On the tight statistical analysis of a maximum likelihood estimator based on profiles

Yanjun Han (Berkeley Simons)

Acknowledgements:

Jayadev Acharya	Cornell ECE
Moses Charikar	Stanford CS
Jiantao Jiao	Berkeley EECS
Kirankumar Shiragur	Stanford MS&E
Aaron Sidford	Stanford MS&E
Tsachy Weissman	Stanford EE
Yihong Wu	Yale Stats

MAD+ Seminar, Center for Data Science and Courant Institute, NYU

Maximum likelihood estimator

If $x \sim P_\theta$ with $\theta \in \Theta$,

$$\theta^{\text{MLE}} \triangleq \arg\max_{\theta \in \Theta} P_\theta(x)$$

Fundamental method of parameter estimation with numerous success in:

- statistics
- signal processing
- machine learning
- ...

Maximum likelihood estimator

If $x \sim P_\theta$ with $\theta \in \Theta$,

$$\theta^{\text{MLE}} \triangleq \arg\max_{\theta \in \Theta} P_\theta(x)$$

Fundamental method of parameter estimation with numerous success in:

- statistics
- signal processing
- machine learning
- ...

"The appeal of maximum likelihood stems from its universal applicability, good mathematical properties, ..., and generally good track record as a tool in applied statistics, a record accumulated over fifty years of heavy usage."

—— [Efron, 1980]

Suboptimality of MLE under group transformation

Theorem (Cai and Low, 2011)

For $X \sim \mathcal{N}(\theta, I_p)$, it holds that

$$\inf_{T(\cdot)} \sup_{\|\theta\|_\infty \le 1} \mathbb{E}_\theta |T(X) - \|\theta\|_1| \asymp p \cdot \frac{\log\log p}{\log p},$$

$$\sup_{\|\theta\|_\infty \le 1} \mathbb{E}_\theta | \|\theta^{\mathsf{MLE}}\|_1 - \|\theta\|_1| \asymp p.$$

Suboptimality of MLE under group transformation

Theorem (Cai and Low, 2011)

For $X \sim \mathcal{N}(\theta, I_p)$, it holds that

$$\inf_{T(\cdot)} \sup_{\|\theta\|_\infty \le 1} \mathbb{E}_\theta |T(X) - \|\theta\|_1| \asymp p \cdot \frac{\log\log p}{\log p},$$

$$\sup_{\|\theta\|_\infty \le 1} \mathbb{E}_\theta |\|\theta^{\mathsf{MLE}}\|_1 - \|\theta\|_1| \asymp p.$$

Theorem (H., Jiao, and Weissman, 2018)

For $X = (X_1, \cdots, X_n)$ with i.i.d. $X_i \sim p = (p_1, \cdots, p_k)$, it holds that

$$\inf_{\widehat{p}} \sup_{p} \mathbb{E}_p \|\widehat{p} - p\|_{1,\text{sorted}} \asymp \sqrt{\frac{k}{n \log n}} + \min\left\{\sqrt{\frac{k}{n}}, n^{-1/3}\right\},$$

$$\sup_{p} \mathbb{E}_p \|p^{\mathsf{MLE}} - p\|_{1,\text{sorted}} \asymp \sqrt{\frac{k}{n}}.$$

Profile

A group action G on a set $\mathcal{X}$ partitions $\mathcal{X}$ into several equivalence classes: for $x, x' \in \mathcal{X}$,

$$x \sim_G x' \iff \exists g \in G : gx = x'$$

Profile

A group action G on a set $\mathcal{X}$ partitions $\mathcal{X}$ into several equivalence classes: for $x, x' \in \mathcal{X}$,

$$x \sim_G x' \iff \exists g \in G : gx = x'$$

Definition (Profile, Orlitsky et al. 2004)

For an observation $x \in \mathcal{X}$, its profile ϕ with respect to the group action G is defined as the equivalence class of x in $\mathcal{X}$:

$$\phi(x) = \{x' \in \mathcal{X} : x' \sim_G x\} = Gx.$$

Profile

A group action G on a set $\mathcal{X}$ partitions $\mathcal{X}$ into several equivalence classes: for $x, x' \in \mathcal{X}$,

$$x \sim_G x' \iff \exists g \in G : gx = x'$$

Definition (Profile, Orlitsky et al. 2004)

For an observation $x \in \mathcal{X}$, its profile ϕ with respect to the group action G is defined as the equivalence class of x in $\mathcal{X}$:

$$\phi(x) = \{x' \in \mathcal{X} : x' \sim_G x\} = Gx.$$

Lemma (Hájek, 1967)

If for all $g \in G$, we have $P_{g\theta}(gx) = P_\theta(x)$ and $L(\theta, T) = L(g\theta, T)$, then $\phi(x)$ is "sufficient" for estimating θ under loss L.

Examples of profiles

Group action: throughout we consider the action of $G = S_p$ on $\mathbb{R}^p$, i.e. for $\pi \in S_p$ and $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$,

$$\pi x \triangleq (x_{\pi(1)}, \cdots, x_{\pi(p)}).$$

Examples of profiles

Group action: throughout we consider the action of $G = S_p$ on $\mathbb{R}^p$, i.e. for $\pi \in S_p$ and $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$,

$$\pi x \triangleq (x_{\pi(1)}, \cdots, x_{\pi(p)}).$$

Example (permutation invariance)

- for a p-dim observation vector $x = (x_1, \cdots, x_p)$, the profile $\phi(x) = (x_{(1)}, x_{(2)}, \cdots, x_{(p)}) \in \mathbb{R}^p$ is the order statistic

Examples of profiles

Group action: throughout we consider the action of $G = S_p$ on $\mathbb{R}^p$, i.e. for $\pi \in S_p$ and $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$,

$$\pi x \triangleq (x_{\pi(1)}, \cdots, x_{\pi(p)}).$$

Example (permutation invariance)

- for a p-dim observation vector $x = (x_1, \cdots, x_p)$, the profile $\phi(x) = (x_{(1)}, x_{(2)}, \cdots, x_{(p)}) \in \mathbb{R}^p$ is the order statistic
- if in addition $x \sim P_\theta$, permutation invariance of the model requires that $P_{\pi\theta}(\pi x) = P_\theta(x)$

Examples of profiles

Group action: throughout we consider the action of $G = S_p$ on $\mathbb{R}^p$, i.e. for $\pi \in S_p$ and $x = (x_1, \cdots, x_p) \in \mathbb{R}^p$,

$$\pi x \triangleq (x_{\pi(1)}, \cdots, x_{\pi(p)}).$$

Example (permutation invariance)

- for a p-dim observation vector $x = (x_1, \cdots, x_p)$, the profile $\phi(x) = (x_{(1)}, x_{(2)}, \cdots, x_{(p)}) \in \mathbb{R}^p$ is the order statistic
- if in addition $x \sim P_\theta$, permutation invariance of the model requires that $P_{\pi\theta}(\pi x) = P_\theta(x)$
- if in addition $L(\theta, T) = L(\pi\theta, T)$, Hájek sufficiency implies that $\phi(x)$ is sufficient for estimating θ under loss L

The Profile MLE

Likelihood of a profile: for $x \sim P_\theta$,

$$\mathbb{P}(\theta, \phi) = \sum_{x \in \mathcal{X} : \phi(x) = \phi} P_\theta(x)$$

The Profile MLE

Likelihood of a profile: for $x \sim P_\theta$,

$$\mathbb{P}(\theta, \phi) = \sum_{x \in \mathcal{X} : \phi(x) = \phi} P_\theta(x)$$

Definition (Profile MLE, Orlitsky et al. 2004)

Given samples with profile ϕ, the PMLE is defined as

$$\theta^{\mathsf{PMLE}}(\phi) = \arg\max_{\theta \in \Theta} \mathbb{P}(\theta, \phi)$$

The Profile MLE

Likelihood of a profile: for $x \sim P_\theta$,

$$\mathbb{P}(\theta, \phi) = \sum_{x \in \mathcal{X} : \phi(x) = \phi} P_\theta(x)$$

Definition (Profile MLE, Orlitsky et al. 2004)

Given samples with profile ϕ, the PMLE is defined as

$$\theta^{\text{PMLE}}(\phi) = \arg\max_{\theta \in \Theta} \mathbb{P}(\theta, \phi)$$

Example: if $x \sim P_\theta = \prod_{j=1}^{p} p_{\theta_j}(x_j)$:

$$\theta^{\text{PMLE}} = \arg\max_{\theta} \mathbb{P}(\theta, (x_{(1)}, x_{(2)}, \cdots, x_{(p)})) = \arg\max_{\theta} \sum_{\pi \in S_p} \prod_{j=1}^{p} p_{\theta_j}(x_{\pi(j)})$$

Questions

- Is there an analogy between MLE and PMLE?
- How to analyze the statistical property of PMLE, where both the zeroth-order and first-order conditions look complicated?
- For permutation-invariant models, is PMLE statistically optimal in estimating permutation-invariant targets of θ?
- Is PMLE subject to certain limitations as well?

Questions

- Is there an analogy between MLE and PMLE?
  Yes - MLE is rate-optimal in parameter estimation, and PMLE is rate-optimal in parameter estimation up to permutation.
- How to analyze the statistical property of PMLE, where both the zeroth-order and first-order conditions look complicated?
- For permutation-invariant models, is PMLE statistically optimal in estimating permutation-invariant targets of θ?
- Is PMLE subject to certain limitations as well?

Questions

- Is there an analogy between MLE and PMLE?
  Yes - MLE is rate-optimal in parameter estimation, and PMLE is rate-optimal in parameter estimation up to permutation.
- How to analyze the statistical property of PMLE, where both the zeroth-order and first-order conditions look complicated?
  Using competitive analysis.
- For permutation-invariant models, is PMLE statistically optimal in estimating permutation-invariant targets of θ?
- Is PMLE subject to certain limitations as well?

Questions

- Is there an analogy between MLE and PMLE?
  Yes - MLE is rate-optimal in parameter estimation, and PMLE is rate-optimal in parameter estimation up to permutation.
- How to analyze the statistical property of PMLE, where both the zeroth-order and first-order conditions look complicated?
  Using competitive analysis.
- For permutation-invariant models, is PMLE statistically optimal in estimating permutation-invariant targets of θ?
  Universally true - when the target error is large.
- Is PMLE subject to certain limitations as well?

Questions

- Is there an analogy between MLE and PMLE?
  Yes - MLE is rate-optimal in parameter estimation, and PMLE is rate-optimal in parameter estimation up to permutation.
- How to analyze the statistical property of PMLE, where both the zeroth-order and first-order conditions look complicated?
  Using competitive analysis.
- For permutation-invariant models, is PMLE statistically optimal in estimating permutation-invariant targets of θ?
  Universally true - when the target error is large.
- Is PMLE subject to certain limitations as well?
  Yes - when the target error is small.

PMLE in discrete distribution model

Discrete distribution model

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p = (p_1, \cdots, p_k)$
  - n: sample size
  - k: support size

Discrete distribution model

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p = (p_1, \cdots, p_k)$
  - n: sample size
  - k: support size
- histogram $h = (h_1, \cdots, h_k) \sim \mathrm{Multinomial}(n; p)$ is sufficient, where $h_j = \sum_{i=1}^n 1(X_i = j)$

Discrete distribution model

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p = (p_1, \cdots, p_k)$
  - n: sample size
  - k: support size
- histogram $h = (h_1, \cdots, h_k) \sim \mathrm{Multinomial}(n; p)$ is sufficient, where $h_j = \sum_{i=1}^{n} 1(X_i = j)$
- profile $\phi = \{\pi h : \pi \in S_k\}$ could be represented by a vector $(\phi_1, \cdots, \phi_n)$ with

$$\phi_i = \# \text{ of domain elements appearing exactly } i \text{ times}$$

  - for example, if $x^n =$ "*abaac*", then $\phi = (2, 0, 1, 0, 0)$
  - "histogram of the histogram" with $h = (3, 1, 1)$

Discrete distribution model

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p = (p_1, \cdots, p_k)$
  - n: sample size
  - k: support size
- histogram $h = (h_1, \cdots, h_k) \sim \mathrm{Multinomial}(n; p)$ is sufficient, where $h_j = \sum_{i=1}^n 1(X_i = j)$
- profile $\phi = \{\pi h : \pi \in S_k\}$ could be represented by a vector $(\phi_1, \cdots, \phi_n)$ with

$$\phi_i = \# \text{ of domain elements appearing exactly } i \text{ times}$$

  - for example, if $x^n =$ "$abaac$", then $\phi = (2, 0, 1, 0, 0)$
  - "histogram of the histogram" with $h = (3, 1, 1)$
- since $\pi h \sim \mathrm{Multinomial}(n; \pi p)$, ϕ is sufficient in estimating the sorted version of p and any symmetric functional $\sum_{j=1}^k f(p_j)$

Discrete distribution model

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p = (p_1, \cdots, p_k)$
  - n: sample size
  - k: support size
- histogram $h = (h_1, \cdots, h_k) \sim \mathrm{Multinomial}(n; p)$ is sufficient, where $h_j = \sum_{i=1}^n 1(X_i = j)$
- profile $\phi = \{\pi h : \pi \in S_k\}$ could be represented by a vector $(\phi_1, \cdots, \phi_n)$ with
$$\phi_i = \#\text{ of domain elements appearing exactly } i \text{ times}$$
  - for example, if $x^n =$ "*abaac*", then $\phi = (2, 0, 1, 0, 0)$
  - "histogram of the histogram" with $h = (3, 1, 1)$
- since $\pi h \sim \mathrm{Multinomial}(n; \pi p)$, ϕ is sufficient in estimating the sorted version of p and any symmetric functional $\sum_{j=1}^k f(p_j)$
- PMLE:
$$p^{\mathsf{PMLE}} = \arg\max_p \sum_{\pi \in S_k} \prod_{j=1}^k p_j^{h_{\pi(j)}}$$

Some PMLE Examples

Example I: $X^n = aba$ with $n = 3$ and $k = 2$

Some PMLE Examples

Example I: $X^n = aba$ with $n = 3$ and $k = 2$

- MLE: $p^{\text{MLE}} = (2/3, 1/3)$

Some PMLE Examples

Example I: $X^n = aba$ with $n = 3$ and $k = 2$

- MLE: $p^{\mathsf{MLE}} = (2/3, 1/3)$
- PMLE: $p^{\mathsf{PMLE}} = (1/2, 1/2)$

Some PMLE Examples

Example I: $X^n = aba$ with $n = 3$ and $k = 2$

- MLE: $p^{\text{MLE}} = (2/3, 1/3)$
- PMLE: $p^{\text{PMLE}} = (1/2, 1/2)$

Example II: $X^n = abac$ with $n = 4$ and $k = 5$

Some PMLE Examples

Example I: $X^n = aba$ with $n = 3$ and $k = 2$

- MLE: $p^{\mathsf{MLE}} = (2/3, 1/3)$
- PMLE: $p^{\mathsf{PMLE}} = (1/2, 1/2)$

Example II: $X^n = abac$ with $n = 4$ and $k = 5$

- MLE: $p^{\mathsf{MLE}} = (1/2, 1/4, 1/4, 0, 0)$

Some PMLE Examples

Example I: $X^n = aba$ with $n = 3$ and $k = 2$

- MLE: $p^{\text{MLE}} = (2/3, 1/3)$
- PMLE: $p^{\text{PMLE}} = (1/2, 1/2)$

Example II: $X^n = abac$ with $n = 4$ and $k = 5$

- MLE: $p^{\text{MLE}} = (1/2, 1/4, 1/4, 0, 0)$
- PMLE: $p^{\text{PMLE}} = (1/5, 1/5, 1/5, 1/5, 1/5)$

Computational burden

$$p^{\text{PMLE}} = \arg\max_p \sum_{\pi \in S_k} \prod_{j=1}^{k} p_j^{h_{\pi(j)}}$$

- non-concave, sum of exponentially many terms
- very hard to compute or even approximate PMLE in general

Computational burden

$$p^{\mathrm{PMLE}} = \arg\max_p \sum_{\pi \in S_k} \prod_{j=1}^{k} p_j^{h_{\pi(j)}}$$

- non-concave, sum of exponentially many terms
- very hard to compute or even approximate PMLE in general

Heuristic algorithms:

- [Orlitsky et al., 2004]: EM-type algorithm
- [Acharya et al., 2010]: symmetric polynomial evaluation
- [Vontobel, 2012, 2014]: Bethe/Sinkhorn approximation of permanent
- [Pavlichin, Jiao, and Weissman, 2019]: dynamic programming

Computational burden

$$p^{\mathrm{PMLE}} = \arg\max_p \sum_{\pi \in S_k} \prod_{j=1}^{k} p_j^{h_{\pi(j)}}$$

- non-concave, sum of exponentially many terms
- very hard to compute or even approximate PMLE in general

Heuristic algorithms:

- [Orlitsky et al., 2004]: EM-type algorithm
- [Acharya et al., 2010]: symmetric polynomial evaluation
- [Vontobel, 2012, 2014]: Bethe/Sinkhorn approximation of permanent
- [Pavlichin, Jiao, and Weissman, 2019]: dynamic programming

Provable approximate algorithms: $\mathbb{P}(\widehat{p}, \phi) \geq \beta \cdot \mathbb{P}(p^{\mathrm{PMLE}}, \phi)$

- [Charikar, Shiragur, and Sidford, 2019]: $\beta = \exp(-n^{2/3} \log n)$
- [Anari et al., 2020a, 2020b]: $\beta = \exp(-\min\{\sqrt{n}, k\} \log n)$

Statistical guarantee

Challenge: very few properties of PMLE could be said except for its defining property

Statistical guarantee

Challenge: very few properties of PMLE could be said except for its defining property

A recent breakthrough:

Theorem (Acharya, Das, Orlitsky, and Suresh, 2017)

For any metric d and accuracy level $\varepsilon > 0$,

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(p^{\mathrm{PMLE}}, p) > 2\varepsilon) \le e^{3\sqrt{n}} \cdot \inf_{\widehat{p}(\phi)} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon)$$

Statistical guarantee

Challenge: very few properties of PMLE could be said except for its defining property

A recent breakthrough:

Theorem (Acharya, Das, Orlitsky, and Suresh, 2017)

For any metric d and accuracy level $\varepsilon > 0$,

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(p^{\text{PMLE}}, p) > 2\varepsilon) \le e^{3\sqrt{n}} \cdot \inf_{\widehat{p}(\phi)} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon)$$

Corollary: as in many examples we have

$$\inf_{\widehat{p}(\phi)} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon) \lesssim \exp\left(-n(\varepsilon - \varepsilon_{n,k})_+^2\right),$$

if n is the minimax sample complexity of achieving accuracy $\varepsilon/2$, the PMLE attains the rate-optimal sample complexity if $\varepsilon \gg n^{-1/4}$.

Improving the exponent

- [Charikar, Shiragur, and Sidford, 2019, Hao and Orlitsky, 2019]: exponent polylog(n) for a (very) restricted class of d and modified PMLE
- [Hao and Orlitsky, 2020]: distribution-dependent exponent $H_n(p)$ with $\sup_p H_n(p) \asymp \sqrt{n}$

Improving the exponent

- [Charikar, Shiragur, and Sidford, 2019, Hao and Orlitsky, 2019]: exponent polylog(n) for a (very) restricted class of d and modified PMLE
- [Hao and Orlitsky, 2020]: distribution-dependent exponent $H_n(p)$ with $\sup_p H_n(p) \asymp \sqrt{n}$

An open question

What is the tight exponent for the competitive analysis of the PMLE?

Main results

Result I: improved competitive analysis of PML

Theorem (H. and Shiragur, 2021)

For any metric d, accuracy level $\varepsilon > 0$ and constant $c \in (0,1)$, we have

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(p^{\mathrm{PMLE}}, p) > 2\varepsilon) \le \exp\left(c' n^{1/3+c}\right) \cdot \inf_{\widehat{p}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon)^{1-c},$$

for some constant c' depending only on c.

Result I: improved competitive analysis of PML

Theorem (H. and Shiragur, 2021)

For any metric d, accuracy level $\varepsilon > 0$ and constant $c \in (0, 1)$, we have

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(p^{\mathsf{PMLE}}, p) > 2\varepsilon)$$
$$\leq \exp\left(c' n^{1/3+c}\right) \cdot \inf_{\widehat{p}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon)^{1-c},$$

for some constant c' depending only on c.

- exponent improved from $O(\sqrt{n})$ to $O(n^{1/3+c})$
- for any β-approximate PMLE, the competitive factor becomes $\exp(c' n^{1/3+c})/\beta$

Result II: optimality of exponent

Theorem (H., 2021)

For any $c, c', c_1, c_2 > 0$, there exists a metric d and accuracy level $\varepsilon > 0$ such that

$$\sup_{p\in\mathcal{M}_k} \mathbb{P}_p(d(p^{\mathrm{PMLE}}, p) > c_1\varepsilon)$$

$$\gg \exp\left(c' n^{1/3-c}\right) \cdot \inf_{\widehat{p}} \sup_{p\in\mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon)^{1-c_2}.$$

Result II: optimality of exponent

Theorem (H., 2021)

For any $c, c', c_1, c_2 > 0$, there exists a metric d and accuracy level $\varepsilon > 0$ such that

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(p^{\mathrm{PMLE}}, p) > c_1 \varepsilon)$$
$$\gg \exp\left(c' n^{1/3-c}\right) \cdot \inf_{\widehat{p}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon)^{1-c_2}.$$

- the exponent $O(n^{1/3-c})$ is not generically attainable for PMLE
- the competitive factor $\exp(O(n^{1/3}))$ is optimal and not superfluous

Result III: PMLE estimates sorted distribution optimally

Theorem (H. and Shiragur, 2021)

The PMLE satisfies that

$$\sup_{p \in \mathcal{M}_k} \mathbb{E}_p \| p^{\mathrm{PML}} - p \|_{1,\mathrm{sorted}} \lesssim \sqrt{\frac{k}{n \log n}} + \widetilde{O}\left(n^{-1/3} \wedge \sqrt{\frac{k}{n}} \right).$$

Result III: PMLE estimates sorted distribution optimally

Theorem (H. and Shiragur, 2021)

The PMLE satisfies that

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

- minimax rate-optimal for estimating sorted distribution
- attains optimal phase transition at $k \asymp n^{1/3}$
- [Acharya et al., 2012]: requires $k \gtrsim n$
- [Hao and Orlitsky, 2019]: requires $k \gtrsim n^{0.8}$
- [Hao and Orlitsky, 2020]: requires $k \gtrsim n^{0.75}$

Application in symmetric functional estimation

Symmetric 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

Symmetric 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 $f(x) = \mathbb{1}(x \neq 0)$

Symmetric 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 $f(x) = \mathbb{1}(x \neq 0)$

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

Symmetric 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 $f(x) = \mathbb{1}(x \neq 0)$

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

Generalization: non-symmetric, multivariate and nonparametric functionals

Ad-hoc estimation

Plug-in of MLE: $\widehat{F} = F(p^{\mathrm{MLE}})$

Ad-hoc estimation

Plug-in of MLE: $\widehat{F} = F(p^{\mathrm{MLE}})$

Effective sample size enlargement

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

Ad-hoc estimation

Plug-in of MLE: $\widehat{F} = F(p^{\mathrm{MLE}})$

Effective sample size enlargement

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)
- ...

Universal estimation

Target

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

Universal estimation

Target

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

$$X_1, \cdots, X_n$$

Universal estimation

Target

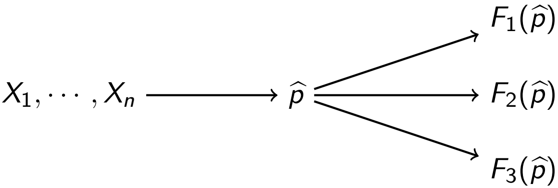
Find a single distribution estimator $\widehat{p}$ such that the plugging $\widehat{p}$ into the functional is universally optimal for "many" functionals

$$X_1, \cdots, X_n \longrightarrow \widehat{p}$$

Universal estimation

Target

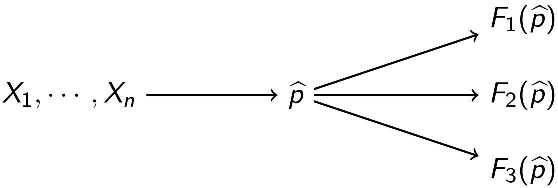
Find a single distribution estimator $\widehat{p}$ such that the plugging $\widehat{p}$ into the functional is universally optimal for "many" functionals



Universal estimation

Target

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

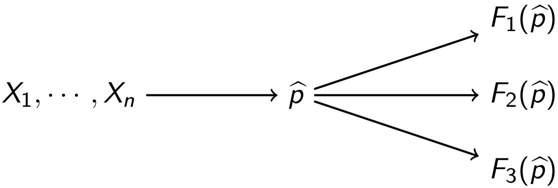


Too good to be true?

Universal estimation

Target

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



Too good to be true? No!

Result IV: universal optimality of PMLE

Theorem (H. and Shiragur, 2021)

For symmetric functionals including:

- Shannon entropy;
- support size;
- support coverage;
- distance to uniformity and general 1-Lipschitz functionals,

the plug-in approach of the PMLE universally attains the optimal sample complexity of achieving an accuracy level $\varepsilon \gg n^{-1/3}$.

Result IV: universal optimality of PMLE

Theorem (H. and Shiragur, 2021)

For symmetric functionals including:

- Shannon entropy;
- support size;
- support coverage;
- distance to uniformity and general 1-Lipschitz functionals,

the plug-in approach of the PMLE universally attains the optimal sample complexity of achieving an accuracy level $\varepsilon \gg n^{-1/3}$.

- Proof: choose $d(p, q) = |F(p) - F(q)|$, and construct minimax rate-optimal estimator for F

Result V: limitation of PMLE

Theorem (H., 2021)

There exists a 1-Lipschitz functional F such that

$$\sup_{p \in \mathcal{M}_k} \mathbb{E}_p |F(p^{\mathrm{PMLE}}) - F(p)| \asymp \begin{cases} \sqrt{\frac{k}{n \log n}} & \text{if } k \gg n^{1/3} \\ \sqrt{\frac{k}{n}} & \text{if } 1 \ll k \ll n^{1/3} \end{cases}$$

Result V: limitation of PMLE

Theorem (H., 2021)

There exists a 1-Lipschitz functional F such that

$$\sup_{p\in\mathcal{M}_k} \mathbb{E}_p|F(p^{\mathrm{PMLE}}) - F(p)| \asymp \begin{cases} \sqrt{\frac{k}{n\log n}} & \text{if } k \gg n^{1/3} \\ \sqrt{\frac{k}{n}} & \text{if } 1 \ll k \ll n^{1/3} \end{cases}$$

In contrast, [Hao and Orlitsky, 2019] shows that for every 1-Lipschitz functional F,

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

Result V: limitation of PMLE

Theorem (H., 2021)

There exists a 1-Lipschitz functional F such that

$$\sup_{p\in\mathcal{M}_k} \mathbb{E}_p|F(p^{\text{PMLE}}) - F(p)| \asymp \begin{cases} \sqrt{\frac{k}{n\log n}} & \text{if } k \gg n^{1/3} \\ \sqrt{\frac{k}{n}} & \text{if } 1 \ll k \ll n^{1/3} \end{cases}$$

In contrast, [Hao and Orlitsky, 2019] shows that for every 1-Lipschitz functional F,

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

- PMLE fails to be optimal when $k \ll n^{1/3}$, or equivalently, $\varepsilon \ll n^{-1/3}$

Result VI: optimality among universal approaches

Theorem (H., 2021)

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

Result VI: optimality among universal approaches

Theorem (H., 2021)

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

- not only the limitation of PMLE, but also the limitation of all possible universal approaches!

Result VI: optimality among universal approaches

Theorem (H., 2021)

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

- not only the limitation of PMLE, but also the limitation of all possible universal approaches!
- a smaller quantity [Hao and Orlitsky, 2019]:

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

Result VI: optimality among universal approaches

Theorem (H., 2021)

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

- not only the limitation of PMLE, but also the limitation of all possible universal approaches!
- a smaller quantity [Hao and Orlitsky, 2019]:

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

- A larger quantity [H., Jiao, and Weissman, 2018]:

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

Summary of approaches

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

Summary of approaches

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

Tight statistical analysis of PML: optimality and limitation

Proof sketch of improved competitive analysis

Review: idea of [Acharya et al., 2017]

Notations:

- Φ_n: the set of all possible profiles with sample size n
- ϕ: a particular profile in Φ_n
- p_ϕ: the PMLE associated with ϕ
- $\mathbb{P}(p, \phi)$: probability of observing ϕ under the true distribution p

Review: idea of [Acharya et al., 2017]

Notations:

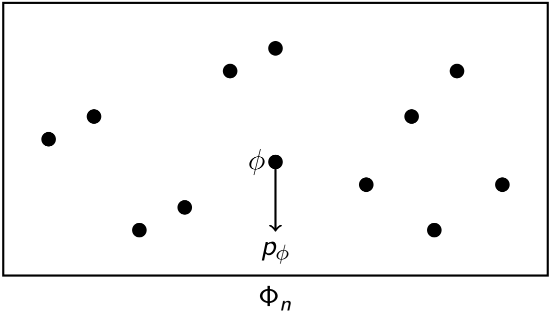
- Φ_n: the set of all possible profiles with sample size n
- ϕ: a particular profile in Φ_n
- p_ϕ: the PMLE associated with ϕ
- $\mathbb{P}(p, \phi)$: probability of observing ϕ under the true distribution p

Technical goal: using only the defining property $\mathbb{P}(p_\phi, \phi) \ge \mathbb{P}(p, \phi)$, find an upper bound of

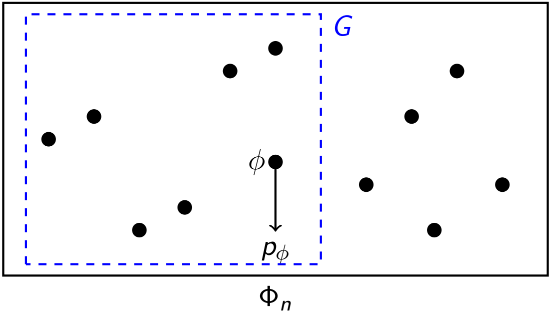
$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(p_\phi, p) > 2\varepsilon)$$

given an estimator $\widehat{p}(\phi)$ with $\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(d(\widehat{p}, p) > \varepsilon) \le \delta$.

Analysis in [Acharya et al., 2017]



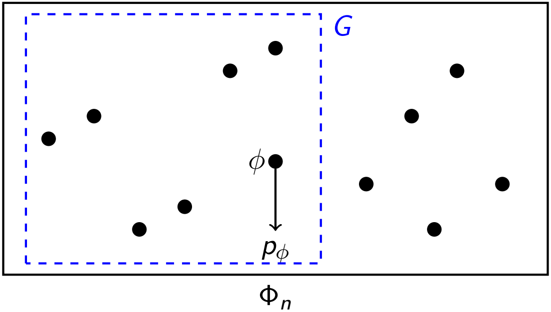
Analysis in [Acharya et al., 2017]



Good profile:

$$G = \{\phi \in \Phi_n : d(\widehat{p}(\phi), p) \leq \varepsilon\}$$

Analysis in [Acharya et al., 2017]

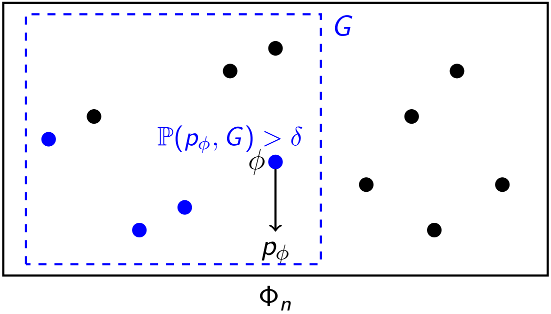


Good profile:

$$G = \{\phi \in \Phi_n : d(\widehat{p}(\phi), p) \leq \varepsilon\}$$

Clearly $\mathbb{P}(p, G) \geq 1 - \delta$.

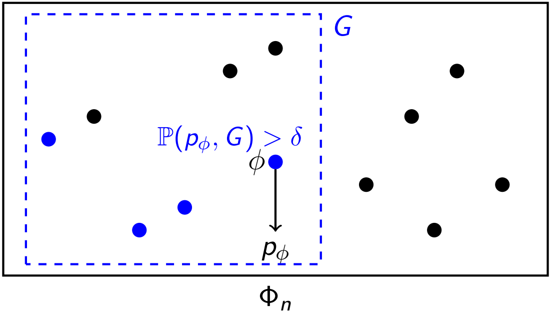
Analysis in [Acharya et al., 2017]



Lemma

For any $\phi \in G$ satisfying $\mathbb{P}(p_\phi, G) > \delta$, we have $d(p_\phi, p) \leq 2\varepsilon$.

Analysis in [Acharya et al., 2017]

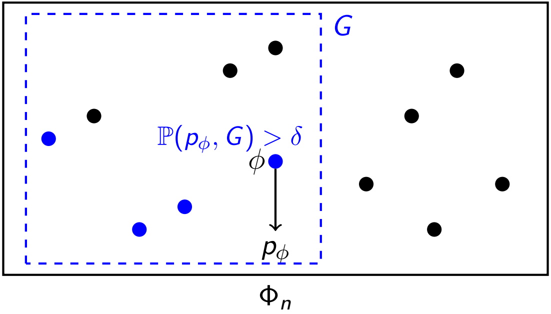


Lemma

For any $\phi \in G$ satisfying $\mathbb{P}(p_\phi, G) > \delta$, we have $d(p_\phi, p) \leq 2\varepsilon$.

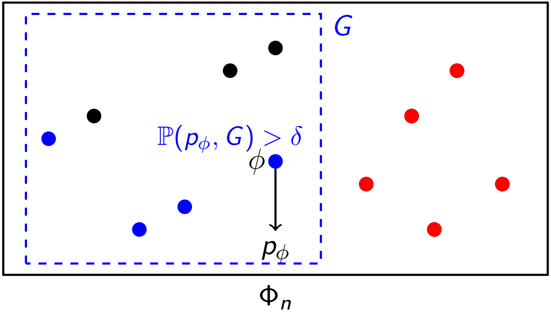
Proof: $\mathbb{P}(p_\phi, G) > \delta \implies d(\widehat{p}(\phi'), p_\phi) \leq \varepsilon$ for some $\phi' \in G$. Also, definition of $G \implies d(\widehat{p}(\phi'), p) \leq \varepsilon$. □

Analysis in [Acharya et al., 2017]



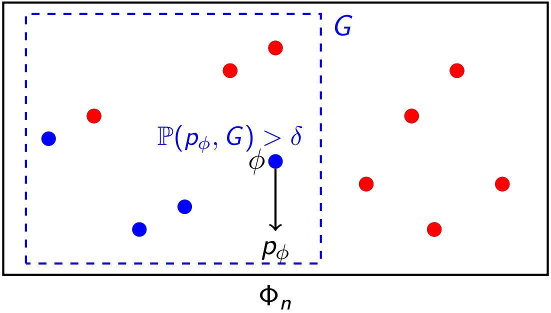
$\mathbb{P}_p(d(p_\phi, p) > 2\varepsilon)$

Analysis in [Acharya et al., 2017]



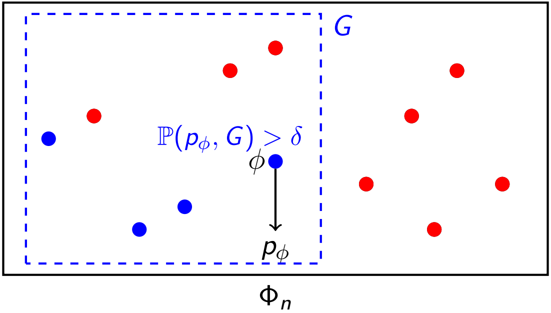
$\mathbb{P}_p(d(p_\phi, p) > 2\varepsilon) \leq \mathbb{P}(p, G^c)$

Analysis in [Acharya et al., 2017]



$$\mathbb{P}_p(d(p_\phi, p) > 2\varepsilon) \leq \mathbb{P}(p, G^c) + \sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p_\phi, G) \leq \delta)$$

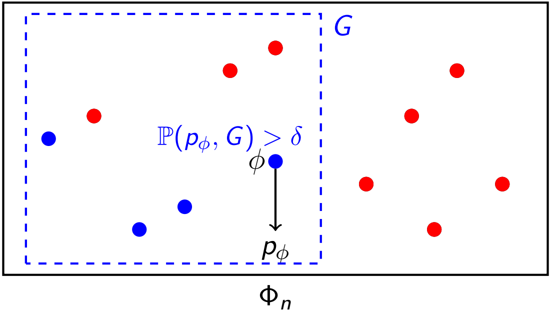
Analysis in [Acharya et al., 2017]



$$\mathbb{P}_p(d(p_\phi, p) > 2\varepsilon) \le \mathbb{P}(p, G^c) + \sum_{\phi \in G} \mathbb{P}(p, \phi)\mathbb{1}(\mathbb{P}(p_\phi, G) \le \delta)$$
$$\le \delta + \sum_{\phi \in G} \mathbb{P}(p, \phi)\mathbb{1}(\mathbb{P}(p, \phi) \le \delta)$$

for $\mathbb{P}(p_\phi, G) \ge \mathbb{P}(p_\phi, \phi) \ge \mathbb{P}(p, \phi)$.

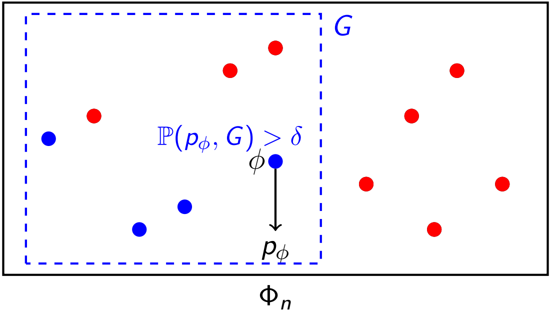
Analysis in [Acharya et al., 2017]



$$\begin{aligned}\mathbb{P}_p(d(p_\phi, p) > 2\varepsilon) &\leq \mathbb{P}(p, G^c) + \sum_{\phi \in G} \mathbb{P}(p, \phi)\mathbb{1}(\mathbb{P}(p_\phi, G) \leq \delta) \\ &\leq \delta + \sum_{\phi \in G} \mathbb{P}(p, \phi)\mathbb{1}(\mathbb{P}(p, \phi) \leq \delta) \\ &\leq (1 + |\Phi_n|) \cdot \delta\end{aligned}$$

for $\mathbb{P}(p_\phi, G) \geq \mathbb{P}(p_\phi, \phi) \geq \mathbb{P}(p, \phi)$.

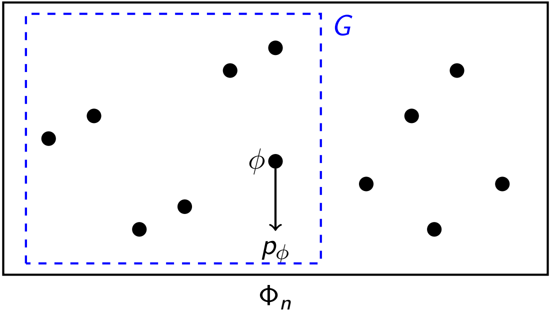
Analysis in [Acharya et al., 2017]



$$\begin{aligned}
\mathbb{P}_p(d(p_\phi, p) > 2\varepsilon) &\le \mathbb{P}(p, G^c) + \sum_{\phi \in G} \mathbb{P}(p, \phi)\mathbb{1}(\mathbb{P}(p_\phi, G) \le \delta) \\
&\le \delta + \sum_{\phi \in G} \mathbb{P}(p, \phi)\mathbb{1}(\mathbb{P}(p, \phi) \le \delta) \\
&\le (1 + |\Phi_n|) \cdot \delta \le \exp(3\sqrt{n}) \cdot \delta,
\end{aligned}$$

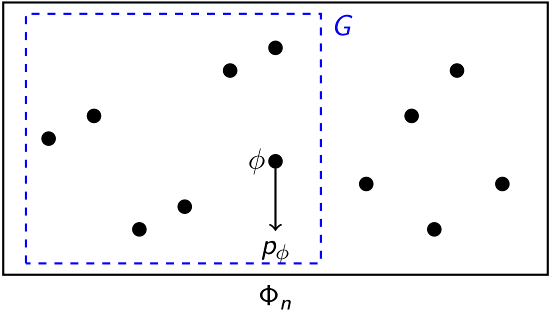
for $\mathbb{P}(p_\phi, G) \ge \mathbb{P}(p_\phi, \phi) \ge \mathbb{P}(p, \phi)$.

Our proof idea



A potentially loose inequality: $\mathbb{P}(p_\phi, G) \geq \mathbb{P}(p_\phi, \phi)$ for $\phi \in G$

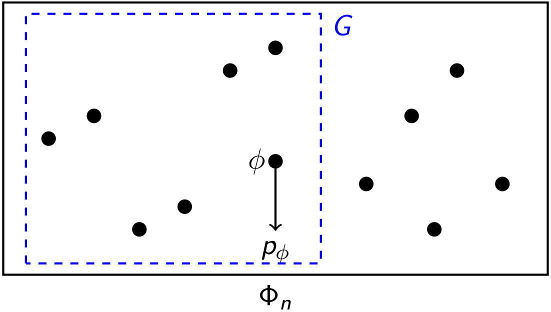
Our proof idea



A potentially loose inequality: $\mathbb{P}(p_\phi, G) \geq \mathbb{P}(p_\phi, \phi)$ for $\phi \in G$

- could be tight when p_ϕ is essentially supported on ϕ

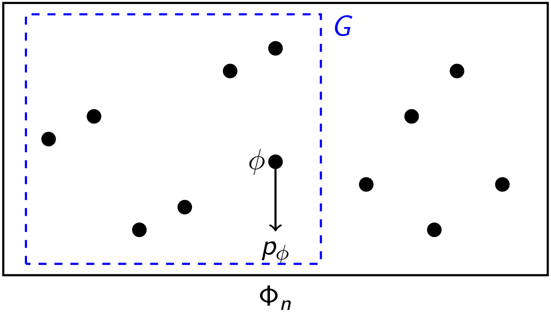
Our proof idea



A potentially loose inequality: $\mathbb{P}(p_\phi, G) \geq \mathbb{P}(p_\phi, \phi)$ for $\phi \in G$

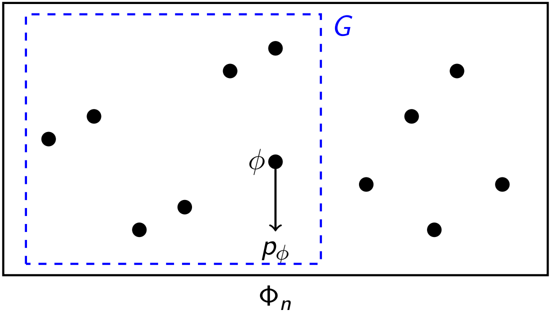
- could be tight when p_ϕ is essentially supported on ϕ
- in that case, $\mathbb{P}(p_{\phi'}, \phi) \ll \mathbb{P}(p_\phi, \phi)$

Our proof idea



Q: What if we could have $\mathbb{P}(p_\phi, \phi) \approx \mathbb{P}(p_{\phi'}, \phi)$ for all $\phi, \phi' \in G$?

Our proof idea

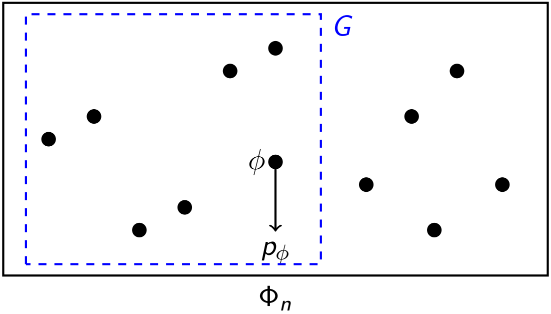


Q: What if we could have $\mathbb{P}(p_\phi, \phi) \approx \mathbb{P}(p_{\phi'}, \phi)$ for all $\phi, \phi' \in G$?
A: Then we are in a great shape, for if $\mathbb{P}(p_{\phi'}, G) < \delta$ for some $\phi' \in G$, then

$$\delta > \mathbb{P}(p_{\phi'}, G) = \sum_{\phi \in G} \mathbb{P}(p_{\phi'}, \phi) \approx \sum_{\phi \in G} \mathbb{P}(p_\phi, \phi) \geq \sum_{\phi \in G} \mathbb{P}(p, \phi) = \mathbb{P}(p, G),$$

a contradiction to $\mathbb{P}(p, G) \geq 1 - \delta$.

Our proof idea



Idea

Improved bound if we could show certain "continuity" property of $\phi \mapsto p_\phi$.

Key covering lemma

Covering lemma

Let $0 < s < r < 1/2$ be any fixed constants. There exists a discrete set of profiles $\Phi \subseteq \Phi_n$ such that:

- the new set Φ has a smaller cardinality $|\Phi| \leq \exp(n^r \log n)$;
- every profile $\phi \in \Phi_n$ could be approximated by some profile $\phi' \in \Phi$ in the following sense: for all $S \subseteq \Phi_n$,

$$\mathbb{P}(p_\phi, S) \geq \mathbb{P}(p_{\phi'}, S)^{1/(1-n^{-s})} \cdot \exp\left(-cn^{1-2r+s}\right),$$

$$\mathbb{P}(p_{\phi'}, S) \geq \mathbb{P}(p_\phi, S)^{1/(1-n^{-s})} \cdot \exp\left(-cn^{1-2r+s}\right),$$

  where $c = c(r, s) > 0$.

Key covering lemma

Covering lemma

Let $0 < s < r < 1/2$ be any fixed constants. There exists a discrete set of profiles $\Phi \subseteq \Phi_n$ such that:

- the new set Φ has a smaller cardinality $|\Phi| \le \exp(n^r \log n)$;
- every profile $\phi \in \Phi_n$ could be approximated by some profile $\phi' \in \Phi$ in the following sense: for all $S \subseteq \Phi_n$,

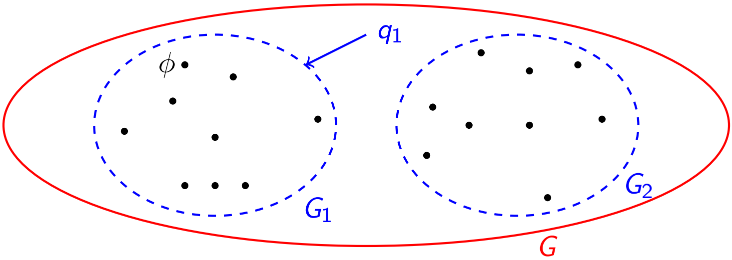
$$\mathbb{P}(p_\phi, S) \ge \mathbb{P}(p_{\phi'}, S)^{1/(1-n^{-s})} \cdot \exp\left(-cn^{1-2r+s}\right),$$
$$\mathbb{P}(p_{\phi'}, S) \ge \mathbb{P}(p_\phi, S)^{1/(1-n^{-s})} \cdot \exp\left(-cn^{1-2r+s}\right),$$

  where $c = c(r, s) > 0$.

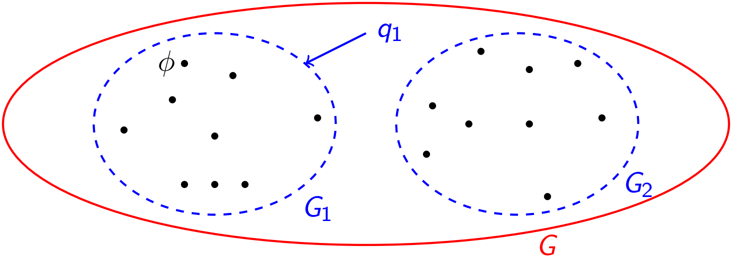
A covering property of PML distributions $\{p_\phi : \phi \in \Phi_n\}$

- $r \uparrow$: the cardinality $\uparrow$, approximation exponent $\downarrow$
- $s \uparrow$: probability exponent $\downarrow$, multiplicative exponent $\uparrow$

Applying the covering lemma with $r = 3/8, s = 1/8$



Applying the covering lemma with $r = 3/8, s = 1/8$

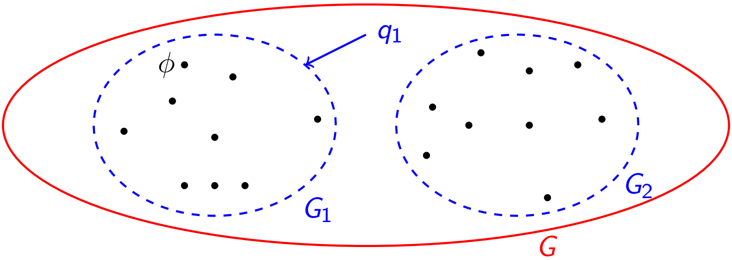


If $\mathbb{P}(p_\phi, G_1) \leq \delta$, then

$$\delta \geq \mathbb{P}(p_\phi, G_1) \geq \mathbb{P}(q_1, G_1)^{1/(1-n^{-1/8})} \cdot \exp(-cn^{3/8})$$

$$\implies \mathbb{P}(q_1, G_1) \leq \delta^{1-o(1)} \cdot \exp(cn^{3/8})$$

Applying the covering lemma with $r = 3/8, s = 1/8$

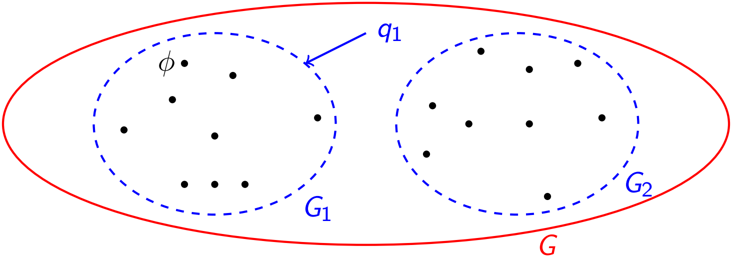


If $\mathbb{P}(p_\phi, G_1) \leq \delta$, then

$$\delta \geq \mathbb{P}(p_\phi, G_1) \geq \mathbb{P}(q_1, G_1)^{1/(1-n^{-1/8})} \cdot \exp(-cn^{3/8})$$
$$\Longrightarrow \mathbb{P}(q_1, G_1) \leq \delta^{1-o(1)} \cdot \exp(cn^{3/8})$$

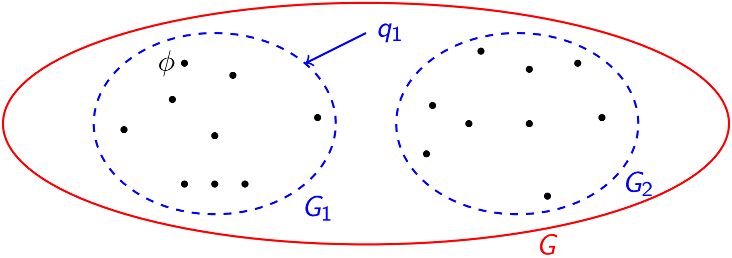
"going-down process"

Applying the covering lemma with $r = 3/8, s = 1/8$



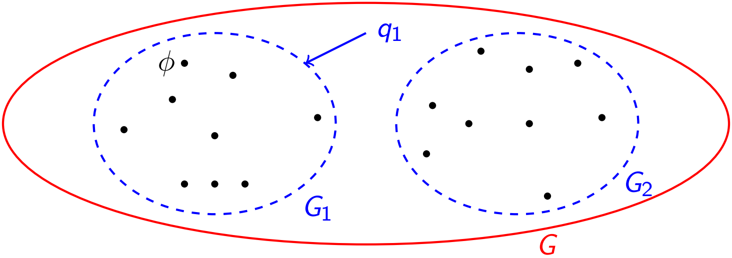
$\mathbb{P}(q_1, G_1)$

Applying the covering lemma with $r = 3/8, s = 1/8$



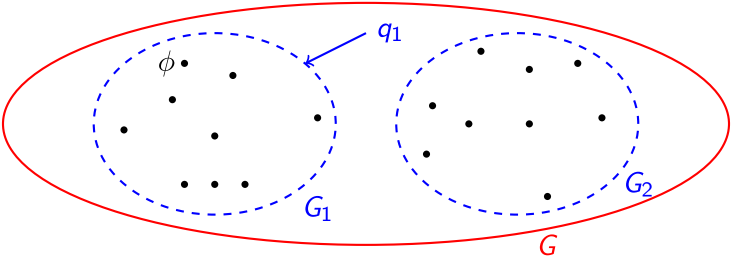
$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi)$$

Applying the covering lemma with $r = 3/8, s = 1/8$



$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi) \geq \exp(-cn^{3/8}) \sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi)^{1/(1-n^{-1/8})}$$

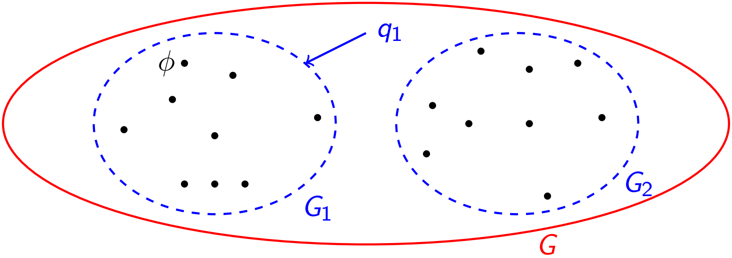
Applying the covering lemma with $r = 3/8, s = 1/8$



$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi) \geq \exp(-cn^{3/8}) \sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi)^{1/(1-n^{-1/8})}$$

$$\geq \exp(-cn^{3/8}) \left(\sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi) \right)^{1/(1-n^{-1/8})} \cdot |G_1|^{-n^{-1/8}/(1-n^{-1/8})}$$

Applying the covering lemma with $r = 3/8, s = 1/8$

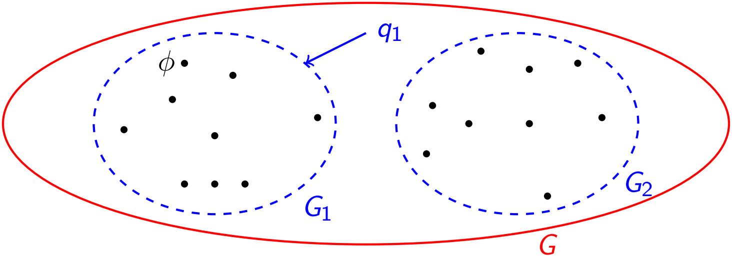


$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi) \geq \exp(-cn^{3/8}) \sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi)^{1/(1-n^{-1/8})}$$

$$\geq \exp(-cn^{3/8}) \left(\sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi) \right)^{1/(1-n^{-1/8})} \cdot |G_1|^{-n^{-1/8}/(1-n^{-1/8})}$$

$$\geq \mathbb{P}(p, G_1)^{1+o(1)} \cdot \exp(-cn^{3/8})$$

Applying the covering lemma with $r = 3/8, s = 1/8$



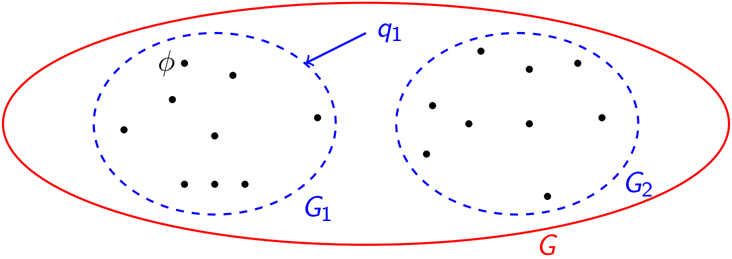
$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi) \geq \exp(-cn^{3/8}) \sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi)^{1/(1-n^{-1/8})}$$

$$\geq \exp(-cn^{3/8}) \left(\sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi) \right)^{1/(1-n^{-1/8})} \cdot |G_1|^{-n^{-1/8}/(1-n^{-1/8})}$$

$$\geq \mathbb{P}(p, G_1)^{1+o(1)} \cdot \exp(-cn^{3/8})$$

"going-up" process

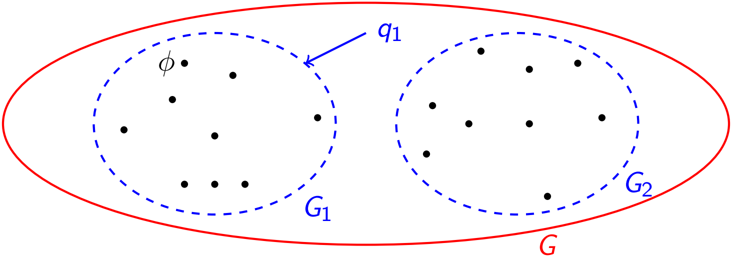
Applying the covering lemma with $r = 3/8, s = 1/8$



Conclusion: if $\mathbb{P}(p_\phi, G_1) \le \delta$ for some $\phi \in G_1$, then

$$\mathbb{P}(p, G_1) \le \delta^{1-o(1)} \cdot \exp(cn^{3/8}).$$

Applying the covering lemma with $r = 3/8, s = 1/8$



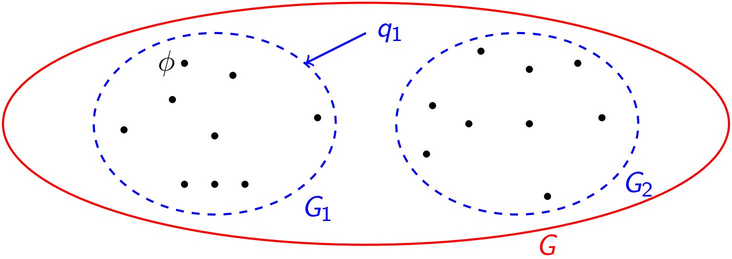
Conclusion: if $\mathbb{P}(p_\phi, G_1) \le \delta$ for some $\phi \in G_1$, then

$$\mathbb{P}(p, G_1) \le \delta^{1-o(1)} \cdot \exp(cn^{3/8}).$$

Using $|\Phi| \le \exp(n^{3/8} \log n)$, we have

$$\sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p_\phi, G) \le \delta) \le \delta^{1-o(1)} \cdot \exp(cn^{3/8} \log n).$$

Applying the covering lemma with $r = 3/8, s = 1/8$



Conclusion: if $\mathbb{P}(p_\phi, G_1) \leq \delta$ for some $\phi \in G_1$, then

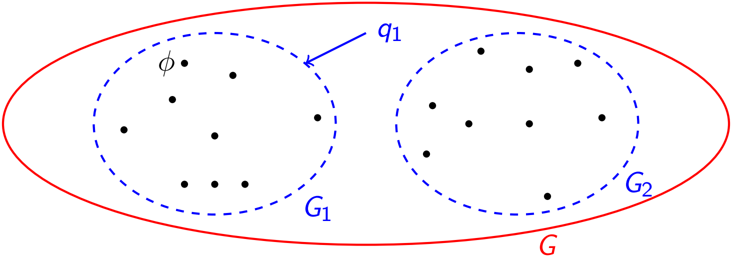
$$\mathbb{P}(p, G_1) \leq \delta^{1-o(1)} \cdot \exp(cn^{3/8}).$$

Using $|\Phi| \leq \exp(n^{3/8} \log n)$, we have

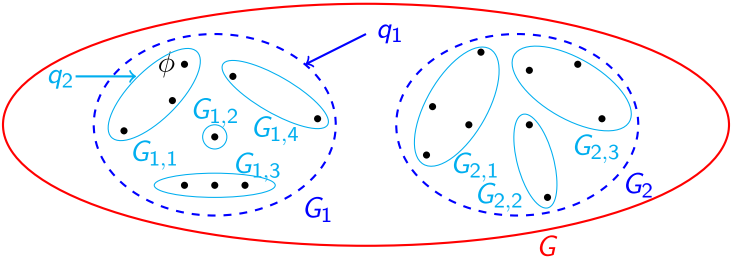
$$\sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p_\phi, G) \leq \delta) \leq \delta^{1-o(1)} \cdot \exp(cn^{3/8} \log n).$$

Already improves over $\exp(3\sqrt{n})$!

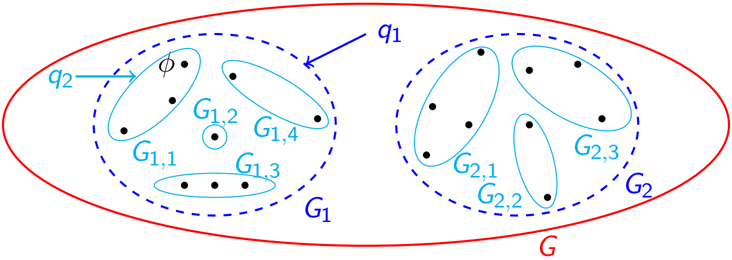
General case: chaining



General case: chaining

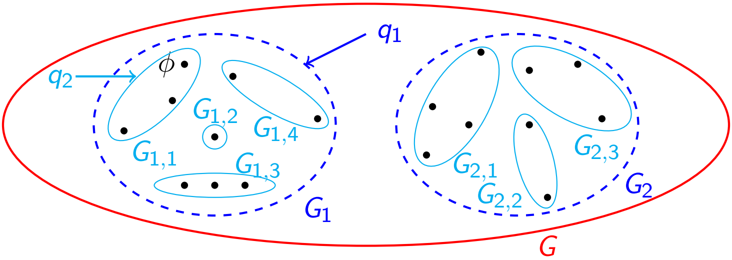


General case: chaining



- "going-down": move along $\mathbb{P}(p_\phi, G_1) \to \mathbb{P}(q_2, G_1) \to \mathbb{P}(q_1, G_1)$

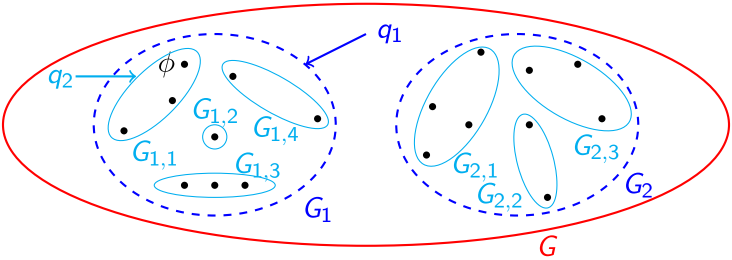
General case: chaining



- "going-down": move along $\mathbb{P}(p_\phi, G_1) \to \mathbb{P}(q_2, G_1) \to \mathbb{P}(q_1, G_1)$
- "going-up": move along

$$\mathbb{P}(q_1, G_1) \to \sum \mathbb{P}(q_2, G_{1,1}) \to \sum \sum \mathbb{P}(p_\phi, \phi) \to \mathbb{P}(p, G_1)$$

General case: chaining



- "going-down": move along $\mathbb{P}(p_\phi, G_1) \to \mathbb{P}(q_2, G_1) \to \mathbb{P}(q_1, G_1)$
- "going-up": move along

$$\mathbb{P}(q_1, G_1) \to \sum \mathbb{P}(q_2, G_{1,1}) \to \sum\sum \mathbb{P}(p_\phi, \phi) \to \mathbb{P}(p, G_1)$$

- choice of parameters: choose $(r_1, s_1), (r_2, s_2), \cdots$ to obtain exponents

$$\frac{3}{8} \to \frac{7}{20} \to \frac{15}{44} \to \cdots \to \frac{1}{3}$$

Generalization to Gaussian location model

PMLE in Gaussian location model

Theorem

For $X \sim \mathcal{N}(\theta, I_p)$, the PMLE satisfies

$$\sup_{\|\theta\|_\infty \le M} \mathbb{P}_\theta(d(\theta^{\mathsf{PMLE}}, \theta) \ge 2\varepsilon)$$

$$\le \exp\left(\widetilde{O}(p^{1/3} M^{2/3})\right) \cdot \inf_{\widehat{\theta}(\phi)} \sup_{\|\theta\|_\infty \le M} \mathbb{P}_\theta(d(\widehat{\theta}, \theta) \ge \varepsilon)^{1-o(1)} + \frac{1}{\mathrm{poly}(p)}$$

PMLE in Gaussian location model

Theorem

For $X \sim \mathcal{N}(\theta, I_p)$, the PMLE satisfies

$$\sup_{\|\theta\|_\infty \le M} \mathbb{P}_\theta(d(\theta^{\mathsf{PMLE}}, \theta) \ge 2\varepsilon)$$

$$\le \exp\left(\widetilde{O}(p^{1/3} M^{2/3})\right) \cdot \inf_{\widehat{\theta}(\phi)} \sup_{\|\theta\|_\infty \le M} \mathbb{P}_\theta(d(\widehat{\theta}, \theta) \ge \varepsilon)^{1-o(1)} + \frac{1}{\mathrm{poly}(p)}$$

- main technical challenge: continuous values of X

Implication on sorted parameter estimation

Corollary

It holds that

$$\sup_{\|\theta\|_\infty \le 1} \mathbb{E}_\theta \|\theta^{\mathsf{PMLE}} - \theta\|_{1,\mathsf{sorted}} \lesssim p \cdot \frac{\log\log p}{\log p}.$$

- matching the minimax risk obtained in [Niles-Weed and Rigollet, 2019]

Concluding remarks

- Is there an analogy between MLE and PMLE?
  Yes - MLE is rate-optimal in parameter estimation, and PMLE is rate-optimal in parameter estimation up to permutation.
- How to analyze the statistical property of PMLE, where both the zeroth-order and first-order conditions look complicated?
  Using competitive analysis.
- For permutation-invariant models, is PMLE statistically optimal in estimating permutation-invariant targets of θ?
  Universally true - when the target error is large.
- Is PMLE subject to certain limitations as well?
  Yes - when the target error is small.

Concluding remarks

- Is there an analogy between MLE and PMLE?
  Yes - MLE is rate-optimal in parameter estimation, and PMLE is rate-optimal in parameter estimation up to permutation.
- How to analyze the statistical property of PMLE, where both the zeroth-order and first-order conditions look complicated?
  Using competitive analysis.
- For permutation-invariant models, is PMLE statistically optimal in estimating permutation-invariant targets of θ?
  Universally true - when the target error is large.
- Is PMLE subject to certain limitations as well?
  Yes - when the target error is small.

Future directions:

- tightness of exponent in Gaussian location model?
- direct analysis of PMLE?
- relationships to nonparametric MLE?
  $\pi^{\text{NPMLE}} = \arg\max_{\pi} \sum_{i=1}^{n} \log \int p_{\theta}(x_i)\pi(d\theta)$

References

- Y. Han, J. Jiao, and T. Weissman. "Local moment matching: a unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance." Conference on Learning Theory (COLT), 2018.
- Y. Han and K. Shiragur. "On the competitive analysis and high-accuracy optimality of profile maximum likelihood." Symposium on Discrete Algorithms (SODA), 2021.
- Y. Han. "On the high-accuracy limitation of adaptive property estimation." International Conference on Artificial Intelligence and Statistics (AISTATS), 2021.

Thank you!