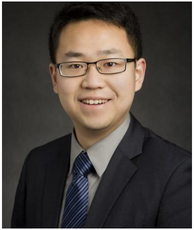
Permutation Mixtures and Empirical Bayes

Yanjun Han (NYU Courant Math and CDS)

Joint work with:

Jonathan Niles-Weed
(NYU)

Yandi Shen
(CMU)



Yihong Wu
(Yale)

MILD Seminar, UBC ECE
December 13, 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 $\pi \sim \text{Unif}(S_n)$;
- $\rightarrow$ $\mathbb{P}_n$ is the joint distribution of $(X_1, \ldots, X_n)$ with $X_i = Z_{\pi(i)}$;
- $\rightarrow$ in mathematical terms:

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

Setup

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

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

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

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

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 random permutations
- $\rightarrow$ de Finetti-style theorems
- $\rightarrow$ compound decisions in empirical Bayes

Motivation

Later in the talk:

→ statistical and IT applications involving random permutations

→ de Finetti-style theorems

→ compound decisions in empirical Bayes

Bigger picture:

→ general mean-field approximation

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

- $\rightarrow$ $\mathbb{P}_n = \nu_{\mathbb{P}} \star \mathcal{N}(0, I_n)$, 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;

Our result

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

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

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

Failure of 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, ...]

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

Failure of 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

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

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

Failure of 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

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

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

Failure of method of moments (cont'd)

In general dimensions:

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

→ $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)$ 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

Failure of 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

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 $\ell \geq 2$

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

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

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

Failure of method of cumulants

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

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

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

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

- $\rightarrow$ a better behavior for certain cross terms
- $\rightarrow$ 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

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

Main result

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: $\mathrm{D}_{\chi^2}(\mathcal{P}) = \sup_{P_1, P_2 \in \mathcal{P}} \chi^2(P_1 \| P_2)$

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

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 (H., Niles-Weed'24)

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

→ $\mathbb{P}_n$ 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 (H., Niles-Weed'24)

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

Examples

Theorem (H., Niles-Weed'24)

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

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 (H., Niles-Weed'24)

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

Example I (Two-component Gaussian)

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

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

Example II (Bounded Gaussian)

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

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

Applications

Permutation prior

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

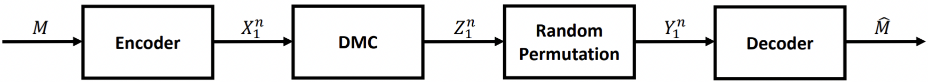
- $\rightarrow$ statisticians would like to prove lower bounds on the estimation error of θ
- $\rightarrow$ a prevalent strategy is to impose a prior distribution on θ, and a permutation prior is sometimes preferred: $\theta = (v_{\pi(1)}, \ldots, v_{\pi(n)})$ for a known vector v and a random permutation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

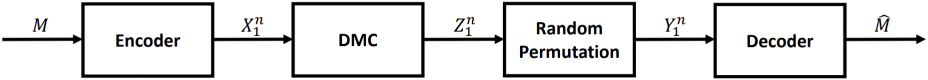
$$I_{\mathbb{Q}_n}(\theta; X^n) - O_{\mathcal{P}}(1) \le I_{\mathbb{P}_n}(\theta; X^n) \le I_{\mathbb{Q}_n}(\theta; X^n)$$

Permutation channel



The noisy permutation channel introduced in [Makur'20]

Permutation channel



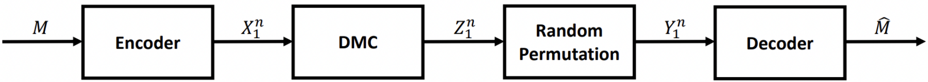
The noisy permutation channel introduced in [Makur'20]

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

$\rightarrow$ 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 $(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$

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 $(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$

Using the 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 the 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 $k \leq n$:

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

Using the 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 $k \leq n$:

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

Sketch of the first upper bound

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 $\mathbb{Q}_n$

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

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 $\mathbb{Q}_n$

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

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 $\mathcal{N}(0,1)$.

$\{H_0(x), H_1(x), \ldots\}$ 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 $\theta \sim \mathrm{Unif}(\{\pm\mu\})$

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

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 $L^2(\varphi)$

$\mathbb{E}[\theta^k]$ 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 $\theta \sim \mathrm{Unif}(\{\pm\mu\})$

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

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:

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

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

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

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

$\rightarrow$ 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 $|S| = \ell$,

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

$\rightarrow$ piecing everything together:

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

Importance of zero-mean: a Maclaurin-type inequality

For a vector $x = (x_1, \ldots, x_n)$, define the elementary symmetric polynomial

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

Importance of zero-mean: a Maclaurin-type inequality

For a vector $x = (x_1, \ldots, x_n)$, define the elementary symmetric polynomial

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

Theorem (Upper bound on ESPs for centered vector)

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

- $\to$ If $x \in \mathbb{R}^n$, then $|e_\ell(x)|^2 \le 10\binom{n}{\ell}$;
- $\to$ If $x \in \mathbb{C}^n$, a weaker upper bound holds:

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

Importance of zero-mean: a Maclaurin-type inequality

For a vector $x = (x_1, \dots, x_n)$, define the elementary symmetric polynomial

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

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 $|e_\ell(x)|^2 \le 10\binom{n}{\ell}$;
- → If $x \in \mathbb{C}^n$, a weaker upper bound holds:

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

- → 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]:

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

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

Proof of the 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:

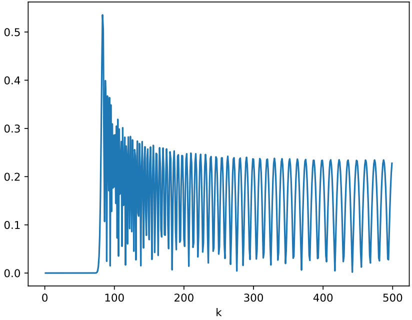
$$x^{(k)} = \Big(\underbrace{\sqrt{\frac{k}{n-k}}, \dots, \sqrt{\frac{k}{n-k}}}_{n-k \text{ copies}}, \underbrace{-\sqrt{\frac{n-k}{k}}, \dots, -\sqrt{\frac{n-k}{k}}}_{k \text{ copies}} \Big)$$

Proof of the 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:

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

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



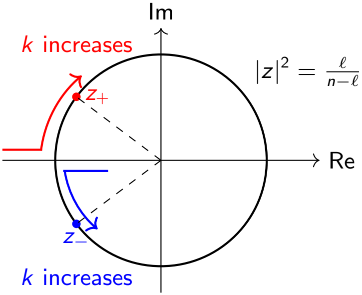
The quantity $|e_\ell(x^{(k)})|^2 / \binom{n}{\ell}$ vs. k for $n = 1000, \ell = 300$.

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

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

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

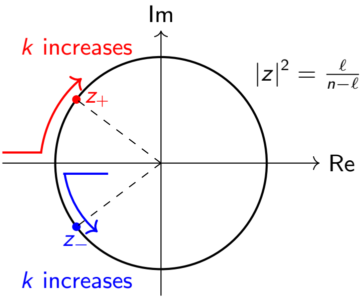
$$\text{Saddle point equation :} \quad \frac{\ell}{z} = \sum_{i=1}^n \frac{x_i}{1 + x_i z}$$



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

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

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



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

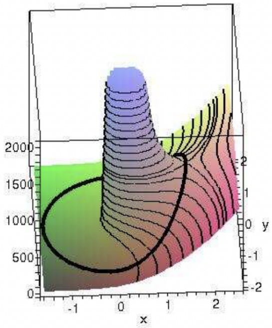


Illustration of saddle point method

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 $r = \sqrt{\frac{\ell}{n-\ell}}$:

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

Use AM-GM:

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

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 $r = \sqrt{\frac{\ell}{n-\ell}}$:

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

Use AM-GM:

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

This proves the inequality for the complex case.

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

Compound decisions in empirical Bayes

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]:

→ 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]:

→ compound decision 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:

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

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]:

- → compound decision setting: independent $X_i \sim P_{\theta_i}$, aim to estimate $\theta = (\theta_1, \dots, \theta_n)$
- → 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:

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

Question

Do these oracles have similar estimation power?

The Gaussian case

- $\rightarrow$ observation vector: $X^n \sim \mathcal{N}(\theta^n, I_n)$
- $\rightarrow$ the oracle only knows the multiset $\{\theta_1, \ldots, \theta_n\}$ but not the order
- $\rightarrow$ equivalently, θ^n follows a permutation prior on a given multiset
- $\rightarrow$ under the quadratic loss:

$$\widehat{\theta}_i^{\mathrm{S}} = \mathbb{E}[\theta_i \mid X_i], \qquad \widehat{\theta}_i^{\mathrm{PI}} = \mathbb{E}[\theta_i \mid X^n].$$

The Gaussian case

- $\rightarrow$ observation vector: $X^n \sim \mathcal{N}(\theta^n, I_n)$
- $\rightarrow$ the oracle only knows the multiset $\{\theta_1, \ldots, \theta_n\}$ but not the order
- $\rightarrow$ equivalently, θ^n follows a permutation prior on a given multiset
- $\rightarrow$ under the quadratic loss:

$$\widehat{\theta}_i^{\mathrm{S}} = \mathbb{E}[\theta_i \mid X_i], \qquad \widehat{\theta}_i^{\mathrm{PI}} = \mathbb{E}[\theta_i \mid X^n].$$

Greenshtein and Ritov (2009)

If $|\theta_i| \leq \mu$ for all $i \in [n]$ and $\mu \geq 1$,

$$\mathbb{E}\left[\|\widehat{\theta}^{\mathrm{S}} - \widehat{\theta}^{\mathrm{PI}}\|^2\right] = e^{O(\mu^2)}.$$

- $\rightarrow$ an $O(1)$ upper bound even if the vectors are n-dimensional
- $\rightarrow$ can the dependence on μ be improved for large μ?

A tight upper bound

Theorem ([H., Niles-Weed, Shen, Wu, 24+])

If $|\theta_i| \le \mu$ for all $i \in [n]$ and $\mu \ge 1$,

$$\mathbb{E}\left[\|\widehat{\theta}^{\mathrm{S}} - \widehat{\theta}^{\mathrm{PI}}\|^2\right] = O(\mu \log^2 n).$$

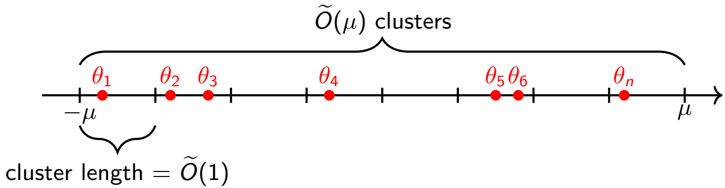
A tight upper bound

Theorem ([H., Niles-Weed, Shen, Wu, 24+])

If $|\theta_i| \leq \mu$ for all $i \in [n]$ and $\mu \geq 1$,

$$\mathbb{E}\left[\|\widehat{\theta}^{\mathrm{S}} - \widehat{\theta}^{\mathrm{PI}}\|^2\right] = O(\mu \log^2 n).$$

The quantity μ represents the number of "clusters":



$\rightarrow$ by the concentration of Gaussian, each cluster roughly corresponds to an independent subproblem

$\rightarrow$ linear dependence on the # of clusters is tight by adding up all clusters

Application: Competitive Distribution Estimation

Competitive distribution estimation

A Poisson sequence model:

$$(N_1, \ldots, N_k) \sim \mathrm{Poi}(np_1) \otimes \cdots \otimes \mathrm{Poi}(np_k)$$

→ n: sample size

→ k: support size

→ $p = (p_1, \ldots, p_k)$: an unknown probability vector

Competitive distribution estimation

A Poisson sequence model:

$$(N_1, \ldots, N_k) \sim \mathsf{Poi}(np_1) \otimes \cdots \otimes \mathsf{Poi}(np_k)$$

- → n: sample size
- → k: support size
- → $p = (p_1, \ldots, p_k)$: an unknown probability vector

Target of competitive distribution estimation: based on observed counts $(N_1, \ldots, N_k)$, devise an estimator $\widehat{p}$ to minimize the expected regret:

$$\mathrm{regret}(\widehat{p}; p) = \mathbb{E}\left[\mathrm{KL}(p \| \widehat{p}) - \mathrm{KL}(p \| \widehat{p}^{\mathrm{PI}})\right],$$

where

$$\widehat{p}^{\mathsf{PI}} = \operatorname*{argmin}_{\widehat{q}} \max_{p' : \{p'\} = \{p\}} \mathbb{E}_{p'}[\mathsf{KL}(p' \| \widehat{q})]$$

is the best permutation-invariant decision rule

Upper bound ([Orlitsky and Suresh'15])

A modified Good–Turing estimator $\widehat{p}^{\mathrm{MGT}}$ achieves

$$\sup_p \mathrm{regret}(\widehat{p}^{\mathrm{MGT}}; p) = \widetilde{O}\left(\min\left\{\frac{k}{n}, \frac{1}{\sqrt{n}}\right\}\right).$$

“Why is Good–Turing Good”

Upper bound ([Orlitsky and Suresh’15])

A modified Good–Turing estimator $\widehat{p}^{\mathrm{MGT}}$ achieves

$$\sup_p \mathrm{regret}(\widehat{p}^{\mathrm{MGT}}; p) = \widetilde{O}\left(\min\left\{\frac{k}{n}, \frac{1}{\sqrt{n}}\right\}\right).$$

The Good–Turing estimator $\widehat{p}^{\mathrm{GT}}$ [Good’53]: for $N_i = y$,

$$\widehat{p}_i^{\mathrm{GT}} = \frac{y+1}{n} \cdot \frac{\sum_{j=1}^k 1(N_j = y+1)}{\sum_{j=1}^k 1(N_j = y)}$$

"Why is Good–Turing Good"

Upper bound ([Orlitsky and Suresh'15])

A modified Good–Turing estimator $\widehat{p}^{\mathrm{MGT}}$ achieves

$$\sup_p \mathrm{regret}(\widehat{p}^{\mathrm{MGT}}; p) = \widetilde{O}\left(\min\left\{\frac{k}{n}, \frac{1}{\sqrt{n}}\right\}\right).$$

The Good–Turing estimator $\widehat{p}^{\mathrm{GT}}$ [Good'53]: for $N_i = y$,

$$\widehat{p}_i^{\mathrm{GT}} = \frac{y+1}{n} \cdot \frac{\sum_{j=1}^k 1(N_j = y+1)}{\sum_{j=1}^k 1(N_j = y)}$$

Lower bound ([Orlitsky and Suresh'15])

$$\inf_{\widehat{p}} \sup_p \mathrm{regret}(\widehat{p}; p) = \Omega\left(\min\left\{\frac{k}{n}, \frac{1}{n^{2/3}}\right\}\right).$$

Better Good–Turing: NPMLE

A different idea based on so-called "g-modeling" [Efron'14]:

→ think of $p_1, \ldots, p_k \overset{\text{i.i.d.}}{\sim} G^\star$, with the empirical measure $G^\star = \frac{1}{k}\sum_{i=1}^k \delta_{p_i}$

→ a natural estimator is the nonparametric MLE (NPMLE):

$$\widehat{G} = \operatorname*{argmax}_{G:\ \mathbb{E}_G[p] \le \frac{1}{k}} \sum_{i=1}^{k} \log \mathbb{E}_G\left[\mathbb{P}(\mathrm{Poi}(np) = N_i)\right]$$

→ the final estimator $\widehat{p}^{\mathrm{NPMLE}}$ mimics the separable oracle:

$$\widehat{p}^{\mathrm{NPMLE}} = \text{normalized version of } (\mathbb{E}_{\widehat{G}}[p_1 \mid N_1], \ldots, \mathbb{E}_{\widehat{G}}[p_k \mid N_k])$$

Better Good–Turing: NPMLE

A different idea based on so-called "g-modeling" [Efron'14]:

$\rightarrow$ think of $p_1, \ldots, p_k \overset{\text{i.i.d.}}{\sim} G^\star$, with the empirical measure $G^\star = \frac{1}{k}\sum_{i=1}^k \delta_{p_i}$

$\rightarrow$ a natural estimator is the nonparametric MLE (NPMLE):

$$\widehat{G} = \underset{G:\ \mathbb{E}_G[p] \le \frac{1}{k}}{\text{argmax}} \sum_{i=1}^k \log \mathbb{E}_G \left[\mathbb{P}(\mathrm{Poi}(np) = N_i)\right]$$

$\rightarrow$ the final estimator $\widehat{p}^{\mathrm{NPMLE}}$ mimics the separable oracle:

$$\widehat{p}^{\mathrm{NPMLE}} = \text{normalized version of } (\mathbb{E}_{\widehat{G}}[p_1 \mid N_1], \ldots, \mathbb{E}_{\widehat{G}}[p_k \mid N_k])$$

Theorem (H., Niles-Weed, Shen, Wu, 24+)

The above estimator $\widehat{p}^{\mathrm{NPMLE}}$ achieves

$$\sup_p \mathrm{regret}(\widehat{p}^{\mathrm{NPMLE}}; p) = \widetilde{O}\left(\min\left\{\frac{k}{n}, \frac{1}{n^{2/3}}\right\}\right).$$

Regret analysis

Part I of regret: $\widehat{p}^{\mathrm{NPMLE}}$ against the separable oracle

$$\widehat{p}^{\mathrm{S}} = \text{normalized version of } (\mathbb{E}_{G^\star}[p_1 \mid N_1], \ldots, \mathbb{E}_{G^\star}[p_k \mid N_k])$$

$\rightarrow$ use the theory of NPMLE to argue that $\mathbb{E}_{\widehat{G}}[p_i \mid N_i] \approx \mathbb{E}_{G^\star}[p_i \mid N_i]$

Part I of regret: $\widehat{p}^{\mathrm{NPMLE}}$ against the separable oracle

$$\widehat{p}^{\mathrm{S}} = \text{normalized version of } (\mathbb{E}_{G^\star}[p_1 \mid N_1], \ldots, \mathbb{E}_{G^\star}[p_k \mid N_k])$$

→ use the theory of NPMLE to argue that $\mathbb{E}_{\widehat{G}}[p_i \mid N_i] \approx \mathbb{E}_{G^\star}[p_i \mid N_i]$

Part II of regret: separable oracle $\widehat{p}^{\mathrm{S}}$ against the PI oracle $\widehat{p}^{\mathrm{PI}}$

→ our EB upper bound in the Poisson case gives

$$(\star) = \mathbb{E}\left[\mathrm{KL}\left(\widehat{p}^{\mathrm{S}} \| \widehat{p}^{\mathrm{PI}}\right)\right] = \frac{\widetilde{O}\,(\#\text{ of clusters in the Poisson model})}{n}$$

→ it turns out that

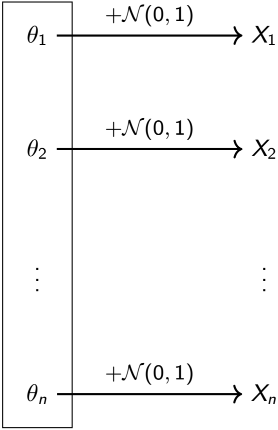
$$\#\text{ of clusters in the Poisson model} = O\left(\min\left\{k, n^{1/3}\right\}\right)$$

Proof for the Gaussian case

An information-theoretic argument

→ Recall that

$$\widehat{\theta}_1^{\mathrm{S}} = \mathbb{E}[\theta_1 \mid X_1], \quad \widehat{\theta}_1^{\mathrm{PI}} = \mathbb{E}[\theta_1 \mid X^n].$$



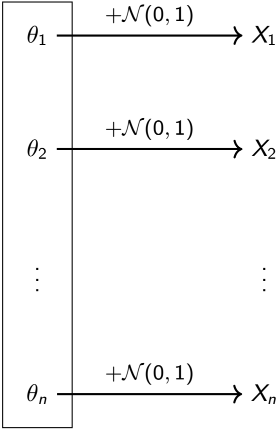
An information-theoretic argument

→ Recall that

$$\widehat{\theta}_1^{\mathrm{S}} = \mathbb{E}[\theta_1 \mid X_1], \quad \widehat{\theta}_1^{\mathrm{PI}} = \mathbb{E}[\theta_1 \mid X^n].$$

→ Tao's inequality:

$$\mathbb{E}\left[(\mathbb{E}[\theta_1 \mid X_1] - \mathbb{E}[\theta_1 \mid X^n])^2\right] = \widetilde{O}(1) \cdot I(\theta_1; X_2^n \mid X_1).$$



An information-theoretic argument

→ Recall that

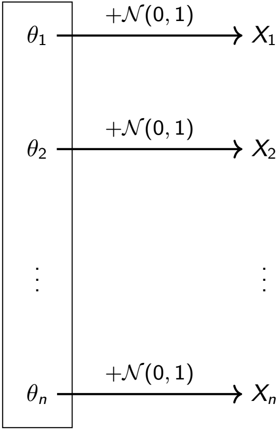
$$\widehat{\theta}_1^{\mathrm{S}} = \mathbb{E}[\theta_1 \mid X_1], \quad \widehat{\theta}_1^{\mathrm{PI}} = \mathbb{E}[\theta_1 \mid X^n].$$

→ Tao's inequality:

$$\mathbb{E}\left[(\mathbb{E}[\theta_1 \mid X_1] - \mathbb{E}[\theta_1 \mid X^n])^2\right] = \widetilde{O}(1) \cdot I(\theta_1; X_2^n \mid X_1).$$

→ A "model-free" upper bound:

$$\begin{aligned} I(\theta_1; X_2^n \mid X_1) &= H(\theta_1 \mid X_1) - H(\theta_1 \mid X^n) \\ &\le H(\theta_1 \mid X_1) - \frac{1}{n} H(\theta^n \mid X^n) \\ &= H(\theta_1) - \frac{H(\theta^n)}{n} - \underbrace{\left(I(\theta_1; X_1) - \frac{I(\theta^n; X^n)}{n}\right)}_{\ge 0 \text{ as } P_{X^n \mid \theta^n} = \prod_i P_{X_i \mid \theta_i}} \\ &\le H(\theta_1) - \frac{H(\theta^n)}{n} = \frac{1}{n} \mathrm{KL}(P_{\theta^n} \| \prod_i P_{\theta_i}) \\ &= \widetilde{O}\left(\frac{|\mathrm{supp}(\{\theta_1, \ldots, \theta_n\})|}{n}\right) \end{aligned}$$



Improvement via "noisy" θ^n

$\rightarrow$ new idea: add a noisy Z_i between θ_i and X_i

$$\begin{tikzcd}
\theta_1 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & Z_1 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & X_1 \\
\theta_2 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & Z_2 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & X_2 \\
\vdots & \vdots & \vdots \\
\theta_n \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & Z_n \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & X_n
\end{tikzcd}$$

Improvement via “noisy” θ^n

→ new idea: add a noisy Z_i between θ_i and X_i

→ key identity:

$$\begin{aligned}
&\mathbb{E}[\theta_1 \mid X_1] - \mathbb{E}[\theta_1 \mid X^n] \\
&= 2\left(\mathbb{E}[Z_1 \mid X_1] - \mathbb{E}[Z_1 \mid X^n]\right)
\end{aligned}$$

$$\begin{tikzcd}
\theta_1 \arrow[r, "+\mathcal{N}(0, \frac{1}{2})"] & Z_1 \arrow[r, "+\mathcal{N}(0, \frac{1}{2})"] & X_1 \\
\theta_2 \arrow[r, "+\mathcal{N}(0, \frac{1}{2})"] & Z_2 \arrow[r, "+\mathcal{N}(0, \frac{1}{2})"] & X_2 \\
\vdots & \vdots & \vdots \\
\theta_n \arrow[r, "+\mathcal{N}(0, \frac{1}{2})"] & Z_n \arrow[r, "+\mathcal{N}(0, \frac{1}{2})"] & X_n
\end{tikzcd}$$

Improvement via "noisy" θ^n

→ new idea: add a noisy Z_i between θ_i and X_i

→ key identity:

$$\begin{aligned}&\mathbb{E}[\theta_1 \mid X_1] - \mathbb{E}[\theta_1 \mid X^n] \\ &= 2\left(\mathbb{E}[Z_1 \mid X_1] - \mathbb{E}[Z_1 \mid X^n]\right)\end{aligned}$$

→ the previous "model-free" bound now gives

$$\mathbb{E}\left[\left(\mathbb{E}[\theta_1 \mid X_1] - \mathbb{E}[\theta_1 \mid X^n]\right)^2\right] \lesssim \frac{1}{n}\mathrm{KL}(P_{Z^n} \| \prod_i P_{Z_i})$$

$$\begin{tikzcd}
\theta_1 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & Z_1 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & X_1 \\
\theta_2 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & Z_2 \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & X_2 \\
\vdots & \vdots & \vdots \\
\theta_n \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & Z_n \arrow[r, "+\mathcal{N}(0,\frac{1}{2})"] & X_n
\end{tikzcd}$$

Improvement via "noisy" θ^n

- $\to$ new idea: add a noisy Z_i between θ_i and X_i
- $\to$ key identity:

$$\begin{aligned}&\mathbb{E}[\theta_1 \mid X_1] - \mathbb{E}[\theta_1 \mid X^n] \\ &= 2\left(\mathbb{E}[Z_1 \mid X_1] - \mathbb{E}[Z_1 \mid X^n]\right)\end{aligned}$$

- $\to$ the previous "model-free" bound now gives

$$\mathbb{E}\left[\left(\mathbb{E}[\theta_1 \mid X_1] - \mathbb{E}[\theta_1 \mid X^n]\right)^2\right] \lesssim \frac{1}{n}\mathrm{KL}(P_{Z^n} \| \prod_i P_{Z_i})$$

- $\to$ the final quantity $\mathrm{KL}(P_{Z^n} \| \prod_i P_{Z_i})$ is now between a Gaussian permutation mixture and its i.i.d. approximation!

$$\begin{aligned}
\theta_1 &\xrightarrow{+\mathcal{N}(0,\frac{1}{2})} Z_1 \xrightarrow{+\mathcal{N}(0,\frac{1}{2})} X_1 \\
\theta_2 &\xrightarrow{+\mathcal{N}(0,\frac{1}{2})} Z_2 \xrightarrow{+\mathcal{N}(0,\frac{1}{2})} X_2 \\
&\quad\vdots \qquad\qquad\quad \vdots \qquad\qquad\quad \vdots \\
\theta_n &\xrightarrow{+\mathcal{N}(0,\frac{1}{2})} Z_n \xrightarrow{+\mathcal{N}(0,\frac{1}{2})} X_n
\end{aligned}$$

Concluding remarks

Take home messages:

$\rightarrow$ permutations induce weak dependency, quantitatively

$\rightarrow$ centered basis is preferred in the method of “moments”

$\rightarrow$ NPMLE outperforms Good–Turing in empirical Bayes

Concluding remarks

Take home messages:

- $\rightarrow$ permutations induce weak dependency, quantitatively
- $\rightarrow$ centered basis is preferred in the method of "moments"
- $\rightarrow$ NPMLE outperforms Good–Turing in empirical Bayes

Further questions:

- $\rightarrow$ 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?
- $\rightarrow$ a better understanding of the noisy Z? stochastic localization?

Concluding remarks

Take home messages:

$\rightarrow$ permutations induce weak dependency, quantitatively

$\rightarrow$ centered basis is preferred in the method of "moments"

$\rightarrow$ NPMLE outperforms Good–Turing in empirical Bayes

Further questions:

$\rightarrow$ 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?

$\rightarrow$ a better understanding of the noisy Z? stochastic localization?

Thank You!