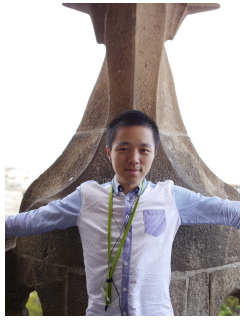
High-accuracy Optimality and Limitation of the Profile Maximum Likelihood



Yanjun Han (Stanford EE)

Kirankumar Shiragur (Stanford MS&E)

Adknowledgements: Jayadev Acharya, Moses Charikar, Aaron Sidford, Tsachy Weissman

IT Forum, September 25, 2020

Profile

Profile: given samples $X_1, X_2, \cdots, X_n$ taking value in a finite domain of size k, its profile ϕ is a vector $(\phi_1, \cdots, \phi_n)$ with

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

Profile

Profile: given samples $X_1, X_2, \cdots, X_n$ taking value in a finite domain of size k, its profile ϕ is 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 $\phi = (2, 0, 1, 0, 0)$

Profile

Profile: given samples $X_1, X_2, \cdots, X_n$ taking value in a finite domain of size k, its profile ϕ is 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 $h = (3, 1, 1)$

Profile maximum likelihood (PML)

Profile maximum likelihood (PML): given samples with profile ϕ, the PML is defined as (Orlitsky et al.'04)

$$p^{\mathsf{PML}}(\phi) = \arg \max_{p \in \mathcal{M}_k} \mathbb{P}(p, \phi)$$

- $\mathcal{M}_k$: set of all discrete distributions with support size k

Profile maximum likelihood (PML)

Profile maximum likelihood (PML): given samples with profile ϕ, the PML is defined as (Orlitsky et al.'04)

$$p^{\mathrm{PML}}(\phi) = \arg \max_{p \in \mathcal{M}_k} \mathbb{P}(p, \phi)$$

- $\mathcal{M}_k$: set of all discrete distributions with support size k
- in the previous example with $\phi = (2, 0, 1, 0, 0)$, $p^{\mathrm{PML}}(\phi)$ solves

$$\max_{(p_1, p_2, p_3)} p_1^3 p_2 p_3 + p_1 p_2^3 p_3 + p_1 p_2 p_3^3.$$

PML examples

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

PML examples

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

- Empirical estimate: $p^{\mathrm{EMP}} = (2/3, 1/3)$

PML examples

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

- Empirical estimate: $p^{\mathrm{EMP}} = (2/3, 1/3)$
- PML estimate: $p^{\mathrm{PML}} = (1/2, 1/2)$

PML examples

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

- Empirical estimate: $p^{\mathrm{EMP}} = (2/3, 1/3)$
- PML estimate: $p^{\mathrm{PML}} = (1/2, 1/2)$

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

PML examples

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

- Empirical estimate: $p^{\mathrm{EMP}} = (2/3, 1/3)$
- PML estimate: $p^{\mathrm{PML}} = (1/2, 1/2)$

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

- Empirical estimate: $p^{\mathrm{EMP}} = (1/2, 1/4, 1/4, 0, 0)$

PML examples

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

- Empirical estimate: $p^{\mathrm{EMP}} = (2/3, 1/3)$
- PML estimate: $p^{\mathrm{PML}} = (1/2, 1/2)$

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

- Empirical estimate: $p^{\mathrm{EMP}} = (1/2, 1/4, 1/4, 0, 0)$
- PML estimate: $p^{\mathrm{PML}} = (1/5, 1/5, 1/5, 1/5, 1/5)$

Computational burden of PML

Very hard to compute or even approximate PML in general

Computational burden of PML

Very hard to compute or even approximate PML in general

- highly non-convex optimization involving exponentially many terms

Computational burden of PML

Very hard to compute or even approximate PML in general

- highly non-convex optimization involving exponentially many terms
- several heuristic algorithms

Computational burden of PML

Very hard to compute or even approximate PML in general

- highly non-convex optimization involving exponentially many terms
- several heuristic algorithms
- provably polynomial-time approximate algorithms not available until very recently (Charikar, Shiragur, and Sidford'19) and (Anari, Charikar, Shiragur, and Sidford'20)

Background: 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

- n: sample size
- k: support size

Background: 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

- n: sample size
- k: support size

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

Background: 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

- n: sample size
- k: support size

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...

Background: 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

- n: sample size
- k: support size

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 estimator (MLE): $\widehat{F} = F(\widehat{p}_n)$, with $\widehat{p}_n$ the empirical distribution

Ad-hoc estimation

Plug-in estimator (MLE): $\widehat{F} = F(\widehat{p}_n)$, with $\widehat{p}_n$ the empirical distribution

Effective sample size enlargement

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

Ad-hoc estimation

Plug-in estimator (MLE): $\widehat{F} = F(\widehat{p}_n)$, with $\widehat{p}_n$ the empirical distribution

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)
- ...

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

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

$$X_1, \cdots, X_n$$

Adaptive estimation

Target

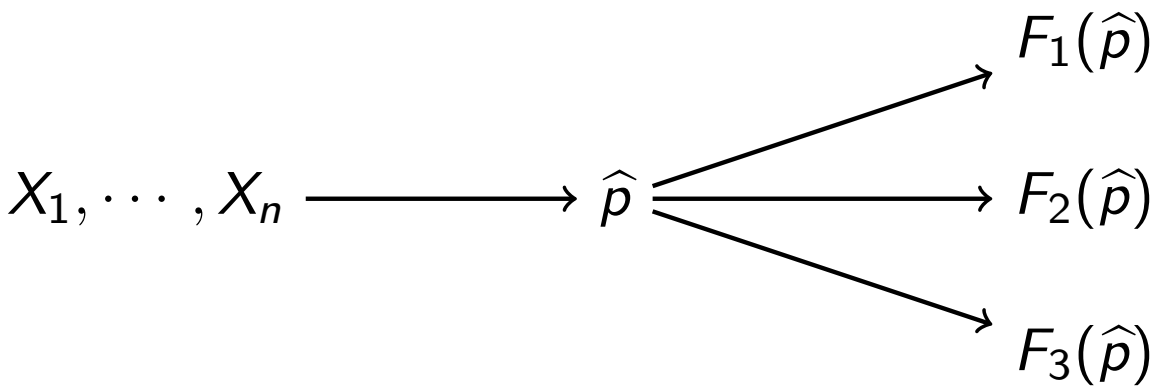
Find a single distribution estimator $\widehat{p}$ such that the plugging $\widehat{p}$ into the functional is universally optimal for "many" functionals

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

Adaptive estimation

Target

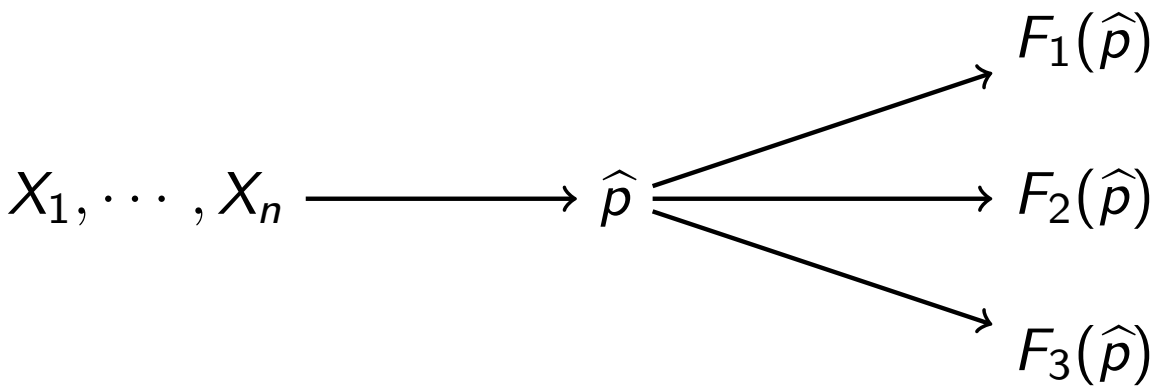
Find a single distribution estimator $\widehat{p}$ such that the plugging $\widehat{p}$ into the functional is universally optimal for "many" functionals



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

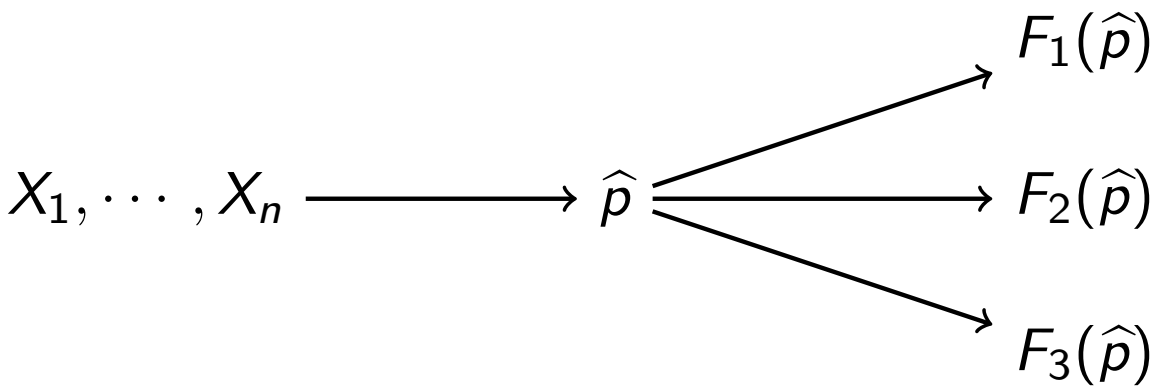


Too good to be true?

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



Too good to be true? No!

First approach: 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 $\varepsilon \gg n^{-1/3}$.

First approach: 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

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

Second approach: PML

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

Second approach: PML

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

A recent breakthrough:

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

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

Second approach: PML

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

A recent breakthrough:

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

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

Corollary: as 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 $\varepsilon \gg n^{-1/4}$.

Summary of approaches

	ad-hoc	LMM	PML
optimality	full: $\varepsilon \gg n^{-1/2}$	if $\varepsilon \gg n^{-1/3}$	if $\varepsilon \gg n^{-1/4}$
complexity	almost linear	polynomial	polynomial*
functional independent	✗	✓	✓
asymmetric functional	✓	✗	✗
free parameter tuning	✗	✗	✓

Summary of approaches

	ad-hoc	LMM	PML
optimality	full: $\varepsilon \gg n^{-1/2}$	if $\varepsilon \gg n^{-1/3}$	if $\varepsilon \gg n^{-1/4}$
complexity	almost linear	polynomial	polynomial*
functional independent	✗	✓	✓
asymmetric functional	✓	✗	✗
free parameter tuning	✗	✗	✓

Open question: is the requirement $\varepsilon \gg n^{-1/4}$ for PML an artifact of the analysis, or a fundamental limitation?

Main result of this talk

	ad-hoc	LMM	PML
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	✗	✗	✓

Main result of this talk

	ad-hoc	LMM	PML
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 analysis of PML: high-accuracy optimality and limitation

Main theorems

Informal Theorem 1

The PML plug-in approach is competitive against all estimators, with an amplification factor at most $\exp(n^{1/3+o(1)})$ on the error probability

Main theorems

Informal Theorem 1

The PML plug-in approach is competitive against all estimators, with an amplification factor at most $\exp(n^{1/3+o(1)})$ on the error probability

Implication: optimality of PML when $\varepsilon \gg n^{-1/3}$

Main theorems

Informal Theorem 1

The PML plug-in approach is competitive against all estimators, with an amplification factor at most $\exp(n^{1/3+o(1)})$ on the error probability

Implication: optimality of PML when $\varepsilon \gg n^{-1/3}$

Informal Theorem 2

When $\varepsilon \ll n^{-1/3}$, the PML plug-in approach (as well as general adaptive approaches) fails to achieve the optimal sample complexity for some 1-Lipschitz functional

Main theorems

Informal Theorem 1

The PML plug-in approach is competitive against all estimators, with an amplification factor at most $\exp(n^{1/3+o(1)})$ on the error probability

Implication: optimality of PML when $\varepsilon \gg n^{-1/3}$

Informal Theorem 2

When $\varepsilon \ll n^{-1/3}$, the PML plug-in approach (as well as general adaptive approaches) fails to achieve the optimal sample complexity for some 1-Lipschitz functional

Implication: strict price of adaptation when $\varepsilon \ll n^{-1/3}$

Part I: High-accuracy optimality of PML

"On the Competitive Analysis and High Accuracy Optimality of Profile Maximum Likelihood"

arXiv: 2004.03166

Review: idea of [ADOS'17]

Notations:

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

Review: idea of [ADOS'17]

Notations:

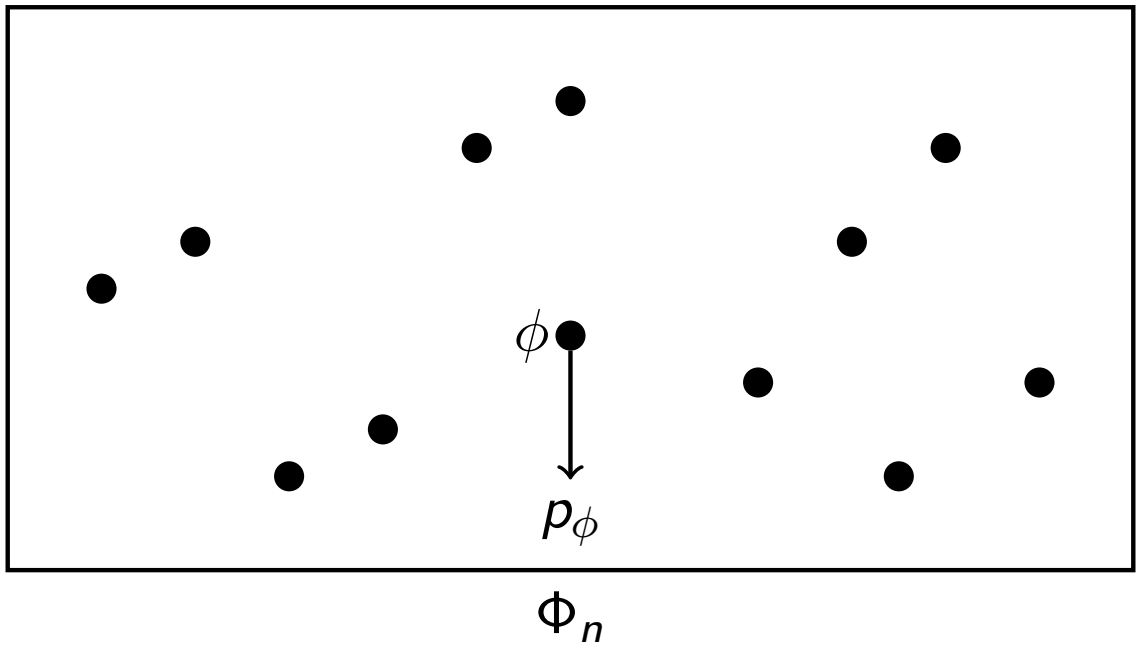
- Φ_n: the set of all possible profiles with sample size n
- ϕ: a particular profile in Φ_n
- p_ϕ: the PML distribution associated with ϕ
- $\mathbb{P}(p, \phi)$: probability of observing ϕ under the true distribution p

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

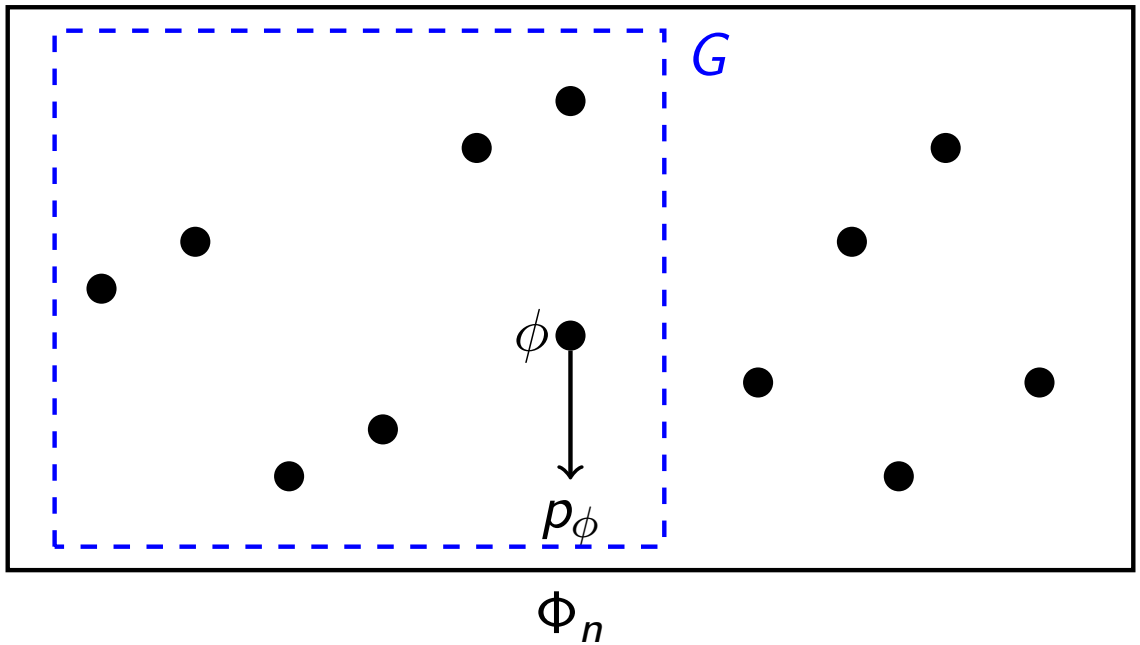
$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p_\phi) - F(p)| > 2\varepsilon)$$

given an estimator $\widehat{F}(\phi)$ with $\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| > \varepsilon) \leq \delta$.

Analysis in [ADOS'17]



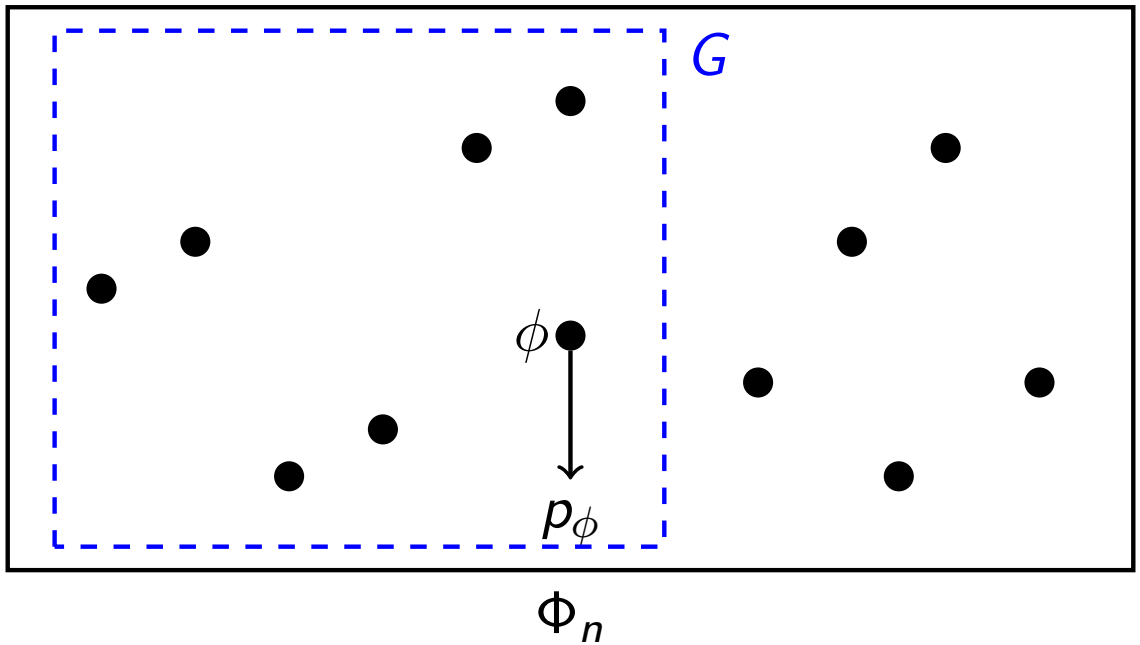
Analysis in [ADOS'17]



Good profile:

$$G = \{\phi \in \Phi_n : |\widehat{F}(\phi) - F(p)| \leq \varepsilon\}$$

Analysis in [ADOS'17]

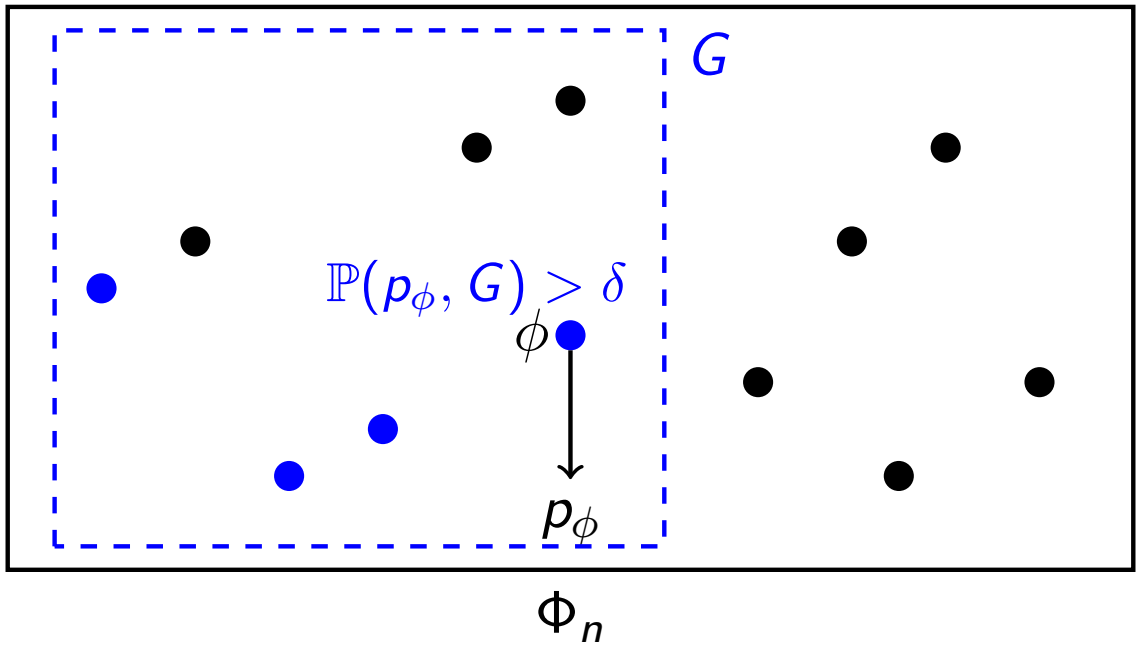


Good profile:

$$G = \{\phi \in \Phi_n : |\widehat{F}(\phi) - F(p)| \leq \varepsilon\}$$

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

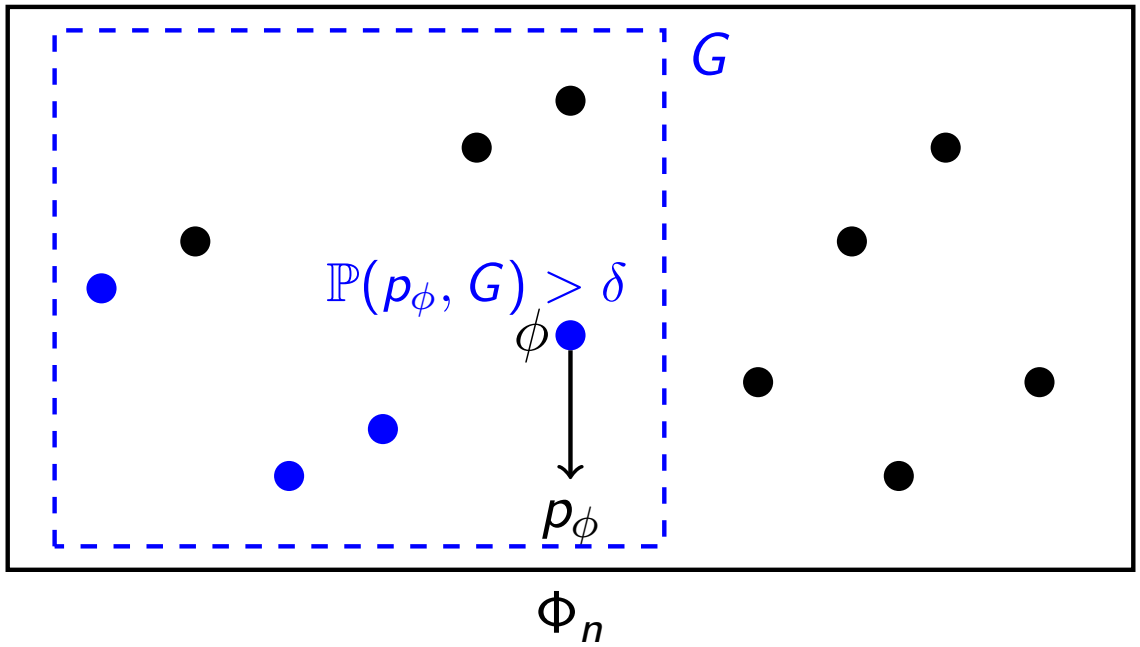
Analysis in [ADOS'17]



Lemma

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

Analysis in [ADOS'17]

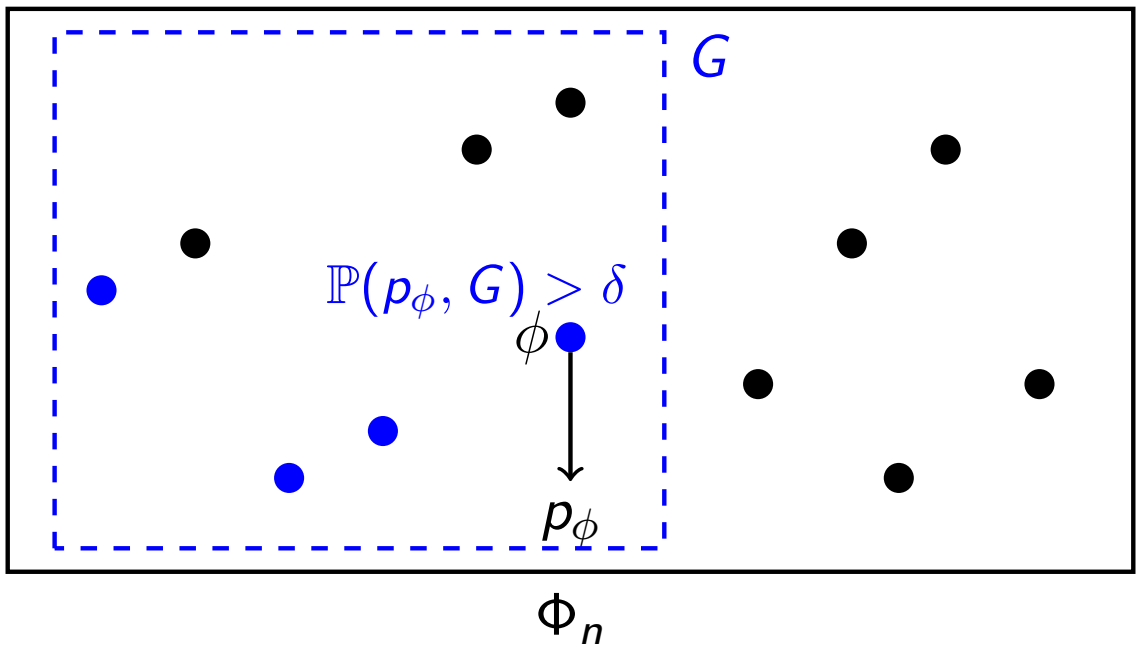


Lemma

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

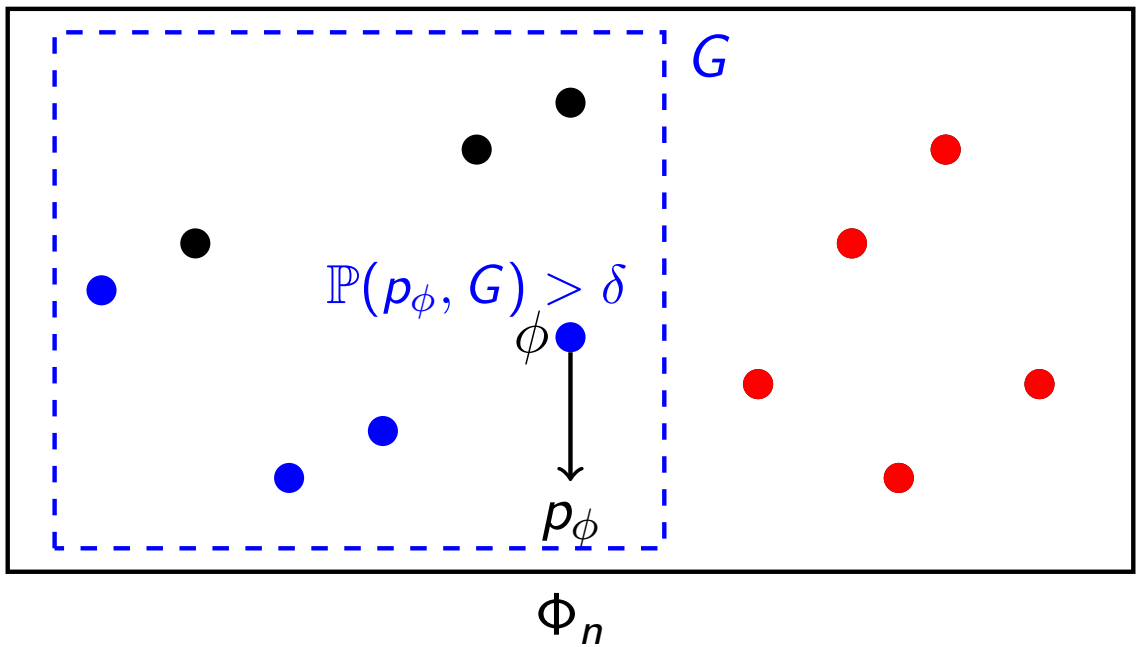
Proof: $\mathbb{P}(p_\phi, G) > \delta \Longrightarrow |\widehat{F}(\phi') - F(p_\phi)| \leq \varepsilon$ for some $\phi' \in G$. Also, definition of $G \Longrightarrow |\widehat{F}(\phi') - F(p)| \leq \varepsilon$. □

Analysis in [ADOS'17]



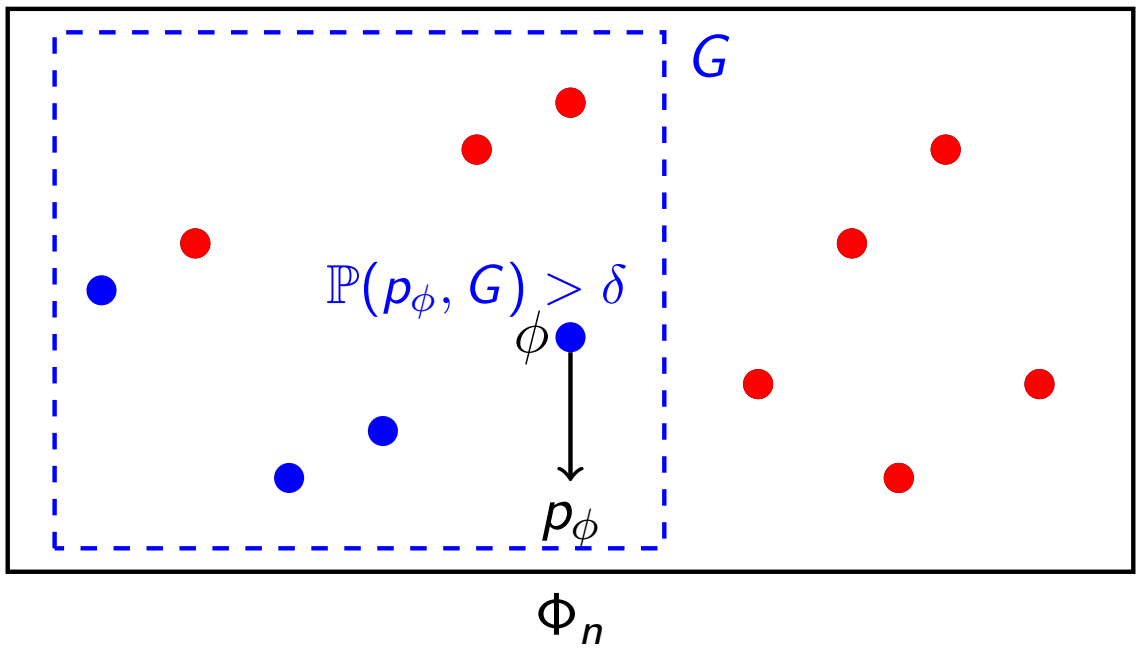
$\mathbb{P}_p(|F(p_\phi) - F(p)| > 2\varepsilon)$

Analysis in [ADOS'17]



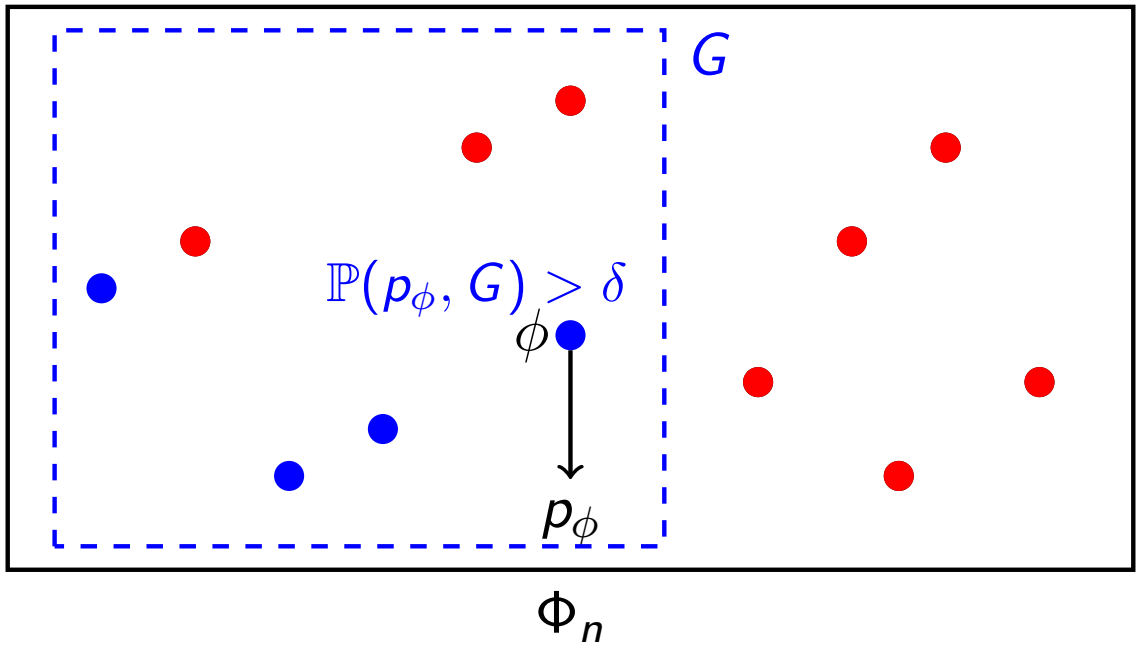
$$\mathbb{P}_p(|F(p_\phi) - F(p)| > 2\varepsilon) \leq \mathbb{P}(p, G^c)$$

Analysis in [ADOS'17]



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

Analysis in [ADOS'17]

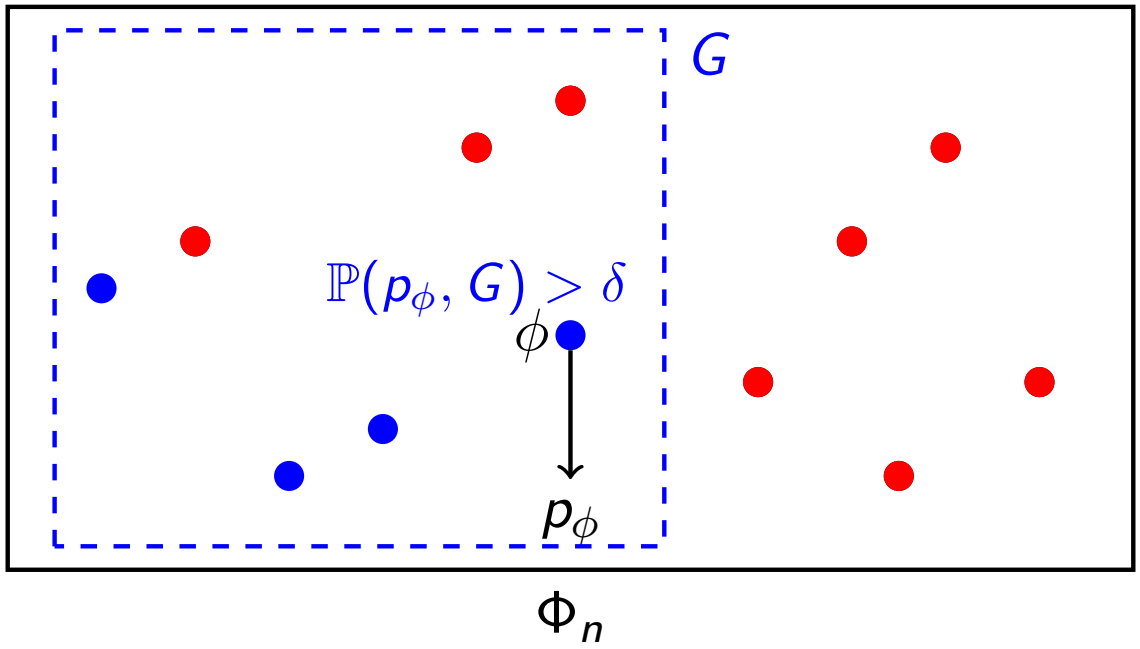


$$\mathbb{P}_p(|F(p_\phi) - F(p)| > 2\varepsilon) \leq \mathbb{P}(p, G^c) + \sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p_\phi, G) \leq \delta)$$

$$\leq \delta + \sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p, \phi) \leq \delta)$$

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

Analysis in [ADOS'17]



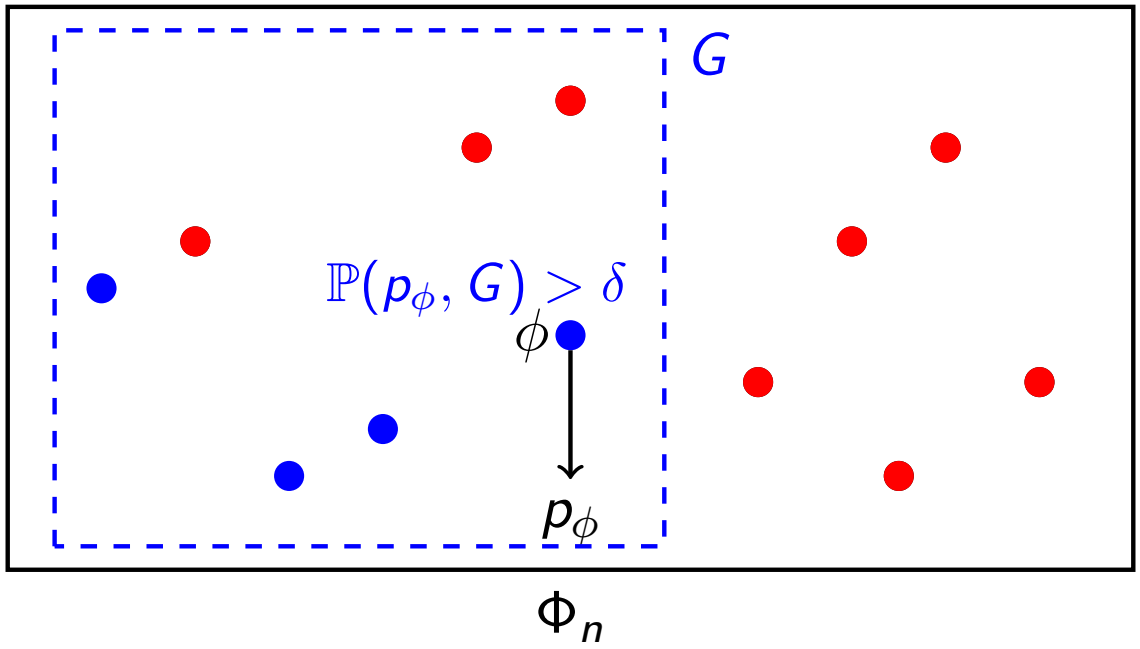
$$\mathbb{P}_p(|F(p_\phi) - F(p)| > 2\varepsilon) \leq \mathbb{P}(p, G^c) + \sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p_\phi, G) \leq \delta)$$

$$\leq \delta + \sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p, \phi) \leq \delta)$$

$$\leq (1 + |\Phi_n|) \cdot \delta$$

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

Analysis in [ADOS'17]



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

$$\le \delta + \sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p, \phi) \le \delta)$$

$$\le (1 + |\Phi_n|) \cdot \delta \le \exp(3\sqrt{n}) \cdot \delta,$$

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

More related work

Propose modifications of PML such that $|\Phi_n|$ is smaller: pseudo/truncated PML (Charikar, Shiragur, and Sidford'19, Hao and Orlitsky'19)

- not the PML anymore
- how to modify depends on the target functional

More related work

Propose modifications of PML such that $|\Phi_n|$ is smaller: pseudo/truncated PML (Charikar, Shiragur, and Sidford'19, Hao and Orlitsky'19)

- not the PML anymore
- how to modify depends on the target functional

Find distribution-dependent bound of the effective cardinality of $|\Phi_n|$: profile entropy (Hao and Orlitsky'20)

- worst-case bound still $\exp(\Omega(\sqrt{n}))$

More related work

Propose modifications of PML such that $|\Phi_n|$ is smaller: pseudo/truncated PML (Charikar, Shiragur, and Sidford'19, Hao and Orlitsky'19)

- not the PML anymore
- how to modify depends on the target functional

Find distribution-dependent bound of the effective cardinality of $|\Phi_n|$: profile entropy (Hao and Orlitsky'20)

- worst-case bound still $\exp(\Omega(\sqrt{n}))$

Still open whether the previous analysis could be improved in general

Our result

Theorem (Han and Shiragur'20)

If there exists an estimator $\widehat{F}(\phi)$ such that

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| > \varepsilon) \leq \delta,$$

then for any $c > 0$, the PML distribution p^{PML} satisfies

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathsf{PML}}) - F(p)| > (2 + o(1))\varepsilon) \leq \delta^{1-c} \cdot \exp\left(c' n^{1/3+c}\right),$$

for some constant c' depending only on c.

Our result

Theorem (Han and Shiragur'20)

If there exists an estimator $\widehat{F}(\phi)$ such that

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| > \varepsilon) \leq \delta,$$

then for any $c > 0$, the PML distribution p^{PML} satisfies

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathsf{PML}}) - F(p)| > (2 + o(1))\varepsilon) \leq \delta^{1-c} \cdot \exp\left(c' n^{1/3+c}\right),$$

for some constant c' depending only on c.

- improve the exponent from $O(\sqrt{n})$ to $O(n^{1/3+c})$ for any $c > 0$

Our result

Theorem (Han and Shiragur'20)

If there exists an estimator $\widehat{F}(\phi)$ such that

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| > \varepsilon) \leq \delta,$$

then for any $c > 0$, the PML distribution p^{PML} satisfies

$$\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathsf{PML}}) - F(p)| > (2 + o(1))\varepsilon) \leq \delta^{1-c} \cdot \exp\left(c' n^{1/3+c}\right),$$

for some constant c' depending only on c.

- improve the exponent from $O(\sqrt{n})$ to $O(n^{1/3+c})$ for any $c > 0$
- work for approximate PML as well

Some corollaries

Corollary 1 (functional estimation)

For many symmetric functionals (e.g. entropy, support size, distance to uniformity), the PML plug-in approach attains the optimal rate of the sample complexity within the accuracy level $\varepsilon \gg n^{-1/3}$.

Some corollaries

Corollary 1 (functional estimation)

For many symmetric functionals (e.g. entropy, support size, distance to uniformity), the PML plug-in approach attains the optimal rate of the sample complexity within the accuracy level $\varepsilon \gg n^{-1/3}$.

Accuracy level improved from $n^{-1/4}$ to $n^{-1/3}$

Some corollaries

Corollary 1 (functional estimation)

For many symmetric functionals (e.g. entropy, support size, distance to uniformity), the PML plug-in approach attains the optimal rate of the sample complexity within the accuracy level $\varepsilon \gg n^{-1/3}$.

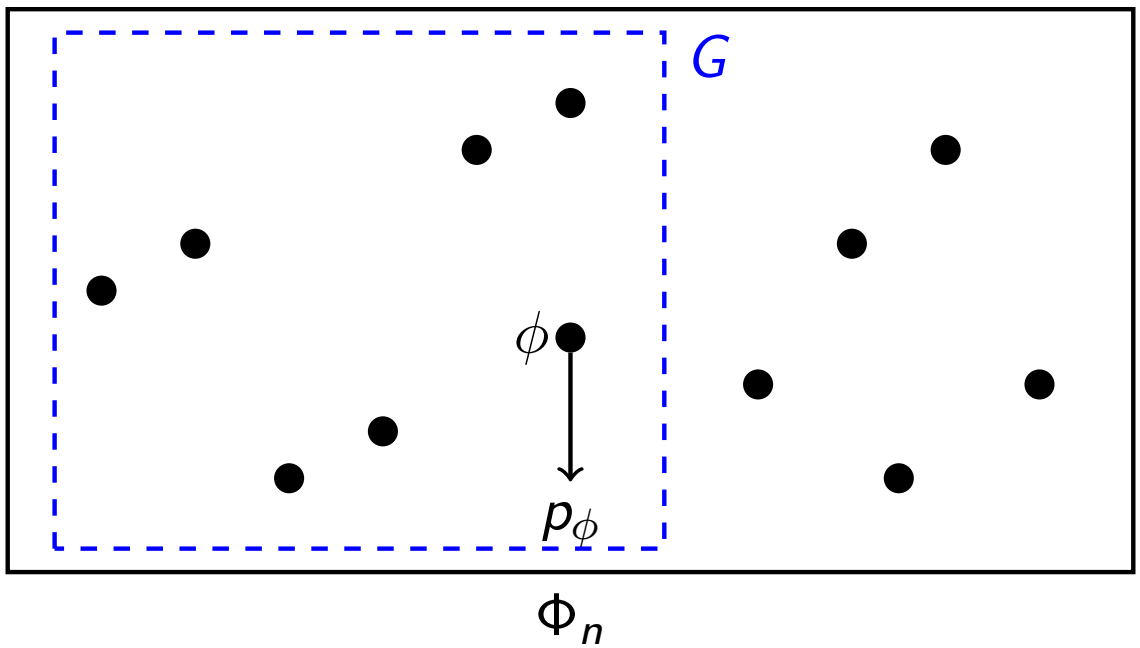
Accuracy level improved from $n^{-1/4}$ to $n^{-1/3}$

Corollary 2 (sorted distribution estimation)

The PML distribution p^{PML} itself is a minimax rate-optimal estimator of the sorted true distribution:

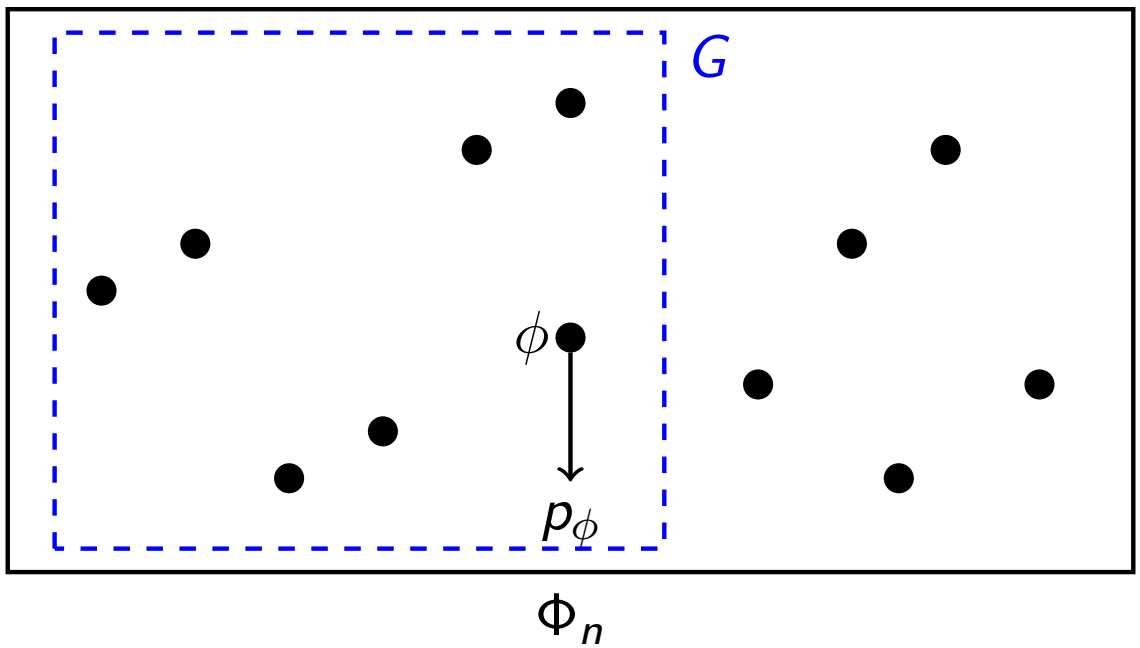
$$\sup_{p \in \mathcal{M}_k} \mathbb{E}_p \| p^{\mathrm{PML}} - p \|_{1,\mathrm{sorted}} \lesssim \sqrt{\frac{k}{n \log n}} + \widetilde{O}\left(n^{-1/3} \wedge \sqrt{\frac{k}{n}} \right).$$

Proof idea



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

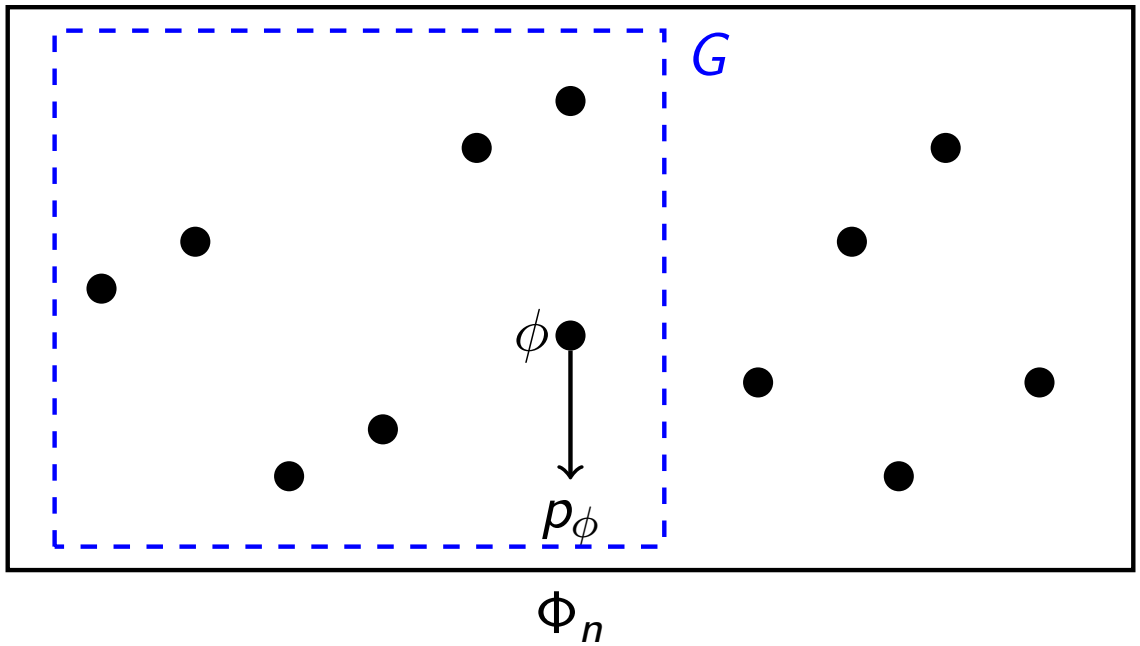
Proof idea



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

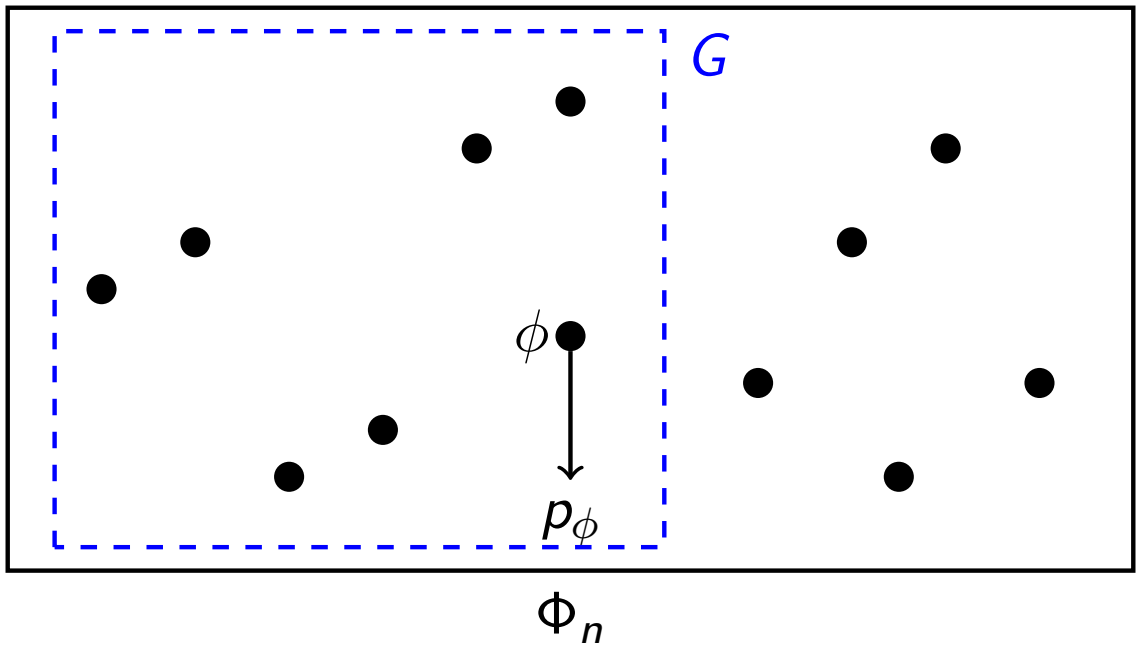
Proof idea



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

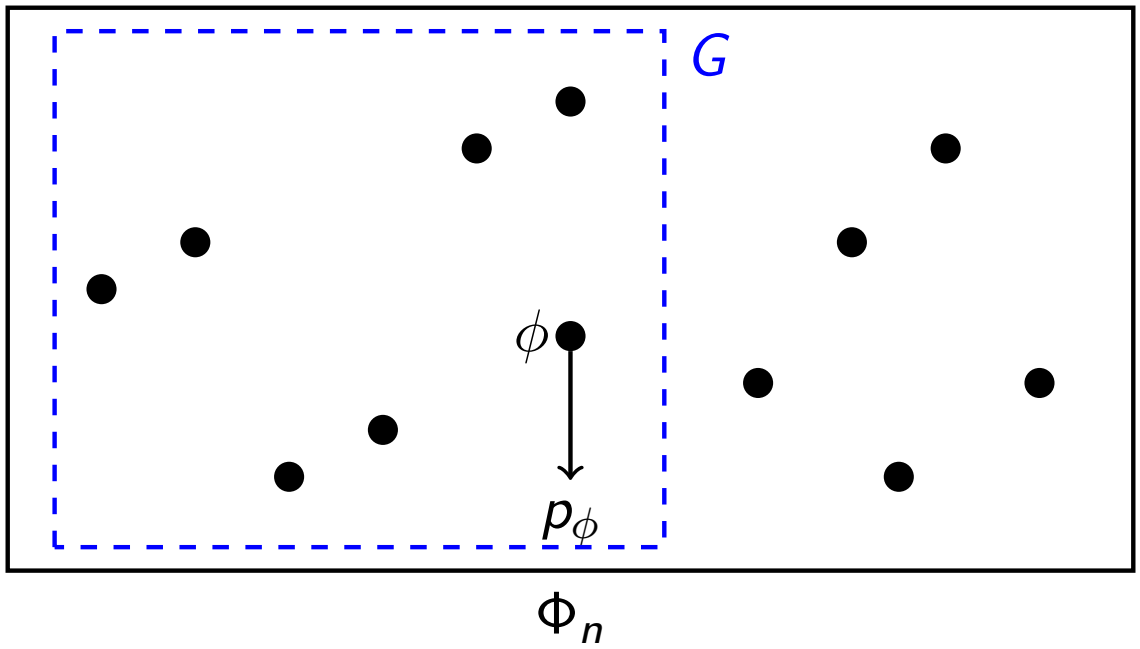
- could be tight when p_ϕ is essentially supported on ϕ
- in that case, $\mathbb{P}(p_{\phi'}, \phi) \ll \mathbb{P}(p_\phi, \phi)$

Proof idea



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

Proof idea

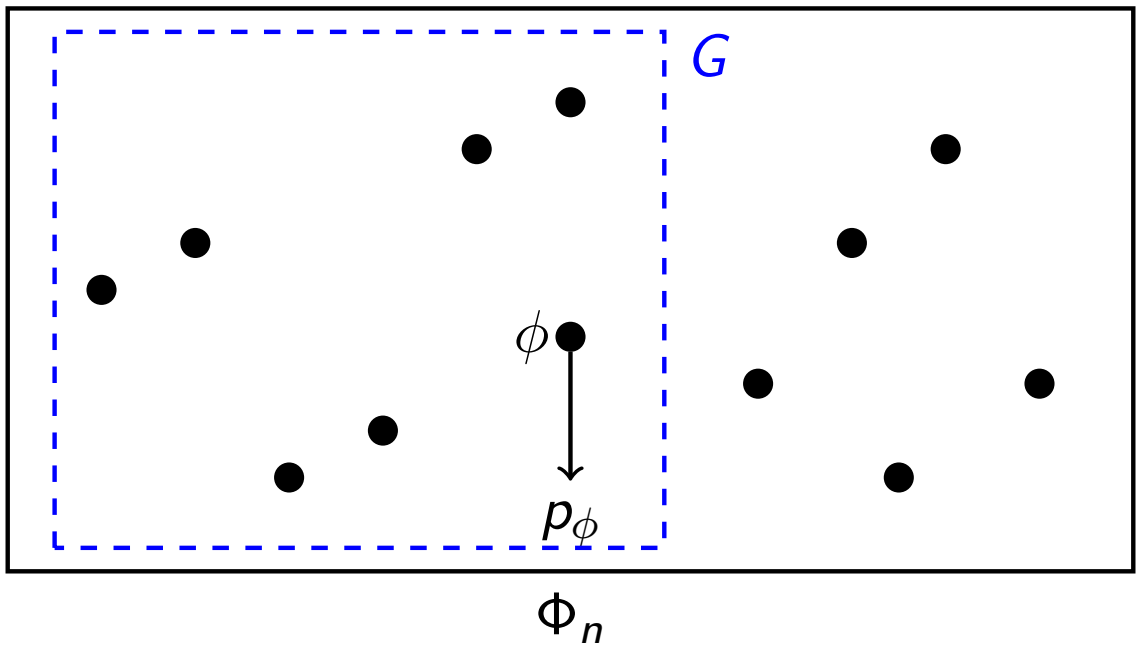


Q: What if we could have $\mathbb{P}(p_\phi, \phi) \approx \mathbb{P}(p_{\phi'}, \phi)$ for all $\phi, \phi' \in G$?
A: Then we are in a great shape, for if $\mathbb{P}(p_{\phi'}, G) < \delta$ for some $\phi' \in G$, then

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

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

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

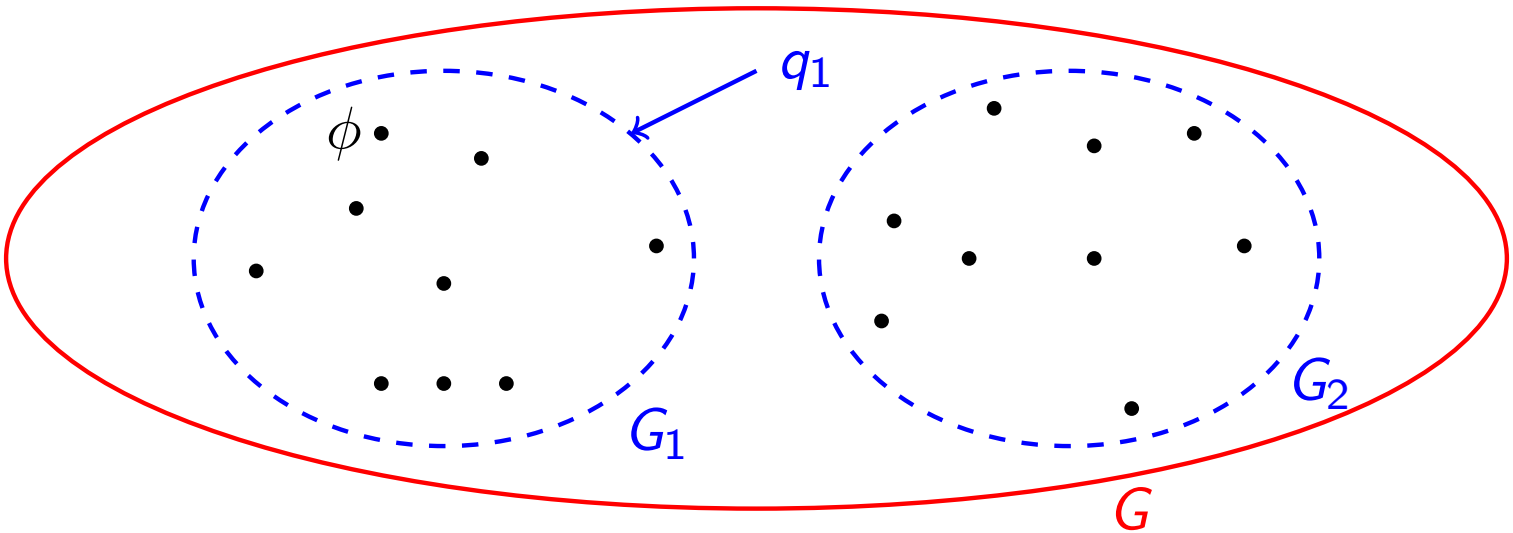
$$\mathbb{P}(p_\phi, S) \geq \mathbb{P}(p_{\phi'}, S)^{1/(1-n^{-s})} \cdot \exp\left(-cn^{1-2r+s}\right),$$
$$\mathbb{P}(p_{\phi'}, S) \geq \mathbb{P}(p_\phi, S)^{1/(1-n^{-s})} \cdot \exp\left(-cn^{1-2r+s}\right),$$

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

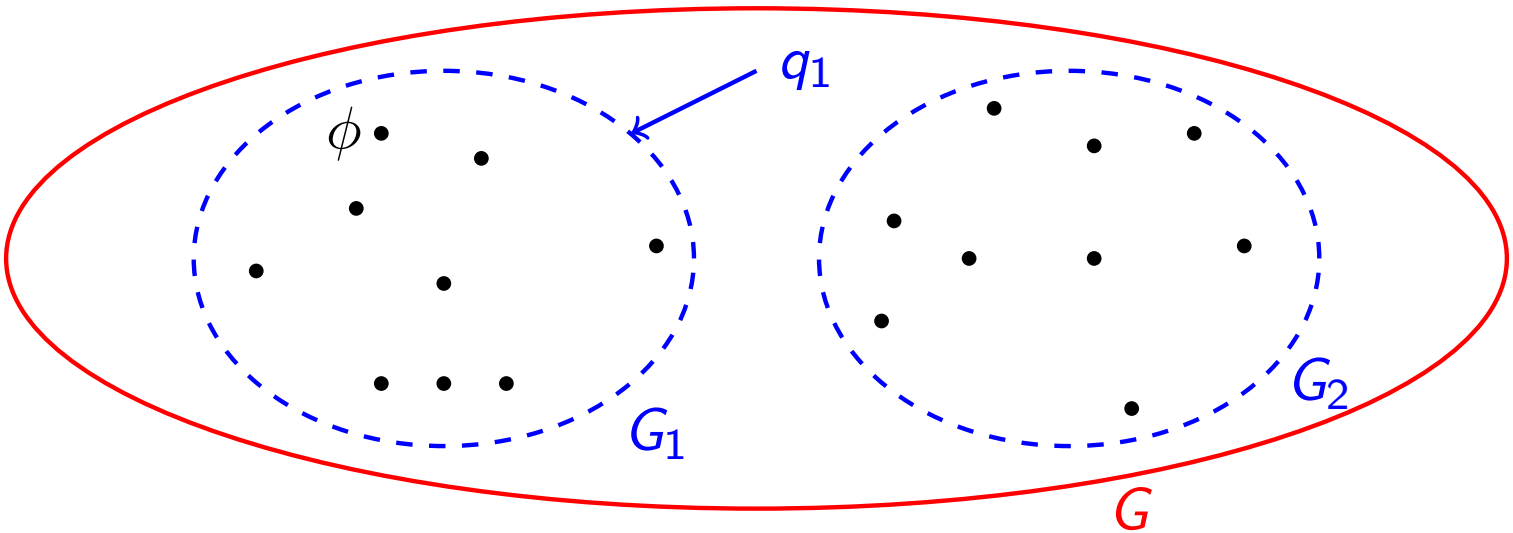
A covering property of PML distributions $\{p_\phi : \phi \in \Phi_n\}$

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

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



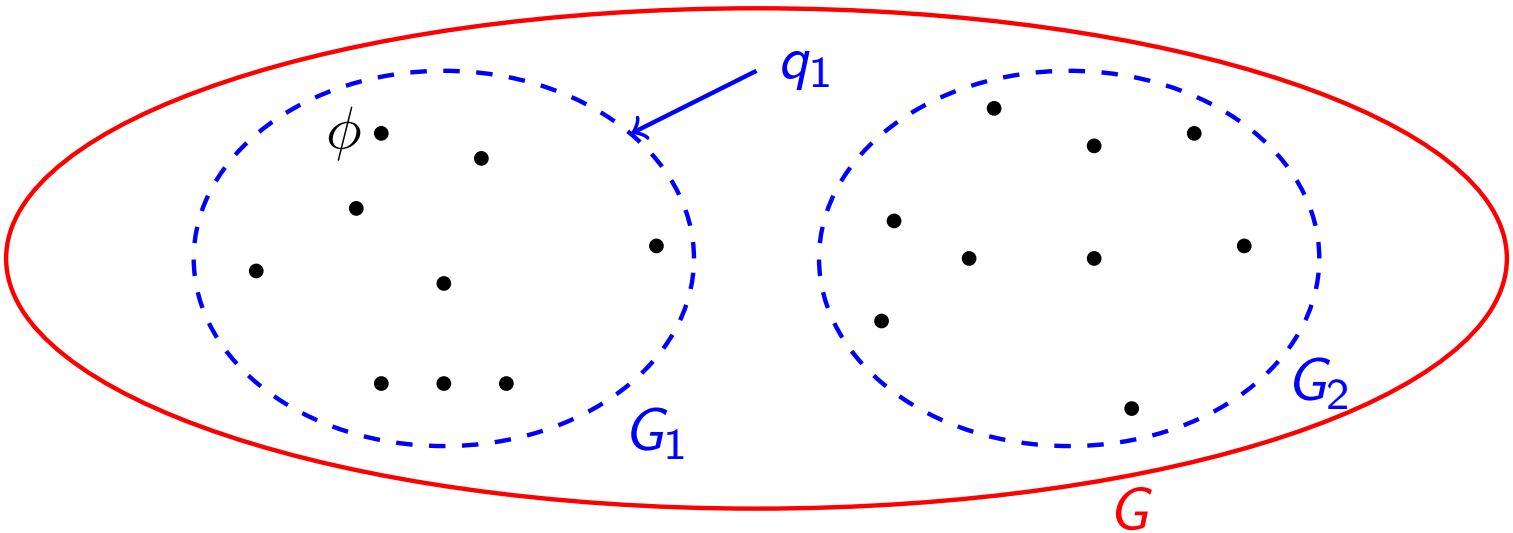
Applying the covering lemma with $r = 3/8, s = 1/8$



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 $r = 3/8, s = 1/8$

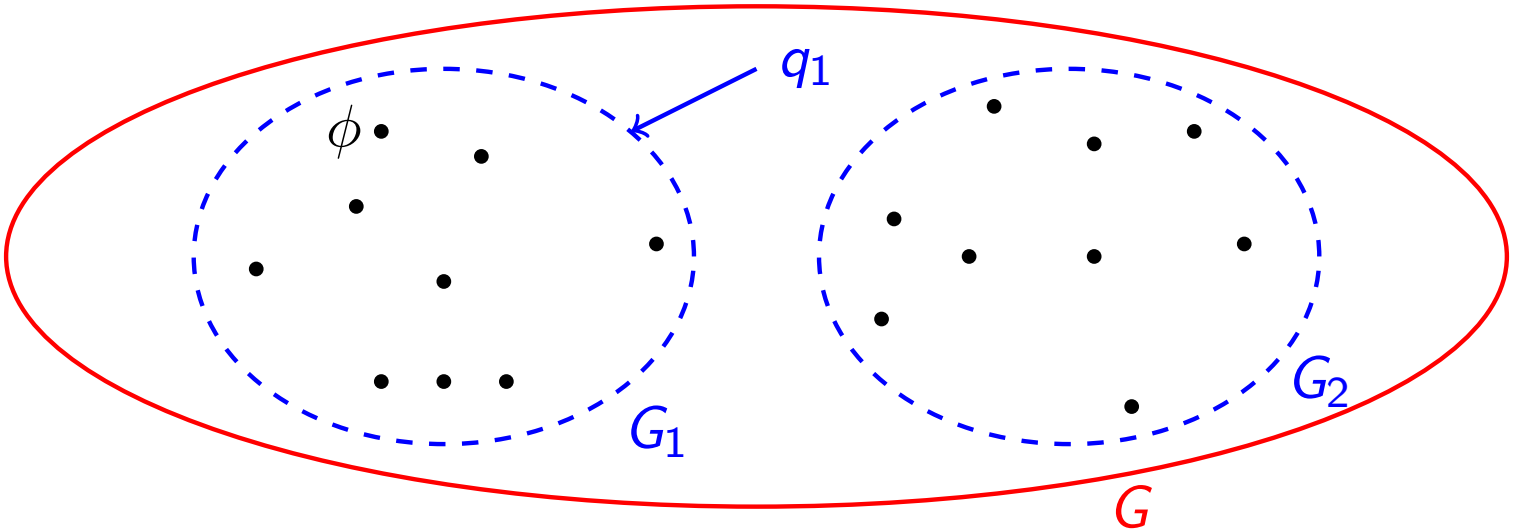


If $\mathbb{P}(p_\phi, G_1) \leq \delta$, then

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

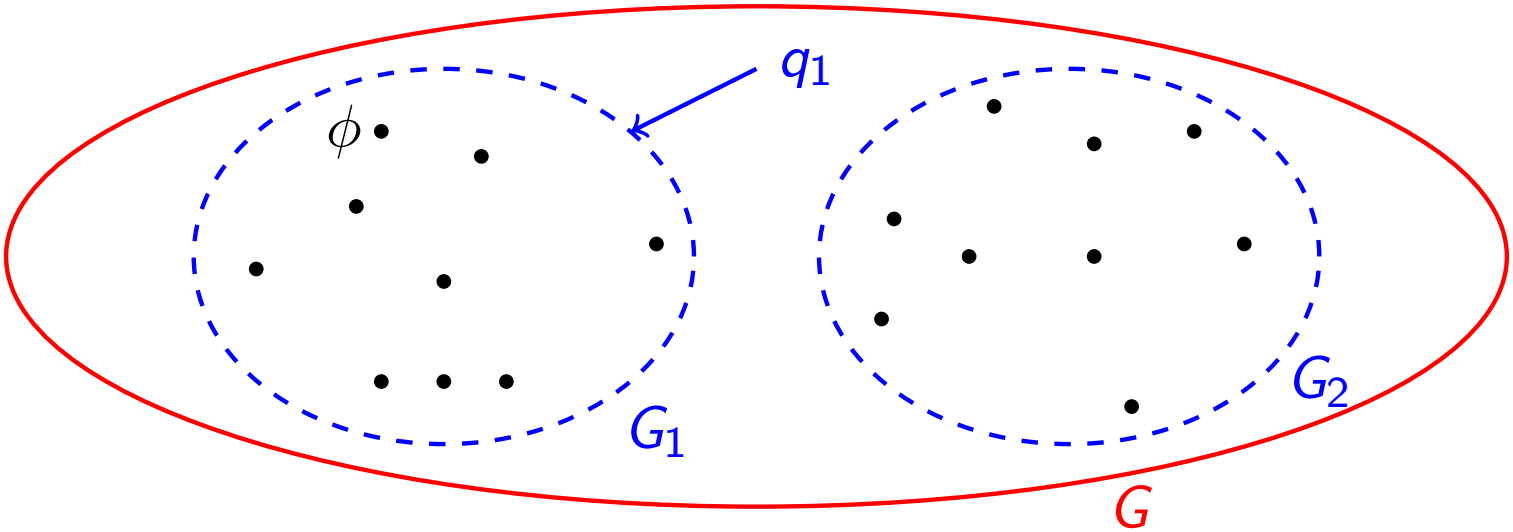
"going-down process"

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



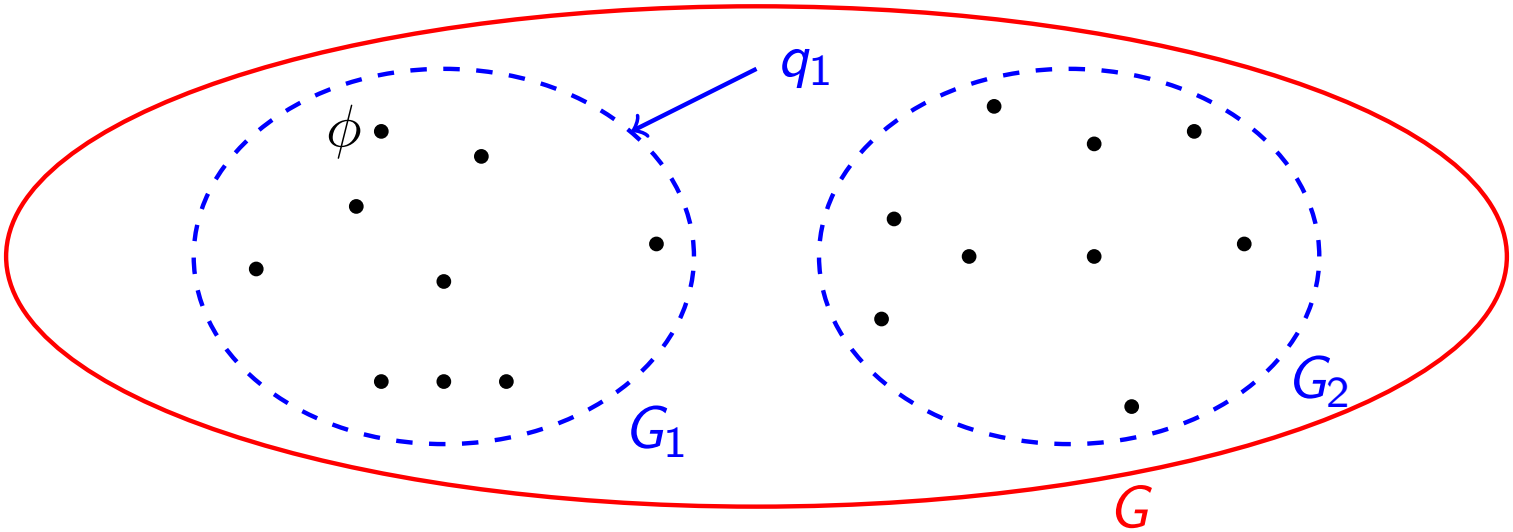
$\mathbb{P}(q_1, G_1)$

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



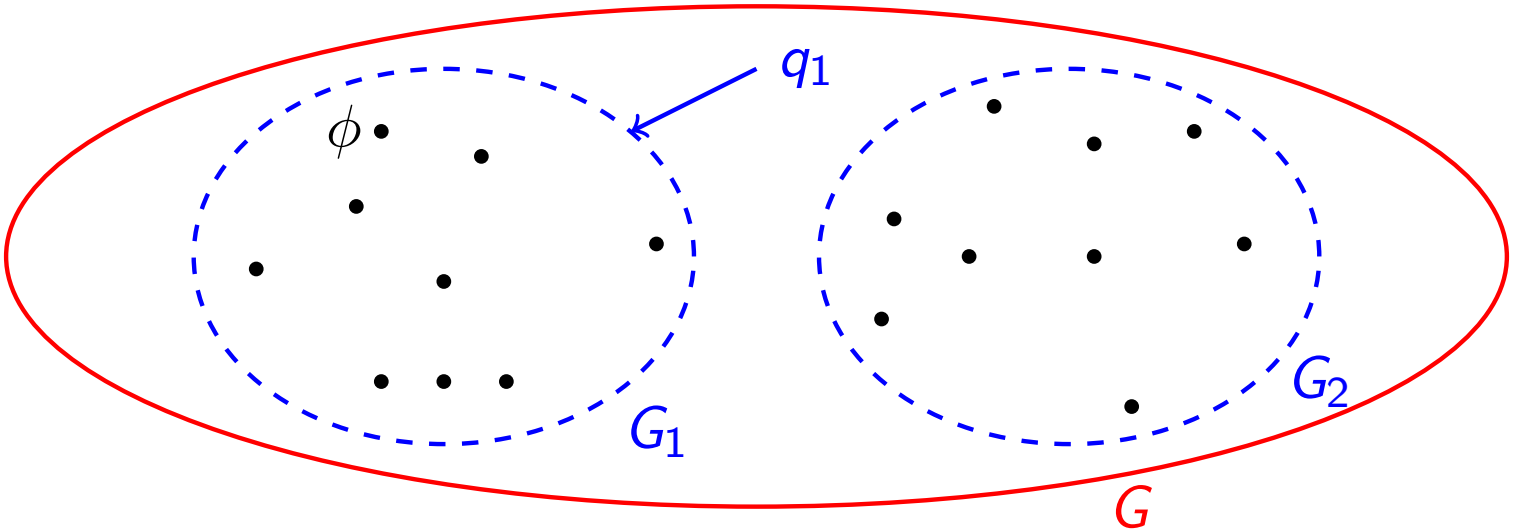
$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi)$$

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



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

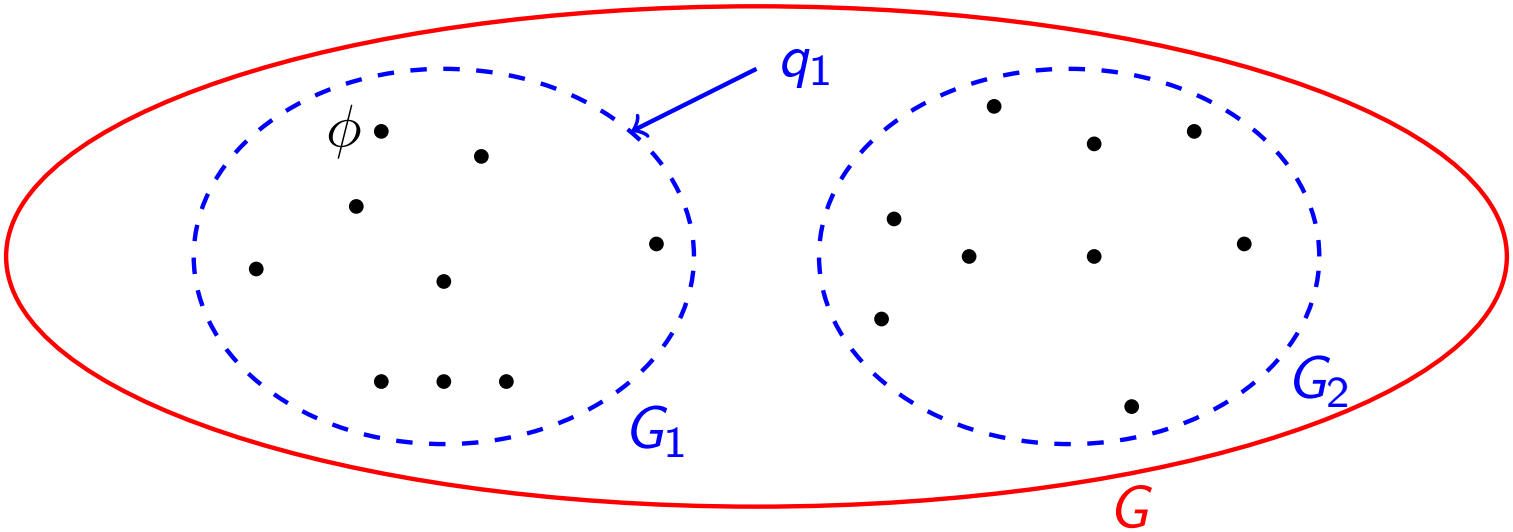
Applying the covering lemma with $r = 3/8, s = 1/8$



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

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

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

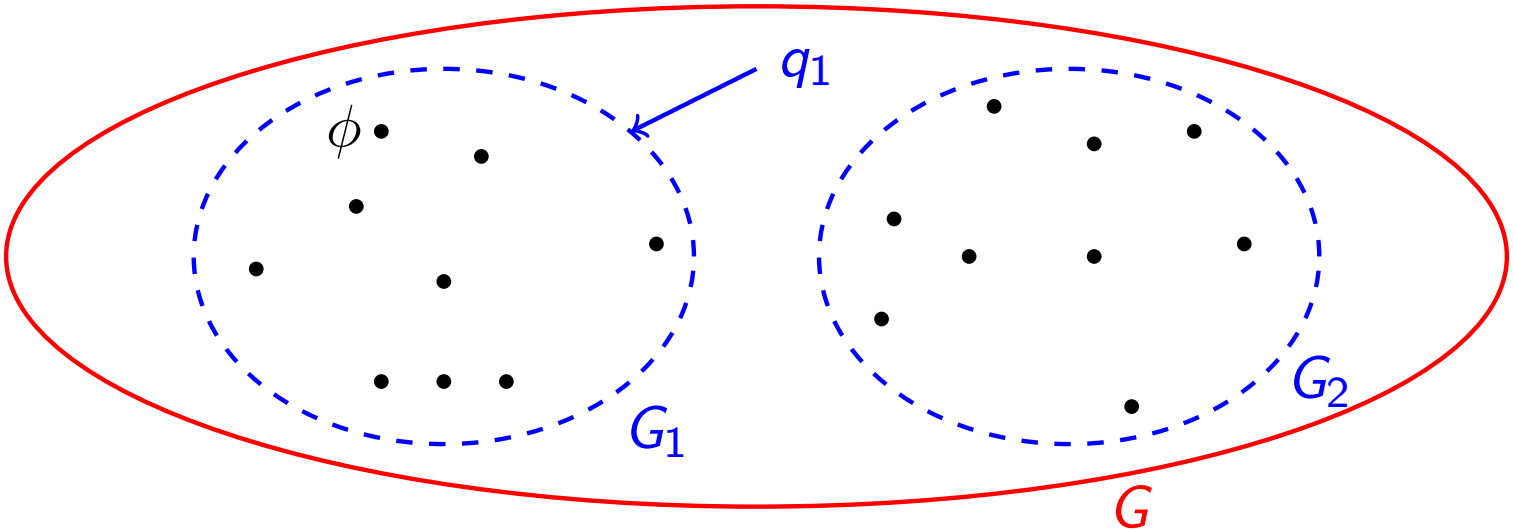


$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi) \geq \exp(-cn^{3/8}) \sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi)^{1/(1-n^{-1/8})}$$

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

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

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



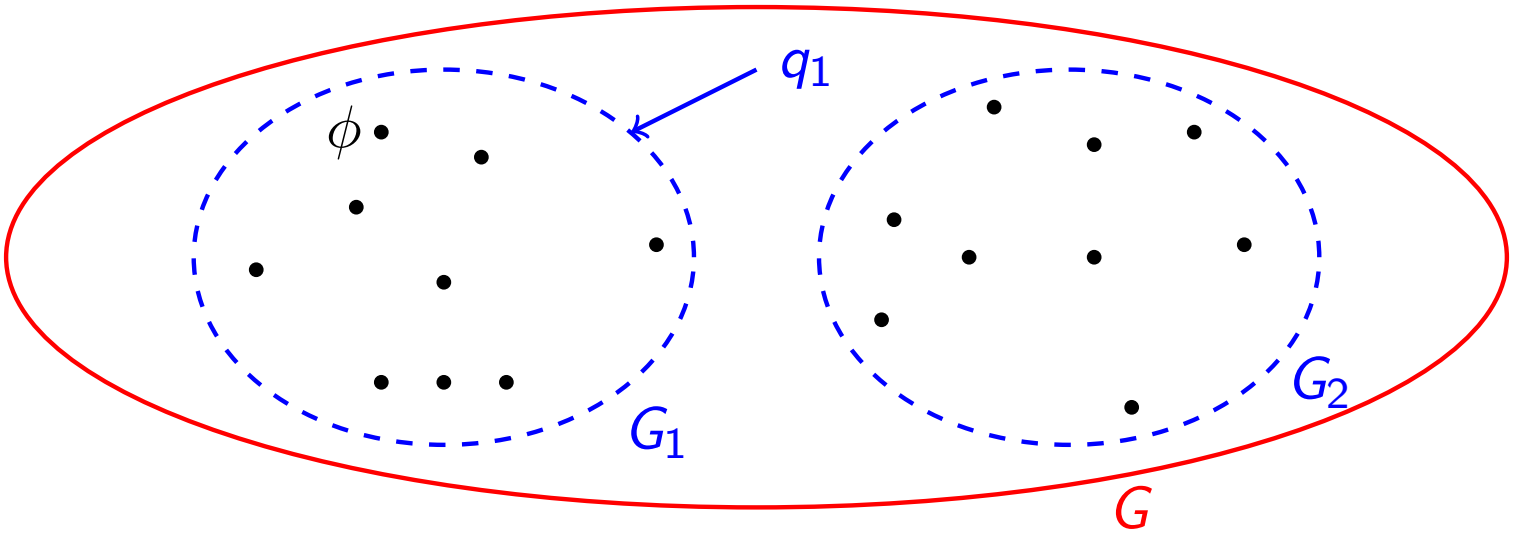
$$\mathbb{P}(q_1, G_1) = \sum_{\phi \in G_1} \mathbb{P}(q_1, \phi) \geq \exp(-cn^{3/8}) \sum_{\phi \in G_1} \mathbb{P}(p_\phi, \phi)^{1/(1-n^{-1/8})}$$

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

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

"going-up" process

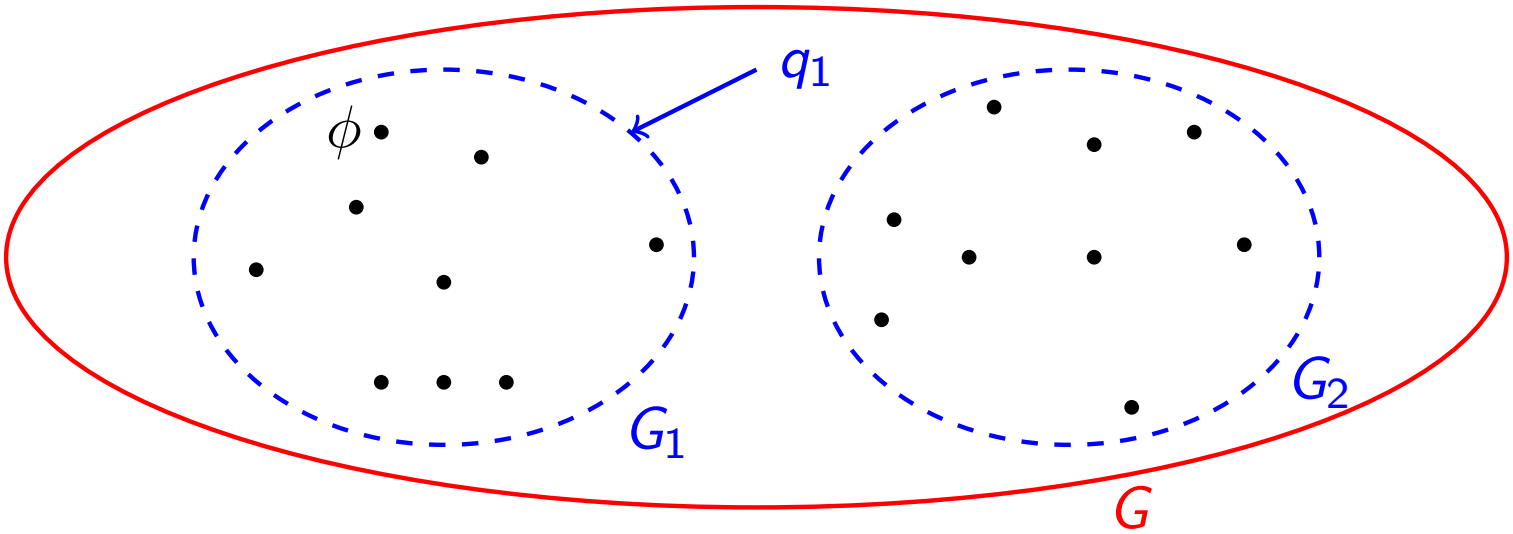
Applying the covering lemma with $r = 3/8, s = 1/8$



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

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

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



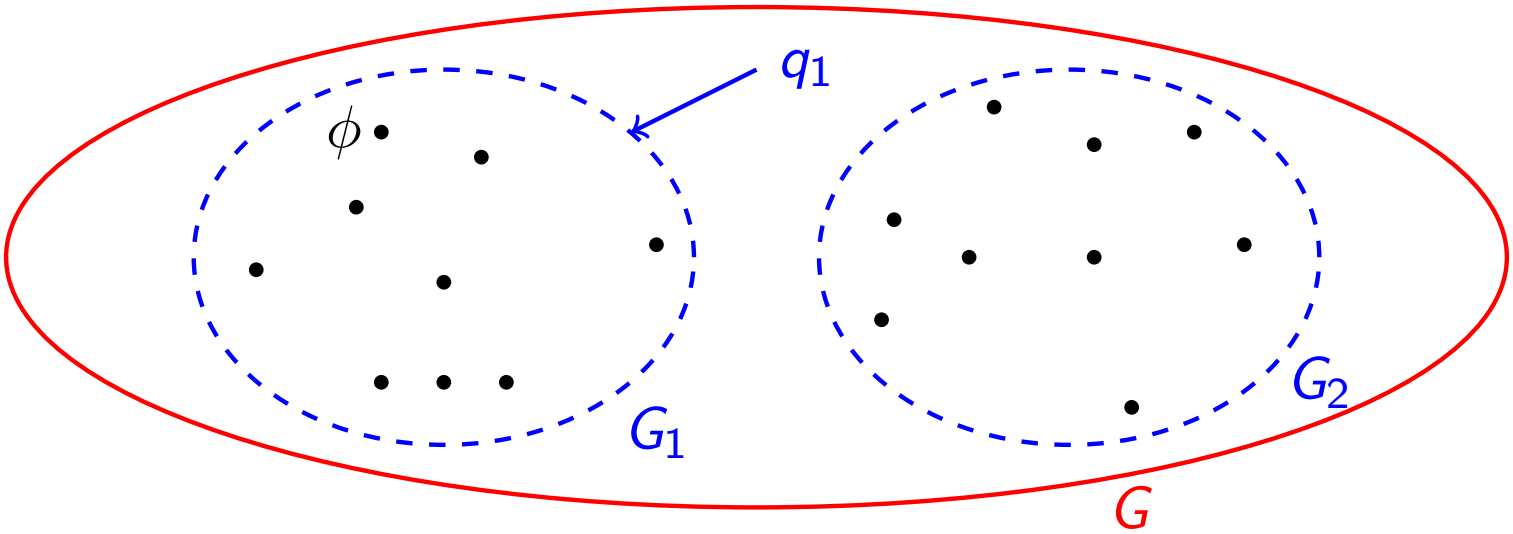
Conclusion: if $\mathbb{P}(p_\phi, G_1) \le \delta$ for some $\phi \in G_1$, then

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

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

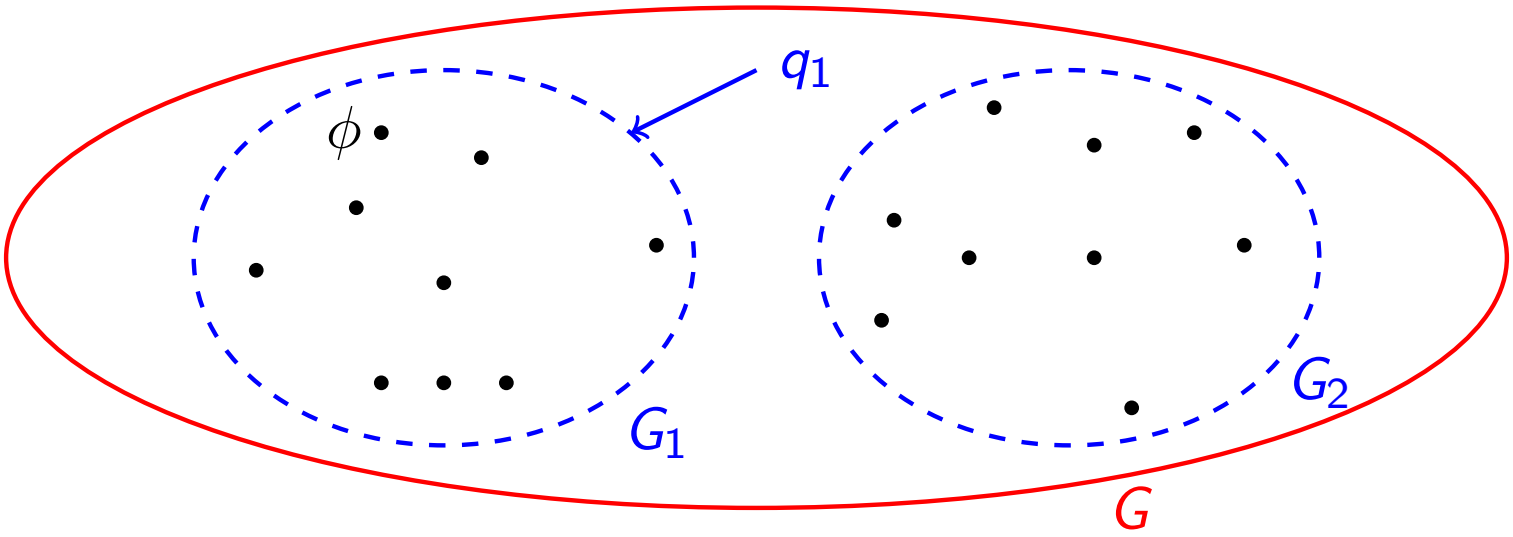
$$\mathbb{P}(p, G_1) \le \delta^{1-o(1)} \cdot \exp(cn^{3/8}).$$

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

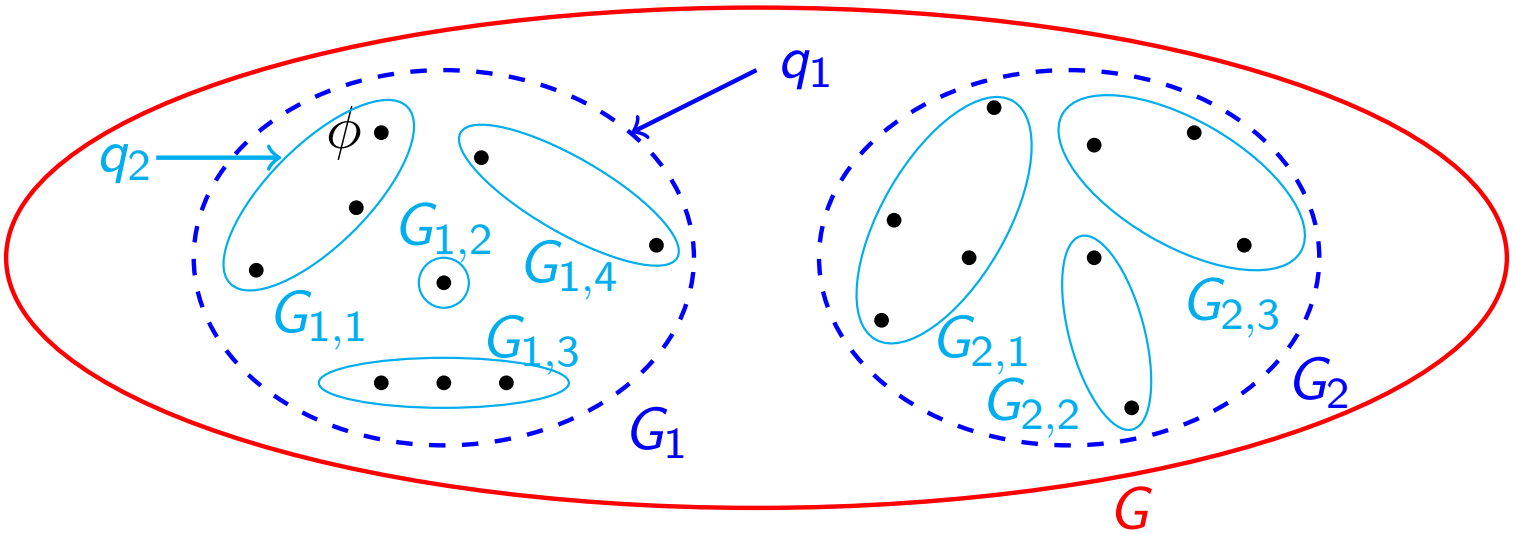
$$\sum_{\phi \in G} \mathbb{P}(p, \phi) \mathbb{1}(\mathbb{P}(p_\phi, G) \le \delta) \le \delta^{1-o(1)} \cdot \exp(cn^{3/8} \log n).$$

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

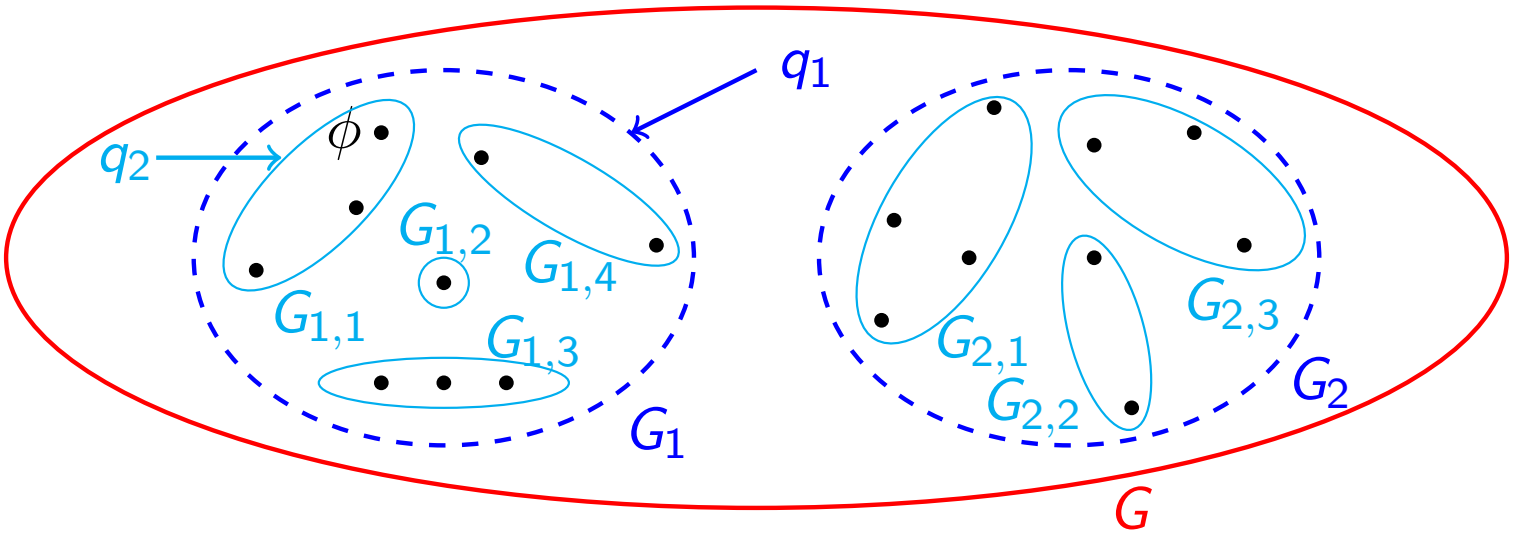
General case: chaining



General case: chaining

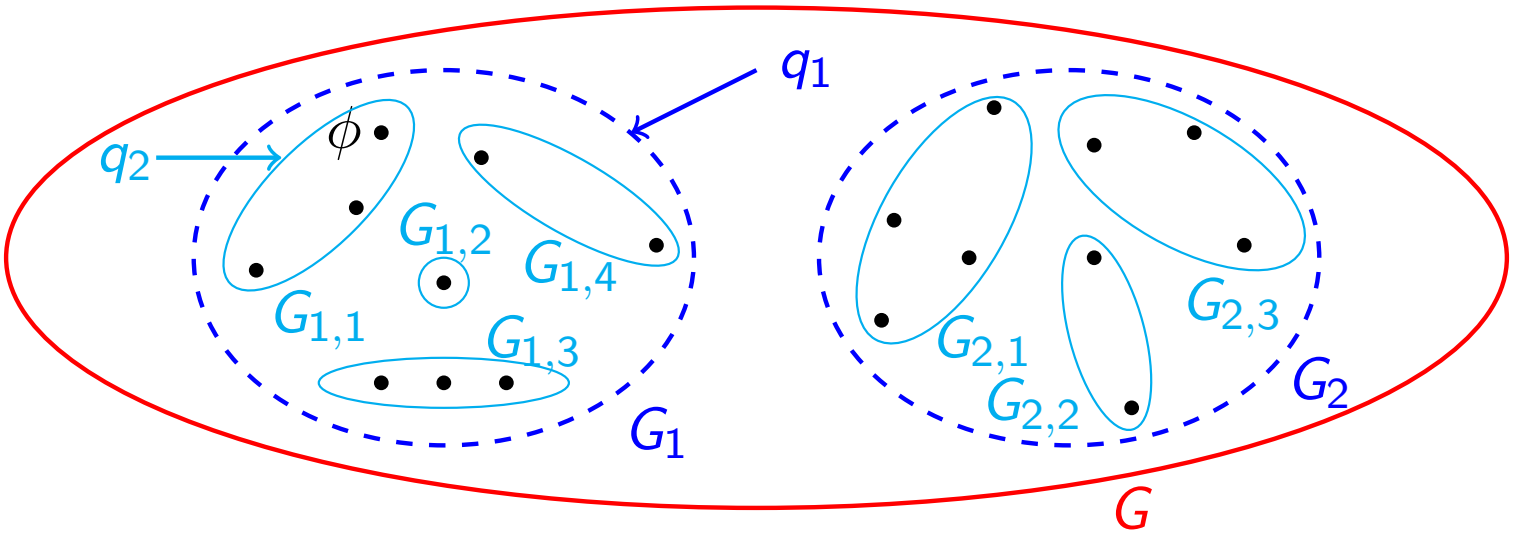


General case: chaining



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

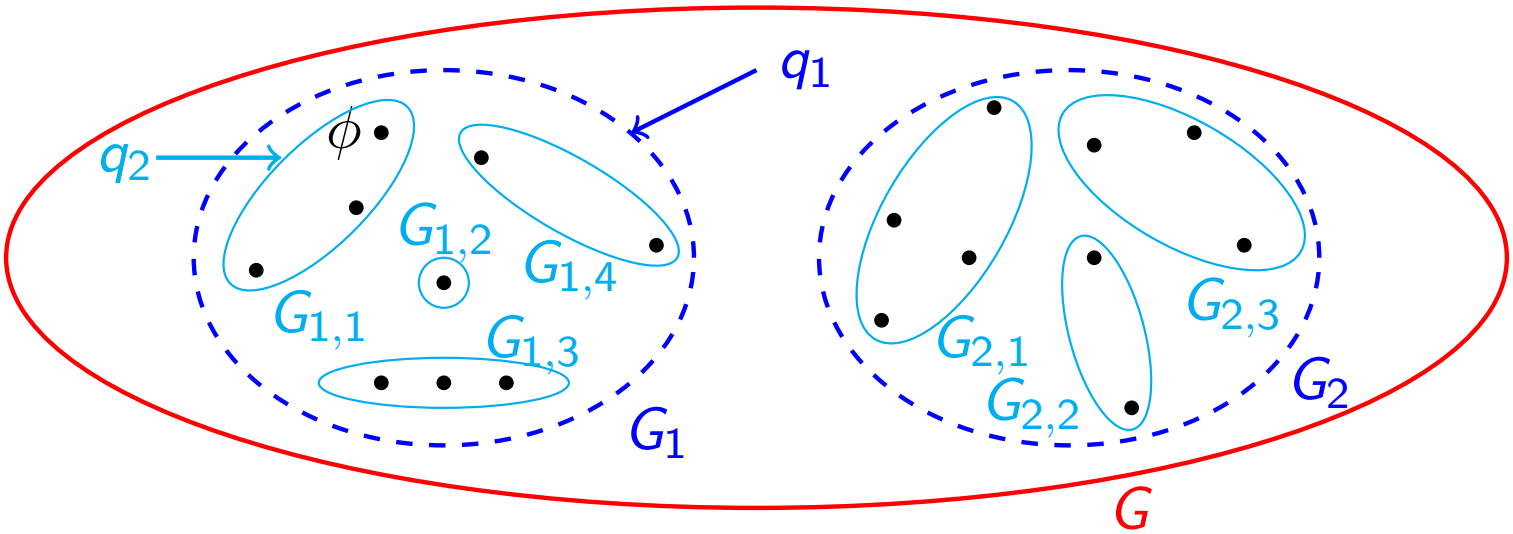
General case: chaining



- "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

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

Summary of Part I

- competitive factor improved from $\exp(3\sqrt{n})$ to $\exp(O(n^{1/3+c}))$
- accuracy threshold improved from $\varepsilon \gg n^{-1/4}$ to $\varepsilon \gg n^{-1/3}$
- covering/continuity property of PML distributions and chaining

Part II: High-accuracy limitation of PML

"On the High Accuracy Limitation of Adaptive Property Estimation"
arXiv: 2008.11964

Motivation

Question

Is the threshold $\varepsilon \gg n^{-1/3}$ or the competitive factor $\exp(O(n^{1/3+c}))$ tight for the PML?

Motivation

Question

Is the threshold $\varepsilon \gg n^{-1/3}$ or the competitive factor $\exp(O(n^{1/3+c}))$ tight for the PML?

Recall that $\varepsilon \gg n^{-1/3}$ is required for both LMM and PML...

A broader question

Is there any unavoidable price to pay for adaptation?

Adaptive minimax risk

Adaptive minimax risk:

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{M}_k} \sup_{F \in \mathcal{F}} \mathbb{E}_p |F(\widehat{p}) - F(p)|$$

Adaptive minimax risk

Adaptive minimax risk:

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{M}_k} \sup_{F \in \mathcal{F}} \mathbb{E}_p |F(\widehat{p}) - F(p)|$$

- adaptive estimation: find a single estimator $\widehat{p}$ which work for all symmetric functionals in $\mathcal{F}$

Adaptive minimax risk

Adaptive minimax risk:

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{M}_k} \sup_{F \in \mathcal{F}} \mathbb{E}_p |F(\widehat{p}) - F(p)|$$

- adaptive estimation: find a single estimator $\widehat{p}$ which work for all symmetric functionals in $\mathcal{F}$

Choice of $\mathcal{F}$: the set of all 1-Lipschitz functionals, i.e. $F(p) = \sum_{i=1}^{k} f(p_i)$ with f being 1-Lipschitz

What is known?

A smaller quantity: (Hao and Orlitsky'19)

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

for all $\log n \lesssim k \lesssim n \log n$.

What is known?

A smaller quantity: (Hao and Orlitsky'19)

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

for all $\log n \lesssim k \lesssim n\log n$.

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

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

The new result

Theorem (Han'20)

Under mild conditions on $\widehat{p}$ (satisfied by both LMM and PML),

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

The new result

Theorem (Han'20)

Under mild conditions on $\widehat{p}$ (satisfied by both LMM and PML),

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

Implication: a strict penalty of adaptation when $k \ll n^{1/3}$, or equivalently, $\varepsilon \ll n^{-1/3}$

Implication on PML

Corollary

For any $c, c', C > 0$, it holds that

$$\log \left[\sup_{\varepsilon > 0} \sup_{F \in \mathcal{F}_{\mathrm{Lip}}} \frac{\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq C\varepsilon)}{\left(\inf_{\widehat{F}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon) \right)^c} \right] \gtrsim n^{1/3 - c'}.$$

Implication on PML

Corollary

For any $c, c', C > 0$, it holds that

$$\log \left[\sup_{\varepsilon>0} \sup_{F \in \mathcal{F}_{\mathrm{Lip}}} \frac{\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq C\varepsilon)}{\left(\inf_{\widehat{F}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon)\right)^c} \right] \gtrsim n^{1/3-c'}.$$

- the competitive factor $\exp(O(n^{1/3+c}))$ cannot be improved to $\exp(O(n^{1/3-c}))$ in general

Implication on PML

Corollary

For any $c, c', C > 0$, it holds that

$$\log \left[\sup_{\varepsilon>0} \sup_{F \in \mathcal{F}_{\mathrm{Lip}}} \frac{\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq C\varepsilon)}{\left(\inf_{\widehat{F}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon)\right)^c} \right] \gtrsim n^{1/3-c'}.$$

- the competitive factor $\exp(O(n^{1/3+c}))$ cannot be improved to $\exp(O(n^{1/3-c}))$ in general
- the optimality requirement $\varepsilon \gg n^{-1/3}$ of PML is not superfluous

Implication on PML

Corollary

For any $c, c', C > 0$, it holds that

$$\log \left[\sup_{\varepsilon>0} \sup_{F \in \mathcal{F}_{\mathrm{Lip}}} \frac{\sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|F(p^{\mathrm{PML}}) - F(p)| \geq C\varepsilon)}{\left(\inf_{\widehat{F}} \sup_{p \in \mathcal{M}_k} \mathbb{P}_p(|\widehat{F} - F(p)| \geq \varepsilon)\right)^c} \right] \gtrsim n^{1/3-c'}.$$

- the competitive factor $\exp(O(n^{1/3+c}))$ cannot be improved to $\exp(O(n^{1/3-c}))$ in general
- the optimality requirement $\varepsilon \gg n^{-1/3}$ of PML is not superfluous
- caution: does not rule out the possibility that PML could be fully optimal for a given functional

Comparison with usual adaptive estimation

General minimax formulation:

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

Comparison with usual adaptive estimation

General minimax formulation:

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

Many past work: adaptation to a class of parameter sets $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$

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

Comparison with usual adaptive estimation

General minimax formulation:

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

Many past work: adaptation to a class of parameter sets $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$

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

This work: adaptation to a class of loss functions $L \in \mathcal{L}$

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

Comparison with usual adaptive estimation

General minimax formulation:

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

Many past work: adaptation to a class of parameter sets $\Theta_1 \subseteq \Theta_2 \subseteq \cdots$

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

This work: adaptation to a class of loss functions $L \in \mathcal{L}$

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

- in our problem, $L_F(p, \widehat{p}) = |F(p) - F(\widehat{p})|$, and $\mathcal{L} = \{L_F : F \text{ is 1-Lip}\}$

High-level proof idea

Traditional hypothesis testing argument: find $\theta_1, \cdots, \theta_M \in \Theta$ such that the following conditions hold:

High-level proof idea

Traditional hypothesis testing argument: find $\theta_1, \cdots, \theta_M \in \Theta$ such that the following conditions hold:

- separation condition: for all $i \neq j$,

$$\inf_a [L(\theta_i, a) + L(\theta_j, a)] \geq \Delta;$$

High-level proof idea

Traditional hypothesis testing argument: find $\theta_1, \cdots, \theta_M \in \Theta$ such that the following conditions hold:

- separation condition: for all $i \neq j$,

$$\inf_a [L(\theta_i, a) + L(\theta_j, a)] \geq \Delta;$$

- indistinguishability condition: the learner could not distinguish from the individual θ_i's with their mixture.

High-level proof idea

Traditional hypothesis testing argument: find $\theta_1, \cdots, \theta_M \in \Theta$ such that the following conditions hold:

- separation condition: for all $i \neq j$,

$$\inf_a [L(\theta_i, a) + L(\theta_j, a)] \geq \Delta;$$

- indistinguishability condition: the learner could not distinguish from the individual θ_i's with their mixture.

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

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

Summary of Part II

- $\varepsilon \gg n^{-1/3}$ lower bound for general adaptive approaches
- tight lower bound analysis of PML
- a strictly larger adaptive minimax risk for functional estimation

Concluding remarks

Tight optimality and limitation of the PML plug-in approach:

- optimality: improved upper bound from $\varepsilon \gg n^{-1/4}$ to $\varepsilon \gg n^{-1/3}$
- limitation: a novel $\varepsilon \gg n^{-1/3}$ lower bound for general adaptive approaches

Thank you!