Approximate Independence of Permutation Mixtures

Yanjun Han (NYU Courant Math and CDS)

Joint work with:

Jonathan Niles-Weed (NYU Courant Math and CDS)

CSP Seminar, UMich EECS
October 3, 2024

Setup

Let $P_1, \ldots, P_n$ be n probability distributions over the same space.

A permutation mixture $\mathbb{P}_n$:

- $\rightarrow$ draw independent $Z_1 \sim P_1, \ldots, Z_n \sim P_n$;
- $\rightarrow$ draw a uniformly random permutation $\pi \sim \mathrm{Unif}(S_n)$;
- $\rightarrow$ $\mathbb{P}_n$ is the joint distribution of $(X_1, \ldots, X_n)$ with $X_i = Z_{\pi(i)}$;
- $\rightarrow$ in mathematical terms:

$$(X_1, \cdots, X_n) \sim \mathbb{E}_{\pi \sim \mathrm{Unif}(S_n)} \left[\otimes_{i=1}^{n} P_{\pi(i)} \right] \quad \text{under } \mathbb{P}_n.$$

Setup

Let $P_1, \dots, P_n$ be n probability distributions over the same space.

A permutation mixture $\mathbb{P}_n$:

- $\rightarrow$ draw independent $Z_1 \sim P_1, \dots, Z_n \sim P_n$;
- $\rightarrow$ draw a uniformly random permutation $\pi \sim \mathrm{Unif}(S_n)$;
- $\rightarrow$ $\mathbb{P}_n$ is the joint distribution of $(X_1, \dots, X_n)$ with $X_i = Z_{\pi(i)}$;
- $\rightarrow$ in mathematical terms:

$$(X_1, \cdots, X_n) \sim \mathbb{E}_{\pi \sim \mathrm{Unif}(S_n)} \left[\otimes_{i=1}^n P_{\pi(i)} \right] \quad \text{under } \mathbb{P}_n.$$

An i.i.d. (mean-field) approximation $\mathbb{Q}_n$:

$$(X_1, \cdots, X_n) \sim \left(\frac{1}{n} \sum_{i=1}^n P_i \right)^{\otimes n} \quad \text{under } \mathbb{Q}_n.$$

Setup

Let $P_1, \dots, P_n$ be n probability distributions over the same space.

A permutation mixture $\mathbb{P}_n$:

- → draw independent $Z_1 \sim P_1, \dots, Z_n \sim P_n$;
- → draw a uniformly random permutation $\pi \sim \mathrm{Unif}(S_n)$;
- → $\mathbb{P}_n$ is the joint distribution of $(X_1, \dots, X_n)$ with $X_i = Z_{\pi(i)}$;
- → in mathematical terms:

$$(X_1, \cdots, X_n) \sim \mathbb{E}_{\pi \sim \mathrm{Unif}(S_n)} \left[\otimes_{i=1}^n P_{\pi(i)} \right] \quad \text{under } \mathbb{P}_n.$$

An i.i.d. (mean-field) approximation $\mathbb{Q}_n$:

$$(X_1, \cdots, X_n) \sim \left(\frac{1}{n} \sum_{i=1}^n P_i \right)^{\otimes n} \quad \text{under } \mathbb{Q}_n.$$

Target of this work

Show that the i.i.d. approximation $\mathbb{Q}_n$ to $\mathbb{P}_n$ is accurate, i.e. the information divergence (or statistical distance) between $\mathbb{P}_n$ and $\mathbb{Q}_n$ is small (and ideally, independent of n)

Motivation

Later in the talk:

- $\rightarrow$ statistical and IT applications involving permutations
- $\rightarrow$ compound decisions in empirical Bayes
- $\rightarrow$ de Finetti-style theorems

Motivation

Later in the talk:

- $\rightarrow$ statistical and IT applications involving permutations
- $\rightarrow$ compound decisions in empirical Bayes
- $\rightarrow$ de Finetti-style theorems

Bigger picture:

- $\rightarrow$ general mean-field approximation
- $\rightarrow$ information geometry of high-dimensional mixtures

A toy example

Let $P_1 = \cdots = P_{n/2} = \mathcal{N}(\mu, 1)$ and $P_{n/2+1} = \cdots = P_n = \mathcal{N}(-\mu, 1)$

- $\rightarrow$ $\mathbb{P}_n = \nu_{\mathbb{P}} \star \mathcal{N}(0, I_n)$, where $\nu_{\mathbb{P}}$ is the distribution of n uniformly random draws from the multiset $\{-\mu, \ldots, -\mu, \mu, \ldots, \mu\}$ without replacement;
- $\rightarrow$ $\mathbb{Q}_n = \nu_{\mathbb{Q}} \star \mathcal{N}(0, I_n)$, where $\nu_{\mathbb{Q}}$ is the counterpart with replacement;

$\chi^2(P \| Q) := \sum_x \frac{(p_x - q_x)^2}{q_x}$

A toy example

Let $P_1 = \cdots = P_{n/2} = \mathcal{N}(\mu, 1)$ and $P_{n/2+1} = \cdots = P_n = \mathcal{N}(-\mu, 1)$

- → $\mathbb{P}_n = \nu_{\mathbb{P}} \star \mathcal{N}(0, I_n)$, where $\nu_{\mathbb{P}}$ is the distribution of n uniformly random draws from the multiset $\{-\mu, \ldots, -\mu, \mu, \ldots, \mu\}$ without replacement;
- → $\mathbb{Q}_n = \nu_{\mathbb{Q}} \star \mathcal{N}(0, I_n)$, where $\nu_{\mathbb{Q}}$ is the counterpart with replacement;

Our result

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = \begin{cases} O(\mu^4) & \text{if } \mu \leq 1, \\ O(\exp(\mu^2)) & \text{if } \mu > 1. \end{cases}$$

- → χ^2-divergence independent of dimension n
- → smaller than the one-dimensional divergence $\chi^2(\mathcal{N}(\mu, 1) \| \mathcal{N}(-\mu, 1))$
- → existing approaches fail even for this toy example

$\chi^2(P\|Q) := \sum_x \frac{(p_x - q_x)^2}{q_x}$

Failed approach I: reduction to two simple distributions

Apply convexity to reduce to the divergence between two simple distributions:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_n \| \mathbb{Q}_n) &= \mathrm{KL}(\mathbb{E}_{\vartheta \sim \nu_{\mathbb{P}}}[\mathcal{N}(\vartheta, I_n)] \| \mathbb{E}_{\vartheta' \sim \nu_{\mathbb{Q}}}[\mathcal{N}(\vartheta', I_n)]) \\ &\leq \min_{\rho \in \Pi(\nu_{\mathbb{P}}, \nu_{\mathbb{Q}})} \mathbb{E}_{(\vartheta, \vartheta') \sim \rho} \left[\mathrm{KL}(\mathcal{N}(\vartheta, I_n) \| \mathcal{N}(\vartheta', I_n))\right] \\ &= \frac{W_2^2(\nu_{\mathbb{P}}, \nu_{\mathbb{Q}})}{2} \asymp \sqrt{n}\mu^2\end{aligned}$$

$\mathrm{KL}(P \| Q) := \sum_x p(x) \log \frac{p(x)}{q(x)}$

Failed approach I: reduction to two simple distributions

Apply convexity to reduce to the divergence between two simple distributions:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_n \| \mathbb{Q}_n) &= \mathrm{KL}(\mathbb{E}_{\vartheta \sim \nu_{\mathbb{P}}}[\mathcal{N}(\vartheta, I_n)] \| \mathbb{E}_{\vartheta' \sim \nu_{\mathbb{Q}}}[\mathcal{N}(\vartheta', I_n)]) \\ &\leq \min_{\rho \in \Pi(\nu_{\mathbb{P}}, \nu_{\mathbb{Q}})} \mathbb{E}_{(\vartheta, \vartheta') \sim \rho} \left[\mathrm{KL}(\mathcal{N}(\vartheta, I_n) \| \mathcal{N}(\vartheta', I_n))\right] \\ &= \frac{W_2^2(\nu_{\mathbb{P}}, \nu_{\mathbb{Q}})}{2} \asymp \sqrt{n}\mu^2\end{aligned}$$

→ grows with the dimension n

→ wrong dependence on μ

$\mathrm{KL}(P \| Q) := \sum_x p(x) \log \frac{p(x)}{q(x)}$

Failed approach II: reduction to one simple distribution

A more careful coupling:

$$\mathsf{KL}(\mathbb{P}_n \| \mathbb{Q}_n) \le \min_{\{\nu_{\theta'}\}_{\theta' \in \{\pm\mu\}^n}} \mathbb{E}_{\vartheta' \sim \nu_{\mathbb{Q}}} \left[\mathsf{KL}\left(\mathbb{E}_{\vartheta \sim \nu_{\vartheta'}} [\mathcal{N}(\vartheta, I_n)] \, \| \mathcal{N}(\vartheta', I_n) \right) \right],$$

where the minimization is over all possible families of distributions $\{\nu_{\theta'}\}_{\theta' \in \{\pm\mu\}^n}$ such that $\mathbb{E}_{\vartheta' \sim \nu_{\mathbb{Q}}}[\nu_{\vartheta'}] = \nu_{\mathbb{P}}$.

Failed approach II: reduction to one simple distribution

A more careful coupling:

$$\mathrm{KL}(\mathbb{P}_n \| \mathbb{Q}_n) \le \min_{\{\nu_{\theta'}\}_{\theta' \in \{\pm\mu\}^n}} \mathbb{E}_{\vartheta' \sim \nu_{\mathbb{Q}}} \left[\mathrm{KL}\left(\mathbb{E}_{\vartheta \sim \nu_{\vartheta'}} [\mathcal{N}(\vartheta, I_n)] \| \mathcal{N}(\vartheta', I_n) \right) \right],$$

where the minimization is over all possible families of distributions $\{\nu_{\theta'}\}_{\theta' \in \{\pm\mu\}^n}$ such that $\mathbb{E}_{\vartheta' \sim \nu_{\mathbb{Q}}}[\nu_{\vartheta'}] = \nu_{\mathbb{P}}$.

- $\rightarrow$ a judicious choice in [Ding'22] leads to an upper bound $O(\mu^2)$ for small μ
- $\rightarrow$ however, can show that any such upper bound must be $\Omega(\mu^2)$

Failed approach III: method of moments

A powerful approach to upper bound the statistical difference between two mixtures distributions, with many recent applications [Cai and Low'11, Hardt and Price'15, Wu and Yang'20, Han et al.'20, ...]

$\mathrm{TV}(P, Q) := \frac{1}{2} \sum_x |p_x - q_x|$

Failed approach III: method of moments

A powerful approach to upper bound the statistical difference between two mixtures distributions, with many recent applications [Cai and Low'11, Hardt and Price'15, Wu and Yang'20, Han et al.'20, ...]

Idea: express the Gaussian likelihood ratio in terms of Hermite polynomials

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k,$$

$\mathrm{TV}(P, Q) := \frac{1}{2}\sum_x |p_x - q_x|$

Failed approach III: method of moments

A powerful approach to upper bound the statistical difference between two mixtures distributions, with many recent applications [Cai and Low'11, Hardt and Price'15, Wu and Yang'20, Han et al.'20, ...]

Idea: express the Gaussian likelihood ratio in terms of Hermite polynomials

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k,$$

so that

$$\begin{aligned}
\mathrm{TV}(\mu \star \mathcal{N}(0,1), \nu \star \mathcal{N}(0,1))^2 &= \frac{1}{4}\left(\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left|\mathbb{E}_{U\sim\mu}\left[\frac{\varphi(Z-U)}{\varphi(Z)}\right] - \mathbb{E}_{V\sim\nu}\left[\frac{\varphi(Z-V)}{\varphi(Z)}\right]\right|\right)^2 \\
&= \frac{1}{4}\left(\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left|\sum_{k=0}^{\infty}\frac{H_k(Z)}{k!}\left(\mathbb{E}_{U\sim\mu}[U^k] - \mathbb{E}_{V\sim\nu}[V^k]\right)\right|\right)^2 \\
&\overset{\text{C-S}}{\leq} \frac{1}{4}\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left(\sum_{k=0}^{\infty}\frac{H_k(Z)}{k!}\left(\mathbb{E}_{U\sim\mu}[U^k] - \mathbb{E}_{V\sim\nu}[V^k]\right)\right)^2 \\
&= \frac{1}{4}\sum_{k=0}^{\infty}\frac{(\mathbb{E}_{U\sim\mu}[U^k] - \mathbb{E}_{V\sim\nu}[V^k])^2}{k!}
\end{aligned}$$

$\mathrm{TV}(P,Q) := \frac{1}{2}\sum_x |p_x - q_x|$

Failed approach III: method of moments (cont'd)

In general dimensions:

$$\mathrm{TV}(\nu_{\mathbb{P}} \star \mathcal{N}(0, I_n) \| \nu_{\mathbb{Q}} \star \mathcal{N}(0, I_n))^2 \leq \frac{1}{4} \sum_{\vec{\alpha} \in \mathbb{N}^n} \frac{(m_{\vec{\alpha}}(\nu_{\mathbb{P}}) - m_{\vec{\alpha}}(\nu_{\mathbb{Q}}))^2}{\vec{\alpha}!}$$

→ $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is a multi-index, with $\vec{\alpha}! := \alpha_1! \cdots \alpha_n!$

→ $m_{\vec{\alpha}}(\mu) := \mathbb{E}_{\vartheta \sim \mu}[\vartheta_1^{\alpha_1} \cdots \vartheta_n^{\alpha_n}]$ denotes the joint moment

Failed approach III: method of moments (cont'd)

In general dimensions:

$$\mathsf{TV}(\nu_{\mathbb{P}} \star \mathcal{N}(0, I_n) \| \nu_{\mathbb{Q}} \star \mathcal{N}(0, I_n))^2 \le \frac{1}{4} \sum_{\vec{\alpha} \in \mathbb{N}^n} \frac{(m_{\vec{\alpha}}(\nu_{\mathbb{P}}) - m_{\vec{\alpha}}(\nu_{\mathbb{Q}}))^2}{\vec{\alpha}!}$$

→ $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)$ is a multi-index, with $\vec{\alpha}! := \alpha_1! \cdots \alpha_n!$

→ $m_{\vec{\alpha}}(\mu) := \mathbb{E}_{\vartheta \sim \mu}[\vartheta_1^{\alpha_1} \cdots \vartheta_n^{\alpha_n}]$ denotes the joint moment

Application to our toy example:

→ non-zero moment difference starting from $|\vec{\alpha}| = 2$, suggesting an $O(\mu^4)$ dependence

→ however, too many cross terms in high dimensions: the total contributions of $|\vec{\alpha}| = 2\ell$ are at least $\Omega_\ell(\mu^{4\ell} n^{\ell-1})$, which is growing with n for $\ell \ge 2$.

Failed approach IV: method of cumulants

A recent development based on cumulants [Schramm and Wein'22]:

$$\chi^2(\nu_{\mathbb{P}} \star \mathcal{N}(0, I_n) \| \nu_{\mathbb{Q}} \star \mathcal{N}(0, I_n)) \leq \sum_{\vec{\alpha} \in \mathbb{N}^d} \frac{\kappa_{\vec{\alpha}}^2}{\vec{\alpha}!},$$

where $\kappa_{\vec{\alpha}}$ is the joint cumulant

$$\kappa_{\vec{\alpha}} = \kappa_{\nu_{\mathbb{Q}}}\left(\frac{\mathrm{d}\nu_{\mathbb{P}}}{\mathrm{d}\nu_{\mathbb{Q}}}, \vartheta_1, \ldots, \vartheta_1, \vartheta_2, \ldots, \vartheta_2, \ldots, \vartheta_n\right).$$

$\kappa(X_1, \ldots, X_n) := \frac{\partial^n}{\partial t_1 \cdots \partial t_n}\Big|_{t_1 = \cdots = t_n = 0} \log \mathbb{E}\left[\exp\left(\sum_{i=1}^n t_i X_i\right)\right]$

Failed approach IV: method of cumulants

A recent development based on cumulants [Schramm and Wein'22]:

$$\chi^2(\nu_{\mathbb{P}} \star \mathcal{N}(0, I_n) \| \nu_{\mathbb{Q}} \star \mathcal{N}(0, I_n)) \leq \sum_{\vec{\alpha} \in \mathbb{N}^d} \frac{\kappa_{\vec{\alpha}}^2}{\vec{\alpha}!},$$

where $\kappa_{\vec{\alpha}}$ is the joint cumulant

$$\kappa_{\vec{\alpha}} = \kappa_{\nu_{\mathbb{Q}}} \left(\frac{\mathrm{d}\nu_{\mathbb{P}}}{\mathrm{d}\nu_{\mathbb{Q}}}, \vartheta_1, \ldots, \vartheta_1, \vartheta_2, \ldots, \vartheta_2, \ldots, \vartheta_n \right).$$

- $\to$ a better behavior for certain cross terms
- $\to$ however, can show that $\kappa_{(1,\ell,0,\ldots,0)} \asymp C^\ell \ell!$ for odd ℓ, so summing along this subsequence gives a diverging result

---

$\kappa(X_1, \ldots, X_n) := \frac{\partial^n}{\partial t_1 \cdots \partial t_n}\Big|_{t_1 = \cdots = t_n = 0} \log \mathbb{E}\left[\exp\left(\sum_{i=1}^n t_i X_i\right)\right]$

Let $P_1, \ldots, P_n \in \mathcal{P}$. Define the following dimension-independent quantities:

Definition (Quantities of $\mathcal{P}$)

- $\rightarrow$ χ^2 channel capacity: $\mathrm{C}_{\chi^2}(\mathcal{P}) = \sup_{\rho \in \Delta(\mathcal{P})} I_{\chi^2}(P; X)$, with $P \sim \rho$ and $X \sim P$
- $\rightarrow$ χ^2 diameter: $\mathrm{D}_{\chi^2}(\mathcal{P}) = \sup_{P_1, P_2 \in \mathcal{P}} \chi^2(P_1 \| P_2)$

$I_{\chi^2}(X; Y) := \chi^2(P_{XY} \| P_X P_Y)$

Main result

Let $P_1, \ldots, P_n \in \mathcal{P}$. Define the following dimension-independent quantities:

Definition (Quantities of $\mathcal{P}$)

- χ^2 channel capacity: $\mathsf{C}_{\chi^2}(\mathcal{P}) = \sup_{\rho \in \Delta(\mathcal{P})} I_{\chi^2}(P; X)$, with $P \sim \rho$ and $X \sim P$
- χ^2 diameter: $\mathsf{D}_{\chi^2}(\mathcal{P}) = \sup_{P_1, P_2 \in \mathcal{P}} \chi^2(P_1 \| P_2)$

Theorem

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) \le \min \left\{ 10 \sum_{\ell=2}^{n} \mathsf{C}_{\chi^2}(\mathcal{P})^{\ell}, (1 + \mathsf{D}_{\chi^2}(\mathcal{P}))^{1+\mathsf{C}_{\chi^2}(\mathcal{P})} - 1 \right\}$$

- $\mathbb{P}_n$ is contiguous to $\mathbb{Q}_n$: $\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = \mathcal{O}_{\mathcal{P}}(1)$ if $\mathsf{D}_{\chi^2}(\mathcal{P}) < \infty$
- high-probability events under the simpler product measure $\mathbb{Q}_n$ translate to high-probability events under the mixture $\mathbb{P}_n$

$I_{\chi^2}(X; Y) := \chi^2(P_{XY} \| P_X P_Y)$

Examples

Theorem

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) \leq \min\left\{ 10 \sum_{\ell=2}^{n} \mathsf{C}_{\chi^2}(\mathcal{P})^{\ell}, (1 + \mathsf{D}_{\chi^2}(\mathcal{P}))^{1+\mathsf{C}_{\chi^2}(\mathcal{P})} - 1 \right\}$$

Examples

Theorem

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) \le \min\left\{ 10 \sum_{\ell=2}^{n} \mathsf{C}_{\chi^2}(\mathcal{P})^\ell, (1+\mathsf{D}_{\chi^2}(\mathcal{P}))^{1+\mathsf{C}_{\chi^2}(\mathcal{P})} - 1 \right\}$$

Example I (Two-component Gaussian)

$\mathcal{P} = \{\mathcal{N}(\mu, 1), \mathcal{N}(-\mu, 1)\}$: $\mathsf{C}_{\chi^2}(\mathcal{P}) \le 1 - e^{-\mu^2}$, so

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = \begin{cases} O(\mu^4) & \text{if } \mu \le 1, \\ O(\exp(\mu^2)) & \text{if } \mu > 1. \end{cases}$$

Examples

Theorem

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) \le \min\left\{ 10 \sum_{\ell=2}^{n} \mathsf{C}_{\chi^2}(\mathcal{P})^{\ell}, (1 + \mathsf{D}_{\chi^2}(\mathcal{P}))^{1+\mathsf{C}_{\chi^2}(\mathcal{P})} - 1 \right\}$$

Example I (Two-component Gaussian)

$\mathcal{P} = \{\mathcal{N}(\mu, 1), \mathcal{N}(-\mu, 1)\}$: $\mathsf{C}_{\chi^2}(\mathcal{P}) \le 1 - e^{-\mu^2}$, so

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = \begin{cases} O(\mu^4) & \text{if } \mu \le 1, \\ O(\exp(\mu^2)) & \text{if } \mu > 1. \end{cases}$$

Example II (Bounded Gaussian)

$\mathcal{P} = \{\mathcal{N}(\theta, 1) : |\theta| \le \mu)\}$: $\mathsf{C}_{\chi^2}(\mathcal{P}) = O(\mu \wedge \mu^2), \mathsf{D}_{\chi^2}(\mathcal{P}) = \exp(O(\mu^2))$, so

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = \begin{cases} O(\mu^4) & \text{if } \mu \le 1, \\ \exp(O(\mu^3)) & \text{if } \mu > 1. \end{cases}$$

Applications

Permutation prior

Sequence model in statistics: observe $X_i \sim P_{\theta_i}$ with unknown $\theta = (\theta_1, \dots, \theta_n)$

- $\rightarrow$ statisticians would like to prove lower bounds on the estimation error of θ
- $\rightarrow$ a prevalent strategy is to impose a prior distribution on θ, and a permutation prior is sometimes preferred: $\theta = (v_{\pi(1)}, \dots, v_{\pi(n)})$ for a known vector v and a random permutation π
- $\rightarrow$ a key quantity in the analysis: mutual information $I(\theta; X^n)$

Permutation prior

Sequence model in statistics: observe $X_i \sim P_{\theta_i}$ with unknown $\theta = (\theta_1, \ldots, \theta_n)$

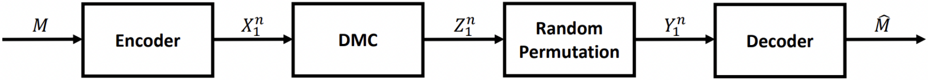
- $\rightarrow$ statisticians would like to prove lower bounds on the estimation error of θ
- $\rightarrow$ a prevalent strategy is to impose a prior distribution on θ, and a permutation prior is sometimes preferred: $\theta = (v_{\pi(1)}, \ldots, v_{\pi(n)})$ for a known vector v and a random permutation π
- $\rightarrow$ a key quantity in the analysis: mutual information $I(\theta; X^n)$

Our result: can pretend as if the coordinates $\theta_i \sim \frac{1}{n}\sum_{j=1}^{n} \delta_{v_j}$ are i.i.d.

Mutual information under a permutation prior

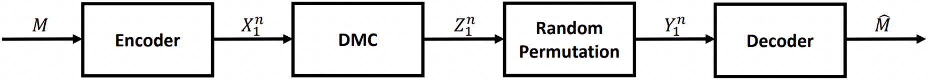
$$I_{\mathbb{Q}_n}(\theta; X^n) - \mathcal{O}_{\mathcal{P}}(1) \le I_{\mathbb{P}_n}(\theta; X^n) \le I_{\mathbb{Q}_n}(\theta; X^n)$$

Permutation channel



The noisy permutation channel introduced in [Makur'20]

Permutation channel



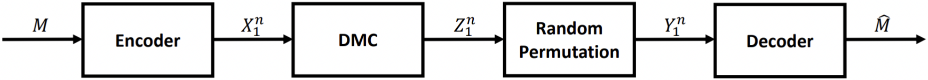
The noisy permutation channel introduced in [Makur'20]

→ target: find the channel capacity $C_n(\mathcal{P}) = \max_{p(x^n)} I(X^n; Y^n)$

→ known achievability [Makur'20] and converse [Tang and Polyanskiy'23]:

$$C_n(\mathcal{P}) \sim \frac{\mathrm{rank}(P_{Z|X}) - 1}{2} \log n \quad \text{for discrete } \mathcal{P}.$$

Permutation channel



The noisy permutation channel introduced in [Makur'20]

→ target: find the channel capacity $C_n(\mathcal{P}) = \max_{p(x^n)} I(X^n; Y^n)$

→ known achievability [Makur'20] and converse [Tang and Polyanskiy'23]:

$$C_n(\mathcal{P}) \sim \frac{\mathrm{rank}(P_{Z|X}) - 1}{2} \log n \quad \text{for discrete } \mathcal{P}.$$

Our result: for general $\mathcal{P}$, can pretend as if Y^n have independent coordinates

Converse for general permutation channels

$$C_n(\mathcal{P}) \leq \mathrm{Red}(\mathrm{conv}(\mathcal{P})^{\otimes n}) + \mathcal{O}_{\mathcal{P}}(1)$$

Theorem (de Finetti)

Any exchangeable distribution P_{X^∞} can be written as an i.i.d. mixture:

$$P_{X^\infty}(x^\infty) = \mathbb{E}_\theta \left[\prod_{i=1}^{\infty} Q_\theta(x_i) \right].$$

The joint distribution of $(X_1, \ldots, X_n)$ is exchangeable if $(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$

finite de Finetti theorems

Theorem (de Finetti)

Any exchangeable distribution P_{X^∞} can be written as an i.i.d. mixture:

$$P_{X^\infty}(x^\infty) = \mathbb{E}_\theta \left[\prod_{i=1}^{\infty} Q_\theta(x_i) \right].$$

Approximately holds for exchangeable distribution P_{X^n} with finite n:

- $\rightarrow$ [Diaconis and Freedman'80]: $\mathrm{KL}(P_{X^k} \| \mathbb{E}_\theta[Q_\theta^{\otimes k}]) \lesssim \frac{k^2}{n}$
- $\rightarrow$ [Stam'78]: for small $|\mathcal{X}|$, $\mathrm{KL}(P_{X^k} \| \mathbb{E}_\theta[Q_\theta^{\otimes k}]) \lesssim \frac{|\mathcal{X}| k^2}{n(n+1-k)}$
- $\rightarrow$ some recent refinements in [Gavalakis and Kontoyiannis'21; Johnson, Gavalakis, and Kontoyiannis'24]

The joint distribution of $(X_1, \ldots, X_n)$ is exchangeable if $(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$

Our extensions

Using first upper bound and $\mathsf{C}_{\chi^2}(\mathcal{P}) \leq |\mathcal{X}|$:

χ^2-type finite de Finetti

For exchangeable distribution P_{X^n} and $k \leq n$:

$$\chi^2\left(P_{X^k} \| \mathbb{E}_\theta[Q_\theta^{\otimes k}]\right) \lesssim \frac{k^2|\mathcal{X}|^2}{n^2} \quad \text{if } k < \frac{n}{|\mathcal{X}|}.$$

Our extensions

Using first upper bound and $\mathsf{C}_{\chi^2}(\mathcal{P}) \le |\mathcal{X}|$:

χ^2-type finite de Finetti

For exchangeable distribution P_{X^n} and $k \le n$:

$$\chi^2\left(P_{X^k} \| \mathbb{E}_\theta[Q_\theta^{\otimes k}]\right) \lesssim \frac{k^2|\mathcal{X}|^2}{n^2} \quad \text{if } k < \frac{n}{|\mathcal{X}|}.$$

Using second upper bound:

Noisy de Finetti

Let P_{Y^n} be the output distribution with an input exchangeable distribution P_{X^n} and a channel $\mathcal{P}$. Then for $k \le n$:

$$\chi^2\left(P_{Y^k} \| \mathbb{E}_\theta[Q_\theta^{\otimes k}]\right) = \mathcal{O}_{\mathcal{P}}\left(\frac{k^2}{n^2}\right) \quad \text{if } \mathsf{D}_{\chi^2}(\mathcal{P}) < \infty.$$

Empirical Bayes

The empirical Bayes framework [Robbins'51; '56]:

- $\rightarrow$ idea: estimate the prior distribution from data
- $\rightarrow$ lots of empirical successes but limited theoretical understanding

Empirical Bayes

The empirical Bayes framework [Robbins'51; '56]:

$\rightarrow$ idea: estimate the prior distribution from data

$\rightarrow$ lots of empirical successes but limited theoretical understanding

A new theoretical paradigm [Hannan and Robbins'55; Greenshtein and Ritov'09]:

$\rightarrow$ a simple setting: independent $X_i \sim P_{\theta_i}$, aim to estimate $\theta = (\theta_1, \dots, \theta_n)$

$\rightarrow$ target: find an estimator with a small regret compared with powerful oracles

$$\mathrm{regret}(\widehat{\theta}) = \mathbb{E}_\theta[L(\theta, \widehat{\theta})] - \inf_{\widehat{\theta}^{\mathrm{oracle}}} \mathbb{E}_\theta[L(\theta, \widehat{\theta}^{\mathrm{oracle}})]$$

$\rightarrow$ simple/separable oracle: $\widehat{\theta}_i^{\mathrm{S}} = f(X_i)$ for a single function f

$\rightarrow$ permutation invariant oracle:

$$\widehat{\theta}^{\mathrm{PI}}_{\pi(i)}(X_{\pi(1)}, \dots, X_{\pi(n)}) = \widehat{\theta}^{\mathrm{PI}}_i(X_1, \dots, X_n)$$

Empirical Bayes

The empirical Bayes framework [Robbins'51; '56]:

- → idea: estimate the prior distribution from data
- → lots of empirical successes but limited theoretical understanding

A new theoretical paradigm [Hannan and Robbins'55; Greenshtein and Ritov'09]:

- → a simple setting: independent $X_i \sim P_{\theta_i}$, aim to estimate $\theta = (\theta_1, \ldots, \theta_n)$
- → target: find an estimator with a small regret compared with powerful oracles

$$\mathrm{regret}(\widehat{\theta}) = \mathbb{E}_\theta[L(\theta, \widehat{\theta})] - \inf_{\widehat{\theta}^{\mathrm{oracle}}} \mathbb{E}_\theta[L(\theta, \widehat{\theta}^{\mathrm{oracle}})]$$

- → simple/separable oracle: $\widehat{\theta}_i^{\mathrm{S}} = f(X_i)$ for a single function f
- → permutation invariant oracle:

$$\widehat{\theta}_{\pi(i)}^{\mathrm{PI}}(X_{\pi(1)}, \ldots, X_{\pi(n)}) = \widehat{\theta}_i^{\mathrm{PI}}(X_1, \ldots, X_n)$$

Question

Do these oracles have similar estimation power?

Our results

Oracle analysis:

- $\rightarrow$ the true parameter $\theta = (\theta_1, \ldots, \theta_n)$ is known to both oracles
- $\rightarrow$ because of the limitations on the oracles, an equivalent formulation is that the oracle only knows the multiset $\{\theta_1, \ldots, \theta_n\}$ but not the order
- $\rightarrow$ equivalently, θ follows a permutation prior

Our results

Oracle analysis:

- → the true parameter $\theta = (\theta_1, \ldots, \theta_n)$ is known to both oracles
- → because of the limitations on the oracles, an equivalent formulation is that the oracle only knows the multiset $\{\theta_1, \ldots, \theta_n\}$ but not the order
- → equivalently, θ follows a permutation prior

Our technical ingredient: upper bound the distance between two permutation mixtures

TV upper bound for two permutation mixtures

$$\mathsf{TV}(\mathbb{P}_n^{-i}, \mathbb{P}_n^{-j}) = \mathcal{O}_{\mathcal{P}}\left(\frac{1}{\sqrt{n}}\right).$$

Our results

Oracle analysis:

- → the true parameter $\theta = (\theta_1, \ldots, \theta_n)$ is known to both oracles
- → because of the limitations on the oracles, an equivalent formulation is that the oracle only knows the multiset $\{\theta_1, \ldots, \theta_n\}$ but not the order
- → equivalently, θ follows a permutation prior

Our technical ingredient: upper bound the distance between two permutation mixtures

TV upper bound for two permutation mixtures

$$\mathsf{TV}(\mathbb{P}_n^{-i}, \mathbb{P}_n^{-j}) = \mathcal{O}_{\mathcal{P}}\left(\frac{1}{\sqrt{n}}\right).$$

Application to empirical Bayes:

Simple oracles are as powerful as permutation-invariant oracles

For bounded separable loss $L(\theta, \widehat{\theta}) = \sum_{i=1}^{n} L_i(\theta_i, \widehat{\theta}_i)$:

$$\inf_{\widehat{\theta}^{\mathrm{S}}} \frac{1}{n}\mathbb{E}_\theta[L(\theta, \widehat{\theta}^{\mathrm{S}})] - \inf_{\widehat{\theta}^{\mathrm{PI}}} \frac{1}{n}\mathbb{E}_\theta[L(\theta, \widehat{\theta}^{\mathrm{PI}})] = \mathcal{O}_{\mathcal{P}}\left(\frac{1}{\sqrt{n}}\right).$$

First upper bound via a new basis expansion

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!} \theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

$(\theta_1, \ldots, \theta_n) = (\mu, \ldots, \mu, -\mu, \ldots, -\mu)$.

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

Hyperbolic basis?

$$\frac{\varphi(x-\theta)}{\varphi_0(x)} = 1 + \tanh(\mu x)\frac{\theta}{\mu}, \quad \theta \in \{\pm\mu\}$$

where $\varphi_0(x) = \frac{\varphi(x-\mu)+\varphi(x+\mu)}{2}$ is the common marginal distribution of $\mathbb{P}_n$ and $\mathbb{Q}_n$

$(\theta_1, \ldots, \theta_n) = (\mu, \ldots, \mu, -\mu, \ldots, -\mu)$.

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

$\{H_0(x), H_1(x), \dots\}$ are orthogonal in $L^2(\varphi)$

Hyperbolic basis?

$$\frac{\varphi(x-\theta)}{\varphi_0(x)} = 1 + \tanh(\mu x)\frac{\theta}{\mu}, \quad \theta \in \{\pm\mu\}$$

where $\varphi_0(x) = \frac{\varphi(x-\mu)+\varphi(x+\mu)}{2}$ is the common marginal distribution of $\mathbb{P}_n$ and $\mathbb{Q}_n$

$(\theta_1, \dots, \theta_n) = (\mu, \dots, \mu, -\mu, \dots, -\mu)$.

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

$\{H_0(x), H_1(x), \dots\}$ are orthogonal in $L^2(\varphi)$

Hyperbolic basis?

$$\frac{\varphi(x-\theta)}{\varphi_0(x)} = 1 + \tanh(\mu x)\frac{\theta}{\mu}, \quad \theta \in \{\pm\mu\}$$

where $\varphi_0(x) = \frac{\varphi(x-\mu)+\varphi(x+\mu)}{2}$ is the common marginal distribution of $\mathbb{P}_n$ and $\mathbb{Q}_n$

$\{1, \tanh(\mu x)\}$ are orthogonal in $L^2(\varphi_0)$

$(\theta_1, \dots, \theta_n) = (\mu, \dots, \mu, -\mu, \dots, -\mu)$.

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

$\{H_0(x), H_1(x), \dots\}$ are orthogonal in $L^2(\varphi)$

Hyperbolic basis?

$$\frac{\varphi(x-\theta)}{\varphi_0(x)} = 1 + \tanh(\mu x)\frac{\theta}{\mu}, \quad \theta \in \{\pm\mu\}$$

where $\varphi_0(x) = \frac{\varphi(x-\mu)+\varphi(x+\mu)}{2}$ is the common marginal distribution of $\mathbb{P}_n$ and $\mathbb{Q}_n$

$\{1, \tanh(\mu x)\}$ are orthogonal in $L^2(\varphi_0)$

$\mathbb{E}[\frac{\theta}{\mu}] = 0$ for $\theta \sim \mathrm{Unif}(\{\pm\mu\})$

$(\theta_1, \dots, \theta_n) = (\mu, \dots, \mu, -\mu, \dots, -\mu)$.

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

$\{H_0(x), H_1(x), \dots\}$ are orthogonal in $L^2(\varphi)$

$\mathbb{E}[\theta^k]$ possibly non-zero for $\theta \sim \text{Unif}(\{\pm\mu\})$

Hyperbolic basis?

$$\frac{\varphi(x-\theta)}{\varphi_0(x)} = 1 + \tanh(\mu x)\frac{\theta}{\mu}, \quad \theta \in \{\pm\mu\}$$

where $\varphi_0(x) = \frac{\varphi(x-\mu)+\varphi(x+\mu)}{2}$ is the common marginal distribution of $\mathbb{P}_n$ and $\mathbb{Q}_n$

$\{1, \tanh(\mu x)\}$ are orthogonal in $L^2(\varphi_0)$

$\mathbb{E}[\frac{\theta}{\mu}] = 0$ for $\theta \sim \text{Unif}(\{\pm\mu\})$

$(\theta_1, \dots, \theta_n) = (\mu, \dots, \mu, -\mu, \dots, -\mu)$.

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

$\{H_0(x), H_1(x), \dots\}$ are orthogonal in $L^2(\varphi)$

$\mathbb{E}[\theta^k]$ possibly non-zero for $\theta \sim \text{Unif}(\{\pm\mu\})$

Hyperbolic basis?

$$\frac{\varphi(x-\theta)}{\varphi_0(x)} = 1 + \tanh(\mu x)\frac{\theta}{\mu}, \quad \theta \in \{\pm\mu\}$$

where $\varphi_0(x) = \frac{\varphi(x-\mu)+\varphi(x+\mu)}{2}$ is the common marginal distribution of $\mathbb{P}_n$ and $\mathbb{Q}_n$

$\{1, \tanh(\mu x)\}$ are orthogonal in $L^2(\varphi_0)$

$\mathbb{E}[\frac{\theta}{\mu}] = 0$ for $\theta \sim \text{Unif}(\{\pm\mu\})$

Under the new basis:

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}(x^n) = \mathbb{E}_\pi\left[\prod_{i=1}^n \frac{\varphi(x_i - \theta_{\pi(i)})}{\varphi_0(x_i)}\right] = \mathbb{E}_\pi\left[\prod_{i=1}^n \left(1 + \tanh(\mu x_i)\frac{\theta_{\pi(i)}}{\mu}\right)\right]$$

$(\theta_1, \dots, \theta_n) = (\mu, \dots, \mu, -\mu, \dots, -\mu)$.

Toy example: a different basis

Hermite basis:

$$\frac{\varphi(x-\theta)}{\varphi(x)} = \sum_{k=0}^{\infty} \frac{H_k(x)}{k!}\theta^k$$

where φ is the density of $\mathcal{N}(0,1)$.

$\{H_0(x), H_1(x), \dots\}$ are orthogonal in $L^2(\varphi)$

$\mathbb{E}[\theta^k]$ possibly non-zero for $\theta \sim \mathrm{Unif}(\{\pm\mu\})$

Hyperbolic basis?

$$\frac{\varphi(x-\theta)}{\varphi_0(x)} = 1 + \tanh(\mu x)\frac{\theta}{\mu}, \quad \theta \in \{\pm\mu\}$$

where $\varphi_0(x) = \frac{\varphi(x-\mu)+\varphi(x+\mu)}{2}$ is the common marginal distribution of $\mathbb{P}_n$ and $\mathbb{Q}_n$

$\{1, \tanh(\mu x)\}$ are orthogonal in $L^2(\varphi_0)$

$\mathbb{E}[\frac{\theta}{\mu}] = 0$ for $\theta \sim \mathrm{Unif}(\{\pm\mu\})$

Under the new basis:

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}(x^n) = \mathbb{E}_\pi\left[\prod_{i=1}^n \frac{\varphi(x_i-\theta_{\pi(i)})}{\varphi_0(x_i)}\right] = \mathbb{E}_\pi\left[\prod_{i=1}^n\left(1+\tanh(\mu x_i)\frac{\theta_{\pi(i)}}{\mu}\right)\right]$$

$$= \sum_{S\subseteq[n]} \mathbb{E}_\pi\left[\prod_{i\in S}\frac{\theta_{\pi(i)}}{\mu}\right]\prod_{i\in S}\tanh(\mu x_i)$$

$(\theta_1, \dots, \theta_n) = (\mu, \dots, \mu, -\mu, \dots, -\mu)$.

Toy example: full analysis

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}(x^n) = \sum_{S \subseteq [n]} \mathbb{E}_\pi \left[\prod_{i \in S} \frac{\theta_{\pi(i)}}{\mu} \right] \prod_{i \in S} \tanh(\mu x_i)$$

Toy example: full analysis

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}(x^n) = \sum_{S\subseteq[n]} \mathbb{E}_\pi \left[\prod_{i\in S} \frac{\theta_{\pi(i)}}{\mu}\right] \prod_{i\in S} \tanh(\mu x_i)$$

$\rightarrow$ orthogonality of $\{1, \tanh(\mu x)\}$ under $L^2(\varphi_0)$:

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S\subseteq[n]} \left(\mathbb{E}_\pi \left[\prod_{i\in S} \frac{\theta_{\pi(i)}}{\mu}\right]\right)^2 \mathsf{C}_{\chi^2}(\mathcal{P})^{|S|}$$

Toy example: full analysis

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}(x^n) = \sum_{S\subseteq[n]} \mathbb{E}_\pi \left[\prod_{i\in S} \frac{\theta_{\pi(i)}}{\mu}\right] \prod_{i\in S} \tanh(\mu x_i)$$

$\rightarrow$ orthogonality of $\{1, \tanh(\mu x)\}$ under $L^2(\varphi_0)$:

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S\subseteq[n]} \left(\mathbb{E}_\pi \left[\prod_{i\in S} \frac{\theta_{\pi(i)}}{\mu}\right]\right)^2 \mathsf{C}_{\chi^2}(\mathcal{P})^{|S|}$$

$\rightarrow$ the inner expectation: for $|S| = \ell$,

$$\left(\mathbb{E}_\pi \left[\prod_{i\in S} \frac{\theta_{\pi(i)}}{\mu}\right]\right)^2 \le \frac{\mathbb{1}_{\ell \text{ is even}}}{\binom{n}{\ell}}$$

Toy example: full analysis

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}(x^n) = \sum_{S\subseteq[n]} \mathbb{E}_\pi \left[\prod_{i\in S} \frac{\theta_{\pi(i)}}{\mu}\right] \prod_{i\in S} \tanh(\mu x_i)$$

→ orthogonality of $\{1, \tanh(\mu x)\}$ under $L^2(\varphi_0)$:

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S\subseteq[n]} \left(\mathbb{E}_\pi\left[\prod_{i\in S}\frac{\theta_{\pi(i)}}{\mu}\right]\right)^2 \mathsf{C}_{\chi^2}(\mathcal{P})^{|S|}$$

→ the inner expectation: for $|S| = \ell$,

$$\left(\mathbb{E}_\pi\left[\prod_{i\in S}\frac{\theta_{\pi(i)}}{\mu}\right]\right)^2 \le \frac{\mathbb{1}_{\ell \text{ is even}}}{\binom{n}{\ell}}$$

→ piecing everything together:

$$\chi^2(\mathbb{P}_n\|\mathbb{Q}_n) = \mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] - 1 \le \mathsf{C}_{\chi^2}(\mathcal{P})^2 + \mathsf{C}_{\chi^2}(\mathcal{P})^4 + \cdots + \mathsf{C}_{\chi^2}(\mathcal{P})^n$$

General case: doubly centered expansion

Doubly centered expansion:

$$\frac{\mathrm{d}P_i}{\mathrm{d}\overline{P}}(x) = 1 + \Psi_i(x)$$

$\overline{P} := \frac{1}{n} \sum_{i=1}^{n} P_i$

General case: doubly centered expansion

Doubly centered expansion:

$$\frac{\mathrm{d}P_i}{\mathrm{d}\overline{P}}(x) = 1 + \Psi_i(x)$$

Centered in the x direction: $\mathbb{E}_{x \sim \overline{P}}[\Psi_i(x)] = 0$ for all i

$\rightarrow$ this leads to

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S \subseteq [n]} \mathbb{E}_{\mathbb{Q}_n}\left[\mathbb{E}_\pi\left(\prod_{i \in S} \Psi_{\pi(i)}(X_i)\right)\right]^2$$

$\overline{P} := \frac{1}{n}\sum_{i=1}^n P_i$

General case: doubly centered expansion

Doubly centered expansion:

$$\frac{\mathrm{d}P_i}{\mathrm{d}\overline{P}}(x) = 1 + \Psi_i(x)$$

Centered in the x direction: $\mathbb{E}_{x\sim\overline{P}}[\Psi_i(x)] = 0$ for all i

$\rightarrow$ this leads to

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S\subseteq[n]} \mathbb{E}_{\mathbb{Q}_n}\left[\mathbb{E}_\pi\left(\prod_{i\in S}\Psi_{\pi(i)}(X_i)\right)\right]^2$$

Centered in the i direction: $\mathbb{E}_{I\sim\mathrm{Unif}(\{1,\ldots,n\})}[\Psi_I(x)] = 0$ for all x

$\rightarrow$ how does this lead to a small value of $\left|\mathbb{E}_\pi\left(\prod_{i\in S}\Psi_{\pi(i)}(X_i)\right)\right|$?

$\overline{P} := \frac{1}{n}\sum_{i=1}^n P_i$

Importance of centering

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S\subseteq[n]} \mathbb{E}_{\mathbb{Q}_n}\left[\mathbb{E}_\pi\left(\prod_{i\in S}\Psi_{\pi(i)}(X_i)\right)\right]^2$$

$\mathrm{Perm}(A) := \sum_{\pi\in S_n}\prod_{i=1}^n A_{i,\pi(i)}$ is the matrix permanent, and A_T selects the columns of A with indices in T

Importance of centering

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S\subseteq[n]} \mathbb{E}_{\mathbb{Q}_n}\left[\mathbb{E}_\pi\left(\prod_{i\in S}\Psi_{\pi(i)}(X_i)\right)\right]^2$$

A key technical lemma:

An inequality for centered matrix

Let $A = (a_{ij}) \in \mathbb{R}^{\ell\times n}$ be a real matrix with $1 \le \ell \le n$ with all row sums being zero, and normalized properly with $\sum_{j=1}^n a_{ij}^2 = n$ for all $i \in [\ell]$. Then the following inequality holds:

$$\left|\frac{1}{\ell!}\sum_{T\subseteq[n],|T|=\ell} \mathrm{Perm}(A_T)\right| \le \sqrt{10\binom{n}{\ell}}$$

$\mathrm{Perm}(A) := \sum_{\pi\in S_n}\prod_{i=1}^n A_{i,\pi(i)}$ is the matrix permanent, and A_T selects the columns of A with indices in T

Importance of centering

$$\mathbb{E}_{\mathbb{Q}_n}\left[\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)^2\right] = \sum_{S\subseteq[n]} \mathbb{E}_{\mathbb{Q}_n}\left[\mathbb{E}_\pi\left(\prod_{i\in S}\Psi_{\pi(i)}(X_i)\right)\right]^2$$

A key technical lemma:

An inequality for centered matrix

Let $A = (a_{ij}) \in \mathbb{R}^{\ell\times n}$ be a real matrix with $1 \le \ell \le n$ with all row sums being zero, and normalized properly with $\sum_{j=1}^n a_{ij}^2 = n$ for all $i \in [\ell]$. Then the following inequality holds:

$$\left|\frac{1}{\ell!}\sum_{T\subseteq[n],|T|=\ell} \mathrm{Perm}(A_T)\right| \le \sqrt{10\binom{n}{\ell}}$$

- $\rightarrow$ centering is important: without it, the quantity is $\binom{n}{\ell}$ for the all-ones matrix A
- $\rightarrow$ this squared root saving crucially prevents the final coefficients from growing with n
- $\rightarrow$ the proof is the main technical challenge (see following slides)

$\mathrm{Perm}(A) := \sum_{\pi\in S_n}\prod_{i=1}^n A_{i,\pi(i)}$ is the matrix permanent, and A_T selects the columns of A with indices in T

Step I: symmetric multilinear forms

A deep result due to S. Banach [Banach'38]:

Banach's Theorem

Let $L(x_1, \ldots, x_n)$ be a symmetric multilinear form from a Hilbert space to either $\mathbb{R}$ or $\mathbb{C}$. Then

$$\sup\{|L(x_1, x_2, \ldots, x_n)| : |x_1| \leq 1, \ldots, |x_n| \leq 1\} = \sup\{|L(x, x, \ldots, x)| : |x| \leq 1\}.$$

Step I: symmetric multilinear forms

A deep result due to S. Banach [Banach'38]:

Banach's Theorem

Let $L(x_1, \ldots, x_n)$ be a symmetric multilinear form from a Hilbert space to either $\mathbb{R}$ or $\mathbb{C}$. Then

$$\sup\left\{|L(x_1, x_2, \ldots, x_n)| : |x_1| \leq 1, \ldots, |x_n| \leq 1\right\} = \sup\left\{|L(x, x, \ldots, x)| : |x| \leq 1\right\}.$$

→ the target quantity

$$L(r_1, \ldots, r_\ell) := \frac{1}{\ell!} \sum_{T \subseteq [n], |T| = \ell} \mathrm{Perm}(A_T)$$

is symmetric and multilinear in the rows $r_1, \ldots, r_\ell \in \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$

→ Banach's theorem shows that it suffices to consider A with identical rows $x \in \mathbb{R}^n$

→ the target quantity then becomes the elementary symmetric polynomial (ESP) $e_\ell(x) = \sum_{|S|=\ell} \prod_{i \in S} x_i$

Step II: a Maclaurin-type inequality

We are done once we prove the following inequality for ESPs:

Theorem (Upper bound on ESPs for centered vector)

Let $\sum_{i=1}^{n} x_i = 0$ and $\sum_{i=1}^{n} |x_i|^2 = n$.

→ If $x \in \mathbb{R}^n$, then $|e_\ell(x)|^2 \leq 10\binom{n}{\ell}$;

→ If $x \in \mathbb{C}^n$, a weaker upper bound holds:

$$|e_\ell(x)|^2 \leq \frac{n^n}{\ell^\ell (n-\ell)^{n-\ell}} < 3\sqrt{\ell+1}\binom{n}{\ell}.$$

$e_\ell(x) := \sum_{|S|=\ell} \prod_{i \in S} x_i$

Step II: a Maclaurin-type inequality

We are done once we prove the following inequality for ESPs:

Theorem (Upper bound on ESPs for centered vector)

Let $\sum_{i=1}^n x_i = 0$ and $\sum_{i=1}^n |x_i|^2 = n$.

→ If $x \in \mathbb{R}^n$, then $|e_\ell(x)|^2 \leq 10\binom{n}{\ell}$;

→ If $x \in \mathbb{C}^n$, a weaker upper bound holds:

$$|e_\ell(x)|^2 \leq \frac{n^n}{\ell^\ell (n-\ell)^{n-\ell}} < 3\sqrt{\ell+1}\binom{n}{\ell}.$$

→ similar problems have been recently studied in [Gopalan and Yehudayoff'14; Meka, Reingold, and Tal'19; Doron, Hatami, and Hoza'20; Tao'23]

→ best known bound due to [Tao'23]:

$$|e_\ell(x)|^2 \leq \binom{n}{\ell}^2 \left(\frac{\ell-1}{n-1}\right)^\ell \leq e^\ell \binom{n}{\ell}$$

→ we crucially need to improve the base e to the best possible constant 1

$e_\ell(x) := \sum_{|S|=\ell} \prod_{i \in S} x_i$

Proof of the key inequality

For the real case, can argue via the method of Lagrangian multipliers that the maximizer $x^\star$ is only supported on two points, i.e. it suffices to consider $x = x^{(k)}$ for some k:

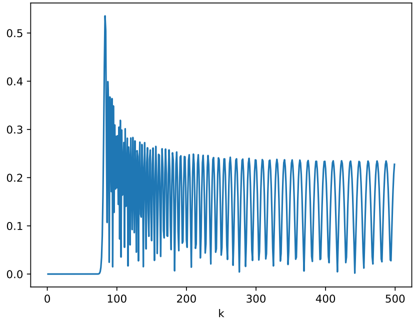
$$x^{(k)} = \Big(\underbrace{\sqrt{\frac{k}{n-k}}, \ldots, \sqrt{\frac{k}{n-k}}}_{n-k \text{ copies}}, \underbrace{-\sqrt{\frac{n-k}{k}}, \ldots, -\sqrt{\frac{n-k}{k}}}_{k \text{ copies}}\Big)$$

Proof of the key inequality

For the real case, can argue via the method of Lagrangian multipliers that the maximizer $x^\star$ is only supported on two points, i.e. it suffices to consider $x = x^{(k)}$ for some k:

$$x^{(k)} = \Big(\underbrace{\sqrt{\frac{k}{n-k}}, \dots, \sqrt{\frac{k}{n-k}}}_{n-k \text{ copies}}, \underbrace{-\sqrt{\frac{n-k}{k}}, \dots, -\sqrt{\frac{n-k}{k}}}_{k \text{ copies}}\Big)$$

However, upper bounding $|e_\ell(x^{(k)})|$ is still very challenging!!



The quantity $|e_\ell(x^{(k)})|^2 / \binom{n}{\ell}$ vs. k for $n = 1000, \ell = 300$.

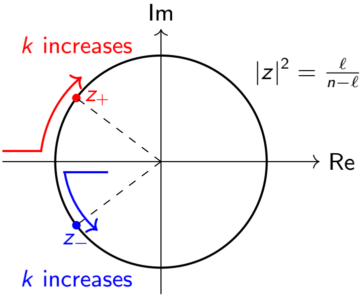
Saddle point analysis

$$\text{Cauchy's formula :} \quad e_\ell(x) = \frac{1}{2\pi \mathrm{i}} \oint_\Gamma \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \frac{\mathrm{d}z}{z}$$

$$\text{Saddle point equation :} \quad \frac{\ell}{z} = \sum_{i=1}^n \frac{x_i}{1 + x_i z}$$

$$\text{Cauchy's formula :} \quad e_\ell(x) = \frac{1}{2\pi i} \oint_\Gamma \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \frac{dz}{z}$$

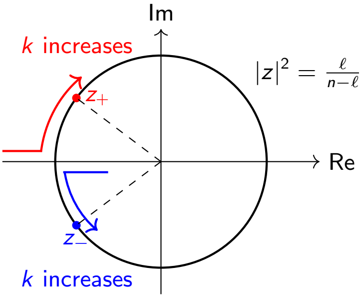
$$\text{Saddle point equation :} \quad \frac{\ell}{z} = \sum_{i=1}^n \frac{x_i}{1 + x_i z}$$



Saddle points for $x = x^{(k)}$

$$\text{Cauchy's formula :}\quad e_\ell(x) = \frac{1}{2\pi \mathrm{i}} \oint_\Gamma \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \frac{\mathrm{d}z}{z}$$

$$\text{Saddle point equation :}\quad \frac{\ell}{z} = \sum_{i=1}^n \frac{x_i}{1 + x_i z}$$



Saddle points for $x = x^{(k)}$

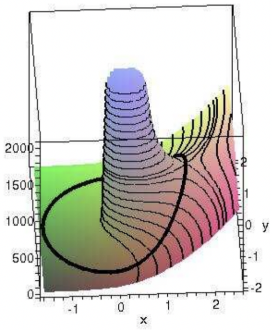


Illustration of saddle point method

Application of saddle point method

Saddle points suggest the contour choice of $\Gamma = \{z : |z| = r\}$ with $r = \sqrt{\frac{\ell}{n-\ell}}$:

$$|e_\ell(x)| = \left| \frac{1}{2\pi \mathrm{i}} \oint_\Gamma \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \frac{\mathrm{d}z}{z} \right| \leq \max_{|z|=r} \left| \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \right|$$

Application of saddle point method

Saddle points suggest the contour choice of $\Gamma = \{z : |z| = r\}$ with $r = \sqrt{\frac{\ell}{n-\ell}}$:

$$|e_\ell(x)| = \left| \frac{1}{2\pi \mathrm{i}} \oint_\Gamma \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \frac{\mathrm{d}z}{z} \right| \le \max_{|z|=r} \left| \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \right|$$

Use AM-GM:

$$\begin{aligned} \prod_{i=1}^n |1 + x_i z|^2 &= \prod_{i=1}^n (1 + 2\Re(x_i z) + |x_i|^2 r^2) \\ &\le \left(\frac{1}{n} \sum_{i=1}^n (1 + 2\Re(x_i z) + |x_i|^2 r^2) \right)^n = (1 + r^2)^n. \end{aligned}$$

This proves the inequality for the complex case.

Application of saddle point method

Saddle points suggest the contour choice of $\Gamma = \{z : |z| = r\}$ with $r = \sqrt{\frac{\ell}{n-\ell}}$:

$$|e_\ell(x)| = \left| \frac{1}{2\pi \mathrm{i}} \oint_\Gamma \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \frac{\mathrm{d}z}{z} \right| \le \max_{|z|=r} \left| \frac{\prod_{i=1}^n (1 + x_i z)}{z^\ell} \right|$$

Use AM-GM:

$$\begin{aligned} \prod_{i=1}^n |1 + x_i z|^2 &= \prod_{i=1}^n (1 + 2\Re(x_i z) + |x_i|^2 r^2) \\ &\le \left(\frac{1}{n} \sum_{i=1}^n (1 + 2\Re(x_i z) + |x_i|^2 r^2) \right)^n = (1 + r^2)^n. \end{aligned}$$

This proves the inequality for the complex case.

Real case: a more careful saddle point analysis for $x = x^{(k)}$.

Second upper bound via matrix permanent

An alternative view from matrix permanent

Drawbacks of the first upper bound:

- $\rightarrow$ meaningless when $\mathsf{C}_{\chi^2}(\mathcal{P}) \geq 1$
- $\rightarrow$ why loose: Banach's inequality may overlook the benefits from different rows

$\mathrm{Perm}(A) := \sum_{\pi \in S_n} \prod_{i=1}^n A_{i,\pi(i)}$, $\overline{P} := \frac{1}{n} \sum_{i=1}^n P_i$

An alternative view from matrix permanent

Drawbacks of the first upper bound:

→ meaningless when $C_{\chi^2}(\mathcal{P}) \geq 1$

→ why loose: Banach's inequality may overlook the benefits from different rows

An observation thanks to permutations:

χ^2 divergence as matrix permanents

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = \frac{n^n}{n!}\mathrm{Perm}(A) - 1,$$

where $A \in \mathbb{R}^{n\times n}$ is given by $A_{i,j} = \mathbb{E}_{\overline{P}}\left[\frac{\mathrm{d}P_i}{\mathrm{d}\overline{P}}\frac{\mathrm{d}P_j}{\mathrm{d}\overline{P}}\right]$.

$\mathrm{Perm}(A) := \sum_{\pi\in S_n}\prod_{i=1}^n A_{i,\pi(i)}$, $\overline{P} := \frac{1}{n}\sum_{i=1}^n P_i$

An alternative view from matrix permanent

Drawbacks of the first upper bound:

→ meaningless when $\mathsf{C}_{\chi^2}(\mathcal{P}) \geq 1$

→ why loose: Banach's inequality may overlook the benefits from different rows

An observation thanks to permutations:

χ^2 divergence as matrix permanents

$$\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = \frac{n^n}{n!}\mathrm{Perm}(A) - 1,$$

where $A \in \mathbb{R}^{n\times n}$ is given by $A_{i,j} = \mathbb{E}_{\overline{P}}\left[\frac{\mathrm{d}P_i}{\mathrm{d}\overline{P}}\frac{\mathrm{d}P_j}{\mathrm{d}\overline{P}}\right]$.

The famous van der Waerden conjecture (proven in 1980's) states that $\mathrm{Perm}(A) \geq \frac{n!}{n^n}$ for all doubly stochastic matrices, so showing $\chi^2(\mathbb{P}_n \| \mathbb{Q}_n) = O(1)$ essentially means that $\mathrm{Perm}(A)$ is nearly as small as possible

---

$\mathrm{Perm}(A) := \sum_{\pi\in S_n}\prod_{i=1}^n A_{i,\pi(i)}$, $\overline{P} := \frac{1}{n}\sum_{i=1}^n P_i$

Properties of matrix A

Properties of A

- $\rightarrow$ A is PSD and doubly stochastic;
- $\rightarrow$ $\mathrm{Tr}(A) \leq \mathsf{C}_{\chi^2}(\mathcal{P}) + 1$;
- $\rightarrow$ its spectral gap satisfies $1 - \lambda_2(A) \geq \frac{1}{\mathsf{D}_{\chi^2}(\mathcal{P})+1}$.

Properties of matrix A

Properties of A

- $\rightarrow$ A is PSD and doubly stochastic;
- $\rightarrow$ $\mathrm{Tr}(A) \leq \mathsf{C}_{\chi^2}(\mathcal{P}) + 1$;
- $\rightarrow$ its spectral gap satisfies $1 - \lambda_2(A) \geq \frac{1}{\mathsf{D}_{\chi^2}(\mathcal{P})+1}$.

Suggests to use the eigendecomposition $A = UDU^\top$ and expand

$$\frac{n^n}{n!}\mathrm{Perm}(UDU^\top) = \sum_{\ell=0}^{n} S_\ell(\lambda_2, \ldots, \lambda_n),$$

with homogeneous polynomials S_ℓ of total degree ℓ

Properties of A

→ A is PSD and doubly stochastic;

→ $\mathrm{Tr}(A) \leq \mathsf{C}_{\chi^2}(\mathcal{P}) + 1$;

→ its spectral gap satisfies $1 - \lambda_2(A) \geq \frac{1}{\mathsf{D}_{\chi^2}(\mathcal{P})+1}$.

Suggests to use the eigendecomposition $A = UDU^\top$ and expand

$$\frac{n^n}{n!}\mathrm{Perm}(UDU^\top) = \sum_{\ell=0}^{n} S_\ell(\lambda_2, \ldots, \lambda_n),$$

with homogeneous polynomials S_ℓ of total degree ℓ

Key idea: express S_ℓ using complex normal random variables

Expressing the sum $\sum_{\ell=0}^{n} S_\ell$

Complex normal random variable:

- $\rightarrow$ $z \sim \mathcal{CN}(0,1)$ iff $z = x + \mathrm{i}y$ with independent $x, y \sim \mathcal{N}(0, \frac{1}{2})$
- $\rightarrow$ moment condition: $\mathbb{E}[z^m \bar{z}^n] = n! \mathbb{1}_{m=n}$ for $z \sim \mathcal{CN}(0,1)$

Expressing the sum $\sum_{\ell=0}^{n} S_\ell$

Complex normal random variable:

- → $z \sim \mathcal{CN}(0,1)$ iff $z = x + \mathrm{i}y$ with independent $x, y \sim \mathcal{N}(0, \frac{1}{2})$
- → moment condition: $\mathbb{E}[z^m \bar{z}^n] = n!\mathbb{1}_{m=n}$ for $z \sim \mathcal{CN}(0,1)$

Fact I (observed in [Gurvit'03])

$$\sum_{\ell=0}^{n} S_\ell \propto \mathbb{E}\left[\prod_{i=1}^{n} \left|\left(UD^{1/2}z\right)_i\right|^2\right], \qquad z_1, \ldots, z_n \sim \mathcal{CN}(0,1).$$

Expressing the sum $\sum_{\ell=0}^{n} S_\ell$

Complex normal random variable:

- → $z \sim \mathcal{CN}(0,1)$ iff $z = x + \mathrm{i}y$ with independent $x, y \sim \mathcal{N}(0, \frac{1}{2})$
- → moment condition: $\mathbb{E}[z^m \bar{z}^n] = n! \mathbb{1}_{m=n}$ for $z \sim \mathcal{CN}(0,1)$

Fact I (observed in [Gurvit'03])

$$\sum_{\ell=0}^{n} S_\ell \propto \mathbb{E}\left[\prod_{i=1}^{n} \left|\left(UD^{1/2}z\right)_i\right|^2\right], \qquad z_1, \ldots, z_n \sim \mathcal{CN}(0,1).$$

Applying AM-GM to the product gives

$$\sum_{\ell=0}^{n} S_\ell \leq \sum_{\ell_2 + \cdots + \ell_n \leq n} \lambda_2^{\ell_2} \cdots \lambda_n^{\ell_n} \leq \prod_{i=2}^{n} \frac{1}{1 - \lambda_i}$$

- → the trace and spectral gap properties lead to the second upper bound

Expressing the individual term S_ℓ

Fact II

$$S_\ell \propto \mathbb{E}\left[\left|e_\ell\left((\widetilde{U}\widetilde{D}^{1/2}z)_1, \dots, (\widetilde{U}\widetilde{D}^{1/2}z)_n\right)\right|^2\right], \qquad z_1, \dots, z_{n-1} \sim \mathcal{CN}(0,1),$$

where $(\widetilde{U}, \widetilde{D})$ takes out the leading eigenvector/eigenvalue in (U, D).

---

recall that $e_\ell(x_1, \dots, x_n) = \sum_{|S|=\ell} \prod_{i \in S} x_i$.

Expressing the individual term S_ℓ

Fact II

$$S_\ell \propto \mathbb{E}\left[\left|e_\ell\left((\widetilde{U}\widetilde{D}^{1/2}z)_1,\dots,(\widetilde{U}\widetilde{D}^{1/2}z)_n\right)\right|^2\right], \qquad z_1,\dots,z_{n-1}\sim\mathcal{CN}(0,1),$$

where $(\widetilde{U},\widetilde{D})$ takes out the leading eigenvector/eigenvalue in (U,D).

→ can show that the vector $\widetilde{U}\widetilde{D}^{1/2}z$ sums into zero

→ using our key inequality eventually leads to

$$S_\ell \le 3\sqrt{\ell+1}\sum_{\ell_2+\cdots+\ell_n=\ell}\lambda_2^{\ell_2}\cdots\lambda_n^{\ell_n}$$

→ useful in empirical Bayes applications

recall that $e_\ell(x_1,\dots,x_n)=\sum_{|S|=\ell}\prod_{i\in S}x_i$.

Concluding remarks

Take home messages:

→ permutations induce weak dependency, quantitatively

→ centered basis is preferred in the method of "moments"

→ lifting to complex domains makes a theorist's life less complex

Concluding remarks

Take home messages:

- $\rightarrow$ permutations induce weak dependency, quantitatively
- $\rightarrow$ centered basis is preferred in the method of "moments"
- $\rightarrow$ lifting to complex domains makes a theorist's life less complex

Further questions:

- $\rightarrow$ remove the $\mathcal{O}(\sqrt{\ell})$ factor for centered complex vectors?
- $\rightarrow$ for bounded Gaussian case, improve the χ^2 upper bound $\exp(O(\mu^3))$ to $\exp(O(\mu^2))$?
- $\rightarrow$ method of "moments" for two high-dimensional mixtures?

Concluding remarks

Take home messages:

- $\rightarrow$ permutations induce weak dependency, quantitatively
- $\rightarrow$ centered basis is preferred in the method of "moments"
- $\rightarrow$ lifting to complex domains makes a theorist's life less complex

Further questions:

- $\rightarrow$ remove the $\mathcal{O}(\sqrt{\ell})$ factor for centered complex vectors?
- $\rightarrow$ for bounded Gaussian case, improve the χ^2 upper bound $\exp(O(\mu^3))$ to $\exp(O(\mu^2))$?
- $\rightarrow$ method of "moments" for two high-dimensional mixtures?

Thank You!
arXiv: 2408.09341