IMS Le Cam Lecture

Maximum Likelihood in modern times: the ugly, the bad, and the good

Jon A. Wellner

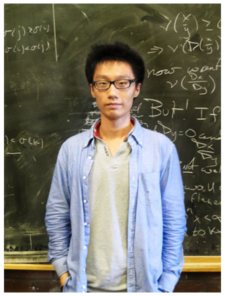
University of Washington, Seattle

Joint Statistical Meetings, Seattle, August 10, 2015

IMS Meeting, Seattle

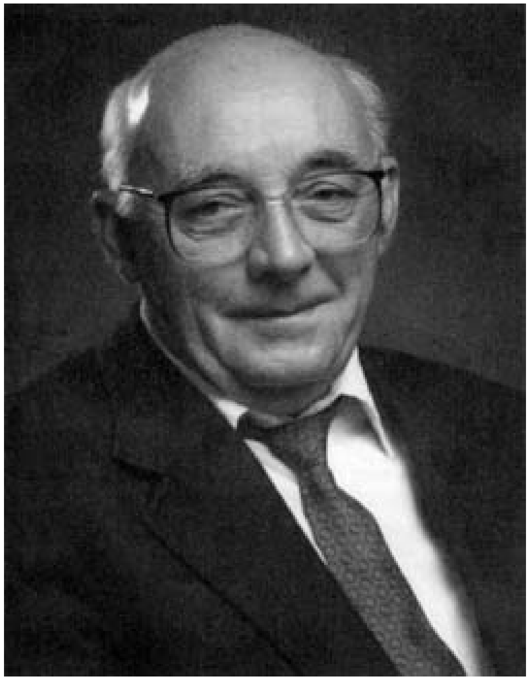
Based (in part) on joint work with:

- **Qiyang (Roy) Han**
- **Charles Doss**
- **Hanna Jankowski**
- **Marloes Maathuis**



- Kaspar Rufibach
- Arseni Seregin

Some photos of Lucien Le Cam:



Lucien Le Cam 1987:

Oberwolfach, 1980, with daughter Linda (Oberwolfach photo collection)

Oberwolfach, 1980, with daughter Linda (from Günther Sawitzki, Heidelberg)

Outline

- 1. Starting points: two papers
- 2: ML in the Modern Age (preliminary examples)
- 3: Existence and Non-existence of MLEs
- 4: Consistency and Inconsistency of MLEs
- 5: Convergence rates: Optimal and Sub-optimal
- 6: Stability and Instability under model misspecification
- 7: Summary and Conclusions

1. Starting points: Two Papers

- **L. Le Cam (1990):** *Maximum likelihood: An Introduction*. Lots can go wrong, even in parametric models:
  - ▷ Contaminated normal mixture model; Kiefer and Wolfowitz (1956): MLE does not exist.
  - ▷ Shifted log-normal; Hill (1963); MLE does not exist; likelihood blows up.
  - ▷ Spiked Gaussians; Kemperman (noted in Le Cam (1970); MLE does not exist; likelihood blows up.
  - ▷ Neyman Scott models, Ferguson and Bahadur examples; MLE exists but is inconsistent.
  - ▷ MLE does not always have minimum risk: Stein's inadmissibility theorem!
  - ▷ Dose - binary response model: Berkson says "minimum chi-square, not maximum likelihood".

- **S. Stigler (2007):** *The Epic Story of Maximum Likelihood*.
  - ▷ History of ML from “well before Fisher” (earliest reference is Lambert (1760)) to “Le Cam’s dissertation” (1953).
  - ▷ Nasty ugly fact due to Joe Hodges (1951): super-efficient estimators, with asymptotically smaller variance than MLE’s exist even for the simplest parametric models, contradicting Fisher’s (1922) claims of asymptotic optimality of MLE’s.
  - ▷ Stigler closes his paper with:

    > “We now understand the limitations of maximum likelihood better than Fisher did, but far from well enough to guarantee safety in its application in complex situations where it is most needed. Maximum Likelihood remains a truly beautiful theory, even though tragedy may lurk around a corner.”

Ugly fact of super efficiency:

- Context: $X_1, \ldots, X_n$ i.i.d. $N(\theta, 1)$.
- MLE of θ is $\widehat{\theta}_n = \overline{X}_n$.
- Hodges estimator T_n: for $0 < a < 1$,

$$T_n = \begin{cases} \overline{X}_n, & \text{if} \quad |\overline{X}_n| \geq n^{-1/4}, \\ (1-a) \cdot 0 + a\overline{X}_n, & \text{if} \quad |\overline{X}_n| < n^{-1/4} \end{cases}$$

- $\sqrt{n}(\overline{X} - \theta) \overset{d}{=} N(0, 1)$ under P_θ for all θ, $n \geq 1$.
-

$$\sqrt{n}(T_n - \theta) \to_d \begin{cases} N(0,1), & \text{if } \theta \neq 0, \\ N(0, a^2), & \text{if } \theta = 0. \end{cases}$$

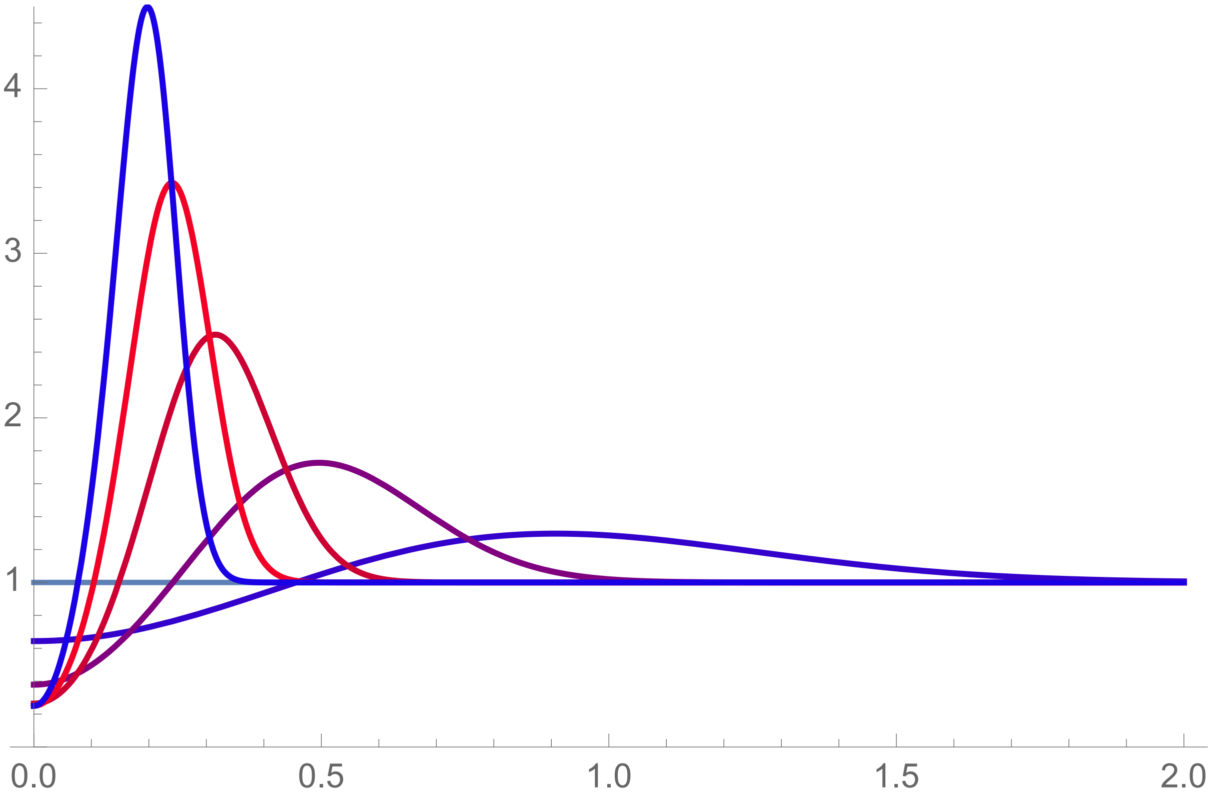
- Under $\theta_n = t/\sqrt{n}$,

$$\sqrt{n}(T_n - \theta_n) \to_d aZ + t(a-1) \sim N(t(a-1), a^2).$$

Risk (MSE) of Hodges estimator:

$$R_n(\theta) = E_\theta\{n(T_n - \theta)^2\}$$

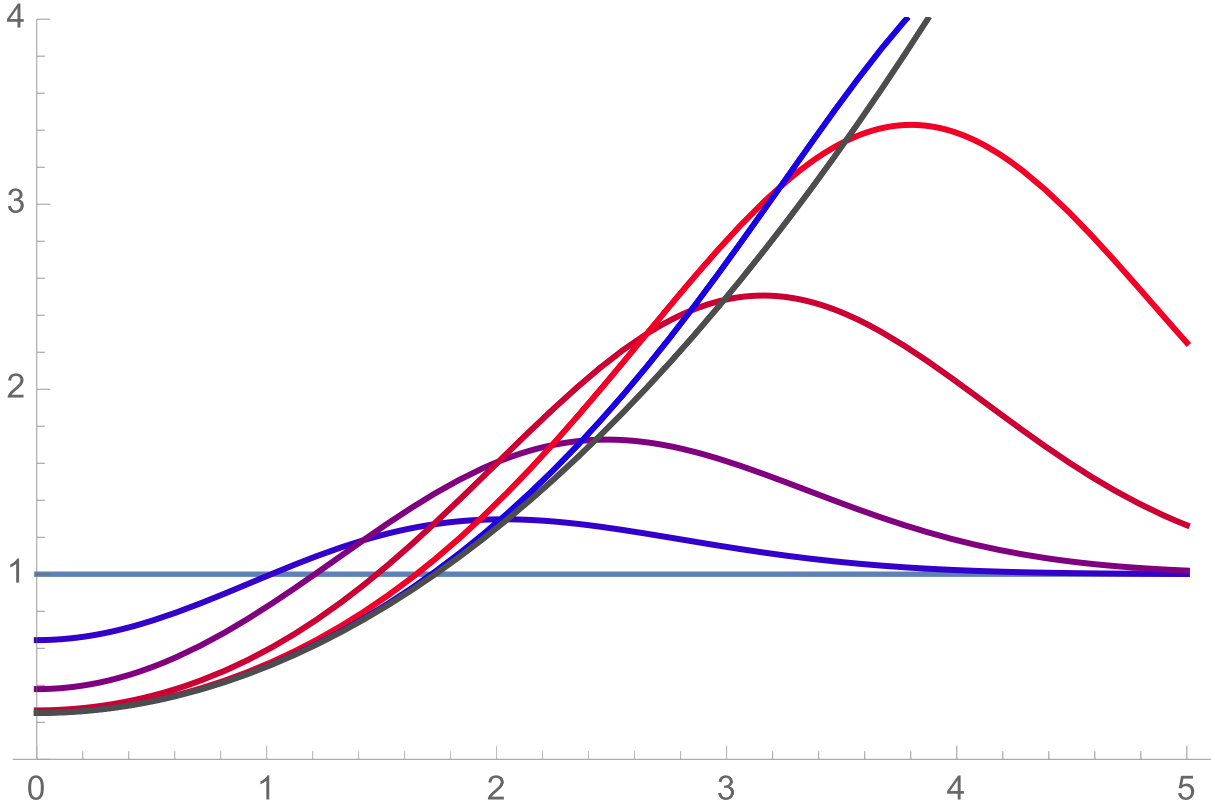
with $a = .5$ and $n \in \{5, 25, 100, 250, 500\}$.



Local Risk of Hodges estimator:

$$R_n^{loc}(t) = R_n(t/\sqrt{n}) = E_{t/\sqrt{n}}\{n(T_n - t/\sqrt{n})^2\}$$

and limiting risk $R_{asymp}(t) = a^2 + (1-a)^2 t^2$ with $a = .5$, $\theta = t/\sqrt{n}$, and $n \in \{5, 25, 100, 250, 500\}$.



2. Maximum Likelihood in the modern age: 1953 - present

- Two (rough) periods:
  - "Completion" of the parametric story: 1953 - 1972.
    - parametric MLE: Le Cam (1970).
    - Hájek - Le Cam convolution and (local) asymptotic minimax theorems (1970 - 1972)
  - Infinite dim. parameter spaces: 1972 - 1993
    - Semiparametric models:
      Cox proportional hazards model (1972); Efron (1977);
      Breslow estimator (1972);
      efficient estimators for the symmetric location model.
    - General rate theory:
      Le Cam (1973, 1975), Birgé (1983).
    - Developing interaction between rate theory and empirical process methods:
      Strassen & Dudley (1969),
      Vapnik and Chervonenkis (1971), Dudley (1978)

- Questions:
  - ▲ What about nonparametric and semiparametric models?
  - ▲ What about models with dimension of the parameter space increasing with sample size?
  - ▲ When are parameter spaces “too large” for MLE’s (or any minimum contrast estimator)?
  - ▲ What properties do we want to require of our procedures?
    - Existence and uniqueness?
    - Consistency?
    - Efficiency or rate efficiency?
    - Stability under model misspecification?
    - Easily (or efficiently) computable?
    - “Objectivity” or “reproducibility”
      (Not too many tuning parameters!)

Good or desirable properties of the MLE for a given model $\mathcal{P}$:

- Existence (or existence for $n \geq N_0$).
- Consistent: $\widehat{\theta}_n \rightarrow_p \theta_0$ as $n \rightarrow \infty$.
- Efficient (finite-dimensional parameters), or ...
- Rate efficient (infinite-dimensional parameters): MLE converges at the "optimal" (global) rate.
- Stable under model misspecification: if $P_0 \notin \mathcal{P}$, then

$$\widehat{\theta}_n \rightarrow_p \text{argmin}_{\theta \in \Theta} d(P_0, P_\theta)$$

  for some metric or divergence on probability measures, e.g. $d(P, Q) = K(P, Q)$, the Kullback-Leibler divergence.

- When these desirable properties hold, then we describe the pair (MLE, $\mathcal{P}$) or the situation as "good".

- When these desirable properties hold, then we describe the pair (MLE, $\mathcal{P}$), or the situation as "good".
- When these various properties fail, then we describe them as follows:
  - If the MLE $\widehat{\theta}_n$ does not exist, then the situation is "bad": i.e. the pair (MLE, $\mathcal{P}$) is bad.
  - If the MLE $\widehat{\theta}_n$ exists, but is inconsistent, then the situation is "ugly"; i.e. the pair (MLE, $\mathcal{P}$) is ugly.
  - If the MLE $\widehat{\theta}_n$ is consistent but rate inefficient, then the situation is "bad".
  - If the MLE $\widehat{\theta}_n$ is consistent but unstable under model misspecification, then the situation, or the pair (MLE, $\mathcal{P}$) is "bad".

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 $p_{\theta_0} \in \mathcal{P} = \{p_\theta : \ \theta \in \Theta\}$.

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 $\widehat{\theta}_n \in \Theta$ satisfying

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

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

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

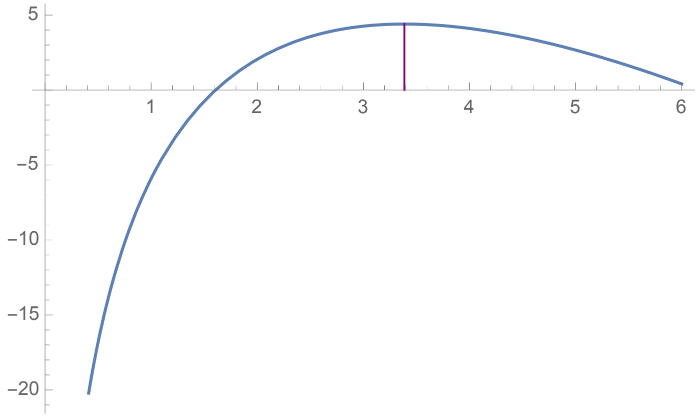
Example 1. (A “regular” parametric model) Exponential (θ). If $X_1, \ldots, X_n$ are i.i.d. p_{θ_0} where $p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x)$. Then

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i) = \theta^n \exp(-\theta \sum_1^n X_i),$$

so

$$n^{-1} \log L_n(\theta) = \mathbb{P}_n \log p_\theta(X) = \log(\theta) - \theta \overline{X}_n,$$

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



Example 2. (A "nonparametric" model) Monotone decreasing densities on $[0, \infty)$. Suppose $X_1, \ldots, X_n$ are i.i.d. $p_0 \in \mathcal{P}$ where

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

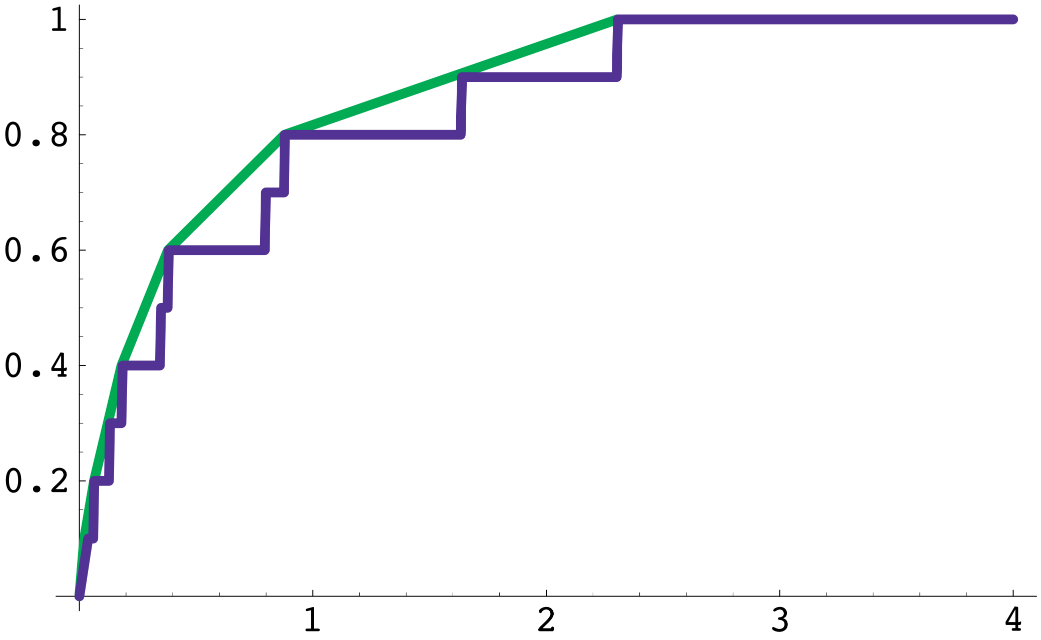
This is a nonparametric model defined by a shape constraint. Then

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

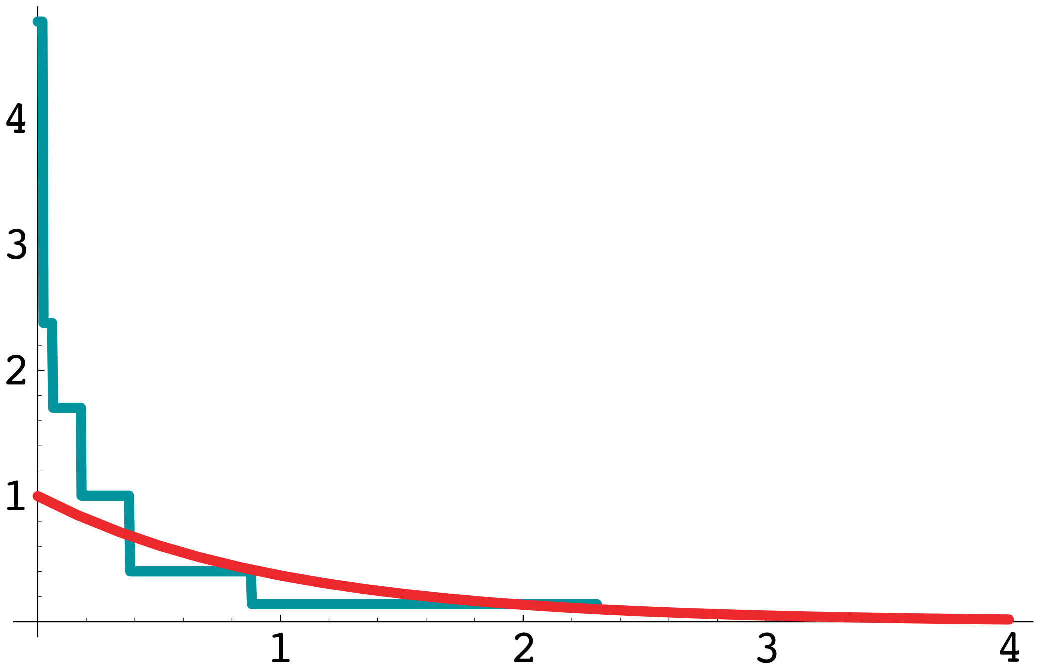
is maximized by the Grenander estimator:

$$\begin{aligned} \widehat{p}_n(x) \ &= \ \text{left derivative at } x \text{ of the} \\ & \qquad \text{Least Concave Majorant} \\ & \qquad \mathbb{C}_n \ \text{ of } \ \mathbb{F}_n \end{aligned}$$

where $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^{n} 1\{X_i \le x\}$. This is due to Grenander (1956). Consistency, limiting distributions at points, and rates of convergence due to Prakasa Rao, Groeneboom (1985), Birgé (1987,1989), and van de Geer (1993).



$\mathbb{F}_n$ (dark blue), empirical distribution function, $n = 10$
$\mathbb{C}_n$ (green), least concave majorant of $\mathbb{F}_n$



MLE $\widehat{p}_n$, the Grenander estimator (green);
truth $p_0(x) = e^{-x}$ (red)
No tuning parameter!

Example 3. (A semiparametric model) $X_1, \ldots, X_n$ are i.i.d. $p_0 \in \mathcal{P}$ where

$$\mathcal{P} = \{p_{\theta,G} : \ \theta > 0, \ G \text{ a distribution function on } \mathbb{R}^+\} \ \text{ with}$$

$$p_{\theta,G}(x,y) = \int_0^\infty v^2\theta \cdot \exp(-v(x+\theta y))dG(v)$$

This paired exponential, or frailty, mixture model is contained in the class of semiparametric mixture models considered by Kiefer and Wolfowitz (1956).

- The MLE $(\widehat{\theta}, \widehat{G})$ of (θ, G) exists and is unique: Lindsay (1980, 1983a,b, 1995).
- The MLE is consistent (under compactness and envelope conditions): Wald (1949); Kiefer and Wolfowitz (1956).
- The MLE of θ is asymptotically efficient (in certain cases): van der Vaart (1996). In particular, efficiency holds for Example 3.

- In general the (K-W, 1956) class of semiparametric mixture models is a **success story** for the (semiparametric) MLE!

  This class of models gives a way around the inconsistency examples of Neyman and Scott (1948).

  Hence the pair (MLE, $\mathcal{P}$) is GOOD, at least at the level of consistency.
- But . . . asymptotic efficiency is known only for a few cases!

  van der Vaart (1996)! (More theory needed.)
- Lack of general computational implementations and algorithms has slowed and impeded progress.

  Recent work by Koenker and Mizera (2013, 2014).

Setting 2: non-dominated families (a slight detour) 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 empirical likelihood of $X_1, \ldots, X_n$ is defined to be

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

Then a (Nonparametric) 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

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

if it exists. A more sophisticated version of this definition is given by Kiefer and Wolfowitz (1956); see also Barlow (1968) and Scholz (1980).

Example 4. (A completely nonparametric model) If $\mathcal{P}$ =all probability measures on $(\mathcal{X}, \mathcal{A})$, then the NonParametric MLE is

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

where $\delta_x(A) = 1_A(x)$, so $\mathbb{P}_n(A) = n^{-1}\#\{i \le n : \ X_i \in A\}$.

Further examples where this approach "works":

- Right censored survival data: the nonparametric MLE is the Kaplan-Meier estimator.
- The Cox (1972) proportional hazards model: partial likelihood = profile likelihood, and the maximum profile likelihood estimator is asymptotically efficient: Efron (1977), Begun, Hall, Huang, and W (1983), Bickel, Klaassen, Ritov, and W (1993).
- Significant theory via Gill (1989), Gill and van der Vaart (1993), and van der Vaart (1995)

3. Existence and Nonexistence of MLE's

Existence and uniqueness:

- Cramér (1946)
- Mäkeläinen, Schmidt, Styan (1981).
- Lindsay (1983a, 1983b, 1995), Lindsay and Roeder (1993)
- log-concave densities: Dümbgen and Rufibach (2009); Cule, Samworth, and Stewart (2010)

Nonexistence or non-uniqueness: examples

- Kiefer and Wolfowitz (1956)
- Kraft and Le Cam (1956)
- Hill (1963)
- Barnett (1966)
- Reeds (1985): (multiple roots of Cauchy likelihood equations)
- Drton and Richardson (2004); Drton (2006)
- Unimodal densities, unknown mode: MLE does not exist

  Thus (MLE, $\mathcal{P}_{unimodal}$) is **bad**.

  Wegman (1968, 1969, 1970a,b), Reiss (1973, 1976), Bickel and Fan (1996), Birgé (1997)
- $s-$concave densities? Existence OK, but uniqueness not clear.

4. Consistency and Inconsistency of MLE's

Consistency:

- Wald (1949)
- Kiefer and Wolfowitz (1956)
- Huber (1967); Pollard (1985, 1989)
- Perlman (1972)
- Reiss (1973, 1978); Pfanzagl (1988)
- Wang (1985)
- van de Geer (1993)

Inconsistency: counterexamples

- Neyman and Scott (1948)
- Bahadur (1958), Ferguson (1982)
- Le Cam (1975), (1990)
- Barlow, Bartholomew, Bremner, and Brunk (1972)
- Boyles, Marshall, and Proschan (1985)
- Pan and Chappell (1999)
- Maathuis and Wellner (2008)

Inconsistency Example 1

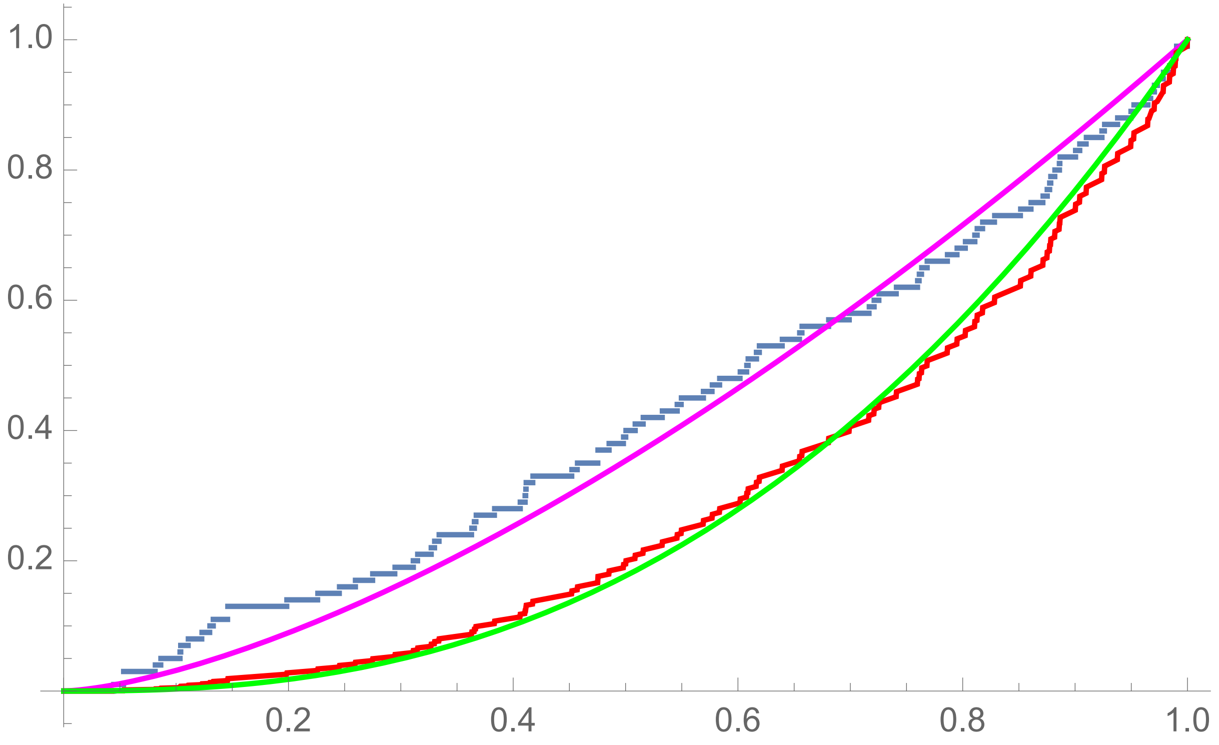
- A distribution function F on $[0, b)$ is star-shaped if $F(x)/x$ is non-decreasing on $[0, b)$.

If $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{star}$ then Barlow, Bartholomew, Bremner, and Brunk (1972) show that the (nonparametric) MLE $\widehat{F}_n$ of $F \in \mathcal{F}_{star}$ is

$$\widehat{F}_n(x) = \left\{ \mathbb{F}_n(x) \cdot \frac{x}{X_{(n)}} \ \wedge \ 1 \right\}$$

where $\mathbb{F}_n$ is the empirical distribution function of the sample and $X_{(n)} = \max_{i \le n} X_i$. Thus

$$\widehat{F}_n(x) \rightarrow_{a.s.} F(x) \cdot \frac{x}{F^{-1}(1)} \ \neq F(x).$$



- True F: $F(x) = x^{3/2}$ (magenta); Empirical $\mathbb{F}_n$ (blue); $n = 100$
- $F_{asymp}(x) = xF(x) = x^{5/2}$ (green); MLE $\widehat{F}_n$ (red).
- Thus the pair $(MLE, \mathcal{P}_{starshaped})$ is ugly .
- Repairs and alternatives:

  Jongbloed (2009); Groeneboom and Jongbloed (2014).

Inconsistency Example 2

- $\Lambda(x) \equiv -\log(1-F(x)) = \int_0^x (1-F(y))^{-1} dF(y)$
  $= \text{ cumulative hazard function of } F.$
- A distribution function F on $[0,b)$ has

Increasing Failure Rate Average if $\lambda(x) \equiv x^{-1}\Lambda(x)$ is non-decreasing. If $X_1, \ldots, X_n$ are i.i.d. $F \in \mathcal{F}_{IFRA}$ then Boyles, Marshall, and Proschan (1985) show that the (nonparametric) MLE $\widehat{F}_n$ of $F \in \mathcal{F}_{IFRA}$ is given by

$$\widehat{\lambda}_n(x) = \begin{cases} \widehat{\lambda}_j, & X_{(j)} \le x < X_{(j+1)}, \ j = 0, \ldots, n-1, \\ \infty, & x \ge X_{(n)}, \end{cases}$$

where

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

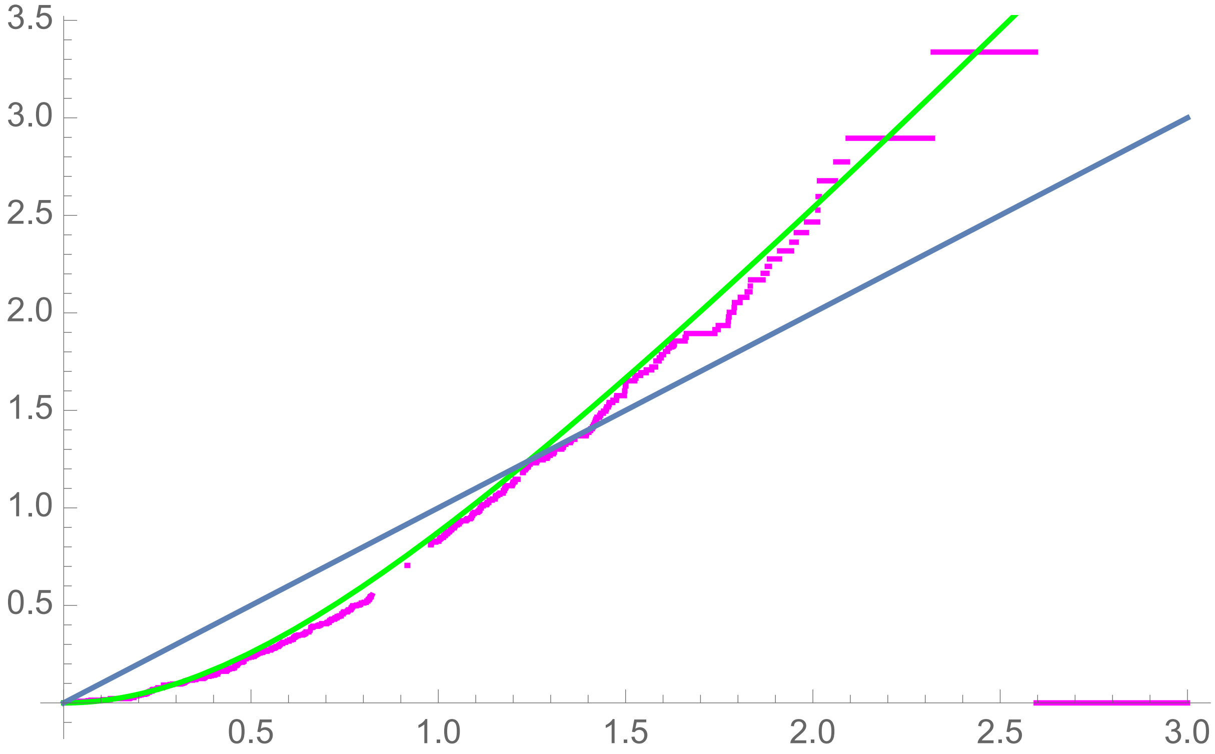
Furthermore, BMP (1985) show that ...

$$\begin{aligned}\widehat{\lambda}_n(x) \ \to_{a.s.} \ & \int_0^x \left(\int_y^\infty z dF(z) \right)^{-1} dF(y) \\ \neq \quad & x^{-1} \int_0^x \left(\int_y^\infty dF(z) \right)^{-1} dF(y) = x^{-1}\Lambda(x).\end{aligned}$$

For example, if $1 - F(x) = e^{-x}$ (so $X_j \sim$ exp(1)), then

$$\begin{aligned} x^{-1}\widehat{\Lambda}_n(x) \to_{a.s.} \log(1+x) \neq 1, \quad \text{or} \\ 1 - \widehat{F}_n(x) \to_{a.s.} (1+x)^{-x} \neq \exp(-x). \end{aligned}$$

This situation is just plain ugly!



True λ: $1 - F(x) = \exp(-x^2)$; $\lambda(x) = x$, (blue)
$\widehat{\lambda}_{asymp}(x)$ (green);
MLE $\widehat{\lambda}_n$ (magenta)

- Alternatives and repairs:

  Rojo and Samaniego (1994); ?? .

Inconsistency Example 3 (Maathuis & W, 2008)

- $X =$ a survival time
  (time of individual becoming HIV positive.)
- $Y =$ a real-valued "mark variable"
  (measure of genetic distance between infecting HIV virus & virus in vaccine.)
- $(X, Y) \sim F$ on $(0, \infty) \times \mathbb{R}$.
- $\mathbf{T} = (T_1, \ldots, T_k) =$ a vector of "observation times", $0 < T_1 < T_2 < \cdots < T_k$, independent of (X, Y).
- Observations (per individual): $W = (\mathbf{T}, \mathbf{\Delta}, Z)$ where
  $\Delta_j = 1\{T_{j-1} < X \le T_j\}$, $j = 1, \ldots, k+1$,
  with $T_0 \equiv 0$, $T_{k+1} \equiv \infty$
  $Z = \Delta_+ Y$, $\Delta_+ = \sum_{j=1}^{k} \Delta_j$.
- Estimate F_0, the joint distribution of (X, Y) based on $W_1, \ldots, W_n$ i.i.d. as W.

Inconsistency Example 3, cont'd:

- Maathuis and W (2008) show that $\widehat{F}_n$ exists and is (essentially) unique.
- $\widehat{F}_n(x, y) \to_{a.s.} F_\infty(x, y) \neq F(x, y)$. **ugly!**
- Alternatives and repairs: (via sieves or smoothing)
  Hudgens, Maathuis, and Gilbert (2007);
  Groeneboom, Jongbloed and Witte (2010, 2012a,b)

More Inconsistency Examples:

- Left truncated and interval censored data:
  Pan and Chappell (1999).
- Bivariate right censoring: Tsai, Leurgans, and Crowley (1986);
  van de Laan (1996);
  repair by Prentice (2014)?!
- . . .

5. Convergence Rates:

Optimal and Sub-optimal

- Le Cam (1973, 1975)
- Birgé (1983, 1986)
- Birgé and Massart (1993)

Covering numbers & bracketing numbers: (T, ρ) a metric space
Kolmogorov and Tikhomirov (1959), ...

$$N(\epsilon, T, \rho) = \min\{k \in \mathbb{N} : \ T \subset \cup_{j=1}^{k} B(t_j, \epsilon)\}$$

$=$ minimum number of $\rho -$ balls of radius ϵ needed to cover T.

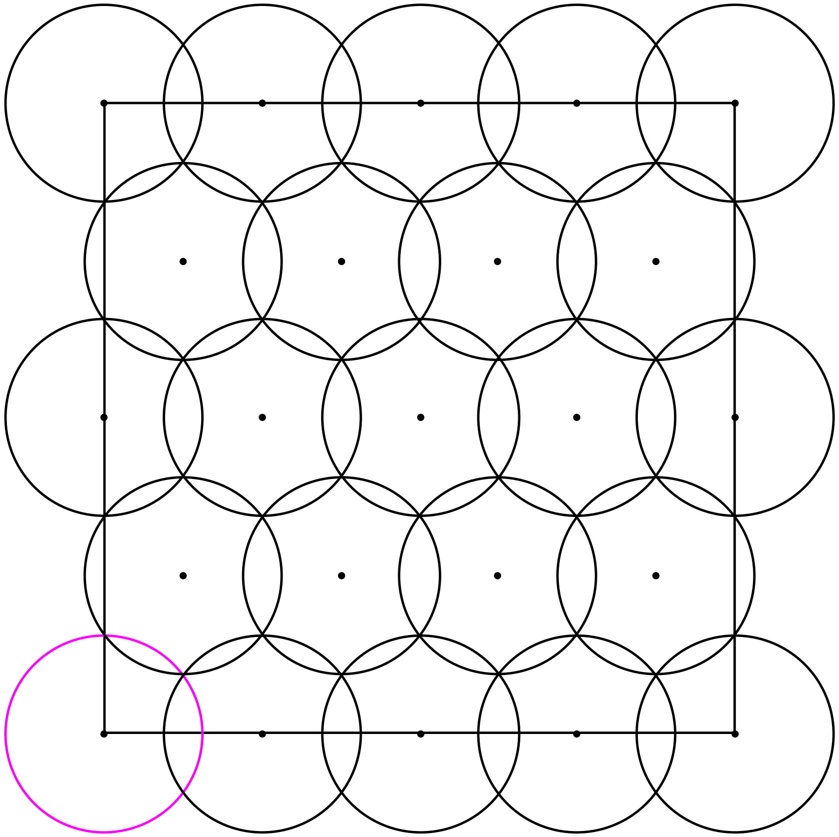
Proposition. For $B(0, R) \subset \mathbb{R}^d$, $\rho = \| \cdot \| =$ Euclidean distance,

$$N(\epsilon, B(0, R), \| \cdot \|) \le \left(\frac{6R}{\epsilon}\right)^d$$

Theorem. (Kolmogorov 1955) If $\mathcal{X}$ is a bounded convex subset of $\mathbb{R}^d$ and $C_1^\alpha(\mathcal{X})$ is the subset of the collection of all $\alpha-$smooth functions $C^\alpha(\mathcal{X})$ on $\mathcal{X}$ with $\|f\|_\alpha \le 1$, then

$$\log N(\epsilon, C_1^\alpha(\mathcal{X}), \| \cdot \|_\infty) \le \frac{K}{\epsilon^{d/\alpha}}$$

for every $\epsilon > 0$ where $K = K_{\alpha, d, \mathcal{X}}$.



An $\epsilon = 5/32$ covering of the unit square, $T = [0,1]^2$;
$N(\epsilon, T, \|\cdot\|) = 23$.

From Le Cam (1973,1975) and Birgé (1983):
Minimax optimal rate of convergence $\delta_n = \delta_n^{opt}$ is determined by

$$n\delta_n^2 = \log N_{[\,]}(\delta_n, \mathcal{P}, d). \tag{1}$$

If

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

then (1) leads to the optimal rate of convergence

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

Thus

$$d(T_n, \theta) = O_p(\delta_n^{opt}) \qquad \text{or} \qquad E_\theta d(T_n, \theta) = O(\delta_n^{opt});$$

On the other hand, the bounds from Birgé and Massart (1993) yield achieved rates of convergence for MLE's (and other minimum contrast estimators) $\delta_n = \delta_n^{ach}$ determined by

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

so that

$$d(\widehat{\theta}_n, \theta) = O_p(\delta_n^{ach}) \qquad \text{or} \qquad E_\theta d(\widehat{\theta}_n, \theta) = O(\delta_n^{ach}).$$

Note that since $N_{[\,]}(\epsilon, \mathcal{P}, d)$ is a decreasing function of ϵ,

$$\int_{c\delta_n^2}^{\delta_n} \sqrt{\log N_{[\,]}(\epsilon, \mathcal{P}, d)} d\epsilon \ge \delta_n \cdot \sqrt{\log N_{[\,]}(\delta_n, \mathcal{P}, d)},$$

where

$$\sqrt{n}\delta_n^2 = \delta_n \sqrt{\log N_{[\,]}(\delta_n, \mathcal{P}, d)}$$

leads to (1).

... and if (2) holds, this leads to the rate of convergence

$$\delta_n^{ach} = \left\{ \begin{array}{ll} n^{-\gamma/(2\gamma+1)}, & \text{if } \ \gamma > 1/2, \\ n^{-\gamma/2}, & \text{if } \ \gamma < 1/2, \end{array} \right\} \neq \delta_n^{opt}$$

$$\text{if} \quad \left\{ \begin{array}{ll} 1/\gamma < 2, & \text{the ``Donsker'' domain,} \\ 1/\gamma > 2, & \text{the ``trans-Donsker'' domain} \end{array} \right\}.$$

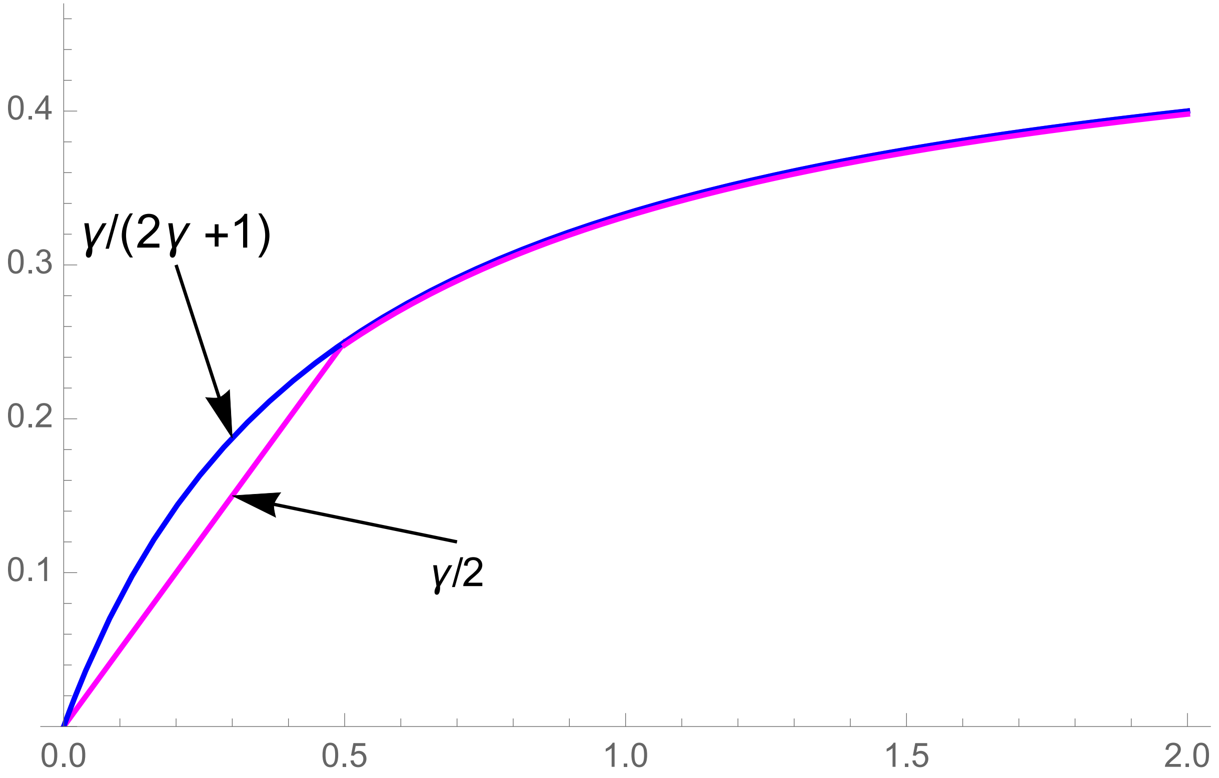
See van der Vaart & W (1996), sections 3.2 and 3.4; van de Geer (2000), chapter 7. Thus there is the possibility that Maximum Likelihood is not (rate-)optimal when $\gamma < 1/2$. Thus when

$$\gamma^{-1} = \frac{d}{\alpha},$$

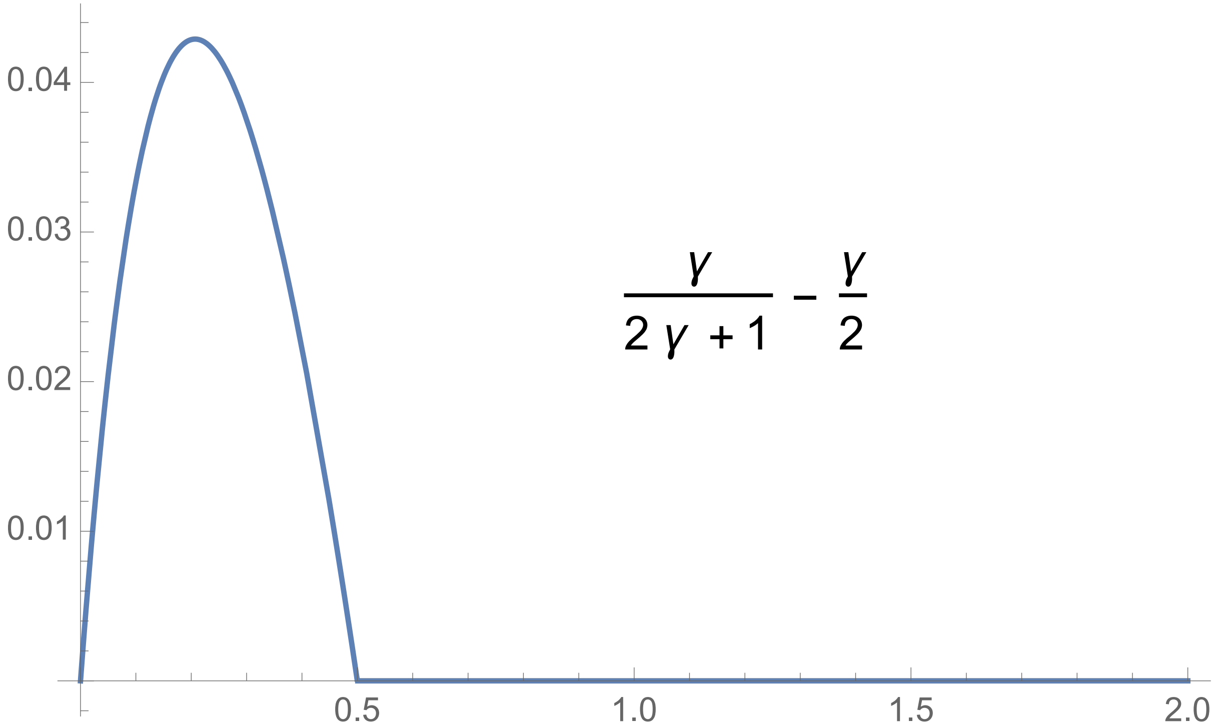
$$\left\{ \begin{array}{llll} \gamma^{-1} < 2 & \text{if } \ \alpha > d/2, & \text{MLE is rate optimal,} & \text{(good!),} \\ \gamma^{-1} > 2 & \text{if } \ \alpha < d/2, & \text{MLE is rate sub-optimal,} & \text{(bad!)} \end{array} \right\}.$$

van de Geer (2000), page 122:
“Because there do exist other estimators with better convergence rates (see ...), one should not use the maximum likelihood estimator when the entropy integral diverges.”

- Many interesting models with $\gamma^{-1} = d/\alpha < 2$:
  - ▷ Monotone functions on $\mathbb{R}$: $\alpha = 1$, $d = 1$, so $\gamma^{-1} = 1$.
  - ▷ Convex functions on $\mathbb{R}$ or $\mathbb{R}^2$: $\alpha = 2$, $d \in \{1, 2, 3\}$, so $\gamma^{-1} \in \{1/2, 1, 3/2\}$.
- Some interesting models with smoothness increasing “naturally” with dimension:
  - ▷ Functions of bounded variation on $\mathbb{R}^d$.
    $\alpha = d$, $\gamma = d$, so $\gamma^{-1} = d/d = 1$ (but $(\log n)^\tau$ factors appear! Gao and W, 2013)
- Many interesting models with $\gamma^{-1} = d/\alpha > 2$:
  - ▷ Coordinatewise monotone functions on $(\mathbb{R}^+)^d$:
    $\alpha = 1$, $d \geq 3$, so $\gamma^{-1} = d/\alpha > 2$
  - ▷ Convex functions on $\mathbb{R}^d$: $\alpha = 2$, $d > 4$, so $\gamma^{-1} = d/\alpha > 2$.



Rate exponents: optimal (blue); achieved (magenta)



Difference of rate exponents: $\frac{\gamma}{2\gamma+1} - \frac{\gamma}{2} = \frac{\gamma(1-2\gamma)}{2(2\gamma+1)}$,
maximum difference $= (3/4) - 1/\sqrt{2} \doteq .043\ldots$,
achieved at $\gamma = (\sqrt{2} - 1)/2 \doteq .207\ldots$.

- **Example:** Birgé (unpublished?), S. Chatterjee (2014)

Suppose we observe $\underline{Y} = \underline{\theta} + \underline{Z}$ with $\underline{Z} \sim N_{n+1}(\underline{0}, I)$, where $\underline{\theta} = (\theta_0, \underline{\theta}') \in \Theta$ where

$$\Theta = \{\underline{\theta} \in \mathbb{R}^{n+1} : \ |\theta_0| \le n^{1/4}, \ \|\underline{\theta}'\| \le 2(1 - n^{-1/4}|\theta_0|)\}.$$

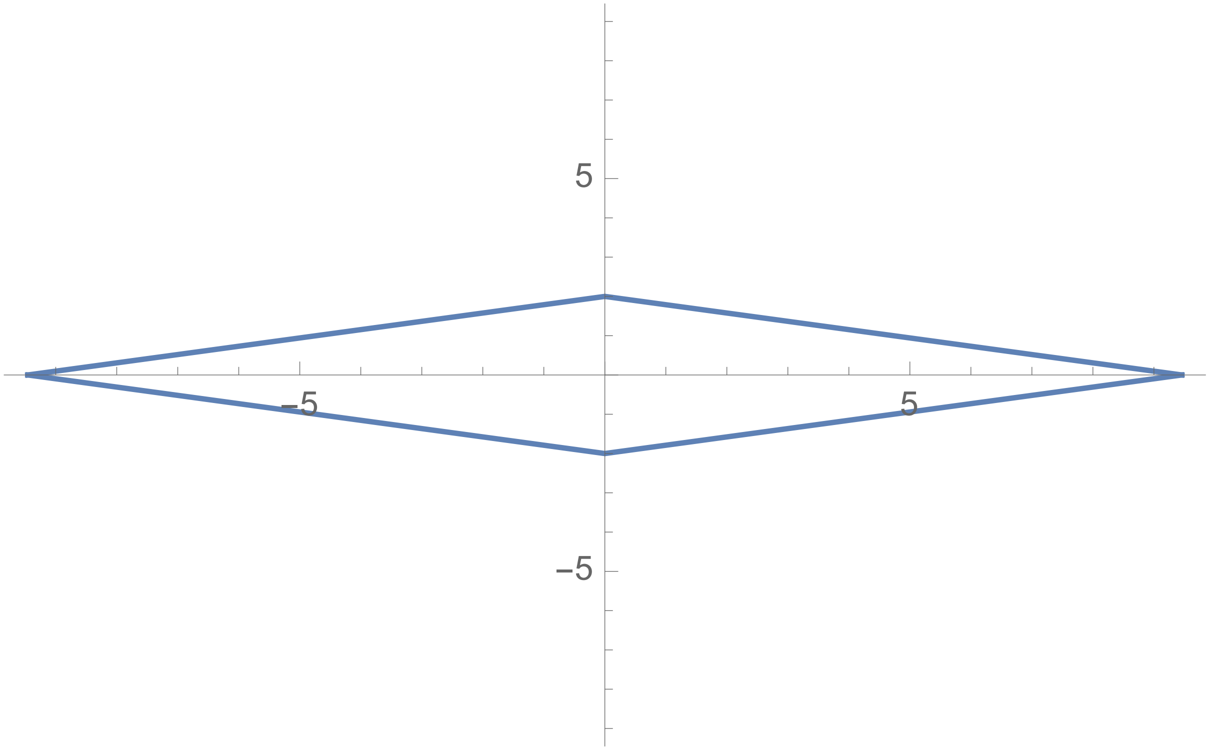
Then for $n \ge 128$, $\widehat{\underline{\theta}}_n =$ LSE $=$ MLE $= \ \prod(\underline{Y}|\Theta)$,

$$\sup_{\theta\in\Theta} E_\theta \|\widehat{\underline{\theta}}_n - \underline{\theta}\|^2 \ge (3/4)\sqrt{n} + 3.$$

On the other hand, if $\widetilde{\underline{\theta}}_n \equiv (Y_0, 0, \ldots, 0)$, then

$$\sup_{\theta\in\Theta} E_\theta \|\widetilde{\underline{\theta}}_n - \underline{\theta}\|^2 \le 5.$$

Hence the MLE is rate suboptimal for this model.



A section of the set Θ for Birgé's example, $n = 8000$.

6. Stability or instability under Model misspecification

- Stability or Instablility

Le Cam (1990), page 168, writes:
"...if possible the method or procedure, or optimality principle used to select the estimation procedure should preferably have some sort of stability ...".

Examples from shape constrained estimation:

- ML estimation of a decreasing density on $\mathbb{R}^+$: (Good; Patilea 2001, Jankowski 2014)
- ML estimation of a log-concave density on $\mathbb{R}^d$. (Good; Cule-Samworth 2010)
- ML estimation of $s-$ concave densities on $\mathbb{R}^d$. (Bad, Seregin & W 2010)
- Rényi - divergence estimation of an $s-$ concave density on $\mathbb{R}^d$. (Good; Han & W, 2015)

6. Outline

- A: The Grenander estimator off the model.
- B: Log-concave and $s-$concave densities on $\mathbb{R}$ and $\mathbb{R}^d$
- C: $s-$concave densities on $\mathbb{R}$ and $\mathbb{R}^d$
- D: Maximum Likelihood for log-concave and $s-$concave densities
  - ▷ 1: Basics
  - ▷ 2: On the model
  - ▷ 3: Off the model
- E. An alternative to ML: Rényi divergence estimators
  - ▷ 1. Basics
  - ▷ 2. On the model
  - ▷ 3. Off the model

6A. The Grenander estimator off the model

- Suppose that $X_1, \ldots, X_n$ are i.i.d. Q where the distribution function $F_Q(x) = Q(-\infty, x]$ is not concave, and hence $Q \notin \mathcal{P}_{mon}$.
- Under very mild conditions, it has been shown by Patilea (2001) that the Grenander estimator, i.e. MLE $\widehat{p}_n$ for $\mathcal{P}_{mon}$ satisfies:

$$\int_{\mathbb{R}+} |\widehat{p}_n(x) - p^*(x)| dx \rightarrow_{a.s.} 0$$

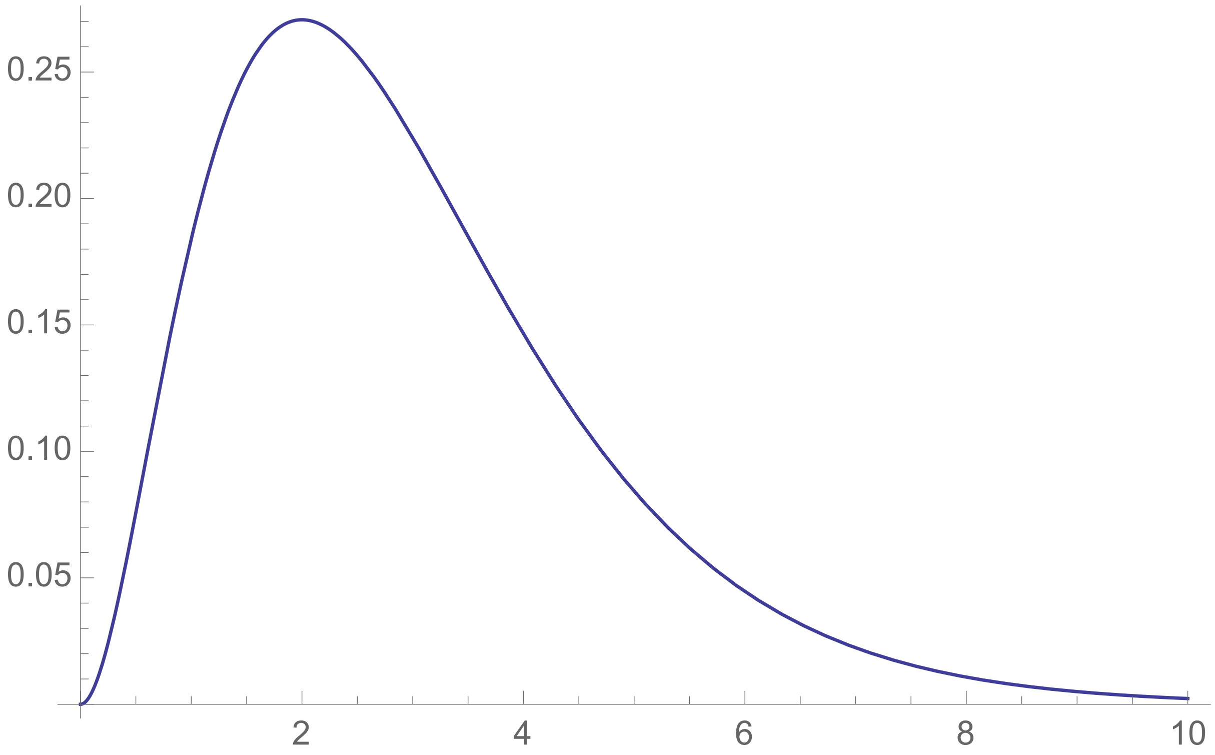
where, for the Kullback-Leibler divergence

$$K(Q, P) = \int \text{log}(dQ/dP)(x) dQ(x),$$

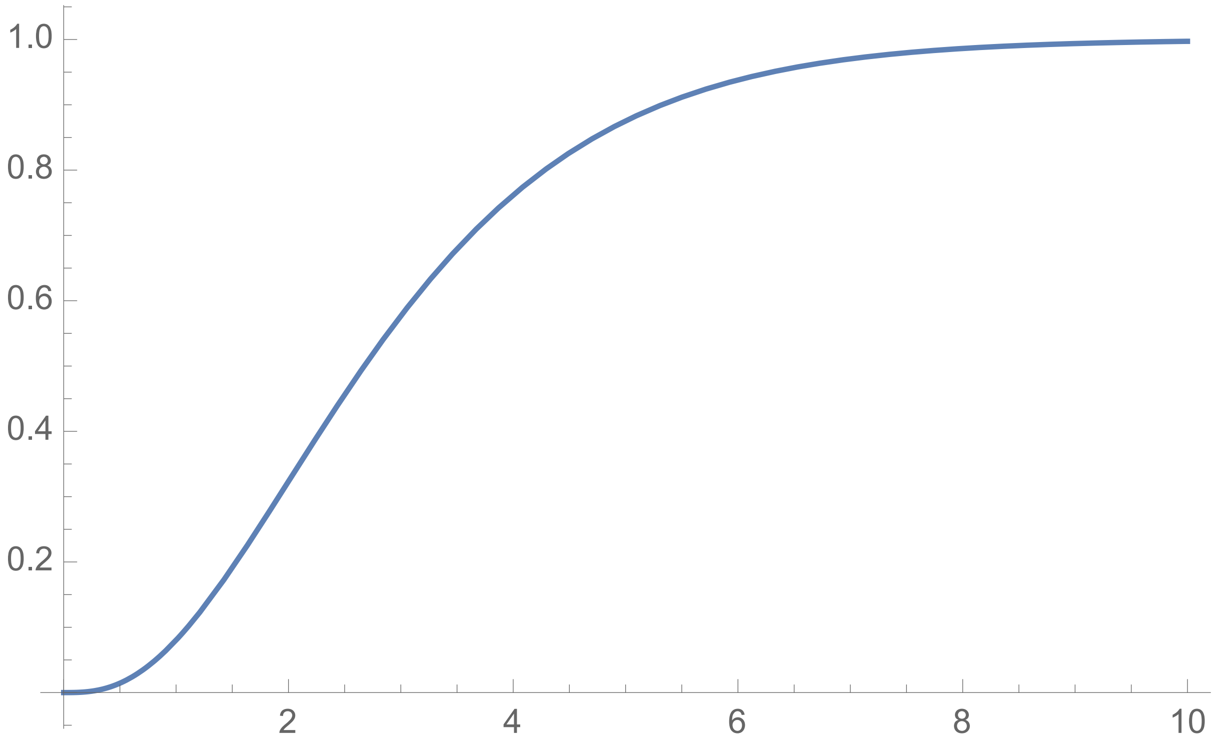
$$p^* \equiv p_Q^* = \text{argmin}_{p \in \mathcal{P}_{mon}(\mathbb{R}+)} K(Q, P)$$

is the "pseudo-true" density in $\mathcal{P}_{mon}(\mathbb{R}^+)$ corresponding to Q. Patilea (2001) also shows that

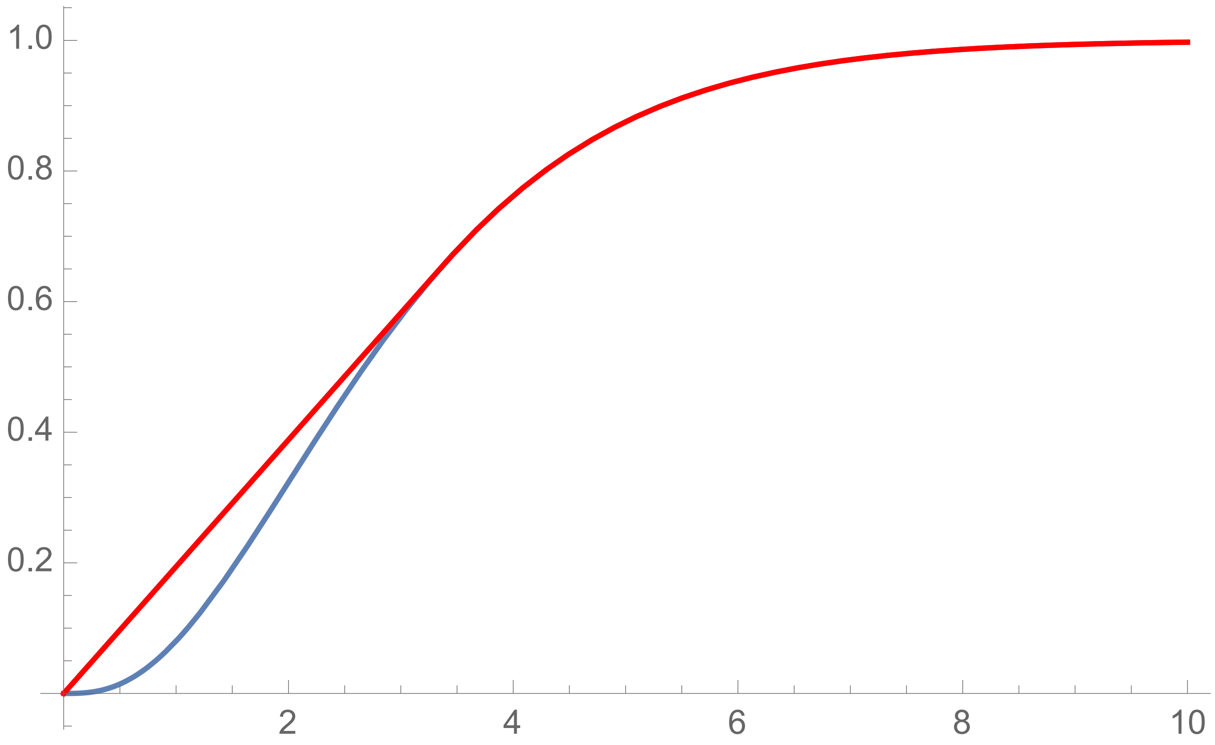
$$\begin{aligned} p^*(x) &= \text{left derivative at } x \text{ of the} \\ &\qquad \text{Least Concave Majorant} \\ &\qquad C \quad \text{of} \quad F_Q(x) \equiv \int_0^x dQ(y). \end{aligned}$$



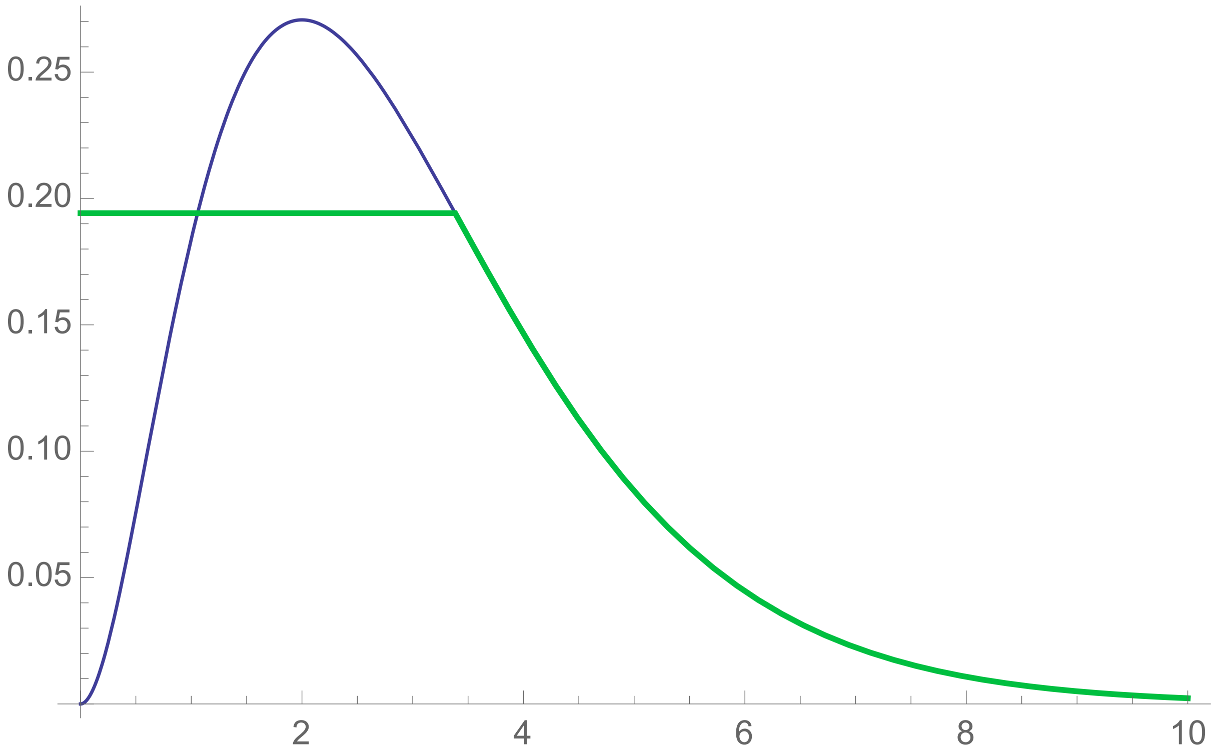
The Gamma$(3, 1)$ density $q(x) = 2^{-1}x^2\text{exp}(-x)1_{(0,\infty)}(x)$.



The Gamma$(3,1)$ distribution function,
$F_Q(x) = 1 - \exp(-x)\left(1 + x + (1/2)x^2\right) 1_{(0,\infty)}(x)$.



The Gamma$(3,1)$ distribution function and its Least Concave Majorant.



The pseudo true density p_Q^* for the Grenander estimator at the Gamma$(3, 1)$ distribution.

This stability property under model misspecification is (very) good! Further results for smooth functionals by Jankowski (2014).

6B. Log-concave densities on $\mathbb{R}$ and $\mathbb{R}^d$

If a density p on $\mathbb{R}^d$ is of the form

$$p(x) \equiv p_\varphi(x) = \exp(\varphi(x)) = \exp\left(-(-\varphi(x))\right)$$

where φ is concave (so $-\varphi$ is convex), then f is **log-concave**. The class of all densities p on $\mathbb{R}^d$ of this form is called the class of *log-concave* densities, $\mathcal{P}_{log-concave} \equiv \mathcal{P}_0$.

Properties of log-concave densities:

- Every log-concave density p is unimodal (quasi concave).
- $\mathcal{P}_0$ is closed under convolution.
- $\mathcal{P}_0$ is closed under marginalization.
- $\mathcal{P}_0$ is closed under weak limits.
- A density p on $\mathbb{R}$ is log-concave if and only if its convolution with any unimodal density is again unimodal (Ibragimov, 1956).

- Many parametric families are log-concave, for example:
  - ▷ Normal (μ, σ^2)
  - ▷ Uniform(a, b)
  - ▷ Gamma(r, λ) for $r \geq 1$
  - ▷ Beta(a, b) for $a, b \geq 1$
- t_r densities with $r > 0$ are not log-concave.
- Tails of log-concave densities are necessarily sub-exponential.
- $\mathcal{P}_{log-concave} =$ the class of "Polyá frequency functions of order 2", PF_2, in the terminology of Schoenberg (1951) and Karlin (1968). See Marshall and Olkin (1979), chapter 18, and Dharmadhikari and Joag-Dev (1988), page 150. for nice introductions.

6C. $s-$ concave densities on $\mathbb{R}$ and $\mathbb{R}^d$

Let $s \in \mathbb{R}$. If a density p on $\mathbb{R}^d$ is of the form

$$p(x) \equiv p_\varphi(x) = \begin{cases} (\varphi(x))^{1/s}, & \varphi \ \ \mathit{convex}, \text{ if } s < 0 \\ \exp(-\varphi(x)), & \varphi \ \ \mathit{convex}, \text{ if } s = 0 \\ (\varphi(x))^{1/s}, & \varphi \ \ \mathit{concave}, \text{ if } s > 0, \end{cases}$$

then f is **s-concave**.

The classes $\mathcal{P}_s$ of all densities p on $\mathbb{R}^d$ of these forms are the classes of $s-$*concave* densities.
The following inclusions hold: if $-\infty < s < 0 < r < \infty$, then

$$\mathcal{P}_r \subset \mathcal{P}_0 \subset \mathcal{P}_s \subset \mathcal{P}_{-\infty}$$

Properties of s-concave densities:

- Every $s-$concave density p is quasi-concave: $\mathcal{P}_s \subset \mathcal{P}_{-\infty}$.
- The Student t_ν density, $t_\nu \in \mathcal{P}_s$ for $s \le -1/(1+\nu)$. Thus the Cauchy density ($= t_1$) is in $\mathcal{P}_{-1/2} \subset \mathcal{P}_s$ for $s \le -1/2$.
- The classes $\mathcal{P}_s$ have interesting closure properties under convolution and marginalization which follow from the Borell-Brascamp-Lieb inequality: let $0 < \lambda < 1$, $-1/d \le s \le \infty$, and let $p, q, h : \mathbb{R}^d \to [0, \infty)$ be integrable functions such that

  $$h((1-\lambda)x + \lambda y) \ge M_s(p(x), q(x), \lambda) \quad \text{for all} \quad x, y \in \mathbb{R}^d$$

  where

  $$M_s(a, b, \lambda) = ((1-\lambda)a^s + \lambda b^s)^{1/s}, \qquad M_0(a, b, \lambda) = a^{1-\lambda} b^\lambda.$$

  Then

  $$\int_{\mathbb{R}^d} h(x)dx \ge M_{s/(sd+1)}\left(\int_{\mathbb{R}^d} p(x)dx, \int_{\mathbb{R}^d} q(x)dx, \lambda \right).$$

D. Maximum Likelihood:

0-concave and s-concave densities

MLE of p and φ: Let $\mathcal{C}$ denote the class of all concave functions $\varphi : \mathbb{R}^d \to [-\infty, \infty)$. The estimator $\widehat{\varphi}_n$ based on $X_1, \ldots, X_n$ i.i.d. as p_0 is the maximizer of the "adjusted criterion function"

$$\begin{aligned} \ell_n(\varphi) &= \int \log p_\varphi(x) d\mathbb{P}_n(x) - \int p_\varphi(x) dx \\ &= \begin{cases} \int \varphi(x) d\mathbb{P}_n(x) - \int e^{\varphi(x)} dx, & s = 0, \\ \int (1/s) \log(-\varphi(x))_+ d\mathbb{P}_n(x) - \int (-\varphi(x))_+^{1/s} dx, & s < 0, \end{cases} \end{aligned}$$

over $\varphi \in \mathcal{C}$.

1. Basics

- The MLE's for $\mathcal{P}_0$ exist and are unique when $n \ge d+1$. Dümbgen and Rufibach (2009);
  Cule, Samworth, and Stewart (2010)

- The MLE for $\mathcal{P}_s$ does not exist if $s < -1/d$. (Doss & W (2013), well known for $s = -\infty$ and $d = 1$.)

- The MLE's for $\mathcal{P}_s$ exist for $s \in (-1/d, 0)$ when

$$n \ge d\left(\frac{r}{r-d}\right)$$

  where $r = -1/s$. Thus the MLE exists only if $n \to \infty$ as $-1/s = r \searrow d$. (Seregin & W (2010))

- MLE $\widehat{\varphi}_n$ is piecewise affine for $-1/d < s \le 0$.

- Uniqueness of MLE's for $\mathcal{P}_s$?

2. On the model

- The MLE's are Hellinger and L_1- consistent.
- The log-concave MLE's $\widehat{p}_{n,0}$ satisfy
$$\int e^{a|x|}|\widehat{p}_{n,0}(x) - p_0(x)|dx \rightarrow_{a.s.} 0.$$
for $a < a_0$ where $p_0(x) \le \exp(-a_0|x| + b_0)$.
- The $s-$concave MLE's are computationally awkward; log is "too aggressive" a transform for $s-$concave densities. [Note that ML has difficulties even for location $t-$ families: multiple roots of the likelihood equations.]
- Global rates? $H(\widehat{p}_{n,s}, p_0) = O_p(n^{-2/5})$ for $-1 < s \le 0$, $d = 1$. (Doss & W 2013, 2015).
- Pointwise distribution theory for $\widehat{p}_{n,0}$ when $d = 1$.
Pointwise distribution theory for MLE when $s < 0$? $d > 1$?

3. Off the model (Cule and Samworth (2010), Dümbgen, Samworth, and Schumacher (2011))

Now suppose that $X \sim Q$ is an arbitrary probability measure on $\mathbb{R}^d$ with density q, $E_Q|X| < \infty$, $X_1, \ldots, X_n$ are i.i.d. q.

- The MLE $\widehat{p}_{n,0}$ for $\mathcal{P}_0$ satisfies:

$$\int_{\mathbb{R}^d} |\widehat{p}_{n,0}(x) - p_0^*(x)| dx \rightarrow_{a.s.} 0$$

where, for the Kullback-Leibler divergence

$$K(q,p) = \int q(x) \text{log}(q(x)/p(x)) dx,$$

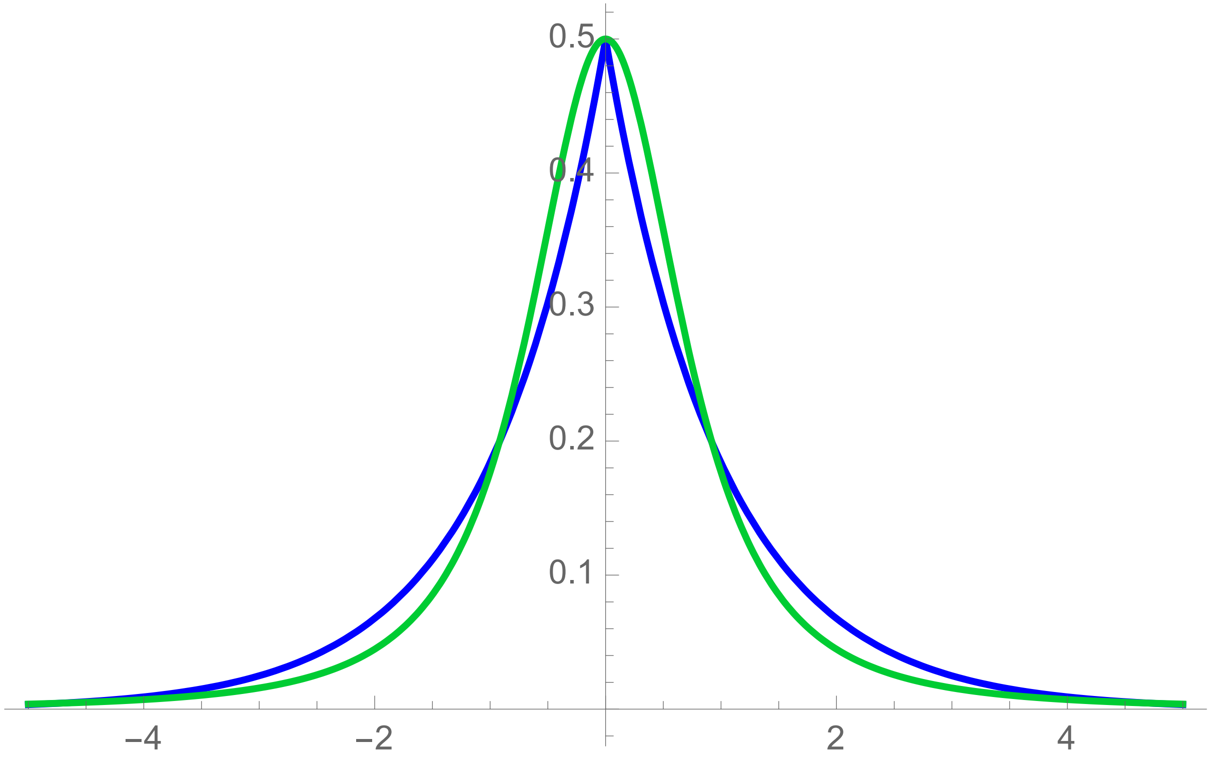
$$p_0^* \equiv p_Q^* = \text{argmin}_{p \in \mathcal{P}_0(\mathbb{R}^d)} K(q,p)$$

is the "pseudo-true" density in $\mathcal{P}_0(\mathbb{R}^d)$ corresponding to q. In fact: for any $a < a_0$ where $p_0^*(x) \le \exp(-a_0\|x\| + b_0)$,

$$\int_{\mathbb{R}^d} e^{a\|x\|} |\widehat{p}_{n,0}(x) - p_0^*(x)| dx \rightarrow_{a.s.} 0$$

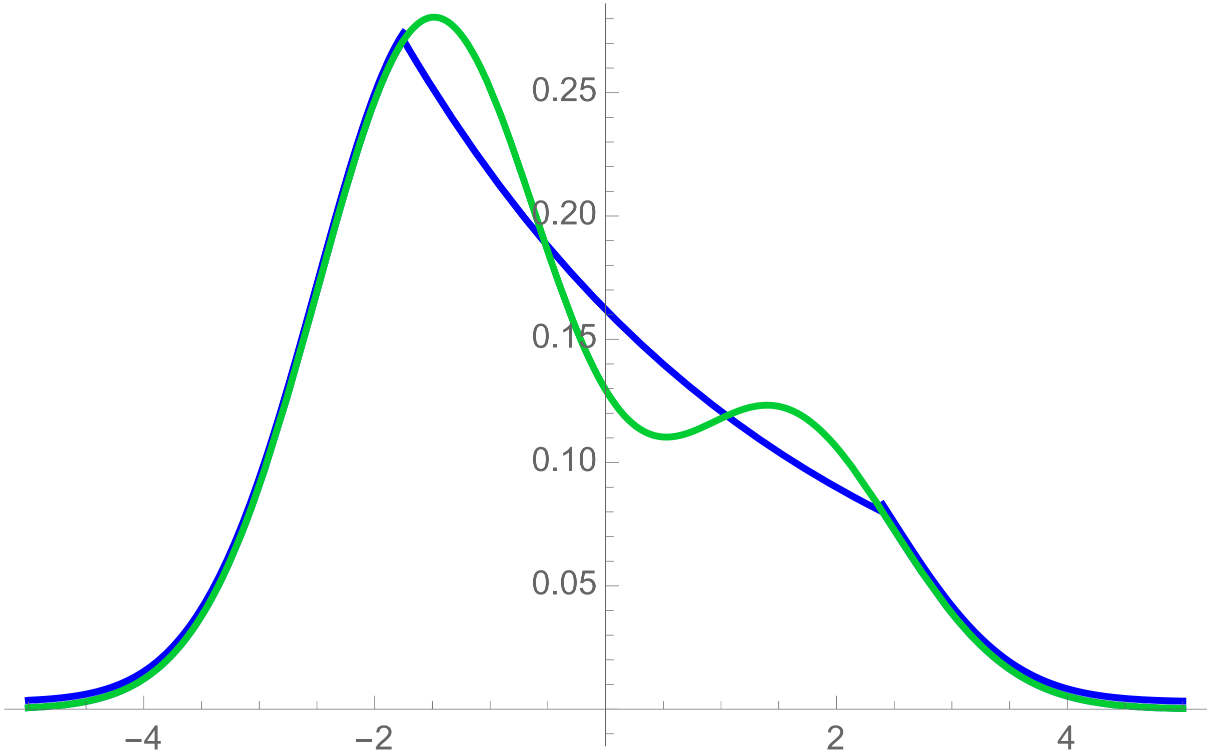
- Example 1. (Dümbgen, Samworth, and Schumacher (2011)). If

$$q(x) = \frac{1}{2}\frac{1}{(1+x^2)^{3/2}}, \qquad \text{then} \qquad p^*(x) = \frac{1}{2}\text{exp}(-|x|).$$



- Example 2. (Dümbgen, Samworth, and Schumacher (2011)). If

$$q(x) = .7 \cdot N(-1.5, 1) + .3N(1.5, 1), \qquad \text{then} \qquad p^* \quad \text{is :}$$



- This stability property of the MLE $\widehat{p}_{n,0}$ for $\mathcal{P}_0$ is (very!) good!
- In contrast, the MLE $\widehat{p}_{n,s}$ for $\mathcal{P}_s$ does not behave well off the model. Retracing the basic arguments of Cule and Samworth (2010) leads to negative conclusions. (How negative remains to be pinned down!)

Conclusion: Investigate **alternative methods** for estimation in the larger classes $\mathcal{P}_s$ with $s < 0$! This leads to the proposals by Koenker and Mizera (2010).

E. An alternative to ML:
Rényi divergence estimators

0. Notation and Definitions

- $\beta = 1 + 1/s < 0$, $\alpha^{-1} + \beta^{-1} = 1$.
- $\mathcal{C}(\underline{X}) =$ all continuous functions on conv($\underline{X}$).
- $\mathcal{C}^*(\underline{X}) =$ all signed Radon measures on $\mathcal{C}(\underline{X})$ = dual space of $\mathcal{C}(\underline{X})$.
- $\mathcal{G}(\underline{X}) =$ all closed convex (lower s.c.) functions on conv($\underline{X}$).
- $\mathcal{G}(\underline{X})_+ =$ all non-negative $g \in \mathcal{G}(\underline{X})$
- $\mathcal{G}(\underline{X})^\circ = \{G \in \mathcal{C}^*(\underline{X}) : \ \int g dG \le 0$ for all $g \in \mathcal{G}(\underline{X}\}$, the polar (or dual) cone of $\mathcal{G}(\underline{X})$;
- $\mathcal{G}(\underline{X})_+^\circ$

Primal problems: $\mathcal{P}_0$ and $\mathcal{P}_s$:

- $\mathcal{P}_0$: $\min_{g \in \mathcal{G}(\underline{X})} L_0(g, \mathbb{P}_n)$ where

$$L_0(g, \mathbb{P}_n) = \mathbb{P}_n g + \int_{\mathbb{R}^d} \exp(-g(x))dx.$$

- $\mathcal{P}_s$: $\min_{g \in \mathcal{G}(\underline{X})_+} L_s(g, \mathbb{P}_n)$ where

$$L_s(g, \mathbb{P}_n) = \mathbb{P}_n g + \frac{1}{|\beta|} \int_{\mathbb{R}^d} g(x)^\beta dx.$$

Dual problems: $\mathcal{P}_0$ and $\mathcal{P}_s$:

- $\mathcal{D}_0$: $\max_p \{-\int p(y)\log p(y)dy\}$ subject to
$$p(y) = \frac{d(\mathbb{P}_n - G)}{dy} \quad \text{for some} \quad G \in \mathcal{G}(\underline{X})^{\circ}.$$
- $\mathcal{D}_s$: $\max_p \int \frac{p(y)^{\alpha}}{\alpha}dy$ subject to
$$p(y) = \frac{d(\mathbb{P}_n - G)}{dy} \quad \text{for some} \quad G \in \mathcal{G}(\underline{X})^{\circ}_{+}.$$

Why do these make sense?

- Population version of $\mathcal{P}_0$: $\min_{g\in\mathcal{G}} L_0(g,p_0)$ where

  $$L_0(g,p_0) = \int \{g(x)p_0(x) + e^{-g(x)}\}dx.$$

  Minimizing the integrand pointwise in $g = g(x)$ for fixed $p_0(x)$ yields $p_0(x) - e^{-g} = 0$ if $e^{-g} = e^{-g(x)} = p_0(x)$.

- Population version of $\mathcal{P}_s$: $\min_{g\in\mathcal{G}} L_s(g,p_0)$ where

  $$L_s(g,p_0) = \int \{g(x)p_0(x) + \frac{1}{|\beta|}g^{\beta}(x)\}dx.$$

  Minimizing the integrand pointwise in $g = g(x)$ for fixed $p_0(x)$ yields $f_0(x) + (\beta/|\beta|)g^{\beta-1} = p_0(x) - g^{\beta-1} = 0$, and hence $g^{1/s} = g^{1/s}(x) = p_0(x)$.

1. **Basics for the Rényi divergence estimators:**

- (Koenker and Mizera, 2010) If conv($\underline{X}$) has non-empty interior (true a.s. if $n \geq d+1$), then strong duality between $\mathcal{P}_s$ and $\mathcal{D}_s$ holds. The dual optimal solution exists, is unique, and $\widehat{p}_n = \widehat{g}_n^{1/s}$.
- (Koenker and Mizera, 2010) The solution $p = g^{1/s}$ in the population version of the problem when $Q = P_0$ has density $p_0 \in \mathcal{P}_s$ is Fisher-consistent; i.e. $p = p_0$.

2. Off the model: Han & W (2015)

Let

$$\mathcal{Q}_1 \equiv \{Q \text{ on } (\mathbb{R}^d, \mathcal{B}^d) : \int \|x\| dQ(x) < \infty\},$$
$$\mathcal{Q}_0 \equiv \{Q \text{ on } (\mathbb{R}^d, \mathcal{B}^d) : \text{int}(\text{csupp}(Q)) \neq \emptyset\}.$$

- Theorem (Han & W, 2015): If $-1/(d+1) < s < 0$ and $Q \in \mathcal{Q}_0 \cap \mathcal{Q}_1$, then the primal problem $\mathcal{P}_s(Q)$ has a unique solution $\tilde{g} \in \mathcal{G}$ which satisfies $\tilde{p} = \tilde{g}^{1/s}$ where $\tilde{g}$ is bounded away from 0 and $\tilde{p}$ is a bounded density.

- Theorem (Han & W, 2015): Suppose that:
  (i) $d \geq 1$,
  (ii) $-1/(d+1) < s < 0$, and
  (iii) $Q \in \mathcal{Q}_0 \cap \mathcal{Q}_1$.
  If $p_{Q,s}$ denotes the (pseudo-true) solution to the primal problem $\mathcal{P}_s(Q)$, then for any $\kappa < r - d = (-1/s) - d$,

$$\int (1 + |x|)^\kappa |\widehat{p}_{n,s}(x) - p_{Q,s}(x)| dx \to_{a.s.} 0 \quad \text{as} \quad n \to \infty.$$

- Theorem (Han & W, 2015): Let $d = 1$. If $\widehat{p}_{n,s}$ denotes the solution to the primal problem $\mathcal{P}_s$ and $\widehat{p}_{n,0}$ denotes the solution to the primal problem $\mathcal{P}_0$, then for any $\kappa > 0$, $p \geq 1$,

$$\int (1 + |x|)^\kappa |\widehat{p}_{n,s}(x) - \widehat{p}_{n,0}(x)|^p dx \to 0 \quad \text{as} \quad s \nearrow 0.$$

3. On the model: Q has density $p \in \mathcal{P}_s$; $p = g^{1/s}$ for some g convex.

- Consistency: Suppose that: (i) $d \geq 1$ and $-1/(d+1) < s < 0$. Then for any $\kappa < r - d = (-1/s) - d$,

  $$\int (1 + |x|)^{\kappa} |\widehat{p}_{n,s}(x) - p(x)| dx \rightarrow_{a.s.} 0 \quad \text{as} \quad n \rightarrow \infty.$$

  Thus $H(\widehat{p}_{n,s}, p) \rightarrow_{a.s.} 0$ as well.

- Pointwise limit theory: (paralleling the results of Balabdaoui, Rufibach, and W (2009) for $s = 0$)

See Han & W (2015) and W (2015), EMS talk, Amsterdam

- Summary for log-concave MLE and $s-$concave MLE:
  - For log-concave densities $\mathcal{P}_0$: MLE is consistent and stable under model misspecification. Good!
  - For $s-$concave densities $\mathcal{P}_s$ with $s < 0$:
    - ▷ MLE does not exist for $s < -1/d$; exists for $-1/d < s < 0$ but only for $n \geq rd/(r-d)$ with $r = -1/s > d$. Bad!
    - ▷ MLE is consistent for $-1/d < s < 0$, but unstable under model misspecification. Bad!

  For $s-$concave densities $\mathcal{P}_s$ with $s < 0$:
    - ▷ Rényi divergence estimators exist for $-1/(d+1) < s < 0$ and $n \geq d+1$. Good!
    - ▷ Rényi divergence estimators are stable under model misspecification. Good!

7. Summary and Conclusions:

ugly, bad, and good

- Bad and ugly:

Maximum likelihood remains useful in many nonparametric and semiparametric models, but it also has potential difficulties and shortcomings, including:

- The MLE may not exist.
- If the MLE exists, it may only exist for sample sizes which grow with the size of the model.
- If the MLE exists, it may be inconsistent.
- If the MLE is consistent, it may be rate - inefficient for models which are "trans-Donsker" (i.e. with divergent entropy integrals). (This behavior persists for other minimum contrast estimators!)
- If the MLE exists and is consistent, it may be unstable with respect to model misspecification.

- Good: Maximum likelihood estimation and likelihood methods more broadly have been very successful in many semiparametric and nonparametric models, including:
    - Nonparametric models: Right censoring: the Kaplan-Meier estimator: (consistent, asymptotically efficient).
    - Semiparametric models: The Cox proportional hazards model: (consistent, asymptotically efficient).
    - Semiparametric models: Kiefer-Wolfowitz semiparametric mixture models: (consistent, asymptotically efficient in some cases)
    - Nonparametric, shape constrained: Grenander's MLE of a decreasing density: (consistent, stable under model misspecification)
    - Nonparametric, shape constrained: Log-concave densities: (consistent, stable under model misspecification).

- Many problems remain:
  - How can we make use of likelihood methods in settings involving high-dimensions?
  - How to find alternatives to ML with desirable statistical (and computational) properties?
  - Other divergences or contrast functions may have better behavior?
  - Can we have both rate efficiency and stability under model misspecification?

Maximum Likelihood & variants: current activity levels, MathSciNet

search term	total hits	hits 2010-2015	% recent
Maximum likelihood	13003	2782	21%
Partial likelihood	308	71	23%
Nonparametric ML	306	189	62%
Semiparametric ML	283	111	39%
Profile likelihood	255	90	35%
Pseudo likelihood	217	81	37%
Restricted ML	202	42	21%
Conditional ML	144	33	23%
Penalized ML	106	87	82%
Approximate likelihood	102	27	26%
Composite likelihood	89	73	82%
Generalized ML	79	12	15%
ML + consistency	1163	254	22%
ML + efficiency	977	198	20%
ML + computation	752	232	31%
ML + smooth	204	48	24%
ML + sieve	82	18	22%

From van der Vaart (2002), “The statistical work of Lucien Le Cam”.

> “In Section 11 we noted that Le Cam was critical of the method of maximum likelihood, which in his view looks for a “peculiarity” of the likelihood function. This did not prevent him from investigating conditions under which the MLE does have the properties that are usually ascribed to it.”

From Le Cam (1990), “An introduction to maximum likelihood”.

> “If the hallowed principle of maximum likelihood leads us to difficulties, maybe some other principle will save us. There is such a principle. It is as follows:
>
> - *Basic Principle 0:* **Do not trust any principle.**

Thank you!