On the High Accuracy Limitation of Adaptive Property Estimation

Yanjun Han
Department of Electrical Engineering, Stanford University
Email: yjhan@stanford.edu

Objective

Target: characterize the following adaptive minimax risk:

$$R^\star_{\text{adaptive}}(n,k) = \inf_{\widehat{p}} \sup_{p\in\mathcal{M}_k} \sup_{F\in\mathcal{F}_{\text{Lip}}} \mathbb{E}_p|F(\widehat{p}) - F(p)|$$

- n: sample size;
- k: support size;
- p: unknown true distribution;
- $\widehat{p}$: a distribution estimator based on n iid observations from p;
- $\mathcal{M}_k$: all discrete distributions with support size k;
- F: symmetric functional/property defined as $F(p) = \sum_{i=1}^k f(p_i)$;
- $\mathcal{F}_{\text{Lip}}$: class of all functionals F such that f is 1-Lipschitz.

Related work

Two similar quantities:

- A smaller quantity (Hao and Orlitsky'19):

$$\sup_{F\in\mathcal{F}_{\text{Lip}}} \inf_{\widehat{p}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p|F(\widehat{p}) - F(p)| \asymp \sqrt{\frac{k}{n\log n}}, \quad \log n \lesssim k \lesssim n\log n.$$

- A larger quantity (Han, Jiao, and Weissman'18):

$$\inf_{\widehat{p}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p\left[\sup_{F\in\mathcal{F}_{\text{Lip}}} |F(\widehat{p}) - F(p)|\right] \asymp \begin{cases} \sqrt{\frac{k}{n\log n}} & \text{if } k \gg n^{1/3} \\ \sqrt{\frac{k}{n}} & \text{if } 1 \ll k \ll n^{1/3} \end{cases}.$$

Motivation: functional estimation

Problem: Given n i.i.d. observations $X_1,\cdots,X_n \sim p = (p_1,\cdots,p_k)$, aim to estimate the quantity $F(p) = \sum_{i=1}^k f(p_i)$ for a given f

Example: Shannon entropy when $f(x) = -x\log x$, support size when $f(x) = 1(x\neq 0)$

Applications: genetics, image processing, computer vision, secrecy, ecology, physics...

Generalization: non-symmetric, multivariate and nonparametric functionals

Ad-hoc estimation

Optimal estimator with n samples $\Longleftrightarrow$ MLE with $n\log n$ samples

Supported in lots of recent literature:

- Shannon entropy (VV11a, VV11b, VV13, JVHW15, WY16)
- Rényi entropy (AOST14, AOST17)
- distance to uniformity (VV13, JHW18)
- divergences (HJW16, JHW18, BZLV18)
- nonparametrics (HJM17, HJWW17)
- general 1-Lipschitz functional (HO19a, HO19b)

Adaptive estimation

Target: find a single distribution estimator $\widehat{p}$ such that the plugging $\widehat{p}$ into the functional is universally optimal for "many" functionals

$$X_1,\cdots,X_n \longrightarrow \widehat{p} \begin{cases} \longrightarrow F_1(\widehat{p}) \\ \longrightarrow F_2(\widehat{p}) \\ \longrightarrow F_3(\widehat{p}) \end{cases}$$

Too good to be true? No!

Idea I: local moment matching (LMM)

Theorem (Han, Jiao, and Weissman'18): There exists a single estimator $\widehat{p}$, efficiently computable, which achieves the optimal sample complexity for a large class of symmetric functionals whenever $\varepsilon \gg n^{-1/3}$.

In particular, it solves the minimax problem

$$\inf_{\widehat{p}} \sup_{p\in\mathcal{M}_k} \mathbb{E}_p\|\widehat{p} - p\|_{1,\text{sorted}} \asymp \sqrt{\frac{k}{n\log n}} + \left(\widetilde{\Theta}(n^{-1/3}) \wedge \sqrt{\frac{k}{n}}\right).$$

Idea II: profile maximum likelihood (PML)

Theorem (Acharya, Das, Orlitsky, and Suresh'17):

$$\sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|F(p^{\text{PML}}) - F(p)| > 2\varepsilon) \le e^{3\sqrt{n}} \cdot \inf_{\widehat{F}} \sup_{p\in\mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| > \varepsilon).$$

Theorem (Han and Shiragur'21): Improved competitive factor from $e^{3\sqrt{n}}$ to $\exp(n^{1/3+o(1)})$.

Corollary: since the tail probability on the RHS is typically $\exp(-n\varepsilon^2)$ when n exceeds the sample complexity of acheving error ε, the PML plug-in approach attains the rate-optimal sample complexity if $\varepsilon \gg n^{-1/3}$.

Table of comparison

	ad-hoc	LMM	PML
optimality	full: $\varepsilon \gg n^{-1/2}$	if $\varepsilon \gg n^{-1/3}$	if $\varepsilon \gg n^{-1/3}$
complexity	almost linear	polynomial	polynomial*
functional independent	✗	✓	✓
asymmetric functional	✓	✗	✗
free parameter tuning	✗	✗	✓

Question: is there a fundamental discrepancy between non-adaptive and adaptive approaches?

Main results

Theorem:

$$R^\star_{\text{adaptive}} \asymp \begin{cases} \sqrt{\frac{k}{n\log n}} & \text{if } k \gg n^{1/3} \\ \sqrt{\frac{k}{n}} & \text{if } 1 \ll k \ll n^{1/3} \end{cases}.$$

Corollary: The competitive factor in the PML analysis cannot be improved from $\exp(cn^{1/3+o(1)})$ to $\exp(cn^{1/3-o(1)})$.

Implication:

- phase transition for the adaptive minimax risk
- strict penalty of adaptation iff $\varepsilon \ll n^{-1/3}$
- LMM and PML both optimal in the class of adaptive estimators

Comparison with classical adaptive estimation

General minimax formulation:

$$\inf_T \sup_{\theta\in\Theta} \mathbb{E}_\theta[L(\theta, T)].$$

Classical adaptive estimation: adapting to parameter sets

- a nested class of parameter sets $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$;
- penalty of adaptation:

$$\inf_T \max_{m\ge 1} \frac{\sup_{\theta\in\Theta_m} \mathbb{E}_\theta[L(\theta,T)]}{\inf_{T_m} \sup_{\theta\in\Theta_m} \mathbb{E}_\theta[L(\theta,T_m)]}.$$

Our adaptive estimation: adapting to loss functions

- a class of loss functions $L \in \mathcal{L}$;
- in our example, $L_F(p,\widehat{p}) = |F(p) - F(\widehat{p})|$, and $\mathcal{L} = \{L_F : F \text{ is 1-Lip}\}$;
- adaptive minimax risk:

$$\inf_T \sup_{\theta\in\Theta} \sup_{L\in\mathcal{L}} \mathbb{E}_\theta[L(\theta,T)].$$

Proof technique

Find $\theta_1,\cdots,\theta_M \in \Theta$ and $L_1,\cdots,L_M \in \mathcal{L}$ with the indistinguishability condition and a new separation condition: for all $i\neq j$,

$$\inf_a [L_i(\theta_i, a) + L_j(\theta_j, a)] \ge \Delta.$$

References

- Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pages 11–21, 2017.
- Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. In *Conference On Learning Theory*, pages 3189–3221, 2018.
- Yi Hao and Alon Orlitsky. Unified sample-optimal property estimation in near-linear time. In *Advances in Neural Information Processing Systems*, pages 11104–11114, 2019.
- Yanjun Han and Kirankumar Shiragur. On the competitive analysis and high accuracy optimality of profile maximum likelihood. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1317–1336. SIAM, 2021.