Minimax Rate for Learning From Pairwise Comparisons in the BTL Model

Anonymous Authors[1]

Abstract

We consider the problem of learning the qualities $w_1, \ldots, w_n$ of a collection of items by performing noisy comparisons among them. A standard assumption is that there is a fixed "comparison graph" and every neighboring pair of items is compared k times. We will study the popular Bradley-Terry-Luce model, where the probability that item i wins a comparison against j equals $w_i/(w_i + w_j)$. The goal is to understand how the expected error in estimating the vector $w = (w_1, \ldots, w_n)$ behaves in the regime when the number of comparisons k is large.

Our contribution is the determination of the minimax rate up to a constant factor. We show that this rate is achieved by a simple algorithm based on weighted least squares, with weights determined from the empirical outcomes of the comparisons. This algorithm can be implemented in nearly linear time in the total number of comparisons.

1. Introduction

Estimation of item qualities from user preferences is a common problem across multiple domains in e-commerce, health care, and social science. The dominant approach is to rely on raw scores provided by users; for instance, Amazon asks customers for ratings on a scale ranging from 1-5 stars, which are then aggregated to produce an average rating for each item.

Unfortunately, such user-provided scores can be poorly calibrated. Users could differ substantially in how they reach the decision to assign scores; worse, different items could be popular among different classes of users, and these user classes could have statistical differences in the way they assign ratings. It is challenging to deal with this disparity in a principled manner.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

An alternative line of research has explored data fusion based on better calibrated measures, such as the outcomes of comparisons among items. In many contexts, comparison data is readily available. When a user chooses to purchase one of several items recommended by a webpage, it is natural to view this as the outcome of an implicit comparison. The outcome of a sports game can be viewed as the result of a noisy comparison of the strengths of the two teams. Finally, when users click on a particular webpage in response to a list of sites provided by a search engine, this may be viewed as the outcome of a comparison between user estimates of the informativeness of the corresponding webpages. Many additional examples can be given and we refer the reader to (Cattelan, 2012) for an extensive overview of comparison models and their uses.

The simplest and most common model is the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 2012) which posits n items with quality measures $w_1, \ldots, w_n$, with item i winning each comparison against item j independently with probability $w_i/(w_i + w_j)$. All comparisons in this model are pairwise. The BTL model is extremely well-studied; for a sampling of its uses, we mention its applications to an empirical analysis of sports tournaments (Cattelan *et al.*, 2013), measurements of pain among patients (Matthews & Morris, 1995), estimating driver crash risks (Li & Kim, 2000), and testing the power of arguments in referendums (Loewen *et al.*, 2012), among many others.

This paper is concerned with estimating the vector $w = (w_1, \ldots, w_n)$ from the outcomes of comparisons carried out according to the BTL model. It is standard to assume that there is a given undirected graph $G = (\{1, \ldots, n\}, E)$, and every pair of neighbors in G is compared k times. The goal is to recover the vector of true weights vector w. Note, however, that since scaling every entry of w does not change the probability distribution of the outcomes under the BTL model, what can actually be recovered is a normalized version of w.

This problem formulation is standard in the literature; in particular, its analysis has been the subject of a number of recent papers, e.g., (Negahban *et al.*, 2012; Rajkumar & Agarwal, 2014; Negahban *et al.*, 2016; Agarwal *et al.*, 2018; Hendrickx *et al.*, 2019). One can, of course, introduce a number of complicating factors (e.g., more general com-

parisons models, active comparisons, different numbers of comparisons across each edge, simultaneous comparisons of multiple items, etc), and we below survey a number of works analyzing these extensions. However, surprisingly it turns out that, despite literature on the BTL model dating back to the 1950s, many fundamental questions in this simplest setting remain open.

In this paper, we address one of those questions, namely understanding the rate at which the error in the recovery of w decays with the number of comparisons per edge k in terms of the graph G and the true weight vector w.

We will propose an algorithm for the recovery of w based on nonlinearly scaled weighted least-squares. Our main contribution is to show that, up to a constant factor, this algorithm achieves the asymptotic minimax rate for this problem, which we characterize in terms of the trace of a certain matrix depending both on the graph G and the weights w.

1.1. Previous work

The earliest references on the BTL model are (Bradley & Terry, 1952; Rao & Kupper, 1967; Davidson, 1970; Beaver & Gokhale, 1975) dating back to 1950s-1970s. These works focused on maximum likelihood estimation and hypothesis testing. We mention in particular (Beaver, 1977), which proposed doing so with a least squares approach, which is in the same spirit as the method proposed in this paper. The problem was first introduced in the context of internet search in the now-classic paper (Dwork *et al.*, 2001). Several methods for the general class of problems of rank aggregation were proposed in (Dwork *et al.*, 2001), particularly a method based on encoding qualities as the stationary distribution of a Markov chain built from the outcomes of comparisons.

An extremely large literature on analysis of pairwise comparisons has sprung within the statistics and machine learning literature in the past two decade and, a a result, it is not possible to survey all the work that has been done. There are many variations of the problem that have been studied, from more sophisticated models such as Thurstone and Placket-Luce (Hajek *et al.*, 2014; Maystre & Grossglauser, 2015), to online or bandit versions (Szörényi *et al.*, 2015; Yue *et al.*, 2012), to models with active learning (Jamieson & Nowak, 2011; Ailon, 2012), to models with multiple users with potentially different preferences among items (Wu *et al.*, 2015). We next focus only on papers most directly related to our work, namely papers concerned with rates for recovery of the true weights w in the BTL model.

The first rigorous analysis of the error rate in the pairwise case appeared in (Negahban *et al.*, 2012) in the case of a random comparison graph and in (Negahban *et al.*, 2016) for an arbitrary graph. The underlying method recovered an estimate $\hat{W}$ from the stationary distribution of a Markov chain constructed based on the outcomes of the comparisons. By construction, the elements of $\hat{W}$ summed to one, which made it natural to compare $\hat{W}$ with the normalized version of the true weights $w/||w||_1$.

It was shown in (Negahban *et al.*, 2016) that, for a number of comparisons k large enough as a function of the graph G, assuming that the weight imbalance is bounded as

$$\max_{i,j} \frac{w_i}{w_j} \leq b, \tag{1}$$

then with high probability we have that

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

where $d_{\max}, d_{\min}$ are the largest/smallest degrees in the comparison graph and $1-\rho$ is the spectral gap of the random walk on the comparison graph G.

To understand how this scales in terms of the number of nodes n, we can use the results of (Landau & Odlyzko, 1981) which show that $1/(1-\rho)$ for a simple random walk on any graph will have worst-case scaling of $O(n^3)$. Thus the right-hand side above has a worst-case scaling of $O(n^7 \log n)/k$.

To our knowledge (Negahban *et al.*, 2012; 2016) represent the first understanding of how error bounds for w scale in terms of the corresponding graph. A consequence of those results is that a good approximation to the (scaled) true weights w can be found using a polynomial number of samples. Moreover, the results of (Negahban *et al.*, 2016) suggest a natural open problem: to understand just how fast the error decays for the best possible method.

Recently the bounds of (Negahban *et al.*, 2016) were recently improved in (Agarwal *et al.*, 2018), resulting in a better scaling with b and replacing $d_{\text{avg}}/d_{\min}$ with $d_{\text{avg}}/d_{\min}$, among other improvements. Moreover, improved bounds in the somewhat more restrictive setting when comparisons are made over the complete graph, but with each pair of edges sampled independently (at positive rates which could differ across edges) were obtained in (Rajkumar & Agarwal, 2014).

Considerably more general models of ranking are quite common in the literature; in particular, we mention the papers (Rajkumar & Agarwal, 2016; Shah *et al.*, 2016; Negahban *et al.*, 2018), discussed next. In (Rajkumar & Agarwal, 2016), the class of ranking models learnable from a random comparison graph G with average degree that scales as $\log(n)$ was studied, and it was shown that this possible under a certain "low-rank" condition on the underlying model. In (Shah *et al.*, 2016) namely estimating w under a general ranking model parametrized by a nonlinear function which

included the BTL model was a special case. Adopting the normalization condition $\sum_{i=1}^n \log w_i = 0$, upper and lower bounds were shown in (Shah *et al.*, 2016) after m comparisons for $E\left|\left|\hat{W} - \log w\right|\right|_2^2$; the upper bound scaled with $(n/m)\lambda_2(L)^{-1}$, where L is the Laplacian of the comparison graph, and the lower bound had a complicated dependence on the Laplacian spectrum. In (Negahban *et al.*, 2018) upper and lower bounds depending on the Laplacian spectrum were derived for the multinomial logit model, which is much more general than the BTL model.

For the BTL model specifically, progress towards the best rate was made in the recent paper (Hendrickx *et al.*, 2019). The error measure considered in that paper was the sine of the angle made by $\hat{W}$ and w, which can be expressed as

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

The sine of the angle is a standard way to measure distance between subspaces and, as the above identity suggests, it can be thought of as the relative distance between w and the best normalized version of $\hat{W}$. Moreover, because $\sin(\theta) \approx \theta$ for small θ, this error measure is essentially the same (provided the number of samples is large) as measuring the angle between $\hat{W}$ and w. Additionally, as remarked in (Hendrickx *et al.*, 2019) it can be shown that

$$\frac{1}{\sqrt{2}}\left|\left|\frac{x}{||x||_2} - \frac{y}{||y||_2}\right|\right|_2 \leq |\sin(x,y)| \leq \left|\left|\frac{x}{||x||_2} - \frac{y}{||y||_2}\right|\right|_2$$

so that the sine is, up to a constant, the error in the two-norm after normalization. Finally, the sine is also equivalent, up to polynomial factors of b, to previous metrics used in this problem. In particular, it was shown in (Hendrickx *et al.*, 2019) that the sine is within a $\sqrt{b}$ multiplicative factor of the norm used in (Negahban *et al.*, 2012) (see left-hand side of Eq. (2) and it can be shown it is within a polynomial factor of $E\left|\left|\hat{W} - \log w\right|\right|_2^2$ used in (Shah *et al.*, 2016).

Upper and lower bounds were established (Hendrickx *et al.*, 2019) on $\sin^2(\hat{W}, w)$, both holding when the number of samples per edge k is large enough. As far as upper bounds, it was shown that, for large enough k as a function of G and δ, with probability $1-\delta$ we have the bounds

$$\sin^2(\hat{W}, w) = O\left(\frac{b^2 R_{\max}(1 + \log(1/\delta))}{k}\right) \quad (3)$$

$$\sin^2(\hat{W}, w) = O\left(\frac{b^4 R_{\text{avg}}(1 + \log(1/\delta))}{k}\right), \quad (4)$$

where $R_{\text{avg}}, R_{\max}$ are, respectively, the average and largest electrical resistance[1] of the comparison graph G. A corresponding lower bound was proved showing that, for large enough k as a function of the graph G,

$$E\left[\sin^2(\hat{W}, w)\right] \geq \frac{R_{\text{avg}}}{k}. \quad (5)$$

These results come close, but do not quite characterize, the asymptotic minimax rate. Putting all the bounds together, it becomes clear that the electrical resistance is the key graph-theoretic quantity. However, there are gaps between the upper and lower bounds, both in terms of scaling with b and in terms of the difference between average and maximum resistance[2].

1.2. Our contribution

The purpose of the present paper is to present a new algorithm, coupled with new upper and lower bounds, which characterize the minimax rate for this problem (using the sine as a measure of distance). We will need the following definition: we set

$$\gamma(i,j) = \frac{1}{(w_i + w_j)^2},$$

and we use L_γ to mean the Laplacian of the graph G where edge (i,j) has weight $\gamma(i,j)$. We next state two theorems, the first providing an upper bound and the second providing a lower bound, which are the main results of this paper.

Theorem 1. *For large enough k, there is a polynomial-time method which produces a estimate $\hat{W}$ which satisfies*

$$E\left[\sin^2(\hat{W}, w)\right] \leq O\left(\frac{1}{k}\right)\frac{\text{Tr}\left(L_\gamma^\dagger\right)}{||w||_2^2}, \quad (6)$$

where $L_\gamma^\dagger$ refers to the Moore-Penrose pseudoinverse of L_γ. The method which accomplishes this is the WLSM described in Section 2.

Theorem 2. *For any algorithm which constructs an estimate $\hat{W}$ only from the outcomes of k comparisons across each edge[3] of G, we have that for large enough k*

$$E\left[\sin^2(\hat{W}, w)\right] \geq \Omega\left(\frac{1}{k}\right)\frac{\text{Tr}\left(L_\gamma^\dagger\right)}{||w||_2^2}, \quad (7)$$

where $L_\gamma^\dagger$ refers to the Moore-Penrose pseudoinverse of L_γ.

[1] Resistances are defined in terms of the circuit obtained by replacing every edge in a graph by a resistor of unit resistance.

[2] There is also a gap in terms of the $\log(1/\delta)$ factor present in Eq. (3) and Eq. (4) but not in Eq. (5). However, this gap is not important, as can be expected to go away when integrating the high-probability bounds of Eq. (3) and Eq. (4) over δ to obtain a bound on the expectation.

[3] For a formal definition of what it means for an estimator to only depend on the outcomes of the comparisons, see Chapter 8.7 of (Van der Vaart, 2000); the rate proved in Theorem 2 is a "local minimax rate" in the sense of that chapter.

Up to the differences between the constants in $O(\cdot)$ and $\Omega(\cdot)$ notation, these two results characterize the minimax rate. We note these are absolute constants, i.e., they do not depend on any of the problem parameters, and in particular, they do not depend on b. Finally, we remark that it is easy to derive both the upper and lower bounds of (Hendrickx *et al.*, 2019) from these theorems using the well-known fact that the average graph resistance is proportional to the trace of the Laplacian pseudoinverse (see e.g., (Vishnoi, 2013))[4].

1.3. Remainder of this paper

We give an informal presentation of our algorithm in Section 2. In the following Section 3 we conduct some simulations which lead to two further conjectures regarding our algorithms and the previous literature. We draw some conclusions in Section 4. Proofs of Theorem 1 and Theorem 2 are provided in the supplementary information.

2. Our approach

The underlying intuition of approach is best explained by using a series of non-rigorous approximations. While our method will be formally analyzed in the supplementary information, in this section we make free use of such approximations.

For every pair of neighbors i, j in G, we will use F_{ij} to denote the fraction of times node i wins the comparisons against its neighbor j. It will be helpful sometimes to turn G into a directed graph by orienting every edge arbitrarily; we will use $\overrightarrow{E}$ to refer to the edge set of this directed graph.

Across each edge, we also define the ratio $R_{ij} = F_{ij}/F_{ji}$ which captures the imbalance between item qualities across the edge (i, j); indeed, by the strong law of large numbers,

$$R_{ij} \to \frac{w_i/(w_i + w_j)}{w_j/(w_i + w_j)} = \frac{w_i}{w_j}, \tag{8}$$

where the convergence would happen with probability one if we were to take the number of comparisons $k \to \infty$.

Our goal is to figure out the weights w_i from knowledge of the quantities R_{ij} for large but nevertheless finite k. One approach is to take the logarithm of both sides of Eq. (8) to obtain that

$$\log R_{ij} \approx \log w_i - \log w_j, \quad \text{for all edges } (i, j) \in \overrightarrow{E}.$$

The $\approx$ symbol hides the error that occurs from taking k finite. This is now a linear system of equations in the quantities $\log w_i$, so a natural approach is to solve the collection of equations

$$\log R_{ij} = z_i - z_j, \quad \text{for all edges } (i, j) \in \overrightarrow{E},$$

in the least-squares sense. In other words, we can try to find

$$z^* = \arg \min_{z_1,\dots,z_n} \sum_{(i,j)\in E} (\log R_{ij} - (z_i - z_j))^2. \tag{9}$$

We can then build an estimator $\hat{W}$ of the item quality vector w by setting $\hat{W}_i = e^{z_i^*}$. This is exactly what is done in (Beaver, 1977; Hendrickx *et al.*, 2019) and broadly similar to the approach taken earlier in (Jiang *et al.*, 2011).

One disadvantage of this algorithm is that it does not take into consideration the differences in variance in comparisons across different edges. Indeed, observe that the variance in outcomes in comparing items i and j depends on the weights w_i and w_j. Furthermore, if the variance across an edge (i, j) is relatively low, then the corresponding squared term in Eq. (9) should have higher weight. This motivates a weighted least squares approach: we will divide each term in Eq. (9) by the standard deviation of $\log R_{ij}$.

In general, the standard deviation of $\log R_{ij}$ does not have a simple formula, but when k is large we can repeatedly write Taylor expansions of all quantities involved to turn everything approximately linear. The calculation is relatively simple and we perform it in the next few paragraphs; the uninterested reader may feel free to skip ahead to Eq. (11) to see the outcome.

Defining $\rho_{ij} = w_i/w_j$ to be the true ratio between qualities of items i and j, we can use the fact that $(\log x)' = 1/x$ to

[4] Indeed, the relation referred to is

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

where L is the plain (unweighted) graph Laplacian (see e.g., (Vishnoi, 2013)). Thus taking $w = (1, \dots, 1)$ in Theorem (2) we immediately recover Eq. (5). Similarly, rescaling w so that $\min_i w_i = 1$ we have that Eq. (1) implies that $\max_i w_i \le b$, and using the implications $||w||_2^2 \ge n$ and $(w_i + w_j)^2 = O(b^2)$, Theorem 1 immediately implies an upper bound of $O(b^2 R_{\text{avg}}/k)$, actually improving upon both Eq. (3) and Eq. (4).

write

$$\begin{aligned}\log R_{ij} &\approx \log \rho_{ij} + \frac{1}{\rho_{ij}}(R_{ij} - \rho_{ij}) \\ &= \log \rho_{ij} + \frac{1}{\rho_{ij}}\left(\frac{F_{ij}}{F_{ji}} - \rho_{ij}\right) \\ &= \log \rho_{ij} + \frac{1}{\rho_{ij}}\left(\frac{1 - F_{ji}}{F_{ji}} - \rho_{ij}\right) \\ &= \log \rho_{ij} + \frac{1}{\rho_{ij}}\left(\frac{1}{F_{ji}} - 1 - \rho_{ij}\right) \\ &\approx \log \rho_{ij} + \frac{1}{\rho_{ij}}\left(\frac{1}{p_{ji}} - \frac{1}{p_{ji}^2}(F_{ji} - p_{ji}) - \rho_{ij}\right),\end{aligned}$$

where $p_{ji} = w_j/(w_i + w_j)$ is the correct probability of j winning against i, and the final step takes the linear Taylor approximation of $1/F_{ji}$ around its limit of $1/p_{ji}$.

The advantage of this string of manipulations is that it implies

$$\begin{aligned}\mathrm{var}(\log R_{ij}) &\approx \frac{1}{\rho_{ij}^2}\frac{1}{p_{ji}^4}\mathrm{var}(F_{ji} - p_{ji}), \\ &= \frac{w_i/w_j + w_j/w_i + 2}{k},\end{aligned}$$

where the last step follows by some simple algebraic manipulations. For simplicity, let us define

$$v_{ij} = \frac{w_i}{w_j} + \frac{w_j}{w_i} + 2. \tag{10}$$

Then what we really should do is solve the weighted least squares problem

$$\arg\min_{z_1,\ldots,z_n} \sum_{(i,j)\in\vec{E}} \frac{(\log R_{ij} - (z_i - z_j))^2}{\sqrt{v_{ij}/k}} \tag{11}$$

which properly accounts for the different variances of different comparisons. Indeed, observe each term in Eq. (11) now has the same variance as k gets large. Naturally, we can omit $1/\sqrt{k}$ from the denominator since it multiplies every term.

The big problem with this approach, of course, is that the quantities v_{ij} are actually unknown to us because we (obviously) do not know the true weights $w_1, \ldots, w_n$ a-priori. Thus as written Eq. (11) cannot be implemented.

Nevertheless, even though we do not know the quantities v_{ij}, we can construct estimates of them based on the data. Glancing at Eq. (10), a natual approach is to define

$$\hat{V}_{ij} = \frac{F_{ij}}{F_{ji}} + \frac{F_{ji}}{F_{ij}} + 2. \tag{12}$$

Indeed, if we consider what happens if we were to take $k \to \infty$, it follows from the strong law of large numbers that $\hat{V}_{ij} \to v_{ij}$ with probability one. Thus we simply replace each v_{ij} in Eq. (11) by its estimated counterpart:

$$z^* = \arg\min_{z_1,\ldots,z_n} \sum_{(i,j)\in\vec{E}} \frac{(\log R_{ij} - (z_i - z_j))^2}{\sqrt{\hat{V}_{ij}}}. \tag{13}$$

As before, constructing the estimator $\hat{W}$ will be done by setting $\hat{W}_i = e^{z_i^*}$.

We need to take one final step to have a well-defined algorithm. Clearly, we could have a problem when some $F_{kl} = 0$ because then we might run into the problem of using $\log R_{kl} = \log 0 = -\infty$ in our least squares objective. We resolve this problem by setting F_{kl} to be some small positive number (specifically, $F_{kl} = (1/2)/k$) in this case. Intuitively, when k is sufficiently large compared to b, the probability that some $F_{kl} = 0$ is exponentially small, so it doesn't really matter what we do; nevertheless, we need to do something in order to have a well-defined method.

This is the algorithm we will analyze in the remainder of this paper. We state it formally in the algorithm box below. We will refer to it as the Weighted Least Squares Method, or the WLSM for short.

Algorithm 1 Weighted Least Squares Method

Input: results of k independent comparisons across each edge in E.
for all $(i, j) \in E$ **do**
  Compute F_{ij}, the fraction of times item i wins.
  if $F_{ij} = 0$ **then**
    Set $F_{ij} = (1/2)/k$.
  else if $F_{ij} = 1$ **then**
    Set $F_{ij} = 1 - (1/2)/k$.
  end if
  Set $R_{ij} = F_{ij}/F_{ji}$.
end for
Compute the quantities $\hat{V}_{ij}$ using Eq. (12).
Solve Eq. (13) for the vector z^*.
For all $i = 1, \ldots, n$, set $\hat{W}_i = e^{z_i^*}$.

Finally, as we discuss in the Supplementary Information, this algorithm can be implemented in nearly linear time in the number of edges of G.

Unfortunately, the final procedure we have ended up with involves taking the ratio of two random variables constructed from the data (i.e., the quantities $\log R_{ij}$ and $\hat{V}_{ij}$), which will make analysis of the error in *expectation* challenging. To preview the analsysis of this method (which is available in the supplementary information) we will need to perform a large-deviations analysis of the outcome of the WLSM, which will then need to be integrated to obtain a bound on the expectation of $\sin^2(\hat{W}, w)$.

2.1. Linear time solvability

We now rewrite our algorithm in compact form; this rewriting will be needed later section to discuss the technical novelty in the proof, and will also, as a consequence, show that our algorithm can be implemented in (nearly) linear time.

First we discuss some notation. We let $\overrightarrow{E}$ denote the set of directed edges obtained by orienting every edge in E arbitrarily. We let M be the edge-vertex incidence matrix of the resulting directed graph $(\{1,\dots,n\},\overrightarrow{E})$; note that the graph Laplacian L satisfies $L = MM^T$. The quantities L_V and $L_{\hat{V}}$ correspond to weighted graph Laplacians, where the edge $(i,j) \in E$ is weighted by v_{ij}^{-1} or $\hat{V}_{ij}^{-1}$, respectively. We will omit the subscripts when we stack the above quantities into vectors. For example, the notation R represents the vector in $\mathbb{R}^{|\overrightarrow{E}|}$ obtained by stacking up the quantities $R_{ij}, (i,j) \in \overrightarrow{E}$.

With this notation in place, inspecting Eq. (13), we see that z^* is a least squares solution to the system of equations

$$\hat{V}^{-1/2} M^T z = \hat{V}^{-1/2} \log R.$$

Recall that $\hat{W}$ is our notation for $\hat{W} = e^{z^*}$; thus $\log \hat{W}$ is the least squares solution of

$$\hat{V}^{-1/2} M^T \log W = \hat{V}^{-1/2} \log R.$$

Writing out the least-squares solution explicitly, we have that $\log \hat{W}$ is an exact solution of the equation

$$(\hat{V}^{-1/2}M^T)^T \hat{V}^{-1/2} M^T \log \hat{W} = (\hat{V}^{-1/2}M^T)^T \hat{V}^{-1/2} \log R,$$

that is, of

$$L_{\hat{V}} \log \hat{W} = M \hat{V}^{-1} \log R. \tag{14}$$

Thus we have that one solution to Eq. (14) is

$$\log \hat{W} = L_{\hat{V}}^{\dagger} M \hat{V}^{-1} \log R. \tag{15}$$

Observe that since, by connectivity of G the null space of $L_{\hat{V}}$ is just $\mathrm{span}\{\mathbf{1}\}$, this picks up the solution of Eq. (14) which satisfies $\sum_{i=1}^n \log \hat{W}_i = 0$ or $\prod_{i=1}^n \hat{W}_i = 1$.

Concluding, we see that Eq. (15) is one way to represent a solution $\hat{W}$ we seek to compute. We can now observe that this is a Laplacian linear system, i.e., it requires multiplication by the pseudoinverse of a weighted graph Laplacian. We can now directly apply the results of (Spielman & Teng, 2014), which showed that it is possible to solve Eq. (15) in nearly-linear time, specifically in $O(n \log^c |E| \log(1/\epsilon))$ to accuracy ϵ.

Linear time solvability is important in the context of ranking from comparisons because it allows the underlying algorithm to potentially scale up to very large data-sets, such as those built from counts of web activity (i.e., clicks) or from systems with millions of users and many times that comparisons.

2.2. Main innovation in the proof

At a general level, there is a natural way to try to prove the main results of this paper: on the one hand, there are a variety of "two point estimates" which lower bound the expected error by finding pairs of weights that are as different as possible while giving rise to similar distributions on outcomes; more sophisticated approaches do the same over a distribution of weights. In the reverse direction, we can do a large deviations analysis of Eq. (15), which will involve having an accurate analysis of the behavior of a pseudoinverse of a random matrix. Once the lower and upper bounds obtained this way match, the optimal error rate will have been found. Most of the previous literature on the subject, e.g., (Negahban *et al.*, 2016; Shah *et al.*, 2016; Hendrickx *et al.*, 2019) used such an approach to derive upper or lower bounds.

Unfortunately, this appears to be difficult to carry out directly. Our analysis relies on a "trick" of analyzing a suitably regularized version of the problem, which we informally describe next; the full proof is of course available in the supplementary information.

Our starting point is Eq. (14), which shows that $\log \hat{W}$ is a solution of

$$L_{\hat{V}} \log W = M \hat{V}^{-1} \log R. \tag{16}$$

Of course, this equation has many solutions as $\mathbf{1}$ belongs to the null space of $L_{\hat{V}}$. The previous section defined a solution $\hat{W}$ of this equation, which was the solution with the elements of $\log \hat{W}$ summing to zero. For our analysis, we will find it convenient to "pick out" a different solution of Eq. (16). We proceed as follows.

First, we multiply both sides of Eq. (16) by $\mathrm{diag}(w)^{-1}$:

$$\mathrm{diag}(w)^{-1} L_{\hat{V}} \log W = \mathrm{diag}(w)^{-1} M \hat{V}^{-1} \log R.$$

We next introduce a new variable Y and reparametrize $\log W = \mathrm{diag}(w)^{-1} Y$ so that the last equation can be rewritten more symmetrically as

$$\mathrm{diag}(w)^{-1} L_{\hat{V}} \mathrm{diag}(w)^{-1} Y = \mathrm{diag}(w)^{-1} M \hat{V}^{-1} \log R.$$

As before, this equation has many solutions and we pick one arguably the most "natural" one by setting

$$\hat{Y} = \left(\mathrm{diag}(w)^{-1} L_{\hat{V}} \mathrm{diag}(w)^{-1}\right)^{\dagger} \mathrm{diag}(w)^{-1} M \hat{V}^{-1} \log R, \tag{17}$$

Our analysis will proceed by analyzing the quantity $\hat{Y}$. Naturally, we can use the relation $\log W = \mathrm{diag}(w)^{-1} Y$ to obtain that the quantity $\mathrm{diag}(w)^{-1}\hat{Y}$ is a solution of Eq. (16). It will be helpful to introduce new notation for the latter quantity:

$$\log \hat{W}^r = \mathrm{diag}(w)^{-1} \hat{Y}. \tag{18}$$

The quantity $\hat{W}^r$ is, of course, a rescaled version of $\hat{W}$ (because all solutions of Eq. (15) are rescaled versions of $\hat{W}$). It is possible to be more precise and observe that since the null space of $\mathrm{diag}(w)^{-1}L_{\hat{Y}}\mathrm{diag}(w)^{-1}$ is span of w, we have that $\hat{Y}$ is orthogonal to w; which implies that $\log \hat{W}^r$ is orthogonal to w^2 (where the square is understood elementwise) or

$$\prod_{i=1}^{n}(\hat{W}^r)_i^{w_i^2} = 1. \tag{19}$$

Observe that, because we do not know the true weights w, we cannot compute $\hat{W}^r$. Nevertheless, we can still consider it and analyze its properties, and whatever upper and lower bounds we obtain for the sine of the angle between $\hat{W}^r$ and w will apply to the solutions we can actually compute, since the angle between two vectors is unchanged if one of them is scaled. *It turns out that minimax optimal bounds come out of the analysis only after analyzing the solution $\hat{W}^r$ defined in Eq. (19).* Attempts to match the most straightforward upper and lower bounds for other solutions of Eq. (16) result in upper and lower bounds that do not match. This is the main proof ingredient present in this paper that was not used in earlier works.

Note that analyzing the quantity $\hat{W}^r$ is the same as analyzing the solution $\log \widehat{W}$ of the underlying least-squares problem of Eq. (16) with smallest norm relative to the inner product $\langle x, x\rangle_w = \sum_{i=1}^n w_i^2 x_i^2$. Our approach may thus be viewed as part of a long line of research suggesting that the key is often to choose a metric that is natural for the problem. It is analysis with respect to this (scaled) inner product that ultimately leads to the weighted Laplacian L_γ appearing in our main results and not the ordinary Laplacian L.

3. Simulations and Two Conjectures

We perform a number of experiments designed to gauge the accuracy of the WLSM relative to competing methods. Because we are not aware of any real data sets involving comparisons where the true weights are known, we will use synthetic data. As we will see shortly, two conjectures are suggested by our results. We simulate five different methods:

1. The least-squares method. This is the method that solves Eq. (9) for z^* and then sets $\hat{W}_i = e^{z_i^*}$. In the figures below, it is abbreviated as "LS."

2. Least squares with artificial weights. This solves for z^* using Eq. (11) and then sets $\hat{W}_i = e^{z_i^*}$ as above. It cannot be implemented in practice because we do not know the true variances v_{ij} used in Eq. (11), but it can be used as a useful benchmark to measure degradation in performance from using estimates of these variances. This is abbreviated "artif weight" in the figures.

3. Iterative least squares. This method begins by solving Eq. (11) by setting $v_{ij} = 1$. It then uses the computed w_{ij} to compute v_{ij} using Eq. (10), and then proceeds to re-solve Eq. (11). This cycle (new w_{ij} leading to new v_{ij} then leading to new w_{ij}) is then repeated. This is abbreviated by "iter weight" in the figures.

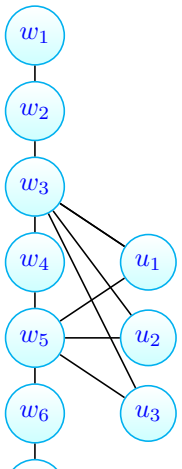
4. Our main algorithm, the WLSM method, which is abbreviated with "emp weight" in the figures.

5. The eigenvector-based algorithm of (Negahban *et al.*, 2012; 2016).

In general, we do not see much of a difference between any of the methods on simple graphs. Representative results are shown in Figure 1 for the 2D grid, the 3D grid, and the Erdos-Renyi random graph. While the method we propose in this paper is usually the best, the gains are extremely modest in the neighborhood of a few percent, as can be eyeballed from the figures. Only three graphs are shown because the pattern is the same on all graphs we have simulated.

However, with some experimentations we have found that the WLSM (along with other least-squares methods) has a significant advantage as compared against the eigenvector-based method in terms of accurately recovering *all* the weights, especially when there are many nodes of small weight. We give one example of such a graph in Figure 2. We take a line graph, pick two nodes that are a neighbor appart, and connect them through a complete bipartite graph with newly introduced nodes (on the right-hand side of the figure). The key idea is that the nodes on the right-hand side (labeled u_1, u_2, u_3 in the figure) will be assigned weight w_i of of 1, while the nodes on the left hand side will have weights that increase geometrically from 1 to b. Thus, for large b, the nodes $u_1, \ldots, u_3$ are not very relevant to (any notion of) distance between normalized versions of $\hat{W}$ and w due to their comparatively small weights. However, neglecting them has the effect of neglecting a large number of paths between w_3 and w_5 which can be used to help estimate the weights on the left-hand side.

Figure 3 shows the difference between $\hat{W}_3 - \hat{W}_5$ when $w_3 = w_5$ and there are approximately 50 nodes u_i on the right-hand side. We compare the difference $\hat{W}_3 - \hat{W}_5$ for both the WLSM and the eigenvector-based method of (Negahban *et al.*, 2012;



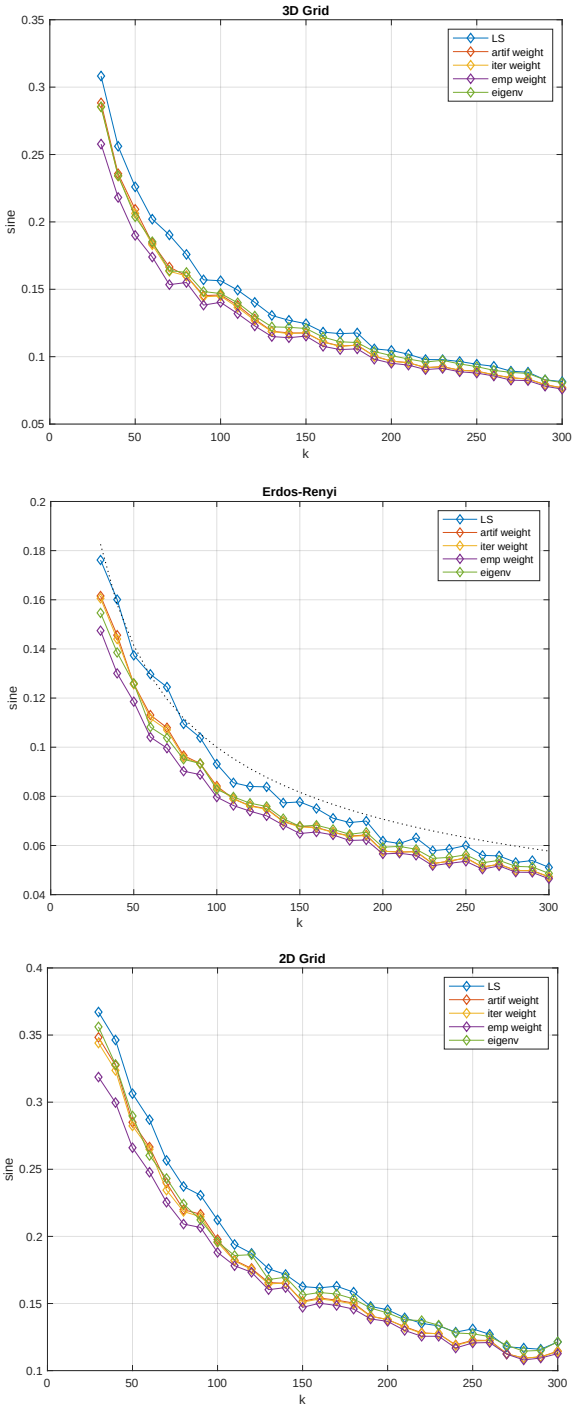


Figure 1. Performance on the 2D grid, 3D grid, and Erdos-Renyi graph. All three plots show $|\sin(\hat{W}, w)|$ on the y-axis vs the number of samples per edge on the x-axis. For the plots, the weights were generated randomly in the interval $[1, 20]$. The 2D and the E-R graph have 100 nodes, while the 3D grid has 125 nodes; the average degree of the E-R graph is 10. Each data point is the average of 50 simulations.

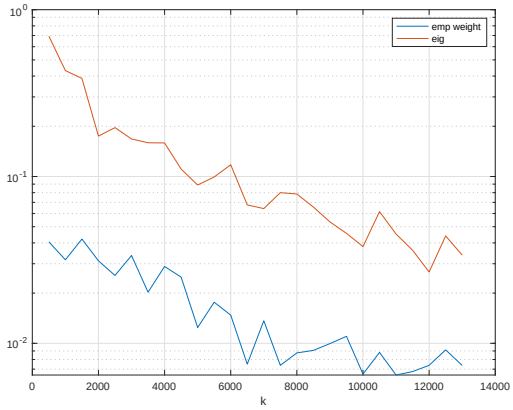


Figure 3. $\hat{W}_3 - \hat{W}_5$ for the eigenvector method in red and the WLSM in blue on the graph of Figure 2.

2016). Each number represents a single run of the algorithm with new random comparisons. We see that the WLSM outperforms by about an order of magnitude.

Our simulations thus point to two conjectures which can be the subject of further work. The first conjecture is that the earlier eigenvector based methods also achieve either the minimax scalings we have identified here, or something very close to them, as our simulations do not appear to detect any significant difference in performance. Indeed, note that the 3D grid has a very strong divergence between average resistance (constant) and spectral gap ($\simeq n^{2/3}$), and yet our simulation on the 3D grid showed no difference between the eigenvector based method (which has been upper bounded in terms of scaling with the spectral gap) and the WLSM (which we know to scale with resistance).

Moreover, a plausible conjecture is that the methods in question achieve optimal performance not just in distance between the vectors $\hat{W}, w$ but also among $\hat{W}_i - w_i$ for each node i (after appropriate normalization). We conjecture this is indeed the case for the WLSM. However, our simulation suggests this may not be the case for the eigenvector method, as we have constructed an example (Figures 2 and 3) where it underperforms in this metric.

4. Conclusions

Our main contribution is the determination of the asymptotic minimax rate for inference from pairwise comparisons. In contrast to previous work, our result is exact up to constant factors.

Besides the conjectures discussed in Section 3, the most natural open question raised by our work is to understand how big the number of samples per edge k has to be for the minimax rate derived in this paper to kick in. A bound can be obtained by tracing out the argument in our proof, but it will scale at least with $\max(|E|, n^2/R_{\mathrm{avg}})$. It is not hard to see that this is tight on some graphs, such as the line graph, but will not be tight on others. We would actually conjecture that $\mathrm{tr}(L_\gamma^\dagger)/||w||_2^2$ is, up to constant factors, not only the minimax rate but also the sample complexity of recovering (a scaled version of) w.

Another direction for future work is to attempt to extend the results presented here to more general comparison models: Thurstone, Mallows, multinomial models, and even hybrid models which combine comparisons with absolute scores. In principle, one could use the mean-value theorem to bound the error in making linear approximations to such models, leading to an error bound for the weighted least squares approach pursued in this work.

References

Agarwal, Arpit, Patil, Prathamesh, & Agarwal, Shivani. 2018. Accelerated spectral ranking. *Pages 70–79 of: International Conference on Machine Learning*.

Ailon, Nir. 2012. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research*, **13**(Jan), 137–164.

Beaver, Robert J. 1977. Weighted least-squares analysis of several univariate Bradley-Terry models. *Journal of the American Statistical Association*, **72**(359), 629–634.

Beaver, Robert J, & Gokhale, DV. 1975. A model to incorporat within-pair order effects in paired comparisons. *Communications in statistics-theory and methods*, **4**(10), 923–939.

Bradley, Ralph Allan, & Terry, Milton E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**(3/4), 324–345.

Brualdi, Richard A, & Ryser, Herbert John. 1991. *Combinatorial matrix theory*. Vol. 39. Springer.

Bubeck, Sébastien. 2011. Introduction to online optimization. *Lecture Notes*, 1–86.

Cattelan, Manuela. 2012. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 412–433.

Cattelan, Manuela, Varin, Cristiano, & Firth, David. 2013. Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**(1), 135–150.

Davidson, Roger R. 1970. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, **65**(329), 317–328.

Dwork, Cynthia, Kumar, Ravi, Naor, Moni, & Sivakumar, Dandapani. 2001. Rank aggregation methods for the web. *Pages 613–622 of: Proceedings of the 10th international conference on World Wide Web*. ACM.

Foster, Ronald M. 1949. The average impedance of an electrical network. *Contributions to Applied Mechanics (Reissner Anniversary Volume)*, 333–340.

Golub, Gene H, & Pereyra, Victor. 1973. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, **10**(2), 413–432.

Hajek, Bruce, & Raginsky, Maxim. 2019. *Statistical Learning Theory*. http://maxim.ece.illinois.edu/teaching/SLT/SLT.pdf. Book draft.

Hajek, Bruce, Oh, Sewoong, & Xu, Jiaming. 2014. Minimax-optimal inference from partial rankings. *Pages 1475–1483 of: Advances in Neural Information Processing Systems*.

Hendrickx, Julien M, Olshevsky, Alex, & Saligrama, Venkatesh. 2019. Graph Resistance and Learning from Pairwise Comparisons. *In: International Conference on Machine Learning*.

Hsu, Daniel, Kakade, Sham, & Zhang, Tong. 2012. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, **17**.

Jamieson, Kevin G, & Nowak, Robert. 2011. Active ranking using pairwise comparisons. *Pages 2240–2248 of: Advances in Neural Information Processing Systems*.

Jiang, Xiaoye, Lim, Lek-Heng, Yao, Yuan, & Ye, Yinyu. 2011. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, **127**(1), 203–244.

Landau, Henry, & Odlyzko, Andrew. 1981. Bounds for eigenvalues of certain stochastic matrices. *Linear algebra and its Applications*, **38**, 5–15.

Li, Lel, & Kim, Karl. 2000. Estimating driver crash risks based on the extended Bradley–Terry model: an induced exposure method. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **163**(2), 227–240.

Loewen, Peter John, Rubenson, Daniel, & Spirling, Arthur. 2012. Testing the power of arguments in referendums: A Bradley–Terry approach. *Electoral Studies*, **31**(1), 212–221.

Luce, R Duncan. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Matthews, JNS, & Morris, KP. 1995. An Application of Bradley-Terry-Type Models to the Measurement of Pain. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **44**(2), 243–255.

Maystre, Lucas, & Grossglauser, Matthias. 2015. Fast and accurate inference of Plackett–Luce models. *Pages 172–180 of: Advances in neural information processing systems*.

Negahban, Sahand, Oh, Sewoong, & Shah, Devavrat. 2012. Iterative ranking from pair-wise comparisons. *Pages 2474–2482 of: Advances in Neural Information Processing Systems*.

Negahban, Sahand, Oh, Sewoong, & Shah, Devavrat. 2016. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, **65**(1), 266–287.

Negahban, Sahand, Oh, Sewoong, Thekumparampil, Kiran K, & Xu, Jiaming. 2018. Learning from comparisons and choices. *The Journal of Machine Learning Research*, **19**(1), 1478–1572.

Rajkumar, Arun, & Agarwal, Shivani. 2014. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. *Pages 118–126 of: International Conference on Machine Learning*.

Rajkumar, Arun, & Agarwal, Shivani. 2016. When can we rank well from comparisons of O (n\log (n)) non-actively chosen pairs? *Pages 1376–1401 of: Conference on Learning Theory*.

Rao, PV, & Kupper, Lawrence L. 1967. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, **62**(317), 194–204.

Shah, Nihar B, Balakrishnan, Sivaraman, Bradley, Joseph, Parekh, Abhay, Ramchandran, Kannan, & Wainwright, Martin J. 2016. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *The Journal of Machine Learning Research*, **17**(1), 2049–2095.

Spielman, Daniel A, & Teng, Shang-Hua. 2014. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, **35**(3), 835–885.

Szörényi, Balázs, Busa-Fekete, Róbert, Paul, Adil, & Hüllermeier, Eyke. 2015. Online rank elicitation for plackett-luce: A dueling bandits approach. *Pages 604–612 of: Advances in Neural Information Processing Systems*.

Tetali, Prasad. 1994. An extension of Foster's network theorem. *Combinatorics, Probability and Computing*, **3**(3), 421–427.

Van der Vaart, Aad W. 2000. *Asymptotic statistics*. Vol. 3. Cambridge university press.

Vishnoi, Nisheeth. 2013. Lx=b. *Foundations and Trends in Theoretical Computer Science*, **8**(1–2), 1–141.

Wu, Rui, Xu, Jiaming, Srikant, Rayadurgam, Massoulié, Laurent, Lelarge, Marc, & Hajek, Bruce. 2015. Clustering and inference from pairwise comparisons. *Pages 449–450 of: ACM SIGMETRICS Performance Evaluation Review*, vol. 43. ACM.

Yue, Yisong, Broder, Josef, Kleinberg, Robert, & Joachims, Thorsten. 2012. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, **78**(5), 1538–1556.

Proof of Theorem 1

As briefly mentioned in the body of the paper, our proof will proceed by doing a large-deviation analysis of the estimation error, and then integrating it to obtain an upper bound on the expectation. We cannot proceed directly by attempting to take expectation of $\sin^2(\hat{W}, w)$ because the WLSM ends up dividing two random variables (specifically, Eq. (13) has random variables in both the numerator and denominator under the $\arg\min$).

To preview what is to come, our algorithm suffers from three sources of error:

- Difference between p_{ij}, the true probability that i wins a coin toss, and F_{ij} due to randomness of the comparisons.
- Error in taking Taylor expansions.
- Error introduced by replacing v_{ij}, which is proportional to the asymptotic variance of $\log R_{ij}$, by the empirical estimate $\hat{V}_{ij}$.

Our analysis will need to bound the effect of each of these factors.

4.1. Notation

We begin by reiterating all the notation we have introduced:

$$
\begin{aligned}
w_i &= \text{true weight of item } i \\
G = (\{1,\dots,n\}, E) &= \text{the comparison graph} \\
\overrightarrow{E} &= \text{set of directed edges obtained by orienting every edge in } E \text{ arbitrarily} \\
k &= \text{number of comparisons across each edge of } G \\
F_{ij} &= \text{proportion of comparisons item } i \text{ wins against item } j \\
p_{ij} &= \frac{w_i}{w_i + w_j}, \text{ true probability that item } i \text{ wins a comparison against item } j \\
R_{ij} &= \frac{F_{ij}}{F_{ji}} \\
\rho_{ij} &= \frac{w_i}{w_j} \\
v_{ij} &= \frac{w_i}{w_j} + \frac{w_j}{w_i} + 2 \\
\hat{V}_{ij} &= \frac{F_{ij}}{F_{ji}} + \frac{F_{ji}}{F_{ij}} + 2
\end{aligned}
$$

We follow the convention that capitalized entries are either random variables or matrices, while lower-case letters correspond to scalars or vectors that are not random. Next, we introduce some new notation:

$$\begin{aligned}
\hat{V} &= \mathrm{diag}(\hat{V}_{ij}) \in \mathbb{R}^{|\overrightarrow{E}|\times|\overrightarrow{E}|} \\
V &= \mathrm{diag}(v_{ij}) \in \mathbb{R}^{|\overrightarrow{E}|\times|\overrightarrow{E}|} \\
M &= \text{Edge-vertex incidence matrix of} \\
&\quad \text{the graph } (\{1,\dots,n\}, \overrightarrow{E}). \\
&\quad \text{Note that } M \in \mathbb{R}^{n\times|\overrightarrow{E}|} \\
L_V &= MV^{-1}M^T \in \mathbb{R}^{n\times n} \\
L_{\hat{V}} &= M\hat{V}^{-1}M^T \in \mathbb{R}^{n\times n} \\
X^l_{ij} &= \text{Bernoulli random variable describing the} \\
&\quad \text{outcome of } l\text{'th comparison across edge } (i,j) \\
\mathbf{1} &= \text{all-ones vector} \\
e_i &= i\text{'th basis vector}
\end{aligned}$$

We take the opportunity to remind the reader of our notational conventions. We will omit the subscripts when we stack the above quantities into vectors. For example, the notation R represents the vector in $\mathbb{R}^{|\overrightarrow{E}|}$ obtained by stacking up the quantities $R_{ij}, (i,j) \in \overrightarrow{E}$. Furthermore, the ordinary graph Laplacian L can be written as $L = MM^T$, and he quantities L_V and $L_{\hat{V}}$ correspond to weighted graph Laplacians, where the edge $(i,j) \in E$ is weighted by v_{ij}^{-1} or $\hat{V}_{ij}^{-1}$, respectively.

We begin by defining an appropriate rescalings of the true weights w to which we can compare $\hat{W}$ defined in Eq. (18). The natural approach is to define w^r to be a rescaling of w such that

$$\prod_{i=1}^{n}(w_i^r)^{w_i^2} = 1, \tag{20}$$

and likewise

$$y_i = w_i \log w_i^r. \tag{21}$$

Observing that for each edge $(i,j) \in E$,

$$\log w_i^r - \log w_j^r = \log \rho_{ij}.$$

we can therefore repeat all the same steps that led to the derivation of Eq. (17) to obtain that

$$y = (\mathrm{diag}(w)^{-1}L_{\hat{V}}\mathrm{diag}(w)^{-1})^{\dagger}\mathrm{diag}(w)^{-1}M\hat{V}^{-1}\log\rho.$$

Putting this together with Eq. (15), we obtain

$$\begin{aligned}
\hat{Y} - y =& (\mathrm{diag}(w)^{-1}L_{\hat{V}}\mathrm{diag}(w)^{-1})^{\dagger}\mathrm{diag}(w)^{-1} \\
& M\hat{V}^{-1}(\log R - \log\rho).
\end{aligned} \tag{22}$$

This equation will be the basis of our analysis for the rest of the paper.

To preview what is going to come, it is certainly true that

$$\sin^2(\hat{W}, w) = \sin^2(\hat{W}^r, w^r),$$

since the sine is unaffected by scalings of the underlying vectors. Thus nothing stops us from analyzing $\sin^2(\hat{W}^r, w^r)$ to analyze the WLSM method (even though we do not know either of these vectors). As discussed in the main body of the paper, it turns out that minimax optimal bounds come out of the least squares analysis only after this peculiar choice of normalization.

We conclude this section with one observation which we will need to use repeatedly throughout the remainder of the paper. Eq. (20) implies that there is at least one i with $w_i \geq 1$ and at least one i with $w_i \leq 1$. Appealing to Eq. (1), we can conclude that $\max_i w_i^r \leq b$ and $\min_i w_i^r \geq b^{-1}$.

4.2. Decomposing the sources of error

We will proceed by obtaining a rate at which the right-hand side of Eq. (22) goes to zero. Our first step is to bound the difference $\log R - \log\rho$. We cite a lemma from the previous literature which derives bound on this quantity by applying Chernoff's inequality.

Proposition 3 (Eq. (13) from (Hendrickx *et al.*, 2019))**.** *Let us write*

$$\log R - \log\rho = V(F-p) + \Delta.$$

Then if $\delta \leq e^{-1}$ and $k = \Omega\left(b\log(n/\delta)\right)$, we have that with probability $1-\delta$, the vector $\Delta \in \mathbb{R}^{|E|}$ satisfies

$$||\Delta||_\infty \leq O\left(\frac{b\log(n/\delta)}{k}\right).$$

The interpretation of this lemma is as follows. The first term, $V(F-p)$, comes from the linear Taylor expansion of $\log R$ about its limit of $\log\rho$, while the second term, Δ, comes from bounding the rest of the terms in the Taylor expansion. The above lemma shows that $||\Delta||_\infty$ tends to be on the order of $O(1/k)$. As expected, this is a faster decay as compared to the first term: indeed, since $F-p$ is the average of k independent random variables, one for each comparison, by central-limit considerations we expect $F-p$ to be on the order of $O(1/\sqrt{k})$.

Furthermore we remark that $\log R$ could potentially have an infinite entry (this can happen if one node wins every comparison against a neighbor). The above lemma implies that the probability of that is at most δ under the lower bound $k \geq \Omega(b\log(n/\delta))$.

By taking $\delta = n/e^{k^{3/4}}$ in this proposition, which satisfies $\delta \leq e^{-1}$ and $k \geq \Omega(b\log(n/\delta))$ for k large enough, we obtain the following.

Corollary 4. *Let $f(w, b, G)$ be any function of the weights w, the constant b, and the graph G. Then*

$$P\left(||\Delta||_\infty \geq f(w,b,G)\right) = O\left(ne^{-k^{3/4}}\right).$$

Our next lemma follows up on these observations by decomposing the error from Eq. (22) into three parts.

Lemma 5. *We have*

$$\hat{Y} - y = A + B + C,$$

where

$$\begin{aligned} A &= (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} M(F-p) \\ B &= (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1}\Delta \\ C &= (\mathrm{diag}(w)^{-1} L_{\hat{V}} \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} M\hat{V}^{-1} \\ &\quad -(\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1}) \\ &\quad (\log R - \log \rho), \end{aligned}$$

and the vector Δ is defined through Proposition 3.

Before we proceed to the proof, which is quite short, we discuss where each of the three terms comes from. Observe that the first term, A, is obtained by replacing all instances of $\hat{V}^{-1}$ with V^{-1} in Eq. (22) and further replacing $\log R - \log \rho$ by its first order Taylor expansion $V(F-p)$. Naturally, this replacement is going to create errors, and these are handled by adding the terms B and C. The term B corrects the error from replacing $\log R - \log \rho$ by $V(F-p)$, while the term C corrects the error from replacing $\hat{V}^{-1}$ by V^{-1}.

We next give the proof of this lemma.

Proof of Lemma 5. Indeed, beginning from Eq. (22), we have

$$\begin{aligned} \hat{Y} - y &= (\mathrm{diag}(w)^{-1} L_{\hat{V}} \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} M\hat{V}^{-1}(\log R - \log \rho) \\ &= (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1}(\log R - \log \rho) \\ &\quad + ((\mathrm{diag}(w)^{-1} L_{\hat{V}} \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} M\hat{V}^{-1} - \\ &\quad (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1})(\log R - \log \rho) \\ &= (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1} V(F-p) \\ &\quad + (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1}\Delta \\ &\quad + ((\mathrm{diag}(w)^{-1} L_{\hat{V}} \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} M\hat{V}^{-1} \\ &\quad - (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1})(\log R - \log \rho) \\ &= A + B + C. \end{aligned} \tag{23}$$

□

4.3. Analyzing the sources of error

We next proceed to bound each term in Lemma 5 separately. Our first step is to bound each A_i.

Lemma 6. *With probability $1-\delta$,*

$$||A||_2^2 \leq O\left(\frac{b}{k}\right) \mathrm{Tr}\left[(\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger\right] \left(1 + \log\frac{1}{\delta}\right) \tag{24}$$

Moreover, for any function $f(w, b, G)$,

$$P(||A||_2^2 \geq f(w,b,G)) = O\left(e^{-k/g(w,b,G)}\right), \tag{25}$$

for some function $g(\cdot,\cdot,\cdot)$ of w, b, G.

Proof. By definition,

$$\begin{aligned} A &= (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} M(F-p) \\ &= (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1/2} \\ &\quad V^{1/2}(F-p) \end{aligned}$$

Now let X_{ij}^l to be the Bernoulli random variable denoting the outcome of l'th toss across edge (i,j) (i.e., 1 if i wins, 0 otherwise). Observe that

$$\mathrm{var}(X_{ij}^l - p_{ij}) = p_{ij}(1-p_{ij}) = \frac{w_i w_j}{(w_i+w_j)^2} = v_{ij}^{-1},$$

and therefore

$$\mathrm{var}\left(\sqrt{V_{ij}}(X_{ij}^l - p_{ij})\right) = 1.$$

After this calculation, we can write

$$A = \left[(\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1/2}\right] \left[V^{1/2}(F-p)\right],$$

where the term in the right brackets is a random vector with zero mean and variance $1/k$. We can rewrite this as

$$A = \left[\frac{1}{\sqrt{k}}(\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1/2}\right] \left[\sqrt{k}V^{1/2}(F-p)\right], \tag{26}$$

and now the term in brackets has zero mean and unit variance. Let us introduce the notation

$$P = (\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1} MV^{-1/2}. \tag{27}$$

We have that $\sqrt{k}V^{1/2}(F-p)$ is a subgaussian random variable with subgaussin parameter of $O(\sqrt{b})$ (this follows from observing that the outcome of l'th toss, $X_{ij}^l - p$, has support contained in $[-1,1]$, as well as rules for adding and scaling subgaussian random variables). Via Theorem 2.1 of

(Hsu *et al.*, 2012) we have that[5]

$$P\left(||A||_2^2 \geq b\left(\frac{\mathrm{Tr}(PP^T)}{k}(1+4t)\right)\right) \leq e^{-t}.$$

Choosing $t = \log(1/\delta)$ and using $\mathrm{Tr}(P^TP) = \mathrm{Tr}(PP^T) = \mathrm{Tr}\left[(\mathrm{diag}(w)^{-1}L_V\mathrm{diag}(w)^{-1})^\dagger\right]$, yields Eq. (24).

To prove Eq. (25), observe that, as a consequence of Eq. (24), for large enough k we must have that $\log(1/\delta)$ has to scale linearly with $k/g(w,b,G)$ in order for $||A||_2^2 > f(w,b,G)$, where we spare ourselves the trouble of writing out the function $g(w,b,G)$ in terms of $f(w,b,G)$. This proves Eq. (24). □

We next turn our attention to the second term in Lemma 5, namely the vector B. While Lemma 6 showed that A_i effectively decays at an $O(1/\sqrt{k})$ rate, our next lemma shows that B_i decays at the faster $O(1/k)$ rate. The lemma requires a few definitions from electric circuit theory, which we next provide.

Given a weighted undirected graph, we can talk about the effective resistance between any two nodes in the graph by treating the edge (i,j) as if it has a resistor of resistance equal to the weight of that edge. We will define the effective resistance between nodes i and j by $R_{\mathrm{eff}}(i,j)$. We then define

$$R_{\max}(i) = \max_{j=1,\ldots,n} R_{\mathrm{eff}}(i,j), \tag{28}$$

to be the average effective resistance between node i and the rest of the nodes in the graph. For a formal analysis of the electric theory of graphs, we refer the reader to Chapter 4 of (Vishnoi, 2013).

Lemma 7. *Let $q \in \mathbb{R}^{\vec{E}}$ be a positive vector. Then for all $i = 1,\ldots,n$,*

$$\left|\left|\mathrm{diag}(q)^{-1}M^T\mathrm{diag}(w)^{-1}(\mathrm{diag}(w)^{-1}L_q\mathrm{diag}(w)^{-1})^\dagger e_i\right|\right|_1 \leq w_i\sqrt{SR_{\max}(i)},$$

where

$$S_q = \sum_{(i,j)\in\vec{E}} q_{ij}^{-1},$$

L_q is the Laplacian of the weighted graph with weights q_{ij}^{-1}, and $R_{\max}(i)$ is defined as in Eq. (28).

Proof. As above, consider turning the comparison graph into a circuit, with edge (i,j) having resistance q_{ij}. This allows us to interpret the equation

$$L_q x = i.$$

Indeed, if i is a vector of currents going in and out of nodes in the network (summing to zero), then $x = L_q^\dagger i$ gives a corresponding vector of electric potentials.

We may write this as

$$\mathrm{diag}(w)^{-1}L_q\mathrm{diag}(w)^{-1}\mathrm{diag}(w)x = \mathrm{diag}(w)^{-1}i$$

which gives an interpretation to the operation

$$x = \mathrm{diag}(w)^{-1}(\mathrm{diag}(w)^{-1}L_q\mathrm{diag}(w)^{-1})^\dagger q.$$

Indeed, we can interpret this as follows: if $q = \mathrm{diag}(w)^{-1}i$ where the entries of i add up to zero, then x a vector of electric potentials resulting from the current inputs i.

Coming back to the problem in question, let us adopt the notation

$$f = \mathrm{diag}(q)^{-1}M^T\mathrm{diag}(w)^{-1}(\mathrm{diag}(w)^{-1}L_q\mathrm{diag}(w)^{-1})^\dagger e_i.$$

Since $(\mathrm{diag}(w)^{-1}L_q\mathrm{diag}(w)^{-1})^\dagger w = 0$, we can rewrite this as

$$f = \mathrm{diag}(q)^{-1}M^T\mathrm{diag}(w)^{-1}(\mathrm{diag}(w)^{-1}L_q\mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1}\left(w_ie_i - \frac{w_i}{||w||_2^2}w^2\right),$$

where w^2 means the element-wise square of the entries of w. Finally,

$$\mathrm{diag}(w)^{-1}(\mathrm{diag}(w)^{-1}L_q\mathrm{diag}(w)^{-1})^\dagger\mathrm{diag}(w)^{-1}\left(w_ie_i - \frac{w_i}{||w||_2^2}w^2\right)$$

is the vector of electric potentials when we put w_i units of current into node i and take out $w_i(w_j^2/||w||_2^2)$ out of node j. Moreover, f is then the vector of edge currents corresponding to this setup.

We may further view f as the superposition of n current flows, with the j'th flow f_j obtained by putting $w_j^2w_i/||w||_2^2$ units of current at i and taking the same amount out of j. We will write this as

$$f = f_1 + \cdots + f_n.$$

The advantage of this representation is that we may apply Thompson's principle (Theorem 4.7 of (Vishnoi, 2013)) to

[5] Strictly speaking, the reference (Hsu *et al.*, 2012) only bounds $P(||A||_2^2 \geq u)$ for $A = PZ$ when P is square. In our case, P is rectangular. However, we observe that any concentration bound for A in terms of $\mathrm{tr}(PP^T)$ proved for the case when P is square immediately implies the same bound when P is rectangular. This follows because

$$||A||_2^2 = A^TA = Z^TP^TPZ = ||QZ||_2^2,$$

where Q is the psd square root of P^TP. Thus we can apply the results of (Hsu *et al.*, 2012) to bound $P(||A||_2^2 \geq u) = P(||QZ||_2^2 \geq u)$; and since $\mathrm{tr}(QQ^T) = \mathrm{tr}(PP^T)$, the result will be exactly the same as if we simply ignored the assumption of (Hsu *et al.*, 2012) that P is square.

each flow f_j. Slightly rephrased, that theorem states that the effective resistance between nodes i and j satisfies

$$\frac{w_i^2 w_j^4}{||w||_2^4} R_{\text{eff}}(i,j) = \sum_{e \in \vec{E}} q_e (f_j)_e^2,$$

where for $e = (a,b)$ we use q_e and q_{ab} interchangeably. We may rewrite this as

$$||f_j . * \sqrt{q}||_2 = w_i w_j^2 \frac{\sqrt{R_{\text{eff}}(i,j)}}{||w||_2^2},$$

where we use ".$*$" to denote the elementwise product of two vectors; note that both f_j and v can be viewed as vectors in $\mathbb{R}^{|\vec{E}|}$. We use these inequalities in conjunction with Cauchy-Schwarz to argue that

$$\begin{aligned} ||f||_1 &\leq \sum_{j=1}^n ||f_j||_1 \\ &= \sum_{j=1}^n ||f_j . * \sqrt{q_j} . * \sqrt{q_j}^{-1}||_1 \\ &\leq \sum_{j=1}^n ||f_j . * \sqrt{q_j}||_2 \sqrt{S} \\ &\leq \sum_{j=1}^n w_i w_j^2 \frac{\sqrt{R_{\text{eff}}(i,j)}}{||w||_2^2} \sqrt{S} \\ &\leq w_i \sqrt{S R_{\max}(i)}, \end{aligned}$$

□

With the above definitions in place, we can state our decay bound on the elements B_i of the vector B.

Lemma 8. *If $\delta \leq e^{-1}$ and $k = \Omega(b \log(n/\delta))$, we have that with probability $1-\delta$,*

$$B_i \leq O\left(w_i \frac{b \log(n/\delta) \sqrt{S R_{\text{avg}}(i)}}{k}, \right) \tag{29}$$

for all $i = 1, \ldots, n$, where

$$S = \sum_{(i,j) \in \vec{E}} v_{ij}^{-1},$$

and $R_{\max}(i)$ is defined as in Eq. (28) for the graph where the edge (i,j) has resistance v_{ij}. Moreover, for any function $f(w,b,G)$ of the graph G and the weights w, we have that

$$P\left(B_i \geq f(w,b,G)\right) \leq O\left(n e^{-k^{3/4}}\right). \tag{30}$$

Proof. By definition we have that

$$\begin{aligned} B_i &= [(\text{diag}(w)^{-1} L_V \text{diag}(w)^{-1})^\dagger \text{diag}(w)^{-1} M V^{-1} \Delta]_i \\ &= e_i^T (\text{diag}(w)^{-1} L_V \text{diag}(w)^{-1})^\dagger \text{diag}(w)^{-1} M V^{-1} \Delta \\ &\leq ||\Delta||_\infty \left|\left| V^{-1} M^T \text{diag}(w)^{-1} (\text{diag}(w)^{-1} L_V \text{diag}(w)^{-1})^\dagger e_i \right|\right|_1 \\ &\leq O\left(\frac{b \log(n/\delta)}{k}\right) \left|\left| V^{-1} M^T \text{diag}(w)^{-1} (\text{diag}(w)^{-1} \right. \right. \\ &\qquad \left. \left. L_V \text{diag}(w)^{-1})^\dagger e_i \right|\right|_1, \end{aligned} \tag{31}$$

where the third inequality used Holder's inequality and the last step used Proposition 3. Now using Lemma 7), the proof of Eq. (29) is concluded.

As for Eq. (30), it follows immediately from Eq. (31) by taking $\delta = n/e^{k^{3/4}}$; for large enough k, we will have both $\delta \leq e^{-1}$ and $k \geq \Omega(b \log(n/\delta))$ required for that equation to hold.

□

We next turn to the analysis of the final term in Lemma 5. Unfortunately, this term is the most cumbersome, and will require quite a number of calculations. Our starting point will be to argue that

$$\begin{aligned} C_i &= e_i^T ((\text{diag}(w)^{-1} L_{\hat{V}} \text{diag}(w)^{-1})^\dagger \text{diag}(w)^{-1} M \hat{V}^{-1} \\ &\quad -(\text{diag}(w)^{-1} L_V \text{diag}(w)^{-1})^\dagger \text{diag}(w)^{-1} M V^{-1})(\log R - \log \rho) \\ &\leq ||e_i^T ((\text{diag}(w)^{-1} L_{\hat{V}} \text{diag}(w)^{-1})^\dagger \text{diag}(w)^{-1} M \hat{V}^{-1} - \\ &\quad (\text{diag}(w)^{-1} L_V \text{diag}(w)^{-1})^\dagger \text{diag}(w)^{-1} M V^{-1})||_1 \\ &\quad ||\log R - \log \rho||_\infty \end{aligned} \tag{32}$$

We now proceed to bound both of the terms on the right-hand side. We begin with the second term, as it's analysis is the easiest.

Lemma 9. *If $\delta \leq e^{-1}$ and $k \geq \Omega(b \log(n/\delta))$, then with probability $1 - 2\delta$,*

$$||\log R - \log \rho||_\infty \leq O\left(\sqrt{\frac{b \log(n/\delta)}{k}}\right) + O\left(\frac{b \log(n/\delta)}{k}\right), \tag{33}$$

Moreover, if $f(w,b,G)$ is any function of w, b, and G, then

$$P\left(||\log R - \log \rho||_\infty \geq f(w,b,G)\right) \leq O\left(n e^{-k^{3/4}}\right). \tag{34}$$

Proof. From Proposition 3, we have that

$$\log R - \log \rho = V(F - p) + \Delta, \tag{35}$$

with

$$||\Delta||_\infty \leq O\left(\frac{b \log(n/\delta)}{k}\right) \tag{36}$$

with probability $1 - \delta$. This leads to the second term in the statement of the lemma. For the first term, we will need to bound $||V(F-p)||_\infty$.

To that end, we observe that Lemma 1 in (Hendrickx *et al.*, 2019) proved that if $\delta \leq e^{-1}$ and $k \geq \Omega\left(b \log(n/\delta)\right)$, then

$$P\left(||F-p||_\infty \geq \sqrt{\frac{O\left(\log(n/\delta)\right)}{kb}}\right) \leq \delta. \tag{37}$$

Thus with probability $1-\delta$,

$$||V(F-p)||_\infty \leq O\left(\sqrt{\frac{b\log(n/\delta)}{k}}\right). \tag{38}$$

Putting together the two probability $1-\delta$ bounds of Eq. (36) and Eq. (38) via the union bound proves Eq. (33). As for Eq. (34), it follows from Eq. (35) and Eq. (38), by taking $\delta = n/e^{k^{3/4}}$ which, for large enough k, satisfies the conditiosn $\delta \leq e^{-1}$ and $k = \Omega(b\log(n/\delta))$; and Corollary 4.

□

Having established this lemma, we have a bound on the second term in Eq. (32); we now turn to analyzing the first term the same equation. Let us introduce the following notation for the first term,

$$\begin{aligned}\Gamma_i =& ||e_i^T((\mathrm{diag}(w)^{-1}L_{\hat{V}}\mathrm{diag}(w)^{-1})^\dagger \mathrm{diag}(w)^{-1}M\hat{V}^{-1} - \\ & (\mathrm{diag}(w)^{-1}L_V\mathrm{diag}(w)^{-1})^\dagger\mathrm{diag}(w)^{-1}MV^{-1})||_1,\end{aligned} \tag{39}$$

which will make the ensuing discussion more compact. We need to upper bound Γ_i, which is to say we need to upper bound the change that comes from replacing V by $\hat{V}$; what makes things difficult, however, is that the expression involves the pseudoinverses L_V and $L_{\hat{V}}^\dagger$. We will take the "brute force" approach of writing out the derivative of the expression inside the one-norm in Eq. (39) and integrating it over the path between V and $\hat{V}$.

To that end, let us define the function

$$\begin{aligned}H_i(u) =& e_i^T(\mathrm{diag}(w)^{-1}M\mathrm{diag}(u)M^T\mathrm{diag}(w)^{-1})^\dagger \\ & \mathrm{diag}(w)^{-1}M\mathrm{diag}(u).\end{aligned}$$

As should be clear from matching up dimensions in this equation, H_i maps $\mathbb{R}^{|\overrightarrow{E}|}$ into $\mathbb{R}^{|\overrightarrow{E}|}$. We will slightly abuse notation by writing expressions like $H_i(v_{ab}^{-1})$, which should be understood to mean H_i applied to the vector in $\mathbb{R}^{|\overrightarrow{E}|}$ obtained by stacking up the quantities v_{ab}^{-1} as (a,b) ranges over the edges in $\overrightarrow{E}$.

By the definition of the weighted Laplacian, the expression Γ can be written as

$$\Gamma_i = ||H_i(\hat{V}_{ab}^{-1}) - H_i(v_{ab}^{-1})||_1.$$

To make the connection to the underlying coin tosses more explicit, we can write $\hat{V}_{ab}^{-1}$ as a function of the fractions F_{ab}, and likewise v_{ab}^{-1} can be written as a function of the true probabilities p_{ab}. To spell this out observe that

$$v_{ab}^{-1} = \frac{1}{w_a/w_b + w_b/w_a + 2} = \frac{w_a w_b}{(w_a+w_b)^2} = p_{ab}(1-p_{ab}),$$

and likewise

$$\hat{V}_{ab}^{-1} = F_{ab}(1-F_{ab}).$$

Thus defining

$$U(x_1,\ldots,x_{|\overrightarrow{E}|}) = \left(x_1(1-x_1),\ldots,x_{|\overrightarrow{E}|}(1-x_{|\overrightarrow{E}|})\right), \tag{40}$$

we can now write

$$\Gamma_i = ||H_i(U(F_{ab})) - H_i(U(p_{ab}))||_1. \tag{41}$$

We thus proceed by upper bounding the gradient of the function $H_i(U(x_{ab}))$ by using the chain rule. Our next lemma takes the first step in this direction by giving an explicit expression for $\partial H_i/\partial u_{ab}$ for some fixed indices $a,b \in \{1,\ldots,n\}$.

In the lemma below,

$$L_u = \mathrm{diag}(w)^{-1}M\mathrm{diag}(u)M^T\mathrm{diag}(w)^{-1},$$

denotes the weighted graph Laplacian when $u \in \mathbb{R}^{|\overrightarrow{E}|}$ is the vector of weights, scaled left and right by $\mathrm{diag}(w)^{-1}$.

Lemma 10. *Let a,b be elements of $\{1,\ldots,n\}$. If the vector $u \in \mathbb{R}^{|\overrightarrow{E}|}$ is elementwise positive, then*

$$\begin{aligned}\frac{dH_i}{du_{ab}} =& e_i^T\left[L_u^\dagger \mathrm{diag}(w)^{-1}(e_a-e_b)\right] \\ & \left[-(e_a-e_b)^T\mathrm{diag}(w)^{-1}L_u^\dagger\mathrm{diag}(w)^{-1}M\mathrm{diag}(u) + e_{ab}^T\right]\end{aligned}$$

Here e_{ab} denotes a column vector in $\mathbb{R}^{|\overrightarrow{E}|}$ with a one in the (a,b) entry and zeros elsewhere. Note that the first expression in brackets is a column vector in $\mathbb{R}^n$ while the second expression in brackets is a row vector in $\mathbb{R}^{|\overrightarrow{E}|}$.

Proof. We first compute

$$\begin{aligned}\frac{dL_u}{du_{ab}} &= \mathrm{diag}(w)^{-1}Me_{ab}e_{ab}^TM^T\mathrm{diag}(w)^{-1} \\ &= \mathrm{diag}(w)^{-1}(e_a-e_b)(e_a-e_b)^T\mathrm{diag}(w)^{-1}.\end{aligned} \tag{42}$$

We next use this to find the derivative of $L_u^\dagger$. We use Theorem 4.3 from (Golub & Pereyra, 1973), which provides an expression for $\partial L_U/\partial u_{ab}$ in a neighborhood of a point where the rank of L_u is constant. That formula applies in our case because as long as $u > 0$ and G is a connected graph, we will have that the rank of G equals $n-1$ (see Section 2.5 of (Brualdi & Ryser, 1991)).

The expression from Theorem 4.3 of (Golub & Pereyra, 1973) is

$$\frac{dA^\dagger}{dx} = -A^\dagger \frac{dA}{dx} A^\dagger + A^\dagger A^{\dagger T} \frac{dA^T}{dx}(I - AA^\dagger) + (I - A^\dagger A)\frac{dA^T}{dx} A^{\dagger T} A^\dagger \tag{43}$$

When we plug in $A = L_u$, the expression simplifies considerably because

$$\begin{aligned}(I - L_u^\dagger L_u)\frac{dL_u^T}{du_{ab}} =& (I - L_u^\dagger L_u)\mathrm{diag}(w)^{-1}(e_a - e_b) \\ & (e_a - e_b)^T \mathrm{diag}(w)^{-1} \\ =& \left[(I - L_u^\dagger L_u)\mathrm{diag}(w)^{-1}(e_a - e_b)\right] \\ & (e_a - e_b)^T \mathrm{diag}(w)^{-1} \\ =& 0,\end{aligned}$$

where the last step follows because $w^T \mathrm{diag}(w)^{-1}(e_a - e_b) = 0$ implies the expression in brackets is zero. Therefore the third term in Eq. (43) is zero and a similar argument implies the second term is zero as well. Thus

$$\begin{aligned}\frac{dL_u^\dagger}{du_{ab}} &= -L_u^\dagger \frac{dL_u}{du_{ab}} L_u^\dagger \\ &= -L_u^\dagger \mathrm{diag}(w)^{-1}(e_a - e_b)(e_a - e_b)^T \mathrm{diag}(w)^{-1} L_u^\dagger.\end{aligned}$$

We can now use this to compute the derivative of H using the chain rule. Indeed,

$$\begin{aligned}\frac{dH}{du_{ab}} =& e_i^T \frac{d\left(\mathrm{diag}(w)^{-1} M \mathrm{diag}(u) M^T \mathrm{diag}(w)^{-1}\right)^\dagger}{du_{ab}} \mathrm{diag}(w)^{-1} M \mathrm{diag}(u) \\ & + e_i^T \left(\mathrm{diag}(w)^{-1} M \mathrm{diag}(u) M^T \mathrm{diag}(w)^{-1}\right)^\dagger \mathrm{diag}(w)^{-1} M \frac{d\mathrm{diag}(u)}{du_{ab}} \\ =& -e_i^T L_u^\dagger \mathrm{diag}(w)^{-1}(e_a - e_b)(e_a - e_b)^T \mathrm{diag}(w)^{-1} L_u^\dagger \mathrm{diag}(w)^{-1} M \mathrm{diag}(u) \\ & + e_i^T L_u^\dagger \mathrm{diag}(w)^{-1}(e_a - e_b) e_{ab}^T.\end{aligned}$$

After some rearranging, this gives the statement of the lemma. □

With this explicit expression for the gradient of H in place, we can proceed to upper bound Γ_i. However, we first need the following lemma, which will be used in one of the intermediate steps of the bound.

Lemma 11. *If u is elementwise positive, then for any $i, a, b \in \{1, \dots, n\}$,*

$$|e_i^T L_u^\dagger \mathrm{diag}(w)^{-1}(e_a - e_b)| \quad \leq \quad w_i R_{\text{eff}}(a, b) \tag{44}$$

Proof. We prove this by appealing to the electrical interpretation just as we did in the proof of Lemma 8. Using the observations made in the proof of that lemma, we interpret $e_i^T \mathrm{diag}(w)^{-1} L_u^\dagger \mathrm{diag}(w)^{-1}(e_a - e_b)$ is the potential at node i when a single unit of current is injected at a and taken out at b in the circuit where the resistance of the edge (i, j) is u_{ij}^{-1}. Call this potential v_i. The quantity we are seeking the bound is thus just $w_i v_i$.

Note, however, that the largest potential is at $i = a$ and the smallest potential is at $i = b$. Thus we can instead upper bound $\max(|v_a|, |v_b|)$. Moreover, v_a is positive and v_b is negative because anything in the range of $L_u^\dagger$ is orthogonal to the positive vector w. Moreover, since by Ohm's law $v_a - v_b = R_{\text{eff}}(a, b) \cdot 1$, we have that

$$\max(|v_a|, |v_b|) \leq v_a - v_b = R_{\text{eff}}(a, b).$$

□

We now put all the pieces together and produce a high probability bound on the quantity C_i.

Lemma 12. *When $\delta \leq e^{-1}$ and $k = \Omega\left(b \log(n/b)\right)$, then we have that with probability $1 - 2\delta$, for all $i = 1, \dots, n$,*

$$C_i \leq w_i b(n-1)\log(n/\delta)\left(O\left(\frac{1}{k}\right) + O\left(\frac{\sqrt{b\log(n/\delta)}}{k^{1.5}}\right)\right).$$

Moreover, if $f(w, b, G)$ is any function of w, b, and G, then

$$P\left(C_i \geq f(w, b, G)\right) \leq O\left(n e^{-k^{3/4}}\right). \tag{45}$$

Proof. Indeed, from Eq. (32) and Eq. (39), we have

$$C_i \leq \Gamma_i || \log R - \log \rho ||_\infty.$$

Using the expression form Eq. (41), we can upper bound this as

$$C_i \leq ||H_i(U(F_{ij})) - H_i(U(p_{ij}))||_1 || \log R - \log \rho ||_\infty.$$

As a consequence of Theorem 4.3 of (Golub & Pereyra, 1973), $H_i(U(u_{ij}))$ is differentiable over the set of $u > 0$; we may therefore invoke the mean value theorem to obtain the bound

$$C_i \leq ||\nabla H_i(U(z))(F - p)||_1 || \log R - \log \rho ||_\infty,$$

where z is some point lying on the line interval connecting the vectors F and p. Using the definition of matrix norm, we can in turn upper bound this as

$$\begin{aligned}C_i \leq & ||\nabla H_i(U(z))||_\infty ||F - p||_\infty || \log R - \log \rho ||_\infty \\ =& \left(\max_{j=1,\dots,n} ||e_j^T \nabla H_i(U(z))||_1\right) ||F - p||_\infty || \log R - \log \rho ||_\infty\end{aligned} \tag{46}$$

where we used the standard fact that the infinity norm of a matrix is the largest one-norm of its rows.

The second and third quantities above have been bounded in the previous lemmas; only the quantities $||e_j^T \nabla H_i(U(z))||_1$

needs to be analyzed. However, observe that

$$\begin{aligned}\frac{\partial H_i(U(z))}{\partial z_{ij}} &= \frac{\partial H_i(u)}{\partial u_{ij}}\frac{\partial u_{ij}}{\partial z_{ij}} \\ &= \frac{\partial H_i(u)}{\partial u_{ij}}(1-2z_{ij}),\end{aligned}$$

where the last line used Eq. (40). Thus if $z \in [0,1]$, we have that

$$\left|\frac{\partial H_i(U(z))}{\partial z_{ij}}\right| \leq \left|\frac{\partial H_i(u)}{\partial u_{ij}}\right|,$$

which in turn implies that

$$||e_j^T \nabla H_i(U(z))||_1 \leq ||e_j^T \nabla H_i(u)||_1. \tag{47}$$

Thus what remains to be done is to obtain a bound on the quantity $||e_j^T \nabla H_i(u)||_1$. This can be done using the explicit expression for the partial derivatives of H_i derived in Lemma 10. Indeed, observe that the rows of $\nabla H_i(u)$ are the transposed vectors $\frac{\partial H_i}{\partial u_{ab}}$ from Lemma 10. Thus using that lemma we have

$$\begin{aligned}\left|\left|\max_{j=1,\ldots,n} e_j^T \nabla H(u)\right|\right|_1 &\leq \sum_{(a,b)\in E}\left|\left|\frac{\partial H_i}{\partial u_{ab}}\right|\right|_\infty \\ &\leq \sum_{(a,b)\in E}\left|\left|e_i^T L_u^\dagger \mathrm{diag}(w)^{-1}(e_a-e_b)\right|\right|_\infty \\ &\quad \left|\left|-(e_a-e_b)^T \mathrm{diag}(w)^{-1} L_u^\dagger \mathrm{diag}(w)^{-1} M \mathrm{diag}(u) + e_a^T\right|\right|_\infty \\ &\leq 2\sum_{(a,b)\in E} w_i R_{\mathrm{eff}}(a,b),\end{aligned} \tag{48}$$

where the last line used Eq. (44) as well as the observation that $(e_a-e_b)^T \mathrm{diag}(w)^{-1} L_u^\dagger \mathrm{diag}(w)^{-1} M \mathrm{diag}(u) \in [-1,1]^{|E|}$. This last observation follows because each entry of this vector is the current in the graph where the resistance of $(i,j)\in E$ is u_{ij}^{-1} and a single unit of current is injected at a and taken out at b.

Our next step is to use Foster's identity (Foster, 1949; Tetali, 1994)

$$\sum_{(a,b)\in E} R_{\mathrm{eff}}(a,b) u_{ab} = n-1,$$

to conclude from Eq. (48) that

$$\left|\left|\max_{j=1,\ldots,n} e_j^T \nabla H(u)\right|\right|_1 \leq w_i \frac{2(n-1)}{\min_{(a,b)\in E} u_{ab}}.$$

Plugging this into Eq. (46) we obtain

$$C_i \leq w_i \frac{2(n-1)}{\min_{(a,b)\in E} u_{ab}} ||F-p||_\infty ||\log R - \log \rho||_\infty \tag{49}$$

Finally, we observe that, as a consequence of Eq. (37) and some algebra, when $k \geq \Omega(b\log(n/\delta))$, we have that $|F_{ij}-p_{ij}| < 1/(4b)$ with probability $1-\delta$. We can use this to bound the quantity $\min_{(a,b)\in E} u_{ab}$ appearing above. Indeed, $u = U(z)$ where z lies on the path between F_{ij} and p_{ij}, and thus $z_{ij} \geq 1/(4b)$ and $1-z_{ij} \geq 1/(4b)$ for all $(i,j)\in E$. Then $[U(z)]_{ij} = z_{ij}(1-z_{ij}) \geq \frac{1}{8b}$.

Using this bound in conjunction with Eq. (38) to bound $||F-p||_\infty$ and Eq. (33) to bound $||\log R - \log \rho||_\infty$ we obtain the statement of the lemma.

We now turn to proving Eq. (45). Inspecting Eq. (49), we first argue that $1/\min_{(a,b)\in E} u_{ab}$ is $O(b)$ with probability $O(ne^{-k^{3/4}})$. Indeed, as discussed just above, each z_{ij} lies on the path between F_{ij} and p_{ij}; this means that that $\min_{ij} z_{ij} < 1/(2(b+1))$ by with probability $O(ne^{-k^{3/4}})$ by taking $\delta = n/e^{k^{3/4}}$ in Lemma 1 of (Hendrickx *et al.*, 2019); which implies $\max_{(i,j)} 1/u_{ij} = \max_{(i,j)}(1/z_{ij})(1/(1-z_{ij})) = O(b)$ with the same probability.

Thus returning to Eq. (49), we see that in order for C_i to exceed some function $f(w,b,G)$, it must be that $||\log R - \log \rho||_\infty$ exceeds some function of the w, n and the graph G; but that has been shown to happen with probability $O(ne^{-k^{3/4}})$ in Eq. (34). In summary, the only way for C_i to exceed a fixed function of w, b, G is one of two events to happen, both of which have probability $O(ne^{-k^{3/4}})$. □

We now put together the bounds we have obtained on A_i, B_i, C_i into several general bounds on the error. Our first step will analyze the worst-case possible scalings. Here we have to consider even the scenario when one node wins all the comparisons to a neighbor, resulting in some F_{ij} which is as small as $O(1/k)$ due the lines 5 and 7 of the WLSM.

Lemma 13. *With probability one, we have that for all $i = 1,\ldots,n$,*

$$\begin{aligned}|B_i| &= O_{w,b,G}(\log k) \\ |C_i| &= O_{w,b,G}(k\log k) \\ |\hat{Y}_i| &= O_{w,b,G}(\sqrt{k}\log k) \\ |\hat{W}_i^r| &\leq e^{O_{w,b,G}(\sqrt{k}\log k)}.\end{aligned}$$

where the $O_{w,b,G}(\cdot)$ notation hides factors depending on w, b, G.

Proof. We begin with the upper bound on B_i. Our starting point is the penultimate line of Eq. (31), which implies that $B_i = O_{w,b,G}(||\Delta||_\infty)$. Using

$$\Delta = \log R - \log \rho - V(F-P),$$

we immediately obtain that $||\Delta||_\infty = O(\log k)$ with probability one, since the smallest F_{ij} will be on the order of $1/k$ due the lines 5 and 7 of the WLSM.

Next, we turn to the bound on C_i. Starting from Eq. (49), we use that $||\log R - \log \rho||_\infty = O_{w,b,G}(\log k)$. Since $u_{ij} = z_{ij}(1-z_{ij})$, the quantity z_{ij} is on the path between F_{ij} and p_{ij}, and F_{ij} cannot be smaller than $(1/2)/k$, we have that $1/\min_{(a,b)} u_{a,b} = O(k)$ with probability one. Plugging these observations into Eq. (49) completes the proof.

Finally, we turn to the bound on Y_i. Recall that Eq. (17) states that

$$\hat{Y} = \left(\mathrm{diag}(w)^{-1} L_{\hat{V}} \mathrm{diag}(w)^{-1}\right)^\dagger \mathrm{diag}(w)^{-1} M \hat{V}^{-1} \log R.$$

Moreover, Lemma 7 shows that

$$\left|\left|\mathrm{diag}(q)^{-1} M^T \mathrm{diag}(w)^{-1} (\mathrm{diag}(w)^{-1} L_q \mathrm{diag}(w)^{-1})^\dagger e_i\right|\right|_1 \leq w_i \sqrt{\sum_{(i,j)} q_{ij}^{-1} R_{max}(i)}.$$

Since we have already shown that $\log R = O(\log k)$ with probability one, and since $\sum_{(i,j)} \hat{V}_{ij}^{-1} = O(k)$ with probability one, this proves the bound we need.

Finally, the bound on W_i^r follows from its definition,

$$\log \hat{W}^r = \mathrm{diag}(w)^{-1} \hat{Y},$$

together with the bound on $\hat{Y}_i$.

□

Our next step is to argue that a sufficiently high moment of the quantities B_i, C_i decays fast. We will rely on such moments in the ensuing analysis. It will suffice to use the fourth moment, as in the following lemma.

Lemma 14. *For all $i = 1, \ldots, n$,*

$$E[B_i^4] = O_{w,b,G}\left(\frac{1}{k^4}\right)$$
$$E[C_i^4] = O_{w,b,G}\left(\frac{1}{k^4}\right)$$

where the $O_{w,b,G}(\cdot)$ notation hides all the factors that don't depend on δ and k.

Proof. In Eq. (31) we have shown that with probability $1-\delta$,

$$B_i^4 \leq O_{w,b,G}\left(\frac{\log^4(1/\delta)}{k^4}\right). \tag{50}$$

This has been shown subject to the conditions that $k \geq \Omega(b \log(n/\delta))$ and $\delta \leq e^{-1}$. We will turn this into a bound on the expectation of B_i by using the identity

$$E[B_i^4] = \int_0^{+\infty} P(B_i^4 \geq u)\, du.$$

Our first step is to rephrase Eq. (50) to bound the integrand as follows. Given any u, we solve for the δ we must plug into Eq. (50) in order to bound $P(B_i^4 \geq u)$. This yields

$$P(B_i^4 \geq u) \leq e^{-(k^4 u / f_1(w,b,G))^{1/4}},$$

for some function $f_1(w,b,G)$. However, this does not hold for all u because we have to account for the conditions $\delta \leq e^{-1}$ and $k \geq \Omega(b\log(n/\delta))$. The former holds when $u \geq f_1(w,b,G)/k^4$; and the latter holds if

$$c\frac{1}{b^4} \geq \frac{\log^4(n/\delta)}{k^4}.$$

for some absolute constant c. Since

$$\begin{aligned}\frac{\log^4(n/\delta)}{k^4} &\leq \frac{8\log^4 n + 8\log^4(1/\delta)}{k^4} \\ &\leq \frac{64 \log^4 n \log^4(1/\delta)}{k^4} \\ &= \frac{64(\log^4 n)u}{f_1(w,b,G)},\end{aligned}$$

we can ensure that the condition $k \geq \Omega(b\log(n/\delta))$ holds by taking u satisfying

$$u \leq \frac{c f_1(w,b,G)}{64 b^4 \log^4 n} = c_1 f_1(w,b,G)(b\log n)^{-4}.$$

With all this in mind, we argue as follows:

$$\begin{aligned}E[B_i^4] = &\int_0^{f_1(w,b,G)/k^4} P(B_i^4 \geq u)\, du \\ &+ \int_{f_1(w,b,G)/k^4}^{c_1 f_1(w,b,G)(b\log n)^{-4}} P(B_i^4 \geq u)\, du \\ &+ \int_{c_1 f_1(w,b,G)(b\log n)^{-4}}^{+\infty} P(B_i^4 \geq u)\, du \\ \leq &\frac{f_1(w,b,G)}{k^4} + \int_{f_1(w,b,G)/k^4}^{c_1 f_1(w,b,G)(b\log n)^{-4}} e^{-(k^4 u / f_1(w,b,G))^{1/4}}\, du \\ &+ \int_{c_1 f_1(w,b,G)(b\log n)^{-4}}^{O_{w,b,G}(\log k)} P(B_i^4 \geq u)\, du\end{aligned}$$

where in the last term we used Lemma 13 to replace the upper bound in the final integral. We next use Eq (30) to obtain

$$\begin{aligned}E[B_i^4] \leq &\frac{f_1(w,b,G)}{k^4} \\ &+ \int_{f_1(w,b,G)/k^4}^{+\infty} e^{-(u/(f_1(w,b,G)/k^4))^{1/4}}\, du \\ &+ O\left(n e^{-k^{3/4}}\right) O_{w,b,G}(\log k) \\ = &O_{w,b,G}\left(\frac{1}{k^4}\right) + \int_1^{+\infty} e^{-x^{1/4}} \frac{f_1(w,b,G)}{k^4} dx\end{aligned}$$

where we made the substitution $u = (f_1(w,b,G)/k^4)x$ in the integral. Using the fact that

$$\int_1^{+\infty} e^{-x^{1/4}}\, dx = \frac{64}{e}$$

implies that

$$E[B_i^4] = O_{w,b,G}\left(\frac{1}{k^4}\right),$$

and gives the needed upper bound B_i. The proof for C_i is similar. □

As a consequence of the previous lemma, we can bound the fourth moment of the quantity $\hat{Y}_i - y_i$ in the next lemma.

Lemma 15.

$$E(\hat{Y}_i - y_i)^4 = O\left(E[A_i^4]\right) + O_{w,b,G}\left(\frac{1}{k^4}\right)$$

Proof. Indeed,

$$\begin{aligned} E(\hat{Y}_i - y_i)^4 &= E[(A_i + B_i + C_i)^4] \\ &= E\left[O\left(A_i^4 + B_i^4 + C_i^4\right)\right] \\ &= O\left(E[A_i^4]\right) + O_{w,b,G}\left(\frac{1}{k^4}\right), \end{aligned}$$

where the second step follows by Young's inequality and the last step uses Lemma 14. □

Taking stock at this point, we are proceeding to bound the fourth moment of the quantity $\hat{Y}_i - y_i = A_i + B_i + C_i$. Our previous lemma reduces this to the fourth moment of A_i, up to terms that decay as fast as $O(1/k^4)$. We thus need to analyze the fourth moment of A_i, which is done in the following lemma.

Lemma 16.

$$\sum_{i=1}^n \sqrt{E[A_i^4]} \leq O\left(\frac{\mathrm{Tr}\left[(\mathrm{diag}(w^r)^{-1} L_V \mathrm{diag}(w^r)^{-1})^\dagger\right]}{k}\right)$$

Proof. Our starting point is the equations Eq. (26) and Eq. (27). Those equations allow us to write

$$A = \frac{1}{\sqrt{k}} P\left(\sqrt{k} V^{1/2}(F - p)\right).$$

Here the matrix P is defined in Eq. (27). What we need to show is that

$$\sum_{i=1}^n \sqrt{E[A_i^4]} = O\left(\frac{\mathrm{Tr}(PP^T)}{k}\right).$$

Equivalently, we need to show that

$$\sum_{i=1}^n \sqrt{E[A_i^4]} = O\left(\frac{||P||_F^2}{k}\right),$$

where $||P||_F$ is the Frobenius norm of P. Letting $Z = \sqrt{k} V^{1/2}(F - p)$ we have that

$$A_i = \frac{1}{\sqrt{k}} \sum_{j=1}^n P_{ij} Z_j,$$

and therefore

$$\begin{aligned} E[A_i^4] &= \frac{1}{k^2} E \sum_{j,q,l,m=1}^n P_{ij} P_{iq} P_{il} P_{im} Z_j Z_q Z_l Z_m \\ &= \frac{1}{k^2} \sum_{q,l=1}^n P_{iq}^2 P_{il}^2 O(1), \end{aligned}$$

because $E[Z_m] = 0$ for all $m = 1, \ldots, n$, so that only terms of the form $Z_q^2 Z_l^2$ or Z_q^4 "survive" the expectation. Thus

$$\begin{aligned} \sqrt{E[A_i^4]} &= \frac{1}{k} O\left(\sqrt{\sum_{q,l=1}^N P_{iq}^2 P_{il}^2}\right) \\ &= \frac{1}{k} O\left(\sqrt{\left(\sum_{j=1}^n P_{ij}^2\right)^2}\right) \\ &= \frac{1}{k} O\left(\sum_{j=1}^n P_{ij}^2\right). \end{aligned}$$

It follows that

$$\sum_{i=1}^n \sqrt{E[A_i^4]} = O\left(\frac{||P||_F^2}{k}\right),$$

and we are done.

□

We are almost ready to put together all the pieces and prove the final theorem. It turns out that we need a technical estimate on how big the ratio $\hat{W}_i^r / w_i^r$ can be; this will be needed to bound various worst-case events. Our next lemma shows that the expectation of the fourth power of this quantity is constant. This will be helpful in the proof of our main theorem, where we will at one point need to interchange these quantities.

Lemma 17. *For large enough k and all $i = 1, \ldots, n$, we have that*

$$E\left[\frac{\max\left((w_i^r)^4, (\hat{W}_i^r)^4\right)}{(w_i^r)^4}\right] = O(1).$$

Proof. Indeed, from

$$\hat{Y} - y = A + B + C,$$

we have

$$\sum_{i=1}^{n}(\hat{Y}_i - y_i)^2 \leq 4\sum_{i=1}^{n}(A_i^2 + B_i^2 + C_i^2)$$

Using Eq. (18) and Eq. (21) to obtain

$$\hat{Y}_i - y_i = w_i(\log \hat{W}_i^r - \log w_i^r), \tag{51}$$

we obtain

$$\sum_{i=1}^{n}\log^2\frac{\hat{W}_i^r}{w_i^r} \leq \frac{4}{w_{\min}^2}\sum_{i=1}^{n}A_i^2 + B_i^2 + C_i^2.$$

This means that

$$\begin{aligned}
P\left(\max_i \frac{\hat{W}_i^r}{w_i^r} \geq e^4\right) &= P\left(\max_i \log\frac{\hat{W}_i^r}{w_i^r} \geq 4\right) \\
&\leq P\left(\max_i \log^2\frac{\hat{W}_i^r}{w_i^r} \geq 16\right) \\
&\leq P\left(\sum_i \log^2\frac{\hat{W}_i^r}{w_i^r} \geq 16\right) \\
&\leq P(||A||_2^2 > w_{\min}^2) + P(||B||_2^2 > w_{\min}^2) \\
&\quad + P(||C||_2^2 > w_{\min}^2)
\end{aligned}$$

By Eq. (25), Eq. (30), and Eq. (45), we have that

$$P\left(\max_i \left|\frac{\hat{W}_i^r}{w_i^r}\right| \geq e^4\right) = O\left(e^{-k/g(w,b,G)}\right) + O\left(ne^{-k^{3/4}}\right)$$

Moreover, by Lemma 13 we have that with probability one

$$\max_i W_i^r \leq e^{O_{b,w,G}\left(\sqrt{k}\log k\right)}.$$

Putting this all together,

$$E\max_i\left|\frac{(\hat{W}_i^r)^4}{(w_i^r)^4}\right| \leq (e^4)^4 + e^{O_{b,w,G}\left(\sqrt{k}\log k\right)}\left[O\left(e^{-k/g(w,b,G)}\right) + O\left(ne^{-k^{3/4}}\right)\right]$$

For large enough k, we therefore have

$$E\max_i\left|\frac{(\hat{W}_i^r)^4}{(w_i^r)^4}\right| \leq e^{16} + 1,$$

which implies the lemma. □

With all these results in place, we can finally prove our first main result. As we will see, our concentration results on B_i, C_i will allow us to argue we can essentially ignore them when k is large enough; and our bound on the expectation of A_i will turn out to give us exactly the expression in Theorem 1.

Proof of Theorem 1. Using the inequality

$$\left|e^a - e^b\right| \leq \max(e^a, e^b)|a - b|,$$

we obtain that

$$|\hat{W}_i^r - w_i^r| \leq \max(w_i^r, \hat{W}_i^r)|\log\hat{W}_i^r - \log w_i^r|. \tag{52}$$

We thus have that, using Lemma 17, Eq. (51, and Lemma 16, Eq. (52) and Lemma 15,

$$\begin{aligned}
E||\hat{W}^r - w^r||_2^2 &\leq E\sum_{i=1}^{n}\max(w_i^r, \hat{W}_i^r)^2(\log\hat{W}_i^r - \log w_i^r)^2 \\
&= E\sum_{i=1}^{n}\frac{\max(w_i^r, \hat{W}_i^r)^2}{(w_i^r)^2}(w_i^r)^2(\log\hat{W}_i^r - \log w_i^r)^2 \\
&\leq \sum_{i=1}^{n}\sqrt{E\frac{\max(w_i^r, \hat{W}_i^r)^4}{(w_i^r)^4}} \\
&\quad \sqrt{E(w_i^r)^4(\log\hat{W}_i^r - \log w_i^r)^4} \\
&= \sum_{i=1}^{n}O(1)\sqrt{E(\hat{Y}_i - y_i)^4} \\
&= O\left(\sum_{i=1}^{n}\sqrt{E[A_i^4] + O_{w,b,G}\left(\frac{1}{k^4}\right)}\right) \\
&= O\left(\sum_{i=1}^{n}\sqrt{E[A_i^4]}\right) + O_{w,b,G}\left(\frac{1}{k^2}\right) \\
&= O\left(\frac{\mathrm{Tr}\left[(\mathrm{diag}(w^r)^{-1}L_V\mathrm{diag}(w^r)^{-1})^\dagger\right]}{k}\right) \\
&\quad + O_{w,b,G}\left(\frac{1}{k^2}\right) \\
&= O\left(\frac{\mathrm{Tr}\left[(\mathrm{diag}(w^r)^{-1}L_V\mathrm{diag}(w^r)^{-1})^\dagger\right]}{k}\right),
\end{aligned}$$

where the last step holds for k large enough; and which further implies that, for k large enough,

$$\begin{aligned}
E\left[\sin^2(\hat{W}, w)\right] &\leq E\frac{||W^r - w^r||_2^2}{||w^r||_2^2} \\
&\leq O\left(\frac{\mathrm{Tr}\left[(\mathrm{diag}(w^r)^{-1}L_V\mathrm{diag}(w^r)^{-1})^\dagger\right]}{k||w^r||_2^2}\right)
\end{aligned} \tag{53}$$

where we used the following identity about the sine between two vectors:

$$|\sin(x, y)| = \inf_\alpha\frac{||\alpha x - y||_2}{||y||_2} \leq \frac{||x - y||_2}{||y||_2}.$$

Finally, the final expression on the right-hand side of Eq. (53) is unaffected by replacing w^r with w, since both numerator and denominator are scaled by the same number.

Thus for large enough k,

$$E\left[\sin^2(\hat{W},w)\right] \leq O\left(\frac{\mathrm{Tr}\left[(\mathrm{diag}(w)^{-1}L_V\mathrm{diag}(w)^{-1})^\dagger\right]}{k||w||_2^2}\right).$$

We complete the proof by observing that $L_\gamma = \mathrm{diag}(w)^{-1}L_v\mathrm{diag}(w)^{-1}$. □

Proof of Theorem 2

We will establish that, for large enough k,

$$E\left[\sin^2(w,\hat{w})\right] \geq \Omega\left(\frac{1}{k}\right)\frac{\mathrm{Tr}\left[L_\gamma^\dagger\right]}{||w||_2^2}, \tag{54}$$

for any estimator $\hat{w}$ built from the outcomes of pairwise comparisons. Our proof is a modification of the proof of the lower bound from (Hendrickx *et al.*, 2019), with departures at key steps. As discussed in the main body of the paper, the main departure is to specifically pick out the solution $\hat{W}^r$ in the lower bound analysis.

We will generate w from a distribution μ, to be specified later. We will use $P_w(y)$ to denote the density on the observation space (consisting of k measurements across each edge of the graph) if w was the vector of true weights. We will use the following lemma [(Hajek & Raginsky, 2019) Chap. 13, Corollary 13.2] to obtain a lower bound on the expectation of the sine-squared:

Lemma 4.1. *Let μ be any joint probability distribution of a random pair (w,w'), such that the marginal distributions of both w and w' are equal to π. Then*

$$\mathbb{E}_{\pi,\mathbf{Y}}[d(w,\hat{w}(\mathbf{Y}))] \geq \mathbb{E}_\mu\left[d(w,w')(1-\|P_w - P_{w'}\|_{TV}\right]$$

where $||\cdot||_{\mathrm{TV}}$ represents the total-variation distance between distributions.

It should be clear that under a random choice of w generated according to some distribution π (described later), the expected error is a lower bound on the worst-case estimation error over all possible w. Thus our goal is to massage the right-hand side of Lemma 4.1 to obtain the right-hand side of Eq. (54).

Actually, we need a slight modification of Lemma 4.1: as remarked in (Hendrickx *et al.*, 2019), it is sufficient that $d(w,w')$ satisfies a weak version of triangle inequality, i.e., $\alpha d(w_1,w_2) \leq d(w_1,\hat{w}) + d(w_2,\hat{w})$ for some pre-specified constant α, with the result that the right-hand side in the above lemma is multiplied by α. In particular, our (square) error criterion $\sin^2(\hat{w},w_z)$ satisfies the weak triangle inequity with a factor of $\alpha = 1/2$, see Lemma A.1 from (Hendrickx *et al.*, 2019), so we can apply Lemma 4.1 to it with an extra factor of $1/2$ on the right-hand side.

Let v_i be the eigenvectors of the $\mathrm{diag}(w)^{-1}L_V\mathrm{diag}(w)^{-1}$ with corresponding eigenvalues σ_i. In the next paragraph, we will use these eigenvectors to design the distribution for w which we will use to obtain our lower bound. Note that this is the first point where our argument diverges from the proof of (Hendrickx *et al.*, 2019); the introduction of this rescaling by $\mathrm{diag}(w)^{-1}$ here is motivated by Eq. (17) and Eq. (18), where the quantity $\mathrm{diag}(w)^{-1}L_V\mathrm{diag}(w)^{-1}$ appears, and comes from a desire to lower bound the error associated with the regularized solution $\hat{W}^r$.

Let $z_2,\ldots,z_n$ be i.i.d random variable taking values 1 and -1 with equal probability. We then set

$$w_z = w + \delta\sum_{i=2}^n \frac{z_i}{\sqrt{\sigma_i}}v_i \tag{55}$$

where, the sum starts at $i=2$ to omit the eigenvector of $\mathrm{diag}(w)^{-1}L_V\mathrm{diag}(w)^{-1}$ associated with the zero eigenvalue (which is just w), δ is a suitably small number to be specified later, and also we set $z_1 = 1$. Let V be the unitary matrix which has v_i as columns; we can write

$$w_z = V\Lambda z,$$

where this relation defines the entries of Λ (e.g., $\lambda_i = \delta/\sqrt{\sigma_i}$ for $i=1,\ldots,n$). We note that the norm of w_z's defined this way are equal, i.e.,

$$\begin{aligned}\|w_z\|_2 &= \sqrt{||w||_2^2 + \delta^2\sum_{i=2}^n\frac{1}{\sigma_i}} \\ &= \sqrt{||w||_2^2 + \delta^2\mathrm{Tr}\left[(\mathrm{diag}(w)^{-1}L_v\mathrm{diag}(w)^{-1})^\dagger\right]}.\end{aligned} \tag{56}$$

Intuitively the error in estimating w_z should be lower bounded in terms of the errors in estimating z_i, and indeed (Hendrickx *et al.*, 2019) showed that

$$\min_{\hat{w}(\mathbf{Y})}\mathbb{E}_{\pi,\mathbf{Y}}[\rho(w_z,\hat{w}(\mathbf{Y}))] = \sum_{i=2}^n\min_{\eta_i(\mathbf{Y})}\frac{\lambda_i^2}{\|w_z\|^2}\mathbb{E}_{\pi,\mathbf{Y}}\left(z_i-\eta_i(\mathbf{Y})\right)^2,$$

where $\mathbf{Y}$ is the vector of outcomes of the comparisons. We are now going to apply Lemma 4.1 to each term on the right hand side individually. Following (Hendrickx *et al.*, 2019), we define the distribution $\mu_i(z,z')$ by keeping z uniformly distributed in $\{-1,1\}^n$, and flipping the i^{th} bit to obtain z' (formally, $z'_i = -z_i$ and $z'_j = z_j$ for every $j \neq i$). Clearly, $\mathbb{E}_{\pi,\mathbf{Y}}d_i(z,z') = 4$. Moreover, by Pinsker's inequality

$$\begin{aligned}\|P_w^{\otimes k} - P_{w'}^{\otimes k}\|_{\mathrm{TV}}^2 &\leq \frac{1}{2}D_{KL}(P_w^{\otimes k}\|P_w'^{\otimes k}) \\ &\leq O(k\delta^2)\end{aligned} \tag{57}$$

where the proof of the second inequality (which holds for small enough δ) is somewhat involved and is relegated to Section 4.4 below. Using these facts along with Lemma 4.1, it follows that for every estimator $\eta_i(\mathbf{Y})$ and for such δ,

$$\mathbb{E}_{\pi,\mathbf{Y}}\left(z_i-\eta_i(\mathbf{Y})\right)^2 \geq \frac{1}{2}4\left(1-\sqrt{O(k\delta^2)}\right),$$

and thus

$$\min_{\hat{w}(\mathbf{Y})} \mathbb{E}_{\pi,\mathbf{Y}}[\rho(w,\hat{w}(\mathbf{Y}))] \geq \sum_{i=2}^{n} \frac{\lambda_i^2}{\|w_z\|^2} 2(1-\sqrt{O(k\delta^2)}) \geq \sum_{i=2}^{n} \frac{2\delta^2(1-\sqrt{O(k\delta^2)})}{\sigma_i S}, \quad (58)$$

where

$$S = ||w||_2^2 + \delta^2 \mathrm{Tr}\left[\left(\mathrm{diag}(w)^{-1} L_v \mathrm{diag}(w)^{-1}\right)^{\dagger}\right]$$

Now choosing δ such that the $O(k\delta^2)$ term is at most $1/2$ (which involves choosing $\delta^2 = \Theta(1/k)$), and further k is large enough so that

$$||w||^2 + \delta^2 \mathrm{Tr}\left[\left(\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1}\right)^{\dagger}\right] < 2||w||_2^2, \quad (59)$$

Putting all these bounds into Eq. (58) yields

$$\min_{\hat{w}(\mathbf{Y})} \mathbb{E}_{\pi,\mathbf{Y}}[\rho(w,\hat{w}(\mathbf{Y}))] \geq \Omega\left(\frac{1}{k}\right) \frac{\sum_{i=2}^{n} 1/\sigma_i}{||w||_2^2}.$$

Using the definition of σ_i, we can rewrite this as

$$\min_{\hat{w}(\mathbf{Y})} \mathbb{E}_{\pi,\mathbf{Y}}[\rho(w,\hat{w}(\mathbf{Y}))] \geq \Omega\left(\frac{1}{k}\right) \frac{\mathrm{Tr}\left[\left(\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1}\right)^{\dagger}\right]}{||w||_2^2}.$$

Finally noting that

$$\mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1} = L_\gamma,$$

we have thus proved Eq. (54) (conditionally on Eq. (57) which is considered in the next section).

4.4. Proof of Equation (57)

We will need to begin with several lemmas. Let $B(w_1, w_2)$ be our notation for a Bernoulli that falls on heads with probability of $w_1/(w_1+w_2)$. We will need some bounds on how the KL-divergence evolves as we perturb w_1, w_2.

Lemma 18. *Fix positive w_1, w_2. For small enough δ,*

$$D_{KL}(B(w_1(1+\delta x_1), w_2(1+\delta x_2))||B(w_1(1-\delta x_1), w_2(1-\delta x_2))) \leq \frac{5\delta^2}{V_{12}}(x_1-x_2)^2$$

where we use the notation $V_{12} = \frac{w_1}{w_2} + 2 + \frac{w_2}{w_1}$ (consistent with our previous usage).

Proof. Let us introduce notations for the probabilities associated with the two Bernoulli distribution, the first corresponding to $z=1$ and the second to $z=-1$:

$$p = \frac{w_1(1+\delta x_1)}{w_1(1+\delta x_1)+w_2(1+\delta x_2)}$$
$$p' = \frac{w_1(1-\delta x_1)}{w_1(1-\delta x_1)+w_2(1-\delta x_2)}$$

Applying Lemma 7.2 of (Bubeck, 2011), we have the estimate

$$D_{KL}(p||p') \leq \frac{(p-p')^2}{p'(1-p')}$$

We consider what this is like in the limit as $\delta \to 0$, as this leads to a number of simplifications. First, consider the denominator: we have that

$$\lim_{\delta\to 0} \frac{1}{p'(1-p')} = \frac{1}{(w_1/(w_1+w_2))(w_2/(w_1+w_2))} = V_{12}.$$

We conclude that, for δ small enough,

$$D_{KL}(p||p') \leq \left(\frac{5}{4}\right)^{1/3} V_{12}(p-p')^2.$$

Of course, the constant in front of the right-hand side can be chosen to be any number greater than one.

Next, let us consider the difference of the two probabilities:

$$\begin{aligned} p-p' &= \frac{w_1}{w_1 + w_2\frac{1+\delta x_2}{1+\delta x_1}} - \frac{w_1}{w_1 + w_2\frac{1-\delta x_2}{1-\delta x_1}} \\ &= \frac{w_1 w_2\left(\frac{1-\delta x_2}{1-\delta x_1} - \frac{1+\delta x_2}{1+\delta x_1}\right)}{(w_1 + w_2\frac{1+\delta x_2}{1+\delta x_1})(w_1 + w_2\frac{1-\delta x_2}{1-\delta x_1})} \\ &:= C(w_1,w_2,\delta)\left(\frac{1-\delta x_2}{1-\delta x_1} - \frac{1+\delta x_2}{1+\delta x_1}\right) \end{aligned}$$

Observing that

$$\lim_{\delta\to 0} C(w_1,w_2,\delta) = \frac{1}{V_{12}},$$

we can conclude that, for δ small enough,

$$\begin{aligned} D_{KL}(p||p') &\leq \left(\frac{5}{4}\right)^{1/3} V_{12} \left(\frac{5}{4}\right)^{1/3} \frac{1}{V_{12}^2}\left(\frac{1-\delta x_2}{1-\delta x_1} - \frac{1+\delta x_2}{1+\delta x_1}\right)^2 \\ &= \frac{(5/4)^{2/3}}{V_{12}}\left(\frac{1-\delta x_2}{1-\delta x_1} - \frac{1+\delta x_2}{1+\delta x_1}\right)^2 \end{aligned}$$

Finally, observe that the function

$$f(t) = \frac{1+tx_2}{1+tx_1} = \frac{1+tx_1+t(x_2-x_1)}{1+tx_1} = 1 + t\frac{x_2-x_1}{1+tx_1}$$

clearly satisfies $f'(0) = x_2 - x_1$. Consequently, for small enough δ, we have that

$$\left(\frac{1-\delta x_2}{1-\delta x_1} - \frac{1+\delta x_2}{1+\delta x_1}\right)^2 \leq \left(\frac{5}{4}\right)^{1/3} ((x_2-x_1)2\delta)^2.$$

Putting it all together, we have that

$$D_{KL}(p||p') \leq \frac{5}{V_{12}}\delta^2(x_2-x_1)^2.$$

□

Corollary 19. *Fix u_1, u_2 positive and arbitrary a_1, a_2, x_1, x_2. Consider the same situation as Lemma 18 except that the weights of node $j = 1, 2$ are*

$$w_j = u_j + \delta a_j + \delta z x_j,$$

where z is either $+1$ or -1. The KL divergence between the corresponding Bernoulli random variables is upper bounded by

$$\frac{5\delta^2}{V_{12}} \left(\frac{x_1}{u_1} - \frac{x_2}{u_2} \right)^2 + O(\delta^3),$$

with $V_{12} = \frac{u_1}{u_2} + 2 + \frac{u_2}{u_1}$

Proof. The weight of node i can be rewritten as

$$u_j + \delta a_j + \delta z x_j, = u_j \left(1 + \delta \frac{a_j}{u_j} \right) \left(1 + z\delta \frac{x_j/u_j}{1 + \delta a_j/u_j} \right).$$

We can then apply Lemma 18 with the modified (bounded) parameters $\tilde{w}_j = u_j(1 + \delta a_j/u_j)$ and $\tilde{x}_j = \frac{x_j/u_j}{1+\delta a_j/u_j}$, and we obtain that the KL divergence is (with $V_{w,12}$ the variance for the weights $\tilde{w}_j$)

$$\begin{aligned} D_{KL} &= \frac{5\delta^2}{V_{w,12}} (\tilde{x}_1 - \tilde{x}_2)^2 \\ &= \frac{5\delta^2}{V_{w,12}} \left(\frac{x_1/u_1}{1 + \delta a_1/u_1} - \frac{x_2/u_2}{1 + \delta a_2/u_2} \right)^2 \\ &= \frac{5\delta^2}{V_{w,12}} \left(\frac{x_1}{u_1} - \frac{x_2}{u_2} + O(\delta) \right)^2 \\ &= \frac{5\delta^2}{V_{w,12}} \left(\frac{x_1}{u_1} - \frac{x_2}{u_2} \right)^2 + O(\delta^3). \end{aligned} \tag{60}$$

Besides, observe that

$$\begin{aligned} \frac{\tilde{w}_1}{\tilde{w}_2} &= \frac{u_1}{u_2} \cdot \frac{1 + \delta a_1/u_1}{1 + \delta a_2/u_2} \\ &= \frac{u_1}{u_2} + O(\delta). \end{aligned}$$

Hence

$$V_{w,12} = \frac{\tilde{w}_1}{\tilde{w}_2} + 2 + \frac{\tilde{w}_1}{\tilde{w}_2} = V_{u,12} + O(\delta).$$

The result follows then from (60). □

With these facts in place, we can now prove the equation to which this subsection is dedicated.

Proof of Equation (57). We can apply Corollary 19 across each edge, with $u_j = w_j$, $x_j = v_i$, and $a_j = \sum_{j \neq i} \lambda_j v_j$ to argue as follows:

$$\begin{aligned} D_{KL}(P_w^{\otimes k} \| P_w'^{\otimes k}) &= k D_{KL}(P_w \| P_w') \\ &\leq \sum_{(a,b) \in E} O\left(\frac{k\delta^2}{V_{ab}} \right) \left(\frac{(v_i)_a}{\sqrt{\sigma_i} w_a} - \frac{(v_i)_b}{\sqrt{\sigma_i} w_b} \right)^2 + O(k\delta^3) \\ &= O(k\delta^2) \frac{1}{\sigma_i} v^T \mathrm{diag}(w)^{-1} L_V \mathrm{diag}(w)^{-1} v + O(k\delta^3) \\ &= O(k\delta^2), \end{aligned}$$

where we used that v_i is an eigenvector of $\mathrm{diag}(w)^{-1} L_v \mathrm{diag}(w)^{-1}$ with eigenvalue σ_i.

□