Adaptive Estimation of Shannon Entropy

Yanjun Han[⋆], Jiantao Jiao[†], Tsachy Weissman[†]

[⋆] Department of Electronic Engineering, Tsinghua University
[†] Department of Electrical Engineering, Stanford University

{yjhan, jiantao, tsachy}@stanford.edu

ISIT2015

June 16, 2015

Problem setting

- For a discrete distribution $P = (p_1, p_2, \cdots, p_S)$ with alphabet size S, then given $X_1, X_2, \cdots, X_n \overset{\text{i.i.d}}{\sim} P$

Question

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

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

- A natural answer: the empirical entropy (MLE) $H(P_n)$, where P_n is the empirical distribution

The decision theoretic framework

- Denote by $\mathcal{P}$ a given collection of probability measure P

Question

How to analyze:

$$R_{maximum}(\mathcal{P}; \hat{H}) = \sup_{P \in \mathcal{P}} \mathbb{E}_P (H(P) - \hat{H})^2$$

$$R_{minimax}(\mathcal{P}) = \inf_{\text{all } \hat{H}} \sup_{P \in \mathcal{P}} \mathbb{E}_P (H(P) - \hat{H})^2$$

- Notations:

$$a_n \asymp b_n, a_n = \Theta(b_n) \iff 0 < c \le \frac{a_n}{b_n} \le C < \infty$$

$$a_n \lesssim b_n, a_n = O(b_n) \iff \frac{a_n}{b_n} \le C < \infty$$

Existing literature

- Choosing $\mathcal{P} = \mathcal{M}_S$, the collection of all distributions with support size S, we have (J., Venkat, Han, Weissman'14, J., Han, Weissman'15)

	Minimax L_2 rate	L_2 rate of MLE
$H(P) = \sum_{i=1}^{S} -p_i \ln p_i$	$\frac{S^2}{(n \ln n)^2} + \frac{\ln^2 S}{n}$	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 0 < \alpha \le 1$	$\frac{S^2}{(n \ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 1 < \alpha < 3/2$	$\frac{1}{(n \ln n)^{2(\alpha-1)}}$	$\frac{1}{n^{2(\alpha-1)}}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, \alpha \ge 3/2$	$\frac{1}{n}$	$\frac{1}{n}$
$\ell_H(P, Q) = \sum_{i=1}^{S} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$	$\frac{S}{n \ln n} + \frac{\sqrt{S}}{n}$	$\frac{S}{n}$
$\ell_1(P, Q) = \sum_{i=1}^{S} \lvert p_i - q_i \rvert$	$\frac{S}{n \ln n}$	$\frac{S}{n}$

Effective Sample Enlargement

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

The adaptive framework

- Some statisticians raised interesting questions: *"We may not use this estimator unless you prove it is adaptive."*
- Alleviate the pessimism of minimaxity: adaptive procedure
  1. We want

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

     where $\mathcal{M}_S(H) = \{P \in \mathcal{M}_S : H(P) \le H\}$.
  2. Is there an estimator satisfying all these requirements without knowing S and H?

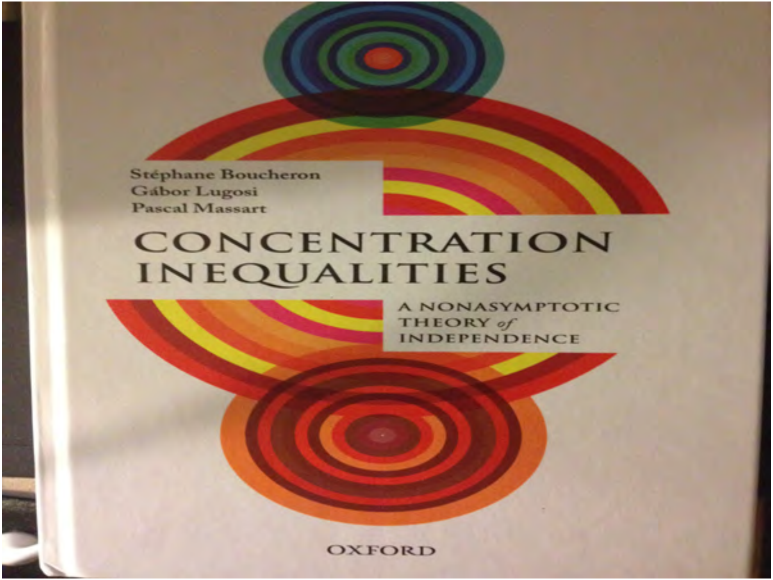
Starting from the MLE

- We can decompose the mean squared error as

$$\text{Mean Squared Error} \;=\; \text{Bias}^2 + \text{Variance}$$
$$\mathbb{E}_P\left(\hat{H} - H(P)\right)^2 = \left(\mathbb{E}_P\hat{H} - H(P)\right)^2 + \mathbb{E}_P\left(\hat{H} - \mathbb{E}_P\hat{H}\right)^2$$

- Consider the MLE $H(P_n)$

Bounding the variance



Bounding the bias

- Given $X \sim \mathrm{B}(n, p), f \in C[0,1]$, the bias of $f(X/n)$ in estimating $f(p)$ is

$$\begin{aligned} B(f,p,n) &= \mathbb{E}_p f(X/n) - f(p) \\ &= \sum_{j=0}^{n} f\left(\frac{j}{n}\right) \cdot \binom{n}{j} p^j (1-p)^{n-j} - f(p) \end{aligned}$$

- We need to bound $B(f,p,n)$ for every f, p, n. Perhaps the first step is to characterize

$$\sup_{p \in [0,1]} |B(f,p,n)|$$

Relationships with positive linear operators

- Say we use $F(\hat{\theta}_n)$ to estimate $F(\theta)$. How to analyze $\mathbb{E}_\theta F(\hat{\theta}_n) - F(\theta)$? We note that $F(\hat{\theta}_n)$
  1. maps a continuous function $F(\theta)$ to another cont. func. of θ
  2. is linear in F
  3. is positive ($F(\theta) \geq 0 \Longrightarrow \mathbb{E}_\theta F(\hat{\theta}_n) \geq 0$)
- Hence,

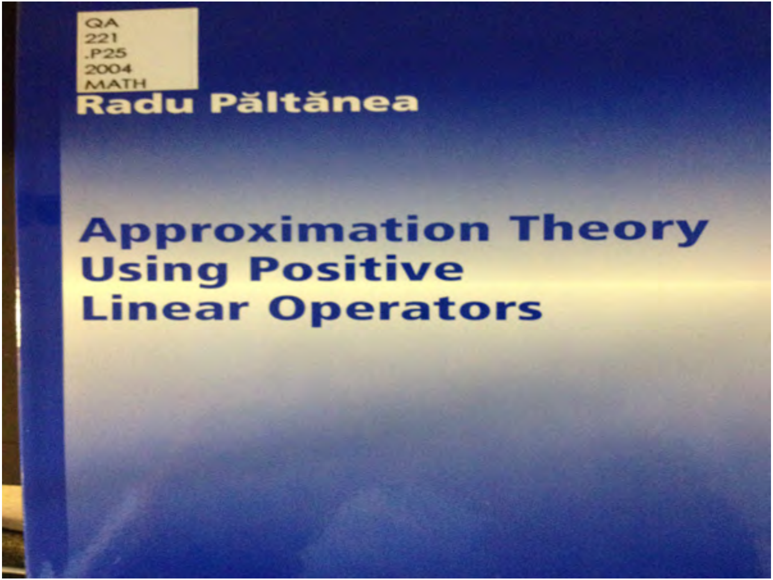
$$\text{Bias of } F(\hat{\theta}_n) \Longleftrightarrow \text{ Approximation error of } \mathbb{E}_\theta F(\hat{\theta}_n)$$

- The answer given by approximation theory (Totik'94, Knoop and Zhou'94)

$$\sup_{p \in [0,1]} |B(f,p,n)| \asymp \omega_\varphi^2(f, n^{-\frac{1}{2}})$$

ω_φ^2: second–order Ditzian–Totik modulus of smoothness

Approximation using positive linear operators



What do we know now?

- Applying the Ditzian–Totik modulus of smoothness to $f(p) = -p \ln p$, we have

$$\sup_{p \in [0,1]} |B(f, p, n)| \lesssim \frac{1}{n}$$

- However, a better pointwise bound can be obtained when p is small:

Theorem (Han, J., Weissman'15)

$$|B(f, p, n)| = \begin{cases} -p \ln(np) + \Theta(1) np^2 & p < 1/n \\ \Theta(1) \frac{1-p}{n} & 1/n \le p < 1 \end{cases}$$

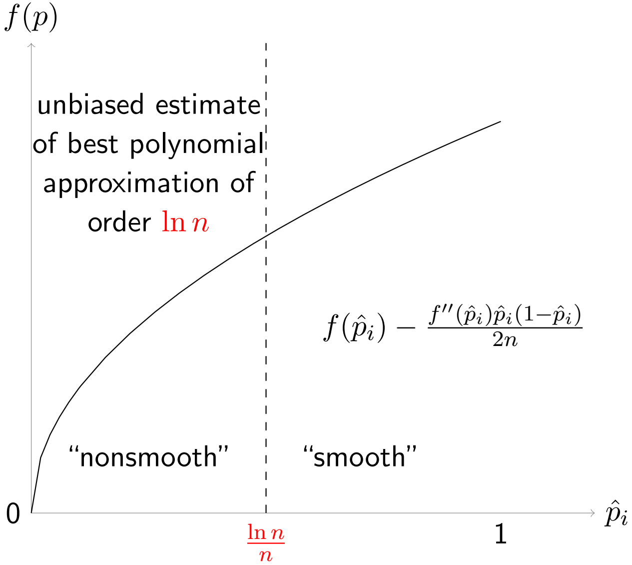
Applying it to entropy estimation

Theorem (Han, J., Weissman'15)

$$\sup_{P\in\mathcal{M}_S(H)} \mathbb{E}_P|H(P_n)-H(P)|^2 = \begin{cases} \Theta(1)\left[\left(\frac{S}{n}\right)^2+\frac{H\ln S}{n}\right] & \textit{if } S\ln S\le e^2 nH,\\ \left[\frac{H}{\ln S}\ln\left(\frac{S\ln S}{nH}\right)+O\left(\frac{H}{\ln S}+\frac{(\ln n)^2}{n}\right)\right]^2 & \textit{otherwise.}\end{cases}$$

- For $\epsilon > \frac{H}{\ln S}$, it requires $\Theta(S^{1-\frac{\epsilon}{H}}\cdot\frac{\ln S}{H})$ samples to achieve L_1 error ϵ

Turn to our minimax estimator



Best polynomial approximation

- Polynomial with degree $\leq n$ can be estimated without bias: for $X \sim \mathrm{B}(n, p)$,

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

- Bias corresponds to the best polynomial approximation error
- Advanced tools from approximation theory: for $f \in C[0,1]$,
  1. norm bound (Ditzian and Totik'87, DeVore and Lorentz'93):

$$\exists p_n, \deg(p_n) \leq n, \|f - p_n\|_\infty \lesssim \omega_\varphi^2(f, n^{-1})$$

     ω_φ^2: second–order Ditzian–Totik modulus of smoothness
  2. pointwise bound (Leviatan'86):

$$\exists p_n, \deg(p_n) \leq n, |f(x) - p_n(x)| \lesssim \omega^2\left(f, \frac{\sqrt{x(1-x)}}{n}\right)$$

     ω^2: second–order modulus of smoothness

Refined pointwise bound

- Applying the preceding result to $f(p) = -p \ln p$:

  norm bound: $\exists p_n, \deg(p_n) \le n, \|f - p_n\|_\infty \lesssim n^{-2}$

  pointwise bound: $\exists p_n, \deg(p_n) \le n, |f(p) - p_n(p)| \lesssim \sqrt{p(1-p)}/n$
- Unsatisfactory for $f(p) = -p \ln p$ and its order-n best approximating polynomial $P_n[f](p)$ (without constant)

Theorem (Han, J., Weissman'15)

$$|f(p) - P_n[f](p)| \begin{cases} = -p \ln(n^2 p) + O(p) & 0 \le p \le n^{-2} \\ \lesssim n^{-2} & n^{-2} < p \le 1 \end{cases}$$

Moreover, there does not exist polynomial p_n such that $\deg(p_n) \le n$ and

$$|f(p) - p_n(p)| \begin{cases} \le -p \ln(n^2 p) - \omega(p) & 0 \le p \le n^{-2} \\ \lesssim n^{-2} & n^{-2} < p \le 1 \end{cases}$$

Applying it to the entropy function

Theorem (Han, J., Weissman'15)

$$\inf_{\hat{H}} \sup_{P\in\mathcal{M}_S(H)} \mathbb{E}_P|\hat{H}-H(P)|^2$$

$$\asymp \begin{cases} \left(\frac{S}{n\ln n}\right)^2 + \frac{H\ln S}{n} & \textit{if } S\ln S \le e^2 nH\ln n, \\ \left[\frac{H}{\ln S}\ln\left(\frac{S\ln S}{nH\ln n}\right) + O\left(\frac{H}{\ln S}+\frac{(\ln n)^2}{n}\right)\right]^2 & \textit{otherwise.} \end{cases}$$

- Adaptivity of our estimator: it requires no knowledge of S or H
- For $\epsilon > \frac{H}{\ln S}$, it requires $\Theta(\frac{S^{1-\frac{\epsilon}{H}}}{H})$ samples to achieve L_1 error ϵ
- $n \to n\ln n$ effective sample enlargement still holds!

Summary

- Adaptive procedure
- Refined pointwise bound in approximation theory
- $n \to n \ln n$ effective sample enlargement

Related work

- Y. Han, J. Jiao, T. Weissman, "Adaptive estimation of Shannon entropy", available on arXiv
- J. Jiao, K. Venkat, Y. Han, T. Weissman, "Minimax estimation of functionals of discrete distributions," IEEE Transactions on Information Theory, vol. 61, no. 5, pp. 2835–2885, 2015
- J. Jiao, K. Venkat, Y. Han, T. Weissman, "Maximum likelihood estimation of functionals of discrete distributrions", available on arXiv
- Y. Han, J. Jiao, T. Weissman, "Is the usual pointwise bound in approximation theory optimal?", in preparation
- J. Jiao, Y. Han, T. Weissman, "Minimax estimation of divergence functions", in preparation

Thank you!