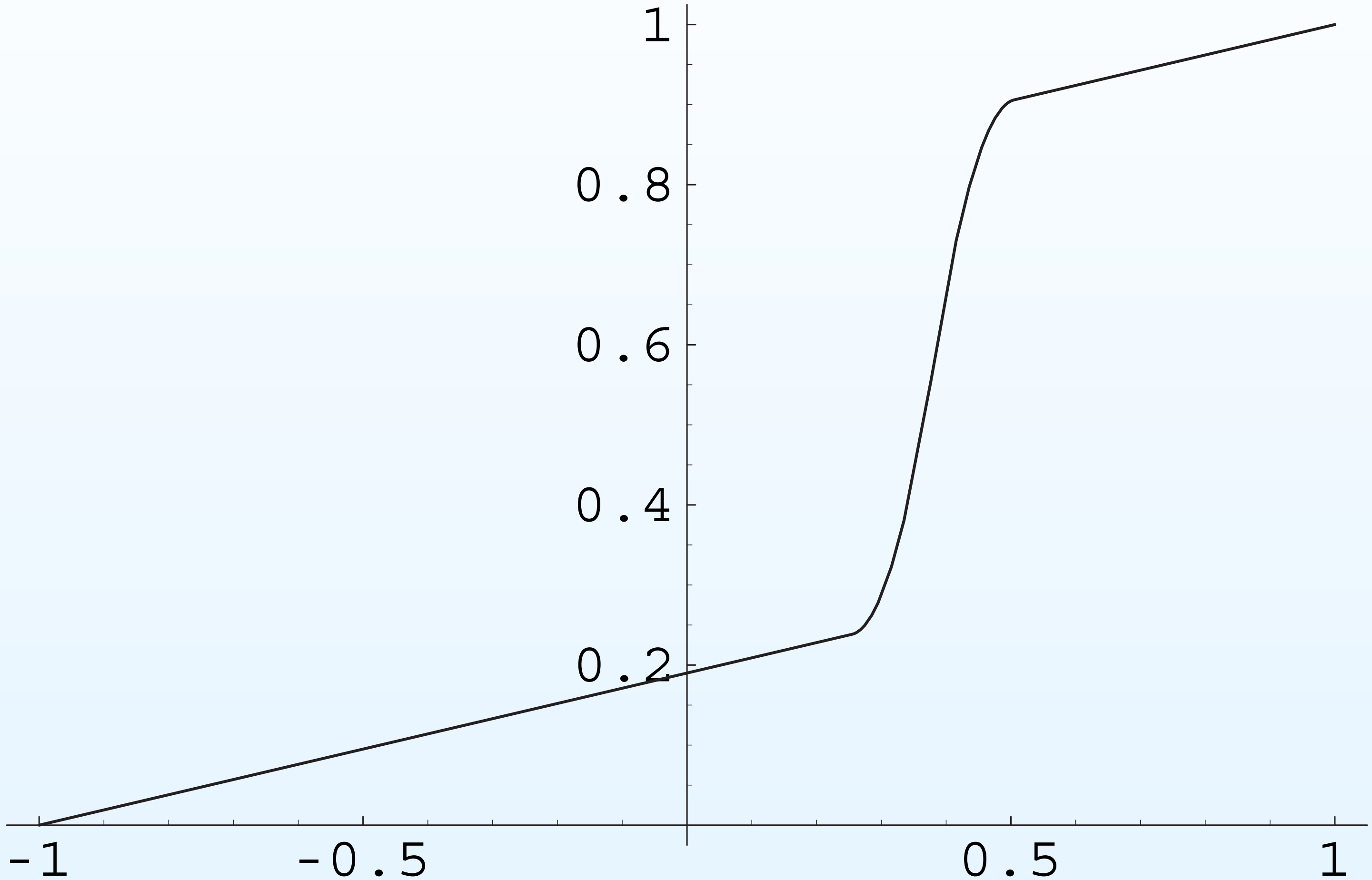
$$f_1(x) = \frac{1}{2} 1_{[-1,1]}(x) \qquad \text{Uniform}[-1,1],$$
$$f_0(x) = (1-|x|) 1_{[-1,1]}(x) \qquad \text{Triangular}[-1,1]$$

and $\delta(\theta)$ satisfies:
  - $\delta(0) = 1$
  - $0 < \delta(\theta) \le 1 - \theta$
  - $\delta(\theta) \to 0$ as $\theta \to 1$.

Density $f_\theta(x)$ for $c = 2$, $\theta = .38$



$F_\theta(x)$ for $c = 2$, $\theta = .38$



- Ferguson (1982) shows that $\hat{\theta}_n \rightarrow_{a.s.} 1$ no matter what θ_0 is true if $\delta(\theta) \rightarrow 0$ “fast enough”.

- Ferguson (1982) shows that $\hat{\theta}_n \to_{a.s.} 1$ no matter what θ_0 is true if $\delta(\theta) \to 0$ "fast enough".
- In fact, the assertion is true if

$$\delta(\theta) = (1-\theta)\exp(-(1-\theta)^{-c} + 1)$$

with $c > 2$. (Ferguson shows that $c = 4$ works.)

- Ferguson (1982) shows that $\hat{\theta}_n \to_{a.s.} 1$ no matter what θ_0 is true if $\delta(\theta) \to 0$ “fast enough”.
- In fact, the assertion is true if

$$\delta(\theta) = (1-\theta)\exp(-(1-\theta)^{-c}+1)$$

with $c > 2$. (Ferguson shows that $c = 4$ works.)

- If $c = 2$, Ferguson’s argument shows that

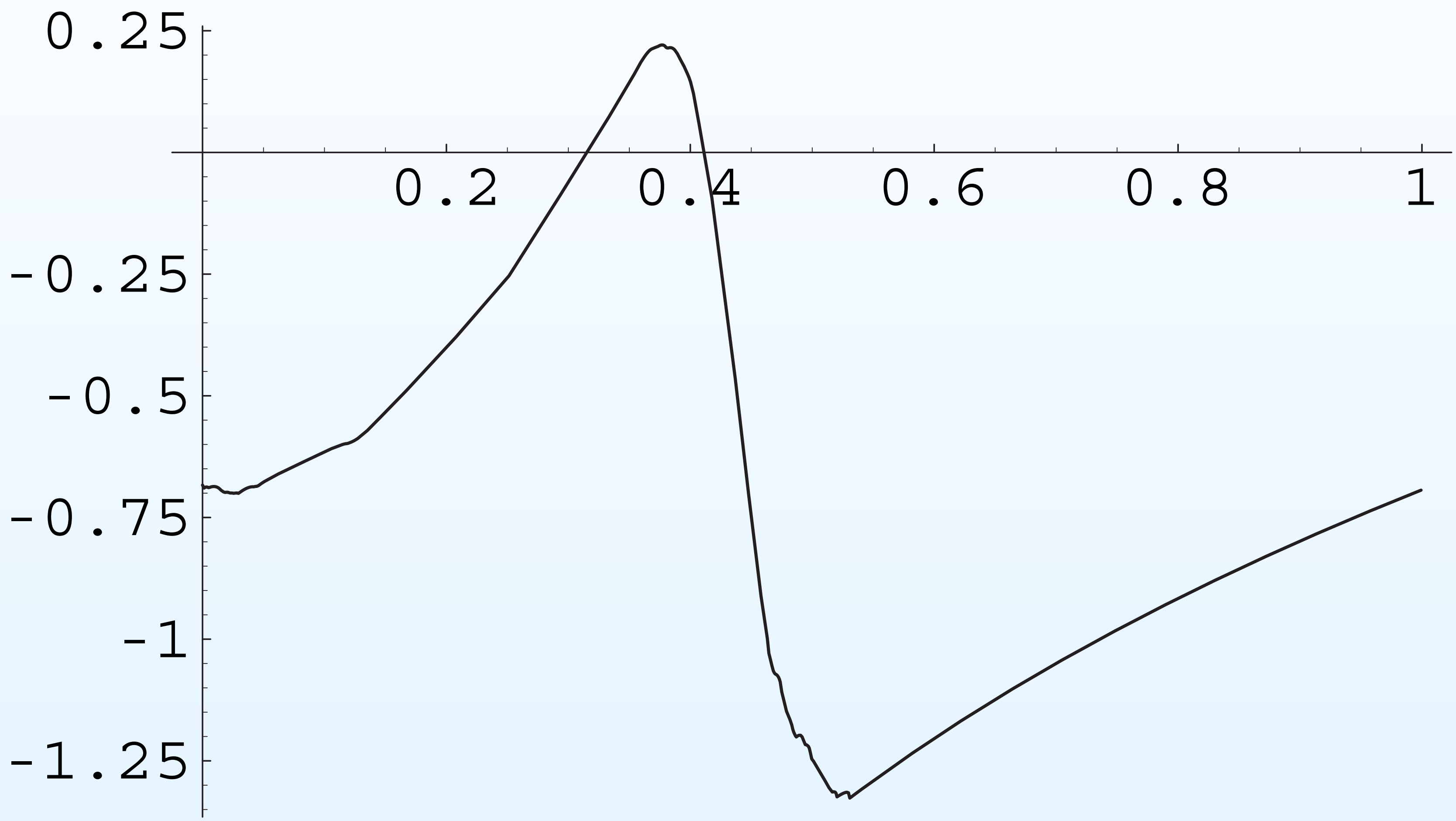
$$\begin{aligned}
&\sup_{0\le\theta\le 1} n^{-1}\log L_n(\theta) \\
\ge\quad & \frac{n-1}{n}\log(M_n/2) + \frac{1}{n}\log\frac{1-M_n}{\delta(M_n)} \\
\to_d\quad & \mathbb{D}
\end{aligned}$$

- where

$$P(\mathbb{D} \leq y) = \exp\left(-\frac{1}{2(y - \log 2)}\right), \quad y \geq \log(2)\,.$$

That is, with E an Exponential(1) random variable

$$\mathbb{D} \stackrel{d}{=} \log 2 + \frac{1}{2E}.$$


0.25
-0.25
-0.5
-0.75
-1
-1.25
0.2
0.4
0.6
0.8
1

- **Counterexample 2.** (4 B's, 1972). A distribution F on $[0, b)$ is star-shaped if $F(x)/x$ is non-decreasing on $[0, b)$. Thus if F has a density f which is increasing on $[0, b)$ then F is star-shaped.

- **Counterexample 2.** (4 B's, 1972). A distribution F on $[0, b)$ is star-shaped if $F(x)/x$ is non-decreasing on $[0, b)$. Thus if F has a density f which is increasing on $[0, b)$ then F is star-shaped.
- Let $\mathcal{F}_{star}$ be the class of all star-shaped distributions on $[0, b)$ for some b.

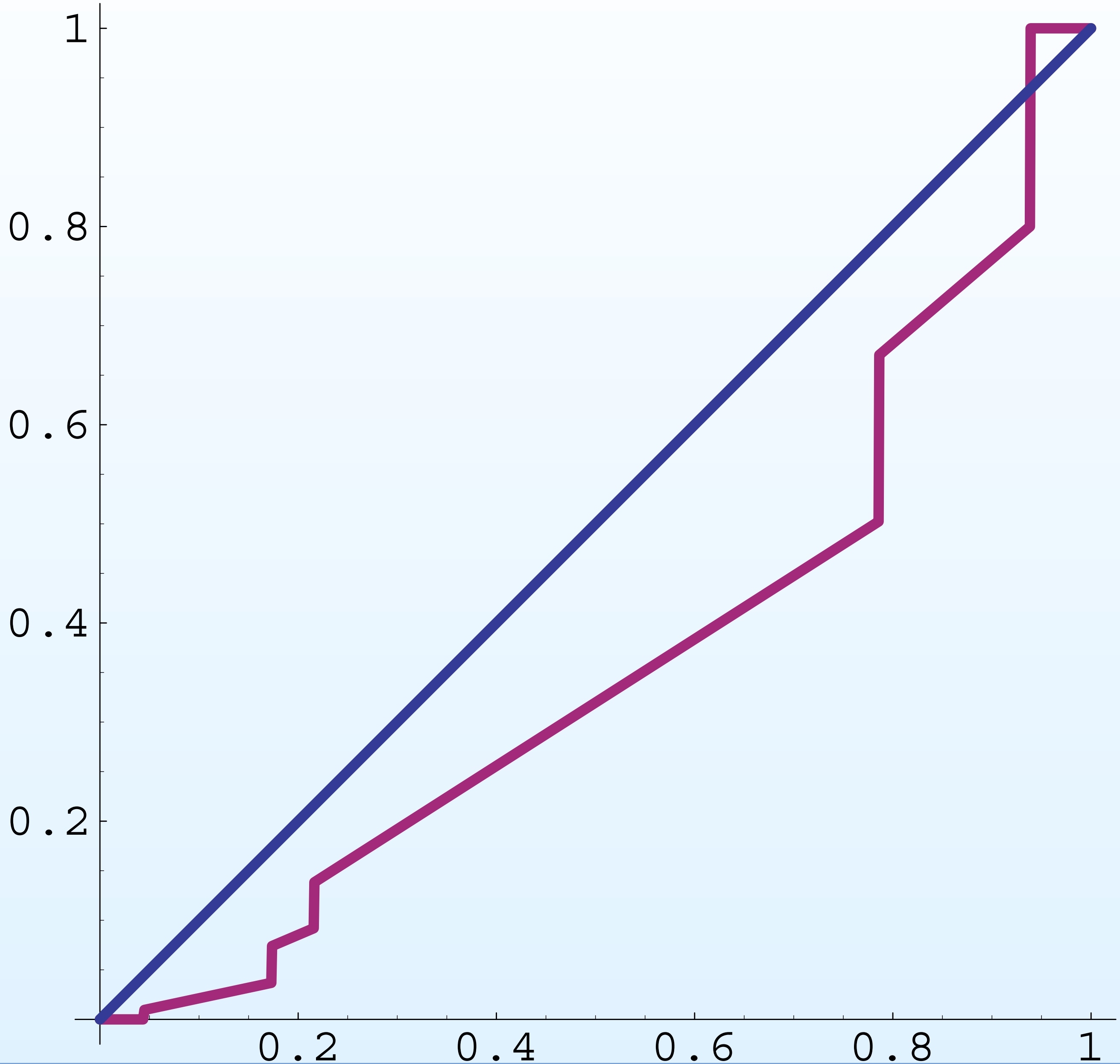
- **Counterexample 2.** (4 B's, 1972). A distribution F on $[0, b)$ is star-shaped if $F(x)/x$ is non-decreasing on $[0, b)$. Thus if F has a density f which is increasing on $[0, b)$ then F is star-shaped.
- Let $\mathcal{F}_{star}$ be the class of all star-shaped distributions on $[0, b)$ for some b.
- Suppose that $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{star}$.

- **Counterexample 2.** (4 B's, 1972). A distribution F on $[0, b)$ is star-shaped if $F(x)/x$ is non-decreasing on $[0, b)$. Thus if F has a density f which is increasing on $[0, b)$ then F is star-shaped.
- Let $\mathcal{F}_{star}$ be the class of all star-shaped distributions on $[0, b)$ for some b.
- Suppose that $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{star}$.
- Barlow, Bartholomew, Bremner, and Brunk (1972) show that the MLE of a star-shaped distribution function F is

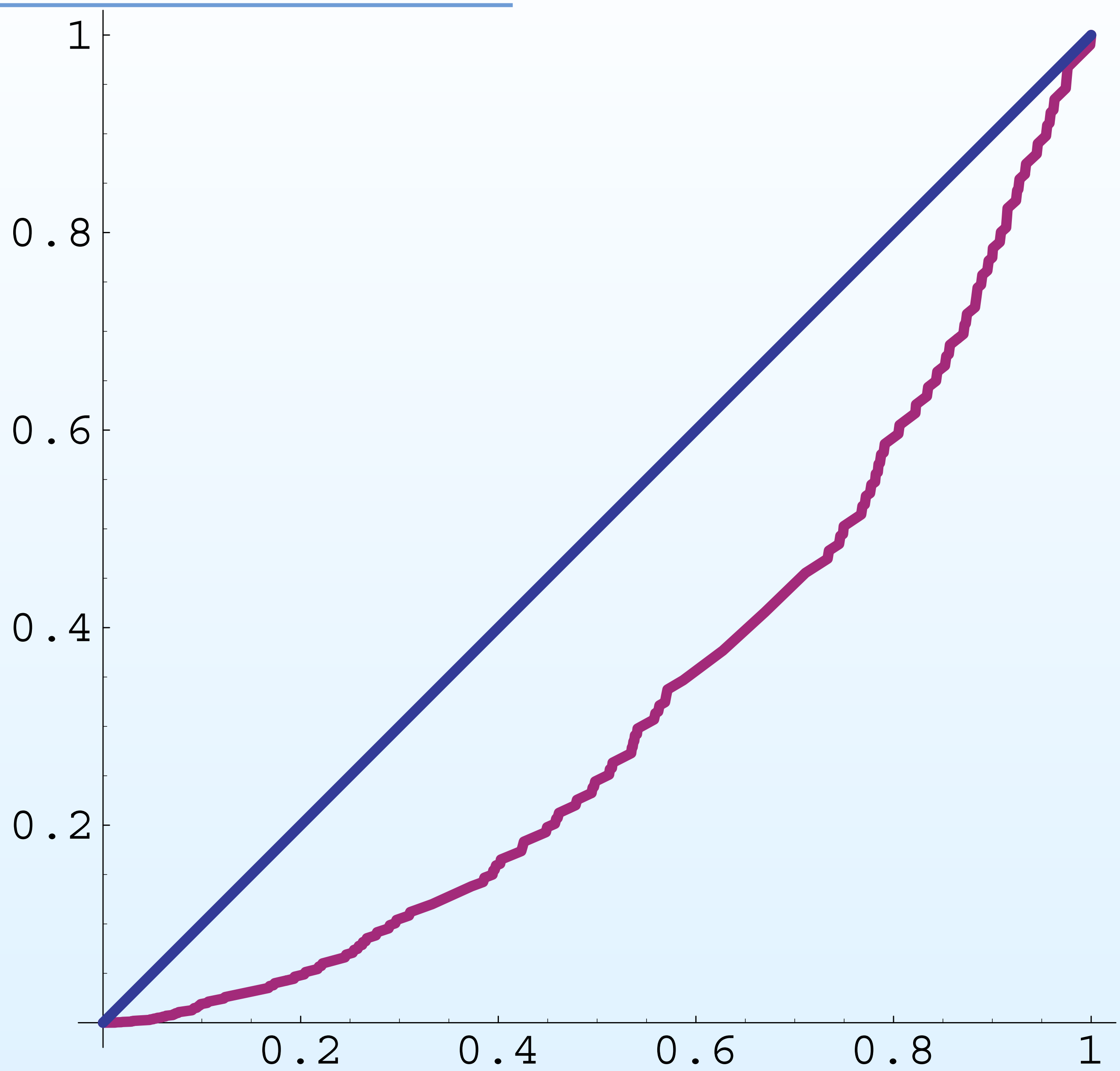
$$\hat{F}_n(x) = \begin{cases} 0, & x < X_{(1)} \\ \frac{ix}{nX_{(n)}}, & X_{(i)} \leq x < X_{(i+1)}, \; i = 1, \ldots, n-1, \\ 1, & x \geq X_{(n)} \, . \end{cases}$$

- Moreover, BBBB (1972) show that if $F(x) = x$ for $0 \le x \le 1$, then
$$\hat{F}_n(x) \rightarrow_{a.s.} x^2 \neq x$$
for $0 \le x \le 1$.

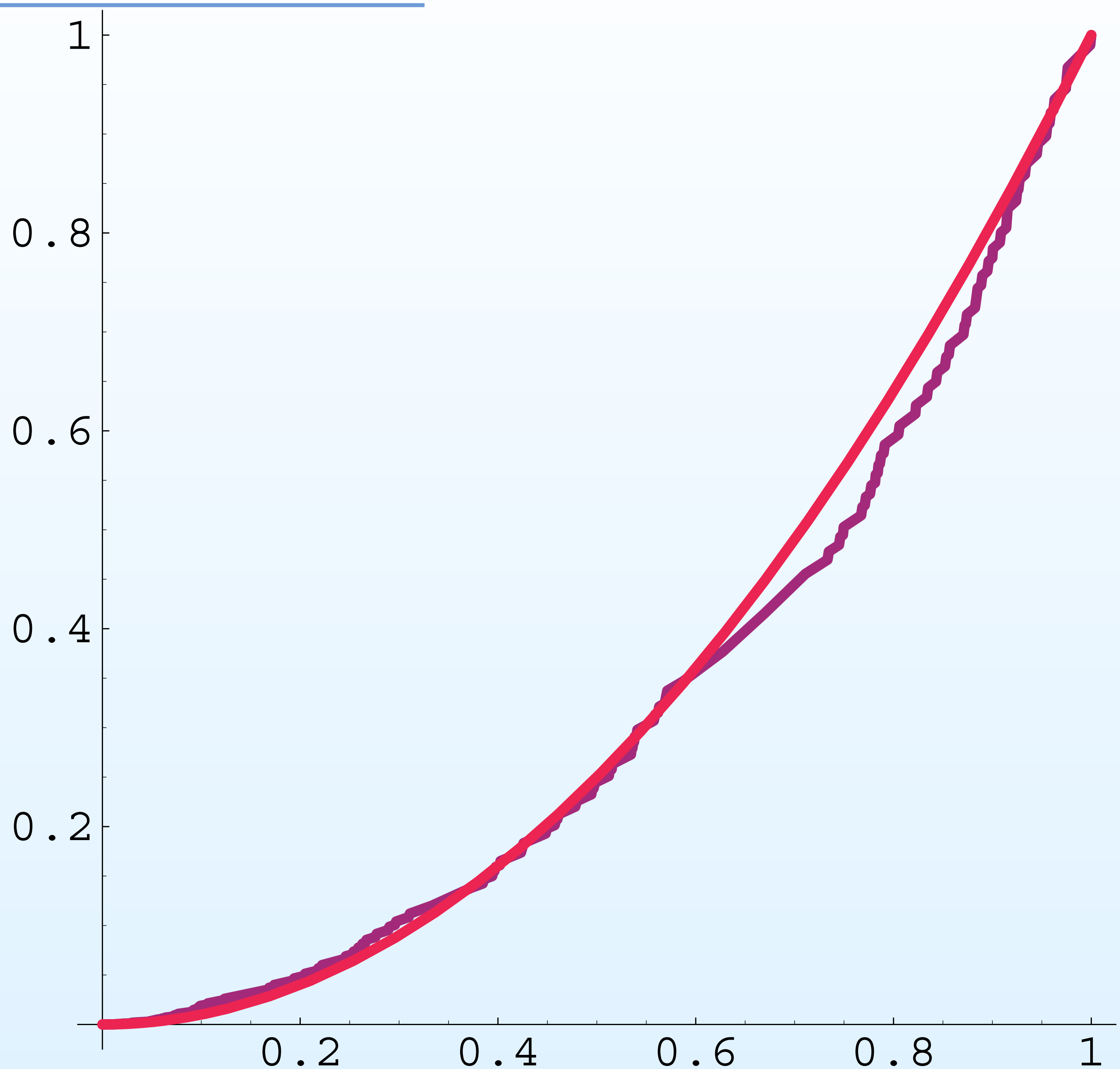
MLE $n = 5$ and true d.f.



MLE $n = 100$ and true d.f.



MLE $n = 100$ and limit



- **Note 1.** Since $X_{(i)} \stackrel{d}{=} S_j/S_{n+1}$ where $S_i = \sum_{j=1}^{i} E_j$ with E_j i.i.d. Exponential(1) rv's, the total mass at order statistics equals

$$\frac{1}{nX_{(n)}}\sum_{i=1}^{n} X_{(i)} \stackrel{d}{=} \sum_{i=1}^{n}\frac{S_i}{nS_n} = \frac{n}{S_n}\frac{1}{n}\sum_{j=1}^{n}\left(1-\frac{j-1}{n}\right)E_j$$

$$\to_p \quad 1\cdot\int_0^1 (1-t)dt = 1/2\,.$$

- **Note 1.** Since $X_{(i)} \stackrel{d}{=} S_j/S_{n+1}$ where $S_i = \sum_{j=1}^{i} E_j$ with E_j i.i.d. Exponential(1) rv's, the total mass at order statistics equals

$$\frac{1}{nX_{(n)}} \sum_{i=1}^{n} X_{(i)} \quad \stackrel{d}{=} \quad \sum_{i=1}^{n} \frac{S_i}{nS_n} = \frac{n}{S_n} \frac{1}{n} \sum_{j=1}^{n} \left(1 - \frac{j-1}{n}\right) E_j$$

$$\to_p \quad 1 \cdot \int_0^1 (1-t)dt = 1/2 \,.$$

- **Note 2.** BBBB (1972) present consistent estimators of F star-shaped via isotonization due to Barlow and Scheurer (1971) and van Zwet.

- **Counterexample 3.** (Boyles, Marshall, Proschan (1985). A distribution F on $[0, \infty)$ is Increasing Failure Rate Average if

$$\frac{1}{x}\{-\log(1 - F(x))\} \equiv \frac{1}{x}\Lambda(x)$$

is non-decreasing; that is, if Λ is star-shaped.

- **Counterexample 3.** (Boyles, Marshall, Proschan (1985). A distribution F on $[0, \infty)$ is Increasing Failure Rate Average if

$$\frac{1}{x}\{-\log(1 - F(x))\} \equiv \frac{1}{x}\Lambda(x)$$

is non-decreasing; that is, if Λ is star-shaped.
- Let $\mathcal{F}_{IFRA}$ be the class of all IFRA- distributions on $[0, \infty)$.

- Suppose that $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{IFRA}$. Boyles, Marshall, and Proschan (1985) showed that the MLE $\hat{F}_n$ of a IFRA-distribution function F is given by

$$-\log(1-\hat{F}_n(x)) = \begin{cases} \hat{\lambda}_j, & X_{(j)} \leq x < X_{(j+1)}, \\ & \quad j = 0, \ldots, n-1 \\ \infty, & x > X_{(n)} \end{cases}$$

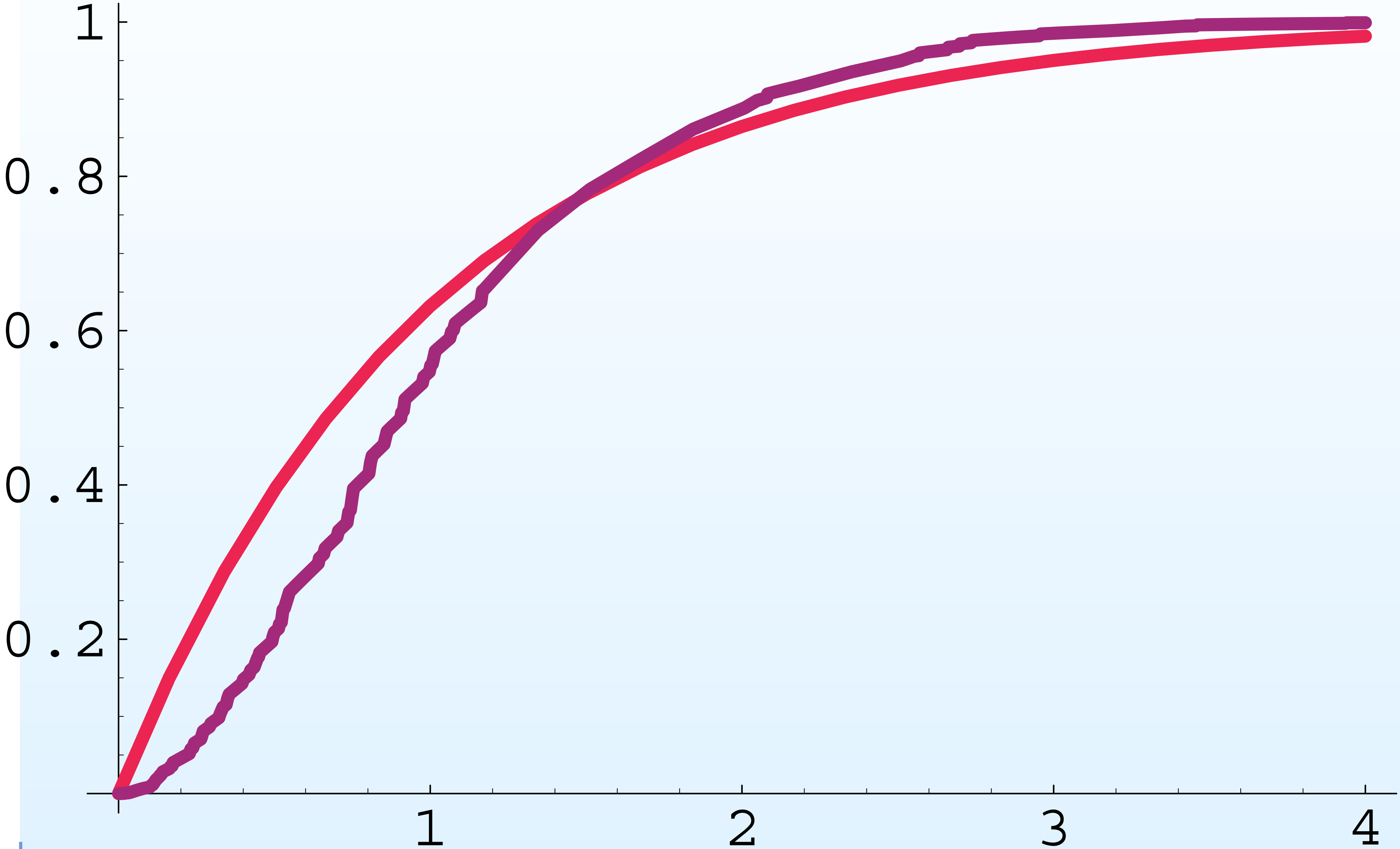
where

$$\hat{\lambda}_j = \sum_{i=1}^{j} X_{(i)}^{-1} \log\left(\frac{\sum_{k=i}^{n} X_{(k)}}{\sum_{k=i+1}^{n} X_{(k)}}\right).$$

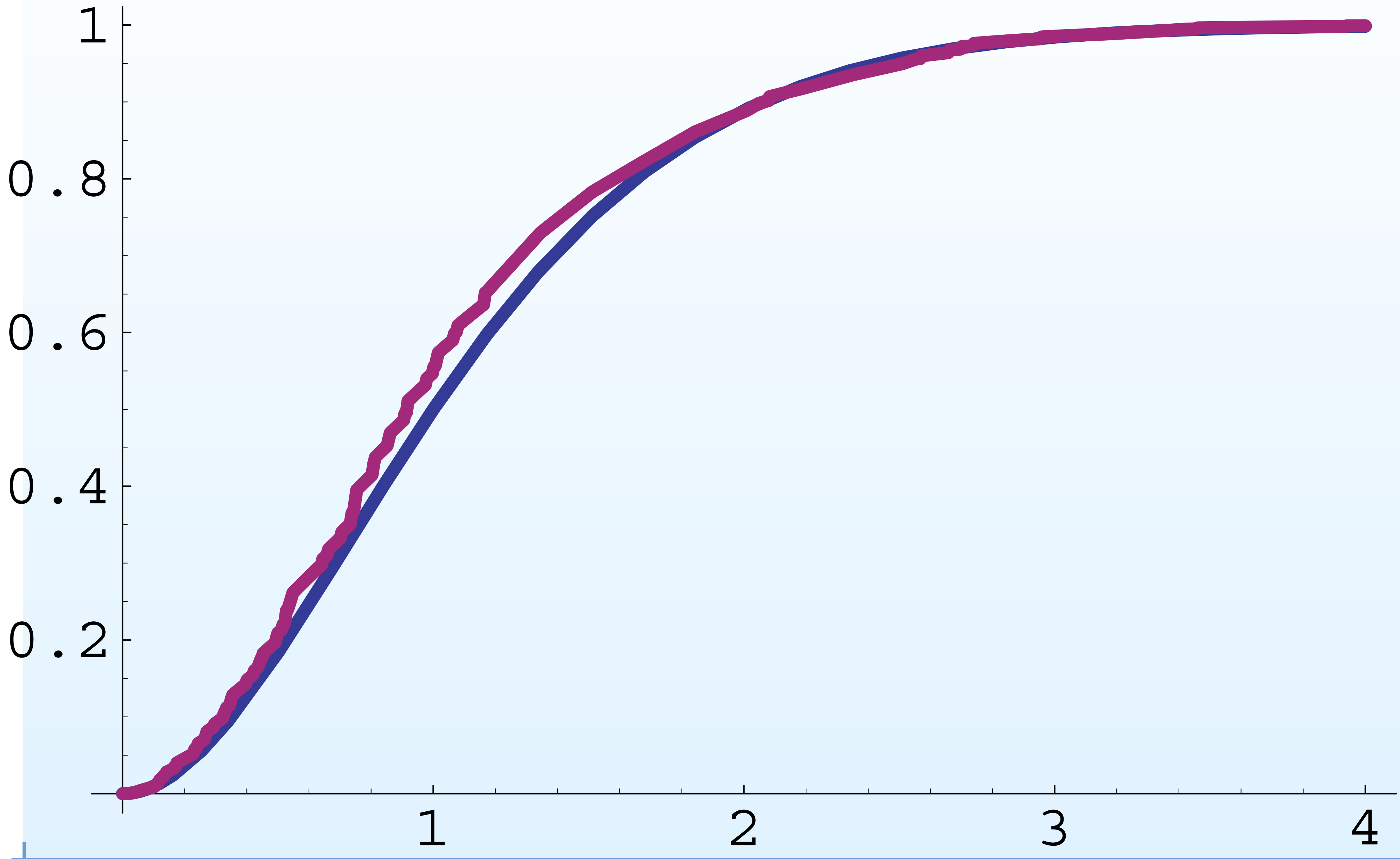
- Moreover, BMP (1985) show that if F is exponential(1), then

$$1 - \hat{F}_n(x) \rightarrow_{a.s.} (1+x)^{-x} \neq \exp(-x), \quad \text{so}$$
$$\frac{1}{x}\hat{\Lambda}_n(x) \rightarrow_{a.s.} \log(1+x) \neq 1\,.$$

MLE $n = 100$ and true d.f. $1 - \exp(-x)$



MLE $n = 100$ and limit d.f. $(1 + x)^{-x}$



More counterexamples:

- bivariate right censoring: Tsai, van der Laan, Pruitt

More counterexamples:

- bivariate right censoring: Tsai, van der Laan, Pruitt
- left truncation and interval censoring: Chappell and Pan (1999)

More counterexamples:

- bivariate right censoring: Tsai, van der Laan, Pruitt
- left truncation and interval censoring:
  Chappell and Pan (1999)
- bivariate interval censoring with a continuous mark:
  Hudgens, Maathuis, and Gilbert (2005)
  Maathuis and Wellner (2005)

3. Beyond consistency: rates and distributions

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

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

- If

$$\log N_{[\,]}(\epsilon, \mathcal{P}) \asymp \frac{K}{\epsilon^{1/\gamma}} \tag{2}$$

  (1) leads to the optimal rate of convergence

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

- On the other hand, bounds (from Birgé and Massart (1993)), yield achieved rates of convergence for maximum likelihood estimators (and other minimum contrast estimators) $r_n = r_n^{ach}$ determined by

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

- If (2) holds, this leads to the rate

$$\begin{cases} n^{\gamma/(2\gamma+1)} & \text{if } \gamma > 1/2 \\ n^{\gamma/2} & \text{if } \gamma < 1/2\,. \end{cases}$$

- Thus there is the possibility that maximum likelihood is not (rate-)optimal when $\gamma < 1/2$.

- 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" of the functions in $\mathcal{P}$.

- 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" of the functions in $\mathcal{P}$.

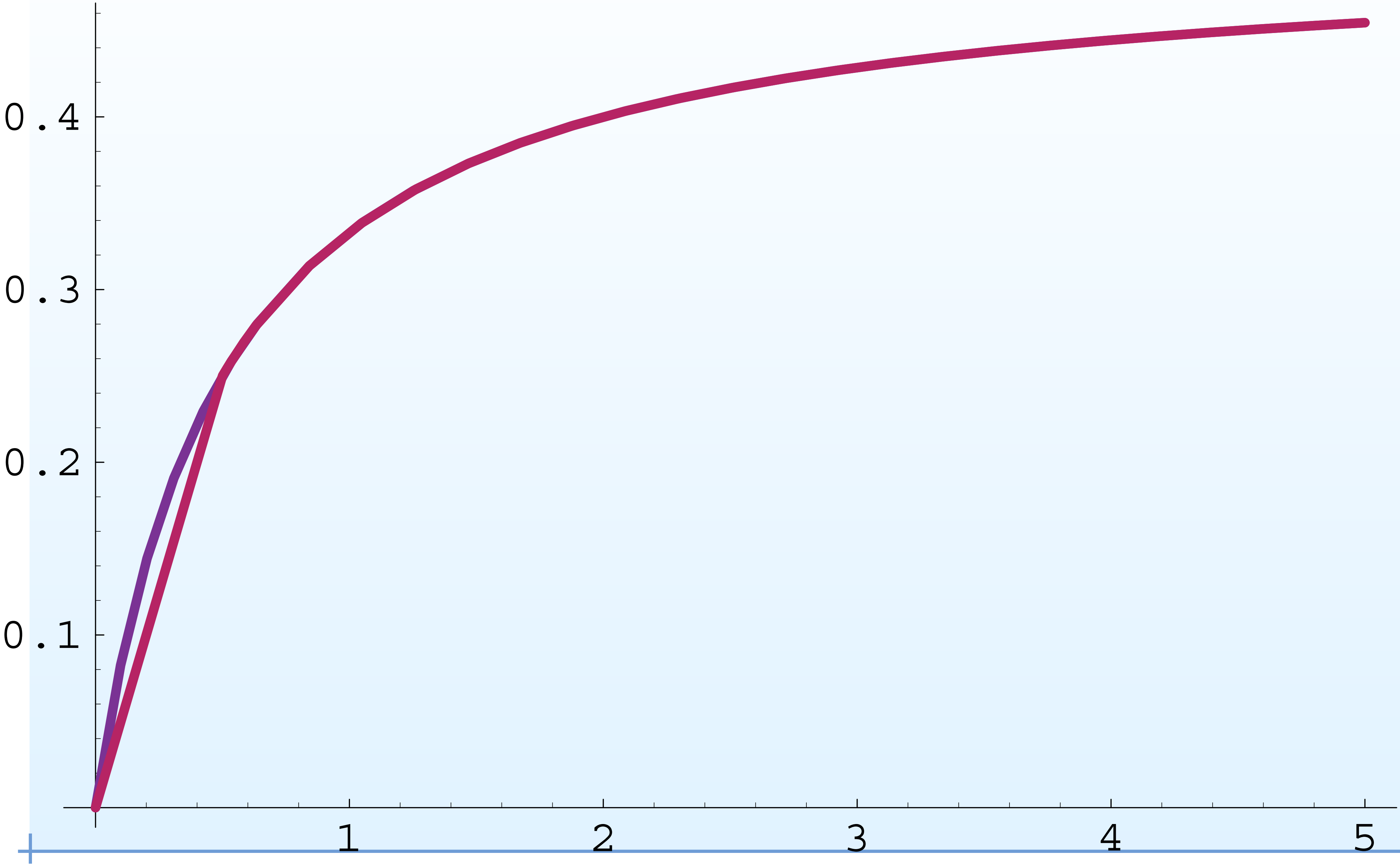
- Hence

$$\alpha < \frac{d}{2}$$

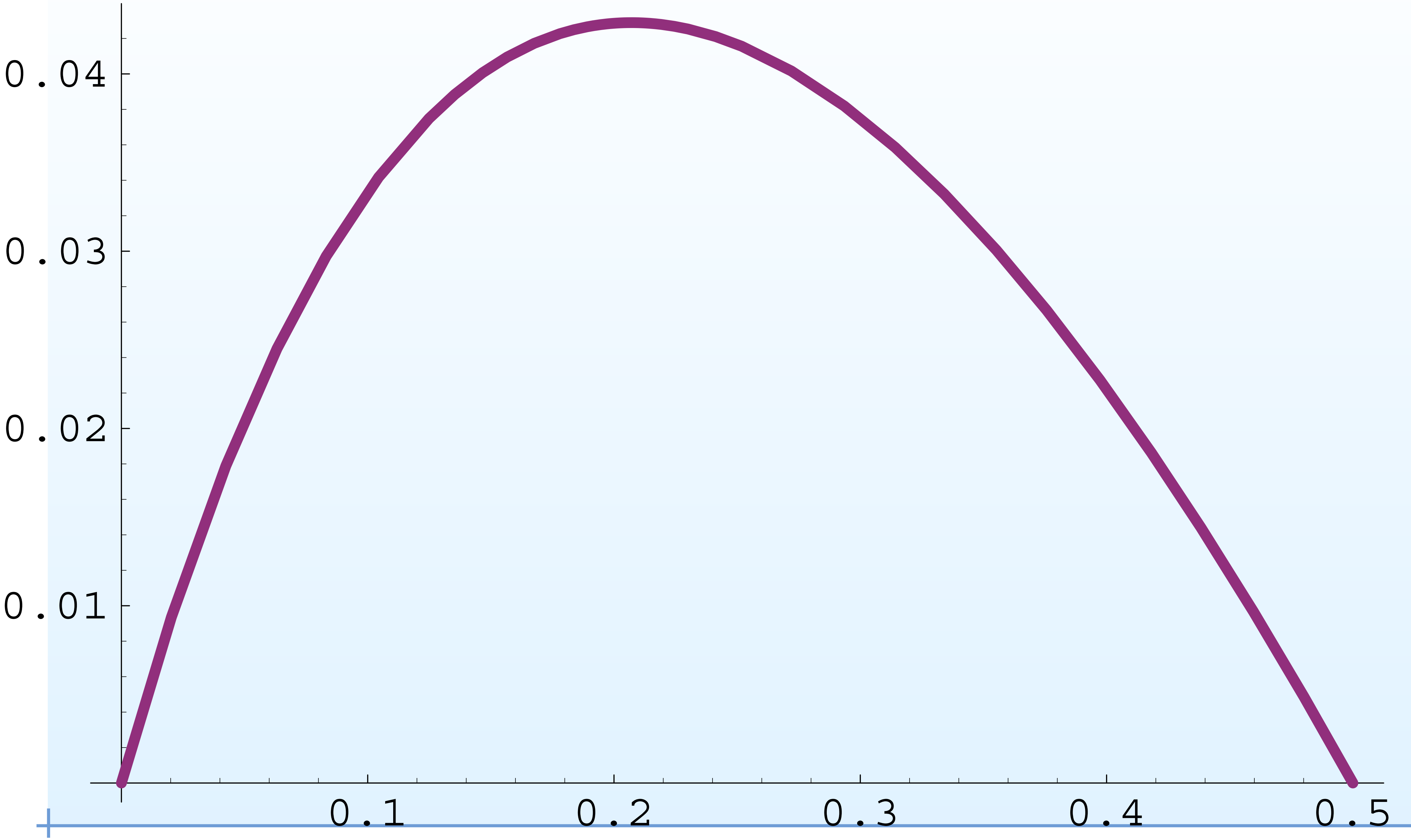
leads to $\gamma < 1/2$.

- 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" of the functions in $\mathcal{P}$.
- Hence
$$\alpha < \frac{d}{2}$$
leads to $\gamma < 1/2$.
- But there are many examples/problem with $\gamma > 1/2$!

Optimal rate and MLE rate as a function of γ



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



4. Positive Examples (some still in progress!)

- Interval censoring (Groeneboom)
  case 1, current status data
  case 2 (Groeneboom)

4. Positive Examples (some still in progress!)

- Interval censoring (Groeneboom)
  case 1, current status data
  case 2 (Groeneboom)
- panel count data
  (Wellner and Zhang, 2000)

4. Positive Examples (some still in progress!)

- Interval censoring (Groeneboom)
  case 1, current status data
  case 2 (Groeneboom)
- panel count data
  (Wellner and Zhang, 2000)
- $k-$monotone densities
  (Balabdaoui and Wellner, 2004)

4. Positive Examples (some still in progress!)

- Interval censoring (Groeneboom)
  case 1, current status data
  case 2 (Groeneboom)
- panel count data
  (Wellner and Zhang, 2000)
- $k-$monotone densities
  (Balabdaoui and Wellner, 2004)
- competing risks current status data
  (Jewell and van der Laan; Maathuis)

- **Example 1.** (interval censoring, case 1)

- **Example 1.** (interval censoring, case 1)
  - $X \sim F$, $Y \sim G$ independent
    Observe $(1\{X \leq Y\}, Y) \equiv (\Delta, Y)$.
    Goal: estimate F. MLE $\hat{F}_n$ exists

- Example 1. (interval censoring, case 1)
  - $X \sim F$, $Y \sim G$ independent
    Observe $(1\{X \leq Y\}, Y) \equiv (\Delta, Y)$.
    Goal: estimate F. MLE $\hat{F}_n$ exists
  - Global rate: $d = 1$, $\alpha = 1$, $\gamma = \alpha/d = 1$.
    $\gamma/(2\gamma + 1) = 1/3$, so $r_n = n^{1/3}$:

$$n^{1/3} h(p_{\hat{F}_n}, p_0) = O_p(1)$$

    and this yields

$$n^{1/3} \int |\hat{F}_n - F_0| dG = O_p(1) \,.$$

- Interval censoring case 1, continued:

- Interval censoring case 1, continued:
  - **Local rate:** (Groeneboom, 1987)

$$n^{1/3}(\hat{F}_n(t_0) - F(t_0)) \\ \to_d \left\{ \frac{F(t_0)(1-F(t_0))f_0(t_0)}{2g(t_0)} \right\}^{1/3} 2\mathbb{Z}$$

where $\mathbb{Z} = \text{argmin}\{W(t) + t^2\}$

- Example 2. (interval censoring, case 2)

- Example 2. (interval censoring, case 2)
  - $X \sim F$, $(U, V) \sim H$, $U \leq V$ independent of X
    Observe i.i.d. copies of (Δ, U, V) where

$$\begin{aligned} \Delta &= (\Delta_1, \Delta_2, \Delta_3) \\ &= (1\{X \leq U\}, 1\{U < X \leq V\}, 1\{V < X\}) \end{aligned}$$

- Example 2. (interval censoring, case 2)
  - $X \sim F$, $(U, V) \sim H$, $U \leq V$ independent of X
    Observe i.i.d. copies of (Δ, U, V) where

$$\begin{aligned}\Delta &= (\Delta_1, \Delta_2, \Delta_3) \\ &= (1\{X \leq U\}, 1\{U < X \leq V\}, 1\{V < X\})\end{aligned}$$

  - Goal: estimate F. MLE $\hat{F}_n$ exists.

- (interval censoring, case 2, continued)

- (interval censoring, case 2, continued)
  - Global rate (separated case): If $P(V - U \geq \epsilon) = 1$, $d = 1$, $\alpha = 1$, $\gamma = \alpha/d = 1$
    $\gamma/(2\gamma + 1) = 1/3$, so $r_n = n^{1/3}$

$$n^{1/3} h(p_{\hat{F}_n}, p_0) = O_p(1)$$

and this yields

$$n^{1/3} \int |\hat{F}_n - F_0| d\mu = O_p(1)$$

where

$$\mu(A) = P(U \in A) + P(V \in A), \quad A \in \mathcal{B}_1$$

- (interval censoring, case 2, continued)

- (interval censoring, case 2, continued)
  - Global rate (nonseparated case): (van de Geer, 1993).

$$\frac{n^{1/3}}{(\log n)^{1/6}} h(p_{\hat{F}_n}, p_0) = O_p(1) \, .$$

Although this looks “worse” in terms of the rate, it is actually better because the Hellinger metric is much stronger in this case.

- (interval censoring, case 2, continued)
  - Global rate (nonseparated case): (van de Geer, 1993).

$$\frac{n^{1/3}}{(\log n)^{1/6}} h(p_{\hat{F}_n}, p_0) = O_p(1) \, .$$

Although this looks “worse” in terms of the rate, it is actually better because the Hellinger metric is much stronger in this case.

(interval censoring, case 2, continued)

- Local rate (separated case): (Groeneboom, 1996)

$$n^{1/3}(\hat{F}_n(t_0) - F_0(t_0)) \to_d \left\{ \frac{f_0(t_0)}{2a(t_0)} \right\}^{1/3} 2\mathbb{Z}$$

where $\mathbb{Z} = \text{argmin}\{W(t) + t^2\}$ and

$$a(t_0) = \frac{h_1(t_0)}{F_0(t_0)} + k_1(t_0) + k_2(t_0) + \frac{h_2(t_0)}{1 - F_0(t_0)}$$

$$k_1(u) = \int_u^M \frac{h(u,v)}{F_0(v) - F_0(u)} dv$$

$$k_2(v) = \int_0^v \frac{h(u,v)}{F_0(v) - F_0(u)} du$$

(interval censoring, case 2, continued)

- Local rate (non-separated case): (conjectured, G&W, 1992)

$$(n\log n)^{1/3}(\hat{F}_n(t_0) - F_0(t_0)) \rightarrow_d \left\{ \frac{3}{4} \frac{f_0(t_0)^2}{h(t_0, t_0)} \right\}^{1/3} 2\mathbb{Z}$$

where $\mathbb{Z} =$ argmin$\{W(t) + t^2\}$

(interval censoring, case 2, continued)

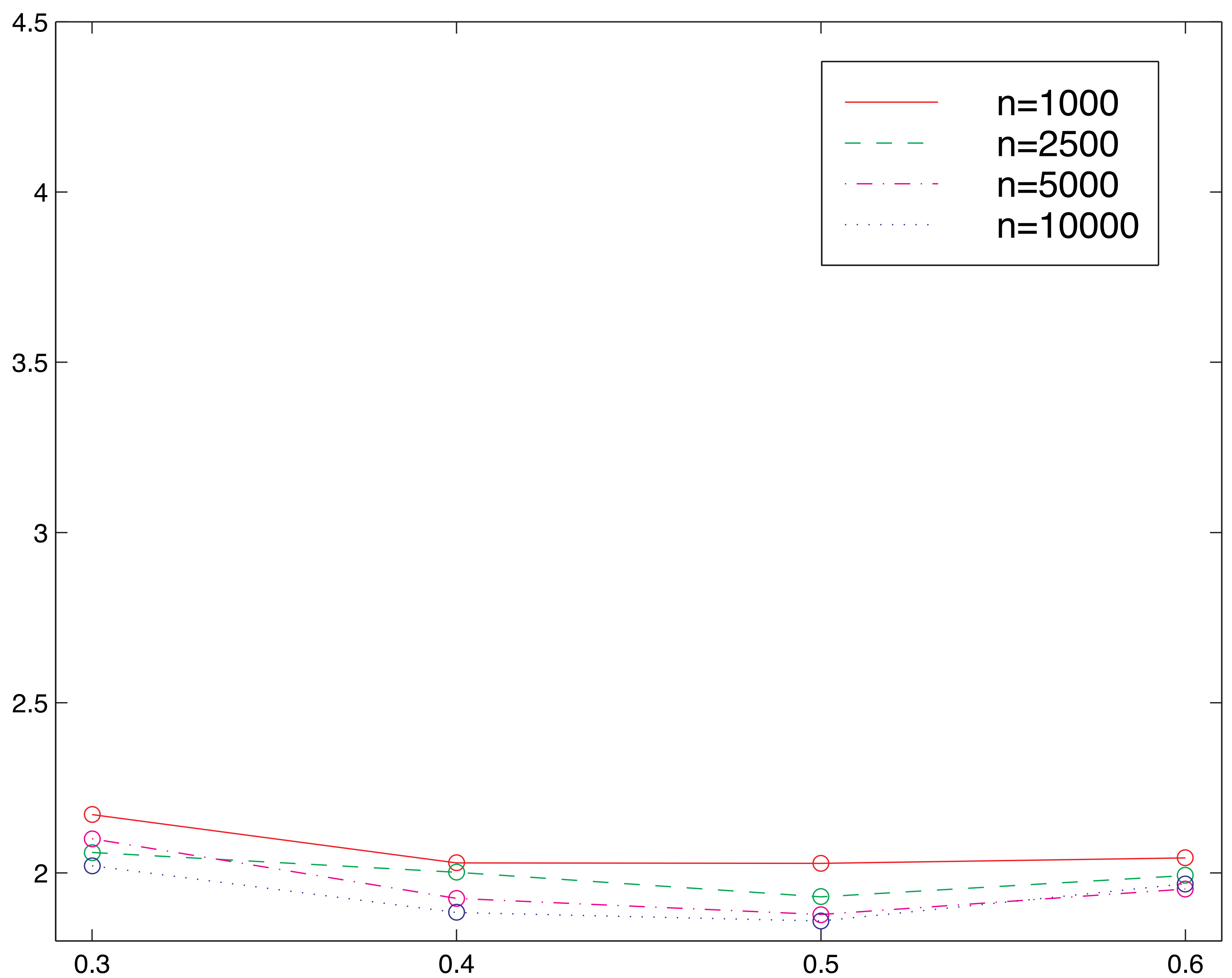
- Local rate (non-separated case): (conjectured, G&W, 1992)

$$(n \log n)^{1/3}(\hat{F}_n(t_0) - F_0(t_0)) \to_d \left\{ \frac{3}{4} \frac{f_0(t_0)^2}{h(t_0,t_0)} \right\}^{1/3} 2\mathbb{Z}$$

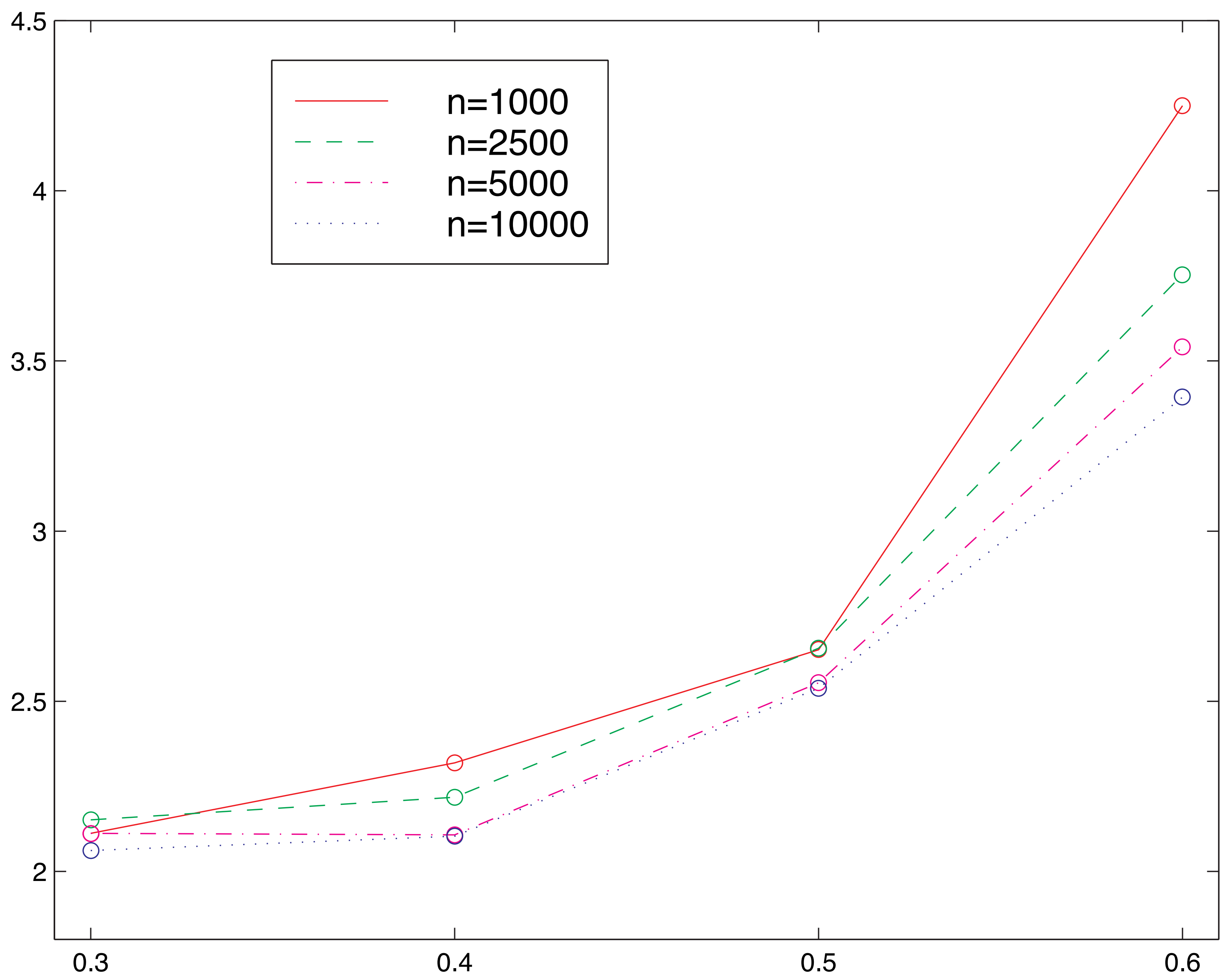
  where $\mathbb{Z} = \text{argmin}\{W(t) + t^2\}$

- Monte-Carlo evidence in support:
  Groeneboom and Ketelaars (2005)

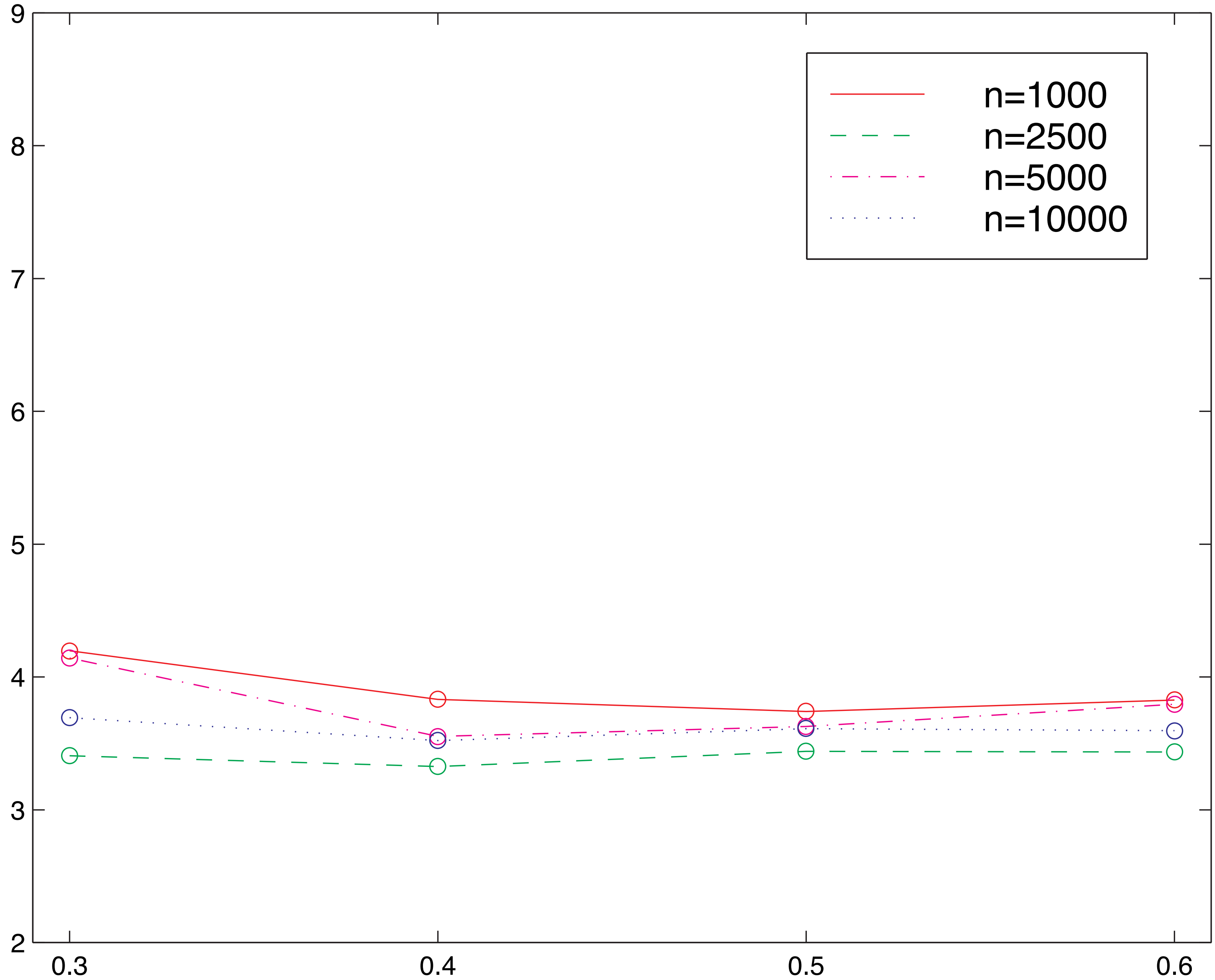
MSE histogram / MSE of MLE $f_0(t) = 1$



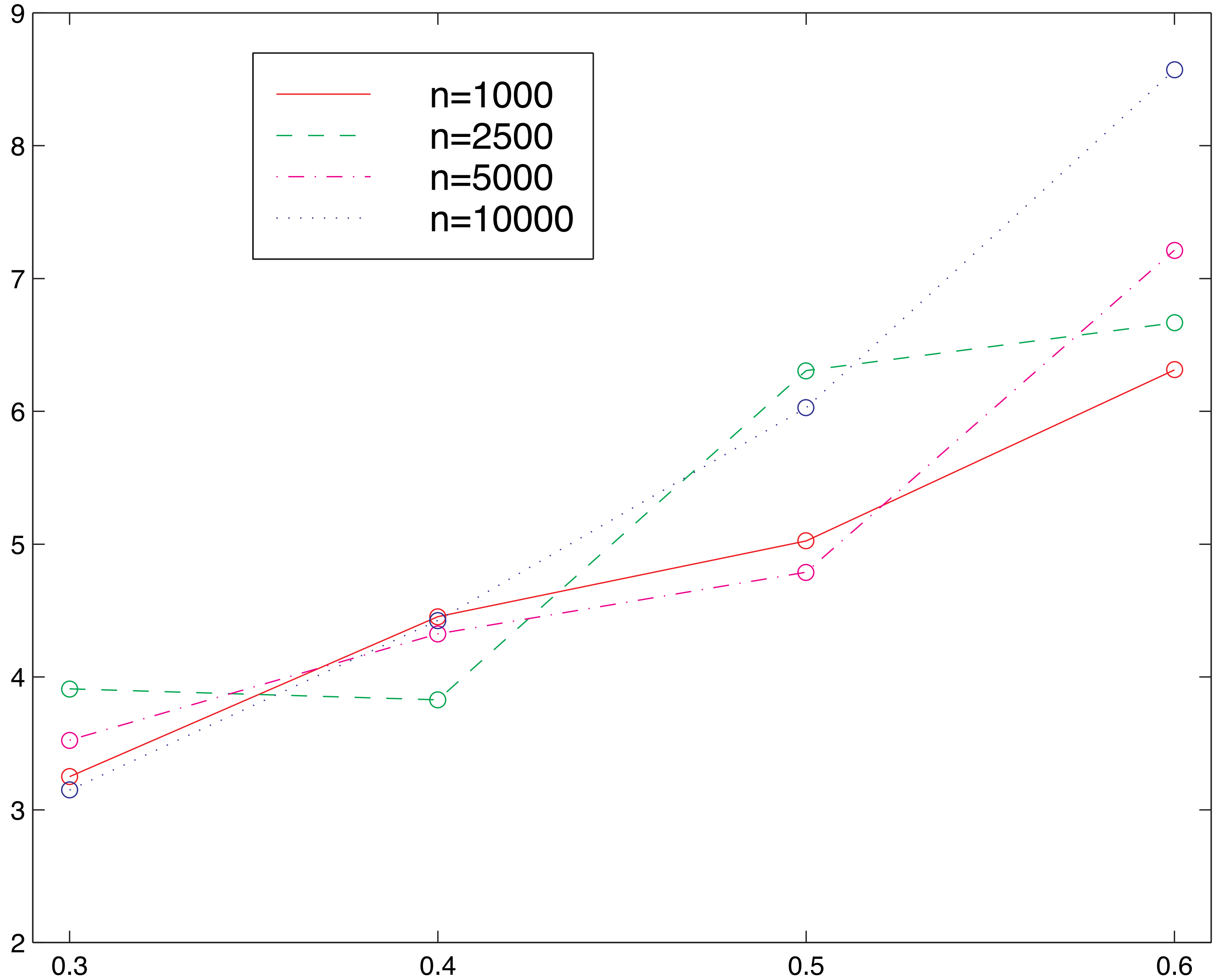
MSE histogram / MSE of MLE $f_0(t) = 4(1-t)^3$



MSE histogram / MSE of MLE $f_0(t) = 1$



MSE histogram / MSE of MLE $f_0(t) = 4(1-t)^3$



- **Example 3.** (k-monotone densities)

- Example 3. (k-monotone densities)
  - A density p on $(0,\infty)$ is $k-$mononontone ($p \in \mathcal{D}_k$) if it is non-negative and nonincreasing when $k=1$; and if $(-1)^j p^{(j)}(x) \geq 0$ for $j=0,\ldots,k-2$ and $(-1)^{k-2}p^{(k-2)}$ is convex for $k \geq 2$.

- Example 3. (k-monotone densities)
  - A density p on $(0,\infty)$ is $k-$monontone ($p \in \mathcal{D}_k$) if it is non-negative and nonincreasing when $k=1$; and if $(-1)^j p^{(j)}(x) \ge 0$ for $j=0,\ldots,k-2$ and $(-1)^{k-2} p^{(k-2)}$ is convex for $k \ge 2$.
  - Mixture representation: $p \in \mathcal{D}_k$ iff

$$p(x) = \int_0^\infty \frac{k}{y^k}(y-x)_+^{k-1} dF(y)$$

for some distribution function F on $(0,\infty)$.

□ $k=1$: monotone decreasing densities on $\mathbb{R}^+$

□ $k=2$: convex decreasing densities on $\mathbb{R}^+$

□ $k \ge 3$: $\ldots$

□ $k=\infty$: completely monotone densities
= scale mixtures of exponential

(k-monotone densities, continued)

- The MLE $\hat{p}_n$ of $p_0 \in \mathcal{D}_k$ exists and is characterized by

$$\int_0^\infty \frac{k}{y^k} \frac{(y-x)_+^k}{\hat{p}_n(x)} d\mathbb{P}_n(x) \begin{cases} \leq 1, & \text{for all } \; y \geq 0 \\ = 1, & \text{if } \; (-1)^k \hat{p}_n^{(k-1)}(y-) > \hat{p}_n^{(k-1)}(y+) \end{cases}$$

(k-monotone densities, continued)

- The MLE $\hat{p}_n$ of $p_0 \in \mathcal{D}_k$ exists and is characterized by

$$\int_0^\infty \frac{k}{y^k} \frac{(y-x)_+^k}{\hat{p}_n(x)} d\mathbb{P}_n(x) \begin{cases} \le 1, & \text{for all } y \ge 0 \\ = 1, & \text{if } (-1)^k \hat{p}_n^{(k-1)}(y-) > \hat{p}_n^{(k-1)}(y+) \end{cases}$$

- $k = 1$ Grenander estimator: $r_n = n^{1/3}$

(k-monotone densities, continued)

- The MLE $\hat{p}_n$ of $p_0 \in \mathcal{D}_k$ exists and is characterized by

$$\int_0^\infty \frac{k}{y^k} \frac{(y-x)_+^k}{\hat{p}_n(x)} d\mathbb{P}_n(x) \begin{cases} \leq 1, & \text{for all } \ y \geq 0 \\ = 1, & \text{if } \ (-1)^k \hat{p}_n^{(k-1)}(y-) > \hat{p}_n^{(k-1)}(y+) \end{cases}$$

- $k = 1$ Grenander estimator: $r_n = n^{1/3}$
  - Global rates and finite n minimax bounds: Birgé (1986), (1987), (1989)

(k-monotone densities, continued)

- The MLE $\hat{p}_n$ of $p_0 \in \mathcal{D}_k$ exists and is characterized by

$$\int_0^\infty \frac{k}{y^k} \frac{(y-x)_+^k}{\hat{p}_n(x)} d\mathbb{P}_n(x) \begin{cases} \leq 1, & \text{for all } \; y \geq 0 \\ = 1, & \text{if } \; (-1)^k \hat{p}_n^{(k-1)}(y-) > \hat{p}_n^{(k-1)}(y+) \end{cases}$$

- $k = 1$ Grenander estimator: $r_n = n^{1/3}$
  - Global rates and finite n minimax bounds: Birgé (1986), (1987), (1989)
  - Local rates: Prakasa Rao (1969), Groeneboom (1985), (1989)

    Kim and Pollard (1990)

$$n^{1/3}(\hat{p}_n(t_0) - p_0(t_0)) \to_d \left\{ \frac{p_0(t_0)|p_0'(t_0)|}{2} \right\}^{1/3} 2\mathbb{Z}$$

(k-monotone densities, continued)

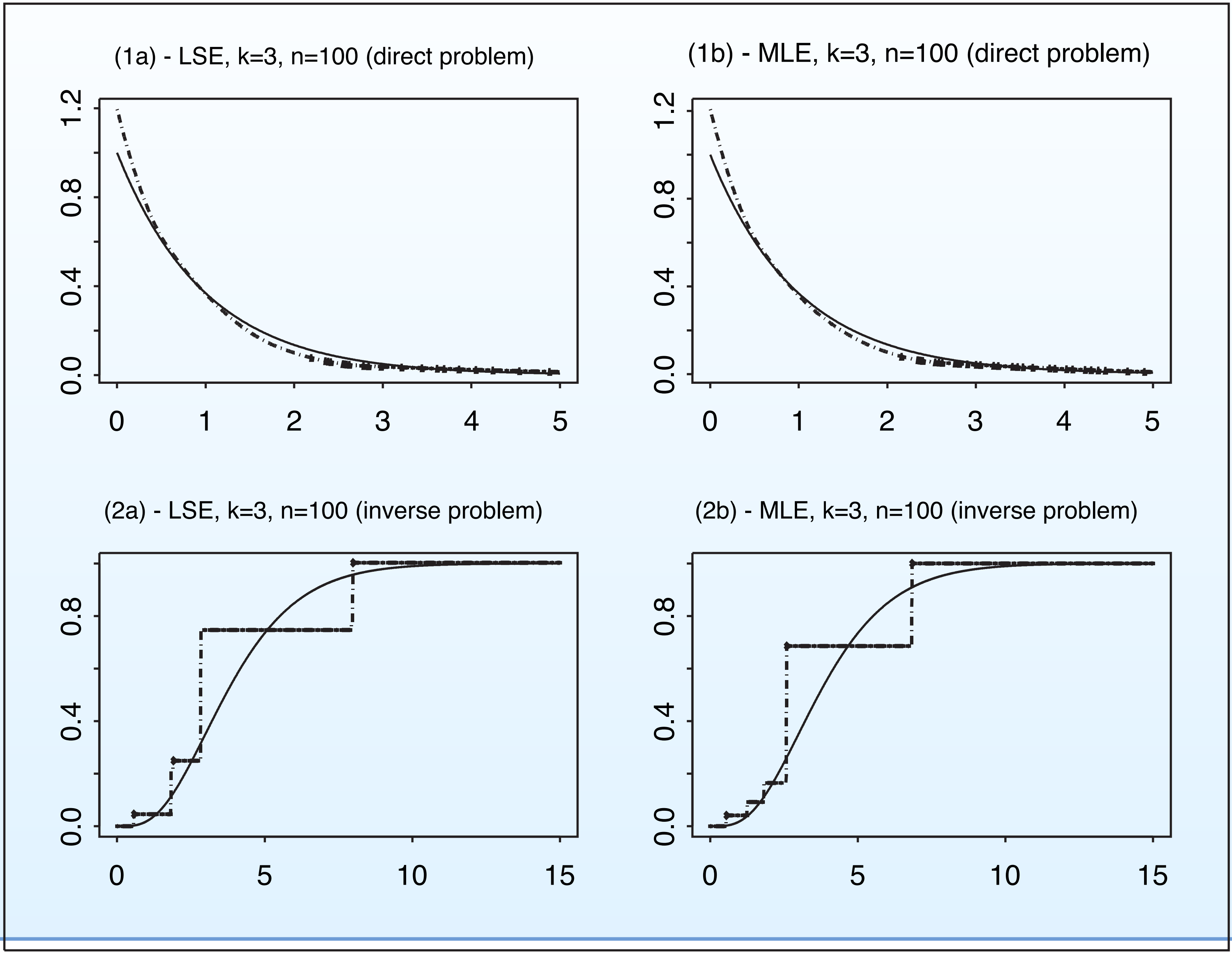
- $k = 2$; convex decreasing density

  $d = 1$, $\alpha = 2$, $\gamma = 2$, $\gamma/(2\gamma + 1) = 2/5$, so $r_n = n^{2/5}$
  (forward problem)
  - ☐ Global rates: nothing yet
  - ☐ Local rates and distributions:
    Groeneboom, Jongbloed, Wellner (2001)

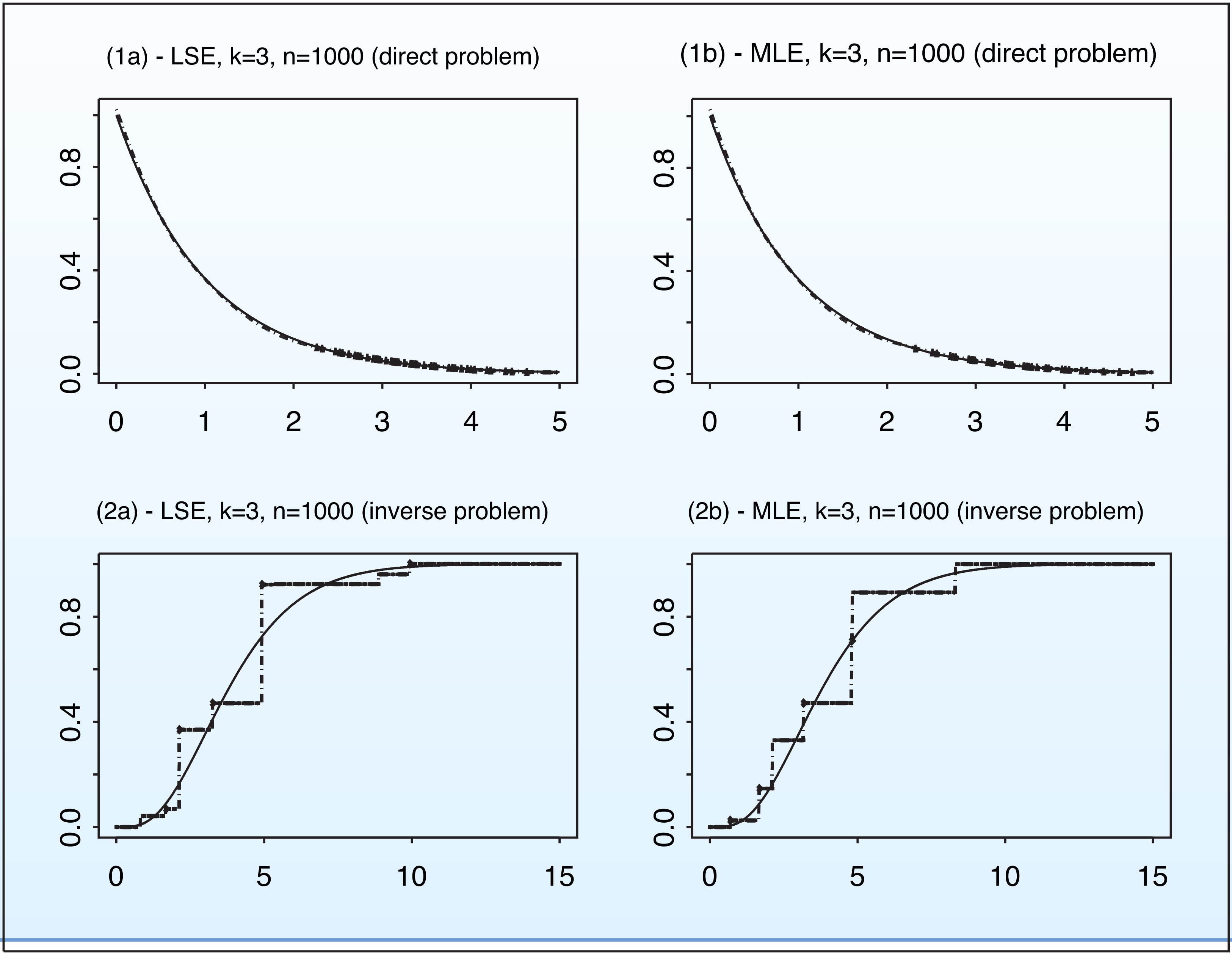
(k-monotone densities, continued)

- $k = 2$; convex decreasing density
  $d = 1$, $\alpha = 2$, $\gamma = 2$, $\gamma/(2\gamma + 1) = 2/5$, so $r_n = n^{2/5}$ (forward problem)
  - ☐ Global rates: nothing yet
  - ☐ Local rates and distributions:
    Groeneboom, Jongbloed, Wellner (2001)
- $k \geq 3$; k - monotone density
  $d = 1$, $\alpha = k$, $\gamma = k$, $\gamma/(2\gamma + 1) = k/(2k + 1)$, so $r_n = n^{k/(2k+1)}$ (forward problem)?
  - ☐ Global rates: nothing yet
  - ☐ Local rates: should be $r_n = n^{k/(2k+1)}$
    progress: Balabdaoui and Wellner (2004)
    local rate is true if a certain conjecture about Hermite interpolation holds

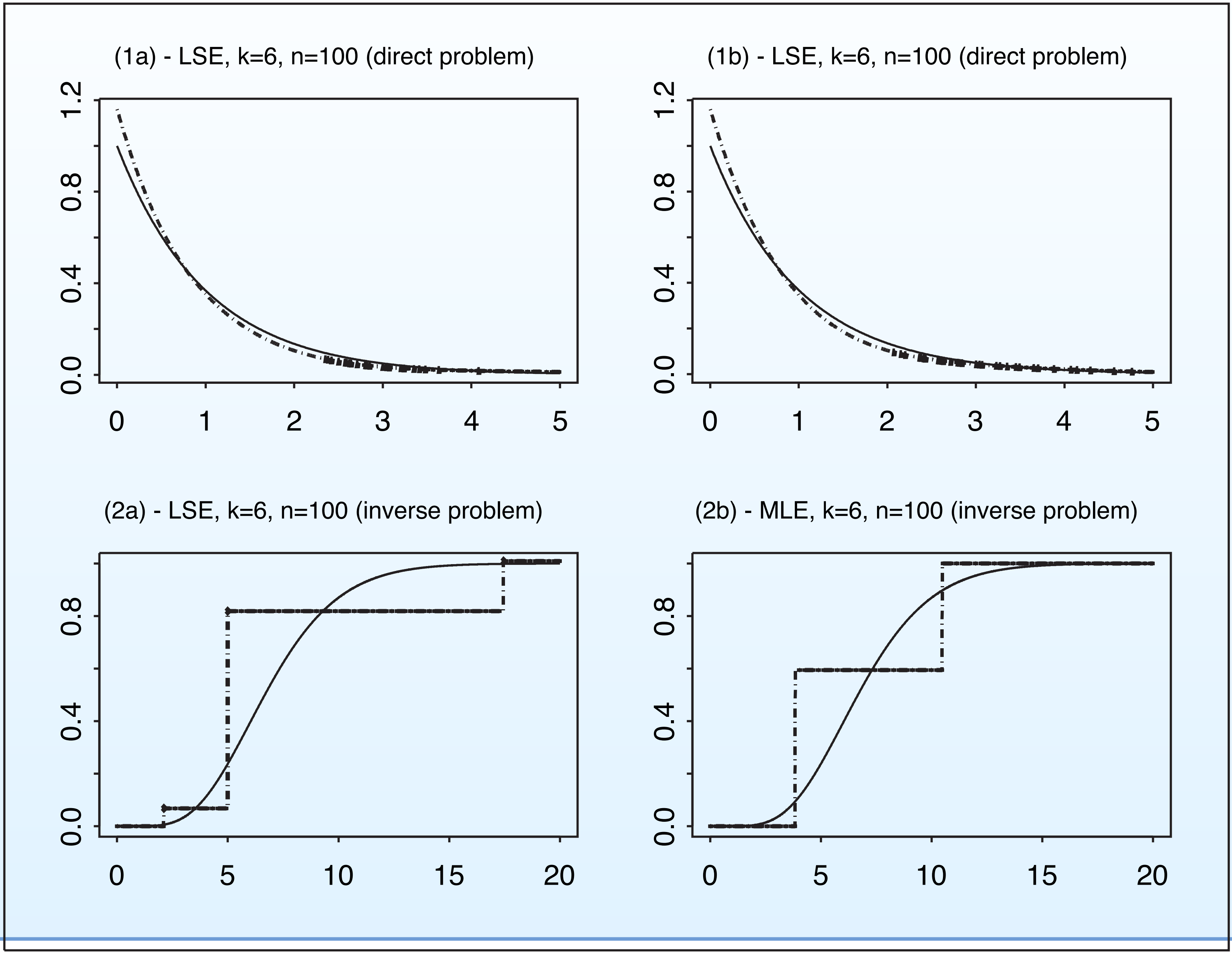
Direct and Inverse estimators $k = 3, n = 100$



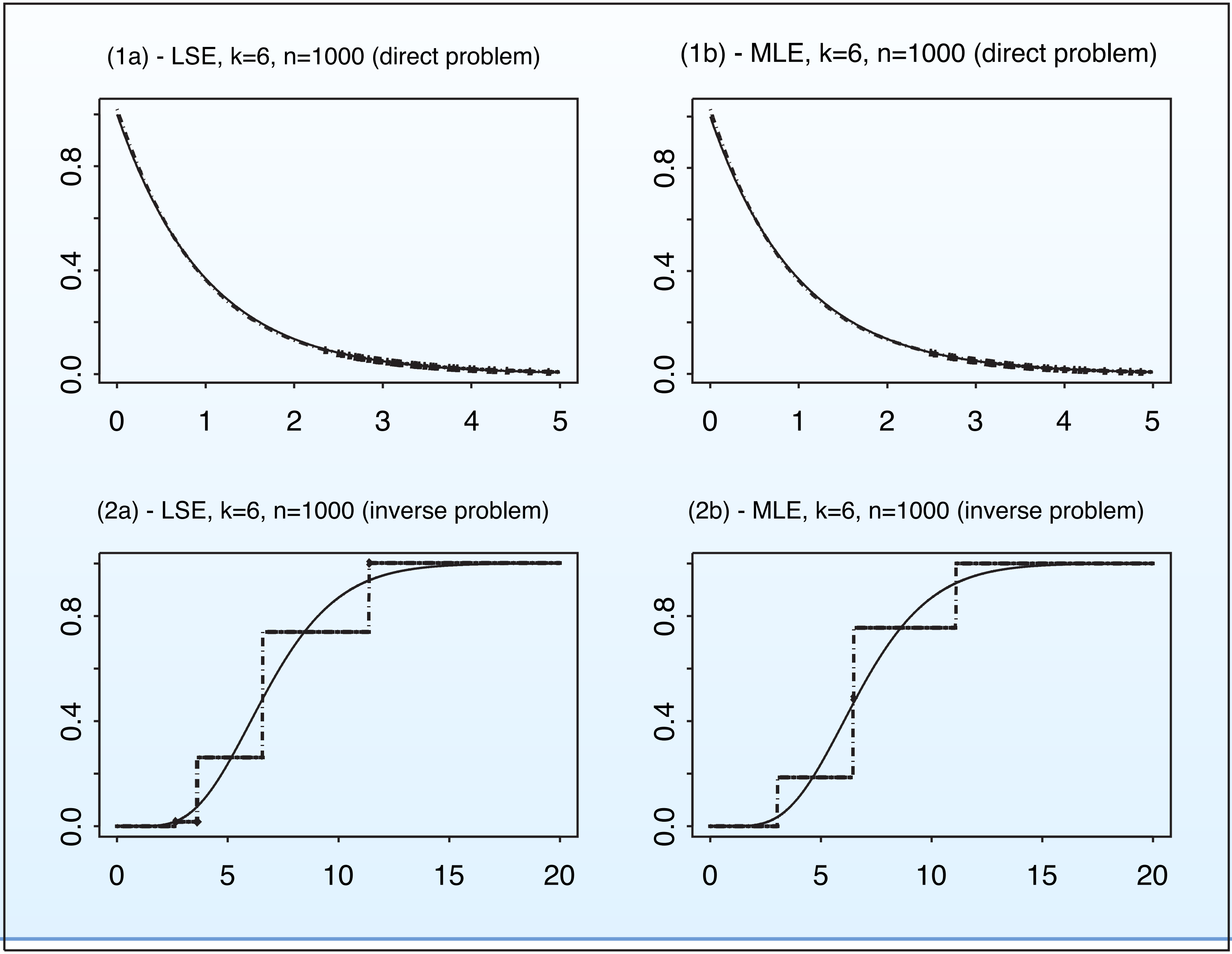
Direct and Inverse estimators $k = 3, n = 1000$



Direct and Inverse estimators $k = 6, n = 100$



Direct and Inverse estimators $k = 6, n = 1000$



- Example 4. (Competing risks with current status data)

- Example 4. (Competing risks with current status data)
  - Variables of interest (X, Y);

    $X =$ failure time; $Y =$ failure cause

    $X \in \mathbb{R}^+$, $Y \in \{1, \ldots, K\}$

    $T =$ an *observation time*, independent of (X, Y)

- Example 4. (Competing risks with current status data)
  - Variables of interest (X, Y);

    $X =$ failure time; $Y =$ failure cause

    $X \in \mathbb{R}^+$, $Y \in \{1, \ldots, K\}$

    $T =$ an *observation time*, independent of (X, Y)
  - Observe: (Δ, T), $\Delta = (\Delta_1, \ldots, \Delta_K, \Delta_{K+1})$ where

$$\Delta_j = 1\{X \le T, Y = j\}, \quad j = 1, \ldots, K$$

$$\Delta_{K+1} = 1\{X > T\}.$$

- Example 4. (Competing risks with current status data)
  - Variables of interest (X, Y);
    $X =$ failure time; $Y =$ failure cause
    $X \in \mathbb{R}^+$, $Y \in \{1, \ldots, K\}$
    $T =$ an *observation time*, independent of (X, Y)
  - Observe: (Δ, T), $\Delta = (\Delta_1, \ldots, \Delta_K, \Delta_{K+1})$ where

$$\Delta_j = 1\{X \le T, Y = j\}, \quad j = 1, \ldots, K$$
$$\Delta_{K+1} = 1\{X > T\}\,.$$

  - Goal: estimate $F_j(t) = P(X \le t, Y = j)$ for $j = 1, \ldots, K$

- Example 4. (Competing risks with current status data)
    - Variables of interest (X, Y);

      $X =$ failure time; $Y =$ failure cause

      $X \in \mathbb{R}^+$, $Y \in \{1, \ldots, K\}$

      $T =$ an *observation time*, independent of (X, Y)
    - Observe: (Δ, T), $\Delta = (\Delta_1, \ldots, \Delta_K, \Delta_{K+1})$ where

$$\Delta_j = 1\{X \le T, Y = j\}, \quad j = 1, \ldots, K$$
$$\Delta_{K+1} = 1\{X > T\}.$$

    - Goal: estimate $F_j(t) = P(X \le t, Y = j)$ for $j = 1, \ldots, K$
    - MLE $\hat{F}_n = (\hat{F}_{n,1}, \ldots, \hat{F}_{n,K})$ exists!

      Characterization of $\hat{F}_n$ involves an *interacting system* of slopes of convex minorants

(competing risks with current status data, continued)

- Global rates. Easy with present methods.

$$n^{1/3}\sum_{k=1}^{K}\int|\hat{F}_{n,k}(t)-F_{0,k}(t)|dG(t)=O_p(1)$$

(competing risks with current status data, continued)

- Global rates. Easy with present methods.

$$n^{1/3}\sum_{k=1}^{K}\int|\hat{F}_{n,k}(t)-F_{0,k}(t)|dG(t)=O_p(1)$$

- Local rates? Conjecture $r_n = n^{1/3}$. New methods needed. Tricky. Maathuis (2006)?

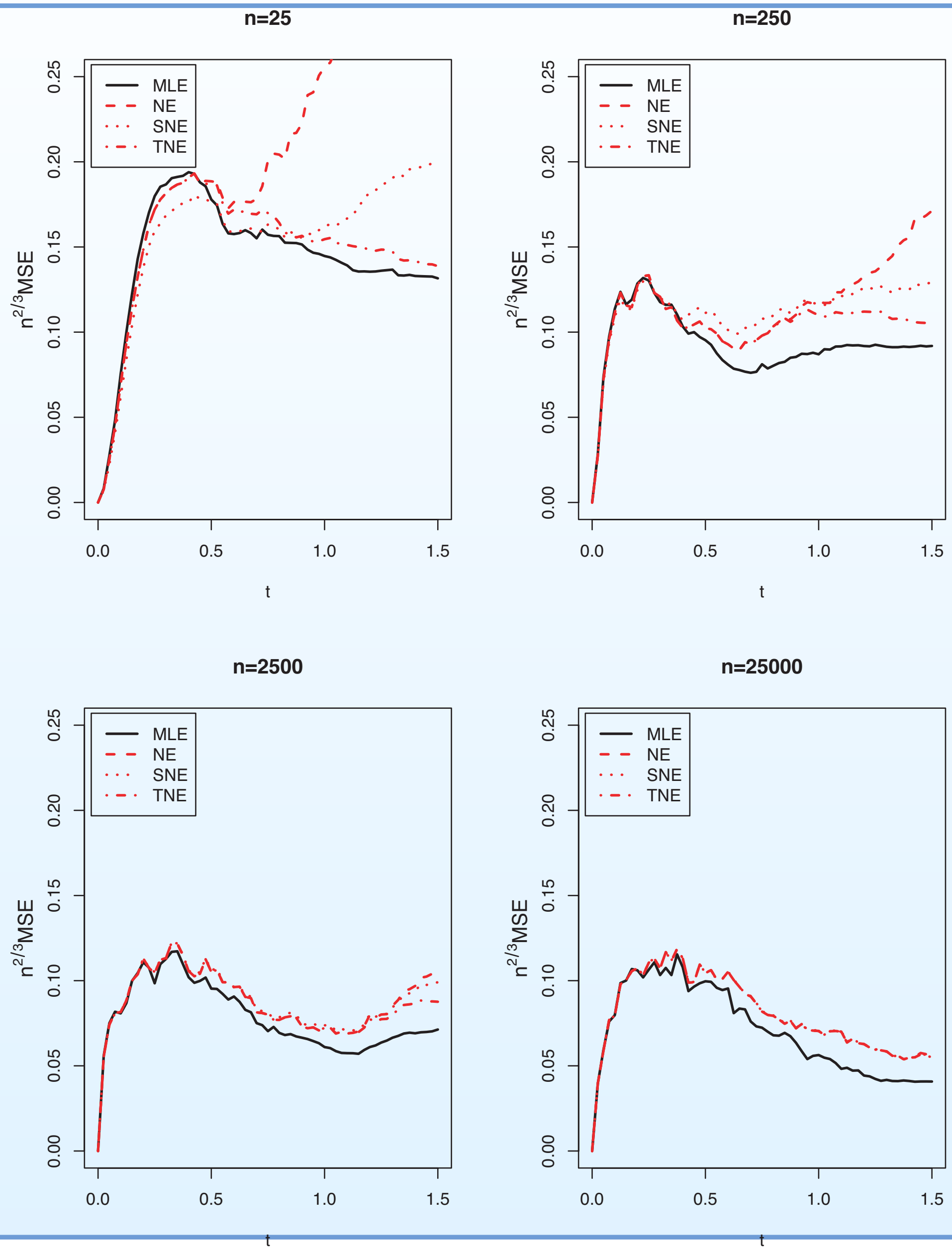
(competing risks with current status data, continued)

- Global rates. Easy with present methods.

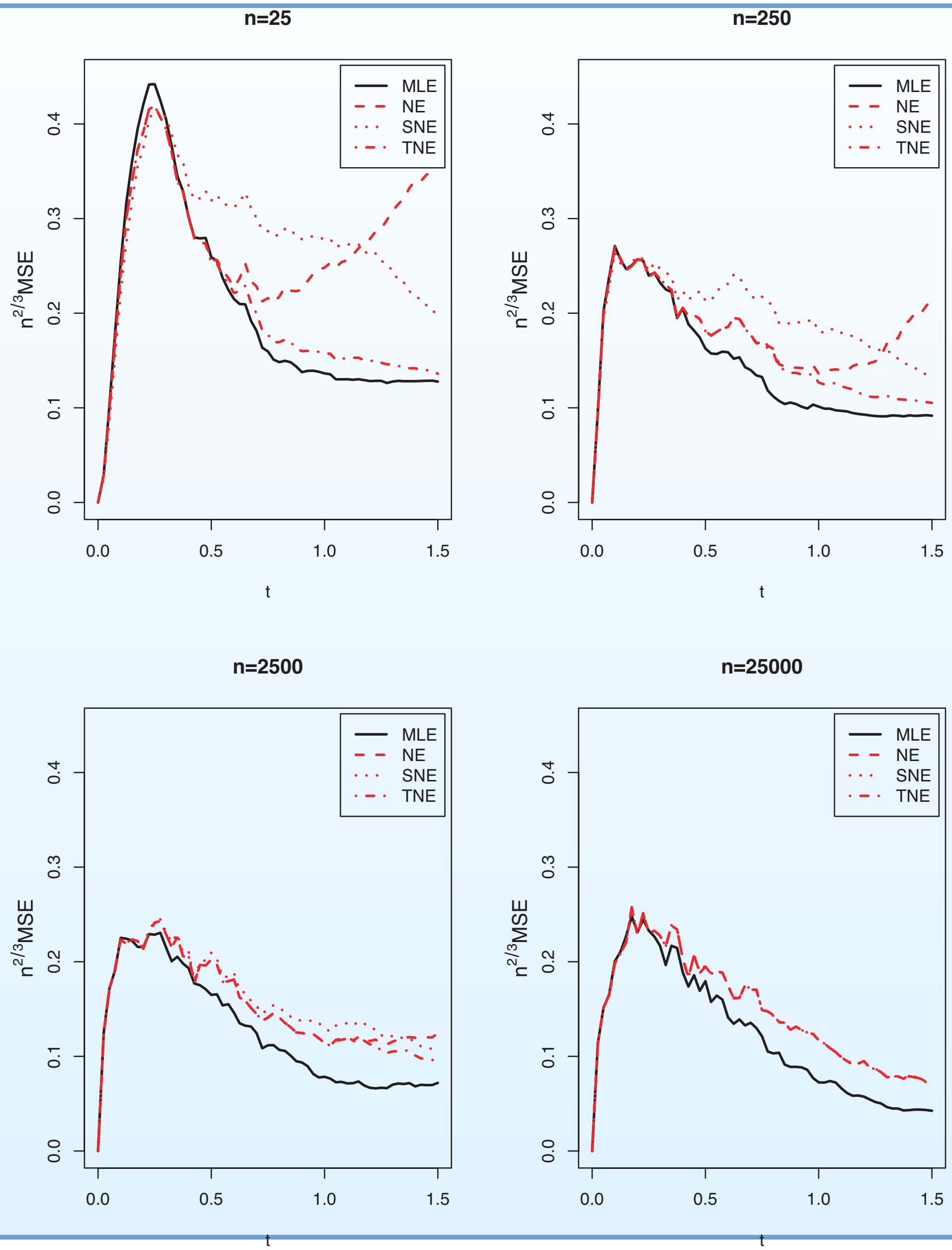
$$n^{1/3}\sum_{k=1}^{K}\int|\hat{F}_{n,k}(t)-F_{0,k}(t)|dG(t)=O_p(1)$$

- Local rates? Conjecture $r_n = n^{1/3}$. New methods needed. Tricky. Maathuis (2006)?
- Limit distribution theory: will involve slopes of an interacting system of greatest convex minorants defined in terms of a vector of dependent two-sided Brownian motions

$n^{2/3} \times$ MSE of MLE and naive estimators of F_1



$n^{2/3} \times$ MSE of MLE and naive estimators of F_2



- **Example 5.** (Monotone densities in $\mathbb{R}^d$)

- Example 5. (Monotone densities in $\mathbb{R}^d$)
  - Entropy bounds? Natural conjectures:
    $\alpha = 1$, d, $\gamma = 1/d$, so $\gamma/(2\gamma+1) = 1/(d+2)$

- **Example 5.** (Monotone densities in $\mathbb{R}^d$)
  - Entropy bounds? Natural conjectures:
    $\alpha = 1$, d, $\gamma = 1/d$, so $\gamma/(2\gamma+1) = 1/(d+2)$
  - Biau and Devroye (2003) show via Assouad's lemma and direct calculations:

$$r_n^{opt} = n^{1/(d+2)}$$

- Example 5. (Monotone densities in $\mathbb{R}^d$)
  - Entropy bounds? Natural conjectures:
    $\alpha = 1$, d, $\gamma = 1/d$, so $\gamma/(2\gamma + 1) = 1/(d + 2)$
  - Biau and Devroye (2003) show via Assouad's lemma and direct calculations:

$$r_n^{opt} = n^{1/(d+2)}$$

  - Rate achieved by the MLE:
    Natural conjecture:

$$r_n^{ach} = n^{1/2d}, \quad d > 2$$

- **Example 5.** (Monotone densities in $\mathbb{R}^d$)
  - Entropy bounds? Natural conjectures:
    $\alpha = 1$, d, $\gamma = 1/d$, so $\gamma/(2\gamma+1) = 1/(d+2)$
  - Biau and Devroye (2003) show via Assouad's lemma and direct calculations:

$$r_n^{opt} = n^{1/(d+2)}$$

  - Rate achieved by the MLE:
    Natural conjecture:

$$r_n^{ach} = n^{1/2d}, \quad d > 2$$

  - Biau and Devroye (2003) construct generalizations of Birgé's (1987) histogram estimators that achieve the optimal rate for all $d \geq 2$.

5. Problems and Challenges

- More tools for local rates and distribution theory?

5. Problems and Challenges

- More tools for local rates and distribution theory?
- Under what additional hypotheses are MLE's globally rate optimal in the case $\gamma > 1/2$?

5. Problems and Challenges

- More tools for local rates and distribution theory?
- Under what additional hypotheses are MLE's globally rate optimal in the case $\gamma > 1/2$?
- More counterexamples to clarify when MLE's do not work?

5. Problems and Challenges

- More tools for local rates and distribution theory?
- Under what additional hypotheses are MLE's globally rate optimal in the case $\gamma > 1/2$?
- More counterexamples to clarify when MLE's do not work?
- What is the limit distribution for interval censoring, case 2? (Does the G&W (1992) conjecture hold?)

5. Problems and Challenges

- More tools for local rates and distribution theory?
- Under what additional hypotheses are MLE's globally rate optimal in the case $\gamma > 1/2$?
- More counterexamples to clarify when MLE's do not work?
- What is the limit distribution for interval censoring, case 2? (Does the G&W (1992) conjecture hold?)
- When the MLE is not rate optimal, is it still preferable from some other perspectives? For example, does the MLE provide efficient estimators of smooth functionals (while alternative rate -optimal estimators fail to have this property)? Compare with Bickel and Ritov (2003).

Problems and challenges, continued

- More rate and optimality theory for Maximum Likelihood Estimators of mixing distributions in mixture models with smooth kernels: e.g. completely monotone densities (scale mixtures of exponential), normal location mixtures (deconvolution problems)

Problems and challenges, continued

- More rate and optimality theory for Maximum Likelihood Estimators of mixing distributions in mixture models with smooth kernels: e.g. completely monotone densities (scale mixtures of exponential), normal location mixtures (deconvolution problems)
- Stable and efficient algorithms for computing MLE's in models where they exist (e.g. mixture models, missing data).

Problems and challenges, continued

- More rate and optimality theory for Maximum Likelihood Estimators of mixing distributions in mixture models with smooth kernels: e.g. completely monotone densities (scale mixtures of exponential), normal location mixtures (deconvolution problems)
- Stable and efficient algorithms for computing MLE's in models where they exist (e.g. mixture models, missing data).
- Results for monotone densities in $\mathbb{R}^d$?

Selected References

- Bahadur, R. R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhya, Ser. A* **20**, 207 - 210.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions.* Wiley, New York.
- Barlow, R. E. and Scheurer, E. M (1971). Estimation from accelerated life tests. *Technometrics* **13**, 145 - 159.
- 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.
- Boyles, R. A., Marshall, A. W., Proschan, F. (1985). Inconsistency of the maximum likelihood estimator of a distribution having increasing failure rate average. *Ann. Statist.* **13**, 413-417.
- Ferguson, T. S. (1982). An inconsistent maximum likelihood estimate. *J. Amer. Statist. Assoc.* **77**, 831–834.
- Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statist. Probab. Lett.* **23**, 165-170.

- Hudgens, M., Maathuis, M., and Gilbert, P. (2005). Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous mark variable. Submitted to *Biometrics*.
- Le Cam, L. (1990). Maximum likelihood: an introduction. *Internat. Statist. Rev.* **58**, 153 - 171.
- Maathuis, M. and Wellner, J. A. (2005). Inconsistency of the MLE for the joint distribution of interval censored survival times and continuous marks. Submitted to *Biometrika*.
- Pan, W. and Chappell, R. (1999). A note on inconsistency of NPMLE of the distribution function from left truncated and case I interval censored data. *Lifetime Data Anal.* **5**, 281-291.