Information Measure Estimation and Applications: Boosting the Effective Sample Size from n to $n \ln n$

Jiantao Jiao (Stanford EE)

Joint work with:

Kartik Venkat	Stanford EE
Yanjun Han	Tsinghua EE
Tsachy Weissman	Stanford EE

Information Theory, Learning, and Big Data Workshop

March 17th, 2015

Problem Setting

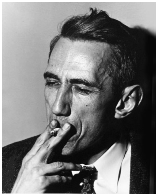
Observe n samples from discrete distribution P with support size S

we want to estimate

$$H(P) = \sum_{i=1}^{S} -p_i \ln p_i \quad \text{(Shannon'48)}$$

$$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, \quad \alpha > 0 \quad \text{(Diversity, Simpson index, Rényi entropy, etc)}$$

What was known?

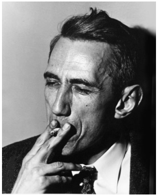


Start with entropy: $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$.

Question

Optimal estimator for $H(P)$ given n samples?

What was known?



Start with entropy: $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$.

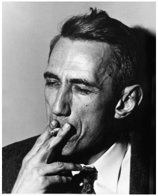
Question

Optimal estimator for $H(P)$ given n samples?

- No unbiased estimator, cannot calculate minimax estimator...

(Lehmann and Casella'98)

What was known?



Start with entropy: $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$.

Question

Optimal estimator for $H(P)$ given n samples?

- No unbiased estimator, cannot calculate minimax estimator...

(Lehmann and Casella'98)

Classical asymptotics

Optimal estimator for $H(P)$ when $n \to \infty$?

Classical problem

Empirical Entropy

$$H(P_n) = \sum_{i=1}^{S} -\hat{p}_i \ln \hat{p}_i$$

where $\hat{p}_i$ is the empirical frequency of symbol i

- $H(P_n)$ is the Maximum Likelihood Estimator (MLE)

Classical problem

Empirical Entropy

$$H(P_n) = \sum_{i=1}^{S} -\hat{p}_i \ln \hat{p}_i$$

where $\hat{p}_i$ is the empirical frequency of symbol i

- $H(P_n)$ is the Maximum Likelihood Estimator (MLE)

Theorem

The MLE $H(P_n)$ is asymptotically efficient. (Hájek–Le Cam theory)

Is there something missing?

Optimal estimator for finite samples unknown :(

Question

How about using MLE when n is finite?

Is there something missing?

Optimal estimator for finite samples unknown :(

Question

How about using MLE when n is finite?

We will show it is a very bad idea in general.

Our evaluation criterion: minimax decision theory

Denote by $\mathcal{M}_S$ distributions with support size S.

Our evaluation criterion: minimax decision theory

Denote by $\mathcal{M}_S$ distributions with support size S.

$$R_n^{\mathsf{max}}(F, \hat{F}) \triangleq \qquad \mathbb{E}_P[(F(P) - \hat{F})^2]$$

Our evaluation criterion: minimax decision theory

Denote by $\mathcal{M}_S$ distributions with support size S.

$$R_n^{\mathsf{max}}(F, \hat{F}) \triangleq \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F})^2]$$

Our evaluation criterion: minimax decision theory

Denote by $\mathcal{M}_S$ distributions with support size S.

$$R_n^{\mathsf{max}}(F, \hat{F}) \triangleq \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F})^2]$$

$$\text{Minimax risk} \triangleq \qquad \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F}))^2]$$

Our evaluation criterion: minimax decision theory

Denote by $\mathcal{M}_S$ distributions with support size S.

$$R_n^{\text{max}}(F, \hat{F}) \triangleq \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F})^2]$$

$$\text{Minimax risk} \triangleq \inf_{\text{all } \hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F}))^2]$$

Our evaluation criterion: minimax decision theory

Denote by $\mathcal{M}_S$ distributions with support size S.

$$R_n^{\max}(F, \hat{F}) \triangleq \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F})^2]$$

$$\text{Minimax risk} \triangleq \inf_{\text{all } \hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F}))^2]$$

Notation:

- $a_n \asymp b_n \Leftrightarrow 0 < c \leq \frac{a_n}{b_n} \leq C < \infty$
- $a_n \gtrsim b_n \Leftrightarrow \frac{a_n}{b_n} \geq C > 0$

Our evaluation criterion: minimax decision theory

Denote by $\mathcal{M}_S$ distributions with support size S.

$$R_n^{\mathsf{max}}(F, \hat{F}) \triangleq \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F})^2]$$

$$\text{Minimax risk} \triangleq \inf_{\text{all } \hat{F}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P[(F(P) - \hat{F}))^2]$$

Notation:

- $a_n \asymp b_n \Leftrightarrow 0 < c \leq \frac{a_n}{b_n} \leq C < \infty$
- $a_n \gtrsim b_n \Leftrightarrow \frac{a_n}{b_n} \geq C > 0$

$$L_2 \text{ Risk} = \text{Bias}^2 + \text{Variance}$$

$$\mathbb{E}_P[(F(P) - \hat{F})^2] = \left(F(P) - \mathbb{E}_P \hat{F}\right)^2 + \mathsf{Var}(\hat{F})$$

Shall we analyze MLE non-asymptotically?

Theorem (J., Venkat, Han, Weissman'14)

$$R_n^{max}(H, H(P_n)) \asymp \underbrace{\frac{S^2}{n^2}}_{Bias^2} + \underbrace{\frac{\ln^2 S}{n}}_{Variance}, \quad n \gtrsim S$$

Shall we analyze MLE non-asymptotically?

Theorem (J., Venkat, Han, Weissman'14)

$$R_n^{max}(H, H(P_n)) \asymp \underbrace{\frac{S^2}{n^2}}_{Bias^2} + \underbrace{\frac{\ln^2 S}{n}}_{Variance}, \quad n \gtrsim S$$

$$n \gg S \Leftrightarrow \text{Consistency}$$

Bias is dominating if n is not too large compared to S

Can we reduce the bias?

- Taylor series? (Taylor expansion of $H(P_n)$ around $P_n = P$)
  - Requires $n \gg S$ (Paninski'03)

Can we reduce the bias?

- Taylor series? (Taylor expansion of $H(P_n)$ around $P_n = P$)
  - Requires $n \gg S$ (Paninski'03)
- Jackknife?
  - Requires $n \gg S$ (Paninski'03)

Can we reduce the bias?

- Taylor series? (Taylor expansion of $H(P_n)$ around $P_n = P$)
  - Requires $n \gg S$ (Paninski'03)
- Jackknife?
  - Requires $n \gg S$ (Paninski'03)
- Bootstrap?
  - Requires at least $n \gg S$ (Han, J., Weissman'15)

Can we reduce the bias?

- Taylor series? (Taylor expansion of $H(P_n)$ around $P_n = P$)
  - Requires $n \gg S$ (Paninski'03)
- Jackknife?
  - Requires $n \gg S$ (Paninski'03)
- Bootstrap?
  - Requires at least $n \gg S$ (Han, J., Weissman'15)
- Bayes estimator under Dirichlet prior?
  - Requires at least $n \gg S$ (Han, J., Weissman'15)

Can we reduce the bias?

- Taylor series? (Taylor expansion of $H(P_n)$ around $P_n = P$)
  - Requires $n \gg S$ (Paninski'03)
- Jackknife?
  - Requires $n \gg S$ (Paninski'03)
- Bootstrap?
  - Requires at least $n \gg S$ (Han, J., Weissman'15)
- Bayes estimator under Dirichlet prior?
  - Requires at least $n \gg S$ (Han, J., Weissman'15)
- Plug-in Dirichlet smoothed distribution?
  - Requires at least $n \gg S$ (Han, J., Weissman'15)

There are many more...

- Coverage adjusted estimator (Chao, Shen'03, Vu, Yu, Kass'07)
- BUB estimator (Paninski'03)
- Shrinkage estimator (Hausser, Strimmer'09)
- Grassberger estimator (Grassberger'08)
- NSB estimator (Nemenman, Shafee, Bialek'02)
- B-Splines estimator (Daub et al. 04)
- Wagner, Viswanath, Kulkarni'11
- Ohannessian, Tan, Dahleh' 11
- ...

There are many more...

- Coverage adjusted estimator (Chao, Shen'03, Vu, Yu, Kass'07)
- BUB estimator (Paninski'03)
- Shrinkage estimator (Hausser, Strimmer'09)
- Grassberger estimator (Grassberger'08)
- NSB estimator (Nemenman, Shafee, Bialek'02)
- B-Splines estimator (Daub et al. 04)
- Wagner, Viswanath, Kulkarni'11
- Ohannessian, Tan, Dahleh' 11
- ...

Recent breakthrough: Valiant and Valiant'11: the exact phase transition of entropy estimation is $n \asymp S/\ln S$

There are many more...

- Coverage adjusted estimator (Chao, Shen'03, Vu, Yu, Kass'07)
- BUB estimator (Paninski'03)
- Shrinkage estimator (Hausser, Strimmer'09)
- Grassberger estimator (Grassberger'08)
- NSB estimator (Nemenman, Shafee, Bialek'02)
- B-Splines estimator (Daub et al. 04)
- Wagner, Viswanath, Kulkarni'11
- Ohannessian, Tan, Dahleh' 11
- ...

Recent breakthrough: Valiant and Valiant'11: the exact phase transition of entropy estimation is $n \asymp S/\ln S$

- Linear Programming based estimator not shown to achieve minimax rates (dependence on ϵ)
- Another one achieves it for $n \lesssim \frac{S^{1.03}}{\ln S}$.
- Not clear about other functionals

Systematic improvement

Question

Can we find a systematic methodology to improve MLE, and achieve the minimax rates for a wide class of functionals?

George Pólya: *"the more general problem may be easier to solve than the special problem".*

Start from first principle

$X \sim \mathsf{B}(n, p)$, if we use $g(X)$ to estimate $f(p)$:

$$\begin{aligned} \mathsf{Bias}(g(X)) &\triangleq f(p) - \mathbb{E}_p g(X) \\ &= f(p) - \sum_{j=0}^{n} g(j) \binom{n}{j} p^j (1-p)^{n-j} \end{aligned}$$

Start from first principle

$X \sim \mathsf{B}(n,p)$, if we use $g(X)$ to estimate $f(p)$:

$$\begin{aligned}\mathsf{Bias}(g(X)) &\triangleq f(p) - \mathbb{E}_p g(X) \\ &= f(p) - \sum_{j=0}^{n} g(j) \binom{n}{j} p^j (1-p)^{n-j}\end{aligned}$$

- Only polynomials of order $\leq n$ can be estimated without bias

$$\mathbb{E}_p \left[\frac{X(X-1)\dots(X-r+1)}{n(n-1)\dots(n-r+1)} \right] = p^r, \quad 0 \leq r \leq n$$

- Bias corresponds to polynomial approximation error

Minimize the maximum bias

What if we choose $g(X)$ such that

$$g = \arg\min_{g} \sup_{p \in [0,1]} |\mathsf{Bias}(g(X))|$$

Minimize the maximum bias

What if we choose $g(X)$ such that

$$g = \arg\min_{g} \sup_{p \in [0,1]} |\mathsf{Bias}(g(X))|$$

Paniński'03

It does not work for entropy. (Variance is too large!)

Minimize the maximum bias

What if we choose $g(X)$ such that

$$g = \arg\min_{g} \sup_{p \in [0,1]} |\mathsf{Bias}(g(X))|$$

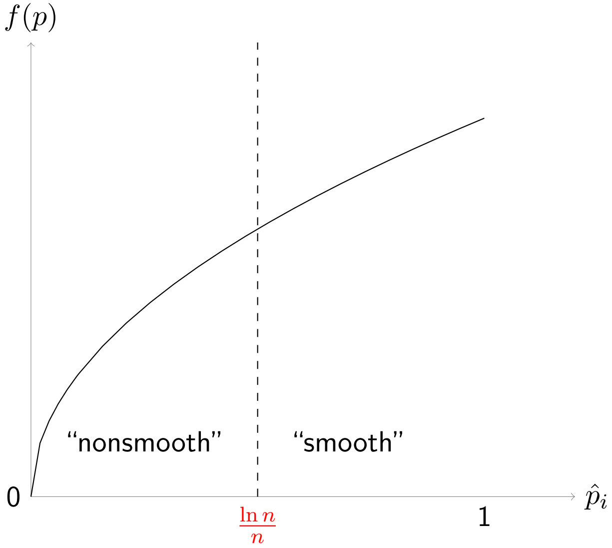
Paniński'03

It does not work for entropy. (Variance is too large!)

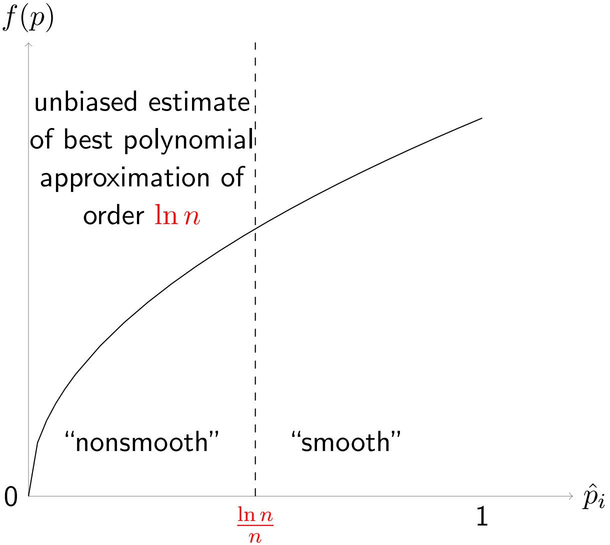
Best polynomial approximation:

$$P_k(x) \triangleq \arg \min_{P_k \in \mathsf{Poly}_k} \sup_{x \in D} |f(x) - P_k(x)|$$

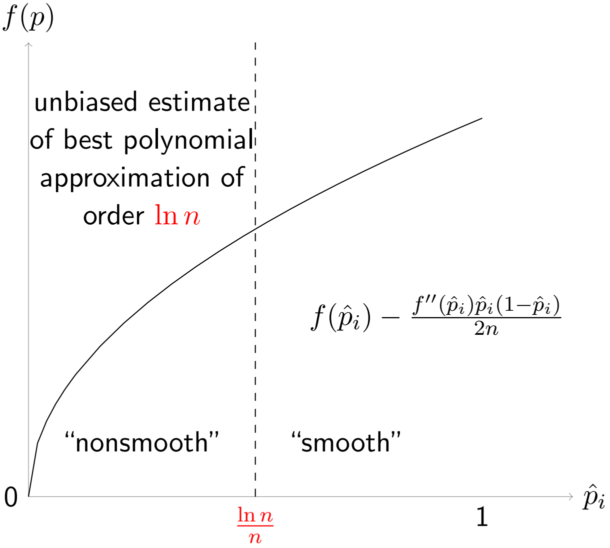
We only approximate when the problem is hard!



We only approximate when the problem is hard!



We only approximate when the problem is hard!



Why $\frac{\ln n}{n}$ threshold, $\ln n$ order?

73

О НАИЛУЧШЕМ ПРИБЛИЖЕНИИ $|x|^p$ ПРИ ПОМОЩИ МНОГОЧЛЕНОВ ВЕСЬМА ВЫСОКОЙ СТЕПЕНИ*

В статье дается асимптотическое значение наилучшего приближения $E_n|x|^p$ на отрезке $[-1, +1]$ при помощи многочленов весьма высокой степени.

1. В настоящей статье дается доказательство того факта, *что наилучшее приближение $E_n|x|^p$ на отрезке $[-1, +1]$ при помощи многочленов весьма высокой степени n удовлетворяет асимптотическому равенству*[1]

$$E_n|x|^p \sim \frac{\mu(p)}{n^p} \quad (n \to \infty), \qquad (1)$$

где $\mu(p)$ не зависит от n.

Исходным пунктом нашего исследования будет формула

Approximation theory in functional estimation

- Lepski, Nemirovski, Spokoiny'99, Cai and Low'11, Wu and Yang'14 (entropy estimation lower bound)
- Duality with shrinkage (J., Venkat, Han, Weissman'14): a new field to be explored!

A class of functionals

Note $F_\alpha(P) = \sum_{i=1}^S p_i^\alpha, \alpha > 0$. [J., Venkat, Han, Weissman'14] showed

	Minimax L_2 rates	L_2 rates of MLE
$H(P)$		$\frac{S^2}{n^2} + \frac{\ln^2 S}{n}$
$F_\alpha(P), 0 < \alpha \le \frac{1}{2}$		$\frac{S^2}{n^{2\alpha}}$
$F_\alpha(P), \frac{1}{2} < \alpha < 1$		$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$
$F_\alpha(P), 1 < \alpha < \frac{3}{2}$		$n^{-2(\alpha-1)}$
$F_\alpha(P), \alpha \ge \frac{3}{2}$	n^{-1}	n^{-1}

A class of functionals

Note $F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, \alpha > 0$. [J., Venkat, Han, Weissman'14] showed

	Minimax L_2 rates	L_2 rates of MLE
$H(P)$	$\frac{S^2}{(n\ln n)^2} + \frac{\ln^2 S}{n}$	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n}$
$F_\alpha(P), 0 < \alpha \le \frac{1}{2}$	$\frac{S^2}{(n\ln n)^{2\alpha}}$	$\frac{S^2}{n^{2\alpha}}$
$F_\alpha(P), \frac{1}{2} < \alpha < 1$	$\frac{S^2}{(n\ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$
$F_\alpha(P), 1 < \alpha < \frac{3}{2}$	$(n\ln n)^{-2(\alpha-1)}$	$n^{-2(\alpha-1)}$
$F_\alpha(P), \alpha \ge \frac{3}{2}$	n^{-1}	n^{-1}

A class of functionals

Note $F_\alpha(P) = \sum_{i=1}^S p_i^\alpha, \alpha > 0$. [J., Venkat, Han, Weissman'14] showed

	Minimax L_2 rates	L_2 rates of MLE
$H(P)$	$\frac{S^2}{(n\ln n)^2} + \frac{\ln^2 S}{n}$	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n}$
$F_\alpha(P), 0 < \alpha \le \frac{1}{2}$	$\frac{S^2}{(n\ln n)^{2\alpha}}$	$\frac{S^2}{n^{2\alpha}}$
$F_\alpha(P), \frac{1}{2} < \alpha < 1$	$\frac{S^2}{(n\ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$
$F_\alpha(P), 1 < \alpha < \frac{3}{2}$	$(n\ln n)^{-2(\alpha-1)}$	$n^{-2(\alpha-1)}$
$F_\alpha(P), \alpha \ge \frac{3}{2}$	n^{-1}	n^{-1}

Effective sample size enlargement

Minimax rate-optimal with n samples $\Leftrightarrow$ MLE with $n\ln n$ samples

Sample complexity perspectives

	MLE	Optimal
$H(P)$	$\Theta(S)$	$\Theta(S/\ln S)$
$F_\alpha(P), 0 < \alpha < 1$	$\Theta(S^{1/\alpha})$	$\Theta(S^{1/\alpha}/\ln S)$
$F_\alpha(P), \alpha > 1$	$\Theta(1)$	$\Theta(1)$

Existing literature on $F_\alpha(P)$ for $0 < \alpha < 2$: minimax rates unknown, sample complexity unknown, optimal schemes unknown..

Unique features of our entropy estimator

- One realization of a general methodology

Unique features of our entropy estimator

- One realization of a general methodology
- Achieve the minimax rate (lower bound due to Wu and Yang'14)

$$\frac{S^2}{(n \ln n)^2} + \frac{\ln^2 S}{n}$$

Unique features of our entropy estimator

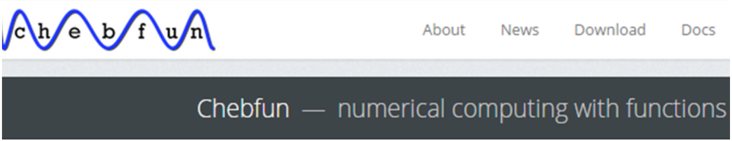
- One realization of a general methodology
- Achieve the minimax rate (lower bound due to Wu and Yang'14)

$$\frac{S^2}{(n \ln n)^2} + \frac{\ln^2 S}{n}$$

- Agnostic to the knowledge of S

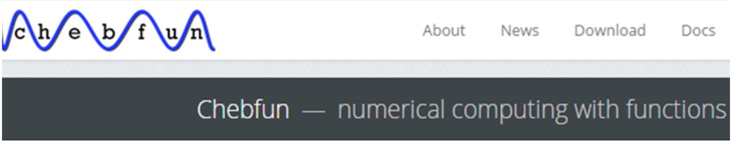
Unique features of our entropy estimator

- Easily implementable: approximation order $\ln n$, 500 order approximation using $0.46s$ on PC. (`Chebfun`)
  Empirical performance outperforms every available entropy estimator
  (Code to be released this week, available upon request)



Unique features of our entropy estimator

- Easily implementable: approximation order $\ln n$, 500 order approximation using $0.46s$ on PC. (`Chebfun`)
  Empirical performance outperforms every available entropy estimator
  (Code to be released this week, available upon request)



- Some statisticians raised interesting questions: *"We may not use this estimator unless you prove it is adaptive."*

Adaptive estimation

Near-minimax estimator:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H}^{\mathsf{Our}} - H(P) \right)^2 \asymp \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2$$

Adaptive estimation

Near-minimax estimator:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H}^{\mathsf{Our}} - H(P) \right)^2 \asymp \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2$$

What if we know a priori $H(P) \leq H$? We want

$$\sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P \left(\hat{H}^{\mathsf{Our}} - H(P) \right)^2 \asymp \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2,$$

where $\mathcal{M}_S(H) = \{P : H(P) \leq H, P \in \mathcal{M}_S\}$.

Adaptive estimation

Near-minimax estimator:

$$\sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H}^{\mathsf{Our}} - H(P) \right)^2 \asymp \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2$$

What if we know a priori $H(P) \leq H$? We want

$$\sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P \left(\hat{H}^{\mathsf{Our}} - H(P) \right)^2 \asymp \inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2,$$

where $\mathcal{M}_S(H) = \{P : H(P) \leq H, P \in \mathcal{M}_S\}$.
Can our estimator satisfy all these requirements without knowing S and H?

Our estimator is adaptive!

Han, J., Weissman'15 showed

Theorem

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P|\hat{H} - H(P)|^2 \asymp \begin{cases} \frac{S^2}{(n \ln n)^2} + \frac{H \ln S}{n} & \text{if } S \ln S \leq enH \ln n, \\ \left[\frac{H}{\ln S} \ln\left(\frac{S \ln S}{nH \ln n}\right)\right]^2 & \text{otherwise.} \end{cases} \quad (1)$$

Our estimator is adaptive!

Han, J., Weissman'15 showed

Theorem

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P |\hat{H} - H(P)|^2 \asymp \begin{cases} \frac{S^2}{(n \ln n)^2} + \frac{H \ln S}{n} & \text{if } S \ln S \leq enH \ln n, \\ \left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right) \right]^2 & \text{otherwise.} \end{cases} \quad (1)$$

- For $\epsilon > \frac{H}{\ln S}$, it requires $n \gtrsim \frac{S^{1-\epsilon/H}}{H}$ (much smaller than $S/\ln S$ for small H!) to achieve root MSE ϵ.

Our estimator is adaptive!

Han, J., Weissman'15 showed

Theorem

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P |\hat{H} - H(P)|^2 \asymp \begin{cases} \frac{S^2}{(n \ln n)^2} + \frac{H \ln S}{n} & \text{if } S \ln S \leq enH \ln n, \\ \left[\frac{H}{\ln S} \ln \left(\frac{S \ln S}{nH \ln n} \right) \right]^2 & \text{otherwise.} \end{cases} \quad (1)$$

- For $\epsilon > \frac{H}{\ln S}$, it requires $n \gtrsim \frac{S^{1-\epsilon/H}}{H}$ (much smaller than $S/\ln S$ for small H!) to achieve root MSE ϵ.
- MLE precisely requires $n \gtrsim \frac{\ln S}{H} S^{1-\epsilon/H}$.

Our estimator is adaptive!

Han, J., Weissman'15 showed

Theorem

$$\inf_{\hat{H}} \sup_{P \in \mathcal{M}_S(H)} \mathbb{E}_P |\hat{H} - H(P)|^2 \asymp \begin{cases} \frac{S^2}{(n \ln n)^2} + \frac{H \ln S}{n} & \textit{if } S \ln S \leq enH \ln n, \\ \left[\frac{H}{\ln S} \ln\left(\frac{S \ln S}{nH \ln n}\right)\right]^2 & \textit{otherwise.} \end{cases} \quad (1)$$

- For $\epsilon > \frac{H}{\ln S}$, it requires $n \gtrsim \frac{S^{1-\epsilon/H}}{H}$ (much smaller than $S/\ln S$ for small H!) to achieve root MSE ϵ.
- MLE precisely requires $n \gtrsim \frac{\ln S}{H} S^{1-\epsilon/H}$.
- The $n \Rightarrow n \ln n$ phenomenon is still true!

Understanding the generality

"Can you deal with cases where the non-differentiable regime is not just a point?"

We consider $L_1(P, Q) = \sum_{i=1}^{S} |p_i - q_i|$.

- Breakthrough by Valiant and Valiant'11: MLE requires $n \gg S$ samples, optimal is $\frac{S}{\ln S}$!
- However, VV'11's construction only proves to be optimal when $\frac{S}{\ln S} \lesssim n \lesssim S$.

Applying our general methodology

- The minimax L_2 rate is $\frac{S}{n \ln n}$, while MLE is $\frac{S}{n}$.
- The $n \Rightarrow n \ln n$ is still true!
- We show that VV'11 cannot achieve it when $n \gtrsim S$.

Applying our general methodology

- The minimax L_2 rate is $\frac{S}{n \ln n}$, while MLE is $\frac{S}{n}$.
- The $n \Rightarrow n \ln n$ is still true!
- We show that VV'11 cannot achieve it when $n \gtrsim S$.

However, is that easy to do? The nonsmooth regime is a whole line! How to cover it?

- Use a small square $\left[0, \frac{\ln n}{n}\right]^2$?
- Use a band?
- Use some other shape?

Bad news

- Best polynomial approximation in 2D is not unique.
- There exists no efficient algorithm to compute it.
- Some polynomial that achieves best approximation rate cannot be used in our methodology.
- All nonsmooth regime design methods mentioned before fail.

Bad news

- Best polynomial approximation in 2D is not unique.
- There exists no efficient algorithm to compute it.
- Some polynomial that achieves best approximation rate cannot be used in our methodology.
- All nonsmooth regime design methods mentioned before fail.
- Solution: hints from our paper "Minimax estimation of functionals of discrete distributions", or wait for soon-be-finished paper "Minimax estimation of divergence functions".

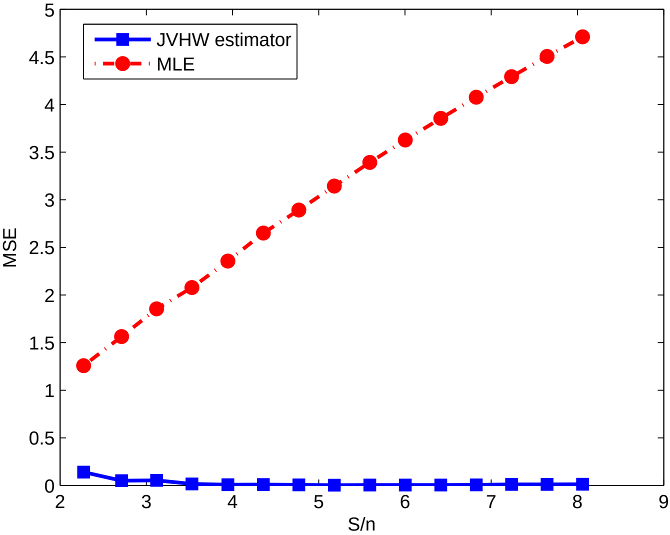
Significant difference in phase transitions

One may think: *"Who cares about a $\ln n$ improvement?"*
Take sequence $n = 2S/\ln S$, S equally (on log) sampled from 10^2 to 10^7.
For each n, S, sample 20 times from a uniform distribution.

The scale of S/n

$$S/n \in [2.2727, 8.0590]$$

Significant improvement!



Mutual Information $I(P_{XY})$

Mutual information:

$$I(P_{XY}) = H(P_X) + H(P_Y) - H(P_{XY})$$

Theorem (J., Venkat, Han, Weissman'14)

$n \Rightarrow n \ln n$ *also holds.*

Learning Tree Graphical Models

d-dimensional random vector with alphabet size S

$$X = (X_1, X_2, \ldots, X_d)$$

Suppose p.m.f. has tree structure

$$P_X = \prod_{i=1}^{d} P_{X_i | X_{\pi(i)}},$$

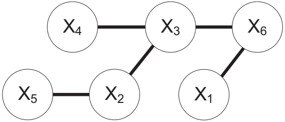
Learning Tree Graphical Models

d-dimensional random vector with alphabet size S

$$X = (X_1, X_2, \ldots, X_d)$$

Suppose p.m.f. has tree structure

$$P_X = \prod_{i=1}^{d} P_{X_i | X_{\pi(i)}},$$



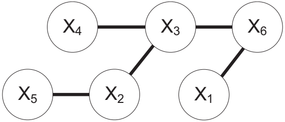
Learning Tree Graphical Models

d-dimensional random vector with alphabet size S

$$X = (X_1, X_2, \ldots, X_d)$$

Suppose p.m.f. has tree structure

$$P_X = \prod_{i=1}^{d} P_{X_i|X_{\pi(i)}},$$



Given n i.i.d. samples from P_X, estimate underlying tree structure

Chow–Liu algorithm (1968)

- Natural approach: Maximum Likelihood!
- Chow–Liu'68 solved it (2000 citations)
  - Maximum Weight Spanning Tree (MWST)
  - Empirical mutual information

Theorem (Chow, Liu'68)

$$E_{\mathsf{MLE}} = \arg \max_{E_Q \textit{ is a tree}} \sum_{e \in E_Q} I(\hat{P}_e)$$

Chow–Liu algorithm (1968)

- Natural approach: Maximum Likelihood!
- Chow–Liu'68 solved it (2000 citations)
  - Maximum Weight Spanning Tree (MWST)
  - Empirical mutual information

Theorem (Chow, Liu'68)

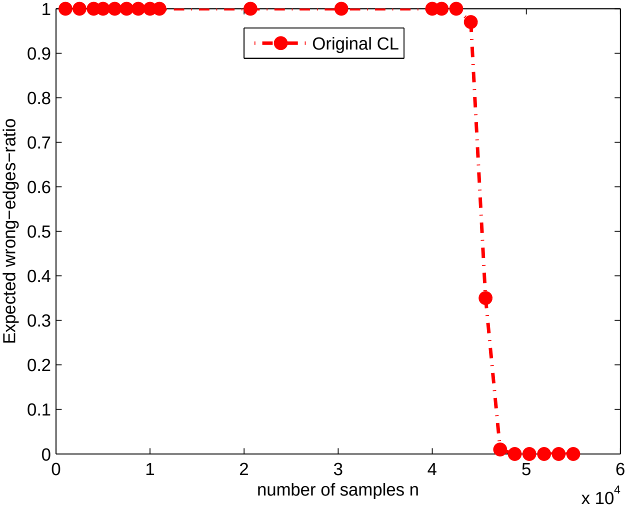
$$E_{\mathsf{MLE}} = \arg \max_{E_Q \textsf{ is a tree}} \sum_{e \in E_Q} I(\hat{P}_e)$$

Our approach

Replace empirical mutual information by better estimates!

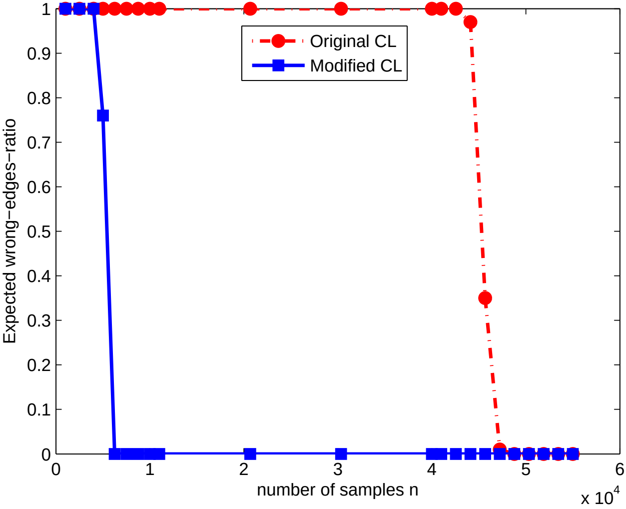
Original CL vs. Modified CL

We set $d = 7, S = 300$, a star graph, sweep n from $1k$ to $55k$.



8 fold improvement! [and even more]

We set $d = 7, S = 300$, a star graph, sweep n from $1k$ to $55k$.



What we did not talk about

- How to analyze the bias of MLE
- Why bootstrap and jackknife fail
- Extensions in multivariate settings (ℓ_p distance)
- Extensions in nonparametric estimation
- Extensions in non-i.i.d. models
- Applications in machine learning, medical imaging, computer vision, genomics, etc

Hindsight

- Unbiased estimation:
  - approximation basis must satisfy unbiased equation (Kolmogorov'50)
  - basis have good approximation properties
  - additional variance incurred by approximation should not be large

Hindsight

- Unbiased estimation:
  - approximation basis must satisfy unbiased equation (Kolmogorov'50)
  - basis have good approximation properties
  - additional variance incurred by approximation should not be large
- Theory of functional estimation $\Leftrightarrow$ Analytical properties of functions
  - The whole field of approximation theory, or even more

Related work

- O. Lepski, A. Nemirovski, and V. Spokoiny'99, "On estimation of the L_r norm of a regression function"
- T. Cai and M. Low'11, "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional",
- Y. Wu and P. Yang'14, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation",
- J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi'14, "The complexity of estimating Renyi entropy"

Our work

- J. Jiao, K. Venkat, Y. Han, T. Weissman, "Minimax estimation of functionals of discrete distributions", to appear in IEEE Transactions on Information Theory
- J. Jiao, K. Venkat, Y. Han, T. Weissman, "Maximum likelihood estimation of functionals of discrete distributrions", available on arXiv
- J. Jiao, K. Venkat, Y. Han, T. Weissman, "Beyond maximum likelihood: from theory to practice", available on arXiv
- J. Jiao, Y. Han, T. Weissman, "Minimax estimation of divergence functions", in preparation.
- Y. Han, J. Jiao, T. Weissman, "Adaptive estimation of Shannon entropy", available on arXiv
- Y. Han, J. Jiao, T. Weissman, "Does Dirichlet prior smoothing solve the Shannon entropy estimation problem?", available on arXiv
- Y. Han, J. Jiao, T. Weissman, "Bias correction using Taylor series, Bootstrap, and Jackknife", in preparation
- Y. Han, J. Jiao, T. Weissman, "How to bound of gap of Jensen's inequality?", in preparation

Thank you!