Optimal Permutation estimation in crowdsourcing problems

Alexandra Carpentier, **Emmanuel Pilliat**, and **Nicolas Verzelen**

Université Potsdam, Université de Montpellier, and INRAE

GESDA - October 6th

Ranking Problems

Crowdsourcing Problems = Aggregation of Experts' opinion

$\rightsquigarrow$ To calibrate the method : need to evaluate the reliability of the experts

Ranking Problems

Crowdsourcing Problems = Aggregation of Experts' opinion

⇝ To calibrate the method : need to evaluate the reliability of the experts

Question / Expert								
True answer	Edible	Toxic	Toxic	Edible	Edible	Edible	Toxic	Edible
Bob	Toxic 0	Toxic 1	Edible 0	Toxic 0	Edible 1	Toxic 0	Toxic 1	Edible 1
Alice	Edible 1	Edible 0	Toxic 1	Edible 1	Edible 1	Edible 1	Toxic 1	Edible 1

1 : Correct answer 0 : Wrong answer

Our Goal

Ranking $\mathbf{n}$ experts according to their ability on $\mathbf{d}$ questions

Statistical Model

n experts and d questions

Observation Model

$Y = M + E \quad \in \mathbb{R}^{n \times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0, 1]$ for all i, k

Statistical Model

n experts and d questions

Observation Model

$Y = M + E \qquad \in \mathbb{R}^{n \times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0,1]$ for all i, k

Parametric Models for M :

- *Questions equally difficult* $\rightsquigarrow M_{ij} = a_i \qquad \approx$ [Dawid and Skene, 1979]

Statistical Model

n experts and d questions

Observation Model

$Y = M + E \qquad \in \mathbb{R}^{n\times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0,1]$ for all i, k

Parametric Models for M :

- *Questions equally difficult* $\rightsquigarrow M_{ij} = a_i \qquad \approx$ [Dawid and Skene, 1979]
- *Ability/difficulty* $\rightsquigarrow M_{ij} = \phi(\alpha_i - \beta_j) \qquad \approx$ [Bradley and Terry, 1952]

Statistical Model

n experts and d questions

Observation Model

$Y = M + E \quad \in \mathbb{R}^{n\times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0,1]$ for all i, k

Parametric Models for M :

- *Questions equally difficult* $\rightsquigarrow M_{ij} = a_i$ ≈ [Dawid and Skene, 1979]
- *Ability/difficulty* $\rightsquigarrow M_{ij} = \phi(\alpha_i - \beta_j)$ ≈ [Bradley and Terry, 1952]

Non-Parametric Models for M ≈ [Mao et al., 2018]

- Increasing columns **up to permutation π^* of rows** : $M_{\pi^{*-1}(i),k} \leq M_{\pi^{*-1}(i+1),k}$

Statistical Model

n experts and d questions

Observation Model

$Y = M + E \quad \in \mathbb{R}^{n\times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0,1]$ for all i, k

Parametric Models for M :

- *Questions equally difficult* $\rightsquigarrow M_{ij} = a_i$ $\approx$ [Dawid and Skene, 1979]
- *Ability/difficulty* $\rightsquigarrow M_{ij} = \phi(\alpha_i - \beta_j)$ $\approx$ [Bradley and Terry, 1952]

Non-Parametric Models for M $\approx$ [Mao et al., 2018]

- Increasing columns **up to permutation** π^* **of rows** : $M_{\pi^{*-1}(i),k} \leq M_{\pi^{*-1}(i+1),k}$
- Rows are increasing : $M_{i,k} \leq M_{i,k+1}$

Statistical Model

n experts and d questions

Observation Model

$Y = M + E \quad \in \mathbb{R}^{n \times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0,1]$ for all i, k

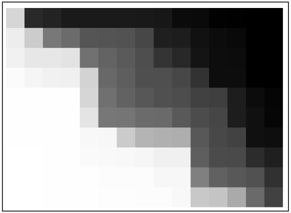
Non-Parametric Models for M $\approx$ [Mao et al., 2018]

- Increasing columns **up to permutation** π^* **of rows** : $M_{\pi^{*-1}(i),k} \leq M_{\pi^{*-1}(i+1),k}$
- Rows are increasing : $M_{i,k} \leq M_{i,k+1}$

Aim

Estimation of π^*.

Partial observation of Y discussed later.



Statistical Model

n experts and d questions

Observation Model

$Y = M + E \quad \in \mathbb{R}^{n \times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0,1]$ for all i, k

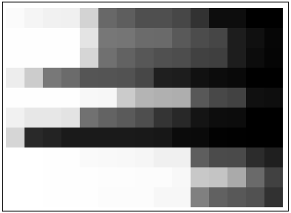
Non-Parametric Models for M $\approx$ [Mao et al., 2018]

- Increasing columns **up to permutation π^* of rows** : $M_{\pi^{*-1}(i),k} \leq M_{\pi^{*-1}(i+1),k}$
- Rows are increasing : $M_{i,k} \leq M_{i,k+1}$

Aim

Estimation of π^*.

Partial observation of Y discussed later.



Statistical Model

n experts and d questions

Observation Model

$Y = M + E \quad \in \mathbb{R}^{n\times d}$

- $(E_{i,k})$ independent and Subgaussian (e.g. Bernoulli)
- $M_{i,k} \in [0,1]$ for all i, k

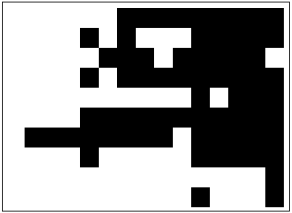
Non-Parametric Models for M $\approx$ [Mao et al., 2018]

- Increasing columns **up to permutation** π^* **of rows** : $M_{\pi^{*-1}(i),k} \leq M_{\pi^{*-1}(i+1),k}$
- Rows are increasing : $M_{i,k} \leq M_{i,k+1}$

Aim

Estimation of π^*.

Partial observation of Y discussed later.



Loss functions

Permutation loss for $\hat{\pi}$

$$l(\hat{\pi}, \pi^*) := \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 = \sum_{i=1}^{n} \sum_{k=1}^{d} (M_{\pi^{-1}(i),k} - M_{\pi^{*-1}(i),k})^2$$

Loss functions

Permutation loss for $\hat{\pi}$

$$l(\hat{\pi}, \pi^*) := \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 = \sum_{i=1}^{n} \sum_{k=1}^{d} (M_{\pi^{-1}(i),k} - M_{\pi^{*-1}(i),k})^2$$

Estimation loss for $\hat{M}$

$$\|\hat{M} - M\|_F^2.$$

Loss functions

Permutation loss for $\hat{\pi}$

$$l(\hat{\pi}, \pi^*) := \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 = \sum_{i=1}^{n}\sum_{k=1}^{d}(M_{\pi^{-1}(i),k} - M_{\pi^{*-1}(i),k})^2$$

Estimation loss for $\hat{M}$

$$\|\hat{M} - M\|_F^2.$$

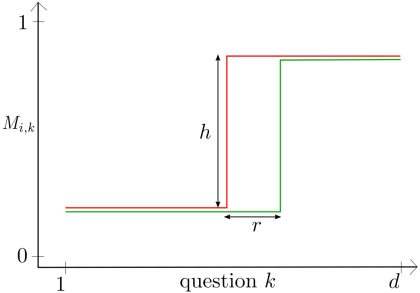
Remark :

- Estimation of π^* is "less demanding" than estimation of M.
- Estimating a bi-isotonic matrix computationally simple.

Interpretation of permutation Loss

Permutation loss for $\hat{\pi}$

$$l(\widehat{\pi}, \pi^*) := \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2.$$



If green and red misclassified : Perm-Loss = $2rh^2$.

Minimax Risk

$$\mathcal{R}^*_{perm}[n,d] = \inf_{\hat{\pi}} \sup_{M,\pi^*} \mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2]$$

$$\mathcal{R}^*_{est}[n,d] = \inf_{\hat{M}} \sup_{M,\pi^*} \mathbb{E}[\|\hat{M} - M\|_F^2]$$

Minimax Risk

$$\mathcal{R}^*_{perm}[n,d] = \inf_{\hat{\pi}} \sup_{M,\pi^*} \mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2]$$

$$\mathcal{R}^*_{est}[n,d] = \inf_{\hat{M}} \sup_{M,\pi^*} \mathbb{E}[\|\hat{M} - M\|_F^2]$$

Recovering π^* is easier then estimating M

$$\mathcal{R}^*_{perm}[n,d] \lesssim \mathcal{R}^*_{est}[n,d]$$

Related Rectangular Problems :

- **Two permutations** [Mao et al., 2018, Shah et al., 2019] :
  M is bi-isotonic up to permutations π_1^* and π_2^* of rows and columns.
  Objective : ranking the experts and the questions.

Related Rectangular Problems :

- **Two permutations** [Mao et al., 2018, Shah et al., 2019] :
  M is bi-isotonic up to permutations π_1^* and π_2^* of rows and columns.
  Objective : ranking the experts and the questions.
- **Column isotony** [Flammarion et al., 2019]

Other ranking and permutations problems

Related Rectangular Problems :

- **Two permutations** [Mao et al., 2018, Shah et al., 2019] :
  M is bi-isotonic up to permutations π_1^* and π_2^* of rows and columns.
  Objective : ranking the experts and the questions.
- **Column isotony** [Flammarion et al., 2019]

Ranking Players in a tournament : M is a $n \times n$ matrix with symmetries.

- **Non-parametric Models** SST [Shah et al., 2016]

Other ranking and permutations problems

Related Rectangular Problems :

- **Two permutations** [Mao et al., 2018, Shah et al., 2019] :
  M is bi-isotonic up to permutations π_1^* and π_2^* of rows and columns.
  Objective : ranking the experts and the questions.
- **Column isotony** [Flammarion et al., 2019]

Ranking Players in a tournament : M is a $n \times n$ matrix with symmetries.

- **Non-parametric Models** SST [Shah et al., 2016]
- **Parametric Models** :
  Bradley-Luce-Terry (e.g. [Chen et al., 2019, Chen et al., 2020])
  Noisy sorting [Braverman and Mossel, 2008]

Related Rectangular Problems :

- **Two permutations** [Mao et al., 2018, Shah et al., 2019] :
  M is bi-isotonic up to permutations π_1^* and π_2^* of rows and columns.
  Objective : ranking the experts and the questions.
- **Column isotony** [Flammarion et al., 2019]

Ranking Players in a tournament : M is a $n \times n$ matrix with symmetries.

- **Non-parametric Models** SST [Shah et al., 2016]
- **Parametric Models** :
  Bradley-Luce-Terry (e.g. [Chen et al., 2019, Chen et al., 2020])
  Noisy sorting [Braverman and Mossel, 2008]

Short story :

- No computational gap for *parametric models* (BLT, noisy sorting)
- mostly unknown for *non-parametric* models computational gaps are conjectured

Main questions

1. Is Estimating π^* much easier than estimating M ?

Main questions

1. Is Estimating π^* much easier than estimating M ?
2. Is there a **computational-statistical gap** ? (as in clustering problems with many groups)

Main questions

1. Is Estimating π^* much easier than estimating M ?
2. Is there a **computational-statistical gap** ? (as in clustering problems with many groups)
3. Is the **non-parametric problem** intrinsically more challenging than the **parametric** one ?

Main questions

1. Is Estimating π^* much easier than estimating M ?
2. Is there a **computational-statistical gap** ? (as in clustering problems with many groups)
3. Is the **non-parametric problem** intrinsically more challenging than the **parametric** one ?

Our Results

- Control of $\mathcal{R}^*_{perm}(n, d)$
- A polynomial-time procedure achieves $\mathcal{R}^*_{perm}(n, d)$

Analysis of least-squares estimator [Shah et al., 2016]

- $\mathbf{\Pi}_n$ collection of all permutations of $[n]$
- Biso collection all bi-isotonic matrices in $[0,1]$

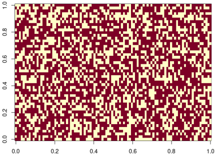
Least-square estimator

$$(\hat{M}^{\mathrm{LS}}, \hat{\pi}^{\mathrm{LS}}) = \underset{\widetilde{M} \in \mathrm{Biso}, \widetilde{\pi} \in \mathbf{\Pi}_n}{\arg\min} \left(\|\widetilde{M}_{\widetilde{\pi}} - Y\|_F^2\right)$$

- $\mathbf{\Pi}_n$ collection of all permutations of $[n]$
- Biso collection all bi-isotonic matrices in $[0,1]$

Least-square estimator

$$(\hat{M}^{\mathrm{LS}}, \hat{\pi}^{\mathrm{LS}}) = \underset{\widetilde{M}\in\mathrm{Biso}, \widetilde{\pi}\in\mathbf{\Pi}_n}{\arg\min} \left(\|\widetilde{M}_{\widetilde{\pi}} - Y\|_F^2\right)$$

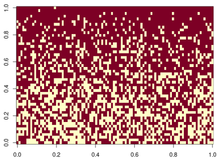


Matrix Y.

- $\mathbf{\Pi}_n$ collection of all permutations of $[n]$
- Biso collection all bi-isotonic matrices in $[0,1]$

Least-square estimator

$$(\hat{M}^{\mathrm{LS}}, \hat{\pi}^{\mathrm{LS}}) = \underset{\widetilde{M}\in \mathrm{Biso}, \widetilde{\pi}\in \mathbf{\Pi}_n}{\arg\min} \left(\|\widetilde{M}_{\widetilde{\pi}} - Y\|_F^2\right)$$



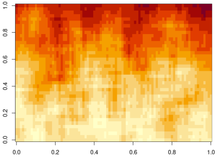
Matrix $Y_{\hat{\pi}^{\mathrm{LS}},.}$

Analysis of least-squares estimator [Shah et al., 2016]

- $\mathbf{\Pi}_n$ collection of all permutations of $[n]$
- Biso collection all bi-isotonic matrices in $[0,1]$

Least-square estimator

$$(\hat{M}^{\mathrm{LS}}, \hat{\pi}^{\mathrm{LS}}) = \underset{\widetilde{M}\in\mathrm{Biso}, \widetilde{\pi}\in\mathbf{\Pi}_n}{\arg\min} \left(\|\widetilde{M}_{\widetilde{\pi}} - Y\|_F^2\right)$$



Matrix $\hat{M}^{\mathrm{LS}}_{\hat{\pi}^{\mathrm{LS}},.}$.

Proposition (e.g.[Shah et al., 2016])

$$\mathbb{E}[\|\widehat{M} - M\|_F^2] \lesssim n + (\sqrt{nd} \wedge nd^{1/3})$$

In this presentation, $\asymp$, $\lesssim$, $\gtrsim$ is up to polylogarithms

Minimax Estimation Rates

Remarks :

- $\hat{M}^{\mathrm{LS}}$ is minimax for the estimation loss

$$\mathcal{R}^*_{est} \asymp n \vee (\sqrt{nd} \wedge nd^{1/3}) \ .$$

Minimax Estimation Rates

Remarks :

- $\hat{M}^{\mathrm{LS}}$ is minimax for the estimation loss

$$\mathcal{R}^*_{est} \asymp n \vee (\sqrt{nd} \wedge nd^{1/3}) \ .$$

- We have $\mathcal{R}^*_{perm} \gtrsim n$.

Minimax Estimation Rates

Remarks :

- $\hat{M}^{\mathrm{LS}}$ is minimax for the estimation loss

$$\mathcal{R}^*_{est} \asymp n \vee (\sqrt{nd} \wedge nd^{1/3}) \ .$$

- We have $\mathcal{R}^*_{perm} \gtrsim n$.

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	? ?	? ?	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n

But the algorithms are not polynomial time.

Global Average Comparison

e.g. [Pananjady and Samworth, 2020, Shah et al., 2019]

A simple ranking method :

- For each expert i, average performances on **all** questions :

$$\overline{Y}_i = \frac{1}{d}\sum_{k=1}^{d} Y_{i,k}$$

- Rank experts according to their average : $\hat{\pi}^{\mathrm{av}}$

Performances and failures of Global Average

Perfect expert on easy questions VS random expert :

$$M_{1,.} = (0.5, 0.5 \ldots 0.5, 0.5, \underbrace{0.9, 0.9, 0.9, 0.9}_{\asymp\sqrt{d}}) \quad ; \qquad M_{2,.} = (0.5, 0.5, \ldots, 0.5, 0.5)$$

Performances and failures of Global Average

Perfect expert on easy questions VS random expert :

$$M_{1,.} = (0.5, 0.5 \dots 0.5, 0.5, \underbrace{0.9, 0.9, 0.9, 0.9}_{\asymp\sqrt{d}}) \quad ; \qquad M_{2,.} = (0.5, 0.5, \dots, 0.5, 0.5)$$

Experts 1 and 2 are not distinguished by $\overline{Y}_i$ $\quad\rightsquigarrow\quad$ Risk($\hat{\pi}^{\mathsf{av}}$) $\asymp \sqrt{d}$

Performances and failures of Global Average

Perfect expert on easy questions VS random expert :

$$M_{1,.} = (0.5, 0.5 \dots 0.5, 0.5, \underbrace{0.9, 0.9, 0.9, 0.9}_{\asymp \sqrt{d}}) \quad ; \qquad M_{2,.} = (0.5, 0.5, \dots, 0.5, 0.5)$$

Experts 1 and 2 are not distinguished by $\overline{Y}_i$ $\quad \rightsquigarrow \quad$ $\text{Risk}(\hat{\pi}^{\mathsf{av}}) \asymp \sqrt{d}$

Guarantees on $\hat{\pi}^{\mathsf{av}}$

$$\sup_M \mathbb{E}\left[l(\hat{\pi}^{\mathsf{av}}, \pi^*)\right] \asymp n\sqrt{d}$$

Performances and failures of Global Average

Perfect expert on easy questions VS random expert :

$$M_{1,.} = (0.5, 0.5 \ldots 0.5, 0.5, \underbrace{0.9, 0.9, 0.9, 0.9}_{\asymp \sqrt{d}}) \quad ; \qquad M_{2,.} = (0.5, 0.5, \ldots, 0.5, 0.5)$$

Experts 1 and 2 are not distinguished by $\overline{Y}_i$ $\quad \rightsquigarrow \quad$ Risk$(\hat{\pi}^{\mathsf{av}}) \asymp \sqrt{d}$

Guarantees on $\hat{\pi}^{\mathsf{av}}$

$$\sup_M \mathbb{E}\left[l(\hat{\pi}^{\mathsf{av}}, \pi^*)\right] \asymp n\sqrt{d}$$

Sup-optimality of Global average :

- comparisons are not localized (similar phenomenon in tournament problems)
- Furthermore, one-to-one comparisons are not sufficient...

Improvements in [Mao et al., 2018] using local averages on bins.

Localization and Hierarchical clustering [Liu and Moitra, 2020]

[Liu and Moitra, 2020] **consider only** $d = n$, and provide a **poly. time** estimator $\hat{\pi}^{(LM)}$

$$\mathbb{E}\left[l(\hat{\pi}^{(LM)}, \pi^*)\right] \lesssim n^{1+o(1)}.$$

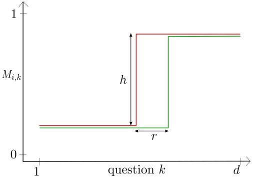
Optimal for $d = n$

Localization and Hierarchical clustering [Liu and Moitra, 2020]

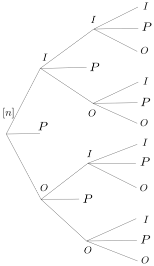
[Liu and Moitra, 2020] **consider only** $d = n$, and provide a **poly. time** estimator $\hat{\pi}^{(LM)}$

$$\mathbb{E}\left[l(\hat{\pi}^{(LM)}, \pi^*)\right] \lesssim n^{1+o(1)}.$$

Optimal for $d = n$



Localization through change-point detection.



Hierarchical sorting.

Summary

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	??	??	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Global average (UB)	$n\sqrt{d}$	$n\sqrt{d}$	$n\sqrt{d}$
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n

Summary

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	? ?	? ?	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Global average (UB)	$n\sqrt{d}$	$n\sqrt{d}$	$n\sqrt{d}$
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n

Remarks :

- Poly. time method of [Liu and Moitra, 2020] minimax for $d = n$
- Known UB for rates in $\mathcal{R}^*_{est}$ and $\mathcal{R}^*_{perm}$ not in polynomial time.

Minimax risks and polynomial time algorithm

Idealized setting : (as in [Liu and Moitra, 2020])
polylog independent full samples $Y^{(1)} = M + E^{(1)}$, $Y^{(2)} = M + E^{(2)}$, ...

Minimax risks and polynomial time algorithm

Idealized setting : (as in [Liu and Moitra, 2020])
polylog independent full samples $Y^{(1)} = M + E^{(1)}$, $Y^{(2)} = M + E^{(2)}$, ...

Theorem [Pilliat, Carpentier, V., 2022]

There exists a estimator $\hat{\pi}$ of π^* which is **poly. time** and **minimax optimal**

$$\mathbb{E}[l(\hat{\pi}, \pi^*)] \lesssim n + (n^{3/4} d^{1/4} \wedge n d^{1/6}) \asymp \mathcal{R}^*_{perm} \quad .$$

Minimax risks and polynomial time algorithm

Idealized setting : (as in [Liu and Moitra, 2020])
polylog independent full samples $Y^{(1)} = M + E^{(1)}$, $Y^{(2)} = M + E^{(2)}$, ...

Theorem [Pilliat, Carpentier, V., 2022]

There exists a estimator $\hat{\pi}$ of π^* which is **poly. time** and **minimax optimal**

$$\mathbb{E}[l(\hat{\pi}, \pi^*)] \lesssim n + (n^{3/4} d^{1/4} \wedge n d^{1/6}) \asymp \mathcal{R}^*_{perm} \quad .$$

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n

Minimax risks and polynomial time algorithm

Idealized setting : (as in [Liu and Moitra, 2020])
polylog independent full samples $Y^{(1)} = M + E^{(1)}$, $Y^{(2)} = M + E^{(2)}$, ...

Theorem [Pilliat, Carpentier, V., 2022]

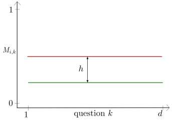
There exists a estimator $\hat{\pi}$ of π^* which is **poly. time** and **minimax optimal**

$$\mathbb{E}[l(\hat{\pi}, \pi^*)] \lesssim n + (n^{3/4} d^{1/4} \wedge n d^{1/6}) \asymp \mathcal{R}^*_{perm} \quad .$$

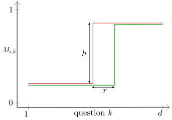
	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n

Consequence : Optimal estimation rate of M achievable in polynomial time.

From global to local averages

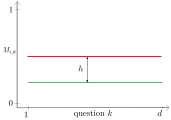


Global average good.

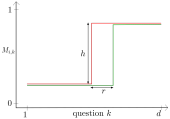


Global average bad → **localize**.

If M_1, M_2 not isotonic or unbounded
undistinguishable if $\|M_{1,.} - M_{2,.}\|_2^2 \lesssim \sqrt{d}$.



Global average good.



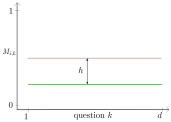
Global average bad → **localize**.

If M_1, M_2 not isotonic or unbounded undistinguishable if $\|M_{1,.} - M_{2,.}\|_2^2 \lesssim \sqrt{d}$.

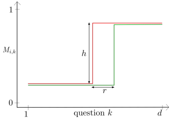
Idea :

- Local difference between experts $\rightsquigarrow$ a high-variation signature
- Variation signatures detectable at larger scale

From global to local averages



Global average good.



Global average bad → **localize**.

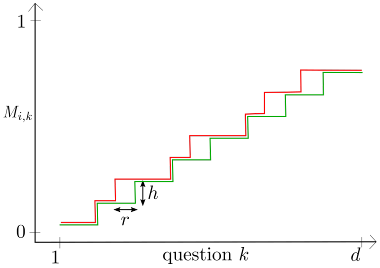
If M_1, M_2 not isotonic or unbounded
undistinguishable if $\|M_{1,.} - M_{2,.}\|_2^2 \lesssim \sqrt{d}$.

Idea :

- Local difference between experts
  $\rightsquigarrow$ a high-variation signature
- Variation signatures detectable at larger scale

Procedure

- Localize areas where any of the two experts varies by more than h...
- ... and compute local averages.



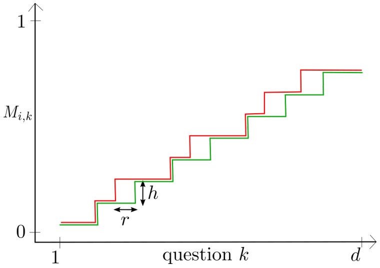
CUSUM Statistic :
$C_{l,r} = \frac{1}{r}\left(\sum_{k=l}^{l+r-1} Y_{1,k} - \sum_{k=l-r}^{l-1} Y_{1,k}\right)$

Pick height $h > 0$ and scale $r > 0$:

Step 1 High-Variation Detection $C_{l,r} \gtrsim h$

$$S_{r,h} = \bigcup\{[l-r, l+r) : C_{l,r} \gtrsim h\}$$

Step 2 Localized comparison
$\Psi(S_{r,h}) = \frac{1}{\sqrt{|S_{r,h}|}} \sum_{k \in S_{r,h}} (Y_{2,k} - Y_{1,k})$



CUSUM Statistic :
$C_{l,r} = \frac{1}{r}\left(\sum_{k=l}^{l+r-1} Y_{1,k} - \sum_{k=l-r}^{l-1} Y_{1,k}\right)$

Pick height $h > 0$ and scale $r > 0$:

Step 1 High-Variation Detection $C_{l,r} \gtrsim h$

$$S_{r,h} = \bigcup\{[l-r, l+r) : C_{l,r} \gtrsim h\}$$

Step 2 Localized comparison
$\Psi(S_{r,h}) = \frac{1}{\sqrt{|S_{r,h}|}} \sum_{k \in S_{r,h}} (Y_{2,k} - Y_{1,k})$

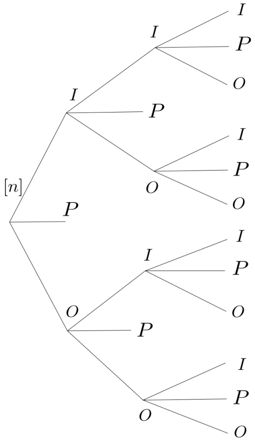
Proposition

Whp valid comparison if $\|M_{1,.} - M_{2,.}\|_2^2 \gtrsim d^{1/6}$

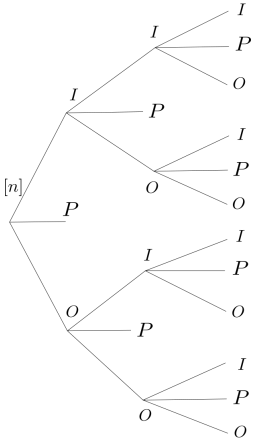
$\rightsquigarrow$ Conversely, optimal for $n = 2$.

Summary

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Global average (UB)	$n\sqrt{d}$	$n\sqrt{d}$	$n\sqrt{d}$
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n



Start from the **complete set** $[n]$ of experts

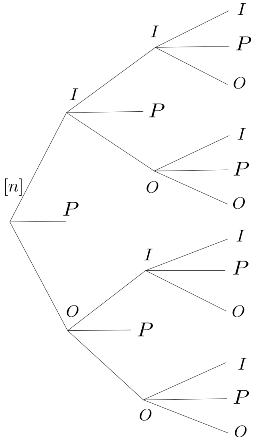


Start from the **complete set** $[n]$ of experts

Build a **Trisection** (O, P, I) of this set where :

1. Experts in O are whp among the $n/2$ worst
2. Experts in I are whp among the $n/2$ best
3. Experts in P are undecided

Hierarchical Sorting Tree



Start from the **complete set** $[n]$ of experts

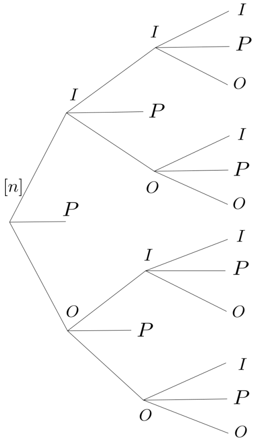
Build a **Trisection** (O, P, I) of this set where :

1. Experts in O are whp among the $n/2$ worst
2. Experts in I are whp among the $n/2$ best
3. Experts in P are undecided

. . . *Iterate on* O, P, I *with* **fresh** *samples*

⇝ **ordered partition** of $[n]$
⇝ Random partition $\widehat{\pi}$

Hierarchical Sorting Tree



Start from the **complete set** $[n]$ of experts

Build a **Trisection** (O, P, I) of this set where :

1. Experts in O are whp among the $n/2$ worst
2. Experts in I are whp among the $n/2$ best
3. Experts in P are undecided

... *Iterate on O, P, I with* ***fresh*** *samples*

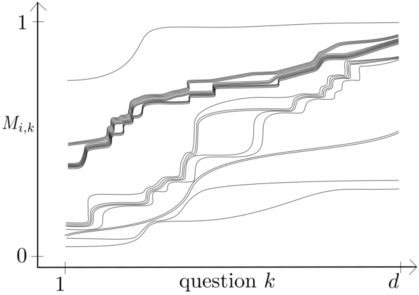
$\rightsquigarrow$ **ordered partition** of $[n]$
  $\rightsquigarrow$ Random partition $\widehat{\pi}$

Lemma

$$l(\widehat{\pi}, \pi^*) \lesssim \sum_{\overline{P}} \|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \ ,$$

where $\overline{P} \supset P$ (slighty larger set)

Partitionning a group G into three blocks

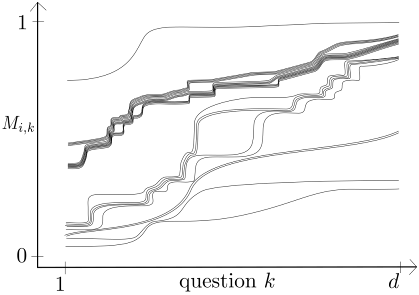


General Strategy :

$O = \varnothing$; $I = \varnothing$
For all heights h, scales r.

1. **Dimension Reduction**
   $\rightsquigarrow$ high-variation regions h of mean expert at scale r

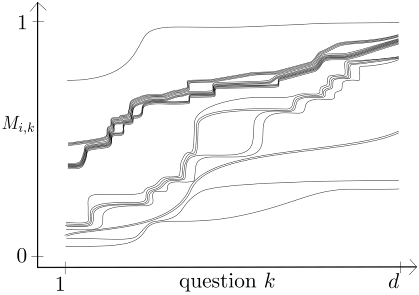
Partitionning a group G into three blocks



General Strategy :

$O = \varnothing$; $I = \varnothing$
For all heights h, scales r.

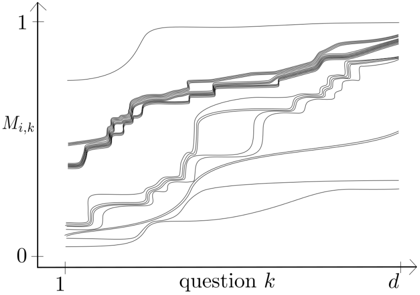
1. **Dimension Reduction**
   $\rightsquigarrow$ high-variation regions h of mean expert at scale r
2. Estimate a **direction** $\omega \in \mathbb{R}_+^d$



General Strategy :

$O = \varnothing$; $I = \varnothing$
For all heights h, scales r.

1. **Dimension Reduction**
   $\rightsquigarrow$ high-variation regions h of mean expert at scale r
2. Estimate a **direction** $\omega \in \mathbb{R}_+^d$
3. Expert **comparison** by
   weighted average $\sum_k Y_{i,k}\omega_k$
   $\rightsquigarrow (L, U) \subset G$
   $G \leftarrow G \setminus (U \cup L)$; $O \leftarrow O \cup L$; $I \leftarrow I \cup U$



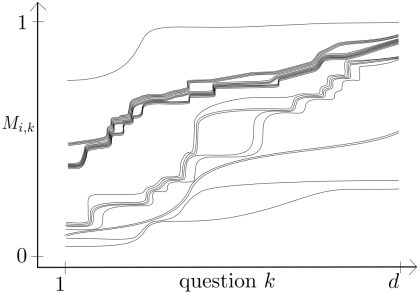
General Strategy :

$O = \varnothing$; $I = \varnothing$
For all heights h, scales r.

1. **Dimension Reduction**
   $\rightsquigarrow$ high-variation regions h of mean expert at scale r
2. Estimate a **direction** $\omega \in \mathbb{R}^d_+$
3. Expert **comparison** by weighted average $\sum_k Y_{i,k}\omega_k$
   $\rightsquigarrow (L, U) \subset G$
   $G \leftarrow G \setminus (U \cup L)$; $O \leftarrow O \cup L$; $I \leftarrow I \cup U$

Iterate *Polylog times*

Partitionning a group G into three blocks



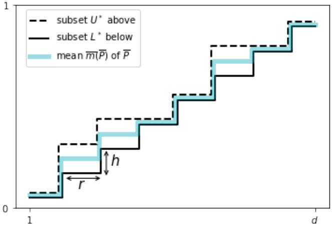
General Strategy :

$O = \varnothing$; $I = \varnothing$
For all heights h, scales r.

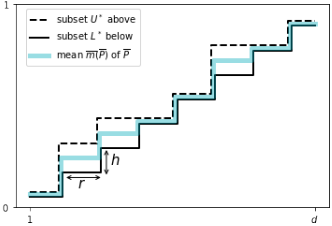
2 Estimate a **direction** $\omega \in \mathbb{R}_+^d$

Iterate *Polylog times*

Toy example (with two pure subgroups)



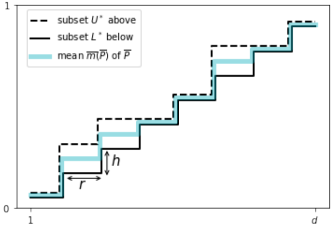
Toy example (with two pure subgroups)



Variation detection :

$\rightsquigarrow$ keeping windows of size r with variation h

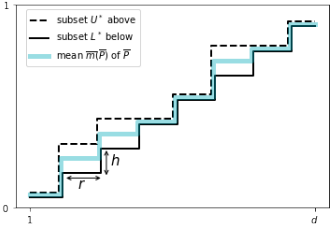
Toy example (with two pure subgroups)



Variation detection :
⇝ keeping windows of size r with variation h
Aggregation :
⇝ rescaled sum of observations on each window :

$$\frac{1}{2}\sqrt{r}h \times \begin{pmatrix} 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Toy example (with two pure subgroups)



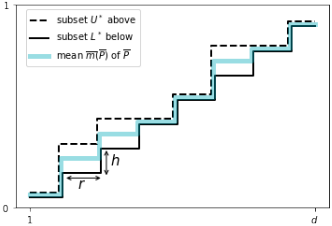
Variation detection :
⇝ keeping windows of size r with variation h
Aggregation :
⇝ rescaled sum of observations on each window :

$$\frac{1}{2}\sqrt{r}h \times \begin{pmatrix} 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Remark : For these two groups of experts

- Ranking = Clustering
- PCA outperforms row sums for large groups (to select active regions).

Toy example (with two pure subgroups)



Variation detection :
$\rightsquigarrow$ keeping windows of size r with variation h
Aggregation :
$\rightsquigarrow$ rescaled sum of observations on each window :

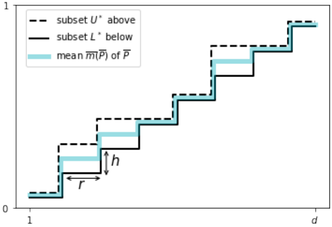
$$\frac{1}{2}\sqrt{r}h \times \begin{pmatrix} 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Remark : For these two groups of experts

- Ranking = Clustering
- PCA outperforms row sums for large groups (to select active regions).

Direction ω selection : right singular vector

Toy example (with two pure subgroups)



Variation detection :
$\rightsquigarrow$ keeping windows of size r with variation h
Aggregation :
$\rightsquigarrow$ rescaled sum of observations on each window :

$$\frac{1}{2}\sqrt{r}h \times \begin{pmatrix} 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Remark : For these two groups of experts

- Ranking = Clustering
- PCA outperforms row sums for large groups (to select active regions).

Direction ω **selection** : ~~right singular vector~~
left singular vector + image thresholding + correction
($\neq$ [Liu and Moitra, 2020])

Suboptimality of the procedure

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n
(modified) PCA+ Hierarchy	$nd^{1/6}$	$n^{2/3}d^{1/3}$	n

Suboptimality of the procedure

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n
(modified) PCA+ Hierarchy	$nd^{1/6}$	$n^{2/3}d^{1/3}$	n

Benefits of hierarchical Sorting :

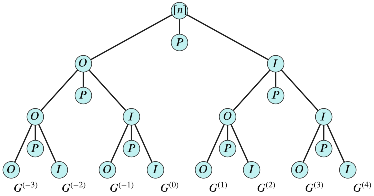
- Allows to localize the differences between subgroup of experts
- Builds upon large groups of close experts

Suboptimality of the procedure

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n
(modified) PCA+ Hierarchy	$nd^{1/6}$	$n^{2/3}d^{1/3}$	n

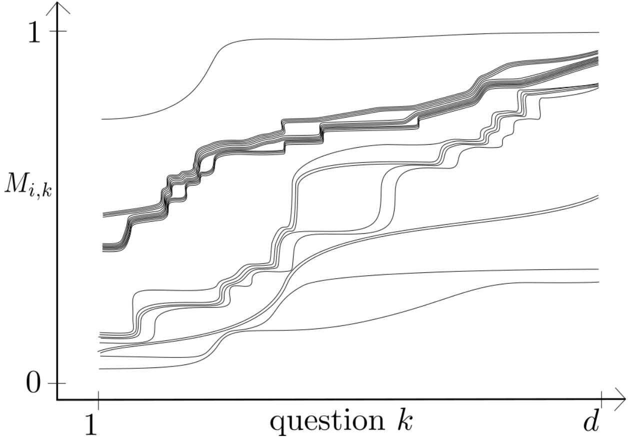
Benefits of hierarchical Sorting :

- Allows to localize the differences between subgroup of experts
- Builds upon large groups of close experts
- . . . but **oblivious** of previous structure found in the data



⇝ Hierarchical Sorting with Memory which is optimal.

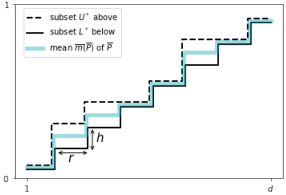
Bi-isotonic M



Each line $M_{i,\cdot}$ represents an expert i

Which information is brought by the tree ?

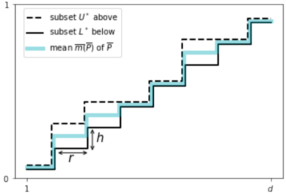
Our vanilla dimension reduction techniques :
Detection of variations of the mean expert in G



Which information is brought by the tree ?

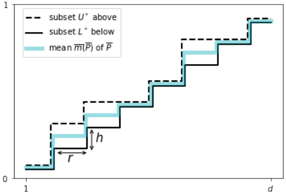
Our vanilla dimension reduction techniques :
Detection of variations of the mean expert in G
. . . but . . .

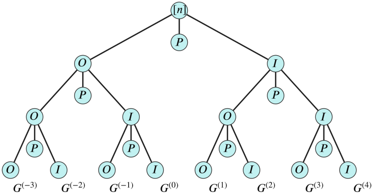
- A large scale r is needed if $|G|$ is small.
- Spurious regions are detected
(those where the width of G is small).



Our vanilla dimension reduction techniques :
Detection of variations of the mean expert in G
. . . but . . .

- A large scale r is needed if $|G|$ is small.
- Spurious regions are detected
(those where the width of G is small).





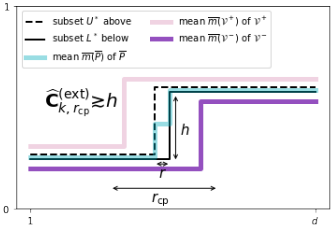
Idea :
Using the partial **ordering** to :

- decrease the variance of the CUSUM
(with $\mathcal{V} \supset G$ experts)
- Estimate the width Δ of G
$\Delta_k = \max_{i \in G} M_{i,k} - \max_{i \in G} M_{i,k}$ of G
by comparing mean experts in groups above and below G.

Fix a height h, and a scale r (possibly too small for G).
Consider expert sets $\mathcal{V}^+$ above G and $\mathcal{V}^-$ below G

Simultaneously check :

1. If **variations** at scale r higher than h
$\widehat{\mathbf{C}}^{(ext)}_{k,r} = \frac{1}{r} \sum_{l=k+1}^{k+r} \overline{y}_l(\mathcal{V}^+ \cup \mathcal{V}^-) - \sum_{l=k+1}^{k+r} \overline{y}_l(\mathcal{V}^+ \cup \mathcal{V}^-)$

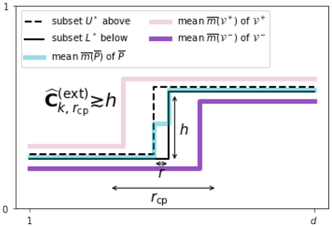


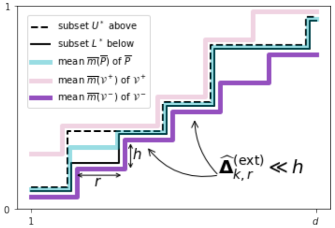
Fix a height h, and a scale r (possibly too small for G).
Consider expert sets $\mathcal{V}^+$ above G and $\mathcal{V}^-$ below G

Simultaneously check :

1. If **variations** at scale r higher than h
$\widehat{\mathbf{C}}^{(ext)}_{k,r} = \frac{1}{r}\sum_{l=k+1}^{k+r} \overline{y}_l(\mathcal{V}^+ \cup \mathcal{V}^-) - \sum_{l=k+1}^{k+r} \overline{y}_l(\mathcal{V}^+ \cup \mathcal{V}^-)$

2. If the **width** of G at scale $\frac{r}{2}$ higher than h.
$\widehat{\Delta}^{(ext)}_{k,r} = \frac{1}{r}\sum_{l=k-r}^{k+r} \overline{y}_l(\mathcal{V}^+) - \overline{y}_l(\mathcal{V}^-)$





Main result

Estimator $\widehat{\pi}^{WM}$ with this new **dimension reduction** step

Theorem

$$\textit{Max-Perm}(\hat{\pi})^{WM} \lesssim \left[nd^{1/6} \wedge (n^{3/4} d^{1/4}) \right] + n \asymp \textit{MiniMax-Perm}$$

Main result

Estimator $\widehat{\pi}^{WM}$ with this new **dimension reduction** step

Theorem

$$\textit{Max-Perm}(\hat{\pi})^{WM} \lesssim \left[nd^{1/6} \wedge (n^{3/4} d^{1/4}) \right] + n \asymp \textit{MiniMax-Perm}$$

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n
(modified) PCA+ Hierarchy+Memory	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n

Estimator $\widehat{\pi}^{WM}$ with this new **dimension reduction** step

Theorem

$$\textit{Max-Perm}(\hat{\pi})^{WM} \lesssim \left[nd^{1/6} \wedge (n^{3/4}d^{1/4})\right] + n \asymp \textit{MiniMax-Perm}$$

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^*_{perm}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^*_{est}$	$nd^{1/3}$	$\sqrt{nd}$	n
Ext. of [Liu and Moitra, 2020] (UB)	d	d	n
(modified) PCA+ Hierarchy+Memory	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n

$\rightsquigarrow$ As a corollary, minimax polynomial-time estimator of M.

Conclusion

- No **computational gap** for this ranking (and estimation) problem
- In comparison to $n = d$, rectangular setting requires **new ideas** :
  $\rightsquigarrow$ side information from partial ranking.
- Results extend to **partial observations** and **general noise** levels.

Conclusion

- No **computational gap** for this ranking (and estimation) problem
- In comparison to $n = d$, rectangular setting requires **new ideas** :
  $\rightsquigarrow$ side information from partial ranking.
- Results extend to **partial observations** and **general noise** levels.

For **two permutations**, existence of a computational gap is not clear.

References I

Bradley, R. A. and Terry, M. E. (1952).
Rank analysis of incomplete block designs : I. the method of paired comparisons.
Biometrika, 39(3/4) :324–345.

Braverman, M. and Mossel, E. (2008).
Noisy sorting without resampling.
In Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, pages 268–276.

Chen, P., Gao, C., and Zhang, A. Y. (2020).
Partial recovery for top-k ranking : Optimality of mle and sub-optimality of spectral method.
arXiv preprint arXiv :2006.16485.

Chen, Y., Fan, J., Ma, C., and Wang, K. (2019).
Spectral method and regularized mle are both optimal for top-k ranking.
Annals of statistics, 47(4) :2204.

Dawid, A. P. and Skene, A. M. (1979).
Maximum likelihood estimation of observer error-rates using the em algorithm.
Journal of the Royal Statistical Society : Series C (Applied Statistics), 28(1) :20–28.

Flammarion, N., Mao, C., and Rigollet, P. (2019).
Optimal rates of statistical seriation.
Bernoulli, 25(1) :623–653.

Liu, A. and Moitra, A. (2020).
Better algorithms for estimating non-parametric models in crowd-sourcing and rank aggregation.
In Abernethy, J. and Agarwal, S., editors, Proceedings of Thirty Third Conference on Learning Theory, volume 125 of Proceedings of Machine Learning Research, pages 2780–2829. PMLR.

Mao, C., Pananjady, A., and Wainwright, M. J. (2018).
Breaking the $1/\sqrt{n}$ barrier : Faster rates for permutation-based models in polynomial time.
In Conference On Learning Theory, pages 2037–2042. PMLR.

Pananjady, A. and Samworth, R. J. (2020).
Isotonic regression with unknown permutations : Statistics, computation, and adaptation.
arXiv preprint arXiv :2009.02609.

References III

Shah, N., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. (2016).
Stochastically transitive models for pairwise comparisons : Statistical and computational issues.
In International Conference on Machine Learning, pages 11–20. PMLR.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. (2019).
Feeling the bern : Adaptive estimators for bernoulli probabilities of pairwise comparisons.
IEEE Transactions on Information Theory, 65(8) :4854–4874.