Spectral Method and Regularized MLE Are Both Optimal for Top-K Ranking

Yuxin Chen

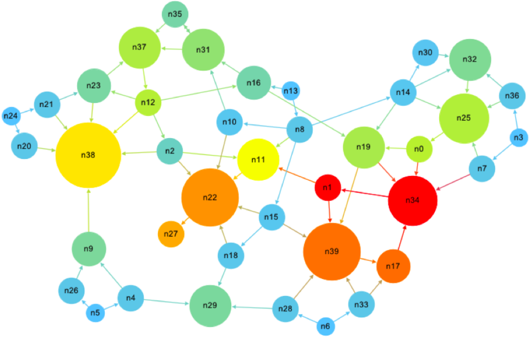
Electrical Engineering, Princeton University

Joint work with Jianqing Fan, Cong Ma and Kaizheng Wang

Ranking

A fundamental problem in a wide range of contexts

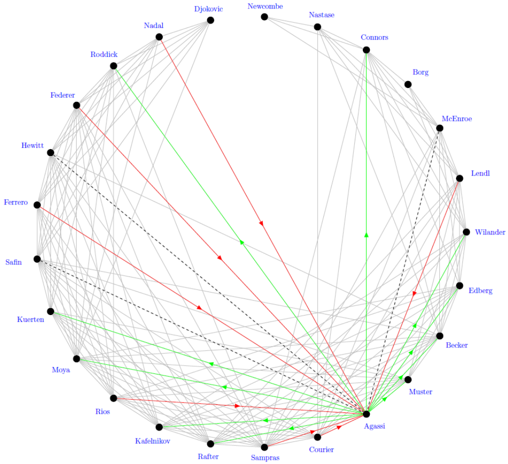
- web search, recommendation systems, admissions, sports competitions, voting, ...



PageRank

figure credit: Dzenan Hamzic

Rank aggregation from pairwise comparisons

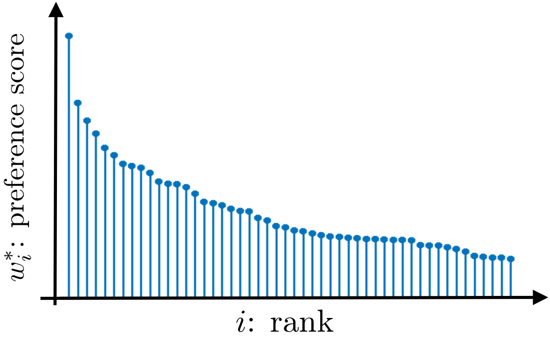


pairwise comparisons for ranking top tennis players

figure credit: Bozóki, Csató, Temesi

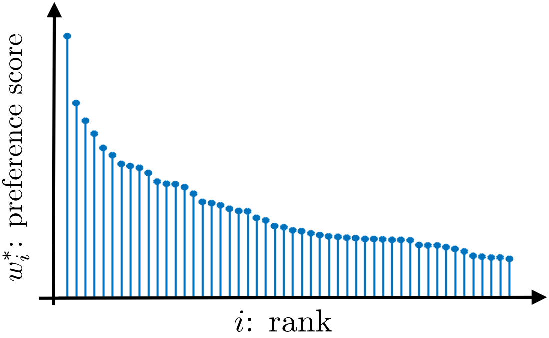
Parametric models

Assign latent score to each of n items $\boldsymbol{w}^* = [w_1^*, \cdots, w_n^*]$



Parametric models

Assign latent score to each of n items $\boldsymbol{w}^* = [w_1^*, \cdots, w_n^*]$



- **This work:** Bradley-Terry-Luce (logistic) model

$$\mathbb{P}\{\text{item } j \text{ beats item } i\} = \frac{w_j^*}{w_i^* + w_j^*}$$

- *Other models: Thurstone model, low-rank model, ...*

Typical ranking procedures

Estimate latent scores

$\longrightarrow$ rank items based on score estimates

Top-K ranking

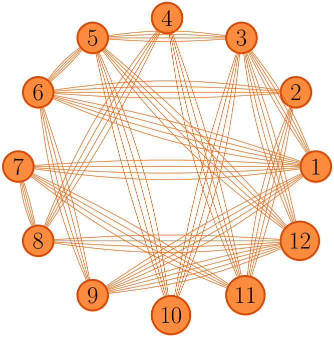
Estimate latent scores

$\longrightarrow$ rank items based on score estimates

Goal: identify the set of top-K items under minimal sample size

Model: random sampling

- Comparison graph: Erdős–Rényi graph $\mathcal{G} \sim \mathcal{G}(n, p)$

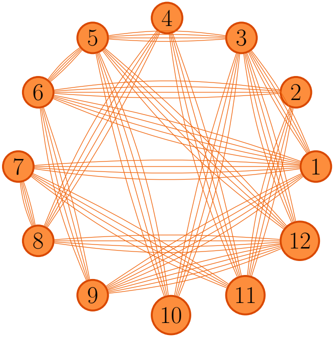


- For each $(i, j) \in \mathcal{G}$, obtain L paired comparisons

$$y_{i,j}^{(l)} \overset{\text{ind.}}{=} \begin{cases} 1, & \text{with prob. } \frac{w_j^*}{w_i^* + w_j^*} \\ 0, & \text{else} \end{cases} \qquad 1 \le l \le L$$

Model: random sampling

- Comparison graph: Erdős–Rényi graph $\mathcal{G} \sim \mathcal{G}(n, p)$



- For each $(i, j) \in \mathcal{G}$, obtain L paired comparisons

$$y_{i,j} = \frac{1}{L} \sum_{l=1}^{L} y_{i,j}^{(l)} \qquad \text{(sufficient statistic)}$$

Prior art

	mean square error for estimating scores	top-K ranking accuracy	
Spectral method	✔	?	Negahban et al. '12
MLE	✔	?	Negahban et al. '12 Hajek et al. '14
Spectral MLE	✔	✔	Chen & Suh. '15

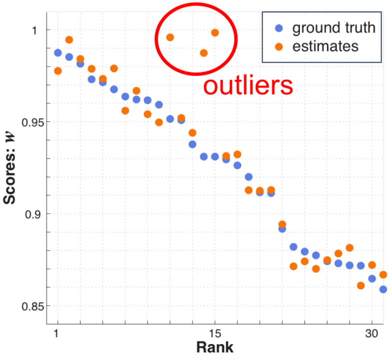
Prior art

"meta metric"

⇧

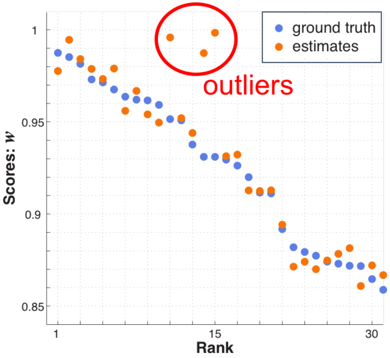
	mean square error for estimating scores	top-K ranking accuracy	
Spectral method	✔	?	Negahban et al. '12
MLE	✔	?	Negahban et al. '12 Hajek et al. '14
Spectral MLE	✔	✔	Chen & Suh. '15

Small ℓ_2 loss $\neq$ high ranking accuracy

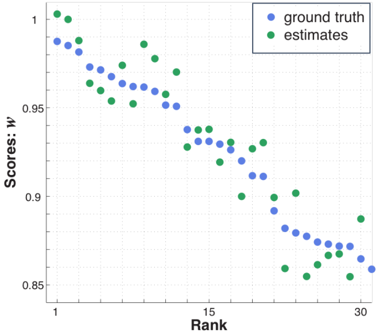


Top 3 : $\{15, 11, 2\}$

Small ℓ_2 loss $\neq$ high ranking accuracy

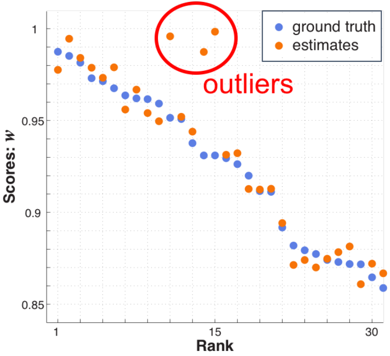


Top 3 : $\{15, 11, 2\}$

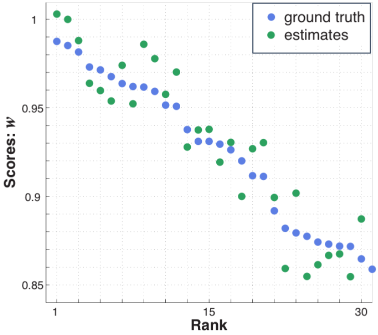


Top 3: $\{1, 2, 3\}$

Small ℓ_2 loss $\neq$ high ranking accuracy



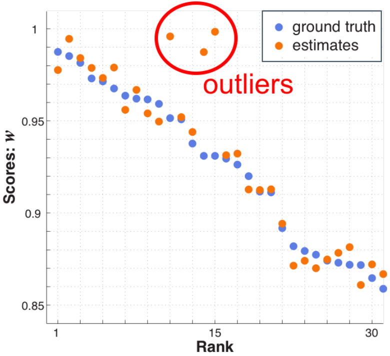
Top 3 : {15, 11, 2}



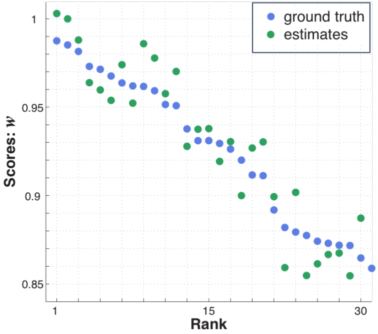
Top 3: {1, 2, 3}

These two estimates have same ℓ_2 loss, but output different rankings

Small ℓ_2 loss $\neq$ high ranking accuracy



Top 3 : $\{15, 11, 2\}$



Top 3: $\{1, 2, 3\}$

These two estimates have same ℓ_2 loss, but output different rankings

Need to control entrywise error!

Optimality?

Is spectral method or MLE alone optimal for top-K ranking?

Optimality?

Is spectral method or MLE alone optimal for top-K ranking?

Partial answer (Jang et al '16):

spectral method works if comparison graph is sufficiently dense

Optimality?

Is spectral method or MLE alone optimal for top-K ranking?

Partial answer (Jang et al '16):

spectral method works if comparison graph is sufficiently dense

This work: affirmative answer for both methods + $\underbrace{\textbf{entire regime}}_{\textbf{inc. sparse graphs}}$

Spectral method (Rank Centrality)

Negahban, Oh, Shah '12

- Construct a probability transition matrix $\boldsymbol{P}$, whose off-diagonal entries obey

$$P_{i,j} \;\propto\; \begin{cases} y_{i,j}, & \text{if } (i,j) \in \mathcal{G} \\ 0, & \text{if } (i,j) \notin \mathcal{G} \end{cases}$$

- Return score estimate as leading left eigenvector of $\boldsymbol{P}$

Rationale behind spectral method

In large-sample case, $\boldsymbol{P} \rightarrow \boldsymbol{P}^*$, whose off-diagonal entries obey

$$P^*_{i,j} \;\propto\; \begin{cases} \frac{w_j^*}{w_i^*+w_j^*}, & \text{if } (i,j) \in \mathcal{G} \\ 0, & \text{if } (i,j) \notin \mathcal{G} \end{cases}$$

- Stationary distribution of $\underbrace{\text{reversible}}_{\text{check detailed balance}}$ $\boldsymbol{P}^*$

$$\boldsymbol{\pi}^* \;\propto\; \underbrace{[w_1^*, w_2^*, \ldots, w_n^*]}_{\text{true score}}$$

Regularized MLE

Negative log-likelihood

$$\mathcal{L}(\boldsymbol{w}) := -\sum_{(i,j)\in\mathcal{G}} \left\{ y_{j,i} \log \frac{w_i}{w_i + w_j} + (1 - y_{j,i}) \log \frac{w_j}{w_i + w_j} \right\}$$

- $\mathcal{L}(\boldsymbol{w})$ becomes convex after reparametrization:

$$\boldsymbol{w} \quad \longrightarrow \quad \boldsymbol{\theta} = [\theta_1, \cdots, \theta_n], \qquad \theta_i = \log w_i$$

Regularized MLE

Negative log-likelihood

$$\mathcal{L}(\boldsymbol{w}) := -\sum_{(i,j)\in\mathcal{G}} \left\{ y_{j,i} \log \frac{w_i}{w_i + w_j} + (1 - y_{j,i}) \log \frac{w_j}{w_i + w_j} \right\}$$

- $\mathcal{L}(\boldsymbol{w})$ becomes convex after reparametrization:

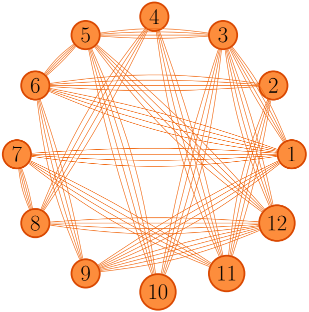
$$\boldsymbol{w} \quad \longrightarrow \quad \boldsymbol{\theta} = [\theta_1, \cdots, \theta_n], \qquad \theta_i = \log w_i$$

(Regularized MLE) minimize$_{\boldsymbol{\theta}}$ $\mathcal{L}_\lambda(\boldsymbol{\theta}) := \mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\lambda\|\boldsymbol{\theta}\|_2^2$

choose $\lambda \asymp \sqrt{\frac{np\log n}{L}}$

Main result

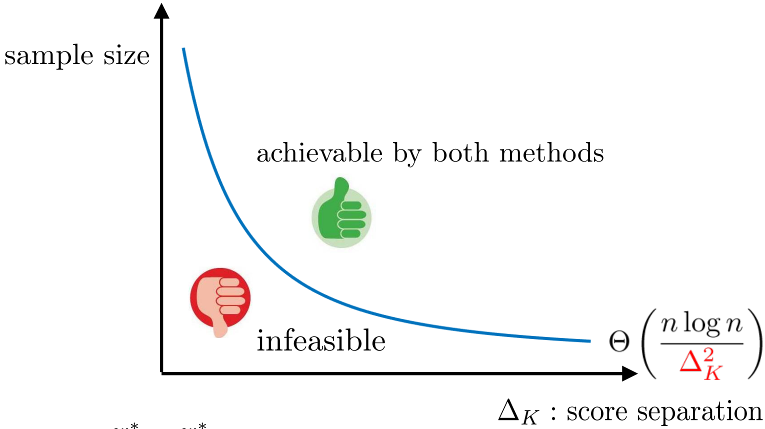
comparison graph $\mathcal{G}(n,p)$; sample size $\asymp pn^2L$



Theorem 1 (Chen, Fan, Ma, Wang '17)

When $p \gtrsim \frac{\log n}{n}$, both spectral method and regularized MLE achieve optimal sample complexity for top-K ranking!

Main result



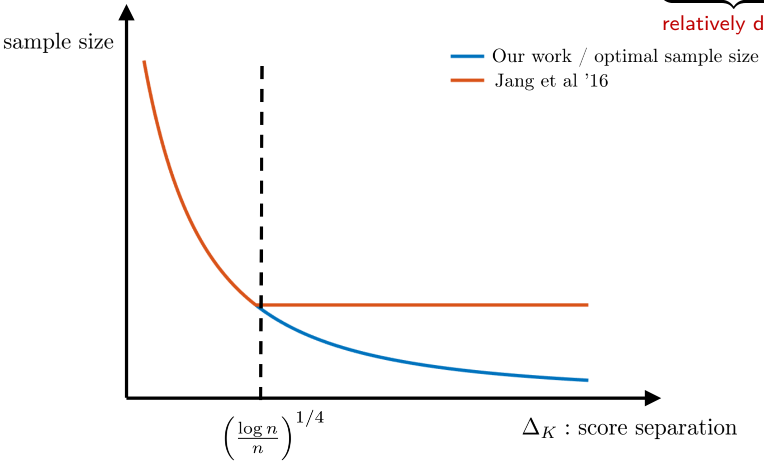
- $\Delta_K := \frac{w^*_{(K)} - w^*_{(K+1)}}{\|\boldsymbol{w}^*\|_\infty}$: score separation

Comparison with Jang et al '16

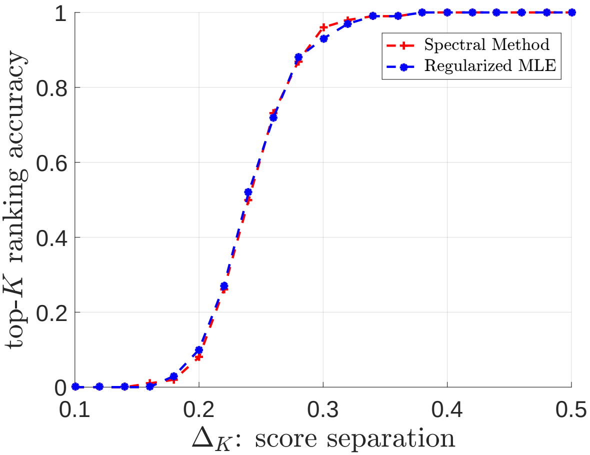
Jang et al '16: spectral method controls entrywise error if $\underbrace{p \gtrsim \sqrt{\frac{\log n}{n}}}_{\text{relatively dense}}$

Comparison with Jang et al '16

Jang et al '16: spectral method controls entrywise error if $\underbrace{p \gtrsim \sqrt{\frac{\log n}{n}}}_{\text{relatively dense}}$

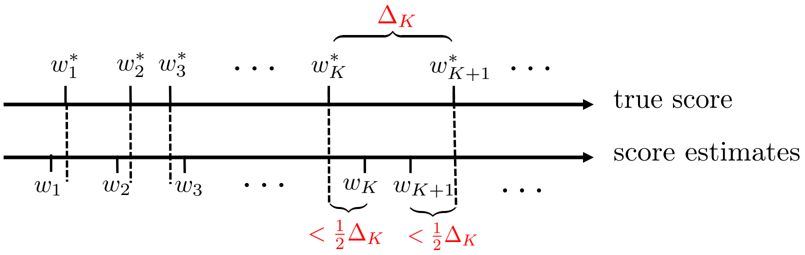


Empirical top-K ranking accuracy



$$n = 200, \;\; p = 0.25, \;\; L = 20$$

Optimal control of entrywise error



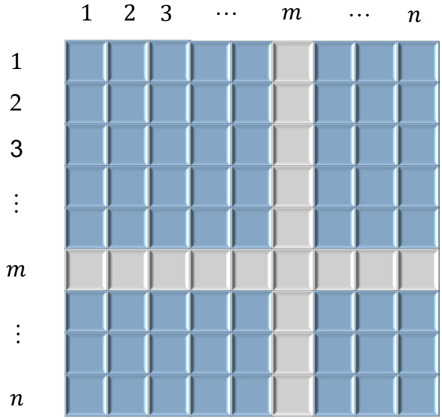
Theorem 2

Suppose $p \gtrsim \frac{\log n}{n}$ and sample size $\gtrsim \frac{n \log n}{\Delta_K^2}$. Then with high prob., the estimates $\boldsymbol{w}$ returned by both methods obey (up to global scaling)

$$\frac{\|\boldsymbol{w} - \boldsymbol{w}^*\|_\infty}{\|\boldsymbol{w}^*\|_\infty} < \frac{1}{2}\Delta_K$$

Key ingredient: leave-one-out analysis

For each $1 \leq m \leq n$, introduce leave-one-out estimate $\boldsymbol{w}^{(m)}$



$$\Longrightarrow \boldsymbol{w}^{(m)}$$

$$\boldsymbol{y} = [y_{i,j}]_{1 \leq i,j \leq n}$$

Key ingredient: leave-one-out analysis

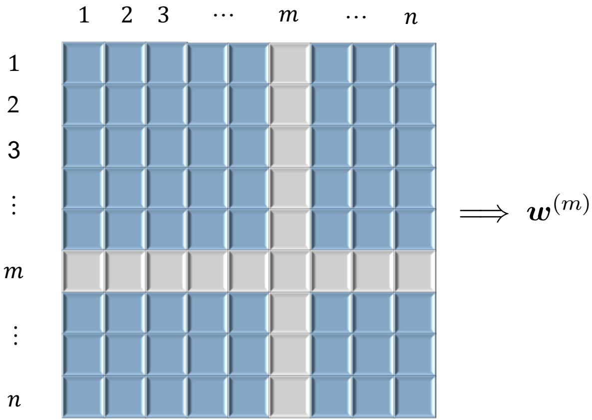
For each $1 \leq m \leq n$, introduce leave-one-out estimate $\boldsymbol{w}^{(m)}$

$$|w_m - w_m^*| \leq \underbrace{\left|w_m^{(m)} - w_m^*\right|}_{\text{Leave-one-out estimation error}} + \underbrace{\|\boldsymbol{w} - \boldsymbol{w}^{(m)}\|_2}_{\text{Leave-one-out perturbation}}$$

$\Downarrow$ statistical independence

$\Downarrow$ stability

Exploit statistical independence



$$\boldsymbol{y} = [y_{i,j}]_{1 \leq i,j \leq n}$$

leave-one-out estimate $\boldsymbol{w}^{(m)}$ $\perp\!\!\!\perp$ all data related to mth item

Leave-one-out stability

leave-one-out estimate $\boldsymbol{w}^{(m)} \approx$ true estimate $\boldsymbol{w}$

Leave-one-out stability

leave-one-out estimate $\boldsymbol{w}^{(m)} \;\approx\;$ true estimate $\boldsymbol{w}$

- Spectral method: eigenvector perturbation bound

$$\|\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}\|_{\boldsymbol{\pi}^*} \lesssim \frac{\big\|\boldsymbol{\pi}(\boldsymbol{P} - \widehat{\boldsymbol{P}})\big\|_{\boldsymbol{\pi}^*}}{\text{spectral-gap}}$$

  - new Davis-Kahan bound for $\underbrace{\text{probability transition matrices}}_{\textit{asymmetric}}$

Leave-one-out stability

leave-one-out estimate $\boldsymbol{w}^{(m)} \approx$ true estimate $\boldsymbol{w}$

- Spectral method: eigenvector perturbation bound

$$\|\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}\|_{\boldsymbol{\pi}^*} \lesssim \frac{\left\|\boldsymbol{\pi}(\boldsymbol{P} - \widehat{\boldsymbol{P}})\right\|_{\boldsymbol{\pi}^*}}{\text{spectral-gap}}$$

  - new Davis-Kahan bound for $\underbrace{\text{probability transition matrices}}_{\textit{asymmetric}}$

- MLE: local strong convexity

$$\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \lesssim \frac{\|\nabla \mathcal{L}_\lambda(\boldsymbol{\theta}; \widehat{\boldsymbol{y}})\|_2}{\text{strong convexity parameter}}$$

A small sample of related works

- **Parametric models**
  - Ford '57
  - Hunter '04
  - Negahban, Oh, Shah '12
  - Rajkumar, Agarwal '14
  - Hajek, Oh, Xu '14
  - Chen, Suh '15
  - Rajkumar, Agarwal '16
  - Jang, Kim, Suh, Oh '16
  - Suh, Tan, Zhao '17
- **Non-parametric models**
  - Shah, Wainwright '15
  - Shah, Balakrishnan, Guntuboyina, Wainwright '16
  - Chen, Gopi, Mao, Schneider '17
- **Leave-one-out analysis**
  - El Karoui, Bean, Bickel, Lim, Yu '13
  - Zhong, Boumal '17
  - Abbe, Fan, Wang, Zhong '17
  - Ma, Wang, Chi, Chen '17
  - Chen, Chi, Fan, Ma '18
  - Chen, Chi, Fan, Ma, Yan '19

Summary

	Optimal sample complexity	Linear-time computational complexity
Spectral method	✔	✔
Regularized MLE	✔	✔

Novel entrywise perturbation analysis for spectral method and convex optimization

Paper: "*Spectral method and regularized MLE are both optimal for top-K ranking*", Y. Chen, J. Fan, C. Ma, K. Wang, *Annals of Statistics*, vol. 47, 2019