Minimax Rate-optimal Estimation of KL Divergence between Discrete Distributions

Yanjun Han (Stanford EE)

Joint work with:

Jiantao Jiao	Stanford EE
Tsachy Weissman	Stanford EE

November 1, 2016

Problem: estimation of information divergence

Given jointly independent samples $X_1, \cdots, X_m \sim P$, $Y_1, \cdots, Y_n \sim Q$, we would like to estimate

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

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

where

- S is the *unknown* support size
- $\frac{p_i}{q_i} \leq u(S)$ is the *unknown* likelihood-ratio bound in the latter case

General 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 estimating non-smooth functionals:

Theorem (Entropy, Jiao, Venkat, H., Weissman'15)

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

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

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

Bias dominates when estimating non-smooth functionals:

Theorem (Entropy, Jiao, Venkat, H., Weissman'15)

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

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

Theorem (Entropy, Jiao, Venkat, H., Weissman'15, Wu and Yang'15)

For $X_1, \cdots, X_n \sim P = (p_1, \cdots, p_S)$ and $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$, if $n \gtrsim \frac{S}{\ln S}$,

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

Effective sample size enlargement

- In estimating functionals of a single distribution P, we have (Jiao, Venkat, H., Weissman'14, Wu and Yang'14, Jiao, H., Weissman'15)

	Minimax L_2 rate	L_2 rate of MLE
$H(P) = \sum_{i=1}^{S} -p_i \ln p_i$	$\frac{S^2}{(n \ln n)^2} + \frac{\ln^2 S}{n}$	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 0 < \alpha < 1/2$	$\frac{S^2}{(n \ln n)^{2\alpha}}$	$\frac{S^2}{n^{2\alpha}}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 1/2 \leq \alpha < 1$	$\frac{S^2}{(n \ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 1 < \alpha < 3/2$	$\frac{1}{(n \ln n)^{2(\alpha-1)}}$	$\frac{1}{n^{2(\alpha-1)}}$

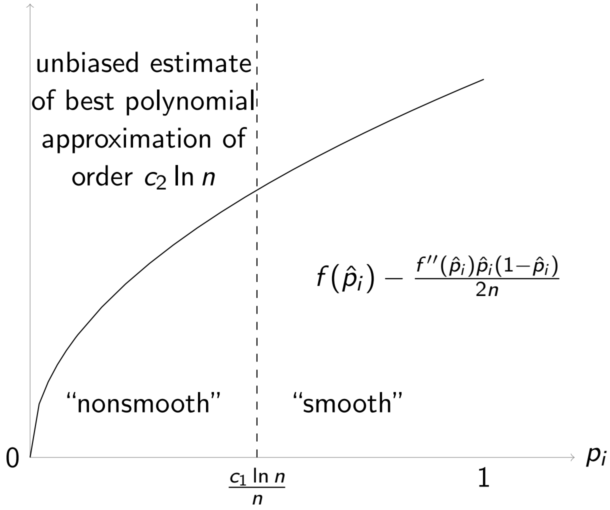
Effective Sample Size Enlargement

Minimax rate-optimal with n samples $\Longleftrightarrow$ Plug-in with $n \ln n$ samples

Similar results also hold for Rényi entropy estimation (Acharya, Orlitsky, Suresh, Tyagi'14), Hellinger divergence and χ^2-divergence estimation (H., Jiao, Weissman'16), L_r norm estimation under Gaussian white noise model (H., Jiao, Mukherjee, Weissman'16)

Optimal estimator for $\sum_{i=1}^{S} f(p_i)$

$f(p_i) = -p_i \ln p_i$ or p_i^α



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 $F_{\text{appr}}(\theta)$.

New challenges

1. Existing work: $I(\cdot)$ is only non-analytic at zero
2. L_1 distance and KL divergence:

$$I_1(p,q) = |p-q|, \qquad I_2(p,q) = p \ln \frac{p}{q}$$

- Bivariate function
- Non-analytic on a segment $p = q \in [0,1]$ or $q = 0, p \in [0,1]$
- $\Theta \neq \hat{\Theta}$ for KL divergence: $\hat{p} > u(S)\hat{q}$ may occur even if $p \leq u(S)q$

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?

Confidence set

Definition (Confidence set)

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 confidence set of significance level $r\in[0,1]$, is a collection of sets $\{U(x)\}_{x\in\hat{\Theta}}$, where $U(x)\subset\Theta$ for any $x\in\hat{\Theta}$, and

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

- Confidence set always exists, but we seek for a small one
- Choice of significance: $r\asymp n^{-A}$

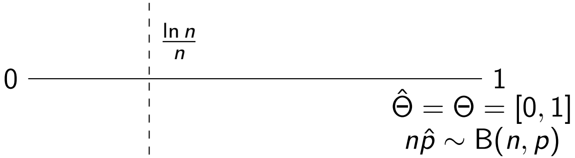
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$

0 ——————————————— 1

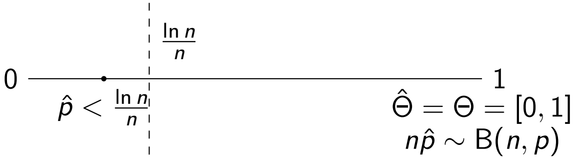
$\hat{\Theta} = \Theta = [0, 1]$

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

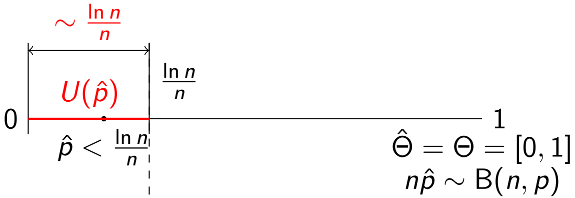
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



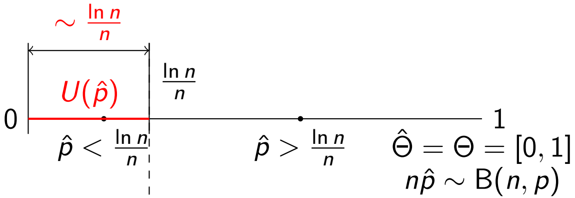
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



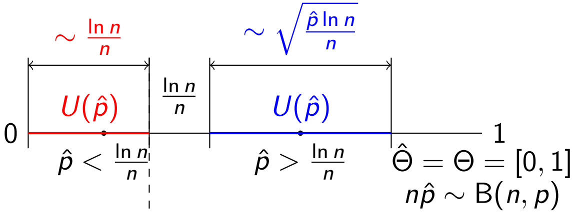
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



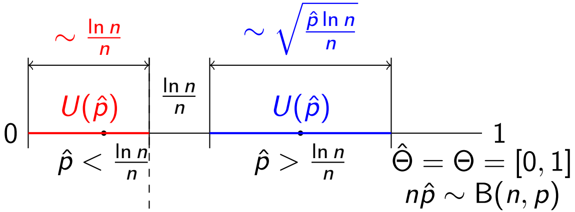
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



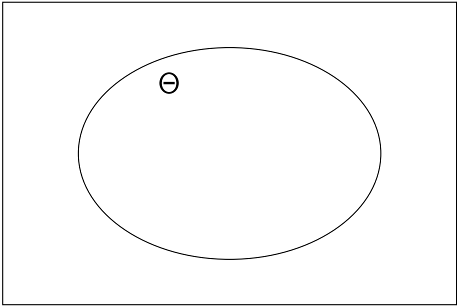
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



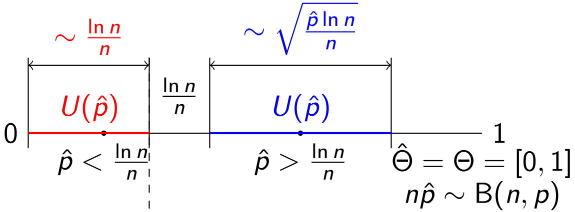
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



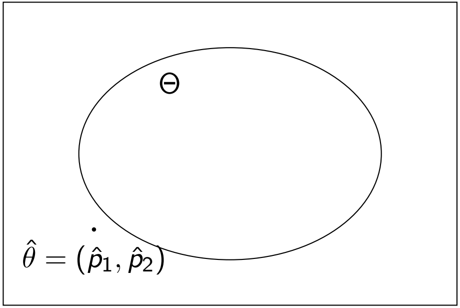
$$\hat{\Theta} = [0, 1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



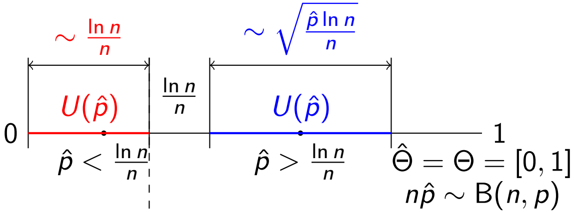
Confidence set in Binomial model: $r \asymp \min\{m,n\}^{-A}$



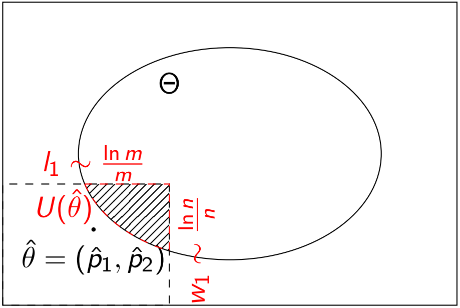
$$\hat{\Theta} = [0,1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m,p_1) \times \mathrm{B}(n,p_2)$$



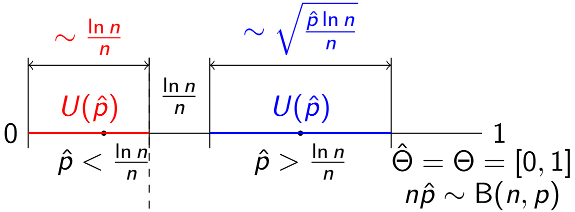
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



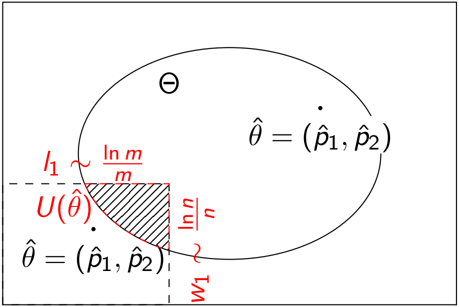
$$\hat{\Theta} = [0, 1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



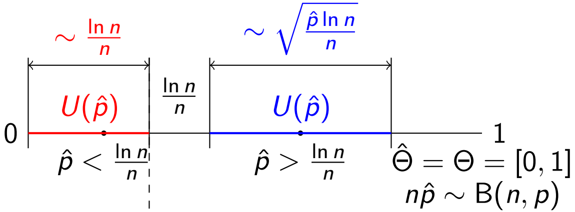
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



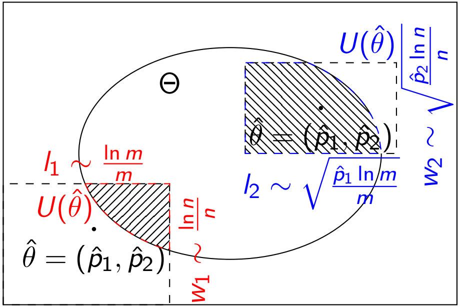
$$\hat{\Theta} = [0, 1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



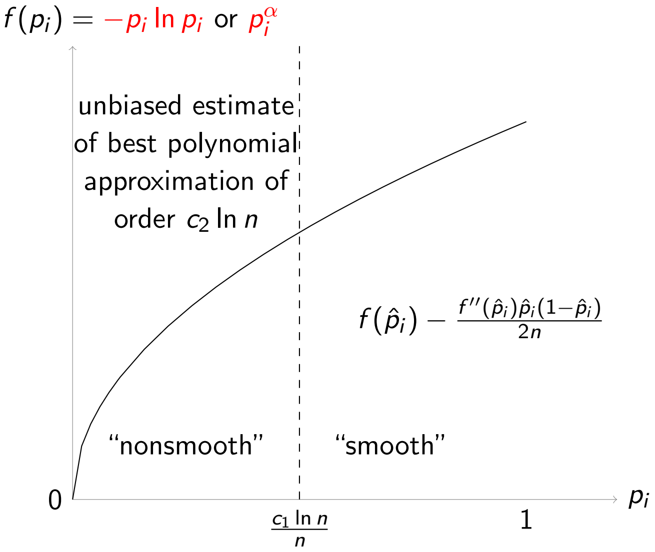
Confidence set in Binomial model: $r \asymp \min\{m, n\}^{-A}$



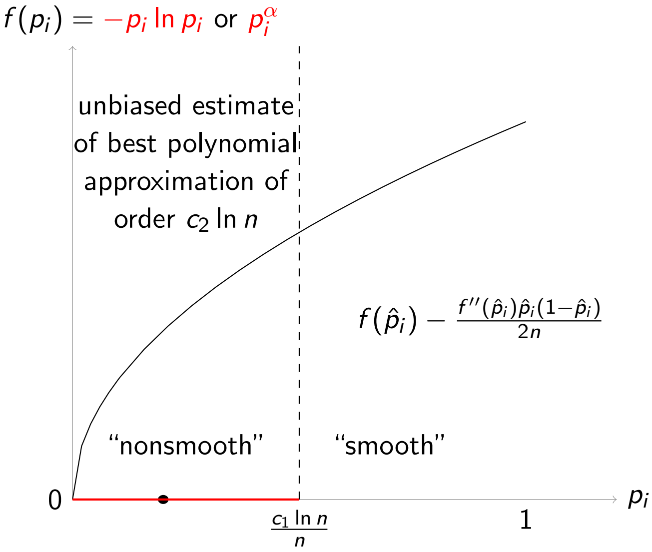
$$\hat{\Theta} = [0,1]^2 : (m\hat{p}_1, n\hat{p}_2) \sim \mathrm{B}(m, p_1) \times \mathrm{B}(n, p_2)$$



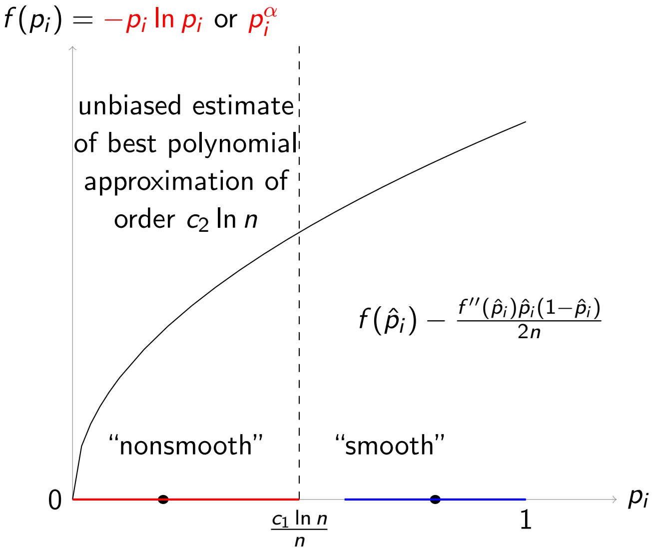
The role of confidence set: entropy estimation



The role of confidence set: entropy estimation



The role of confidence set: entropy estimation



Determine the “non-smooth” regime

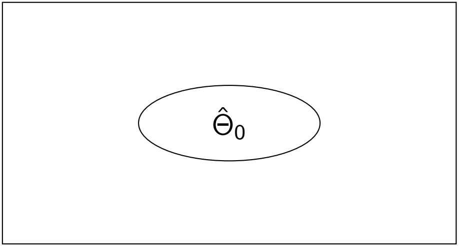
Plug-in works well when $\hat{\theta}_n \notin \hat{\Theta}_0$ (the non-analytic region of $I(\cdot)$)

The criteria

Given a suitable r-confidence set $U(\cdot)$, 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.



Determine the "non-smooth" regime

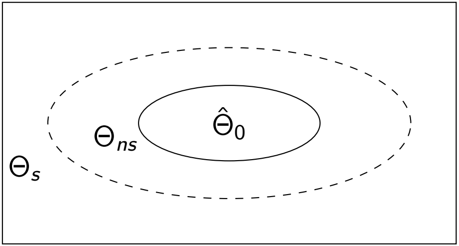
Plug-in works well when $\hat{\theta}_n \notin \hat{\Theta}_0$ (the non-analytic region of $I(\cdot)$)

The criteria

Given a suitable r-confidence set $U(\cdot)$, 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.



Determine the "non-smooth" regime

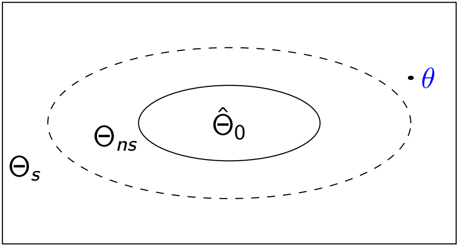
Plug-in works well when $\hat{\theta}_n \notin \hat{\Theta}_0$ (the non-analytic region of $I(\cdot)$)

The criteria

Given a suitable r-confidence set $U(\cdot)$, 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.



Determine the "non-smooth" regime

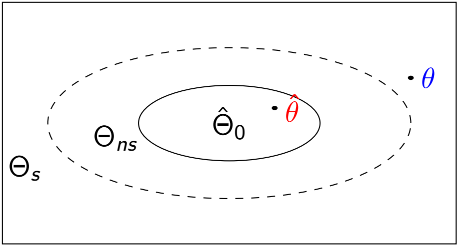
Plug-in works well when $\hat{\theta}_n \notin \hat{\Theta}_0$ (the non-analytic region of $I(\cdot)$)

The criteria

Given a suitable r-confidence set $U(\cdot)$, 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.



Determine the "non-smooth" regime

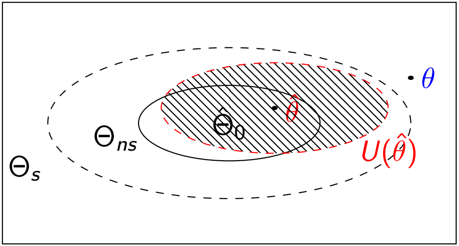
Plug-in works well when $\hat{\theta}_n \notin \hat{\Theta}_0$ (the non-analytic region of $I(\cdot)$)

The criteria

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

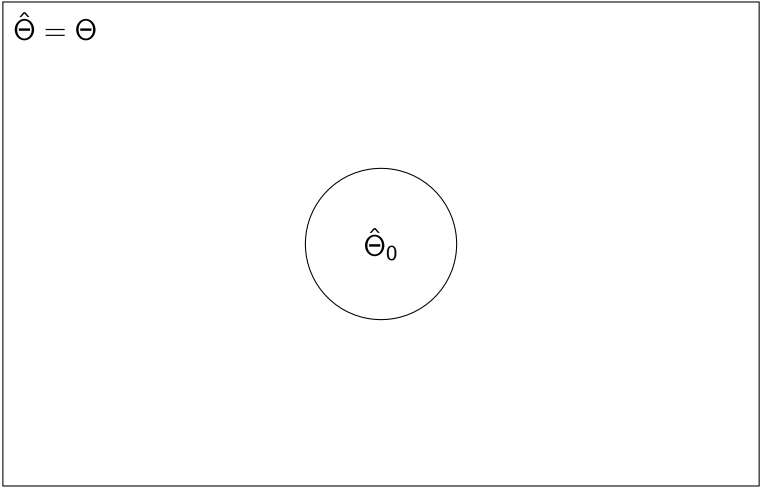
$$\theta \in \cup_{\hat{\theta} \in \hat{\Theta}_0} U(\hat{\theta})$$

and in the "smooth" regime Θ_s otherwise.



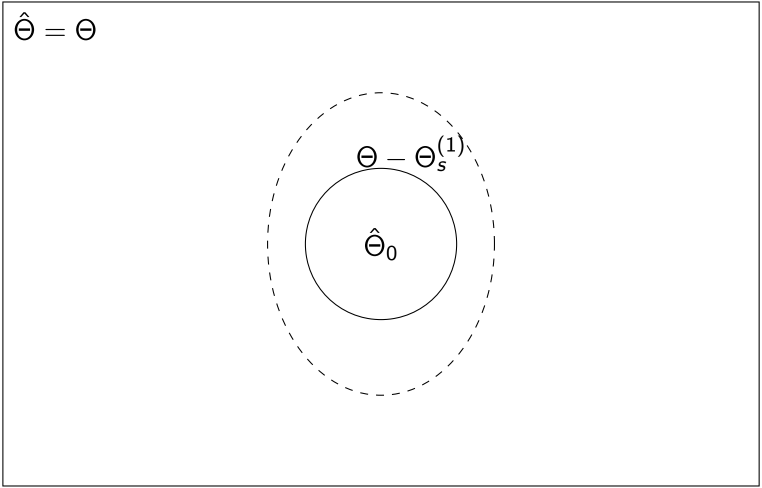
There is something more...

However, we cannot make decisions based on unknown θ!



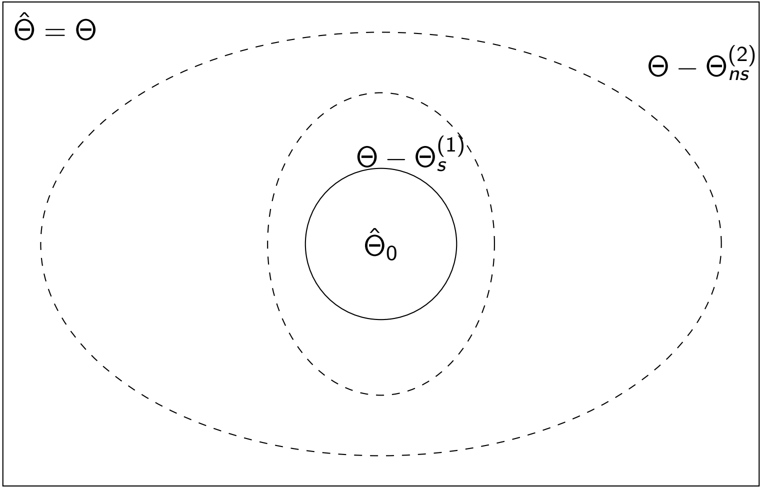
There is something more...

However, we cannot make decisions based on unknown θ!



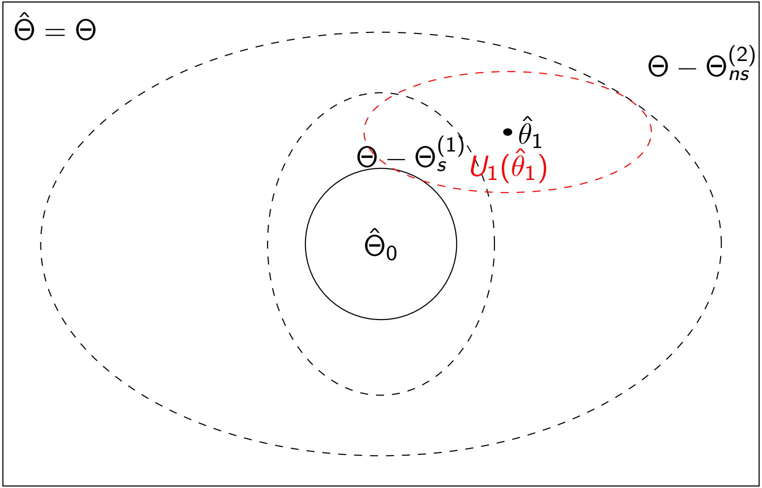
There is something more...

However, we cannot make decisions based on unknown θ!



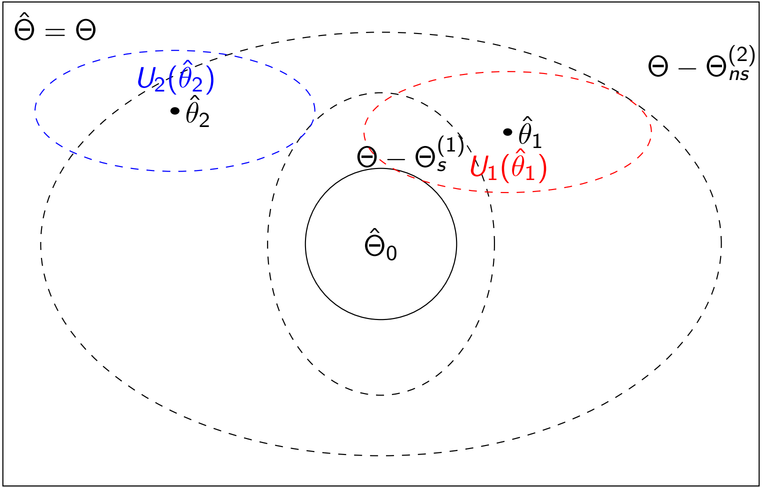
There is something more...

However, we cannot make decisions based on unknown θ!



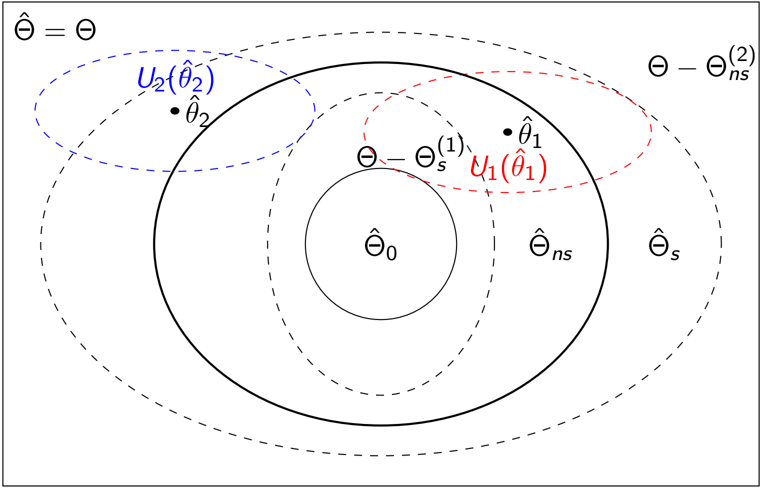
There is something more...

However, we cannot make decisions based on unknown θ!



There is something more...

However, we cannot make decisions based on unknown θ!



Different regimes: approximation and "plug-in"

"Non-smooth" regime: find an approximate functional $I_{\text{appr}}(\theta) \approx I(\theta)$:

- Type: polynomial (admits unbiased estimators)
- Region: confidence set $U(\hat{\theta}_n)$
- Degree: balance bias and variance

Different regimes: approximation and "plug-in"

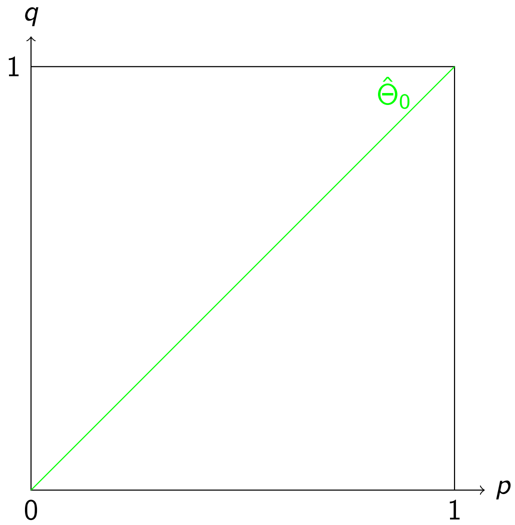
"Non-smooth" regime: find an approximate functional $I_{\mathrm{appr}}(\theta) \approx I(\theta)$:

- Type: polynomial (admits unbiased estimators)
- Region: confidence set $U(\hat{\theta}_n)$
- Degree: balance bias and variance

"Smooth" regime: Taylor-based bias-correction up to any order

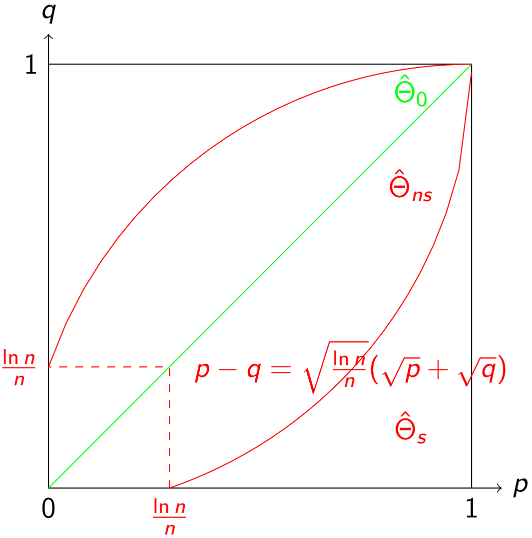
Estimator of ℓ_1 distance

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



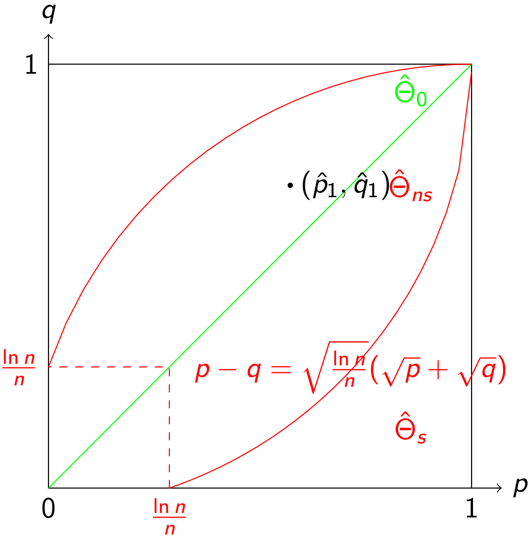
Estimator of ℓ_1 distance

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



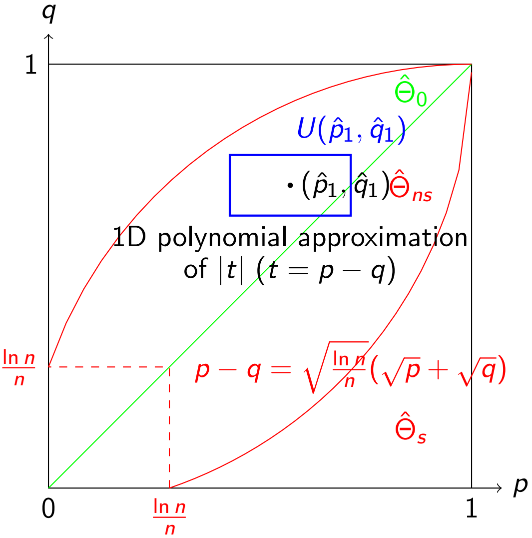
Estimator of ℓ_1 distance

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



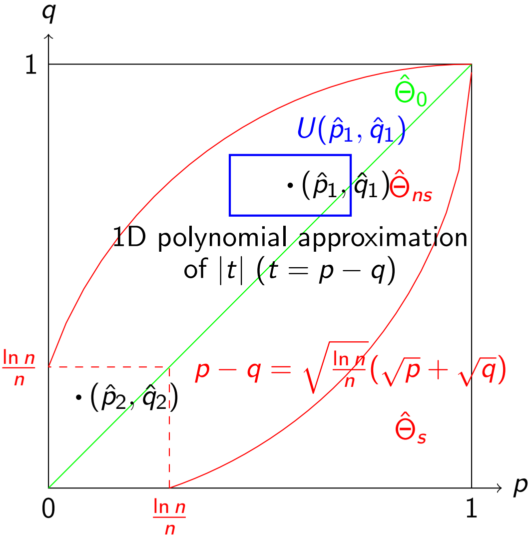
Estimator of ℓ_1 distance

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



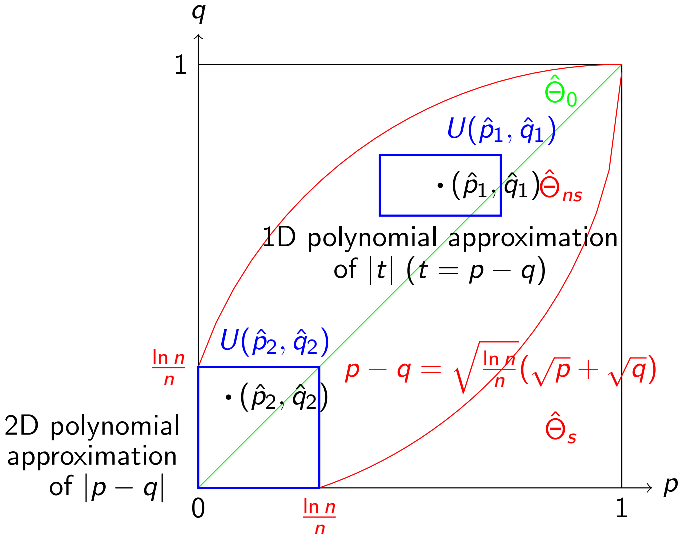
Estimator of ℓ_1 distance

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



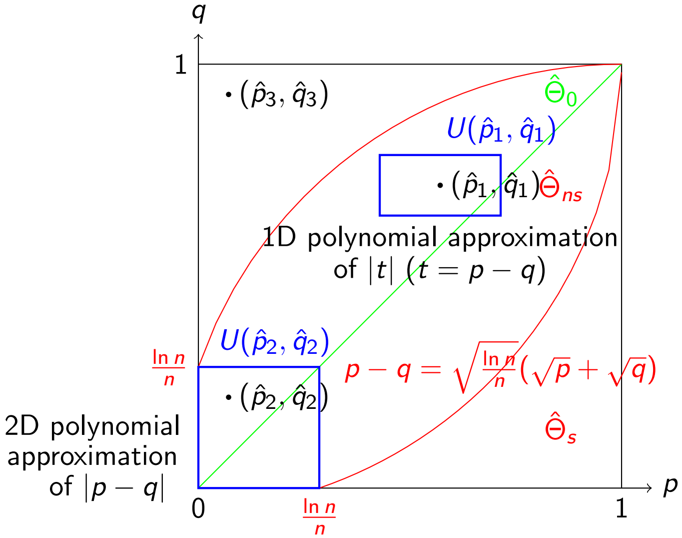
Estimator of ℓ_1 distance

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



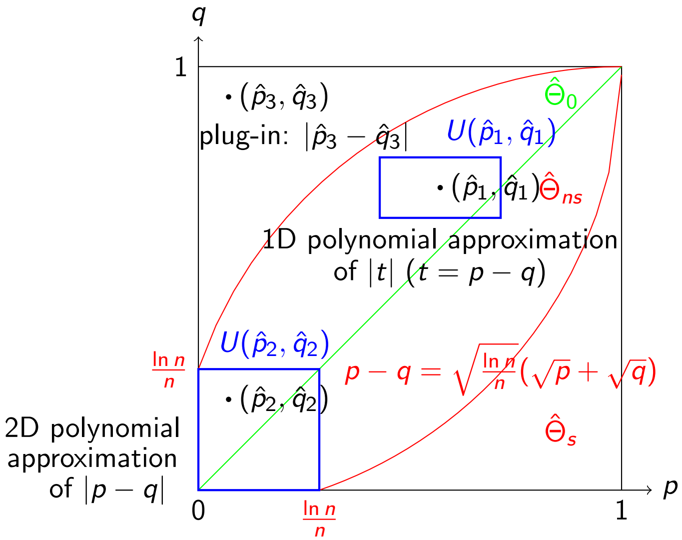
Estimator of ℓ_1 distance

$I(x, y) = |x - y|$, 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

Theorem (Optimal estimator for ℓ_1 distance, Jiao, H., Weissman'16)

The minimax risk in estimating ℓ_1 distance is given by

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

and the previous estimator achieves the upper bound without the knowledge of S.

Effective sample size enlargement:

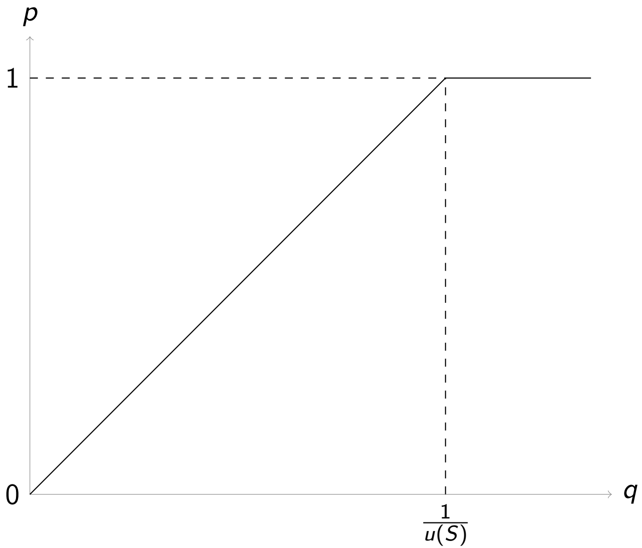
Theorem (Empirical estimator for ℓ_1 distance, Jiao, H., Weissman'16)

The maximum risk of the empirical estimator is given by

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

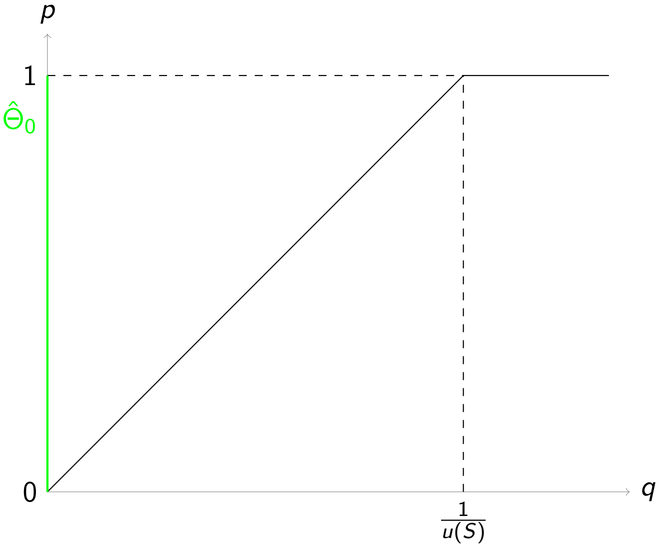
Estimator for KL divergence

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



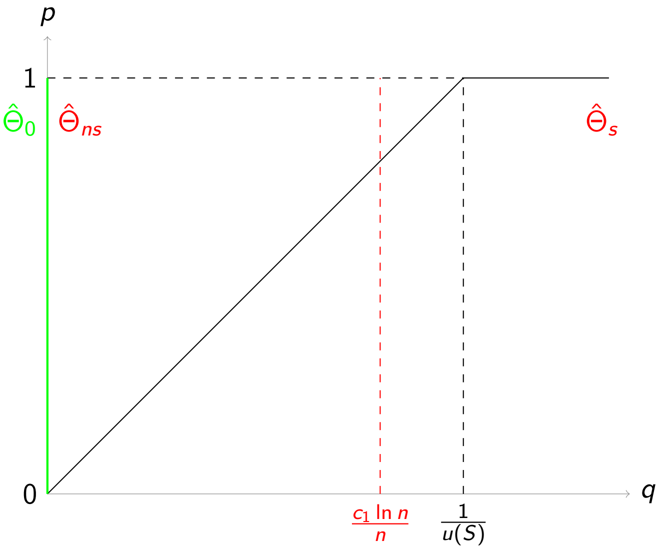
Estimator for KL divergence

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



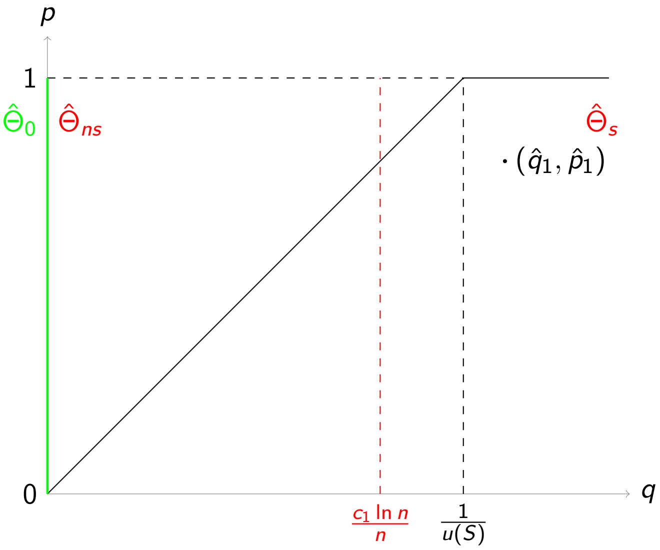
Estimator for KL divergence

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



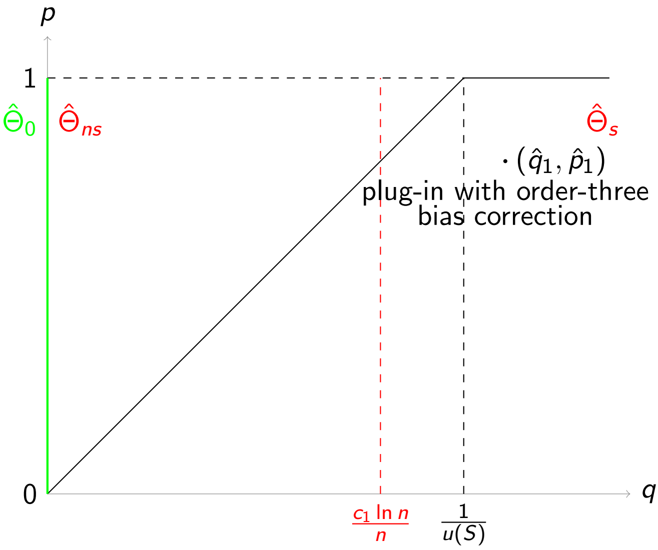
Estimator for KL divergence

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



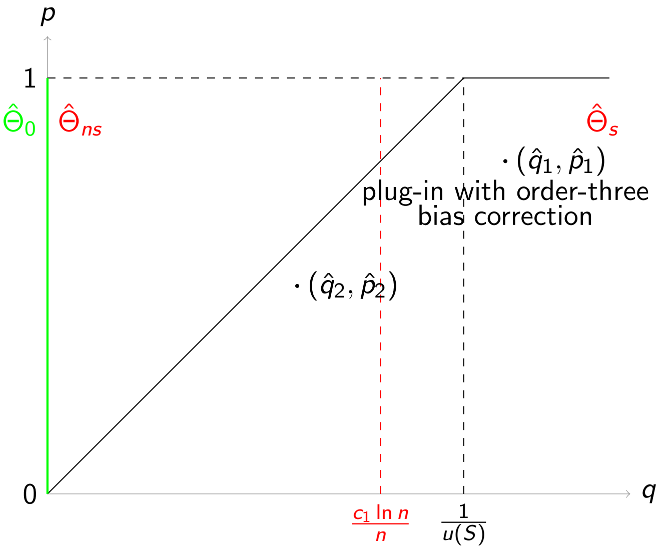
Estimator for KL divergence

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



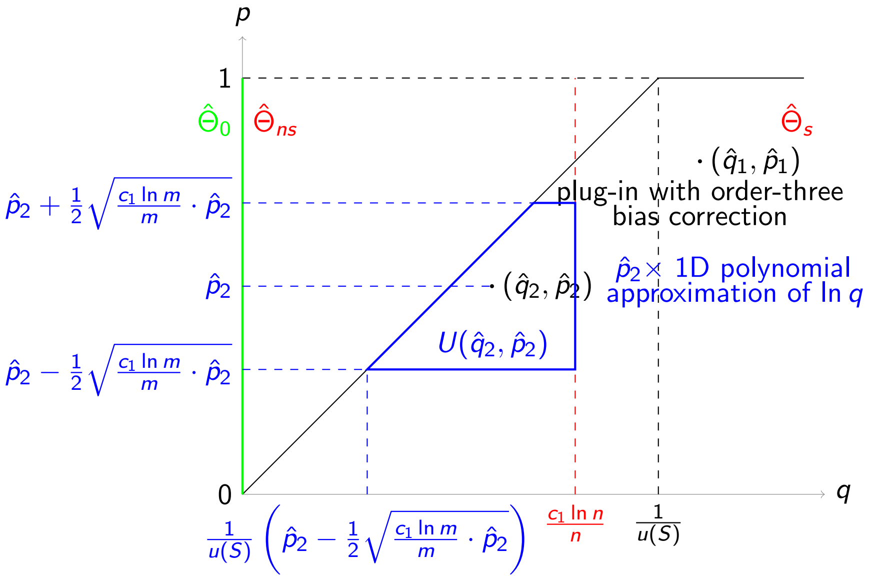
Estimator for KL divergence

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



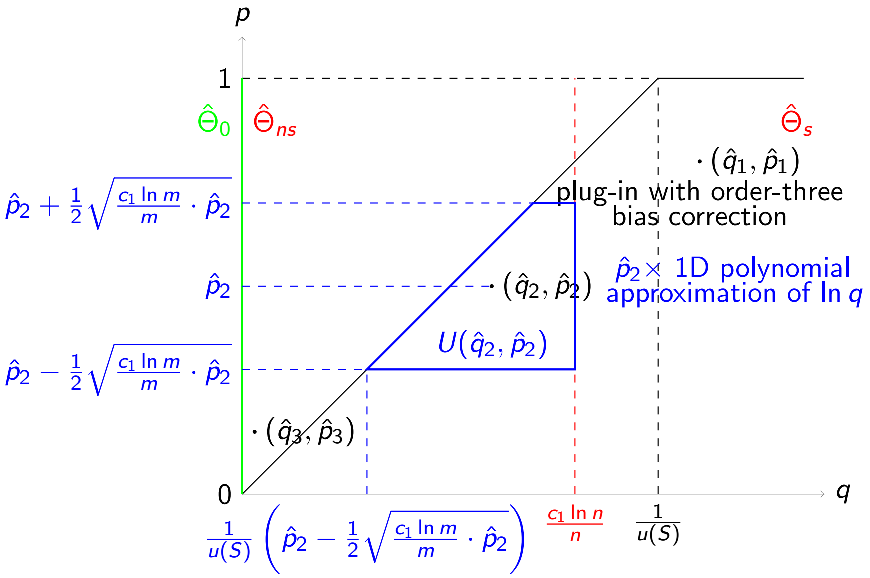
Estimator for KL divergence

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



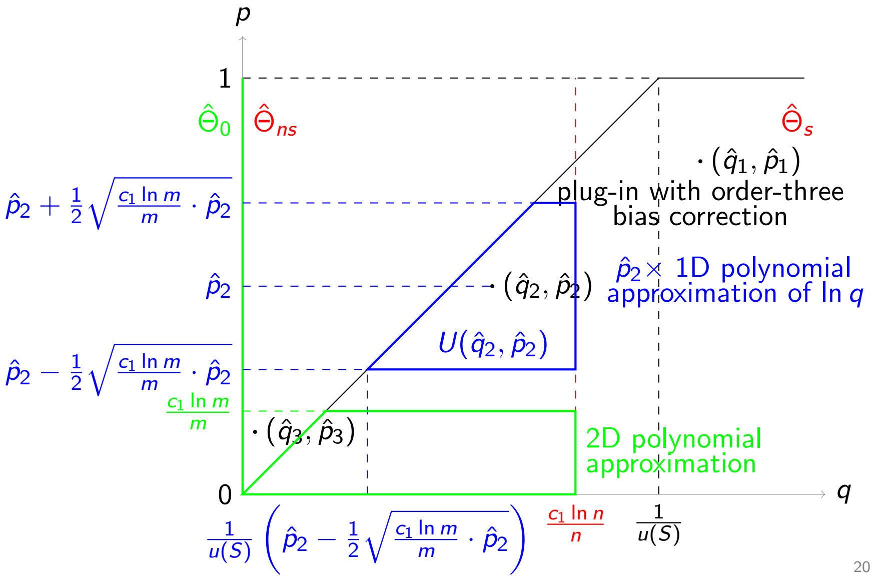
Estimator for KL divergence

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



Estimator for KL divergence

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



Some remarks

Additional remarks:

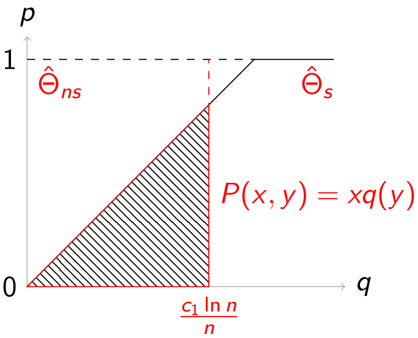
- Best polynomial approximation over general polytopes have not been solved until very recently!

Some remarks

Additional remarks:

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

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



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 the previous estimator attains the upper bound without the knowledge of S *nor* $u(S)$.

Effective sample size enlargement:

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 $\{U(x)\}_{x \in \hat{\Theta}}$ be a satisfactory confidence set.

1. **Classify the Regime**:
   - For the true parameter θ, declare that θ is in the "non-smooth" regime if θ is "close" enough to $\hat{\Theta}_0$ in terms of confidence set. Otherwise declare θ is in the "smooth" regime;
   - Compute $\hat{\theta}_n$, and declare that we are in the "non-smooth" regime if the confidence set 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 $F_{\text{appr}}(\theta)$.

Extensions

Minimax order-optimal estimator and effective sample size enlargement for more non-smooth functionals:

- Other divergences (H., Jiao, Weissman'16):
  - Hellinger distance: $H^2(P,Q) = \sum_{i=1}^{S}(\sqrt{p_i} - \sqrt{q_i})^2$
  - Chi-squared divergence: $\chi^2(P,Q) = \sum_{i=1}^{S}(p_i - q_i)^2/q_i$

Extensions

Minimax order-optimal estimator and effective sample size enlargement for more non-smooth functionals:

- Other divergences (H., Jiao, Weissman'16):
  - Hellinger distance: $H^2(P, Q) = \sum_{i=1}^{S} (\sqrt{p_i} - \sqrt{q_i})^2$
  - Chi-squared divergence: $\chi^2(P, Q) = \sum_{i=1}^{S} (p_i - q_i)^2 / q_i$
- Non-smooth nonparametric functionals (H., Jiao, Mukherjee, Weissman'16):
  - General L_r norm: $I(f) = \|f\|_r = \left(\int_0^1 |f(t)|^r dt \right)^{\frac{1}{r}}$

Extensions

Minimax order-optimal estimator and effective sample size enlargement for more non-smooth functionals:

- Other divergences (H., Jiao, Weissman'16):
  - Hellinger distance: $H^2(P,Q) = \sum_{i=1}^{S}(\sqrt{p_i} - \sqrt{q_i})^2$
  - Chi-squared divergence: $\chi^2(P,Q) = \sum_{i=1}^{S}(p_i - q_i)^2/q_i$
- Non-smooth nonparametric functionals (H., Jiao, Mukherjee, Weissman'16):
  - General L_r norm: $I(f) = \|f\|_r = \left(\int_0^1 |f(t)|^r dt\right)^{\frac{1}{r}}$

Thank you!

Email: yjhan@stanford.edu

Estimating ℓ_1 norm of Gaussian mean

Theorem (ℓ_1 norm of Gaussian mean, Cai and 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

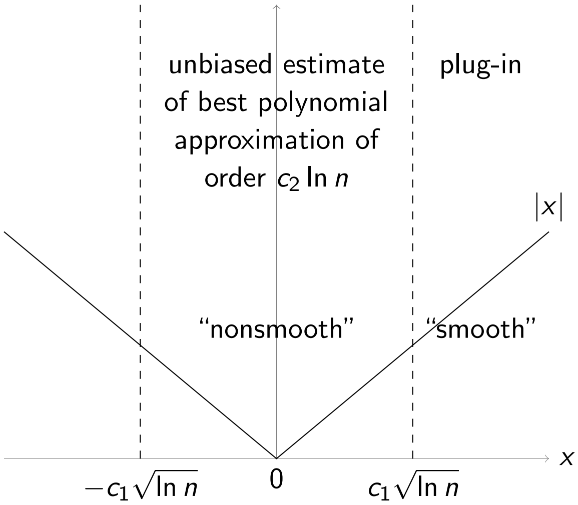
$$\sup_{\boldsymbol{\theta}\in\mathbb{R}^n} \mathbb{E}_{\boldsymbol{\theta}} \left(F(\boldsymbol{y}) - F(\boldsymbol{\theta})\right)^2 \asymp \underbrace{\sigma^2}_{\text{squared bias}} + \underbrace{\frac{\sigma^2}{n}}_{\text{variance}}$$

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

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

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

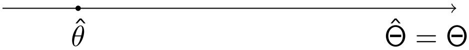
Optimal estimator for ℓ_1 norm



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

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

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

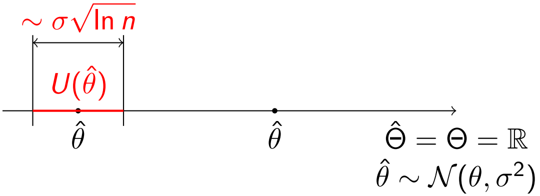


$$\hat{\Theta} = \Theta = \mathbb{R}$$

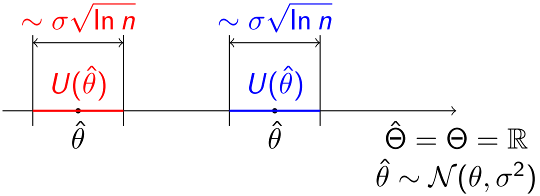
$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2)$$

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

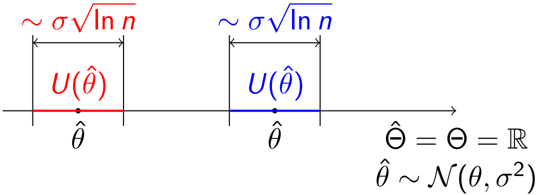
Confidence set in Gaussian model: $r \asymp n^{-A}$

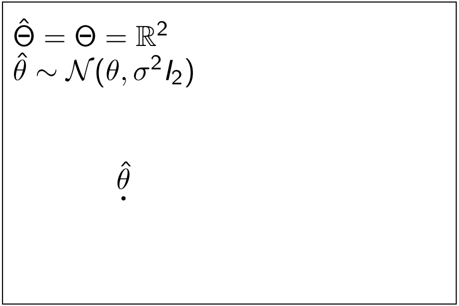


Confidence set in Gaussian model: $r \asymp n^{-A}$

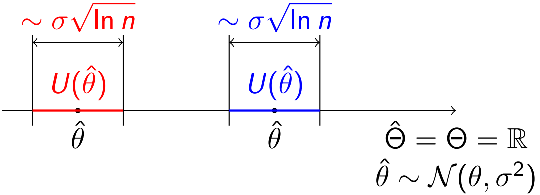


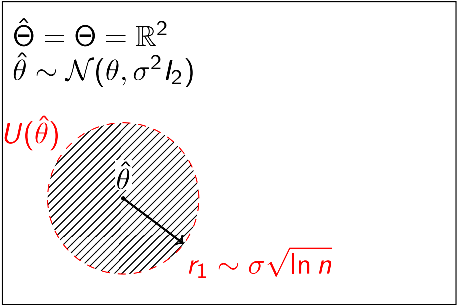
Confidence set in Gaussian model: $r \asymp n^{-A}$



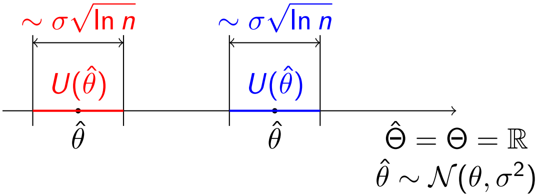


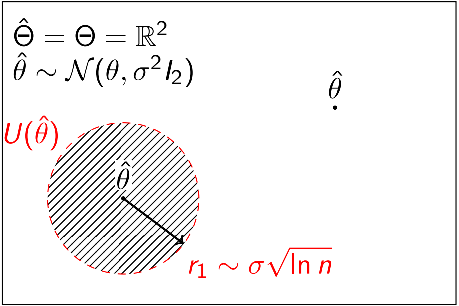
Confidence set in Gaussian model: $r \asymp n^{-A}$



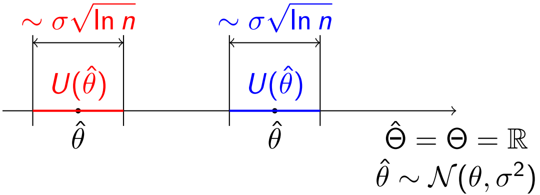


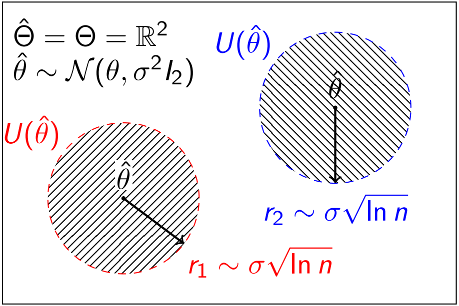
Confidence set in Gaussian model: $r \asymp n^{-A}$



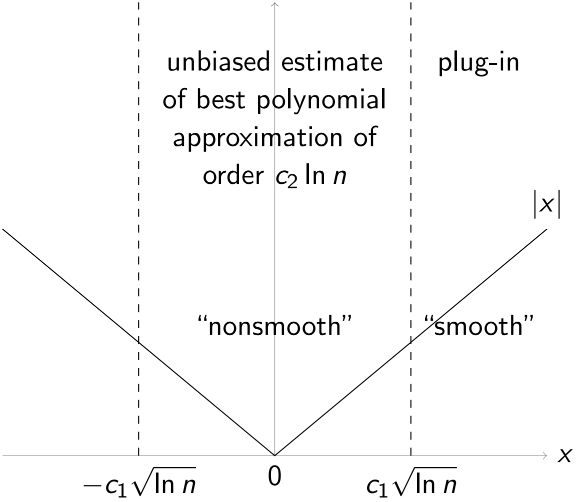


Confidence set in Gaussian model: $r \asymp n^{-A}$

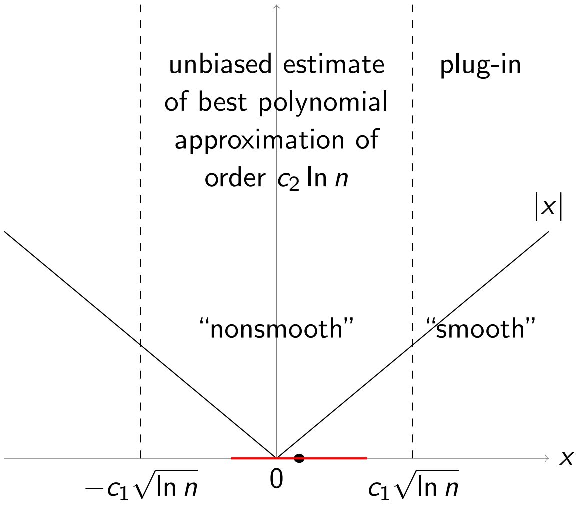




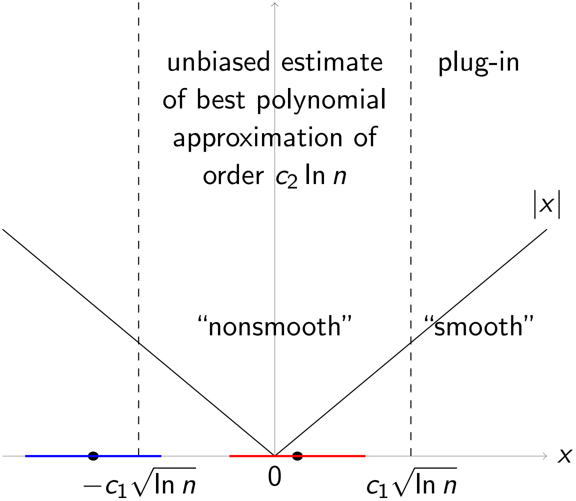
The role of confidence set: ℓ_1 norm estimation



The role of confidence set: ℓ_1 norm estimation



The role of confidence set: ℓ_1 norm estimation



“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., $\mathbb{E} S_j(\hat{\theta}_n^{(2)}) = \theta^j$.

Some remarks on ℓ_1 distance estimation

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