Approximation in high-dimensional and nonparametric statistics: a unified approach

Yanjun Han

Department of Electrical Engineering, Stanford University

Joint work with:

Jiantao Jiao	Stanford EE
Rajarshi Mukherjee	Stanford Stats
Tsachy Weissman	Stanford EE

July 4, 2016

Outline

1. Problem Setup
2. High-dimensional Parametric Setting
3. Infinite Dimensional Nonparametric Setting
   - Upper bound
   - Lower bound
   - General L_r norm

Problem: estimation of functionals

Given i.i.d. samples $X_1, \cdots, X_n \sim P$, we would like to estimate a one-dimensional functional $F(P) \in \mathbb{R}$:

- Parametric case: $P = (p_1, \cdots, p_S)$ is discrete, and

$$F(P) = \sum_{i=1}^{S} I(p_i)$$

  High dimensional: $S \gtrsim n$

- Nonparametric case: P is continuous with density f, and

$$F(P) = \int I(f(x))dx$$

Parametric case: when the functional is smooth...

When $I(\cdot)$ is everywhere differentiable...

Hájek–Le Cam Theory

The plug-in approach $F(P_n)$ is asymptotically efficient, where P_n is the empirical distribution

Nonparametric case: when the functional is smooth...

When $I(\cdot)$ is four times differentiable with bounded $I^{(4)}$, Taylor expansion yields

$$\int I(f(x))dx = \int \Big[I(\hat{f}) + I^{(1)}(\hat{f})(f-\hat{f}) + \frac{1}{2}I^{(2)}(\hat{f})(f-\hat{f})^2 + \frac{1}{6}I^{(3)}(\hat{f})(f-\hat{f})^3 + O((f-\hat{f})^4) \Big] dx$$

where $\hat{f}$ is a "good" estimator of f (e.g., a kernel estimate)

- Key observation: suffice to deal with linear (see, e.g., Nemirovski'00), quadratic (Bickel and Ritov'88, Birge and Massart'95) and cubic terms (Kerkyacharian and Picard'96) separately.
- Require bias reduction

What if $I(\cdot)$ is non-smooth?

Bias dominates when $I(\cdot)$ is non-smooth:

Theorem (ℓ_1 norm of Gaussian mean, Cai–Low'11)

For $y_i \sim \mathcal{N}(\theta_i, \sigma^2), i = 1, \cdots, n$ and $F(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} |\theta_i|$, the plug-in estimator satisfies

$$\sup_{\boldsymbol{\theta} \in \mathbb{R}^n} \mathbb{E}_{\boldsymbol{\theta}} \left(F(\boldsymbol{y}) - F(\boldsymbol{\theta}) \right)^2 \asymp \underbrace{\sigma^2}_{\text{squared bias}} + \underbrace{\frac{\sigma^2}{n}}_{\text{variance}}$$

Theorem (Discrete entropy, Jiao–Venkat–H.–Weissman'15)

For $X_1, \cdots, X_n \sim P = (p_1, \cdots, p_S)$ and $F(P) = \sum_{i=1}^{S} -p_i \ln p_i$, the plug-in estimator satisfies

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P (F(P_n) - F(P))^2 \asymp \underbrace{\frac{S^2}{n^2}}_{\text{squared bias}} + \underbrace{\frac{(\ln S)^2}{n}}_{\text{variance}}$$

The optimal estimator

Theorem (ℓ_1 norm of Gaussian mean, Cai–Low'11)

For $y_i \sim \mathcal{N}(\theta_i, \sigma^2), i = 1, \cdots, n$ *and* $F(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} |\theta_i|$,

$$\inf_{\hat{F}} \sup_{\boldsymbol{\theta} \in \mathbb{R}^n} \mathbb{E}_{\boldsymbol{\theta}} \left(\hat{F} - F(\boldsymbol{\theta}) \right)^2 \asymp \frac{\sigma^2}{\underbrace{\ln n}_{\text{squared bias}}}$$

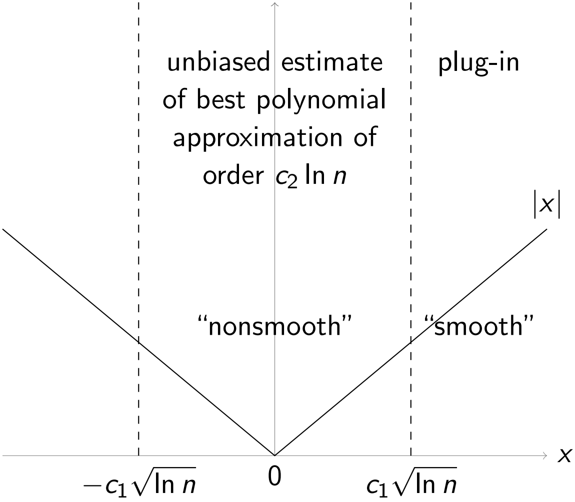
Theorem (Discrete entropy, Jiao–Venkat–H.–Weissman'15)

For $X_1, \cdots, X_n \sim P = (p_1, \cdots, p_S)$ *and* $F(P) = \sum_{i=1}^{S} -p_i \ln p_i$,

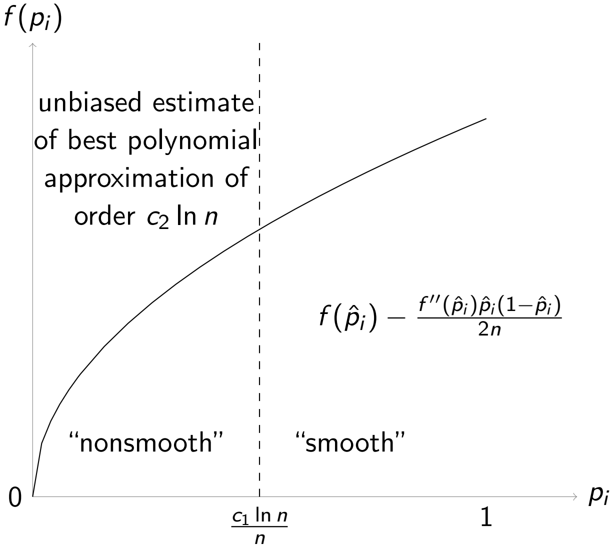
$$\inf_{\hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P (\hat{F} - F(P))^2 \asymp \underbrace{\frac{S^2}{(n \ln n)^2}}_{\text{squared bias}} + \underbrace{\frac{(\ln S)^2}{n}}_{\text{variance}}$$

Effective sample size enlargement: n samples $\longrightarrow$ $n \ln n$ samples

Optimal estimator for ℓ_1 norm



Optimal estimator for entropy



The general recipe

For a statistical model $(P_\theta : \theta \in \Theta)$, consider estimating the functional $F(\theta)$ which is non-analytic at $\Theta_0 \subset \Theta$, and $\hat{\theta}_n$ is a natural estimator for θ.

1. **Classify the Regime**: Compute $\hat{\theta}_n$, and declare that we are in the "non-smooth" regime if $\hat{\theta}_n$ is "close" enough to Θ_0. Otherwise declare we are in the "smooth" regime;
2. **Estimate**:
   - If $\hat{\theta}_n$ falls in the "smooth" regime, use an estimator "similar" to $F(\hat{\theta}_n)$ to estimate $F(\theta)$;
   - If $\hat{\theta}_n$ falls in the "non-smooth" regime, replace the functional $F(\theta)$ in the "non-smooth" regime by an approximation $F_{\text{appr}}(\theta)$ (another functional) which can be estimated without bias, then apply an unbiased estimator for the functional $F_{\text{appr}}(\theta)$.

Questions

- How to determine the "non-smooth" regime?
- In the "smooth" regime, what does " 'similar' to $F(\hat{\theta}_n)$" mean precisely?
- In the "non-smooth" regime, what approximation (including which kind, which degree, and on which region) should be employed?
- What if the domain of $\hat{\theta}_n$ is different from (usually larger than) that of θ?

1. Problem Setup

2. **High-dimensional Parametric Setting**

3. Infinite Dimensional Nonparametric Setting
   - Upper bound
   - Lower bound
   - General L_r norm

Estimation of information divergence

Given joint independent samples $X_1, \cdots, X_m \sim P = (p_1, \cdots, p_S)$ and $Y_1, \cdots, Y_n \sim Q = (q_1, \cdots, q_S)$, we would like to estimate the L_1 distance and the Kullback–Leibler (KL) divergence:

$$\|P - Q\|_1 = \sum_{i=1}^{S} |p_i - q_i|$$

$$D(P\|Q) = \begin{cases} \sum_{i=1}^{S} p_i \ln \frac{p_i}{q_i} & \text{if } P \ll Q, \\ +\infty & \text{otherwise.} \end{cases}$$

In the latter case, we assume a bounded likelihood ratio: $p_i/q_i \leq u(S)$ for any i.

Localization

Definition (Localization)

Consider a statistical model $(P_\theta)_{\theta\in\Theta}$ and an estimator $\hat{\theta} \in \hat{\Theta}$ of θ, where $\Theta \subset \hat{\Theta}$. A localization of level $r \in [0,1]$, or an r-localization, is a collection of sets $\{U(x)\}_{x\in\hat{\Theta}}$, where $U(x) \subset \Theta$ for any $x \in \hat{\Theta}$, and

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\theta \notin U(\hat{\theta})) \leq r.$$

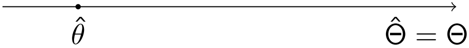
- Naturally induce a reverse localization $V(\theta) = \{\hat{\theta} : U(\hat{\theta}) \ni \theta\}$
- Localization always exists, but we seek for a small one
- Different from confidence set: usually $r \asymp n^{-A}$

Localization in Gaussian model: $r \asymp n^{-A}$

$$\hat{\Theta} = \Theta = \mathbb{R}$$

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$$

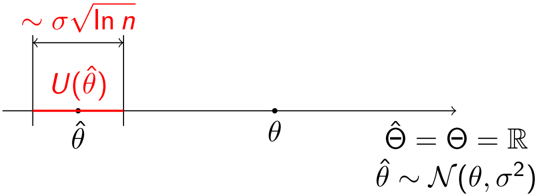
Localization in Gaussian model: $r \asymp n^{-A}$



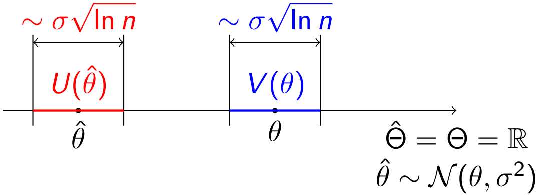
$$\hat{\Theta} = \Theta = \mathbb{R}$$
$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$$

Localization in Gaussian model: $r \asymp n^{-A}$

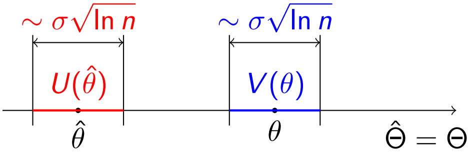
Localization in Gaussian model: $r \asymp n^{-A}$



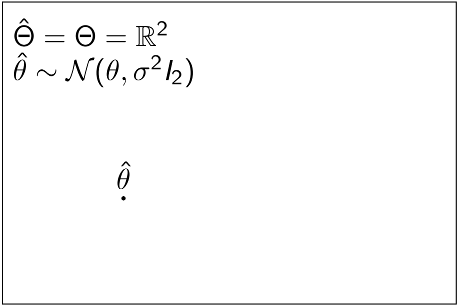
Localization in Gaussian model: $r \asymp n^{-A}$



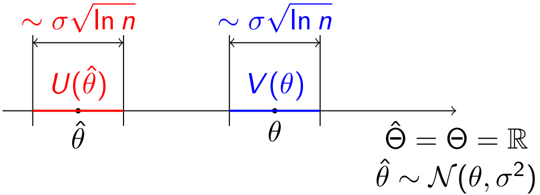
Localization in Gaussian model: $r \asymp n^{-A}$

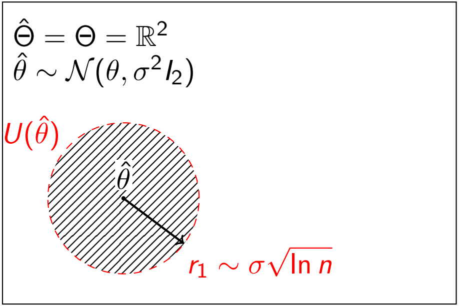


$\hat{\Theta} = \Theta = \mathbb{R}$
$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$

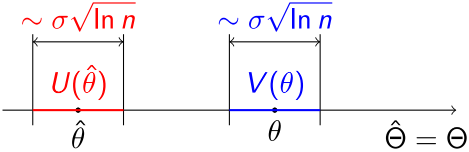


Localization in Gaussian model: $r \asymp n^{-A}$

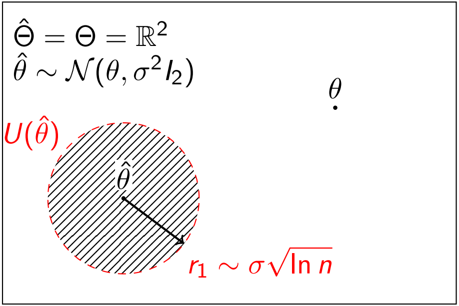




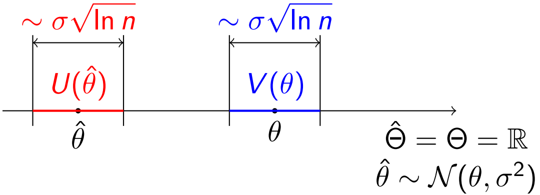
Localization in Gaussian model: $r \asymp n^{-A}$

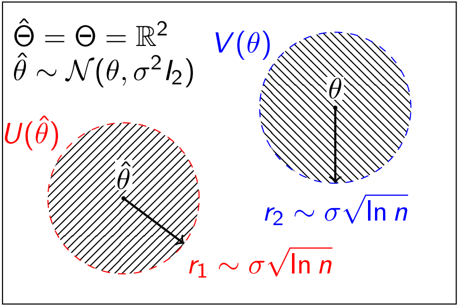


$\hat{\Theta} = \Theta = \mathbb{R}$
$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$



Localization in Gaussian model: $r \asymp n^{-A}$





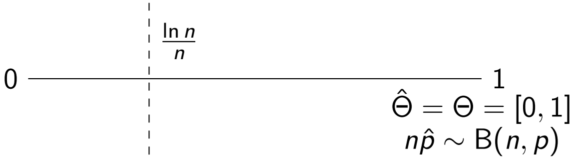
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$

0 ——————————————— 1

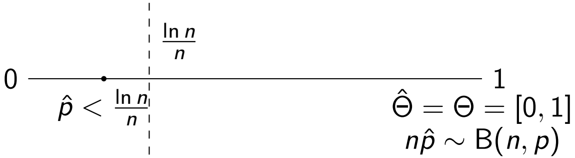
$\hat{\Theta} = \Theta = [0, 1]$

$n\hat{p} \sim \mathrm{B}(n, p)$

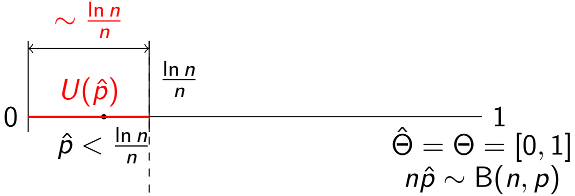
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



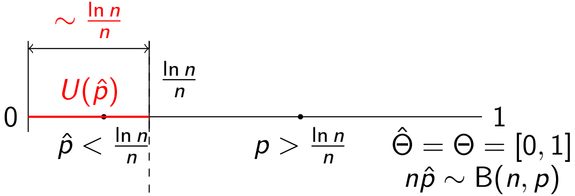
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



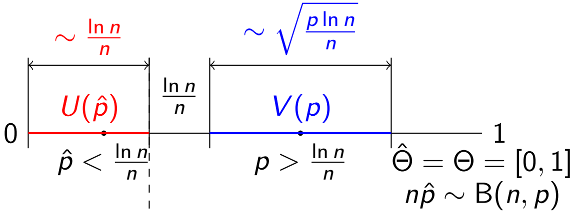
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



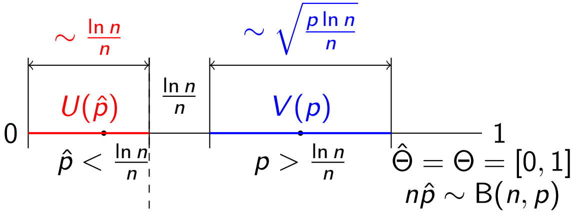
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



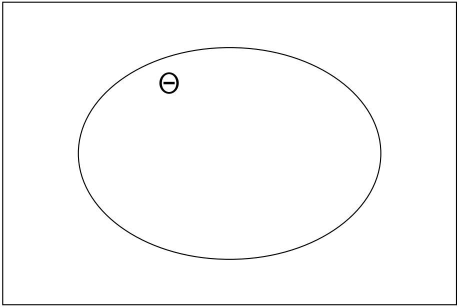
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



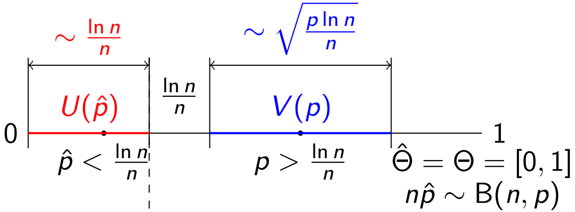
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



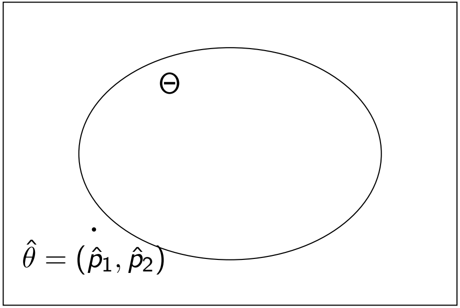
$$\hat{\Theta} = [0,1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



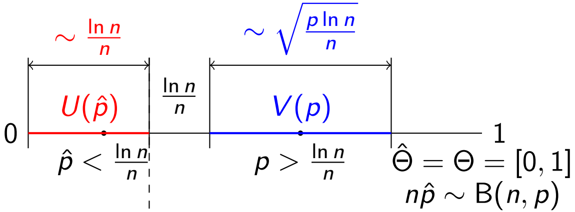
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



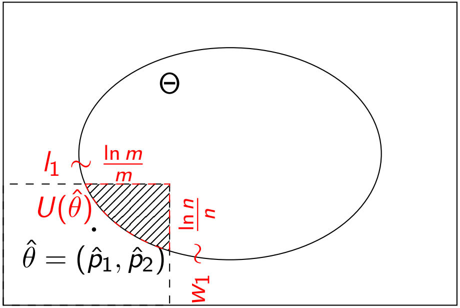
$$\hat{\Theta} = [0, 1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



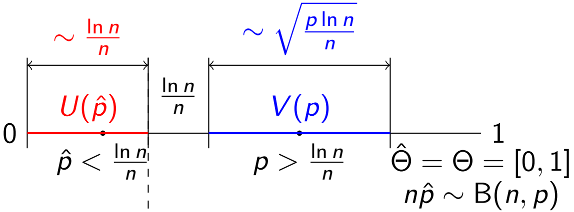
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



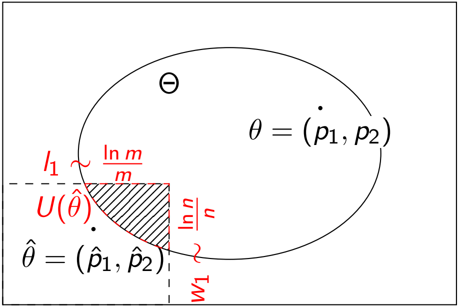
$$\hat{\Theta} = [0, 1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



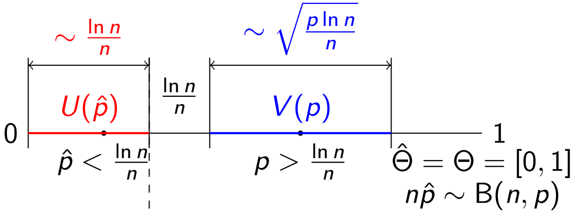
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



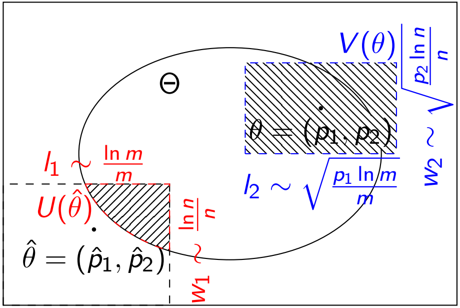
$$\hat{\Theta} = [0, 1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



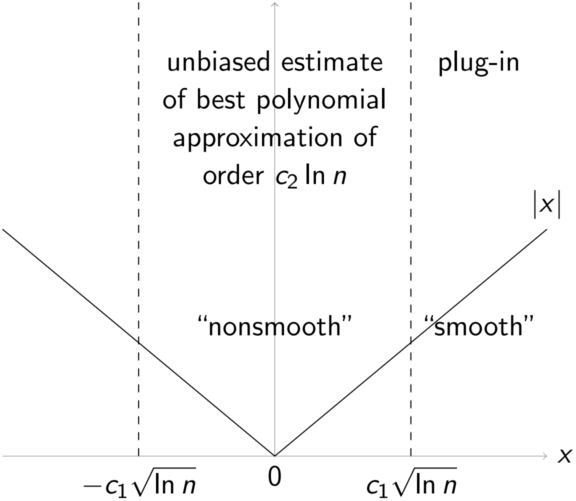
Localization in Binomial model: $r \asymp \min\{m, n\}^{-A}$



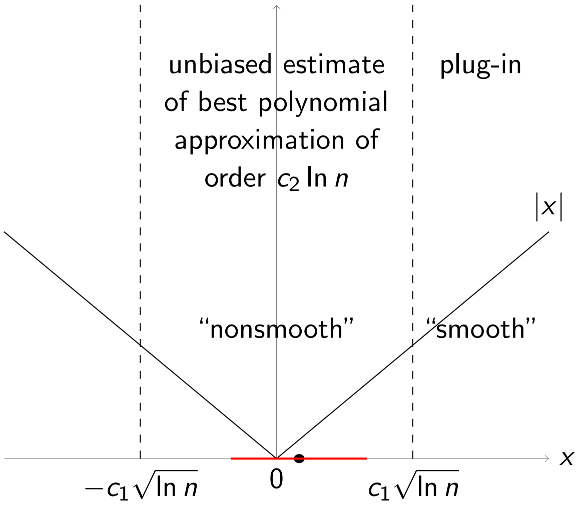
$$\hat{\Theta} = [0, 1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



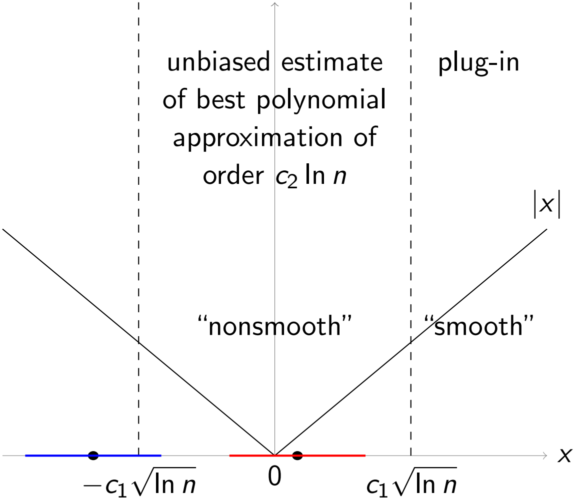
The role of localization: ℓ_1 norm estimation



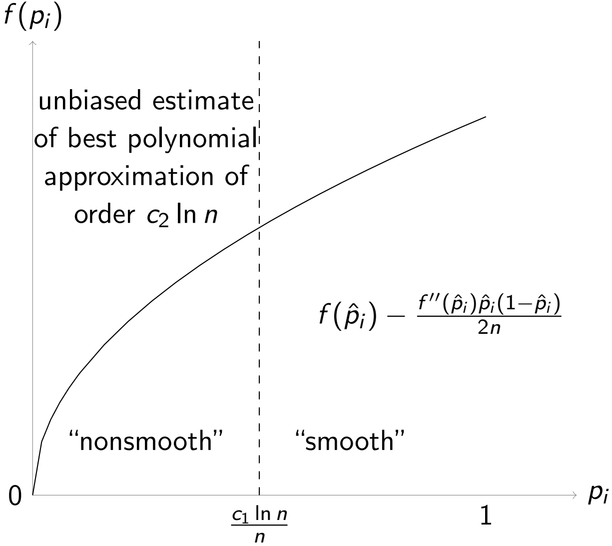
The role of localization: ℓ_1 norm estimation



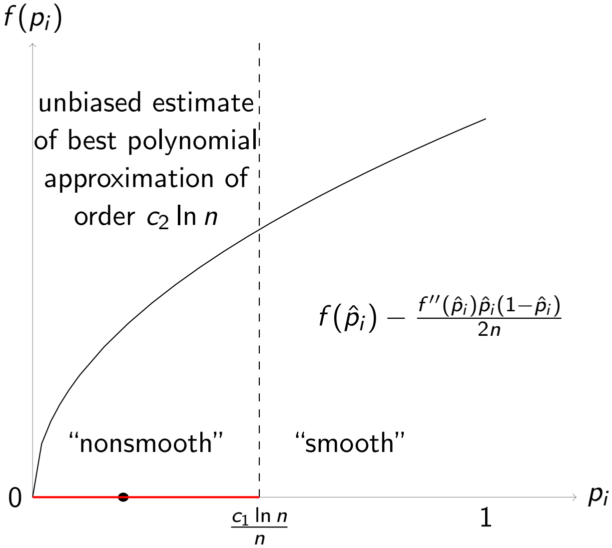
The role of localization: ℓ_1 norm estimation



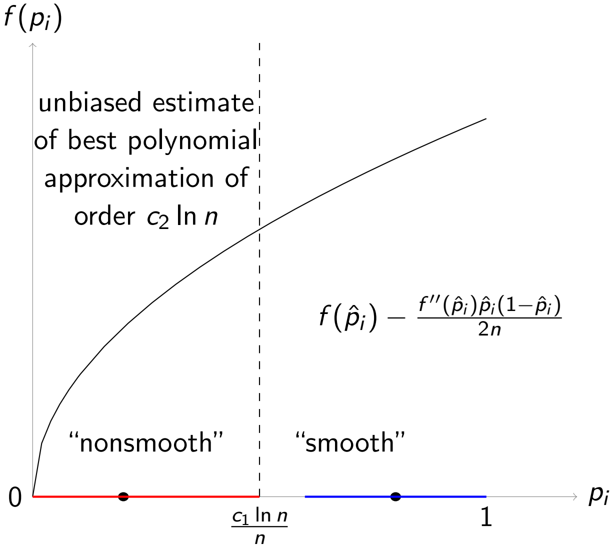
The role of localization: entropy estimation



The role of localization: entropy estimation



The role of localization: entropy estimation



Determine the "non-smooth" regime

Analysis of the plug-in approach:

$$I(\hat{\theta}_n) = I(\theta) + I'(\theta)(\hat{\theta}_n - \theta) + \frac{1}{2}I''(\xi)(\hat{\theta}_n - \theta)^2$$

- Plug-in works well when $\hat{\theta}_n \notin \hat{\Theta}_0$ (recall that $I(\cdot)$ is non-analytic in $\hat{\Theta}_0 \subset \hat{\Theta}$)

Determine the "non-smooth" regime

Analysis of the plug-in approach:

$$I(\hat{\theta}_n) = I(\theta) + I'(\theta)(\hat{\theta}_n - \theta) + \frac{1}{2} I''(\xi)(\hat{\theta}_n - \theta)^2$$

- Plug-in works well when $\hat{\theta}_n \notin \hat{\Theta}_0$ (recall that $I(\cdot)$ is non-analytic in $\hat{\Theta}_0 \subset \hat{\Theta}$)

The criteria

Given a suitable r-localization $U(\cdot)$, we declare that θ falls into the "non-smooth" regime Θ_{ns} if

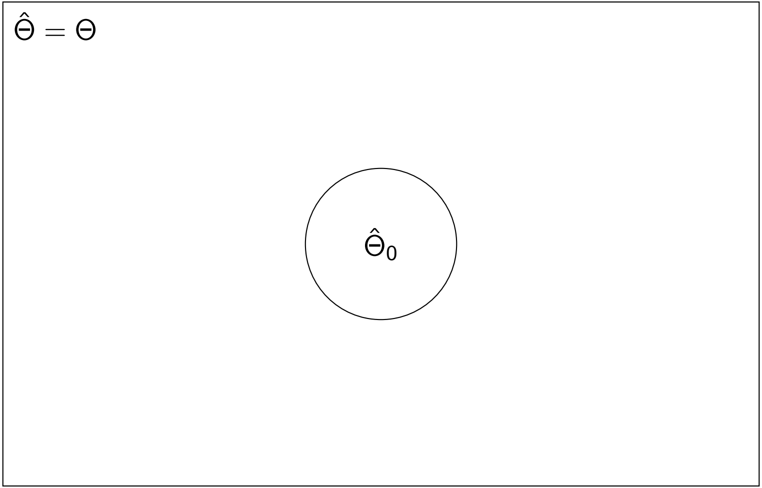
$$\theta \in \cup_{\hat{\theta} \in \hat{\Theta}_0} U(\hat{\theta})$$

and in the "smooth" regime Θ_s otherwise.

Idea: $\sup_{\theta \in \Theta_s} \mathbb{P}_\theta(\hat{\theta}_n \in \hat{\Theta}_0) \le \sup_{\theta \in \Theta_s} \mathbb{P}_\theta(\theta \notin U(\hat{\theta}_n)) \le r$

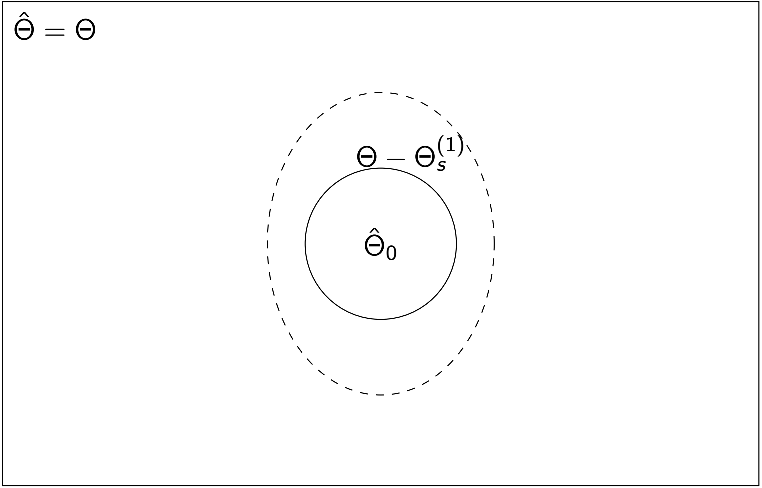
There is something more...

However, we cannot make decisions based on unknown θ!



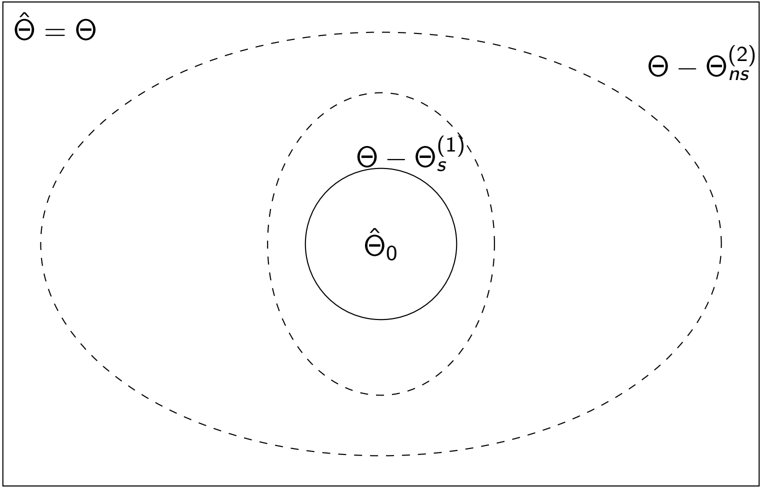
There is something more...

However, we cannot make decisions based on unknown θ!



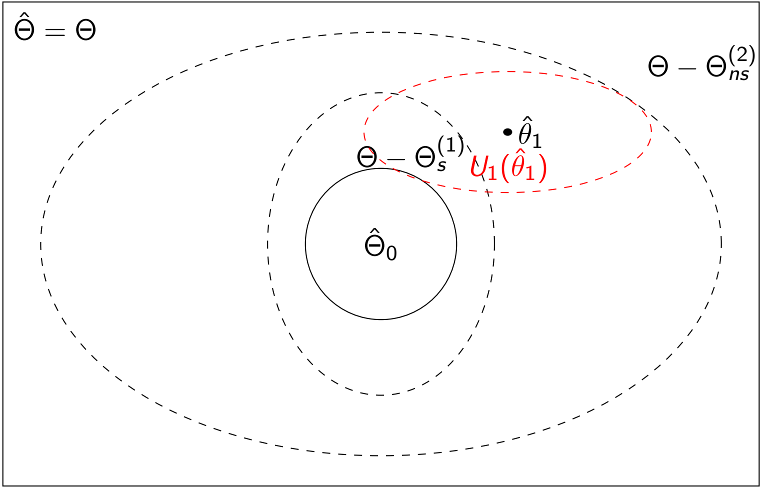
There is something more...

However, we cannot make decisions based on unknown θ!



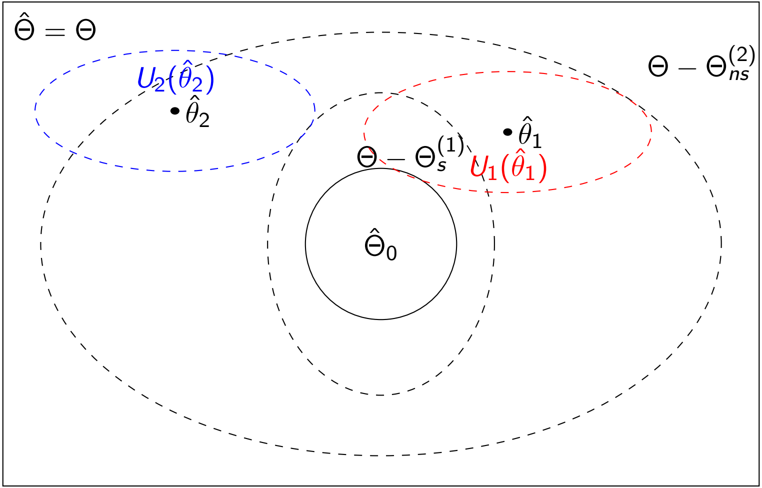
There is something more...

However, we cannot make decisions based on unknown θ!



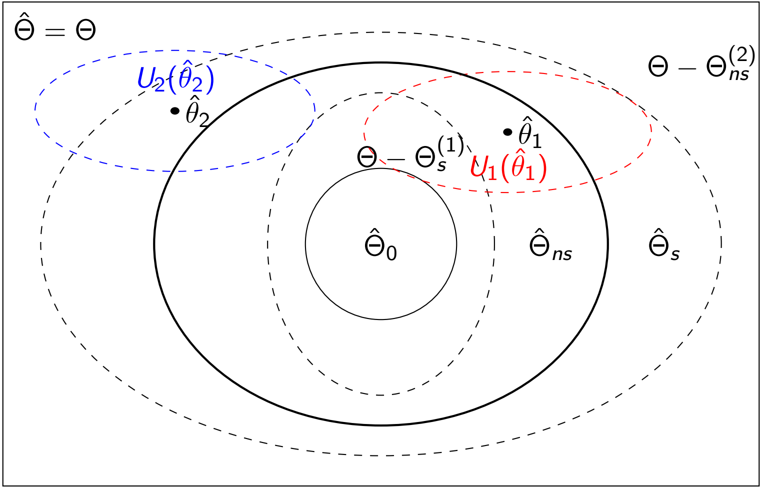
There is something more...

However, we cannot make decisions based on unknown θ!



There is something more...

However, we cannot make decisions based on unknown θ!



“Non-smooth” regime: approximation

Find an approximate functional $I_{\text{appr}}(\theta) \approx I(\theta)$, and use an unbiased estimate $T(\hat{\theta}_n)$, i.e., $\mathbb{E}T(\hat{\theta}_n) = I_{\text{appr}}(\theta)$.

- Type: polynomial in Multinomial, Poisson and Gaussian models (only polynomials have unbiased estimate!)
- Region: suffice to use $U(\hat{\theta}_n)$ ($\theta \in U(\hat{\theta}_n)$ w.h.p.)
- Degree: choose a suitable one to balance bias and variance

“Smooth” regime: bias corrected “plug-in”

Bias correction based on Taylor expansion:

$$\mathbb{E} I(\theta) \approx \mathbb{E} \sum_{k=0}^{r} \frac{I^{(k)}(\hat{\theta}_n)}{k!} (\theta - \hat{\theta}_n)^k$$

Can we find an unbiased estimator for the RHS?

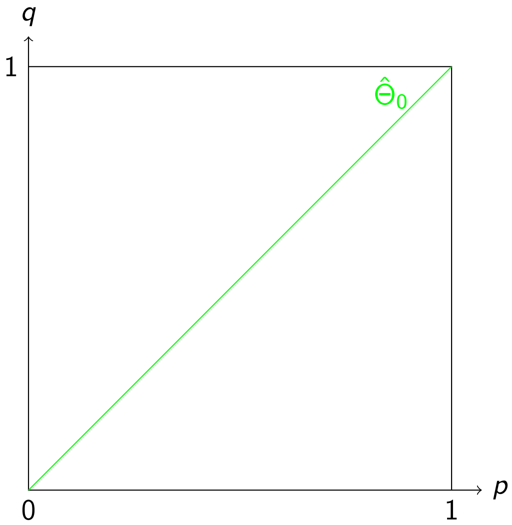
- Solution: sample splitting to obtain independent samples $\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}$
- Use the following estimator:

$$T(\hat{\theta}_n) = \sum_{k=0}^{r} \frac{I^{(k)}(\hat{\theta}_n^{(1)})}{k!} \sum_{j=0}^{k} \binom{k}{j} S_j(\hat{\theta}_n^{(2)})(-\hat{\theta}_n^{(1)})^{k-j}$$

where $S_j(\cdot)$ is an unbiased estimator of θ^j, i.e., $\mathbb{E} S_j(\hat{\theta}_n^{(2)}) = \theta^j$.

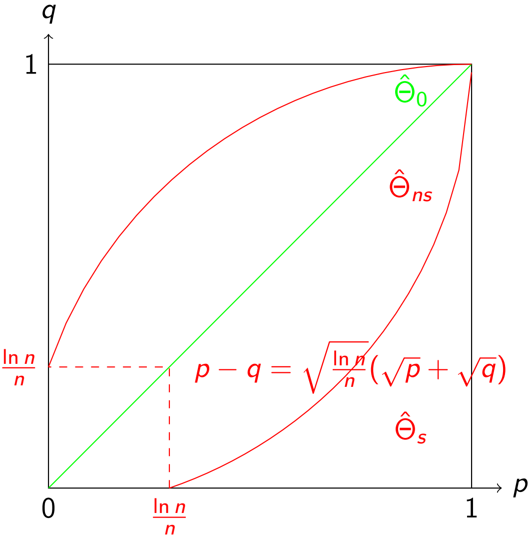
Estimator of ℓ_1 distance

$I(x, y) = |x - y|$, non-analytic regime $\hat{\Theta}_0 = \{(x, y) : x = y \in [0, 1]\}$



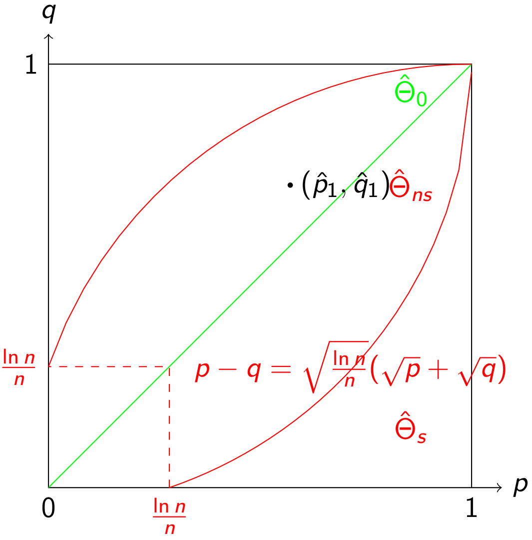
Estimator of ℓ_1 distance

$I(x, y) = |x - y|$, non-analytic regime $\hat{\Theta}_0 = \{(x, y) : x = y \in [0, 1]\}$



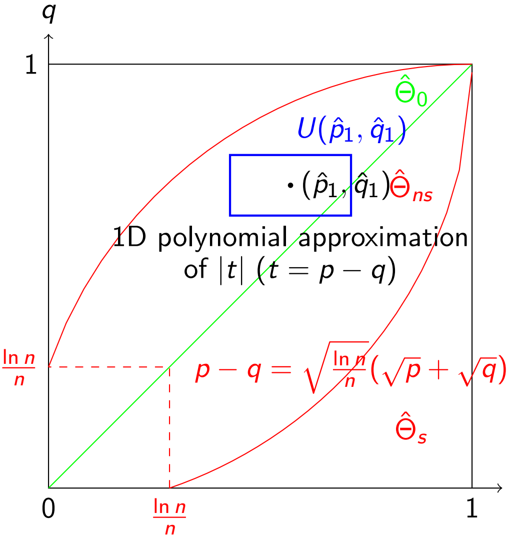
Estimator of ℓ_1 distance

$I(x,y) = |x-y|$, non-analytic regime $\hat{\Theta}_0 = \{(x,y) : x = y \in [0,1]\}$



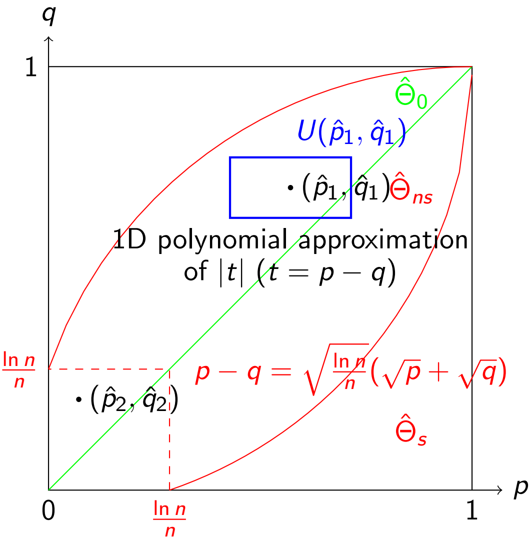
Estimator of ℓ_1 distance

$I(x, y) = |x - y|$, non-analytic regime $\hat{\Theta}_0 = \{(x, y) : x = y \in [0, 1]\}$



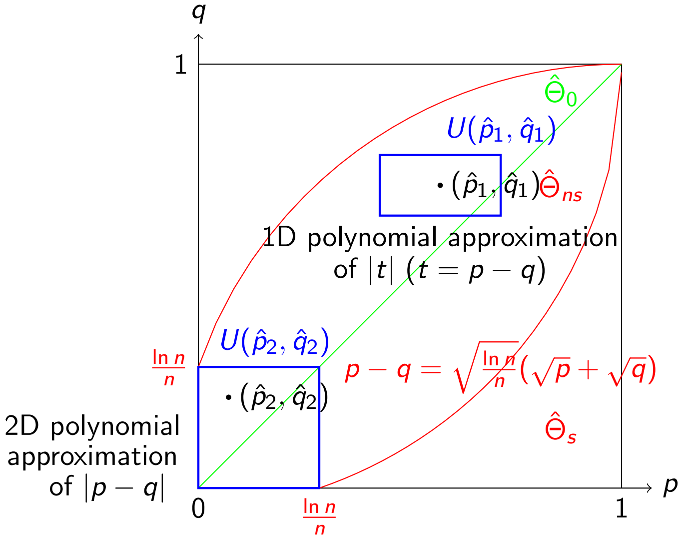
Estimator of ℓ_1 distance

$I(x,y) = |x-y|$, non-analytic regime $\hat{\Theta}_0 = \{(x,y) : x = y \in [0,1]\}$



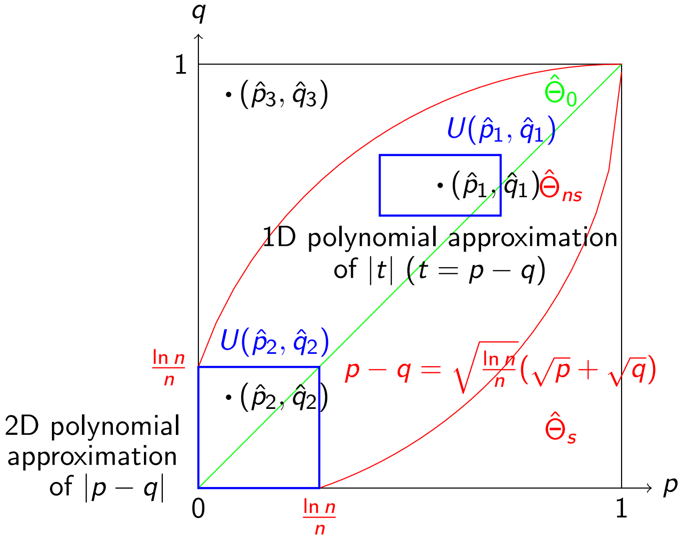
Estimator of ℓ_1 distance

$I(x, y) = |x - y|$, non-analytic regime $\hat{\Theta}_0 = \{(x, y) : x = y \in [0, 1]\}$



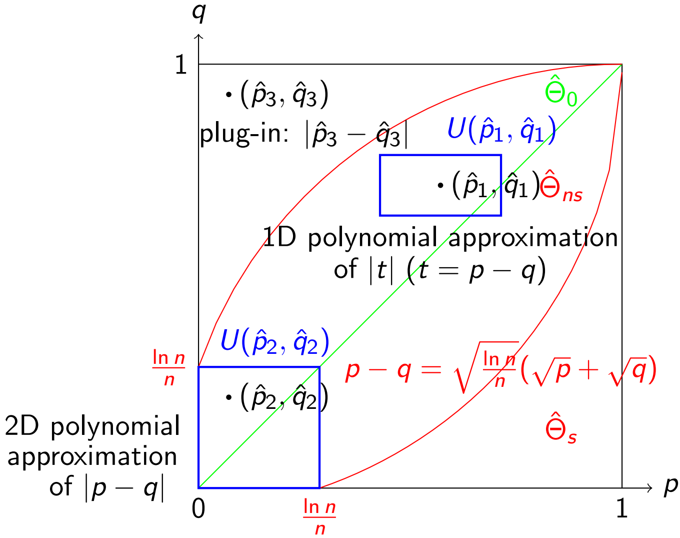
Estimator of ℓ_1 distance

$I(x,y) = |x-y|$, non-analytic regime $\hat{\Theta}_0 = \{(x,y) : x = y \in [0,1]\}$



Estimator of ℓ_1 distance

$I(x, y) = |x - y|$, non-analytic regime $\hat{\Theta}_0 = \{(x, y) : x = y \in [0, 1]\}$



Performance analysis

Let the approximation degree be K, our estimator $\hat{T}$ satisfies

$$\mathbb{E}_{(P,Q)}(\hat{T} - \|P - Q\|_1)^2 \lesssim \frac{S \ln n}{nK^2} + e^{cK} \cdot \frac{SK^2(\ln n)^2}{n^2}$$

Choosing $K \asymp \ln n$, we obtain

Theorem (Optimal estimator for ℓ_1 distance)

The minimax risk in estimating ℓ_1 distance is given by

$$\inf_{\hat{T}} \sup_{P,Q \in \mathcal{M}_S} \mathbb{E}_{(P,Q)}(\hat{T} - \|P - Q\|_1)^2 \asymp \frac{S}{n \ln n}$$

Effective sample size enlargement:

Theorem (Empirical estimator for ℓ_1 distance)

The maximum risk of the empirical estimator is given by

$$\sup_{P,Q \in \mathcal{M}_S} \mathbb{E}_{(P,Q)}(\|P_n - Q_n\|_1 - \|P - Q\|_1)^2 \asymp \frac{S}{n}$$

Some remarks

Additional remarks:

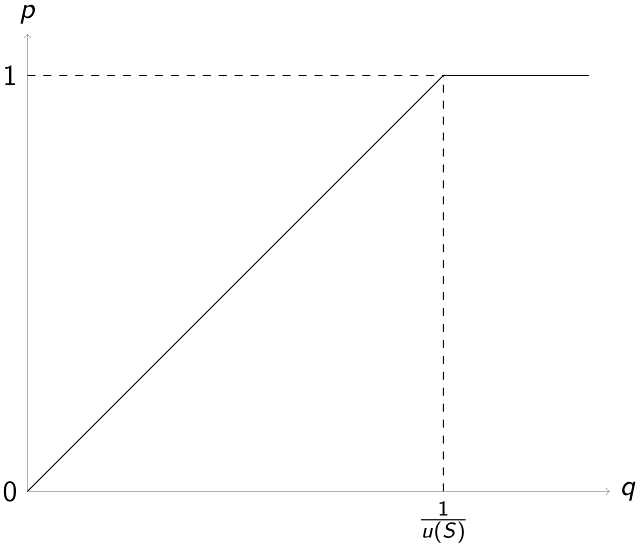
- For large $(\hat{p}, \hat{q})$ in the non-smooth regime, approximating over the whole stripe fails to give the optimal risk
- For small $(\hat{p}, \hat{q})$ in the non-smooth regime, best 2D polynomial approximation is not unique and not all can work:
  - Any 1D polynomial (i.e., $P(x, y) = p(x - y)$) cannot work!
  - We use the decomposition

    $$|x - y| = (\sqrt{x} + \sqrt{y})|\sqrt{x} - \sqrt{y}|$$

    and approximate two terms separately.
  - Still open in general.
- Valiant and Valiant'11 obtains the correct sample complexity $n \gg \frac{S}{\ln S}$, but suboptimal in the convergence rate

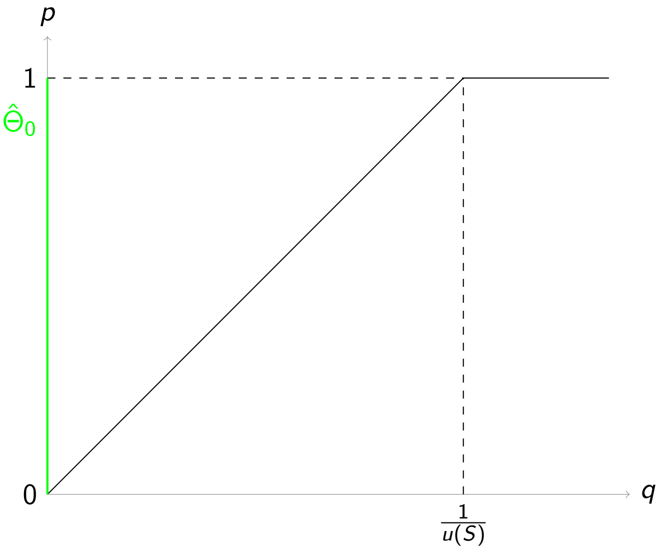
Estimator for KL divergence

$l(p,q) = p \ln q$, $\Theta = \{(p,q) \in [0,1]^2 : p \leq u(S)q\} \subset \hat{\Theta} = [0,1]^2$



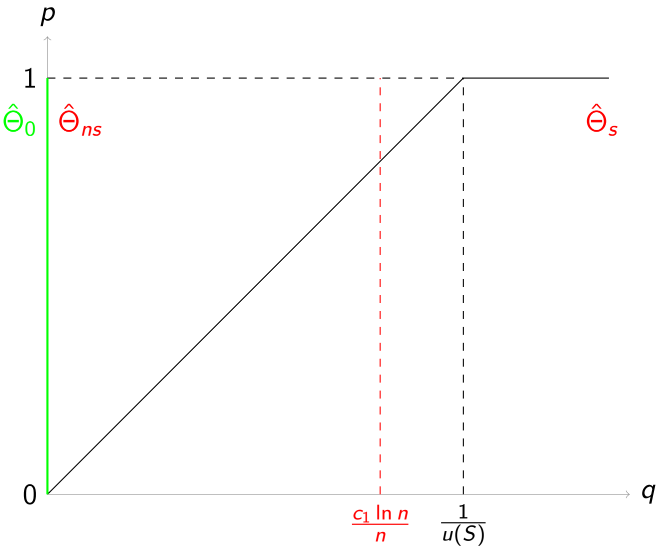
Estimator for KL divergence

$l(p,q) = p \ln q$, $\Theta = \{(p,q) \in [0,1]^2 : p \leq u(S)q\} \subset \hat{\Theta} = [0,1]^2$



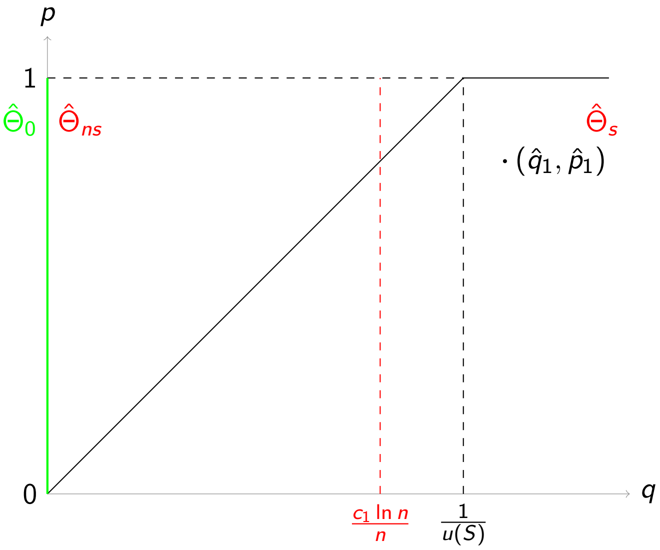
Estimator for KL divergence

$l(p,q) = p \ln q$, $\Theta = \{(p,q) \in [0,1]^2 : p \le u(S)q\} \subset \hat{\Theta} = [0,1]^2$



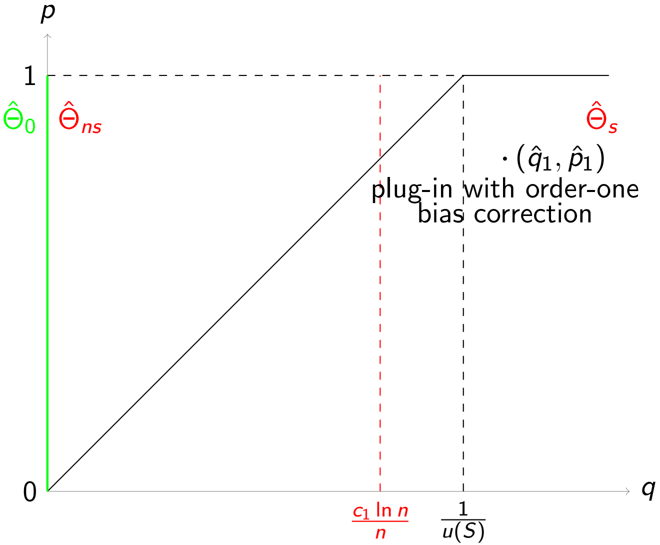
Estimator for KL divergence

$I(p,q) = p \ln q$, $\Theta = \{(p,q) \in [0,1]^2 : p \le u(S)q\} \subset \hat{\Theta} = [0,1]^2$



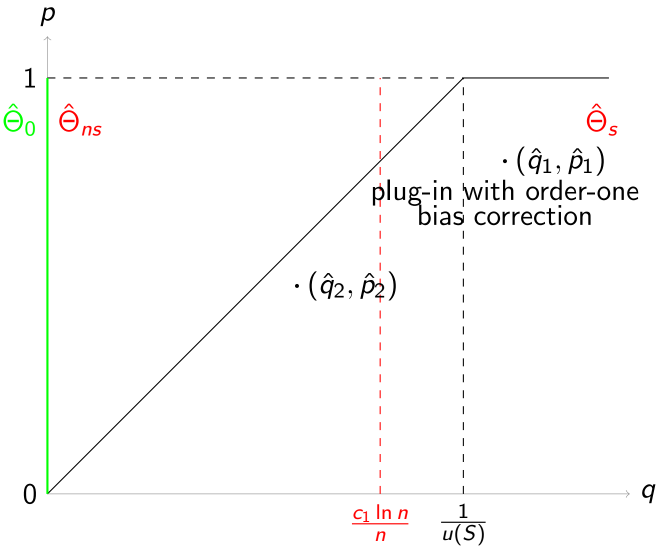
Estimator for KL divergence

$I(p,q) = p \ln q$, $\Theta = \{(p,q) \in [0,1]^2 : p \le u(S) q\} \subset \hat{\Theta} = [0,1]^2$



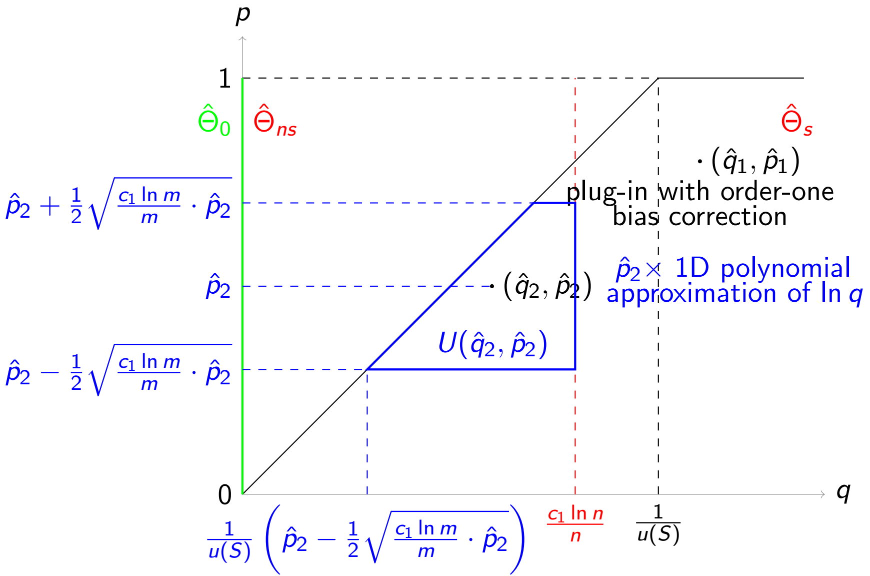
Estimator for KL divergence

$I(p,q) = p \ln q$, $\Theta = \{(p,q) \in [0,1]^2 : p \le u(S) q\} \subset \hat{\Theta} = [0,1]^2$



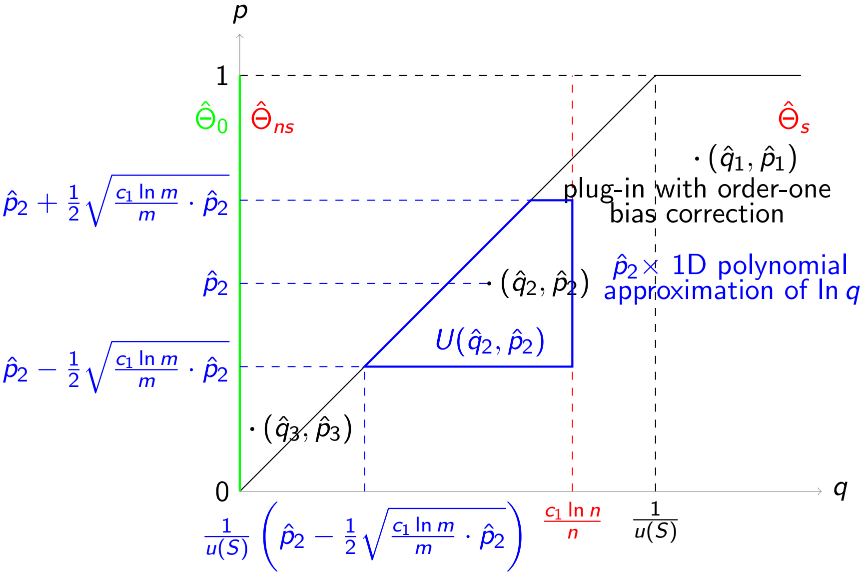
Estimator for KL divergence

$$I(p,q) = p \ln q,\ \Theta = \{(p,q) \in [0,1]^2 : p \le u(S)q\} \subset \hat{\Theta} = [0,1]^2$$



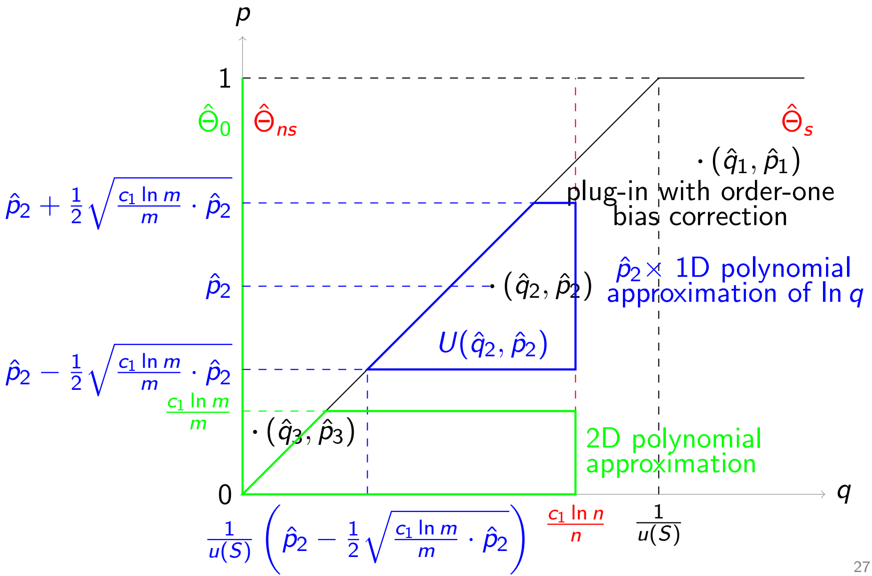
Estimator for KL divergence

$$l(p,q) = p \ln q,\ \Theta = \{(p,q) \in [0,1]^2 : p \le u(S)q\} \subset \hat{\Theta} = [0,1]^2$$



Estimator for KL divergence

$I(p,q) = p \ln q$, $\Theta = \{(p,q) \in [0,1]^2 : p \le u(S)q\} \subset \hat{\Theta} = [0,1]^2$



Some remarks

Additional remarks:

- Best polynomial approximation over general polytopes have not been solved until very recently!
- Room for improvement: use a single polynomial $P(x, y)$ to approximate $x \ln y$ whenever $y \leq \frac{c_1 \ln n}{n}$, where $P(x, y) = xq(y)$, and

$$yq(y) + C = \arg \min_{p \in \mathrm{Poly}_K} \max_{z \in [0, \frac{c_1 \ln n}{n}]} |z \ln z - p(z)|$$

Performance analysis

Theorem (Optimal estimator for KL divergence)

If $m \gtrsim \frac{S}{\ln S}$, $n \gtrsim \frac{Su(S)}{\ln S}$ *and* $u(S) \gtrsim (\ln S)^2$, *we have*

$$\inf_{\hat{T}} \sup_{P,Q \in \mathcal{M}_{S,u(S)}} \mathbb{E}_{P,Q}(\hat{T} - D(P\|Q))^2 \asymp \left(\frac{S}{m \ln m} + \frac{Su(S)}{n \ln n}\right)^2 + \frac{(\ln u(S))^2}{m} + \frac{u(S)}{n}$$

and our estimator attains the upper bound and does not require the knowledge of S *nor* $u(S)$.

The empirical estimator $D(P_m\|Q_n')$ with $Q_n' = \max\{n^{-1}, Q_n\}$:

Theorem (Empirical estimator for KL divergence)

The empirical estimator satisfies

$$\sup_{P,Q \in \mathcal{M}_{S,u(S)}} \mathbb{E}_{P,Q}(D(P_m\|Q_n') - D(P\|Q))^2 \asymp \left(\frac{S}{m} + \frac{Su(S)}{n}\right)^2 + \frac{(\ln u(S))^2}{m} + \frac{u(S)}{n}$$

Summary: the refined general recipe

Let $\{U(x)\}_{x \in \Theta'}$ be a satisfactory localization.

1. **Classify the Regime**:
   - For the true parameter θ, declare that θ is in the "non-smooth" regime if θ is "close" enough to Θ_0' in terms of localization. Otherwise declare θ is in the "smooth" regime;
   - Compute $\hat{\theta}_n$, and declare that we are in the "non-smooth" regime if the localization of $\hat{\theta}_n$ falls into the "non-smooth" regime of θ. Otherwise declare we are in the "smooth" regime;
2. **Estimate**:
   - If $\hat{\theta}_n$ falls in the "smooth" regime, use an estimator "similar" to $F(\hat{\theta}_n)$ to estimate $F(\theta)$;
   - If $\hat{\theta}_n$ falls in the "non-smooth" regime, replace the functional $F(\theta)$ in the "non-smooth" regime by an approximation $F_{\text{appr}}(\theta)$ (another functional which well approximates $F(\theta)$ on $U(\hat{\theta}_n)$) which can be estimated without bias, then apply an unbiased estimator for the functional $F_{\text{appr}}(\theta)$.

The problem

In the Gaussian white noise model

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}} dB_t, \qquad t \in [0,1]$$

with $f \in \mathcal{H}^s(L)$, we would like to estimate the following functional in L_2 risk:

$$\|f\|_r \triangleq \left(\int_0^1 |f(t)|^r dt \right)^{\frac{1}{r}}.$$

Hölder Ball

$f \in C[0,1]$ belongs to the Hölder ball $\mathcal{H}^s(L)$ with $s = m + r > 0$, $m \in \mathbb{N}, r \in (0,1]$, if and only if

$$\sup_{0 \le x < y \le 1} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x-y|^r} \le L$$

Equivalence between nonparametric models

Under certain smoothness conditions ($s > 1/2$ for Hölder balls), Brown et al. proved the asymptotic equivalence between the following models:

- Gaussian white noise model:

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}} dB_t, \qquad t \in [0, 1]$$

- Regression model: for iid $\mathcal{N}(0, 1)$ noise $\{\xi_i\}_{i=1}^n$,

$$y_i = f(i/n) + \sigma\xi_i, \qquad i = 1, 2, \cdots, n$$

- Poisson process: generate $N = \text{Poi}(n)$ iid samples from common density g ($g = f^2, \sigma = 1/2$)
- Density estimation model: generate n iid samples from common density g ($g = f^2, \sigma = 1/2$)

Lepski's result

Theorem (Lepski's result on L_r norm)

For even r, we have

$$\left(\inf_{\hat{T}} \sup_{f \in \mathcal{H}^s(L)} \mathbb{E}_f \left(\hat{T} - \|f\|_r \right)^2 \right)^{\frac{1}{2}} \asymp n^{-\frac{s}{2s+1-1/r}}$$

while for non-even r, we have the lower bound

$$\left(\inf_{\hat{T}} \sup_{f \in \mathcal{H}^s(L)} \mathbb{E}_f \left(\hat{T} - \|f\|_r \right)^2 \right)^{\frac{1}{2}} \gtrsim \frac{(n \ln n)^{-\frac{s}{2s+1}}}{(\ln n)^r}$$

and for $r = 1$, we have the upper bound

$$\left(\inf_{\hat{T}} \sup_{f \in \mathcal{H}^s(L)} \mathbb{E}_f \left(\hat{T} - \|f\|_1 \right)^2 \right)^{\frac{1}{2}} \lesssim (n \ln n)^{-\frac{s}{2s+1}}$$

The Besov ball setting

Instead of the Hölder ball $\mathcal{H}^s(L)$, we use the following Besov ball (generalized Lipschitz class)

$$\mathcal{B}^s_{p,\infty}(L) \triangleq \{f \in L^p[0,1] : |f|_{\mathcal{B}^s_{p,\infty}} \leq L\}$$

with $1 \leq p < \infty$. Properties:

- $\mathcal{B}^s_{p,\infty} \supset \mathcal{B}^s_{p',\infty}$ for $p < p'$
- $\mathcal{B}^s_{\infty,\infty} = \mathcal{H}^s$ for non-integer s

Intuition of Besov ball

$f \in \mathcal{B}^s_{p,\infty}$ if and only if $\|f^{(s)}\|_p < \infty$.

Main result for L_1 norm

Theorem (Minimax risk for estimating L_1 norm)

For any $s > 0$ *and* $1 \leq p < \infty$, *we have*

$$\left(\inf_{\hat{T}} \sup_{f \in \mathcal{B}^s_{p,\infty}(L)} \mathbb{E}_f \left(\hat{T} - \int_0^1 |f(t)| dt \right)^2 \right)^{\frac{1}{2}} \asymp (n \ln n)^{-\frac{s}{2s+1}}$$

Natural estimator for f: the kernel estimate for $s \le 1$

If $f \in \mathcal{H}^s(L)$ with $0 < s \le 1$, consider the simple averaging (rectangle window) with bandwidth $2h$:

$$\tilde{f}_h(x) = \frac{1}{2h} \int_{x-h}^{x+h} dY_t$$

Bias-variance analysis:

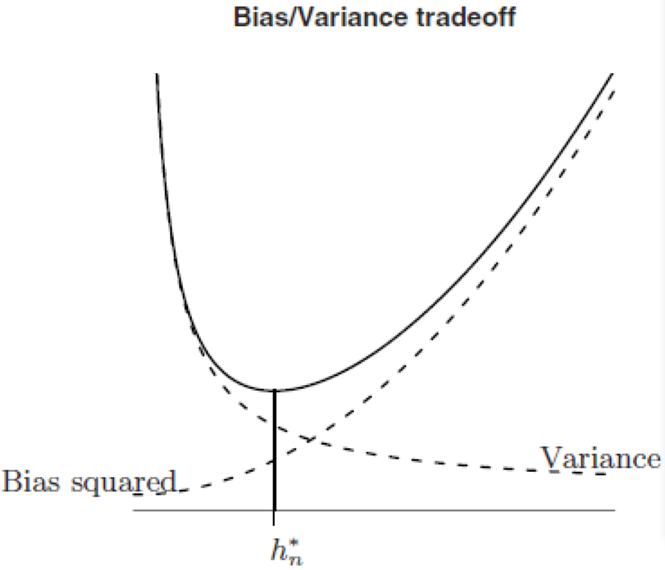
- Bias: $|\mathbb{E}\tilde{f}_h(x) - f(x)| = |\frac{1}{2h} \int_{x-h}^{x+h} (f(t) - f(x))|dt \le Lh^s$
- Variance: $\mathrm{Var}(\tilde{f}_h(x)) = \frac{1}{4h^2} \int_{x-h}^{x+h} \frac{1}{n} dt = \frac{1}{2nh}$
- Optimal bandwidth: $h \asymp n^{-\frac{1}{2s+1}}$

General Besov ball

For general Besov ball $f \in \mathcal{B}^s_{p,\infty}(L)$, the wavelet basis is the optimal basis (attains the Kolmogorov n-width), and the associated kernel K_h with bandwidth h satisfies

$$\|f - K_h f\|_p \lesssim Lh^s$$

Bias-variance tradeoff



First-stage approximation: approximation of function

Consider a kernel estimate of f with bandwidth h:

$$f_h(t) = \int_0^1 \frac{1}{h} K(\frac{t-u}{h}) \cdot f(u) du$$

$$\tilde{f}_h(t) = \int_0^1 \frac{1}{h} K(\frac{t-u}{h}) \cdot dY_u$$

- For suitable kernel, we have $|f(t) - f_h(t)| \lesssim h^s$.
- The observation model becomes

$$\tilde{f}_h(t) = f_h(t) + \lambda_h \xi_h(t)$$

  with $\lambda_h \asymp \frac{1}{\sqrt{nh}}$ and $\xi_h(t) \sim \mathcal{N}(0,1)$.
- $\xi_h(s)$ and $\xi_h(t)$ are independent whenever $|s-t| > h$.

Idea: estimate $\|f_h\|_1$ instead of $\|f\|_1$, i.e., approximate $\|f\|_1$ by $\|f_h\|_1$

Second-stage approximation: approximation of functional

Second-stage approximation:

- Polynomial approximation of $|x|$ on $[-c_1\lambda_h\sqrt{\ln n}, c_1\lambda_h\sqrt{\ln n}]$:

$$|x| \approx \sum_{k=0}^{K} a_k x^k$$

- Let $Q(\tilde{f}_h(t))$ be the unbiased estimator of $\sum_{k=0}^{K} a_k f_h(t)^k$
- Split samples, and estimate $|f_h(t)|$ via

$$T(t) = Q(\tilde{f}_{h,1})\mathbb{1}(|\tilde{f}_{h,2}| \leq c_1\lambda_h\sqrt{\ln n}) + |\tilde{f}_{h,1}|\mathbb{1}(|\tilde{f}_{h,2}| > c_1\lambda_h\sqrt{\ln n})$$

Estimator construction:

$$\hat{T} = \int_0^1 T(t)dt$$

Error analysis

Three types of errors:

- Approximation error I: $|\|f\|_1 - \|f_h\|_1| \le \|f - f_h\|_1 \le \|f - f_h\|_p \lesssim h^s$
- Approximation error II (bias): the bias at a point corresponds to the polynomial approximation error, which is of order $\frac{\lambda_h \sqrt{\ln n}}{K}$. Hence, the integrated bias is upper bounded by

$$|\mathbb{E}\hat{T} - \|f_h\|_1| \lesssim \frac{1}{K}\sqrt{\frac{\ln n}{nh}}$$

- Variance: the standard deviation at a point is upper bounded by $\lambda_h \cdot \exp(cK)$. Since we have h^{-1} "independent samples", the total variance is upper bounded by

$$\sqrt{\mathrm{Var}(\hat{T})} \lesssim \sqrt{h} \cdot \lambda_h \cdot \exp(cK) \asymp n^{-\frac{1}{2}} e^{cK}$$

Choice of parameters: $h \asymp (n \ln n)^{-\frac{1}{2s+1}}, K \asymp \ln n$.

Fuzzy hypothesis testing

Suppose we want to estimate $T(\theta)$ with $\theta \in \Theta$ based on observation $\mathbf{X}$.

Lemma (Tsybakov'08)

Suppose there exist $\zeta \in \mathbb{R}, s > 0, 0 \le \beta_0, \beta_1 < 1$ *and two priors* σ_0, σ_1 *on* Θ *such that*

$$\sigma_0(\theta : T(\theta) \le \zeta - s) \ge 1 - \beta_0 \tag{1}$$

$$\sigma_1(\theta : T(\theta) \ge \zeta + s) \ge 1 - \beta_1. \tag{2}$$

If $TV(F_1, F_0) \le \eta < 1$*, then*

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(|\hat{T} - T(\theta)| \ge s \right) \ge \frac{1 - \eta - \beta_0 - \beta_1}{2}, \tag{3}$$

where $F_i, i = 0, 1$ *are the marginal distributions of* $\mathbf{X}$ *when the priors are* $\sigma_i, i = 0, 1$*, respectively.*

Reduction to parametric model

Fix some smooth g on $[0,1]$. Consider the parametric submodel with

$$f_{\boldsymbol{\theta}}(t) = L' \sum_{i=1}^{N} \theta_i \sqrt{\ln N} \cdot h^s g\left(\frac{t-(i-1)/N}{h}\right)$$

where $h \asymp (n \ln n)^{-\frac{1}{2s+1}}$ is the size of each subinterval, and $N = h^{-1}$.

- Functional value: $\|f_{\boldsymbol{\theta}}\|_1 \asymp h^s \sqrt{\ln N} \cdot \frac{1}{N} \sum_{i=1}^N |\theta_i|$
- Besov ball condition: $(\frac{1}{N} \sum_{i=1}^N |\theta_i|^p)^{\frac{1}{p}} \lesssim \frac{1}{\sqrt{\ln N}}$

Claim

It suffices to prove that in the Gaussian sequence model $y_i = \theta_i + \xi_i$, $i = 1, \cdots, N$, $\xi_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, we have

$$\inf_{\hat{T}} \sup_{\boldsymbol{\theta}:(\frac{1}{N}\sum_{i=1}^N |\theta_i|^p)^{\frac{1}{p}} \lesssim \frac{1}{\sqrt{\ln N}}} \mathbb{E}_{\boldsymbol{\theta}}(\hat{T} - \frac{1}{N}\sum_{i=1}^N |\theta_i|)^2 \gtrsim \frac{1}{\ln N}$$

Construction of two measures

These two measures σ_0, σ_1 should satisfy the following conditions:

- supported on $[-\sqrt{\ln N}, \sqrt{\ln N}]$ (measure concentration)
- large difference in functional value:

$$\int |t|\sigma_0(dt) - \int |t|\sigma_1(dt) \gtrsim \frac{1}{\sqrt{\ln N}}$$

- matching moments ($\Rightarrow$ small total variation distance):

$$\int t^l \sigma_0(dt) = \int t^l \sigma_1(dt), \qquad l = 0, 1, \cdots, c \ln N$$

- constrained moment:

$$\left(\int |t|^p \sigma_i(dt)\right)^{\frac{1}{p}} \lesssim \frac{1}{\sqrt{\ln N}}, \qquad i = 0, 1.$$

Dual: polynomial approximation

The key duality:

$$\sup_{\substack{\mu:\|\mu\|_{\mathsf{TV}}\le 1 \\ \int t^l \mu(dt)=0, l=0,\cdots,K}} \int f(t)\mu(dt) = \inf_{p\in \mathsf{Poly}_K} \|f-p\|_\infty$$

Claim

To construct such measures, it is sufficient (and also necessary) to prove that, for some integer $q \ge p/2$, we have

$$\inf_{\{a_k\}_{k=-q+1}^n} \sup_{x\in[cn^{-2},1]} \left| x^{-q+\frac{1}{2}} - \sum_{k=-q+1}^{n} a_k x^k \right| \gtrsim n^{2q-1}$$

Still a non-trivial question (involving approximation using x^k with $k<0$), but can be solved using approximation theory.

Our result on L_r norm estimation

Theorem (Main result on L_r norm estimation)

In Besov balls $\mathcal{B}^s_{p,\infty}(L)$ with $s > 0$ and $r \le p < \infty$, the minimax risk is given by

$$\left(\inf_{\hat{T}} \sup_{f \in \mathcal{B}^s_{p,\infty}(L)} \mathbb{E}_f \left(\hat{T} - \|f\|_r\right)^2\right)^{\frac{1}{2}} \asymp \begin{cases} n^{-\frac{s}{2s+1-1/r}} & r \text{ even} \\ (n \ln n)^{-\frac{s}{2s+1}} & r \text{ odd or non-integer} \end{cases}$$

The upper bound is attained by polynomial approximation.

- Note: if r is even, $|x|^r = x^r$ is itself a polynomial!

Summary

The general recipe in the nonparametric setting:

- Stage-one approximation: approximate $I(f)$ by $I(f_h)$, where we essentially have a parametric model
- Stage-two approximation: apply the approximation-based method in the parametric case to reduce bias
- Choose the optimal bandwidth h and approximation degree K

Discussions and questions

- What about the Hölder ball case ($p = \infty$)?
- Can our estimator be adaptive in smoothness parameter s? (Lepski's trick)
- Adaptive confidence interval in general L_r norm (Risk estimation)
- Other non-smooth functionals (e.g., differential entropy $\int -f(t) \ln f(t) dt$)

References

- J. Jiao, Y. Han and T. Weissman. "Minimax estimation of L_1 distance." *to appear in ISIT* (Best student paper finalist), 2016.
- Y. Han, J. Jiao and T. Weissman. "Minimax estimation of KL divergence between discrete distributions." *available on Arxiv*, 2016.
- Y. Han, J. Jiao, R. Mukherjee and T. Weissman. "Optimal estimation of L_r norm of functions in Gaussian white noise." *to be submitted*, 2016.

Thank you!

Email: yjhan@stanford.edu