Recent Advances in Ranking: Adversarial Respondents and Lower Bounds on the Bayes Risk

Vincent Y. F. Tan

Department of Electrical and Computer Engineering,
Department of Mathematics,
National University of Singapore

McMaster University
June 15, 2018

Outline

Outline

Ranking

- A fundamental problem in a wide range of contexts

Ranking

- A fundamental problem in a wide range of contexts
- Applications: web search, recommendation systems, social choice, sports competitions, voting, etc.

Ranking

- A fundamental problem in a wide range of contexts
- Applications: web search, recommendation systems, social choice, sports competitions, voting, etc.
- Efforts in developing various ranking algorithms

Ranking

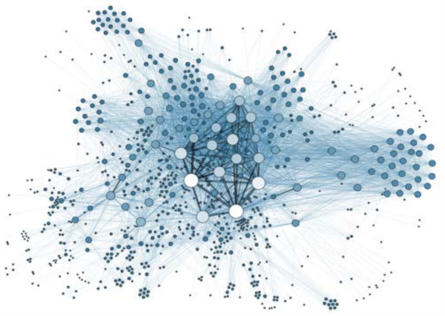
- A fundamental problem in a wide range of contexts
- Applications: web search, recommendation systems, social choice, sports competitions, voting, etc.
- Efforts in developing various ranking algorithms
- A variety of statistical models introduced for evaluating ranking algorithms

Ranking: An Example and Difficulties

Example: Web search

Ranking: An Example and Difficulties

Example: Web search



- $n = 10^9$ websites
- $\binom{n}{2} \approx n^2 = 10^{18}$ comparisons
- Do we really need $\Theta(n^2)$ comparisons?

Large-Scale Ranking

- Suppose that
  - we want a total ordering
  - pairwise comparisons are randomly given (probabilistically).

This indeed requires $\Theta(n^2)$ comparisons

Large-Scale Ranking

- Suppose that
  - we want a total ordering
  - pairwise comparisons are randomly given (probabilistically).

  This indeed requires $\Theta(n^2)$ comparisons
- No way to identify the ordering between 1 and 2 without a direct comparison, i.e., comparison must be made w.p. 1

Large-Scale Ranking

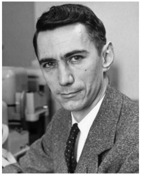
- Suppose that
  - we want a total ordering
  - pairwise comparisons are randomly given (probabilistically).

  This indeed requires $\Theta(n^2)$ comparisons
- No way to identify the ordering between 1 and 2 without a direct comparison, i.e., comparison must be made w.p. 1
- Worse with noisy data

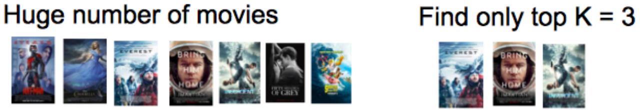
Large-Scale Ranking

- Suppose that
  - we want a total ordering
  - pairwise comparisons are randomly given (probabilistically).

  This indeed requires $\Theta(n^2)$ comparisons
- No way to identify the ordering between 1 and 2 without a direct comparison, i.e., comparison must be made w.p. 1
- Worse with noisy data
- Adopt a Shannon-theoretic approach in our analyses

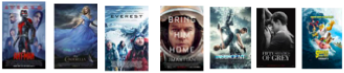


Top-K Ranking Usually Suffices

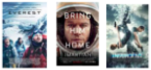


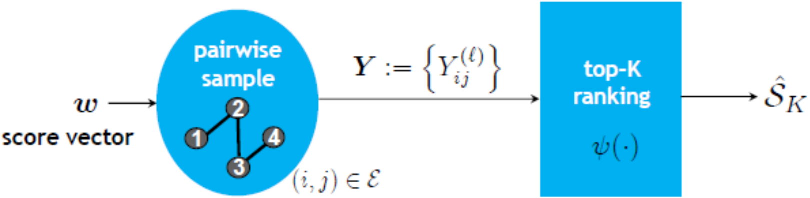
Top-K Ranking Usually Suffices

Huge number of movies



Find only top K = 3





Statistical Model for Top-K Ranking: Part I

- Adopt the Bradley-Terry-Luce or BTL model in which there is an underlying unknown score vector

$$\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n_{++},$$

where w_i is the likeability of movie i.

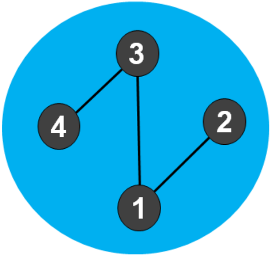
Statistical Model for Top-K Ranking: Part I

- Adopt the Bradley-Terry-Luce or BTL model in which there is an underlying unknown score vector

$$\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n_{++},$$

where w_i is the likeability of movie i.

- Decide which items to compare via a comparison graph



Statistical Model for Top-K Ranking: Part II

- The outcome of the comparison between item 1 and 2 is

$$Y_{12} = \mathbb{I}\{\text{item 1} \succ \text{item 2}\} \sim \mathrm{Bern}\left(\frac{w_1}{w_1 + w_2}\right).$$

Statistical Model for Top-K Ranking: Part II

- The outcome of the comparison between item 1 and 2 is

$$Y_{12} = \mathbb{I}\{\text{item 1} \succ \text{item 2}\} \sim \mathrm{Bern}\left(\frac{w_1}{w_1 + w_2}\right).$$

- E.g., $w_1 = 0.9$ and $w_2 = 0.1$, then item 1 beats item 2 w.p. 90%.

Statistical Model for Top-K Ranking: Part II

- The outcome of the comparison between item 1 and 2 is

$$Y_{12} = \mathbb{I}\{\text{item 1} \succ \text{item 2}\} \sim \mathrm{Bern}\left(\frac{w_1}{w_1 + w_2}\right).$$

- E.g., $w_1 = 0.9$ and $w_2 = 0.1$, then item 1 beats item 2 w.p. 90%.
- We have L independent copies

$$Y_{ij}^{(1)}, \ldots, Y_{ij}^{(L)}$$

for each observed edge $\{i, j\} \in \mathcal{E}$ of the observation graph.

Statistical Model for Top-K Ranking: Part II

- The outcome of the comparison between item 1 and 2 is

$$Y_{12} = \mathbb{I}\{\text{item 1} \succ \text{item 2}\} \sim \mathrm{Bern}\left(\frac{w_1}{w_1 + w_2}\right).$$

- E.g., $w_1 = 0.9$ and $w_2 = 0.1$, then item 1 beats item 2 w.p. 90%.
- We have L independent copies

$$Y_{ij}^{(1)}, \ldots, Y_{ij}^{(L)}$$

  for each observed edge $\{i, j\} \in \mathcal{E}$ of the observation graph.
- Determine fundamental limits on L (as a function of n and other parameters) so that recovery of top-K set is successful.

Outline

Top-K Ranking with Adversaries

Joint work with

Changho Suh (KAIST) Renbo Zhao (NUS)

Top-K Ranking with Adversaries

Joint work with

Changho Suh (KAIST) Renbo Zhao (NUS)

- C. Suh, **VYFT** and R. Zhao "Adversarial Top-K Ranking", IEEE Trans. on Inf. Theory, Apr 2017

Ranking with Adversaries: Crowdsourced Setting

faithful population η portion

adversarial population $1-\eta$ portion

Ranking with Adversaries: Crowdsourced Setting

faithful
population
η
portion

adversarial
population
$1 - \eta$
portion

$$Y_{ij} \sim \mathrm{Bern}\left(\frac{w_i}{w_i + w_j}\right)$$

Ranking with Adversaries: Crowdsourced Setting

faithful population η portion

$$Y_{ij} \sim \text{Bern}\left(\frac{w_i}{w_i + w_j}\right)$$

adversarial population $1-\eta$ portion

$$Y_{ij} \sim \text{Bern}\left(\frac{1/w_i}{1/w_i + 1/w_j}\right)$$

Ranking with Adversaries: Crowdsourced Setting

faithful population η portion

$$Y_{ij} \sim \mathrm{Bern}\left(\frac{w_i}{w_i + w_j}\right)$$

adversarial population $1-\eta$ portion

$$Y_{ij} \sim \mathrm{Bern}\left(\frac{1/w_i}{1/w_i + 1/w_j}\right)$$

$$= \mathrm{Bern}\left(\frac{w_j}{w_i + w_j}\right)$$

Ranking with Adversaries: Crowdsourced Setting

faithful population η portion

$$Y_{ij} \sim \text{Bern}\left(\frac{w_i}{w_i + w_j}\right)$$

adversarial population $1-\eta$ portion

$$Y_{ij} \sim \text{Bern}\left(\frac{1/w_i}{1/w_i + 1/w_j}\right)$$

$$= \text{Bern}\left(\frac{w_j}{w_i + w_j}\right)$$

Spammers provide answers in an adversarial manner

Ranking with Adversaries: Crowdsourced Setting

faithful
population
η
portion

$$Y_{ij} \sim \mathrm{Bern}\left(\frac{w_i}{w_i + w_j}\right)$$

adversarial
population
$1 - \eta$
portion

$$Y_{ij} \sim \mathrm{Bern}\left(\frac{1/w_i}{1/w_i + 1/w_j}\right)$$

$$= \mathrm{Bern}\left(\frac{w_j}{w_i + w_j}\right)$$

Spammers provide answers in an adversarial manner

$$Y_{ij} \sim \mathrm{Bern}\left(\eta \cdot \frac{w_i}{w_i + w_j} + (1 - \eta) \cdot \frac{w_j}{w_i + w_j}\right)$$

Related Work: Crowdsourced BTL [Chen et al. '13][1]

- Given an observed pair, each sample has different distributions

$$Y_{ij}^{(l)} \sim \mathrm{Bern}\left(\eta_l \cdot \frac{w_i}{w_i + w_j} + (1 - \eta_l) \cdot \frac{1/w_i}{1/w_i + 1/w_j} \right)$$

where η_l is a quality parameter of measurement l

[1] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in WSDM, 2013

Related Work: Crowdsourced BTL [Chen et al. '13][1]

- Given an observed pair, each sample has different distributions

$$Y_{ij}^{(l)} \sim \text{Bern}\left(\eta_l \cdot \frac{w_i}{w_i + w_j} + (1 - \eta_l) \cdot \frac{1/w_i}{1/w_i + 1/w_j} \right)$$

where η_l is a quality parameter of measurement l

- Subsumes as a special case our adversarial BTL model when all quality parameters are the same

[1] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in WSDM, 2013

Related Work: Crowdsourced BTL [Chen et al. ’13][1]

- Given an observed pair, each sample has different distributions

$$Y_{ij}^{(l)} \sim \mathrm{Bern}\left(\eta_l \cdot \frac{w_i}{w_i + w_j} + (1 - \eta_l) \cdot \frac{1/w_i}{1/w_i + 1/w_j} \right)$$

where η_l is a quality parameter of measurement l

- Subsumes as a special case our adversarial BTL model when all quality parameters are the same

- The authors developed a ranking algorithm but without theoretical guarantees

[1] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, “Pairwise ranking aggregation in a crowdsourced setting,” in WSDM, 2013

Related Work: Crowdsourced BTL [Chen et al. '13][1]

- Given an observed pair, each sample has different distributions

$$Y_{ij}^{(l)} \sim \mathrm{Bern}\left(\eta_l \cdot \frac{w_i}{w_i + w_j} + (1 - \eta_l) \cdot \frac{1/w_i}{1/w_i + 1/w_j} \right)$$

  where η_l is a quality parameter of measurement l

- Subsumes as a special case our adversarial BTL model when all quality parameters are the same

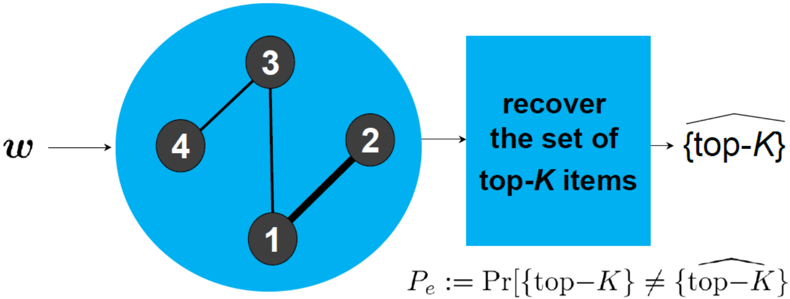
- The authors developed a ranking algorithm but without theoretical guarantees

- More difficult to analyze as there are many more parameters

[1] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in WSDM, 2013

Goal of Adversarial Top-K Ranking

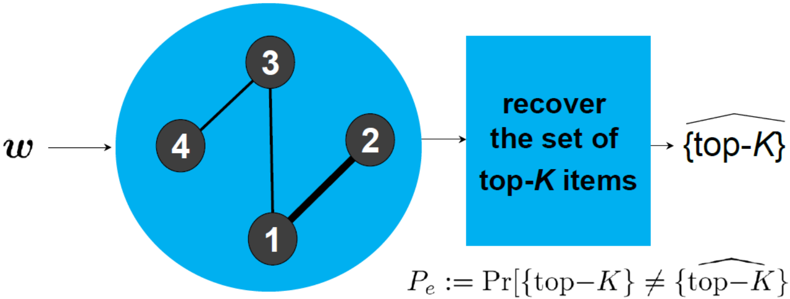
Erdös-Rényi comparison graph



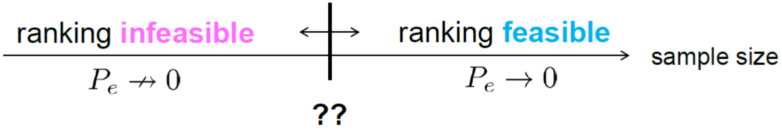
$$P_e := \Pr[\{\text{top}-K\} \neq \widehat{\{\text{top}-K\}}]$$

Goal of Adversarial Top-K Ranking

Erdös-Rényi comparison graph

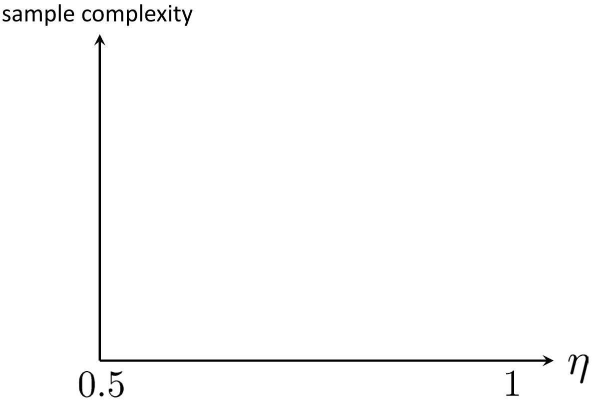


$$P_e := \Pr[\{\text{top}-K\} \neq \widehat{\{\text{top}-K\}}]$$



Contribution #1: $1/2 < \eta < 1$ known

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$

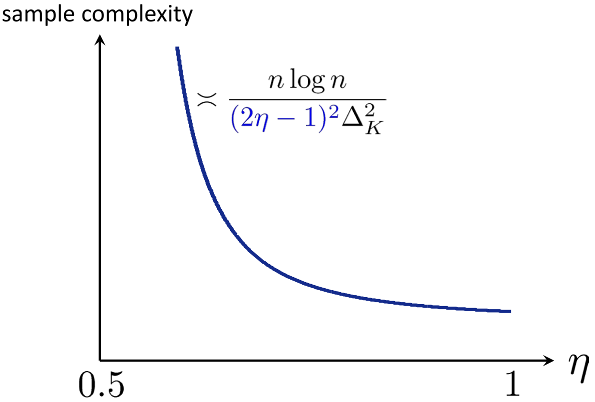


$\eta = 1$ studied by Chen and Suh (2015)[2]

[2] Y. Chen and C. Suh, "Spectral MLE: Top-K rank aggregation from pairwise comparisons," in ICML 2015

Contribution #1: $1/2 < \eta < 1$ known

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$

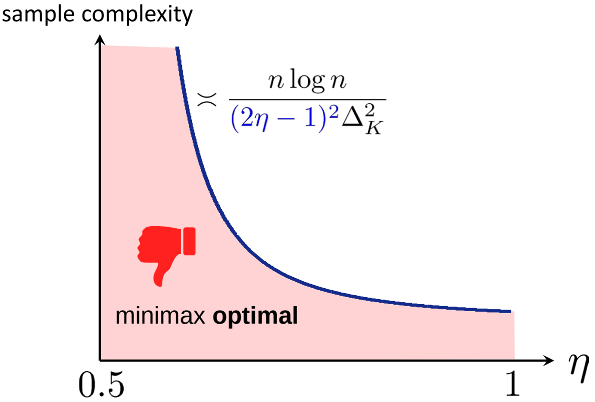


$\eta = 1$ studied by Chen and Suh (2015)[2]

[2] Y. Chen and C. Suh, "Spectral MLE: Top-K rank aggregation from pairwise comparisons," in ICML 2015

Contribution #1: $1/2 < \eta < 1$ known

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$

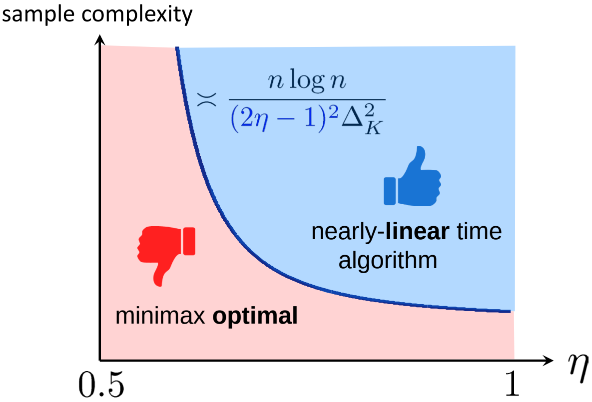


$\eta = 1$ studied by Chen and Suh (2015)[2]

[2] Y. Chen and C. Suh, "Spectral MLE: Top-K rank aggregation from pairwise comparisons," in ICML 2015

Contribution #1: $1/2 < \eta < 1$ known

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$



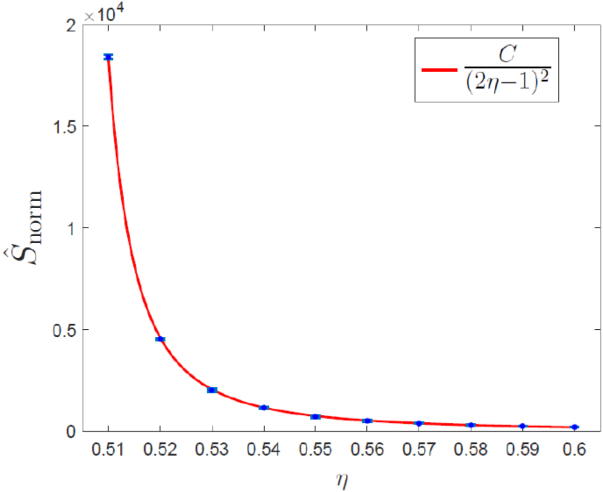
$\eta = 1$ studied by Chen and Suh (2015)[2]

---

[2] Y. Chen and C. Suh, "Spectral MLE: Top-K rank aggregation from pairwise comparisons," in ICML 2015

Contribution #1: $1/2 < \eta < 1$ known

Experimental Results for $n = 1000$ and $K = 10$

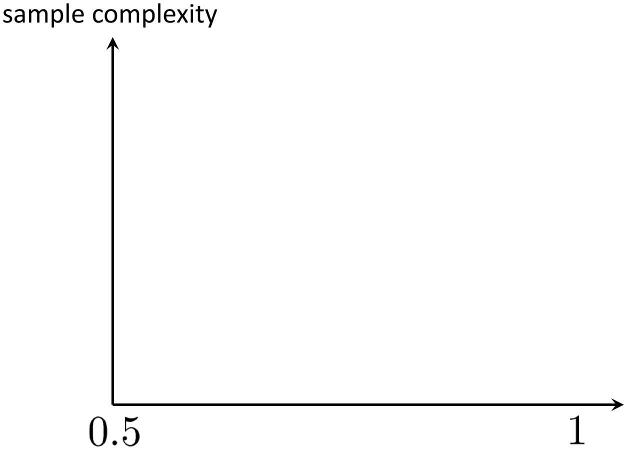


$$\hat{S} = \binom{n}{2} p \hat{L}_{\text{ave}}$$

$$\hat{S}_{\text{norm}} = \frac{\hat{S}}{(n \log n)/\Delta_K^2}$$

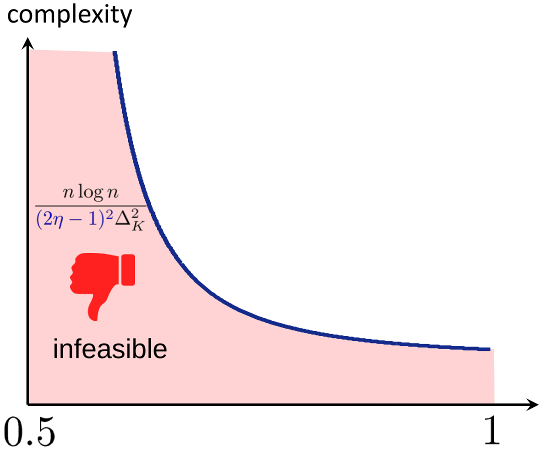
Contribution $\#2$: $1/2 < \eta < 1$ unknown

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$



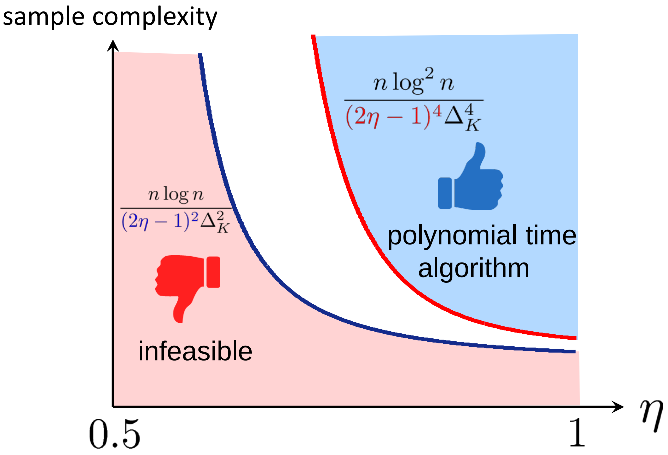
Contribution #2: $1/2 < \eta < 1$ unknown

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$



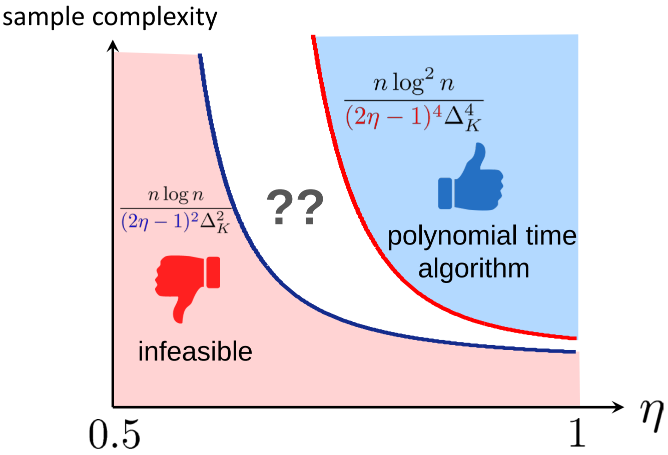
Contribution #2: $1/2 < \eta < 1$ unknown

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$

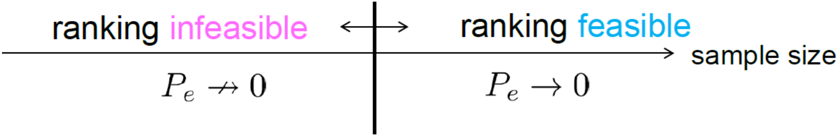


Contribution #2: $1/2 < \eta < 1$ unknown

$\eta =$ Fraction of non-adversaries; $\Delta_K \asymp w_K - w_{K+1}$

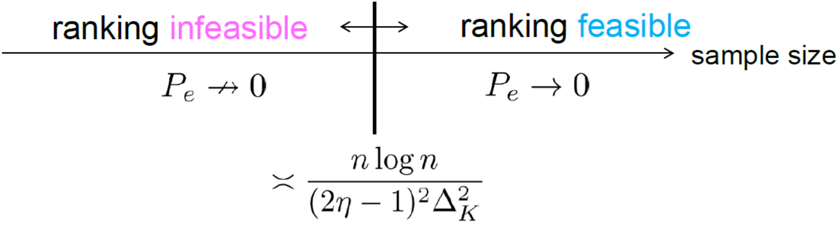


Optimality



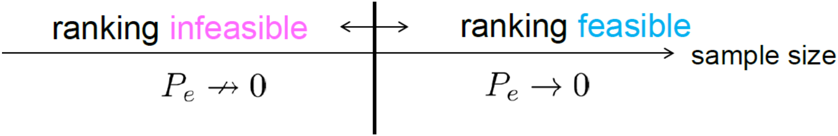
$$\asymp \frac{n \log n}{(2\eta - 1)^2 \Delta_K^2}$$

Optimality



- Minimax optimality: Construct “worst-case” score vectors

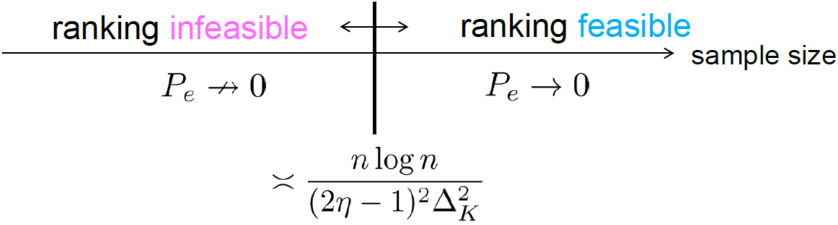
Optimality



$$\asymp \frac{n \log n}{(2\eta - 1)^2 \Delta_K^2}$$

- Minimax optimality: Construct “worst-case” score vectors
- Translation to M-ary hypothesis testing: Construction of multiple hypotheses

Optimality



- Minimax optimality: Construct “worst-case” score vectors
- Translation to M-ary hypothesis testing: Construction of multiple hypotheses
- Information-theoretic ideas applied to statistical learning

Optimality: Tools

- Construction of $M := \min\{K, n-K\} + 1 \le n/2$ hypotheses:

$$\Pr(\sigma([K]) = \mathcal{S}) = \frac{1}{M}, \text{ for } \mathcal{S} = \{2, \ldots, K\} \cup \{i\}, \; i = 1, K+1, \ldots, n$$

Optimality: Tools

- Construction of $M := \min\{K, n-K\} + 1 \le n/2$ hypotheses:

$$\Pr(\sigma([K]) = \mathcal{S}) = \frac{1}{M}, \text{ for } \mathcal{S} = \{2, \ldots, K\} \cup \{i\},\ i = 1, K+1, \ldots, n$$

- Bound mutual info. of permutation and "erased" version of $Y_{ij}^{(l)}$:

$$I(\sigma; \mathbf{Z}) \le \frac{p}{M^2} \sum_{\sigma_1, \sigma_2 \in \mathcal{M}} \sum_{l=1}^{L} \left\{ \sum_{i \ne j} D\left(P_{Y_{ij}^{(l)} | \sigma_1} \,\middle\|\, P_{Y_{ij}^{(l)} | \sigma_2} \right) \right\}$$

Optimality: Tools

- Construction of $M := \min\{K, n-K\} + 1 \le n/2$ hypotheses:

$$\Pr(\sigma([K]) = \mathcal{S}) = \frac{1}{M}, \text{ for } \mathcal{S} = \{2, \ldots, K\} \cup \{i\},\ i = 1, K+1, \ldots, n$$

- Bound mutual info. of permutation and "erased" version of $Y_{ij}^{(l)}$:

$$I(\sigma; \mathbf{Z}) \le \frac{p}{M^2} \sum_{\sigma_1, \sigma_2 \in \mathcal{M}} \sum_{l=1}^{L} \left\{ \sum_{i \ne j} D\left(P_{Y_{ij}^{(l)}|\sigma_1} \middle\| P_{Y_{ij}^{(l)}|\sigma_2} \right) \right\}$$

- Bound the divergence using reverse Pinsker's inequality. Here is where Δ_K comes in

$$\sum_{i \ne j} D\left(P_{Y_{ij}^{(l)}|\sigma_1} \middle\| P_{Y_{ij}^{(l)}|\sigma_2} \right) \le n \cdot (2\eta - 1)^2 \cdot \Delta_K^2$$

Optimality: Tools

- Construction of $M := \min\{K, n-K\} + 1 \leq n/2$ hypotheses:

$$\Pr(\sigma([K]) = \mathcal{S}) = \frac{1}{M}, \text{ for } \mathcal{S} = \{2, \ldots, K\} \cup \{i\}, \; i = 1, K+1, \ldots, n$$

- Bound mutual info. of permutation and "erased" version of $Y_{ij}^{(l)}$:

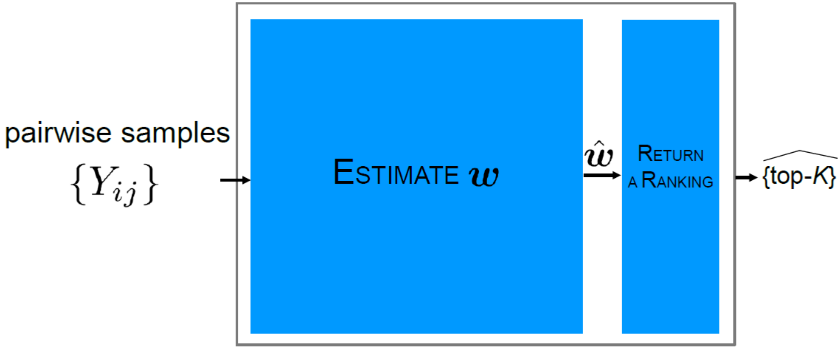
$$I(\sigma; \mathbf{Z}) \leq \frac{p}{M^2} \sum_{\sigma_1, \sigma_2 \in \mathcal{M}} \sum_{l=1}^{L} \left\{ \sum_{i \neq j} D\left(P_{Y_{ij}^{(l)}|\sigma_1} \middle\| P_{Y_{ij}^{(l)}|\sigma_2} \right) \right\}$$

- Bound the divergence using reverse Pinsker's inequality. Here is where Δ_K comes in

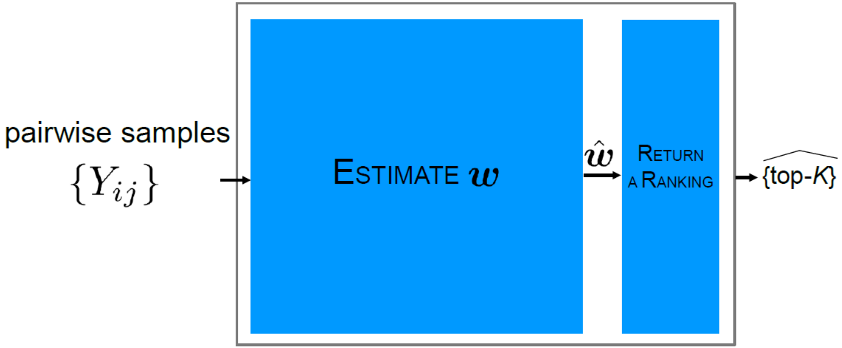
$$\sum_{i \neq j} D\left(P_{Y_{ij}^{(l)}|\sigma_1} \middle\| P_{Y_{ij}^{(l)}|\sigma_2} \right) \leq n \cdot (2\eta - 1)^2 \cdot \Delta_K^2$$

- Fano's inequality

Ranking Algorithm for η Known: Part I

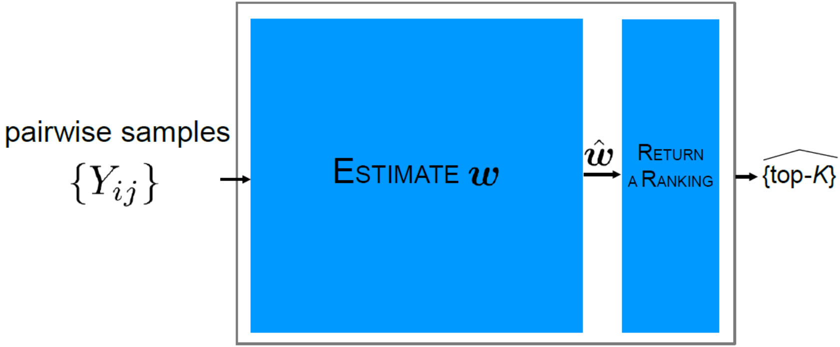


Ranking Algorithm for η Known: Part I



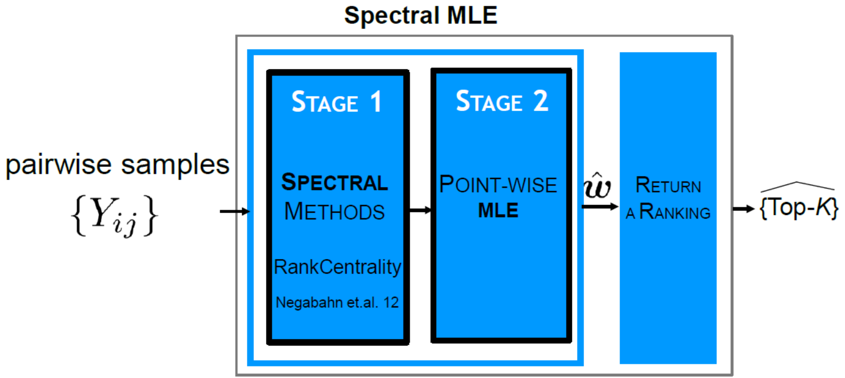
- Scores determine the ranking

Ranking Algorithm for η Known: Part I

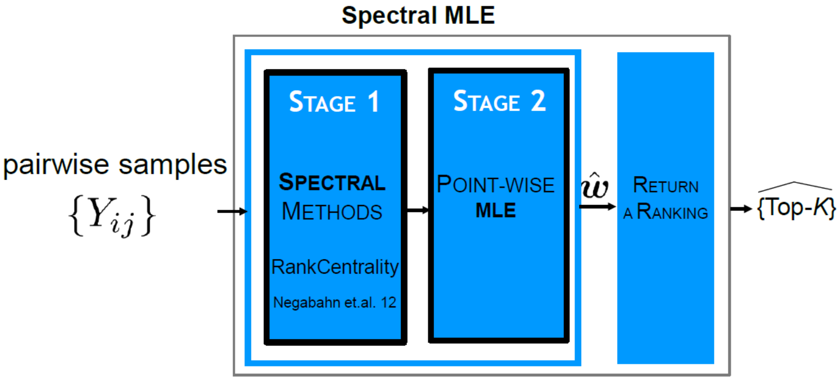


- Scores determine the ranking
- Adopt a two-step approach

Ranking Algorithm for η Known: Part II



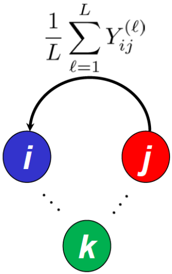
Ranking Algorithm for η Known: Part II



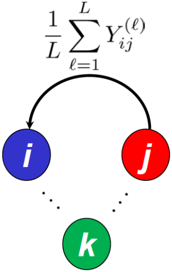
- Key Message:

Small MSE $\Longrightarrow$ Small ℓ_∞ Error of $\hat{\mathbf{w}}$ $\Longrightarrow$ High Ranking Accuracy

How to ensure small MSE for $\eta = 1$?

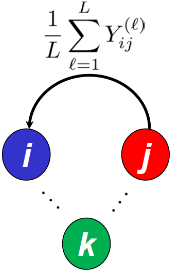


How to ensure small MSE for $\eta = 1$?



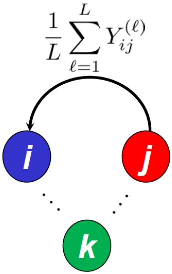
- Recall $\eta = 1$ (no adversaries)

How to ensure small MSE for $\eta = 1$?



- Recall $\eta = 1$ (no adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$

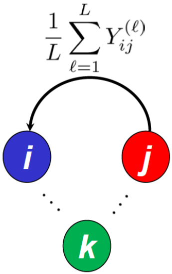
How to ensure small MSE for $\eta = 1$?



- Recall $\eta = 1$ (no adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$
- Convergence to stationary distribution

$$\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} \to \frac{w_i}{w_i + w_j}$$

How to ensure small MSE for $\eta = 1$?



- Recall $\eta = 1$ (no adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$
- Convergence to stationary distribution

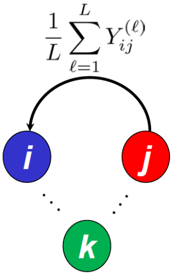
$$\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} \to \frac{w_i}{w_i + w_j}$$

- Detailed balance equation:

$$\pi_i \cdot \frac{w_j}{w_i + w_j} = \pi_j \cdot \frac{w_i}{w_i + w_j}$$

where $\boldsymbol{\pi} := [\pi_1, \pi_2, \ldots, \pi_n]$ is the stat. distn. of the chain.

How to ensure small MSE for $\eta = 1$?



- Recall $\eta = 1$ (no adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$
- Convergence to stationary distribution

$$\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} \to \frac{w_i}{w_i + w_j}$$

- Detailed balance equation:

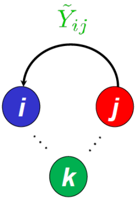
$$\pi_i \cdot \frac{w_j}{w_i + w_j} = \pi_j \cdot \frac{w_i}{w_i + w_j}$$

where $\boldsymbol{\pi} := [\pi_1, \pi_2, \ldots, \pi_n]$ is the stat. distn. of the chain.

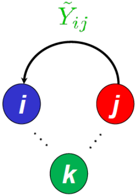
- Stationary distribution converges to $\mathbf{w}$ (up to constant scaling), i.e.,

$$\lim_{L\to\infty} \boldsymbol{\pi}^{(L)} = \alpha \mathbf{w}.$$

How to ensure small MSE for $\eta \in (1/2, 1]$?

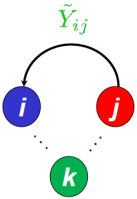


How to ensure small MSE for $\eta \in (1/2, 1]$?



- Arbitrary $\eta \in (1/2, 1]$ (adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$
- Redefine Markov chain

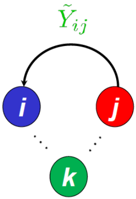
How to ensure small MSE for $\eta \in (1/2, 1]$?



- Arbitrary $\eta \in (1/2, 1]$ (adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$
- Redefine Markov chain

- We instead have the following convergence:

$$\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} \to \eta \frac{w_i}{w_i + w_j} + (1-\eta)\frac{w_j}{w_i + w_j} = (2\eta - 1)\frac{w_i}{w_i + w_j} + (1-\eta)$$



- Arbitrary $\eta \in (1/2, 1]$ (adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$
- Redefine Markov chain

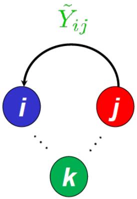
- We instead have the following convergence:

$$\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} \to \eta \frac{w_i}{w_i + w_j} + (1-\eta)\frac{w_j}{w_i + w_j} = (2\eta - 1)\frac{w_i}{w_i + w_j} + (1-\eta)$$

- Redefine "shifted" samples with range scaled by $\mathbf{2\eta - 1}$:

$$\tilde{Y}_{ij} = \frac{1}{\mathbf{2\eta - 1}}\left[\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} - (1-\eta)\right] \to \frac{w_i}{w_i + w_j}$$

How to ensure small MSE for $\eta \in (1/2, 1]$?



- Arbitrary $\eta \in (1/2, 1]$ (adversaries)
- L independent copies $Y_{ij}^{(1)}, Y_{ij}^{(2)}, \ldots, Y_{ij}^{(L)}$
- Redefine Markov chain

- We instead have the following convergence:

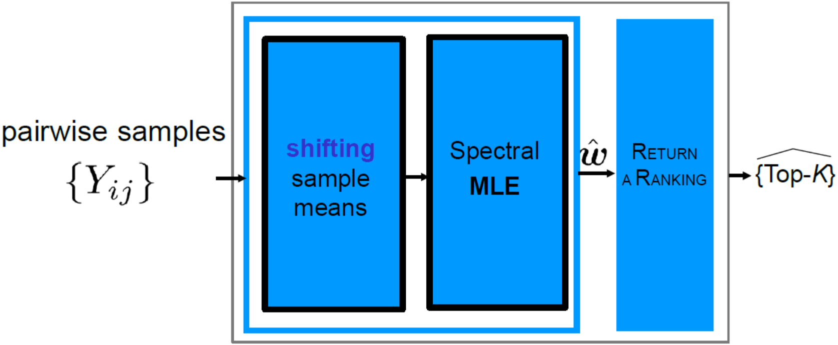
$$\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} \to \eta \frac{w_i}{w_i + w_j} + (1-\eta)\frac{w_j}{w_i + w_j} = (2\eta - 1)\frac{w_i}{w_i + w_j} + (1-\eta)$$

- Redefine "shifted" samples with range scaled by $\mathbf{2\eta - 1}$:

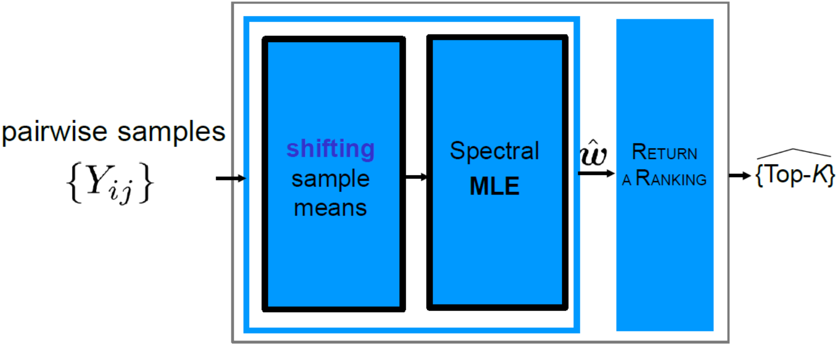
$$\tilde{Y}_{ij} = \frac{1}{\mathbf{2\eta - 1}}\left[\frac{1}{L}\sum_{l=1}^{L} Y_{ij}^{(l)} - (1-\eta)\right] \to \frac{w_i}{w_i + w_j}$$

- Construct Markov chain with transition probabilities $\{\tilde{Y}_{ij}\}$.

Ranking Algorithm for η Known: Summary



Ranking Algorithm for η Known: Summary



- Use several concentraition inequalities (Hoeffding, Bernstein, Tropp, etc.), we can show that if

$$\text{sample size} = L\binom{n}{2}p \succeq \frac{n \log n}{(2\eta - 1)^2 \Delta_K^2} \implies \text{Feasible Top-}K\text{ Ranking}$$

What if η is unknown?

- Adversarial BTL model is a mixture model

[3] P. Jain and S. Oh, "Learning mixtures of discrete product distributions using spectral decompositions," in COLT, 2014

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," JMLR, 2014.

What if η is unknown?

- Adversarial BTL model is a mixture model
- Obtaining global optimality guarantees for mixture model problems is difficult in general

[3] P. Jain and S. Oh, "Learning mixtures of discrete product distributions using spectral decompositions," in COLT, 2014

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," JMLR, 2014.

What if η is unknown?

- Adversarial BTL model is a mixture model
- Obtaining global optimality guarantees for mixture model problems is difficult in general
- Recent developments:
  - Tensor methods: Jain and Oh[3] and Anandkumar et al.[4]
  - Key idea: Exact 2nd and 3rd moments yield sufficient statistics

[3] P. Jain and S. Oh, "Learning mixtures of discrete product distributions using spectral decompositions," in COLT, 2014

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," JMLR, 2014.

What if η is unknown?

- Adversarial BTL model is a mixture model
- Obtaining global optimality guarantees for mixture model problems is difficult in general
- Recent developments:
  - Tensor methods: Jain and Oh[3] and Anandkumar et al.[4]
  - Key idea: Exact 2nd and 3rd moments yield sufficient statistics
- Our setting:
  - Can obtain estimates of 2nd and 3rd moments
  - Can estimate η

[3] P. Jain and S. Oh, "Learning mixtures of discrete product distributions using spectral decompositions," in COLT, 2014

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," JMLR, 2014.

Unknown η: High-Level Algorithm: Part I

1. Turn weights into distribution vectors

$$\pi_0 = \begin{bmatrix} \cdots & \frac{w_i}{w_i + w_j} & \frac{w_j}{w_i + w_j} & \frac{w_{i'}}{w_{i'} + w_{j'}} & \frac{w_{j'}}{w_{i'} + w_{j'}} & \cdots \end{bmatrix}^T$$

Unknown η: High-Level Algorithm: Part I

1. Turn weights into distribution vectors

$$\pi_0 = \left[\cdots \quad \frac{w_i}{w_i + w_j} \quad \frac{w_j}{w_i + w_j} \quad \frac{w_{i'}}{w_{i'} + w_{j'}} \quad \frac{w_{j'}}{w_{i'} + w_{j'}} \quad \cdots \right]^T$$

2. Estimate moments. Ground truth moment matrix and tensor are:

$$M_2 := \eta \pi_0 \otimes \pi_0 + (1 - \eta) \pi_1 \otimes \pi_1,$$
$$M_3 := \eta \pi_0 \otimes \pi_0 \otimes \pi_0 + (1 - \eta) \pi_1 \otimes \pi_1 \otimes \pi_1.$$

Unknown η: High-Level Algorithm: Part I

1. Turn weights into distribution vectors

$$\pi_0 = \begin{bmatrix} \cdots & \frac{w_i}{w_i + w_j} & \frac{w_j}{w_i + w_j} & \frac{w_{i'}}{w_{i'} + w_{j'}} & \frac{w_{j'}}{w_{i'} + w_{j'}} & \cdots \end{bmatrix}^T$$

2. Estimate moments. Ground truth moment matrix and tensor are:

$$M_2 := \eta \pi_0 \otimes \pi_0 + (1-\eta)\pi_1 \otimes \pi_1,$$
$$M_3 := \eta \pi_0 \otimes \pi_0 \otimes \pi_0 + (1-\eta)\pi_1 \otimes \pi_1 \otimes \pi_1.$$

3. Solves a Least Squares Problem

$$\hat{G} \in \operatorname*{arg\,min}_{Z \in \mathbb{R}^{2\times 2\times 2}} \left\| \mathcal{P}_{\Omega_3}\left(Z\left[P_{\hat{M}_2}\right]_3 - \frac{1}{|\mathcal{I}_2|} \sum_{t \in \mathcal{I}_2} \otimes^3 \underline{Y}^{(t)} \right) \right\|_F^2$$

Unknown η: High-Level Algorithm: Part I

1. Turn weights into distribution vectors

$$\pi_0 = \begin{bmatrix} \cdots & \frac{w_i}{w_i + w_j} & \frac{w_j}{w_i + w_j} & \frac{w_{i'}}{w_{i'} + w_{j'}} & \frac{w_{j'}}{w_{i'} + w_{j'}} & \cdots \end{bmatrix}^T$$

2. Estimate moments. Ground truth moment matrix and tensor are:

$$M_2 := \eta \pi_0 \otimes \pi_0 + (1-\eta)\pi_1 \otimes \pi_1,$$
$$M_3 := \eta \pi_0 \otimes \pi_0 \otimes \pi_0 + (1-\eta)\pi_1 \otimes \pi_1 \otimes \pi_1.$$

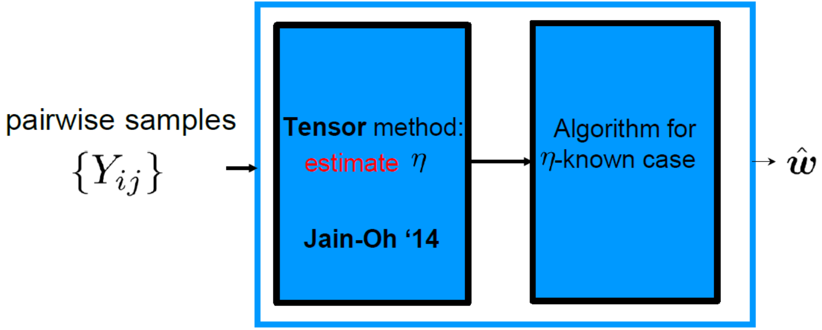
3. Solves a Least Squares Problem

$$\hat{G} \in \operatorname*{arg\,min}_{Z \in \mathbb{R}^{2\times 2\times 2}} \left\| \mathcal{P}_{\Omega_3}\left(Z\left[P_{\hat{M}_2}\right]_3 - \frac{1}{|\mathcal{I}_2|} \sum_{t \in \mathcal{I}_2} \otimes^3 \underline{Y}^{(t)} \right) \right\|_F^2$$

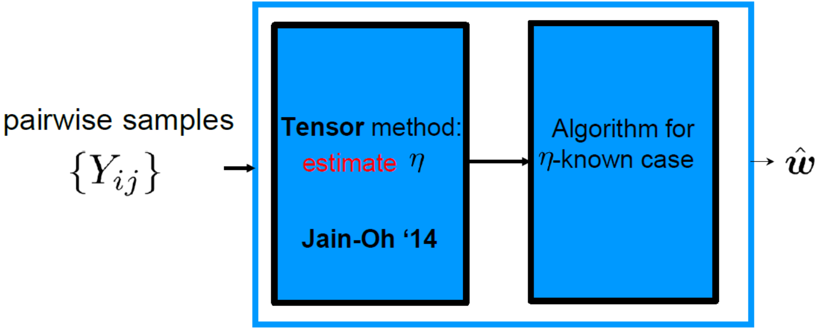
4. Find leading eigenvalue $\lambda_1(\hat{G})$ of $\hat{G}$ which is related to η as follows:

$$\hat{\eta} = \lambda_1(\hat{G})^{-2}$$

Unknown η: High-Level Algorithm: Part II



Unknown η: High-Level Algorithm: Part II



- How does the quality of the estimation of η affect overall sample complexity?

Tradeoff Between $|\hat{\eta} - \eta|$ and Sample Complexity

- With very careful analysis, we can derive a meta-lemma

$$|\hat{\eta} - \eta| \le \epsilon \quad \Longrightarrow \quad \text{Sample size} = L\binom{n}{2}p \succeq \frac{n \log^2 n}{\epsilon^2}$$

Tradeoff Between $|\hat{\eta} - \eta|$ and Sample Complexity

- With very careful analysis, we can derive a meta-lemma

$$|\hat{\eta} - \eta| \le \epsilon \quad \Longrightarrow \quad \text{Sample size} = L\binom{n}{2}p \succeq \frac{n \log^2 n}{\epsilon^2}$$

- This implies that
  - $|\hat{\eta} - \eta| \downarrow$ implies that $\|\hat{\mathbf{w}} - \mathbf{w}\|_\infty \downarrow$ but sample size $\uparrow$

Tradeoff Between $|\hat{\eta} - \eta|$ and Sample Complexity

- With very careful analysis, we can derive a meta-lemma

$$|\hat{\eta} - \eta| \le \epsilon \quad \Longrightarrow \quad \text{Sample size} = L\binom{n}{2}p \succeq \frac{n \log^2 n}{\epsilon^2}$$

- This implies that
  - $|\hat{\eta} - \eta| \downarrow$ implies that $\|\hat{\mathbf{w}} - \mathbf{w}\|_\infty \downarrow$ but sample size $\uparrow$
  - $|\hat{\eta} - \eta| \uparrow$ implies that sample size $\downarrow$ but $\|\hat{\mathbf{w}} - \mathbf{w}\|_\infty \uparrow$

Tradeoff Between $|\hat{\eta} - \eta|$ and Sample Complexity

- With very careful analysis, we can derive a meta-lemma

$$|\hat{\eta} - \eta| \le \epsilon \quad \Longrightarrow \quad \text{Sample size} = L\binom{n}{2}p \succeq \frac{n \log^2 n}{\epsilon^2}$$

- This implies that
  - $|\hat{\eta} - \eta| \downarrow$ implies that $\|\hat{\mathbf{w}} - \mathbf{w}\|_\infty \downarrow$ but sample size $\uparrow$
  - $|\hat{\eta} - \eta| \uparrow$ implies that sample size $\downarrow$ but $\|\hat{\mathbf{w}} - \mathbf{w}\|_\infty \uparrow$

- Find a sweet spot to show that

$$\text{sample size} \succeq \frac{n \log^2 n}{(2\eta - 1)^4 \Delta_K^4}, \quad \Longrightarrow \quad \text{Feasible Top-}K\text{ ranking}$$

Conclusion for Adversarial Top-K Ranking

- Explored a Top-K ranking problem for an adversarial setting

Conclusion for Adversarial Top-K Ranking

- Explored a Top-K ranking problem for an adversarial setting
- Characterized exact order-wise optimal sample complexity for η-known case

Conclusion for Adversarial Top-K Ranking

- Explored a Top-K ranking problem for an adversarial setting
- Characterized exact order-wise optimal sample complexity for η-known case
- Established an upper bound on the sample complexity for the η-unknown case

Conclusion for Adversarial Top-K Ranking

- Explored a Top-K ranking problem for an adversarial setting
- Characterized exact order-wise optimal sample complexity for η-known case
- Established an upper bound on the sample complexity for the η-unknown case
- Developed computationally efficient algorithms for both cases (using state-of-the-art tensor methods for the η-unknown case)

Conclusion for Adversarial Top-K Ranking

- Explored a Top-K ranking problem for an adversarial setting
- Characterized exact order-wise optimal sample complexity for η-known case
- Established an upper bound on the sample complexity for the η-unknown case
- Developed computationally efficient algorithms for both cases (using state-of-the-art tensor methods for the η-unknown case)
- C. Suh, **VYFT** and R. Zhao "Adversarial Top-K Ranking", IEEE Trans. on Inf. Theory, Apr 2017

Outline

Lower Bounds on the Risk of a Bayesian BTL Model

Joint work with

Mine Alsan
(NUS)

Ranjitha Prasad
(TCS Innovation Labs, Delhi)

Lower Bounds on the Risk of a Bayesian BTL Model

Joint work with

Mine Alsan
(NUS)

Ranjitha Prasad
(TCS Innovation Labs, Delhi)

- M. Alsan, R. Prasad and **VYFT**, "Lower Bounds on the Bayes Risk of the Bayesian BTL Model with Applications to Comparison Graphs", IEEE J. on Sel. Topics of Sig. Proc., Oct 2018

Summary of contributions

- Study the fundamental performance limits of ranking algorithms in the Bradley-Terry-Luce model within a Bayesian framework:

Summary of contributions

- Study the fundamental performance limits of ranking algorithms in the Bradley-Terry-Luce model within a Bayesian framework:

  1. Derive lower bounds on the Bayes Risk of estimators.

     - A family of information-theoretic lower bounds for norm-based distortion functions $\|\cdot\|_r^r$, for any $r \geq 1$.

     - The Bayesian Cramér-Rao bound for the MSE, i.e., $r = 2$.

Summary of contributions

- Study the fundamental performance limits of ranking algorithms in the Bradley-Terry-Luce model within a Bayesian framework:

1. Derive lower bounds on the Bayes Risk of estimators.

   - A family of information-theoretic lower bounds for norm-based distortion functions $\|\cdot\|_r^r$, for any $r \geq 1$.

   - The Bayesian Cramér-Rao bound for the MSE, i.e., $r = 2$.

2. Explore optimal comparison graph structures to design experiments minimizing distortion.

Recall the BTL Model

- BTL model: To each item $i \in [n]$, a skill parameter $w_i \in \mathbb{R}_{++}$ s.t.

$$P_{ij} := \Pr[\text{item } i \succ \text{item } j] = \frac{w_i}{w_i + w_j}.$$

Recall the BTL Model

- BTL model: To each item $i \in [n]$, a skill parameter $w_i \in \mathbb{R}_{++}$ s.t.

$$P_{ij} := \Pr[\text{item } i \succ \text{item } j] = \frac{w_i}{w_i + w_j}.$$

$\Rightarrow$ Instead of Top-K ranking, now we want to estimate the vector

$$\mathbf{w} := (w_1, \ldots, w_n) \in \mathbb{R}^n_{++}$$

Recall the BTL Model

- BTL model: To each item $i \in [n]$, a skill parameter $w_i \in \mathbb{R}_{++}$ s.t.

$$P_{ij} := \Pr[\text{item } i \succ \text{item } j] = \frac{w_i}{w_i + w_j}.$$

$\Rightarrow$ Instead of Top-K ranking, now we want to estimate the vector

$$\mathbf{w} := (w_1, \ldots, w_n) \in \mathbb{R}^n_{++}$$

- Given

$$m = \sum_{(i,j):i \neq j} m_{ij} \in \mathbb{N}$$

indep. pairwise comparisons, we count:

Recall the BTL Model

- BTL model: To each item $i \in [n]$, a skill parameter $w_i \in \mathbb{R}_{++}$ s.t.

$$P_{ij} := \Pr[\text{item } i \succ \text{item } j] = \frac{w_i}{w_i + w_j}.$$

$\Rightarrow$ Instead of Top-K ranking, now we want to estimate the vector

$$\mathbf{w} := (w_1, \ldots, w_n) \in \mathbb{R}^n_{++}$$

- Given

$$m = \sum_{(i,j): i \neq j} m_{ij} \in \mathbb{N}$$

indep. pairwise comparisons, we count:

1. m_{ij}: Num. of pairwise comparisons between items i & j,
2. b_{ij}: Num. of comparisons in which i is preferred over j.

$$\Rightarrow \mathbf{M} := \{m_{ij}\} \in \mathbb{N}^{n \times n} \quad \text{and} \quad \mathbf{B} := \{b_{ij}\} \in \mathbb{N}^{n \times n}.$$

Induced Probabilities by BTL Model

- We assume that the matrix $\mathbf{M} = \{m_{ij}\}$ is fixed a priori.

Induced Probabilities by BTL Model

- We assume that the matrix $\mathbf{M} = \{m_{ij}\}$ is fixed a priori.
- The BTL model induces the following distributions:

1. For fixed m_{ij},

$$p(b_{ij}|w_i, w_j) = \mathrm{Bin}(b_{ij}; m_{ij}, P_{ij}).$$

2. For fixed $\mathbf{M}$,

$$p(\mathbf{B}|\boldsymbol{\lambda}) = \prod_{(i,j):i<j} \mathrm{Bin}(b_{ij}; m_{ij}, P_{ij}),$$

Bayesian BTL Model

- Adopt the Bayesian BTL framework by Caron & Doucet[5]:

[5] F. Caron and A. Doucet, “Efficient Bayesian Inference for Generalized Bradley-Terry Models", in JCGS, 2012

Bayesian BTL Model

- Adopt the Bayesian BTL framework by Caron & Doucet[5]:

  1. Prior distribution: They assign

$$p(w_i) = \mathrm{Gam}(w_i; \alpha_i, \beta_i)$$

to each item $i \in [n]$, where $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^n, \boldsymbol{\beta} := \{\beta_i\}_{i=1}^n \in \mathbb{R}_{++}^n$.

[5] F. Caron and A. Doucet, "Efficient Bayesian Inference for Generalized Bradley-Terry Models", in JCGS, 2012

Bayesian BTL Model

- Adopt the Bayesian BTL framework by Caron & Doucet[5]:

  1. Prior distribution: They assign

  $$p(w_i) = \mathrm{Gam}(w_i; \alpha_i, \beta_i)$$

  to each item $i \in [n]$, where $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^n, \boldsymbol{\beta} := \{\beta_i\}_{i=1}^n \in \mathbb{R}_{++}^n$.

  2. Latent random variables: They introduce $\mathbf{Z} := \{Z_{ij}\} \in \mathbb{R}^{n\times n}$

  $$Z_{ij} = Z_{ji} := \sum_{s=1}^{m_{ij}} \min\{Y_{si}, Y_{sj}\},$$

  for $i, j \in [n]$ such that $i < j$, where

  $$Y_i \sim \mathrm{Exp}(w_i) \quad \& \quad Y_j \sim \mathrm{Exp}(w_j) \quad \text{such that} \quad P_{ij} = \Pr[Y_i < Y_j].$$

  Known as Thurstonian interpretation of the BTL model.

[5] F. Caron and A. Doucet, "Efficient Bayesian Inference for Generalized Bradley-Terry Models", in JCGS, 2012

Induced Probabilities by Bayesian BTL Model

1. Prior:

$$p(\mathbf{w}) = \prod_{i=1}^{n} p(w_i) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i, \beta_i),$$

Induced Probabilities by Bayesian BTL Model

1. Prior:

$$p(\mathbf{w}) = \prod_{i=1}^{n} p(w_i) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i, \beta_i),$$

2. Prior $\times$ Likelihood:

$$p(\mathbf{w}, \mathbf{B}) = p(\mathbf{w}) p(\mathbf{B}|\mathbf{w}) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i, \beta_i) \prod_{i<j} \mathrm{Bin}\left(b_{ij}; m_{ij}, \frac{w_i}{w_i + w_j}\right).$$

Induced Probabilities by Bayesian BTL Model

1. Prior:

$$p(\mathbf{w}) = \prod_{i=1}^{n} p(w_i) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i, \beta_i),$$

2. Prior × Likelihood:

$$p(\mathbf{w}, \mathbf{B}) = p(\mathbf{w})p(\mathbf{B}|\mathbf{w}) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i, \beta_i) \prod_{i<j} \mathrm{Bin}\left(b_{ij}; m_{ij}, \frac{w_i}{w_i + w_j}\right).$$

3. Latent Variable:

$$p(Z_{ij}|w_i, w_j) = \mathrm{Gam}(Z_{ij}; m_{ij}, w_i + w_j).$$

Induced Probabilities by Bayesian BTL Model

1. Prior:

$$p(\mathbf{w}) = \prod_{i=1}^{n} p(w_i) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i, \beta_i),$$

2. Prior $\times$ Likelihood:

$$p(\mathbf{w}, \mathbf{B}) = p(\mathbf{w})p(\mathbf{B}|\mathbf{w}) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i, \beta_i) \prod_{i<j} \mathrm{Bin}\left(b_{ij}; m_{ij}, \frac{w_i}{w_i + w_j}\right).$$

3. Latent Variable:

$$p(Z_{ij}|w_i, w_j) = \mathrm{Gam}(Z_{ij}; m_{ij}, w_i + w_j).$$

4. Posterior:

$$p(\mathbf{w}|\mathbf{B}, \mathbf{Z}) = \prod_{i=1}^{n} \mathrm{Gam}(w_i; \alpha_i + b_i, \beta_i + Z_i).$$

where $b_i := \sum_{j \neq i} b_{ij}$ and $Z_i := \sum_{j \neq i} Z_{ij}$.

Bayes Risk

- For any $r \geq 1$, we define the family of Bayes risks for estimating $\mathbf{w}$

Bayes Risk

- For any $r \geq 1$, we define the family of Bayes risks for estimating $\mathbf{w}$

  1. from only $\mathbf{B}$ as

$$R_{\mathrm{B}} := \inf_{\varphi} \mathbb{E}\left[\left\| \mathbf{w} - \varphi(\mathbf{B}) \right\|_r^r \right],$$

  where $\varphi(\mathbf{B})$ is an estimator of $\mathbf{w}$.

Bayes Risk

- For any $r \geq 1$, we define the family of Bayes risks for estimating $\mathbf{w}$

  1. from only $\mathbf{B}$ as
  $$R_{\mathrm{B}} := \inf_{\varphi} \mathbb{E}\left[\left\| \mathbf{w} - \varphi(\mathbf{B}) \right\|_r^r \right],$$
  where $\varphi(\mathbf{B})$ is an estimator of $\mathbf{w}$.

  2. from $\mathbf{B}$ and the latent variable $\mathbf{Z}$ as
  $$R_{\mathrm{B}}^* := \inf_{\varphi^*} \mathbb{E}\left[\left\| \mathbf{w} - \varphi^*(\mathbf{B}, \mathbf{Z}) \right\|_r^r \right],$$
  where $\varphi^*(\mathbf{B}, \mathbf{Z})$ is an estimator of $\mathbf{w}$.

Bayes Risk

- For any $r \geq 1$, we define the family of Bayes risks for estimating $\mathbf{w}$

1. from only $\mathbf{B}$ as

$$R_{\mathrm{B}} := \inf_{\varphi} \mathbb{E}\left[\left\| \mathbf{w} - \varphi(\mathbf{B}) \right\|_r^r \right],$$

where $\varphi(\mathbf{B})$ is an estimator of $\mathbf{w}$.

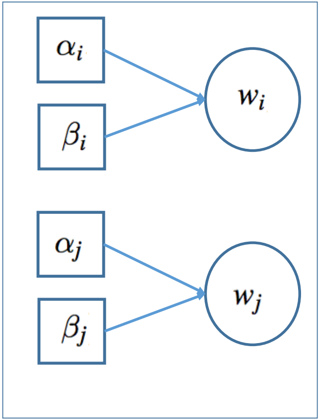
2. from $\mathbf{B}$ and the latent variable $\mathbf{Z}$ as

$$R_{\mathrm{B}}^* := \inf_{\varphi^*} \mathbb{E}\left[\left\| \mathbf{w} - \varphi^*(\mathbf{B}, \mathbf{Z}) \right\|_r^r \right],$$

where $\varphi^*(\mathbf{B}, \mathbf{Z})$ is an estimator of $\mathbf{w}$.

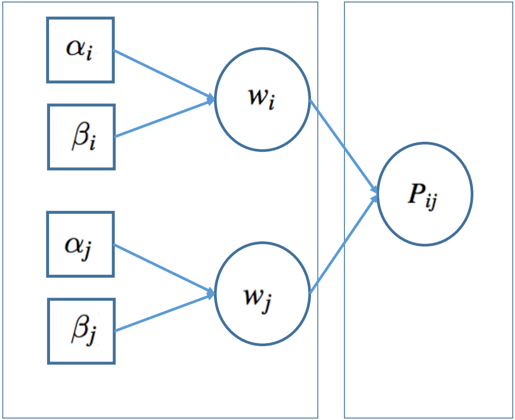
$$R_{\mathrm{B}} \geq R_{\mathrm{B}}^*.$$

Bayesian Network of All Variables



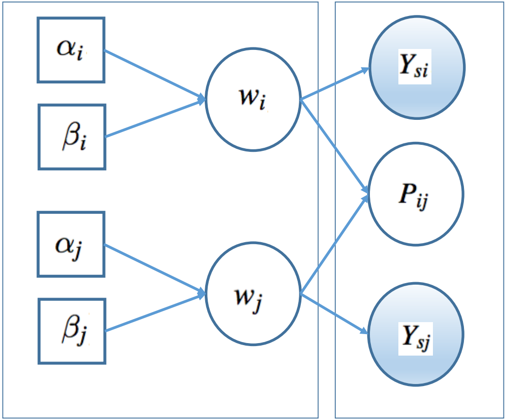
$$w_i \sim \mathrm{Gam}(w_i; \alpha_i, \beta_i) \qquad \text{Prior on } w_i$$

Bayesian Network of All Variables



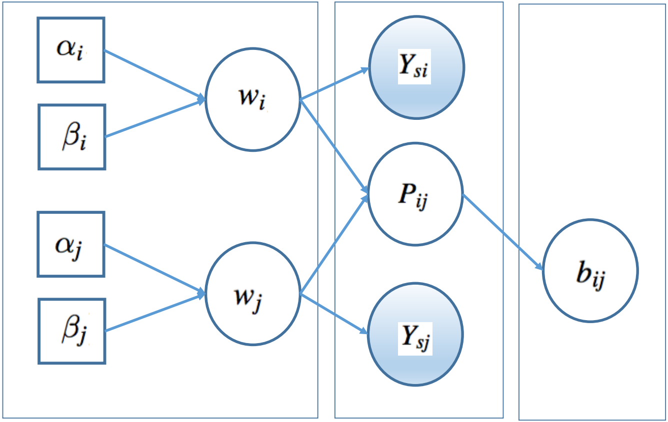
$$P_{ij} = \frac{w_i}{w_i + w_j} \qquad \text{BTL model}$$

Bayesian Network of All Variables



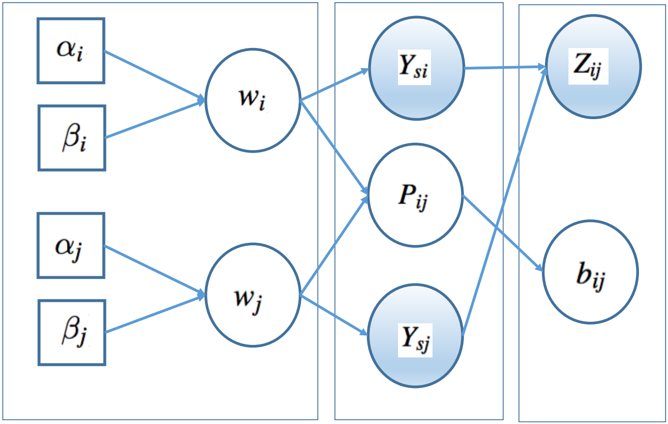
$Y_{si} \sim \text{Exp}(w_i)$ Latent "Arrival Times"

Bayesian Network of All Variables



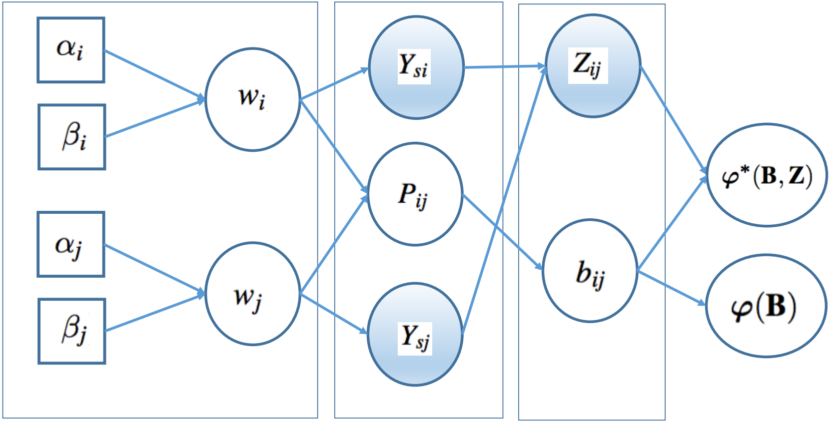
$b_{ij} \sim \text{Bin}\left(b_{ij}; m_{ij}, P_{ij}\right)$ Num of times i beats j out of m_{ij} games

Bayesian Network of All Variables



$$Z_{ij} = \sum_{s=1}^{m_{ij}} \min\{Y_{si}, Y_{sj}\} :$$ Latent variables

Bayesian Network of All Variables



$\varphi(\mathbf{B})$ and $\varphi^*(\mathbf{B}, \mathbf{Z})$: Functions to estimate $\mathbf{w}$

General Lower Bounds on the Bayes Risk

- For $r = 2$, can compute the Bayesian Cramér-Rao bound on R_{B}.

[6] A. Xu and M. Raginsky, "Information-Theoretic Lower Bounds on Bayes Risk in Decentralized Estimation," in IEEE Trans. on Inf. Theory, 2017.

General Lower Bounds on the Bayes Risk

- For $r = 2$, can compute the Bayesian Cramér-Rao bound on R_{B}.
- We compute a family of information-theoretic lower bounds:

[6] A. Xu and M. Raginsky, "Information-Theoretic Lower Bounds on Bayes Risk in Decentralized Estimation," in IEEE Trans. on Inf. Theory, 2017.

General Lower Bounds on the Bayes Risk

- For $r = 2$, can compute the Bayesian Cramér-Rao bound on R_{B}.
- We compute a family of information-theoretic lower bounds:
  1. Theorem 3 of Xu and Raginsky[6] reads: For any $r \geq 1$,

$$R_{\mathrm{B}}^{*} \geq \frac{n}{re}\left(V_n \cdot \Gamma\left(1+\frac{n}{r}\right)\right)^{-r/n} \exp\left[-\frac{r}{n}\big(I(\mathbf{w};\mathbf{B},\mathbf{Z}) - h(\mathbf{w})\big)\right],$$

where V_n is the volume of the unit ball in $(\mathbb{R}^n, \|\cdot\|_r)$.

[6] A. Xu and M. Raginsky, "Information-Theoretic Lower Bounds on Bayes Risk in Decentralized Estimation," in IEEE Trans. on Inf. Theory, 2017.

General Lower Bounds on the Bayes Risk

- For $r = 2$, can compute the Bayesian Cramér-Rao bound on R_{B}.
- We compute a family of information-theoretic lower bounds:
  1. Theorem 3 of Xu and Raginsky[6] reads: For any $r \geq 1$,

$$R_{\mathrm{B}}^{*} \geq \frac{n}{re}\left(V_n \cdot \Gamma\left(1 + \frac{n}{r}\right)\right)^{-r/n} \exp\left[-\frac{r}{n}\left(I(\mathbf{w}; \mathbf{B}, \mathbf{Z}) - h(\mathbf{w})\right)\right],$$

where V_n is the volume of the unit ball in $(\mathbb{R}^n, \|\cdot\|_r)$.

[6] A. Xu and M. Raginsky, "Information-Theoretic Lower Bounds on Bayes Risk in Decentralized Estimation," in IEEE Trans. on Inf. Theory, 2017.

General Lower Bounds on the Bayes Risk

- For $r = 2$, can compute the Bayesian Cramér-Rao bound on R_{B}.
- We compute a family of information-theoretic lower bounds:
  1. Theorem 3 of Xu and Raginsky[6] reads: For any $r \geq 1$,

$$R_{\mathrm{B}}^{*} \geq \frac{n}{re}\left(V_n \cdot \Gamma\left(1+\frac{n}{r}\right)\right)^{-r/n} \exp\left[-\frac{r}{n}\left(I(\mathbf{w};\mathbf{B},\mathbf{Z}) - h(\mathbf{w})\right)\right],$$

     where V_n is the volume of the unit ball in $(\mathbb{R}^n, \|\cdot\|_r)$.
  2. Using Stirling's approximation, we upper bound

$$I(\mathbf{w};\mathbf{B},\mathbf{Z}) - h(\mathbf{w}) = \mathbb{E}\left[\log p(\mathbf{w}|\mathbf{B},\mathbf{Z})\right].$$

[6] A. Xu and M. Raginsky, "Information-Theoretic Lower Bounds on Bayes Risk in Decentralized Estimation," in IEEE Trans. on Inf. Theory, 2017.

Family of Information-Theoretic Lower Bounds

Theorem

For all $i \in [n]$, let

$$m_i := \frac{1}{2} \sum_{j \neq i} m_{ij}. \qquad \text{half the total num. of games } i \text{ plays}$$

Family of Information-Theoretic Lower Bounds

Theorem

For all $i \in [n]$, let

$$m_i := \frac{1}{2} \sum_{j \neq i} m_{ij}. \qquad \textit{half the total num. of games } i \textit{ plays}$$

Then, the Bayes risk is asymptotically lower bounded by

$$R_{\mathrm{B}} \gtrsim \frac{n}{re} \left(V_n \cdot \Gamma\left(1 + \frac{n}{r}\right)\right)^{-r/n} \exp\left[-r\, E(\mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right],$$

where

Family of Information-Theoretic Lower Bounds

Theorem

For all $i \in [n]$, let

$$m_i := \frac{1}{2} \sum_{j \neq i} m_{ij}. \qquad \textit{half the total num. of games } i \textit{ plays}$$

Then, the Bayes risk is asymptotically lower bounded by

$$R_{\mathrm{B}} \gtrsim \frac{n}{re} \left(V_n \cdot \Gamma\left(1 + \frac{n}{r}\right)\right)^{-r/n} \exp\Big[-r\, E(\mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\beta})\Big],$$

where

$$E(\mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\beta}) := \sum_{i=1}^{n} \left(-\frac{1}{2} \log(2\pi) + \log \beta_i - \psi(\alpha_i) + \frac{1}{2} \log(\alpha_i + m_i) \right).$$

Information-Theoretic Lower Bounds

- Take $\alpha_i = \alpha$ and $\beta_i = \beta$, for each $i \in [n]$.

Information-Theoretic Lower Bounds

- Take $\alpha_i = \alpha$ and $\beta_i = \beta$, for each $i \in [n]$.
- For the L^1 norm ($r = 1$),

$$R_{\mathrm{B}}^* \gtrsim \sqrt{\frac{\pi}{2}} \exp\left[-(\log \beta - \psi(\alpha) + 1) \right] \frac{n}{\sqrt{\alpha/n + m}},$$

Information-Theoretic Lower Bounds

- Take $\alpha_i = \alpha$ and $\beta_i = \beta$, for each $i \in [n]$.

- For the L^1 norm ($r = 1$),

$$R_{\mathrm{B}}^* \gtrsim \sqrt{\frac{\pi}{2}} \exp\big[-(\log\beta - \psi(\alpha) + 1)\big] \frac{n}{\sqrt{\alpha/n + m}},$$

- For the squared L^2 norm ($r = 2$),

$$R_{\mathrm{B}} \gtrsim \exp\big[-2(\log\beta - \psi(\alpha)) - 1\big] \frac{n}{\alpha/n + m}.$$

Performance of Lower Bounds: L^1 error

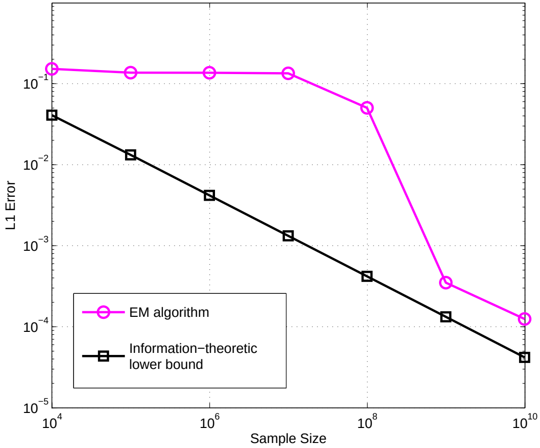


Figure: L^1 error of the EM algo. and the information-theoretic lower bound (for $n = 100$, $\alpha = 5$ and $\beta = \alpha n - 1$).

Perf. of Lower Bounds: MSE (squared L^2 error)

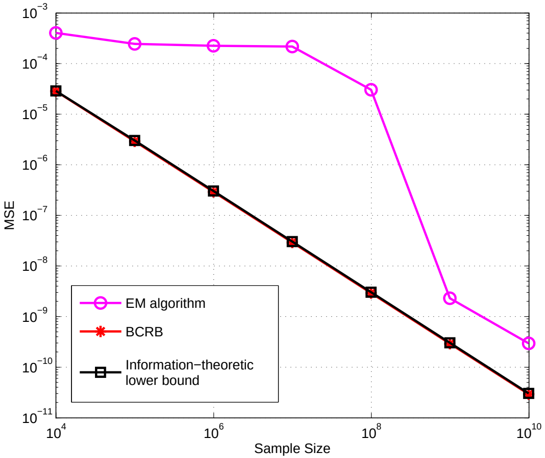


Figure: L^2 error of the EM algo., the IT lower bound and the BCRB (for $n = 100$, $\alpha = 5$, and $\beta = \alpha n - 1$).

Application to General Comparison Graphs

Given a fixed budget of $m = \sum_{i \neq j} m_{ij}$ games,

how to allocate games among n players to minimize the bounds?

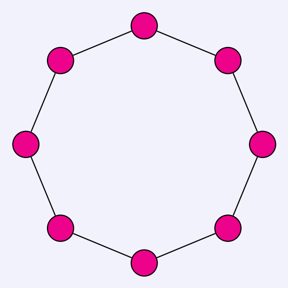
Application to General Comparison Graphs

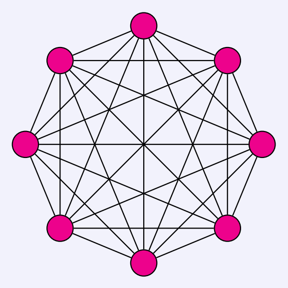
Given a fixed budget of $m = \sum_{i \neq j} m_{ij}$ games,

how to allocate games among n players to minimize the bounds?

Corollary (Optimal Connected Graphs)

Regular Connected Graphs are Optimal!





Application to General Comparison Graphs

Proof:

- Minimizing the lower bound is equivalent to maximizing

$$f(\{m_i\}_{i\in[n]}) := \sum_{i=1}^{n} \frac{1}{2} \log(\alpha_i + m_i)$$

subject to $\sum_{i=1}^{n} m_i = m$ and $m_i \in \mathbb{N}$.

Application to General Comparison Graphs

Proof:

- Minimizing the lower bound is equivalent to maximizing

$$f(\{m_i\}_{i\in[n]}) := \sum_{i=1}^{n} \frac{1}{2} \log\left(\alpha_i + m_i\right)$$

subject to $\sum_{i=1}^{n} m_i = m$ and $m_i \in \mathbb{N}$.

- Solution given by water-filling formula:

$$m_i = |\mu - \alpha_i|_+, \quad \forall\, i \in [n],$$

where $\mu > 0$ is chosen such that

$$\sum_{i=1}^{n} |\mu - \alpha_i|_+ = m.$$

Application to General Comparison Graphs

Proof:

- Minimizing the lower bound is equivalent to maximizing

$$f(\{m_i\}_{i\in[n]}) := \sum_{i=1}^{n} \frac{1}{2} \log\left(\alpha_i + m_i\right)$$

subject to $\sum_{i=1}^{n} m_i = m$ and $m_i \in \mathbb{N}$.

- Solution given by water-filling formula:

$$m_i = |\mu - \alpha_i|_+, \quad \forall\, i \in [n],$$

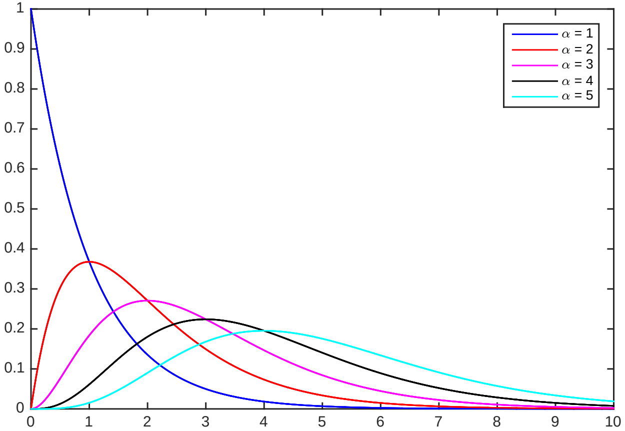
where $\mu > 0$ is chosen such that

$$\sum_{i=1}^{n} |\mu - \alpha_i|_+ = m.$$

- But when $\alpha_i = \alpha$ for all i, m_i are all equal.

The Gamma Distribution with Fixed $\beta = 1$

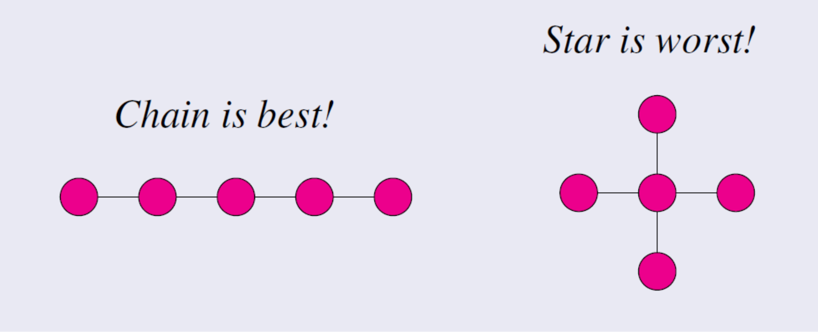
$$\alpha_i \uparrow \implies \text{Greater belief that } w_i \uparrow$$
$$\implies \text{Games } i \text{ plays with others } m_i \downarrow$$



Application to Comparison Graphs: Restricted to Trees

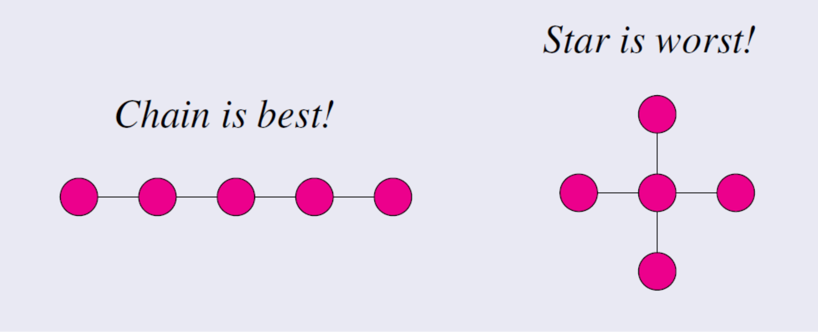
Application to Comparison Graphs: Restricted to Trees

Corollary (Optimal Tree Graphs)



Application to Comparison Graphs: Restricted to Trees

Corollary (Optimal Tree Graphs)



1. Best: Minimizes the (lower bound on the) Bayes Risk
2. Worst: Maximizes the (lower bound on the) Bayes Risk

Application to Comparison Graphs: Restricted to Trees

Proof for Star:

- Maximizing the lower bound on Bayes risk equivalent to

$$\min_{\mathbf{m}:\sum_i m_i = m} g(\mathbf{m}) := \frac{1}{2}\log\left(\alpha + 2m + \sum_{i' \neq i^*} m_{i'}\right) + \sum_i \frac{1}{2}\log(\alpha + m_i)$$

where $\mathbf{m} = \{m_i\}_{i \in [n]}$ and $i^* = 1$ is the central node.

Application to Comparison Graphs: Restricted to Trees

Proof for Star:

- Maximizing the lower bound on Bayes risk equivalent to

$$\min_{\mathbf{m}:\sum_i m_i = m} g(\mathbf{m}) := \frac{1}{2}\log\left(\alpha + 2m + \sum_{i' \neq i^*} m_{i'}\right) + \sum_i \frac{1}{2}\log(\alpha + m_i)$$

  where $\mathbf{m} = \{m_i\}_{i\in[n]}$ and $i^* = 1$ is the central node.

- Shift part of weight of an edge $m_{1j} > 0$, for $j \neq 1$, to create a new edge with weight m_{ji} such that $i \neq 1$. Can show that

$$\frac{\partial g(m_1, \ldots, m_n)}{\partial m_i} > 0$$

  implying that f will be increased by the new configuration.

Effect of Tree Graph Structure on IT Bound

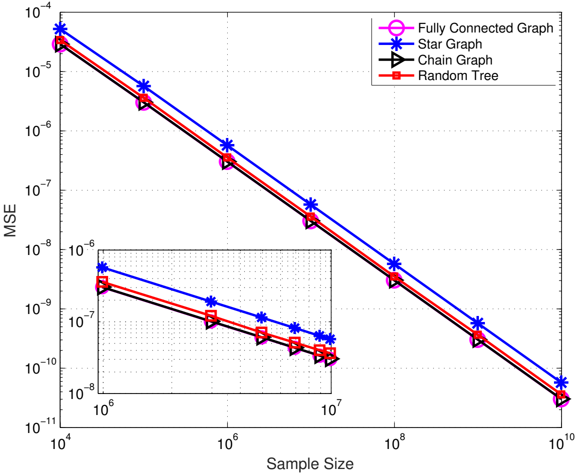


Figure: IT bounds for diff. graph structures (for $n = 100$, $\alpha = 5$, $\beta = \alpha n - 1$).

Effect of Tree Graph Structure on BCRB

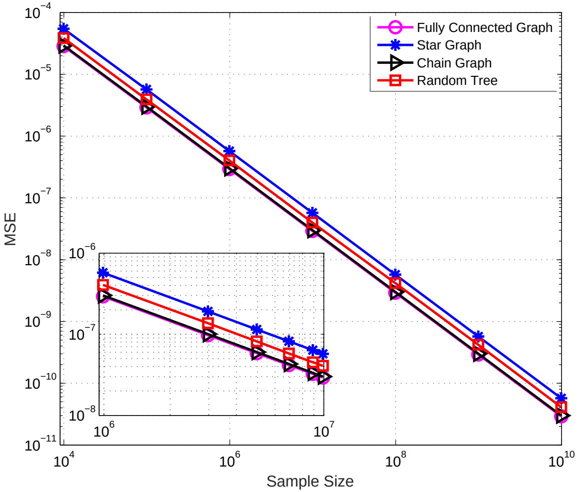


Figure: BCRB for diff. graph structures (for $n = 100$, $\alpha = 5$, $\beta = \alpha n - 1$).

Final Remarks

- Also derived lower bounds for the home-field advantage scenario:

$$P_{ij} = \begin{cases} Q_{ij} := \dfrac{\theta\lambda_i}{\theta\lambda_i + \lambda_j}, & \text{if } i \text{ is home,} \\ \overline{Q}_{ij} := \dfrac{\lambda_i}{\lambda_i + \theta\lambda_j}, & \text{if } j \text{ is home,} \end{cases}$$

where $\theta \in \mathbb{R}_{++}$ models the strength of advantage ($\theta > 1$)

Final Remarks

- Also derived lower bounds for the home-field advantage scenario:

$$P_{ij} = \begin{cases} Q_{ij} := \dfrac{\theta\lambda_i}{\theta\lambda_i + \lambda_j}, & \text{if } i \text{ is home,} \\ \overline{Q}_{ij} := \dfrac{\lambda_i}{\lambda_i + \theta\lambda_j}, & \text{if } j \text{ is home,} \end{cases}$$

  where $\theta \in \mathbb{R}_{++}$ models the strength of advantage $(\theta > 1)$

- Future works: Matching information-theoretic upper bounds

Final Remarks

- Also derived lower bounds for the home-field advantage scenario:

$$P_{ij} = \begin{cases} Q_{ij} := \dfrac{\theta\lambda_i}{\theta\lambda_i + \lambda_j}, & \text{if } i \text{ is home,} \\ \overline{Q}_{ij} := \dfrac{\lambda_i}{\lambda_i + \theta\lambda_j}, & \text{if } j \text{ is home,} \end{cases}$$

  where $\theta \in \mathbb{R}_{++}$ models the strength of advantage $(\theta > 1)$

- Future works: Matching information-theoretic upper bounds
- Other questions related to comparing graph structure, e.g.,

Final Remarks

- Also derived lower bounds for the home-field advantage scenario:

$$P_{ij} = \begin{cases} Q_{ij} := \dfrac{\theta\lambda_i}{\theta\lambda_i + \lambda_j}, & \text{if } i \text{ is home,} \\ \overline{Q}_{ij} := \dfrac{\lambda_i}{\lambda_i + \theta\lambda_j}, & \text{if } j \text{ is home,} \end{cases}$$

  where $\theta \in \mathbb{R}_{++}$ models the strength of advantage $(\theta > 1)$

- Future works: Matching information-theoretic upper bounds
- Other questions related to comparing graph structure, e.g.,

  "Does the fully-connected graph outperform a simple cycle?"

Final Remarks

- Also derived lower bounds for the home-field advantage scenario:

$$P_{ij} = \begin{cases} Q_{ij} := \dfrac{\theta\lambda_i}{\theta\lambda_i + \lambda_j}, & \text{if } i \text{ is home,} \\ \overline{Q}_{ij} := \dfrac{\lambda_i}{\lambda_i + \theta\lambda_j}, & \text{if } j \text{ is home,} \end{cases}$$

  where $\theta \in \mathbb{R}_{++}$ models the strength of advantage $(\theta > 1)$

- Future works: Matching information-theoretic upper bounds
- Other questions related to comparing graph structure, e.g.,

  "Does the fully-connected graph outperform a simple cycle?"

- M. Alsan, R. Prasad and **VYFT**, "Lower Bounds on the Bayes Risk of the Bayesian BTL Model with Applications to Comparison Graphs", IEEE J. on Sel. Topics of Sig. Proc., Oct 2018