Graph Resistance and Learning from Pairwise Comparisons

Alex Olshevsky

Department of ECE, Boston University

Joint work with Julien Hendrickx (UC Louvain) and Venkatesh Saligrama (BU)

Problem Statement

- Given a collection of items with unknown qualities $w_1, \ldots, w_n$, we want to compute $w = (w_1, \ldots, w_n)$ up to scaling from *pairwise* comparisons of items.

Problem Statement

- Given a collection of items with unknown qualities $w_1, \ldots, w_n$, we want to compute $w = (w_1, \ldots, w_n)$ up to scaling from *pairwise* comparisons of items.
- In many contexts, comparisons are the right way to model the available data:

Problem Statement

- Given a collection of items with unknown qualities $w_1, \ldots, w_n$, we want to compute $w = (w_1, \ldots, w_n)$ up to scaling from *pairwise* comparisons of items.

- In many contexts, comparisons are the right way to model the available data:

    - A patient compares how painful or helpful two treatments have been.

Problem Statement

- Given a collection of items with unknown qualities $w_1, \ldots, w_n$, we want to compute $w = (w_1, \ldots, w_n)$ up to scaling from *pairwise* comparisons of items.

- In many contexts, comparisons are the right way to model the available data:

  - A patient compares how painful or helpful two treatments have been.
  - A customer purchases one of several items recommended by an e-commerce site.

Problem Statement

- Given a collection of items with unknown qualities $w_1, \ldots, w_n$, we want to compute $w = (w_1, \ldots, w_n)$ up to scaling from *pairwise* comparisons of items.

- In many contexts, comparisons are the right way to model the available data:

    - A patient compares how painful or helpful two treatments have been.
    - A customer purchases one of several items recommended by an e-commerce site.
    - A user clicks on one of the items suggested by a search engine.

Problem Statement

- Given a collection of items with unknown qualities $w_1, \ldots, w_n$, we want to compute $w = (w_1, \ldots, w_n)$ up to scaling from *pairwise* comparisons of items.

- In many contexts, comparisons are the right way to model the available data:
  - A patient compares how painful or helpful two treatments have been.
  - A customer purchases one of several items recommended by an e-commerce site.
  - A user clicks on one of the items suggested by a search engine.
  - A user chooses one of several movies recommended by a streaming site.

The Simplest Possible Model: BTL over a graph

- Items are compared according to the Bradley-Terry-Luce (BTL) model: probability that item i wins against item j is

$$\frac{w_i}{w_i + w_j}$$

The Simplest Possible Model: BTL over a graph

- Items are compared according to the Bradley-Terry-Luce (BTL) model: probability that item i wins against item j is

$$\frac{w_i}{w_i + w_j}$$

- There are a number of models for item comparisons, and the BTL model is arguably the simplest.

The Simplest Possible Model: BTL over a graph

- Items are compared according to the Bradley-Terry-Luce (BTL) model: probability that item i wins against item j is

$$\frac{w_i}{w_i + w_j}$$

- There are a number of models for item comparisons, and the BTL model is arguably the simplest.
- We assume that there is an underlying "comparison graph" G and if (i, j) is an edge in this graph, items i and j are compared k times.

The Simplest Possible Model: BTL over a graph

- Items are compared according to the Bradley-Terry-Luce (BTL) model: probability that item i wins against item j is

$$\frac{w_i}{w_i + w_j}$$

- There are a number of models for item comparisons, and the BTL model is arguably the simplest.
- We assume that there is an underlying "comparison graph" G and if (i, j) is an edge in this graph, items i and j are compared k times.
- We do not choose the comparison graph.

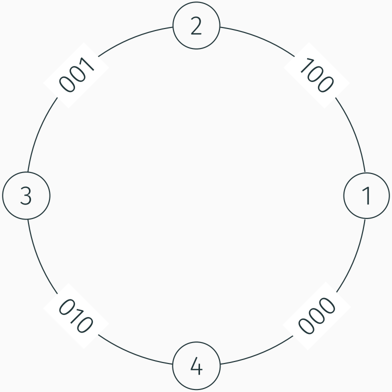
The Simplest Possible Model: BTL over a graph

- Items are compared according to the Bradley-Terry-Luce (BTL) model: probability that item i wins against item j is

$$\frac{w_i}{w_i + w_j}$$

- There are a number of models for item comparisons, and the BTL model is arguably the simplest.
- We assume that there is an underlying “comparison graph” G and if (i, j) is an edge in this graph, items i and j are compared k times.
- We do not choose the comparison graph.
- Goal: understand how fast the error decays with k and G.

Example



- Each edge label represents the outcomes of noisy comparisons.
- Need to compute (scaled versions of) w_1, w_2, w_3, w_4 from these measurements.

Previous Work – I

- The dominant approach has been to construct a Markov chain based on the data whose stationary distribution is an estimate of the true weights.

Previous Work – I

- The dominant approach has been to construct a Markov chain based on the data whose stationary distribution is an estimate of the true weights.
- First proposed by [Dwork, Kumar, Naor, Sivakumar, WWW 2001] and first analyzed [Neghaban, Oh, Shah, NeurIPS 2012]. Under the assumption

$$\max_{i,j} \frac{w_i}{w_j} \leq b,$$

the estimate $\hat{W}$ satisfies

$$\frac{\left\| \frac{w}{||w||_1} - \hat{W} \right\|_2^2}{\left\| \frac{w}{||w||_1} \right\|_2^2} \leq O\left(\frac{1}{k}\right) \frac{b^5 \log n}{\lambda_2^2} \frac{d_{\max}}{d_{\min}^2},$$

Previous Work – I

- The dominant approach has been to construct a Markov chain based on the data whose stationary distribution is an estimate of the true weights.
- First proposed by [Dwork, Kumar, Naor, Sivakumar, WWW 2001] and first analyzed [Neghaban, Oh, Shah, NeurIPS 2012]. Under the assumption

$$\max_{i,j} \frac{w_i}{w_j} \leq b,$$

  the estimate $\hat{W}$ satisfies

$$\frac{\left|\left|\frac{w}{||w||_1} - \hat{W}\right|\right|_2^2}{\left|\left|\frac{w}{||w||_1}\right|\right|_2^2} \leq O\left(\frac{1}{k}\right) \frac{b^5 \log n}{\lambda_2^2} \frac{d_{\max}}{d_{\min}^2},$$

- Worst case scaling is $O(n^7/k)$.

Previous Work – I

- The dominant approach has been to construct a Markov chain based on the data whose stationary distribution is an estimate of the true weights.
- First proposed by [Dwork, Kumar, Naor, Sivakumar, WWW 2001] and first analyzed [Neghaban, Oh, Shah, NeurIPS 2012]. Under the assumption

$$\max_{i,j} \frac{w_i}{w_j} \leq b,$$

  the estimate $\hat{W}$ satisfies

$$\frac{\left|\left| \frac{w}{||w||_1} - \hat{W} \right|\right|_2^2}{\left|\left| \frac{w}{||w||_1} \right|\right|_2^2} \leq O\left(\frac{1}{k}\right) \frac{b^5 \log n}{\lambda_2^2} \frac{d_{\max}}{d_{\min}^2},$$

- Worst case scaling is $O(n^7/k)$.
- Scaling with degrees recently improved by [Agarwal, Patil, Agarwal, ICML 2018].

Previous Work and Motivation

- Computing the maximum likelihood estimator (which can be done in polynomial time) was considered in [Shah, Balakrishnan, Bradley, Parekh, Ramchandran, Wainwright, JMLR 16].

Previous Work and Motivation

- Computing the maximum likelihood estimator (which can be done in polynomial time) was considered in [Shah, Balakrishnan, Bradley, Parekh, Ramchandran, Wainwright, JMLR 16].
- The error bound was

$$O_b\left(\frac{1}{m}\right)\frac{n}{\lambda_2(L)} \geq E\left[\left|\left|\hat{W}-\log w\right|\right|_2^2\right] \geq \Omega_b\left(\frac{1}{m}\right)\max\left(n^2, \max_{l=2,\ldots,n}\sum_{i=\lceil 0.99l\rceil}^{l}\frac{1}{\lambda_i(L)}\right)$$

after m samples, where L is the Laplacian of the comparison graph, and $O_b(\cdot), \Omega_b(\cdot)$ denotes that the constant within the $O(\cdot)$ notation depends on b.

Previous Work and Motivation

- Computing the maximum likelihood estimator (which can be done in polynomial time) was considered in [Shah, Balakrishnan, Bradley, Parekh, Ramchandran, Wainwright, JMLR 16].
- The error bound was

$$O_b\left(\frac{1}{m}\right)\frac{n}{\lambda_2(L)} \geq E\left[\left|\left|\hat{W} - \log w\right|\right|_2^2\right] \geq \Omega_b\left(\frac{1}{m}\right)\max\left(n^2, \max_{l=2,\ldots,n}\sum_{i=\lceil 0.99l\rceil}^{l}\frac{1}{\lambda_i(L)}\right)$$

 after m samples, where L is the Laplacian of the comparison graph, and $O_b(\cdot), \Omega_b(\cdot)$ denotes that the constant within the $O(\cdot)$ notation depends on b.
- Our concern I: we want matching upper and lower bounds.

Previous Work and Motivation

- Computing the maximum likelihood estimator (which can be done in polynomial time) was considered in [Shah, Balakrishnan, Bradley, Parekh, Ramchandran, Wainwright, JMLR 16].
- The error bound was

$$O_b\left(\frac{1}{m}\right)\frac{n}{\lambda_2(L)} \geq E\left[\left|\left|\hat{W} - \log w\right|\right|_2^2\right] \geq \Omega_b\left(\frac{1}{m}\right)\max\left(n^2, \max_{l=2,\ldots,n}\sum_{i=\lceil 0.99l\rceil}^{l}\frac{1}{\lambda_i(L)}\right)$$

  after m samples, where L is the Laplacian of the comparison graph, and $O_b(\cdot), \Omega_b(\cdot)$ denotes that the constant within the $O(\cdot)$ notation depends on b.
- Our concern I: we want matching upper and lower bounds.
- Our concern II: what is the relevant graph-theoretic quantity?

Our results - I

- We give satisfactory answers to these concerns but only when k is large.

Our results - I

- We give satisfactory answers to these concerns but only when k is large.
- The standard way to measure the distance between subspaces is through a sine of the angle:

$$|\sin(\hat{W}, w)| = \inf_{\alpha} \frac{||\alpha \hat{W} - w||_2}{||w||_2}.$$

This same as measures considered above up to factors of b.

Our results - I

- We give satisfactory answers to these concerns but only when k is large.
- The standard way to measure the distance between subspaces is through a sine of the angle:

$$|\sin(\hat{W}, w)| = \inf_{\alpha} \frac{||\alpha \hat{W} - w||_2}{||w||_2}.$$

This same as measures considered above up to factors of b.

- First main result: we give a method such that when $k \geq \Omega\left(|E| \log^2(n/\delta)\right)$, then with probability $1 - \delta$,

$$\begin{aligned} \sin^2(\hat{W}, w) &= O\left(\frac{b^2 R_{\mathrm{max}}(1 + \log(1/\delta))}{k}\right) \\ \sin^2(\hat{W}, w) &= O\left(\frac{b^4 R_{\mathrm{avg}}(1 + \log(1/\delta))}{k}\right), \end{aligned}$$

where $R_{\mathrm{max}}, R_{\mathrm{avg}}$ are, respectively, the maximum and average resistance of the comparison graph.

Our results - II

- Second main result: when $k \geq \sqrt{d_{\mathrm{max}}} n R_{\mathrm{avg}}$,

$$E\left[\sin^2(\hat{W}, w)\right] \geq \frac{R_{\mathrm{avg}}}{k}.$$

Our results - II

- Second main result: when $k \geq \sqrt{d_{\max}} n R_{\text{avg}}$,

$$E\left[\sin^2(\hat{W}, w)\right] \geq \frac{R_{\text{avg}}}{k}.$$

- Punchline: the relevant graph-theoretic quantity is the graph resistance.

Our results - II

- Second main result: when $k \geq \sqrt{d_{\max}} n R_{\text{avg}}$,

$$E\left[\sin^2(\hat{W}, w)\right] \geq \frac{R_{\text{avg}}}{k}.$$

- Punchline: the relevant graph-theoretic quantity is the graph resistance.

- Worst-case for $\sin^2(\hat{W}, w)$ (or other notions of squared distance) is actually $O(n/k)$ when $b = O(1)$.

Our method

- We do the simplest possible thing.

Our method

- We do the simplest possible thing.
- On edge (i, j) let F_{ij} be the fraction of times i wins against j.

Our method

- We do the simplest possible thing.
- On edge (i, j) let F_{ij} be the fraction of times i wins against j.
- Observe that

$$\frac{E[F_{ij}]}{E[F_{ji}]} = \frac{w_i/(w_i + w_j)}{w_j/(w_i + w_j)} = \frac{w_i}{w_j}$$

Our method

- We do the simplest possible thing.
- On edge (i, j) let F_{ij} be the fraction of times i wins against j.
- Observe that

$$\frac{E[F_{ij}]}{E[F_{ji}]} = \frac{w_i/(w_i + w_j)}{w_j/(w_i + w_j)} = \frac{w_i}{w_j}$$

- Our approach: solve the linear system of equations

$$\log \frac{F_{ij}}{F_{ji}} = z_i - z_j,$$

in the least-square sense, and set $\hat{W}_i = e^{z_i}$.

Our method

- We do the simplest possible thing.
- On edge (i, j) let F_{ij} be the fraction of times i wins against j.
- Observe that

$$\frac{E[F_{ij}]}{E[F_{ji}]} = \frac{w_i/(w_i + w_j)}{w_j/(w_i + w_j)} = \frac{w_i}{w_j}$$

- Our approach: solve the linear system of equations

$$\log \frac{F_{ij}}{F_{ji}} = z_i - z_j,$$

  in the least-square sense, and set $\hat{W}_i = e^{z_i}$.
- Can be done in nearly linear time due to work by [Spielman, Teng, 2004].

Why Resistance? The upper bound

- As a toy example, imagine that the comparison graph is a line.

Why Resistance? The upper bound

- As a toy example, imagine that the comparison graph is a line.
- Our method learns something about the ratios $w_1/w_2, w_2/w_3, \ldots, w_{n-1}/w_n$. The squared error in estimating each of these will decay like $1/k$.

Why Resistance? The upper bound

- As a toy example, imagine that the comparison graph is a line.
- Our method learns something about the ratios $w_1/w_2, w_2/w_3, \ldots, w_{n-1}/w_n$. The squared error in estimating each of these will decay like $1/k$.
- Relative errors multiply, e.g.

$$\frac{w_3}{w_1} = \frac{w_2}{w_1}\frac{w_3}{w_2},$$

so if the two quantities on the right are known to some error, those errors will multiply.

Why Resistance? The upper bound

- As a toy example, imagine that the comparison graph is a line.
- Our method learns something about the ratios $w_1/w_2, w_2/w_3, \ldots, w_{n-1}/w_n$. The squared error in estimating each of these will decay like $1/k$.
- Relative errors multiply, e.g.

$$\frac{w_3}{w_1} = \frac{w_2}{w_1}\frac{w_3}{w_2},$$

  so if the two quantities on the right are known to some error, those errors will multiply.
- But $(1+\epsilon)^n \approx 1 + n\epsilon$ when errors are small, the total squared error will scale linearly with n.

Why Resistance? The upper bound

- As a toy example, imagine that the comparison graph is a line.
- Our method learns something about the ratios $w_1/w_2, w_2/w_3, \ldots, w_{n-1}/w_n$. The squared error in estimating each of these will decay like $1/k$.
- Relative errors multiply, e.g.

$$\frac{w_3}{w_1} = \frac{w_2}{w_1}\frac{w_3}{w_2},$$

  so if the two quantities on the right are known to some error, those errors will multiply.
- But $(1+\epsilon)^n \approx 1 + n\epsilon$ when errors are small, the total squared error will scale linearly with n.
- Now imagine an arbitrary graph. Now for any two nodes i and j, we can think about the error over all paths from i to j.

Why Resistance? The upper bound

- As a toy example, imagine that the comparison graph is a line.
- Our method learns something about the ratios $w_1/w_2, w_2/w_3, \ldots, w_{n-1}/w_n$. The squared error in estimating each of these will decay like $1/k$.
- Relative errors multiply, e.g.

$$\frac{w_3}{w_1} = \frac{w_2}{w_1}\frac{w_3}{w_2},$$

  so if the two quantities on the right are known to some error, those errors will multiply.
- But $(1+\epsilon)^n \approx 1 + n\epsilon$ when errors are small, the total squared error will scale linearly with n.
- Now imagine an arbitrary graph. Now for any two nodes i and j, we can think about the error over all paths from i to j.
- Error for each path will scale with length but will decreases when you get to average more paths.

Why Resistance? The upper bound

- As a toy example, imagine that the comparison graph is a line.
- Our method learns something about the ratios $w_1/w_2, w_2/w_3, \ldots, w_{n-1}/w_n$. The squared error in estimating each of these will decay like $1/k$.
- Relative errors multiply, e.g.

$$\frac{w_3}{w_1} = \frac{w_2}{w_1}\frac{w_3}{w_2},$$

  so if the two quantities on the right are known to some error, those errors will multiply.
- But $(1+\epsilon)^n \approx 1 + n\epsilon$ when errors are small, the total squared error will scale linearly with n.
- Now imagine an arbitrary graph. Now for any two nodes i and j, we can think about the error over all paths from i to j.
- Error for each path will scale with length but will decreases when you get to average more paths.
- Clear parallel to resistance.

Why Resistance? The lower bound

- What sort of argument might yield a lower bound of resistance?

Why Resistance? The lower bound

- What sort of argument might yield a lower bound of resistance?
- There is a natural way resistance comes up:

$$R_{\mathrm{avg}} = \frac{\mathrm{Tr}(L^{\dagger})}{n},$$

where L is the graph Laplacian and $L^{\dagger}$ is the Moore-Penrose pseudonverse.

Why Resistance? The lower bound

- What sort of argument might yield a lower bound of resistance?
- There is a natural way resistance comes up:

$$R_{\mathrm{avg}} = \frac{\mathrm{Tr}(L^{\dagger})}{n},$$

where L is the graph Laplacian and $L^{\dagger}$ is the Moore-Penrose pseudonverse.

- One can prove a lower bound by exhibiting $w_1 \neq w_2$ and demonstrating that the expected (total variation) distance between the two distributions on $k|E|$ outcomes is small.

Why Resistance? The lower bound - II

- Choose
$$w = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \frac{1}{\sqrt{k}} \sum_{i=2}^{n} Z_i \frac{v_i}{\sqrt{\lambda_i}},$$
where v_i are the eigenvectors the Laplacian of the comparison graph (normalized so that $||v||_2 = 1$), with λ_i the corresponding eigenvalues, and $Z_i \in \{-1, 1\}$ is a Bernoulli random variable.

Why Resistance? The lower bound - II

- Choose

$$w = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \frac{1}{\sqrt{k}} \sum_{i=2}^{n} Z_i \frac{v_i}{\sqrt{\lambda_i}},$$

  where v_i are the eigenvectors the Laplacian of the comparison graph (normalized so that $||v||_2 = 1$), with λ_i the corresponding eigenvalues, and $Z_i \in \{-1, 1\}$ is a Bernoulli random variable.

- Suppose the error in estimating each Z_i is C, i.e., for any $\widehat{Z}_i$, the error in estimating Z_i satisfies

$$E\left[\left(\hat{Z}_i - Z_i\right)^2\right] \geq C$$

  Then for any $\hat{W}$,

$$E\frac{||\hat{W} - w||_2^2}{||w||_2^2} \geq \frac{C(1/k)\sum_{i=2}^{n} 1/\lambda_i}{n} = \Omega\left(C\frac{\mathrm{Tr}(L^\dagger)}{n}\right) = \Omega\left(CR_{\mathrm{avg}}\right)$$

Why Resistance? The lower bound - II

- Choose

$$w = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \frac{1}{\sqrt{k}} \sum_{i=2}^{n} Z_i \frac{v_i}{\sqrt{\lambda_i}},$$

  where v_i are the eigenvectors the Laplacian of the comparison graph (normalized so that $||v||_2 = 1$), with λ_i the corresponding eigenvalues, and $Z_i \in \{-1, 1\}$ is a Bernoulli random variable.

- Suppose the error in estimating each Z_i is C, i.e., for any $\widehat{Z}_i$, the error in estimating Z_i satisfies

$$E\left[\left(\hat{Z}_i - Z_i\right)^2\right] \geq C$$

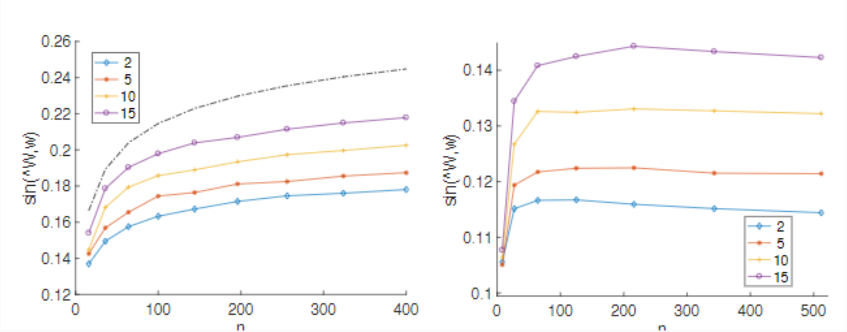
  Then for any $\hat{W}$,

$$E\frac{||\hat{W} - w||_2^2}{||w||_2^2} \geq \frac{C(1/k)\sum_{i=2}^{n} 1/\lambda_i}{n} = \Omega\left(C\frac{\mathrm{Tr}(L^\dagger)}{n}\right) = \Omega\left(CR_{\mathrm{avg}}\right)$$

- Key lemma: C is constant.

Simulations

The following figures show, respectively, evolution on the 2D grid (left, where resistances grows as $O(\log n)$) and 3D grid (right, where resistance is constant).



Conclusion and Future Work

- Our results prove that the squared error decay is $O(R_{\mathrm{avg}}/k)$ for k large enough. Simulations show that this actually seems to be true for all k.

Conclusion and Future Work

- Our results prove that the squared error decay is $O(R_{\mathrm{avg}}/k)$ for k large enough. Simulations show that this actually seems to be true for all k.
- Conjecture: R_{avg} is also the sample complexity of learning in the Bradley-Terry-Luce model.

Conclusion and Future Work

- Our results prove that the squared error decay is $O(R_{\mathrm{avg}}/k)$ for k large enough. Simulations show that this actually seems to be true for all k.
- Conjecture: R_{avg} is also the sample complexity of learning in the Bradley-Terry-Luce model.
- Simulations show that our method performs similarly to Markov chain methods, suggesting that resistance is the right scaling for those methods as well.

Conclusion and Future Work

- Our results prove that the squared error decay is $O(R_{\mathrm{avg}}/k)$ for k large enough. Simulations show that this actually seems to be true for all k.
- Conjecture: R_{avg} is also the sample complexity of learning in the Bradley-Terry-Luce model.
- Simulations show that our method performs similarly to Markov chain methods, suggesting that resistance is the right scaling for those methods as well.
- Getting the correct scaling is still open, as the upper and lower bounds do not match in factors of b as well as in the gap between maximum and average resistance.