ℓ_∞-Bounds of the MLE in the BTL Model under General Comparison Graphs

Abstract

The Bradley-Terry-Luce (BTL) model is a popular statistical approach for estimating the global ranking of a collection of items of interest using pairwise comparisons. To ensure accurate ranking, it is essential to obtain precise estimates of the model parameters in the ℓ_∞-loss. The difficulty of this task depends crucially on the topology of the pairwise comparison graph over the given items. However, beyond very few well-studied cases, such as the complete and Erdös-Rényi comparison graphs, little is known about the performance of the maximum likelihood estimator (MLE) of the BTL model parameters in the ℓ_∞-loss under more general graph topologies. In this paper, we derive novel, general upper bounds on the ℓ_∞ estimation error of the BTL MLE that depend explicitly on the algebraic connectivity of the comparison graph, the maximal performance gap across items and the sample complexity. We demonstrate that the derived bounds perform well and in some cases are sharper compared to known results obtained using different loss functions and more restricted assumptions and graph topologies. We carefully compare our results to Yan et al. (2012), which is closest in spirit to our work. We further provide minimax lower bounds under ℓ_∞-error that nearly match the upper bounds over a class of sufficiently regular graph topologies. Finally, we study the implications of our ℓ_∞-bounds for efficient (offline) tournament design. We illustrate and discuss our findings through various examples and simulations.

1 INTRODUCTION

Simultaneous or 'global' ranking of a set of items is a practical problem that arises naturally in a variety of domains. For example, one may wish to ascertain a 'best player' or 'best team' in a given sports league. Designing a principled statistical approach to global ranking of items is challenging due to data limitations and complex domain-specific relationships between the underlying items to be ranked.

A popular and practicable solution to estimating global ranking is to utilize pairwise comparison information across the items to be ranked, which is easily accessible across many application domains. The BTL model (Bradley and Terry, 1952; Luce, 1959) is a popular statistical model for pairwise comparison data. A similar model was also originally studied in Zermelo (1929). The continued practical and theoretical interest in the BTL model stems from its relatively simple parametric form which provides a good balance between interpretability and tractability for theoretical analysis. The BTL model is domain-agnostic, making it an ideal benchmarking tool across a variety of ranking applications e.g. sports analytics (Fahrmeir and Tutz, 1994; Masarotto and Varin, 2012; Cattelan et al., 2013), and bibliometrics (Stigler, 1994; Varin et al., 2016).

Formally, we can describe the BTL model as follows. Suppose that we have n distinct items, each with a (fixed but unobserved) positive strength or preference score w_i, $i \in [n]$, quantifying item i's propensity to beat other items in pairwise comparisons. The BTL model assumes that the comparisons between different pairs are independent and the outcomes of comparisons between any given pair, say item i and item j, are i.i.d. Bernoulli random variables, with *winning probability* p_{ij}, defined as

$$p_{ij} = \mathbf{Pr}\,(i \text{ beats } j) = \frac{w_i^*}{w_i^* + w_j^*},\ \forall\, i, j \in [n]. \quad (1)$$

A common reparametrization is to set, for each i, $w_i^* = \exp(\theta_i^*)$, where $\boldsymbol{\theta}^* := (\theta_1^*, \ldots, \theta_n^*)^\top \in \mathbb{R}^n$. By convention, we assume that $\sum_{i\in[n]} \theta_i^* = 0$ for parameter identifiability.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

General pairwise comparison graphs

Given n items to be compared, the pairwise comparison scheme among them can be expressed through an undirected simple graph $\mathcal{G}(V, E)$, where the vertex set $V := [n]$ and the edge set $E := \{(i, j) : i \text{ and } j \text{ are compared }\}$ is determined by the comparison scheme. Correspondingly, if we define the directed edge set as $E_d := \{(i, j, k) : (i \text{ beats } j)\ k \text{ times}\}$, then the induced directed simple graph $\mathcal{G}(V, E_d)$ is called a *directed* comparison graph. It is a classical result (Ford, 1957; Simons and Yao, 1999; Hunter, 2004) that the BTL model is identifiable if and only if $\mathcal{G}(V, E)$ is connected, and the maximum likelihood estimator (MLE) of the model parameters exists and is consistent if and only if $\mathcal{G}(V, E_d)$ is strongly connected. Henceforth, *comparison graph* refers to the undirected pairwise comparison graph.

From a statistical theory perspective, much attention in the BTL literature has been paid to two popular estimators, namely the MLE and the spectral method. Typically one is interested in getting sharp bounds for the estimation risk, which could be based on a norm-induced metric $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_p$ or a ranking metric, e.g., Kendall's tau distance (Kendall, 1938). What makes risk analysis of BTL model estimators particularly challenging is a combination of the type of estimation risk loss considered, and the assumptions on the pairwise comparison graph $\mathcal{G}(V, E)$ topology.

Core questions of interest

Among all the metrics measuring uncertainty of estimators of BTL parameters, the ℓ_∞-loss directly connects with ranking metrics, e.g. binary and Hamming top-k (partial) ranking loss (see, e.g. Chen et al., 2019, 2020). Moreover, Chen et al. (2020) show that the MLE attains a sharper minimax rate of the Hamming top-k loss compared to the spectral method.

It is thus natural to study the MLE for the BTL parameters in the ℓ_∞-loss, to better understand the risk optimality of the MLE and further justify its use for practical global and partial ranking problems. In this spirit, Yan et al. (2012) focus specifically on proving ℓ_∞-error bounds for the BTL MLE for general comparison graphs. However a notable limitation in their setting is that they impose a strictly dense comparison graph assumption, which may be impractical in many real world applications. This leaves a gap in the literature, summarized in the following questions:

> ***Core questions:*** *For the BTL model, how does the MLE perform with respect to the ℓ_∞ loss, under much weaker assumptions on the pairwise comparison graph compared to Yan et al. (2012)? That is, assuming only that the comparison graph is connected. Moreover, what are the implications of such bounds in applications?*

Providing a sharp analysis to these questions with a detailed comparison to recent theoretical results in the BTL literature motivates our work in this paper.

Relevant and related literature

We give a brief overview of the work that addresses the challenge of comparison graph topology in ranking. When the comparison graph is a complete graph, Simons and Yao (1999) give a high-probability upper bound for the ℓ_∞ loss, i.e., $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty$ and obtain the asymptotic distribution of the MLE. In the setting where the comparison graph follows the Erdös-Rényi graph model, Chen and Suh (2015), Chen et al. (2019), Chen et al. (2020) and Han et al. (2020) derive high-probability upper bounds for the ℓ_∞ loss. Moreover, Chen et al. (2019) show that both MLE and spectral method are minimax optimal in terms of the binary top-k ranking loss i.e. whether the items with the highest k out of n preference scores are perfectly identified; Chen et al. (2020) consider a Hamming Loss for top-k items and show that the MLE is minimax optimal compared to the spectral method with differences arising in constant factors.

For a broader class of comparison graphs beyond complete and Erdös-Rényi graph, researchers have studied the explicit dependence of the estimation risk on graph topology. In particular, Yan et al. (2012) give a high-probability upper bound for the ℓ_∞-loss for relatively dense graphs. Hajek et al. (2014); Shah et al. (2016) give a high probability upper bound for the ℓ_2 or Euclidean loss $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$, establish upper and lower bounds of $\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ and show the minimax optimality of the constraint MLE across a wide range of graph topologies. Recently, Agarwal et al. (2018) give sharp upper bounds for a novel spectral method in the ℓ_1-loss $\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1$ for $\boldsymbol{\pi}^* = \mathbf{w}^*/\|\mathbf{w}^*\|_1$ instead of $\boldsymbol{\theta}^*$. Hendrickx et al. (2019, 2020) propose a weighted least square method to estimate $\mathbf{w}^*$ and prove a sharp upper bound for their estimator in $\mathbb{E}[\sin^2(\hat{\mathbf{w}}, \mathbf{w}^*)]$ or equivalently in $\mathbb{E}\|\hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|_2 - \mathbf{w}^*/\|\mathbf{w}^*\|_2\|_2^2$, in the sense that this upper bound matches a instance-wise lower bound up to constant factors.

Contributions

Our contributions in this paper are fourfold and are summarized as follows:

- **Upper bounds:** We derive a novel upper bound for the ℓ_∞-error of the regularized MLE in BTL model allowing for general graph topology. Our upper bounds hold under minimal assumptions on graph topologies, i.e., assuming only that the comparison graph is connected. Given such generality, we show our ℓ_∞ bound is tighter than existing results under a broad range of graph topologies, and works well in general. In particular, we carefully compare our work analytically and in simulation to Yan et al. (2012), which is closest in spirit to our work.

A minor corollary of our techniques results in the state of the art ℓ_2-loss bounds for the Erdös-Rényi graph.

- **Lower bounds:** We derive minimax lower bounds for BTL parameter estimation in ℓ_∞-loss. We analyze specific graph topologies satisfying certain regularity connectivity conditions under which the BTL MLE is nearly minimax optimal.

- **Implications for tournament design:** We show that the BTL MLE in ℓ_∞-loss satisfies a unique subadditivity property. We demonstrate how our ℓ_∞ bounds can exploit this property for efficient (offline) tournament design.

- **Extension to the unregularized BTL model:** We also extend our upper bounds under ℓ_∞-loss to the unregularized ('vanilla') BTL MLE, which is also frequently used in practice.

Due to the more complicated form of the vanilla BTL MLE upper bounds and space limitations, we present these analagous results and their proofs separately in Appendix A.7. Henceforth, MLE refers to the regularized BTL MLE unless stated otherwise. In addition to our theoretical contributions a core aspect throughout our paper is to emphasize the interpretability of our results, the associated assumptions, and implications for practical ranking tasks.

Organization of the paper

The rest of the paper is organized as follows. In Section 2 we present our main results for the upper bound in Theorem 1 and an interpretation of the key components of the bound. In Section 3, we discuss minimax lower bounds using the ℓ_∞ risk loss in Theorem 5. In Section 4, we show some practical implications of our results in efficient tournament design from a ranking perspective. In Section 5, we conduct extensive numerical simulations to validate the optimality of our bounds compared to related results in the literature.

Notation

Throughout this paper we use following notational conventions. We typically use lowercase for scalars e.g. $(x, y, z, \ldots)$, boldface lowercase for vectors e.g. $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$, and boldface uppercase for matrices $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots)$. We denote the finite set $\{1, \ldots, n\} =: [n]$. For asymptotics, we denote $x_n \lesssim y_n$ or $x_n = O(y_n)$ and $u_n \gtrsim v_n$ or $u_n = \Omega(v_n)$ if $\forall n$, $x_n \leq c_1 y_n$ and $u_n \geq c_2 v_n$ for some constants $c_1, c_2 > 0$. Similarly, $x_n \asymp y_n$ or $x_n = \Theta(y_n)$ means $\forall n$, $c_1 y_n \leq x_n \leq c_2 y_n$ for some constants $c_1, c_2 > 0$. We denote $\mathbf{e}_i$ as a vector whose entries are all 0 except that the i-th entry is 1. $a_n = o(b_n)$ means $a_n/b_n \to 0$ as $n \to \infty$ and conversely, $a_n = \omega(b_n)$ means $b_n/a_n \to 0$ as $n \to \infty$. We denote $\mathbf{1}_n \in \mathbb{R}^n$ to be a vector of ones.

2 UPPER BOUNDS

Recall that given n items to be compared, the comparison scheme among them defines the comparison graph $\mathcal{G}(V, E)$, where $V = [n]$ and $E = \{(i, j) : i \text{ and } j \text{ are compared }\}$. We denote the corresponding adjacency matrix as $A \in \mathbb{R}^{n \times n}$, and its $(i, j)^{\text{th}}$ entry is $A_{ij} := 1\{(i, j) \in E\}$. The associated (unnormalized) graph Laplacian is the symmetric, positive-semidefinite matrix $\mathcal{L}_{\mathbf{A}} := \mathbf{D} - \mathbf{A}$, where $\mathbf{D} = \text{diag}(n_1, \ldots, n_n)$, with $n_i := \sum_{j=1}^n A_{ij}$ the degree of node i. It is well known that the smallest eigenvalue of $\mathcal{L}_A$ is 0 with an eigenvector $\mathbf{1}_n$. Let $\lambda_2(\mathcal{L}_{\mathbf{A}})$ be the second smallest eigenvalue of $\mathcal{L}_{\mathbf{A}}$, known as the algebraic connectivity of $\mathcal{G}$ (Das, 2004), then $\mathcal{G}$ is connected if and only if $\lambda_2(\mathcal{L}_{\mathbf{A}}) > 0$. Following the standard in the BTL literature we assume a that for each edge (i, j) of the comparison graph, the corresponding items i and j are compared L times, each leading to an independent outcome $y_{ij}^{(l)} \in \{0, 1\}$, where $l \in [L]$. If pairs are compared different number of times, we take L to be the smallest number of pairwise comparisons over the edge set, as a worst-case scenario. The corresponding sample averages are denoted with $\bar{y}_{ij} = \frac{1}{L}\sum_{l=1}^L y_{ij}^{(l)}$ and are sufficient statistics for the model parameters. The ℓ_2-regularized MLE is defined as

$$\hat{\boldsymbol{\theta}}_\rho = \underset{\mathbf{1}_n^\top \boldsymbol{\theta} = 0}{\arg\min}\, \ell_\rho(\boldsymbol{\theta}; \mathbf{y}),\ \ell_\rho(\boldsymbol{\theta}; \mathbf{y}) = \ell(\boldsymbol{\theta}; \mathbf{y}) + \frac{\rho}{2}\|\boldsymbol{\theta}\|_2^2, \quad (2)$$

where $\ell(\boldsymbol{\theta}; \mathbf{y})$ is the negative log-likelihood, given by

$$\begin{aligned}\ell(\boldsymbol{\theta}; \mathbf{y}) := -\sum_{1 \leq i < j \leq n} & A_{ij}\{\bar{y}_{ij} \log \psi(\theta_i - \theta_j) \\ & + (1 - \bar{y}_{ij}) \log[1 - \psi(\theta_i - \theta_j)]\},\end{aligned} \quad (3)$$

and $t \in \mathbb{R} \mapsto \psi(t) = 1/[1 + e^{-t}]$ the sigmoid function.

Under this notational setup, we are ready to state the ℓ_∞ upper bound of the BTL MLE in Theorem 1.

Theorem 1. *Assume the BTL model with parameter $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ such that $\mathbf{1}_n^\top \boldsymbol{\theta}^* = 0$ and a comparison graph $\mathcal{G} = \mathcal{G}([n], E)$ with adjacency matrix $\mathbf{A}$, algebraic connectivity $\lambda_2(\mathcal{L}_{\mathbf{A}})$ and maximum and minimum degrees $n_{\max}$ and $n_{\min}$. Suppose that each pair of items $(i, j) \in E$ are compared L times. Let $\kappa = \max_{i,j} |\theta_i^* - \theta_j^*|$ and $\kappa_E = \max_{(i,j) \in E} |\theta_i^* - \theta_j^*|$ and set $\rho \geq c_\rho \kappa^{-2} e^{-2.5\kappa_E} n^{-4} n_{\max}^{1/2}$. Assume that $\mathcal{G}$ is connected or $\lambda_2(\mathcal{L}_A) > 0$. Then with probability at least $1 - O(n^{-4})$, the regularized MLE $\hat{\boldsymbol{\theta}}_\rho$ from (2) satisfies*

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_\infty \lesssim & \frac{e^{2\kappa_E}}{\lambda_2}\frac{n_{\max}}{n_{\min}}\left(\sqrt{\frac{n+r}{L}} + \rho\kappa\sqrt{\frac{n}{n_{\max}}}\right) \\ & + \frac{e^{\kappa_E}}{\lambda_2}\sqrt{\frac{n_{\max}(\log n + r)}{L}},\end{aligned} \quad (4)$$

$$\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_2 \lesssim \frac{e^{\kappa_E}}{\lambda_2}\left(\sqrt{\frac{n_{\max}(n+r)}{L}} + \rho\kappa\sqrt{n}\right) \quad (5)$$

where $\lambda_2 = \lambda_2(\mathcal{L}_{\mathbf{A}})$, $r = \kappa_E + \log\kappa$ *provided that* $L \leq n^8 e^{5\kappa_E}\max\{1,\kappa\}$, *and* L *is large enough so that the right hand side of Equation* (4) *is smaller than a sufficiently small constant* $C > 0$. *In particular, if we set* $\rho = c_\rho/\kappa\sqrt{n_{\max}/L}$ *for some* $c_\rho > 0$, *then*

$$\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_\infty \lesssim \frac{e^{2\kappa_E}}{\lambda_2}\frac{n_{\max}}{n_{\min}}\sqrt{\frac{n+r}{L}} + \frac{e^{\kappa_E}}{\lambda_2}\sqrt{\frac{n_{\max}(\log n + r)}{L}},$$
$$\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_2 \lesssim \frac{e^{\kappa_E}}{\lambda_2}\sqrt{\frac{n_{\max}(n+r)}{L}}. \tag{6}$$

As a brief sketch, the proof is based on a gradient descent procedure initialized at $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}^*$ and the idea is to control $\|\boldsymbol{\theta}^{(T)} - \hat{\boldsymbol{\theta}}_\rho\|_\infty$ using the linear convergence property and $\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}^*\|_\infty$ using the leave-one-out technique in Chen et al. (2019) and Chen et al. (2020). In fact, our work confirms that such a line of argument extends to more general graph topologies beyond the Erdös-Rényi graph, which is non-trivial. The proof details can be found in Appendix A.2.

Interpretation of key terms

The upper bound in Equation (4) contains several distinct terms, which interact with each other in non-trivial ways and express different aspects of the intrinsic difficulty of the estimation task.

- The factor $\frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_{\mathbf{A}})}$ combines two sources of statistical hardness: the maximal gap in performance κ_E among the ranked items over the edge set E and the algebraic connectivity $\lambda_2(\mathcal{L}_{\mathbf{A}})$ of the comparison graph. It is intuitively clear that the larger the performance gap among the compared items, the more difficult it is to accurately estimate the model parameters. Furthermore, the smaller the algebraic connectivity, the less connected the comparison graph is, due to the presence of bottlenecks[1]. This in turn will increase the chance of obtaining a highly erroneous ranking or of gathering data from which a global ranking cannot be elicited at all. The minimal and maximal degrees $n_{\min}$ and $n_{\max}$ further quantify the impact of the connectivity of the comparison graph.
- We note that the factor $\frac{1}{\lambda_2(\mathcal{L}_{\mathbf{A}})}$ can be equivalently replaced with $\frac{1}{\lambda_2(\mathcal{I})}$ (see Lemma 10 in Appendix A.2). Here, $\mathcal{I} := \nabla^2\ell_0(\boldsymbol{\theta}^*;\mathbf{y})$ is the Fisher information matrix at $\boldsymbol{\theta}^*$ and $\lambda_2(\mathcal{I})$ its smallest non-zero eigenvalue. The fact that the bound depends on the Fisher information is not too surprising. This is so, since this quantity in exponential families quantifies the curvature of the likelihood and the intrinsic difficulty of estimating $\boldsymbol{\theta}^*$.
- Our bounds depend on both κ and κ_E, which is non-standard in the literature. By definition, $\kappa_E \leq \kappa$ and in many cases, κ_E can be much smaller than κ. We discuss this further in Section 5.
- The term $r := \kappa_E + \log\kappa$ shows the impact of large κ and κ_E. When $\kappa \lesssim n$ and $\kappa_E \lesssim \log n$, r is negligible. We will consider this parameter range throughout the paper unless stated otherwise.
- The term $\sqrt{\frac{n}{L}}$ describes explicitly the impact of a high-dimensional parameter space on the estimation problem in relation to L, the number of samples for each comparison, which can be thought of as a measure of the sample size required for each of the n parameters. The inverse root dependence on L is to be expected and, we conjecture, not improvable.

Remark 1. In the case of dense graphs, e.g., complete graphs, $\lambda_2(\mathcal{L}_{\mathbf{A}})$ is large enough so that even $L = 1$ will ensure a consistent estimator as $n \to \infty$. But for sparse graphs, L needs to be larger to compensate for weaker connectivity. The assumption that $L \leq n^8 e^{5\kappa_E}\max\{1,\kappa\}$ is a technical condition. There is nothing special in the exponent for n. Any fixed number larger than 8 can be used which will only affect the constants in the bounds. The condition $L \leq n^8 e^{5\kappa_E}\max\{1,\kappa\}$ may seem counter-intuitive, since it places an upper bound on the sample size. But a control over L is needed because as L gets larger, the optimal choice of the regularization parameter $\rho = c_\rho\frac{1}{\kappa}\sqrt{\frac{n_{\max}}{L}}$ gets smaller and, accordingly, the convergence rate of the gradient descent procedure upon which our proof is based degrades. The optimal choice $\rho = c_\rho\frac{1}{\kappa}\sqrt{\frac{n_{\max}}{L}}$ depends on κ, which is unknown before an estimator is produced, however, one can set $\rho = c_\rho\sqrt{\frac{n_{\max}}{L}}$ and the upper bound will only change by a factor $\max\{1,\kappa\}$ in the first term of Equation (6).

2.1 COMPARISON TO OTHER WORK

To the best of our knowledge, Yan et al. (2012); Hajek et al. (2014); Shah et al. (2016); Negahban et al. (2017); Agarwal et al. (2018); Hendrickx et al. (2019, 2020) are the only papers that study estimation error for the BTL model on a comparison graph with general topology. Since Negahban et al. (2017); Agarwal et al. (2018); Hendrickx et al. (2019, 2020) estimate the the preference scores $\mathbf{w}^*$ rather than $\boldsymbol{\theta}^*$, we cannot directly compare our results with theirs because there is no tight two-sided relationship between their metrics of error and ours. Therefore, here we only compare our results to those in Yan et al. (2012); Hajek et al. (2014); Shah et al. (2016). We include detailed comparison between our bound and the other four papers in Appendix A.1.

ℓ_∞ **loss:** Yan et al. (2012) establish an ℓ_∞-bound of $\frac{e^\kappa}{\min_{i,j} n_{ij}}\sqrt{n_{\max}\log n/L}$ where n_{ij} is the number of common neighbors of item i and item j in the comparison graph, under a strong assumption that $n_{ij} \geq cn$ for some constant $c \in (0,1)$. This constraint on graph topology is stronger than ours since it always implies that the comparison graph

[1]Here, bottlenecks can be formally described as small connected subgraphs with very few edges separating dense portions of the graph.

is dense. In particular, when the comparison graph comes from an Erdös-Rényi model $ER(n,p)$, the conditions in Yan et al. (2012) implies that p is bounded away from 0. Then, $\min_{i,j} n_{i,j} \asymp np^2$, so the bound in Yan et al. (2012) becomes $\frac{e^{\kappa}}{p}\sqrt{\frac{\log n}{npL}}$, while our bound is $\frac{e^{2\kappa_E}}{\sqrt{p}}\sqrt{\frac{\log n}{npL}}$. Our bound is tighter for moderate or small κ_E, and importantly, allows p to vanish. Furthermore, in Section 5, we show by some specific examples that $\min_{i,j} n_{ij}$ could be 0 even for many fairly dense graphs, to illustrate that the Yan et al. (2012) upper bound cannot apply to many realistic settings.

ℓ_2 loss: Hajek et al. (2014); Shah et al. (2016) consider constrained MLE $\hat{\boldsymbol{\theta}} := \min_{\|\boldsymbol{\theta}\|_\infty \le B} \ell_0(\boldsymbol{\theta})$ for a known parameter B such that $\|\boldsymbol{\theta}^*\|_\infty \le B$. In our setting, the upper bounds for ℓ_2-error in Shah et al. (2016) and Hajek et al. (2014) are $e^{4B}\sqrt{n\log n/[\lambda_2(\mathcal{L}_A)L]}$ and $e^{4B}/\lambda_2(\mathcal{L}_A)\sqrt{|E|\log n/L}$. Setting aside the fact that their results require stricter conditions than ours, our ℓ_2 bound of order $e^{\kappa_E}/\lambda_2(\mathcal{L}_{\mathbf{A}})\sqrt{n_{\max} n/L}$ is tighter than theirs for general parameter settings with moderate B, κ and for a broad range of graphs with moderate $\lambda_2(\mathcal{L}_{\mathbf{A}})$, i.e., not too sparse or irregular.

We re-emphasize that Hendrickx et al. (2020) also provide upper bounds for a general fixed comparison graph that matches an instance-wise lower bound, for their parameter of interest $\mathbf{w}^* := \exp(\boldsymbol{\theta}^*)$, instead of $\boldsymbol{\theta}^*$. However, their error metric, i.e., $\sin(\hat{\mathbf{w}}, \mathbf{w}^*)$, is quite different from other similar papers in the BTL literature, including our work. As such, it is not clear how to compare to their results. Furthermore, as noted in Section 1, from the perspective of ranking, an entry-wise metric like $\|\cdot\|_\infty$ is more informative than vector-level metrics like $\|\cdot\|_2$ and $\sin(\cdot,\cdot)$.

2.2 SPECIAL CASES OF GRAPH TOPOLOGIES

We can check some common types of comparison graph topologies and see in what order the necessary sample complexity $N_{\text{comp}} = |E|L$ needs to be to achieve consistency, i.e., $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty = o(1)$. The results are summarized in Table 1. For path and star graphs, we used the specialized bounds in Propositions 3 and 4. As shown in Table 1, our bound now applies to a much broader class of graph topologies under the ℓ_∞-norm compared to Yan et al. (2012).

Graph	$\mathbf{N}_{\text{comp}}$ (Yan et al., 2012)	$\mathbf{N}_{\text{comp}}$ (our work)
Complete	$\Omega(n^2)$	$\Omega(n^2)$
Bipartite	N/A	$\Omega(n^2)$
Path	N/A	$\omega(e^{2\kappa_E} n^2 \log n)$
Star	N/A	$\omega(e^{2\kappa_E} n \log n)$
Barbell	N/A	$\omega(e^{2\kappa_E} n^5 \log n)$

Table 1: Magnitude of N_{comp} to ensure $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty = o(1)$.

Remark 2. For the path graph, star graph, and barbell graph, the necessary sample complexity induced by directly applying our ℓ_∞ bound is larger than the sample complexity induced by the ℓ_2 bound in Shah et al. (2016), though they require more stringent conditions than ours. Thus we provide specialized sharp upper bounds in the case of path and star graph in Proposition 3 and 4. Additionally, in Section 4, we illustrate that by applying a *unique* sub-additivity property of ℓ_∞-loss, we can achieve a much smaller sample complexity in graphs with bottlenecks like the barbell graph.

Erdös-Rényi graph: By applying a union bound on $\lambda_2(\mathcal{L}_{\mathbf{A}})$, $n_{\max}$, and $n_{\min}$ to the sample-wise bounds in Theorem 1, we obtain a corollary in the setting where the comparison graph follows the Erdös-Rényi model $ER(n,p)$.

Corollary 2 (Erdös-Rényi graph)**.** *As a corollary to Theorem 1, suppose that the comparison graph comes from an Erdös-Rényi graph $ER(n,p)$, then under the same conditions, with probability at least $1 - O(n^{-4})$, it holds that*

$$\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_\infty \lesssim e^{2\kappa_E}\sqrt{\frac{\log n}{np^2 L}}, \|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_2 \lesssim e^{\kappa_E}\sqrt{\frac{1}{pL}}.$$

The full form of Corollary 2 with a proof can be found at the end of Appendix A.2. For the Erdös-Rényi comparison graph $ER(n,p)$, the tightest ℓ_∞-norm error bound $e^{2\kappa}\sqrt{\frac{\log n}{npL}}$ is proved in Chen et al. (2019) and Chen et al. (2020). Han et al. (2020) establish an ℓ_∞-norm upper bound of $e^{2\kappa}\sqrt{\frac{\log n}{np}} \cdot \frac{\log n}{\log(np)}$. Negahban et al. (2017) obtain an ℓ_2-norm upper bound of $e^{4\kappa}\frac{\log n}{pL}$ and a lower bound of $e^{-\kappa}\frac{1}{pL}$. Thus the derived ℓ_2-bound in Corollary 2 in Erdös-Rényi case is minimax optimal.

In this case our derived ℓ_∞-bound cannot achieve the rate established in Chen et al. (2019), Chen et al. (2020), though our ℓ_2-bound exhibits the optimal rate proved in Negahban et al. (2017). The reason why our bound does not imply the optimal ℓ_∞-rate under a Erdös-Rényi comparison graph is that our bound is a sample-wise bound and thus cannot leverage some regular property of Erdös-Rényi graph beyond algebraic connectivity and degree homogeneity that is exhibited with high probability.

Tree graphs: For extremely sparse graphs like tree graphs, the general upper bound in Theorem 1 is loose compared to the lower bound in Theorem 5. Therefore, we separately prove some sharp upper bounds for path and star graphs as a complement to our general theory, in these frequently studied cases. For example, single-elimination sports tournaments are commonly designed as a binary tree graph. By the spectral property of path and star graphs (see Appendix A.6), one can verify that the upper bounds in both norms match the ℓ_∞ lower bound in Theorem 5 and the ℓ_2 lower bound in Shah et al. (2016), up to $\sqrt{\log n}$ and $e^{2\kappa_E}$ factors.

Proposition 3 (Path graph)**.** *Suppose the comparison graph is a path graph* $([n], E)$ *with* $E = \{(i, i+1)\}_{i\in[n-1]}$ *and* $L > ce^{2\kappa_E} n \log n$ *for some universal constant c, then with probability at least* $1 - n^{-4}$*, the vanilla MLE* $\hat{\boldsymbol{\theta}}_0$ *satisfies*

$$\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_\infty \lesssim e^{\kappa_E}\sqrt{\frac{n\log n}{L}}, \|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \lesssim e^{\kappa_E} n\sqrt{\frac{\log n}{L}} \tag{7}$$

Proposition 4 (General tree graph)**.** *Suppose the graph is a tree graph* $([n], E)$ *where each item* i *and* j *are compared* L *times such that* $L > ce^{2\kappa_E} n \log n$ *for some universal constant c. Then with probability at least* $1-n^{-4}$*, the vanilla MLE* $\hat{\boldsymbol{\theta}}_0$ *satisfies*

$$\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_\infty \lesssim e^{\kappa_E}\sqrt{\frac{d\log n}{T}}, \|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \lesssim e^{\kappa_E}\sqrt{\frac{dn\log n}{T}}, \tag{8}$$

where $d := \max_{i,j} |\mathrm{path}(i,j)|$ *is the diameter. In particular, for star graph, the upper bound is given by* $d = 1$.

The full form of Proposition 3 and Proposition 4 with proofs are found in Appendix A.2. Briefly, the proofs leverage the closed-form solution of vanilla MLE under the tree graph.

3 LOWER BOUNDS

In this section, we derive a minimax lower bound for the ℓ_∞ loss. Towards that end, we first introduce some new notation. Let N_{comp} be the total number of comparisons that have been observed, so in our setting, $N_{\text{comp}} = |E|L$ where $|E|$ is number of edges in the comparison graph $\mathcal{G}$. Denote the two items involved in the i-th comparison as (i_1, i_2) such that $i_1 < i_2$. Let $\tilde{\mathcal{L}}_A = \frac{1}{N_{\text{comp}}}\sum_{i=1}^{N_{\text{comp}}}(\mathbf{e}_{i_1} - \mathbf{e}_{i_2})(\mathbf{e}_{i_1} - \mathbf{e}_{i_2})^\top$ be the normalized graph Laplacian with pseudo inverse $\tilde{\mathcal{L}}_A^\dagger$ and eigenvalues $0 = \lambda_1(\tilde{\mathcal{L}}_A) \le \lambda_2(\tilde{\mathcal{L}}_A) \le \cdots \le \lambda_n(\tilde{\mathcal{L}}_A)$. With the main notation in place, our minimax lower bound is summarized in the following result.

Theorem 5. *Assume that the comparison graph* $\mathcal{G}$ *is connected and the sample size* $N_{\text{comp}} \ge \frac{c_2 \operatorname{tr}(\tilde{\mathcal{L}}_A^\dagger)}{e^{2\kappa}\kappa^2}$*, any estimator* $\widetilde{\boldsymbol{\theta}}$ *based on* N_{comp} *comparisons with outcomes from the BTL model satisfies*

$$\sup_{\boldsymbol{\theta}^*\in\Theta_\kappa} \mathbb{E}\left[\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty^2\right] \gtrsim \frac{e^{-2\kappa}}{nN_{\text{comp}}} \times \max\left\{n^2, \max_{n'\in\{2,\dots,n\}} \sum_{i=\lceil 0.99n'\rceil}^{n'} [\lambda_i(\tilde{\mathcal{L}}_A)]^{-1}\right\}$$

where $\Theta_\kappa = \{\theta \in \mathbb{R}^n : \mathbf{1}_n^\top \boldsymbol{\theta} = 0,\ \|\boldsymbol{\theta}\|_\infty \le \kappa\}$.

The proof of Theorem 5 largely leverages the lower bound construction from Theorem 2 in Shah et al. (2016). The main modification in adapting it to our setting is to construct an ℓ_∞-packing set. This is done by utilizing the *tight* topological equivalence of ℓ_∞ and ℓ_2 norms in finite dimensions.

We can compare this lower bound with the upper bound in Theorem 1. In our setting, the comparisons distribute evenly over all pairs, so $N_{\text{comp}} = |E|L$, and $\lambda_i(\tilde{\mathcal{L}}_A) = \frac{1}{|E|}\lambda_i(\mathcal{L}_A)$. Thus, given a comparison graph with $\lambda_2(\tilde{\mathcal{L}}_A) \asymp \frac{1}{n}$, the lower bound becomes

$$\sup_{\boldsymbol{\theta}^*\in\Theta_\kappa} \mathbb{E}\left[\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty\right] \gtrsim e^{-\kappa}\sqrt{\frac{n}{N_{\text{comp}}}} \tag{9}$$

In $ER(n,p)$ case, this lower bound becomes $e^{-\kappa}\sqrt{\frac{1}{npL}}$ which matches the upper bound in Chen et al. (2019). For some "regular" graph topology with $\lambda_2(\tilde{\mathcal{L}}_A) \asymp \frac{1}{n}$ like complete graph, expander graph with $\phi = \Omega(n)$ and complete bipartite graph with two partition sets of size $\Omega(n)$, the upper bound becomes

$$\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_\infty \lesssim e^{2\kappa}\sqrt{\frac{n\log n}{N_{\text{comp}}}}.$$

Therefore, when the comparison graph topology is sufficiently regular, our upper bound matches the lower bound up to a $\log n$ factor and a factor of $e^{3\kappa}$.

As a final remark, Negahban et al. (2017) show that the minimax lower bound for ℓ_2-loss and Erdös-Rényi comparison graph $ER(n,p)$ is $e^{-\kappa}\frac{1}{pL}$, which matches our ℓ_2 upper bound up to a factor of $e^{2\kappa}$.

4 IMPLICATIONS FOR TOURNAMENT DESIGN

In this section we discuss how our results can be leveraged to construct more efficient tournament design from a ranking perspective in sports leagues.

As discussed in Section 2.2, for some comparison graphs with small $\lambda_2(\mathcal{L}_{\mathbf{A}})$, the requirement on L and N_{comp} for consistency is stringent. However, as we show next, we can significantly relax the requirement on the sample complexity N_{comp} by an adaptive design that sets different L over edges and uses a *divide and conquer* strategy to estimate $\boldsymbol{\theta}^*$ more efficiently in practice.

Lemma 6 (Subadditivity of ℓ_∞-loss in BTL)**.** *Let* I_1, I_2, I_3 *be three subsets of* $[n]$ *such that* $\cup_{j=1}^3 I_j = [n]$ *and, for each* $j \ne k$, $I_j \not\subseteq I_k$ *and for* $i = 1, 2$, $I_i \cap I_3 \ne \emptyset$*. For any vector* $\boldsymbol{\theta} \in \mathbb{R}^n$*, let* $\boldsymbol{\theta}_{(j)} \in \mathbb{R}^{|I_j|}$ *be the sub-vector of* $\boldsymbol{\theta}$ *consisting of the entries indexed by* I_j, $j = 1, 2, 3$*. Let* $\boldsymbol{\theta}^*$ *be the vector of preference scores in the BTL model over* n *items and* $\hat{\boldsymbol{\theta}}_{(j)}$ *be the MLE of* $\boldsymbol{\theta}^*_{(j)}$ *for the BTL model involving only items in* I_j, $j = 1, 2, 3$*, assuming that the sub-graphs induced by the* I_j*'s are connected and allowing for a different number*

of comparisons in each sub-graph. Finally, let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^n$ *be the ensemble MLE based on* $\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}, \hat{\boldsymbol{\theta}}_{(3)}$*. Then*

$$d_\infty(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \leq 4 \sum_{i=1}^{3} d_\infty(\hat{\boldsymbol{\theta}}_{(i)}, \boldsymbol{\theta}^*_{(i)}), \tag{10}$$

where $d_\infty(\mathbf{v}_1, \mathbf{v}_2) := \|(\mathbf{v}_1 - \mathrm{avg}(\mathbf{v}_1)\mathbf{1}) - (\mathbf{v}_2 - \mathrm{avg}(\mathbf{v}_2)\mathbf{1})\|_\infty$*, where* $\mathrm{avg}(\mathbf{x}) := \frac{1}{n}\boldsymbol{1}_n^\top \mathbf{x}$ *for* $\mathbf{x} \in \mathbb{R}^n$*.*

The proof of Lemma 6 is found in Appendix A.5. As discussed in Section 2.2, for a barbell graph containing two size-$n/2$ complete sub-graphs connected by a single edge, we need $N_{\text{comp}} = \Omega(n^5 \log n)$ for an $o(1)$ error bound. From a practical perspective, we note that such a divide and conquer strategy gives flexibility in the number of comparisons in each sub-graph. For example, if we set $L = 1$ for the two complete sub-graphs to get MLEs $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_2$, and set $L = n$ for the two items linking the two sub-graphs to get an MLE $\hat{\boldsymbol{\theta}}_3$, and combine them by shifting $\hat{\boldsymbol{\theta}}_2$ by the difference of two entries of $\hat{\boldsymbol{\theta}}_3$, then a total sample complexity *reduction* to $N_{\text{comp}} = \Omega(n^2)$ will ensure ℓ_∞-norm error of order $O(e^{2\kappa_E}\sqrt{\log n}/\sqrt{n}) = o(1)$. Because for a complete graph of size m, the ℓ_∞-norm error is $O(e^{2\kappa_E}\sqrt{\log m}/\sqrt{mL})$.

Note that such flexible tournament design is similar to the idea of *active ranking* (Heckel et al., 2019; Ren et al., 2019), but there is still a substantial difference between our setting and active ranking. Active ranking assumes that one can design the tournament in an *online* manner, so that the next pair of items to be compared is determined by the newest outcomes of comparisons. However, in practice many tournaments can only be designed *offline*, i.e., before any games are played. Under this common setting, our ℓ_∞-subadditivity property provides a useful offline approach to efficient tournament design.

5 EXAMPLES AND SIMULATIONS

In this section, we show with numerical experiments on simulated data that our upper bound is tighter than existing results for a broad range of graph topologies, except for some very sparse and very dense cases. This is in-line with our discussions in Sections 2.1 and 2.2. We explicitly focus our comparisons with the ℓ_∞-bound in Yan et al. (2012) here, since their work is closest in spirit to ours. We have also compared our ℓ_∞-upper bound with the ℓ_2-upper bound in Shah et al. (2016) in Appendix A.4 in the appendix. All of our reproducible code is openly accessible[2].

Dependence on κ/κ_E

In all existing work, the dependence of the upper bound on κ or κ_E is exponential (i.e. as in e^κ or e^{κ_E}), which could be abnormally large even for moderate values of κ. Recall that κ and κ_E are the maximal performance gaps over the whole vertex set V, and the edge set E, respectively. When the comparison scheme is carefully designed, we can make $\kappa_E \ll \kappa$ so that our bound is much tighter than others using κ. In BTL model, the maximal (pairwise) winning probability is $p_{\max}(\kappa) = 1/(1 + e^{-\kappa})$. To get a sense, $p_{\max}(2.20) = 0.900$, $p_{\max}(4.59) = 0.990$. A winning probability larger than 0.99 is fairly rare in practice, so it would not be too constraining to set $\kappa = 2.2$ in our simulation. But analytically our result allows κ and κ_E to diverge with n.

In our experiments, we set entries of $\boldsymbol{\theta}^*$ to be $\theta_i^* = \theta_1^* + (i-1)\delta$ with $\delta = \kappa/(n-1)$. Under this setting, for some special graphs, e.g., the island graph in Example 7, κ_E can be much smaller than κ, showing an advantage of our upper bound in representing the dependency on the maximal performance gap κ_E along the edge set, rather than κ the whole vertex set. However, in some cases where the majority of edges has small performance gap and only a few edges have huge gaps, the control purely by κ_E can again be loose. An interesting future direction is to make upper bounds tighter in such cases by including more structural parameters, like the proportion of small-gap edges. We include some illustrative examples in Appendix A.4.

Example 7 (Graph with $\min_{i,j} n_{ij} = 0$)**.** In this case, we intend to illustrate that $\min_{i,j} n_{ij}$ could be 0 or quite close to 0 for even fairly dense graphs, making the upper bound in Yan et al. (2012) less effective. Consider a *3-island* comparison graph $\mathcal{G}$ with n nodes. The induced sub-graphs on node sets V_1, V_2, V_3 with $|V_i| = n_i$ are complete graphs, where $V_1 \cap V_3 = \emptyset$, $V_1 \cup V_2 \cup V_3 = [n]$, and $V_i \cap V_2 \neq \emptyset$ for $i = 1, 3$. There is no edge except for those within V_1, V_2, V_3. This graph $\mathcal{G}$ is connected, and can be fairly dense if we make n_2 large, but $\min_{i,j} n_{i,j} = 0$ always holds since $V_1 \cap V_3 = \emptyset$ and the two induced sub-graphs are complete. See Figure 1 left panel for a visualization of the adjacency matrix of such a graph.

We can also consider more general *island* graphs. A general island graph is determined by n, the size of the graph, n_{island}, the size of island sub-graphs, and n_{overlap}, the number of overlapped nodes between islands. Each island sub-graph is a complete graph, and there are no edges outside islands. For island graphs, it holds that $\min_{i,j} n_{ij} = 0$ and $\kappa_E \approx \kappa \cdot n_{\text{island}}/n$. Figure 1 middle panel shows the adjacency matrix of an island graph with $n_{\text{island}} = 30$, $n_{\text{overlap}} = 5$, $n = 120$. Figure 1 right panel shows the comparion of real error of MLE and our upper bound while varying n_{overlap}. Every point on the lines is the average of 20 trials. It can be seen that representing the upper bound in κ_E rather than κ can make it much tighter when treating κ and κ_E as divergent parameters instead of constants.

In Example 7, we show a common family of graphs which is

[2]Code repository (anonymized): https://anonymous.4open.science/r/experiments-BT-UAI2022

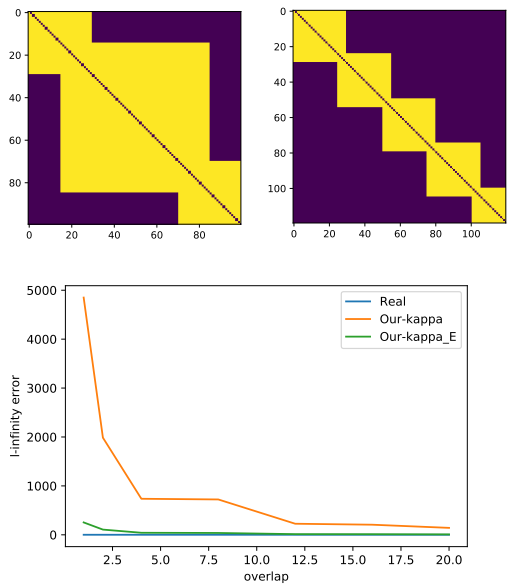

Figure 1: Left: Adjacency matrix of a 3-island graph, with yellow indicating 1 and purple indicating 0; $\lambda_2(\mathcal{L}_{\mathbf{A}}) = 11.92$. Right: Adjacency matrix of a general island graph, with $n_{\text{island}} = 30$, $n_{\text{overlap}} = 5$, $n = 120$; $\lambda_2(\mathcal{L}_{\mathbf{A}}) = 1.19$. Bottom: comparison of real error of MLE and our upper bound using κ and κ_E while varying n_{overlap}. The upper bound of Yan et al. (2012) is not shown because all graphs satisfy $\min_{i,j} n_{ij} = 0$ and the upper bound is always ∞. Note: The error bars are small since the graphs are deterministic and the regularized MLE is very stable.

fairly dense while $\min_{i,j} n_{ij} = 0$, so that the upper bound in Yan et al. (2012) does not hold. Next in Example 8 we consider another family of graphs where their upper bound holds but still looser than our bound.

Example 8 (Barbell graph with random *bridge* edges)**.** Consider a generalized Barbell graph $\mathcal{G}$ containing $n = n_1 + n_2$ nodes, where the induced sub-graph on nodes $\{1, \cdots, n_1\}$ and $\{n_1 + 1, \cdots, n\}$ are complete graphs, and the two sub-graphs are connected by some bridge edges (i, j) for some $1 \le i \le n_1$ and $n_1 + 1 \le j \le n$. Denote the set of bridge edges as E_l, then $|E_l|/(n_1 n_2)$ quantifies the connectivity of $\mathcal{G}$: the larger $|E_l|/(n_1 n_2)$ is, the denser or more regular $\mathcal{G}$ is.

In Figure 2 we show a comparison of real ℓ_∞-loss $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty$, and the upper bounds of ℓ_∞-error in Yan et al. (2012) and our paper. We include this to numerically demonstrate that our bound is in general tighter than Yan et al. (2012), since there is no *known* analytical relationship between $\min_{i,j} n_{ij}$ and $\lambda_2(\mathcal{L}_A)$. In our experiment, we set $n_1 = n_2 = 50$, and randomly link $|E_l|$ edges between the two complete sub-graphs to make $|E_l|/(n_1 n_2)$ vary from 0.004 to 0.96. Every point on the lines is the average of 20 trials. Some points near the left end of the orange line are cut because $\min_{i,j} n_{ij} = 0$ and the corresponding upper bound

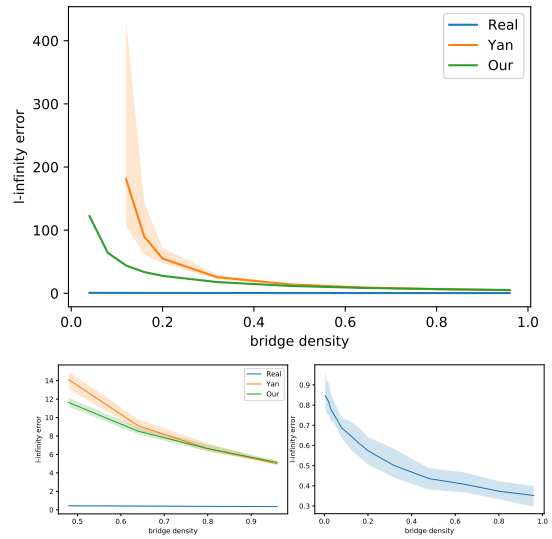

Figure 2: Top: Comparison under the Barbell graph with random bridge edges. Some points near the left end of the orange line are cut because $\min_{i,j} n_{ij} = 0$ and the corresponding upper bound is ∞. Left: zoom-in of for the left panel. Right: the real ℓ_∞-error of MLE. Colored area indicates the upper and lower 0.05 quantiles.

is ∞. It can be seen that our upper bound is strictly tighter than Yan et al. (2012) for a wide range $|E_l|/(n_1 n_2) < 0.8$, and as expected looser in the extremely dense graph topology.

6 DISCUSSION

In this work we provide a sharp risk analysis of the MLE for the BTL global ranking model, under a more general graph topology, in the ℓ_∞-loss. This addresses a major gap in the BTL literature, in extending the comparison graph to more general and thus more practical settings, compared to dense graph setting in Yan et al. (2012). Specifically we derive a novel upper bound for the ℓ_∞ and ℓ_2-loss of the BTL-MLE, showing explicit dependence on the algebraic connectivity of the graph, the sample complexity, and the maximal performance gap between compared items. We also support our analysis by deriving lower bounds for the ℓ_∞-loss and analyze specific topologies satisfying certain regularity conditions under which the MLE is nearly minimax optimal. Additionally we show that the ℓ_∞-loss satisfies a unique subadditivity property for the BTL MLE and utilize our derived bounds for efficient tournament design. We note that our upper bound is suboptimal in the cases where the graph topology is extremely sparse or irregular. Although we provide sharp upper bounds for path and star graphs as separate propositions, we still miss optimality other graph topologies. A good future direction would be to optimize the upper and lower bounds in such comparison graph regimes.

Bibliography

Agarwal, A., Patil, P., and Agarwal, S. (2018). Accelerated Spectral Ranking. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 70–79.

Alon, N., Schwartz, O., and Shapira, A. (2008). An Elementary Construction of Constant-Degree Expanders. *Comb. Probab. Comput.*, 17(3):319–327.

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

Brouwer, A. E. and Haemers, W. H. (2012). *Spectra of graphs.* Springer.

Bubeck, S. (2015). Convex Optimization: Algorithms and Complexity. *Found. Trends Mach. Learn.*, 8(3–4):231–357.

Cattelan, M., Varin, C., and Firth, D. (2013). Dynamic Bradley-Terry modelling of sports tournaments. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 62(1):135–150.

Chen, P., Gao, C., and Zhang, A. Y. (2020). Partial Recovery for Top-k Ranking: Optimality of MLE and Sub-Optimality of Spectral Method. *arXiv:2006.16485*.

Chen, Y., Fan, J., Ma, C., and Wang, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.*, 47(4):2204–2235.

Chen, Y. and Suh, C. (2015). Spectral MLE: Top-K Rank Aggregation from Pairwise Comparisons. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 371–380.

Das, K. (2004). The Laplacian spectrum of a graph. *Computers & Mathematics with Applications*, 48(5):715–724.

Fahrmeir, L. and Tutz, G. (1994). Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison Systems. *Journal of the American Statistical Association*, 89(428):1438–1449.

Ford, L. R. (1957). Solution of a Ranking Problem from Binary Comparisons. *The American Mathematical Monthly*, 64(8):28–33.

Hajek, B. E., Oh, S., and Xu, J. (2014). Minimax-optimal Inference from Partial Rankings. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1475–1483.

Han, R., Ye, R., Tan, C., and Chen, K. (2020). Asymptotic theory of sparse Bradley–Terry model. *Ann. Appl. Probab.*, 30(5):2491–2515.

Heckel, R., Shah, N. B., Ramchandran, K., and Wainwright, M. J. (2019). Active ranking from pairwise comparisons and when parametric assumptions do not help. *Ann. Statist.*, 47(6):3099–3126.

Hendrickx, J. M., Olshevsky, A., and Saligrama, V. (2019). Graph Resistance and Learning from Pairwise Comparisons. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2702–2711. PMLR.

Hendrickx, J. M., Olshevsky, A., and Saligrama, V. (2020). Minimax Rate for Learning From Pairwise Comparisons in the BTL Model. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4193–4202. PMLR.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, 32(1):384–406.

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA.

Masarotto, G. and Varin, C. (2012). The ranking lasso and its application to sport tournaments. *Ann. Appl. Stat.*, 6(4):1949–1970.

Negahban, S., Oh, S., and Shah, D. (2017). Rank Centrality: Ranking from Pairwise Comparisons. *Oper. Res.*, 65(1):266–287.

Ren, W., Liu, J., and Shroff, N. B. (2019). *On Sample Complexity Upper and Lower Bounds for Exact Ranking from Noisy Comparisons*. Curran Associates Inc., Red Hook, NY, USA.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., and Wainwright, M. J. (2016). Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence. *Journal of Machine Learning Research*, 17(58):1–47.

Simons, G. and Yao, Y.-C. (1999). Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *Ann. Statist.*, 27(3):1041–1060.

Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statist. Sci.*, 9:94–108.

Varin, C., Cattelan, M., and Firth, D. (2016). Statistical modelling of citation exchange between statistics journals. *J. Roy. Statist. Soc. Ser. A*, 179(1):1–63.

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices.

Wendland, H. (2018). *Numerical linear algebra*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge. An introduction.

Yan, T., Yang, Y., and Xu, J. (2012). Sparse Paired Comparisons in the Bradley-Terry Model. *Statistica Sinica*, 22(3):1305–1318.

Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.*, 29(1):436–460.

A APPENDIX

A.1 COMPARISON OF RESULTS

This section is a complement to Section 2.1. We will summarize all existing works on the estimation error of Bradley-Terry model in two tables, and then compare our results with the results in Negahban et al. (2017); Agarwal et al. (2018); Hendrickx et al. (2020) in detail.

For simplicity, in Table 2 and Table 3 we use κ to replace B for results in Hajek et al. (2014); Shah et al. (2016) as $\kappa \asymp B$ when $\mathbf{1}^\top \boldsymbol{\theta}^* = 0$, and in Table 3 we omit the lower bound as they are usually in fairly complex forms.

norm	paper	bound
$\|\cdot\|_\infty$	Simons and Yao (1999)	$p=1, \quad \lesssim e^{\kappa}\sqrt{\frac{\log n}{nL}}$
	Yan et al. (2012)	$\lesssim e^{2\kappa}\frac{1}{p}\sqrt{\frac{\log n}{npL}}$
	Han et al. (2020)	$\lesssim e^{2\kappa}\sqrt{\frac{\log n}{np}}\cdot\frac{\log n}{\log(np)}$
	Chen et al. (2019), Chen et al. (2020)	$\lesssim e^{2\kappa}\sqrt{\frac{\log n}{npL}}$
	Our work	$\lesssim e^{2\kappa_E}\sqrt{\frac{\log n}{np^2L}}$
$\|\cdot\|_2^2$	Hajek et al. (2014)	$\lesssim e^{8\kappa}\frac{\log n}{pL}$
	Shah et al. (2016)	$\lesssim e^{8\kappa}\frac{\log n}{pL}, \quad \gtrsim e^{-2\kappa}\frac{1}{pL}$
	Negahban et al. (2017)	$\lesssim e^{4\kappa}\frac{\log n}{pL}, \quad \gtrsim e^{-\kappa}\frac{1}{pL}$
	Chen et al. (2019), Chen et al. (2020)	$\lesssim e^{2\kappa}\frac{1}{pL}$
$\sin^2(\cdot,\cdot)\ (\hat{\mathbf{w}})$	Hendrickx et al. (2020)	$\lesssim e^{2\kappa}\frac{1}{pL\|w\|_2^2}$
$\|\cdot\|_1\ (\hat{\mathbf{w}})$	Agarwal et al. (2018)	$\lesssim e^{\kappa}\sqrt{\frac{\log n}{L}}$

Table 2: Comparison of results under $ER(n,p)$ in literature.

norm	paper	bound
$\|\cdot\|_\infty$	Yan et al. (2012)	$\frac{e^{\kappa}}{\min_{i,j} n_{ij}}\sqrt{\frac{n_{\max}\log n}{L}}$
	Our work	$\frac{e^{2\kappa_E}}{\lambda_2(\mathcal{L}_{\mathbf{A}})}\frac{n_{\max}}{n_{\min}}\sqrt{\frac{n}{L}}+\frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_{\mathbf{A}})}\sqrt{\frac{n_{\max}\log n}{L}}$
$\|\cdot\|_2^2$	Hajek et al. (2014)	$\lesssim e^{8\kappa}\frac{\lvert E\rvert\log n}{\lambda_2(\mathcal{L}_A)^2 L}$
	Shah et al. (2016)	$\lesssim e^{8\kappa}\frac{n\log n}{\lambda_2(\mathcal{L}_A)L}$
	Our work	$\frac{e^{2\kappa_E}}{\lambda_2(\mathcal{L}_{\mathbf{A}})^2}\frac{n_{\max}n}{L}$
$\sin^2(\cdot,\cdot)\ (\hat{\mathbf{w}})$	Hendrickx et al. (2020)	$\lesssim e^{2\kappa}\frac{Tr(\mathcal{L}_A^\dagger)}{L\|w\|_2^2}$
$\|\cdot\|_1\ (\hat{\mathbf{w}})$	Agarwal et al. (2018)	$\lesssim \frac{\eta e^{\kappa}n_{avg}}{\lambda_2(D^{-1}A)n_{\min}}\sqrt{\frac{\log n}{L}}$

Table 3: Comparison of results for a fixed general comparison graph in literature.

In Negahban et al. (2017), they establish an ℓ_2 upper bound for $\|\hat{\boldsymbol{\pi}}-\tilde{\boldsymbol{\pi}}\|_2/\|\tilde{\boldsymbol{\pi}}\|_2$ in the order of $\frac{e^{2.5\kappa}}{\lambda_2(\mathcal{L}_{rw})}\sqrt{\frac{n_{\max}\log n}{L}}$ where $\tilde{\pi}(i) := w_i/\sum_j w_j$ with $w_i = \exp(\theta_i)$, $\hat{\boldsymbol{\pi}}$ is the rank centrality estimamtor of $\tilde{\boldsymbol{\pi}}$, λ_2 refers to the second smallest eigenvalue, $\mathcal{L}_{rw} = D^{-1}A$ (which has the same spectrum as $D^{-1/2}\mathcal{L}_A D^{-1/2}$), and $\mathcal{L}_A = D - A$. Recall that our ℓ_2 upper bound is for $\|\hat{\theta}-\theta\|_2$ and the order is $\frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_A)}\sqrt{\frac{n_{\max}n}{L}}$. We can now see that it's hard to give a general comparison between the two results because 1. for a general graph, there is no precise relationship between $\lambda_2(\mathcal{L}_{rw})$ and $\lambda_2(\mathcal{L}_A)$; 2. more importantly, for a general model parameter θ, there is no tight two-sided bound between $\|\hat{\theta}-\theta\|_2$ and $\|\pi-\tilde{\pi}\|_2/\|\tilde{\pi}\|_2$. Although, it would be a very interesting future work to give a tight description of these two relevant pairs of quantities and make a meaningful comparison.

Agarwal et al. (2018) establish an ℓ_1-norm upper bound for the score parameter $\tilde{\pi}_i := w_i/\sum_j w_j$ with $w_i = \exp(\theta_i)$. Their bound is of the order $\frac{\eta e^{\kappa}n_{avg}}{\lambda_2(D^{-1}A)n_{\min}}\sqrt{\frac{\log n}{L}}$, where $n_{avg} = \sum_{i\in[n]}\tilde{\pi}_i n_i$, $D = \mathrm{diag}(n_1,\cdots,n_n)$, and $\eta := \log\left(\frac{n_{avg}}{n_{\min}\pi_{\min}}\right)$

with $\pi_{\min} = \min_{i\in[n]} \tilde{\pi}_i$.

In Hendrickx et al. (2020), they propose a novel weighted least square method to estimate vector w, with $w_i = \exp(\theta_i)$, and provide delicate theoretical analysis of their method. Their estimator shows a sharp upper bound for $\mathbb{E}[\sin^2(\hat{w}, w)]$ and equivalently for $\mathbb{E}\|\hat{w}/\|\hat{w}\|_2 - w/\|w\|_2\|_2^2$, in the sense that the upper bound for $\mathbb{E}[\sin^2(\hat{w}, w)]$ matches a instance-wise lower bound up to constant factors. Such a universal sharp/optimal bound for general comparison graph (although the lower bound is not in the form of minimax rate) is unique in literature. For convenience of comparison, here we assume w_i's are $O(1)$ (otherwise we can put a factor e^{2B} in the bound $\frac{Tr(\mathcal{L}_A^\dagger)}{L}$). Then their upper bound is of the order $\frac{Tr(\mathcal{L}_A^\dagger)}{L\|w\|_2^2}$, where $\mathcal{L}_A^\dagger$ refers to the Moore-Penrose pseudo inverse of the graph Laplacian of the comparison graph. To correct for their different choice of metric, we need to multiply $\|w\|_2^2$ to their bound, and it becomes $\frac{Tr(\mathcal{L}_A^\dagger)}{L}$. On the other hand, the upper bound for expected ℓ_2 loss in Shah et al. (2016) is $\frac{n}{L\lambda_2(\mathcal{L}_A)}$. Since $\lambda_2(\mathcal{L}_A)$ is the smallest positive eigenvalue of $\mathcal{L}_A$, it holds that $Tr(\mathcal{L}_A^\dagger) < n/\lambda_2(\mathcal{L}_A)$, and hence the upper bound for expected error in Hendrickx et al. (2020) is tighter than the one in Shah et al. (2016). Although, it should be noted that the loss function in Hendrickx et al. (2020) is not directly comparable to a plain ℓ_2 loss, and we are just doing an approximate comparison.

In our paper, however, we provide a high probability bound for $\|\hat{\theta} - \theta\|_2^2$ in the order of $\frac{n_{\max} n}{L\lambda_2^2(\mathcal{L}_A)}$. It's usually hard to make a fair comparison between a high probability bound and a bound for expected metrics, but in Hajek et al. (2014); Shah et al. (2016), they also provide a high probability bound in equation (8b) of Theorem 2. As we discussed in Section 2.2 and Section 5, the ℓ_2 bound is not the primary focus of our paper, and although our theoretical analysis is not optimized for ℓ_2 error, our bound is still tighter than the bound in Hajek et al. (2014); Shah et al. (2016) for moderately dense and regular graphs, and is only worse for fairly sparse and irregular graphs.

A.2 PROOF IN SECTION 2

Proof of Theorem 1. We will use a gradient descent sequence defined recursively by $\theta^{(0)} = \theta^*$ and, for $t = 1, 2, \ldots$,

$$\theta^{(t+1)} = \theta^{(t)} - \eta[\nabla \ell_\rho(\theta^{(t)}) + \rho\theta^{(t)}].$$

Our proof builds heavily on the ideas and techniques developed by Chen et al. (2019) and further extended by Chen et al. (2020), and contains two key steps. In the first step, we control $\|\theta^{(T)} - \hat{\theta}_\rho\|$ for T large enough, by leveraging the convergence property of gradient descent for strong convex functions. In the second step, we control $\|\theta^{(T)} - \theta^*\|$ through a leave-one-out argument. The proof can be sketched as follows:

1. Bound $\|\theta^{(T)} - \hat{\theta}_\rho\|_\infty$, for large T using the linear convergence property of gradient descent for strongly-convex and smooth functions.
2. Bound $\|\theta^{(T)} - \theta^*\|_\infty$ for large T using the leave-one-out argument.
3. Finally, $\|\hat{\theta}_\rho - \theta^*\|_\infty$ is controlled by triangle inequality.

Step 1. Bound $\|\theta^{(T)} - \hat{\theta}_\rho\|_\infty$, for large T.

1. **Linear convergence, orthogonality to $\mathbf{1}_n$.** We say that a function ℓ is α-strongly convex if $\nabla^2 \ell(x) \succeq \alpha I_n$ and β-smooth if $\|\nabla \ell(x) - \nabla \ell(y)\|_2 \leq \beta \|x - y\|_2$ for all $x, y \in \mathrm{dom}(\ell)$. By Lemma 10, we know that $\ell_\rho(\cdot)$ is ρ-strongly convex and $(\rho + n_{\max})$-smooth. By Theorem 3.10 in Bubeck (2015), we have

$$\|\theta^{(t)} - \hat{\theta}_\rho\|_2 \leq (1 - \frac{\rho}{\rho + n_{\max}})^t \|\theta^{(0)} - \hat{\theta}_\rho\|_2. \tag{11}$$

   Besides, as we start with θ^* that satisfies $1_n^\top \theta^* = 0$, it holds that $1_n^\top \theta^{(t)} = 0$ for all $t \geq 0$. To see this, just notice that

$$\nabla \ell_\rho(\theta) = \rho\theta + \sum_{(i,j)\in E} [-\bar{y}_{ij} + \psi(\theta_i - \theta_j)](\boldsymbol{e}_i - \boldsymbol{e}_j)$$

   and $1_n^\top(\boldsymbol{e}_i - \boldsymbol{e}_j)$ for any i, j. Then by $1_n^\top \theta^{(t)} = 0, \forall t$ and (11), we have $1_n^\top \hat{\theta}_\rho = 0$.
2. **Control $\|\theta^* - \hat{\theta}_\rho\|_2$.** By a Taylor expansion, we have that

$$\begin{aligned}\ell_\rho(\hat{\theta}_\rho; y) = \ell_\rho(\theta^*; y) &+ (\hat{\theta}_\rho - \theta^*)^\top \nabla \ell_\rho(\theta^*; y) \\ &+ \frac{1}{2}(\hat{\theta}_\rho - \theta^*)^\top \nabla^2 \ell_\rho(\xi; y)(\hat{\theta}_\rho - \theta^*),\end{aligned}$$

   where ξ is a convex combination of θ^* and $\hat{\theta}_\rho$. By Cauchy-Schwartz inequality,

$$|(\hat{\theta}_\rho - \theta^*)^\top \nabla \ell_\rho(\theta^*; y)| \leq \|\nabla \ell_\rho(\theta^*; y)\|_2 \|\theta^* - \hat{\theta}_\rho\|_2.$$

   The two inequalities above and the fact that $\ell_\rho(\theta^*; y) \geq \ell_\rho(\hat{\theta}_\rho; y)$ yield that

$$\|\theta^* - \hat{\theta}_\rho\|_2 \leq \frac{2\|\nabla \ell_\rho(\theta^*; y)\|_2}{\rho_{\min}(\nabla^2 \ell_\rho(\xi; y))}.$$

   By Lemma 9, $\|\nabla \ell_\rho(\theta^*; y)\|_2 \lesssim \sqrt{\frac{n_{\max}(n+r)}{L}}$. This fact, together with $\rho_{\min}(\nabla^2 \ell_\rho(\xi; y)) \geq \rho$ and $\rho \asymp \frac{1}{\kappa}\sqrt{\frac{n_{\max}}{L}}$, gives that $\|\theta^* - \hat{\theta}_\rho\|_2 \leq c\kappa\sqrt{n + r}$, for some $c > 0$.
3. **Bound $\|\theta^{(T)} - \hat{\theta}_\rho\|_2$.** Take $T = \lfloor \kappa^2 e^{3\kappa_E} n^6 \rfloor$ and remember that $L \leq \kappa^2 e^{4\kappa_E} n^8$. The previous two steps imply that

$$\|\theta^{(T)} - \hat{\theta}_\rho\|_2 \leq c(1 - \frac{\rho}{\rho + n_{\max}})^T \kappa\sqrt{n + r} \leq c \exp\left(-\frac{T\rho}{\rho + n_{\max}}\right) \kappa\sqrt{n + r}.$$

   Let $\tilde{f}_d = \frac{e^{2\kappa_E}}{\lambda_2(\mathcal{L}_A)} \frac{n_{\max}}{n_{\min}} \sqrt{\frac{n+r}{L}} + \frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_A)} \sqrt{\frac{n_{\max}(\log n + r)}{L}}$ and consider inequality $e^{-g}\kappa\sqrt{n + r} \leq \tilde{f}_d$. The solution is given by $g \geq \log\kappa + \frac{1}{2}\log(n + r) - \log\tilde{f}_d$ and the inequality holds as long as $g \geq \kappa + 6\log n + 3\log\kappa$ since $L \leq \max\{1, \kappa\} e^{3\kappa_E} n^8$. Take $g = \kappa + 5\log n + \log\kappa$, then as long as

$$T\rho \geq 2n_{\max} n g,$$

it holds that $T\rho > \frac{1}{2}g(\rho + n_{\max})$, then $\|\theta^{(T)} - \hat{\theta}_\rho\|_2 \le c\exp(-g)\kappa\sqrt{n+r}$ is smaller than $\tilde{C}_d\tilde{f}_d$ for some constant $\tilde{C}_d$. Since $T = \lfloor \kappa^2 e^{3\kappa_E} n^6 \rfloor$ and $\rho \ge \frac{c_\rho}{\kappa}\sqrt{\frac{n_{\max}}{\kappa^2 e^{4\kappa_E} n^8}}$, we have

$$T\rho \ge c_\rho \kappa e^{\kappa_E} n^2 \sqrt{n_{\max}} \ge 2n_{\max} n g.$$

In conclusion, we have

$$\|\theta^{(T)} - \hat{\theta}_\rho\|_\infty \le \|\theta^{(T)} - \hat{\theta}_\rho\|_2 \le \tilde{C}_d \tilde{f}_d.$$

The arguments above also hold with $\tilde{f}_a = \frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_A)}\sqrt{\frac{n_{\max}(n+r)}{L}}$, i.e., we have $\|\theta^{(T)} - \hat{\theta}_\rho\|_2 \le \tilde{C}_a \tilde{f}_a$ for some constant $\tilde{C}_a$.

Step 2. Bound $\|\theta^{(T)} - \theta^*\|_\infty$ by a leave-one-out argument.

Denote $\psi(x) = \frac{1}{1+e^{-x}}$ and $r = \log\kappa + \kappa_E$, and define the leave-one-out negative log-likelihood as

$$\begin{aligned}\ell_n^{(m)}(\theta) = &\sum_{1\le i<j\le n: i,j\ne m} A_{ij}\left[\bar{y}_{ij}\log\frac{1}{\psi(\theta_i-\theta_j)} + (1-\bar{y}_{ij})\log\frac{1}{1-\psi(\theta_i-\theta_j)}\right] \\ &+ \sum_{j\in[n]\setminus\{m\}} A_{mj}\left[\psi\left(\theta_m^*-\theta_j^*\right)\log\frac{1}{\psi(\theta_m-\theta_j)} + \psi\left(\theta_j^*-\theta_m^*\right)\log\frac{1}{\psi(\theta_j-\theta_m)}\right],\end{aligned} \tag{12}$$

so the leave-one-out gradient descent sequence is, for $t = 0, 1, \ldots,$

$$\theta^{(t+1,m)} = \theta^{(t,m)} - \eta\left(\nabla\ell_n^{(m)}\left(\theta^{(t,m)}\right) + \rho\theta^{(t,m)}\right).$$

We initialize both sequences by $\theta^{(0)} = \theta^{(0,m)} = \theta^*$ and use step size $\eta = \frac{1}{\rho+n_{\max}}$. By assumption 2, $\lambda_2(\mathcal{L}_\mathbf{A}) > 0$, so we can let $f_a = C_a\frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_A)}\left[\sqrt{\frac{n_{\max}(n+r)}{L}} + \rho\kappa(\theta^*)\sqrt{n}\right]$, $f_b = 10e^{\kappa_E}\frac{\sqrt{n_{\max}}}{n_{\min}}f_a$, $f_c = C_c\frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_A)}\sqrt{\frac{n_{\max}(\log n+r)}{L}}$, $f_d = f_b + f_c$ with sufficiently large constant $C_c > 0$ and $C_a \gg C_c$. By assumption 1, we have $f_c + f_d \le 0.1$. We will show in Lemma 12, 13, 14, 15 that for all $0 \le t \le T = \lfloor \kappa^2 e^{3\kappa_E} n^6 \rfloor$

$$\begin{aligned}\|\theta^{(t)} - \theta^*\|_2 &\le f_a, \\ \max_{m\in[n]}|\theta_m^{(t,m)} - \theta_m^*| &\le f_b, \\ \max_{m\in[n]}\|\theta^{(t,m)} - \theta^{(t)}\|_2 &\le f_c, \\ \|\theta^{(t)} - \theta^*\|_\infty &\le f_d.\end{aligned} \tag{13}$$

When $t = 0$, (13) holds since $\theta^{(0)} = \theta^{(0,m)} = \theta^*$. By Lemma 12, 13, 14, 15, and a union bound, we know that (13) holds for all $0 \le t \le T = \lfloor \kappa^2 e^{3\kappa_E} n^6 \rfloor$ with probability at least $1 - O(n^{-4})$. Therefore, using the result in step 1, we have

$$\|\hat{\theta}_\rho - \theta^*\|_\infty \le \|\hat{\theta}_\rho - \theta^{(T)}\|_\infty + \|\theta^{(T)} - \theta^*\|_\infty \le 2f_d.$$

As a byproduct, we have

$$\|\hat{\theta}_\rho - \theta^*\|_2 \le \|\hat{\theta}_\rho - \theta^{(T)}\|_2 + \|\theta^{(T)} - \theta^*\|_2 \le 2f_a$$

□

Lemma 9. *With probability at least* $1 - O(\kappa^{-2}e^{-3\kappa_E}n^{-10})$ *the gradient of the regularized log-likelihood satisfies*

$$\|\nabla\ell_\rho(\theta^*)\|_2^2 \lesssim \frac{n_{\max}(n+r)}{L} + \rho\kappa(\theta^*)\sqrt{n}.$$

In particular, for $\rho \asymp \frac{1}{\kappa(\theta^*)}\sqrt{\frac{n_{\max}}{L}}$, *we have* $\|\nabla\ell_\rho(\theta^*)\|_2^2 \lesssim \frac{n_{\max}(n+r)}{L}$.

Proof. Triangle inequality gives

$$\|\nabla\ell_\rho(\theta^*)\|_2 \le \|\nabla\ell_0(\theta^*)\|_2 + \rho\|\theta^*\|_2.$$

By definition of $\kappa(\theta^*)$, we have $\|\theta^*\|_2 \le \sqrt{n}\kappa(\theta^*)$. For the first term, by Lemma 16 we have

$$\|\nabla \ell_0(\theta^*)\|_2^2 = \sum_{i=1}^n \left[\sum_{j\in\mathcal{N}(i)} [\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)] \right]^2 \le C_1 \frac{n_{\max}(n+r)}{L}.$$

□

Lemma 10. *Let* $\kappa_E(x) = \max_{(i,j)\in E} |x_i - x_j|$, *then* $\forall \theta \in \mathbb{R}^n$,

$$\begin{aligned} \lambda_{\max}(\nabla^2 \ell_\rho(\theta; y)) &\le \rho + \frac{1}{2} n_{\max}, \\ \lambda_2(\nabla^2 \ell_\rho(\theta; y)) &\ge \rho + \frac{1}{4e^{\kappa_E(\theta)}} \lambda_2(\mathcal{L}_A). \end{aligned} \tag{14}$$

In particular, we have

$$\lambda_2(\nabla^2 \ell_\rho(\theta; y)) \ge \rho + \frac{1}{4e^{\kappa_E(\theta^*)} e^{2\|\theta-\theta^*\|_\infty}} \lambda_2(\mathcal{L}_A).$$

Proof. Use the fact that

$$\nabla^2 \ell_0(\theta; y) = \sum_{(i,j)\in E} \frac{e^{\theta_i} e^{\theta_j}}{(e^{\theta_i} + e^{\theta_j})^2} (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top,$$

and $\forall (i,j) \in E$, $\frac{1}{4\exp(\kappa_E(\theta))} \le \frac{e^{\theta_i} e^{\theta_j}}{(e^{\theta_i}+e^{\theta_j})^2} \le \frac{1}{4}$, $\kappa_E(x_1) \le \kappa_E(x_2) + 2\|x_1 - x_2\|_\infty$. In addition, the largest eigenvalue of graph Laplacian satisfies (Corollary 3.9.2 in Brouwer and Haemers (2012))

$$\lambda_{\max}(\mathcal{L}_A) \le \max_{(i,j)\in E} (n_i + n_j) \le 2n_{\max}. \tag{15}$$

□

Lemma 11. *Provided that* (13) *holds, then*

$$\begin{aligned} \max_{m\in[n]} \|\theta^{(t+1,m)} - \theta^*\|_\infty &\le f_c + f_d, \\ \max_{m\in[n]} \|\theta^{(t+1,m)} - \theta^*\|_2 &\le f_c + f_a. \end{aligned} \tag{16}$$

Proof. By triangle inequality, we have

$$\begin{aligned} \max_{m\in[n]} \|\theta^{(t,m)} - \theta^*\|_\infty &\le \max_{m\in[n]} \|\theta^{(t,m)} - \theta^{(t)}\|_\infty + \|\theta^{(t)} - \theta^*\|_\infty \le f_c + f_d \\ \max_{m\in[n]} \|\theta^{(t,m)} - \theta^*\|_2 &\le \max_{m\in[n]} \|\theta^{(t,m)} - \theta^{(t)}\|_2 + \|\theta^{(t)} - \theta^*\|_2 \le f_c + f_a. \end{aligned}$$

□

Lemma 12. *Suppose* (13) *holds, and the step size satisfies* $0 < \eta \le \frac{1}{\rho + n_{\max}}$. *If*

$$f_d \le 0.1 \text{ and } f_a \ge C_a \frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)} \left[\sqrt{\frac{n_{\max}(n+r)}{L}} + \rho\kappa(\theta^*)\sqrt{n} \right],$$

for some large constant C_a, *then with probability at least* $1 - O(\kappa^{-2} e^{-3\kappa_E} n^{-10})$ *we have*

$$\|\theta^{(t+1)} - \theta^*\|_2 \le f_a.$$

Proof. By the form of the gradient descent, we have that

$$\begin{aligned}\theta^{(t+1)} - \theta^* &= \theta^{(t)} - \eta\nabla\ell_\rho(\theta^{(t)}) - \theta^* \\ &= \theta^{(t)} - \eta\nabla\ell_\rho(\theta^{(t)}) - [\theta^* - \eta\nabla\ell_\rho(\theta^*)] - \eta\nabla\ell_\rho(\theta^*) \\ &= \left[I_n - \eta\int_0^1 \nabla^2\ell_\rho(\theta(\tau))\mathrm{d}\tau\right](\theta^{(t)} - \theta^*) - \eta\nabla\ell_\rho(\theta^*),\end{aligned}$$

where $\theta(\tau) = \theta^* + \tau(\theta^{(t)} - \theta^*)$. Letting $H = \int_0^1 \nabla^2\ell_\rho(\theta(\tau))\mathrm{d}\tau$, by the triangle inequality,

$$\|\theta^{(t+1)} - \theta^*\|_2 \le \|(I_n - \eta H)(\theta^{(t)} - \theta^*)\|_2 + \eta\|\nabla\ell_\rho(\theta^*)\|_2. \tag{17}$$

Setting $\kappa_E(x) = \max_{(i,j)\in E}|x_i - x_j|$, then, for sufficiently small ϵ, we have that

$$\kappa_E(\theta(\tau)) \le \kappa_E(\theta^*) + 2\|\theta^{(t)} - \theta^*\|_\infty \le \kappa_E(\theta^*) + \epsilon. \tag{18}$$

as long as

$$2f_d \le \epsilon. \tag{19}$$

Then, by Lemma 10 and setting $\epsilon = 0.2$, for any $\tau \in [0, 1]$,

$$\rho + \frac{\lambda_2(\mathcal{L}_A)}{10e^{\kappa_E(\theta^*)}} \le \rho + \frac{\lambda_2(\mathcal{L}_A)}{8e^{\kappa_E(\theta^*)}e^\epsilon} \le \lambda_2(\nabla^2\ell_\rho(\theta(\tau))) \le \lambda_{\max}(\nabla^2\ell_\rho(\theta(\tau))) \le \rho + \frac{1}{2}n_{\max}. \tag{20}$$

Since $1_n^\top(\theta^{(t)} - \theta^*) = 0$, we obtain that

$$\|(I_n - \eta H)(\theta^{(t)} - \theta^*)\|_2 \le \max\{|1 - \eta\lambda_2(H)|, |1 - \eta\lambda_{\max}(H)|\}\|\theta^{(t)} - \theta^*\|_2. \tag{21}$$

By (20) and the fact that $\eta \le \frac{1}{\rho + n_{\max}}$, we get

$$\|(I_n - \eta H)(\theta^{(t)} - \theta^*)\|_2 \le (1 - \frac{\eta\lambda_2(\mathcal{L}_A)}{10e^{\kappa_E(\theta^*)}})\|\theta^{(t)} - \theta^*\|_2. \tag{22}$$

By Lemma 9 and the induction hypothesis, we have

$$\|\theta^{(t+1)} - \theta^*\|_2 \le (1 - \frac{\eta\lambda_2(\mathcal{L}_A)}{10e^{\kappa_E(\theta^*)}})f_a + C\eta\left[\sqrt{\frac{n_{\max}(n+r)}{L}} + \rho\kappa(\theta^*)\sqrt{n}\right] \le f_a \tag{23}$$

as long as

$$f_a \ge C_a\frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)}\left[\sqrt{\frac{n_{\max}(n+r)}{L}} + \rho\kappa(\theta^*)\sqrt{n}\right] \tag{24}$$

for some large constant C_a. □

Lemma 13. *Suppose* (13) *holds and assume that*

1. $f_a = C_a\frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)}\left[\sqrt{\frac{n_{\max}(n+r)}{L}} + \rho\kappa(\theta^*)\sqrt{n}\right]$, $f_c = C_c\frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)}\sqrt{\frac{n_{\max}(\log n+r)}{L}}$ *with* $C_a \gg C_c$.
2. $\frac{n_{\min}}{10e^{\kappa_E(\theta^*)}}f_b \ge \frac{3\sqrt{n_{\max}}}{4}f_a$.
3. $f_c + f_d \le 0.1$.

then as long as the step size satisfies $0 < \eta \le \frac{1}{\rho + n_{\max}}$, *with probability at least* $1 - O(\kappa^{-2}e^{-3\kappa_E}n^{-10})$ *we have*

$$\max_{m\in[n]}\left|\theta_m^{(t+1,m)} - \theta_m^*\right| \le f_b.$$

Proof. Recall that the gradient descent step for leave-one-out estimator $\theta^{(m)}$ is defined as

$$\theta^{(t+1,m)} = \theta^{(t,m)} - \eta \left(\nabla \ell_n^{(m)} \left(\theta^{(t,m)} \right) + \rho \theta^{(t,m)} \right),$$

where

$$\begin{aligned}\ell_n^{(m)}(\theta) = &\sum_{1 \le i < j \le n : i,j \ne m} A_{ij} \left[\bar{y}_{ij} \log \frac{1}{\psi\left(\theta_i - \theta_j\right)} + (1 - \bar{y}_{ij}) \log \frac{1}{1 - \psi\left(\theta_i - \theta_j\right)} \right] \\ &+ \sum_{j \in [n] \setminus \{m\}} A_{mj} \left[\psi\left(\theta_m^* - \theta_j^*\right) \log \frac{1}{\psi\left(\theta_m - \theta_j\right)} + \psi\left(\theta_j^* - \theta_m^*\right) \log \frac{1}{\psi\left(\theta_j - \theta_m\right)} \right].\end{aligned}$$

Direct calculations give

$$\begin{aligned}[\nabla \ell_n^{(m)}(\theta)]_m &= \sum_{j \in [n] \setminus \{m\}} A_{mj} \left[\psi(\theta_m^* - \theta_j^*)(\psi(\theta_m^* - \theta_j^*) - 1) + (1 - \psi(\theta_m^* - \theta_j^*))\psi(\theta_m - \theta_j) \right] \\ &= \sum_{j \in [n] \setminus \{m\}} A_{mj} \left[-\psi(\theta_m^* - \theta_j^*) + \psi(\theta_m - \theta_j) \right].\end{aligned}$$

Thus, we have

$$\theta_m^{(t+1,m)} - \theta_m^* = \left(1 - \eta\rho - \eta \sum_{j \in [n] \setminus \{m\}} A_{mj} \psi'(\xi_j) \right) (\theta_m^{(t,m)} - \theta_m^*) - \rho\eta\theta_m^* + \eta \sum_{j \in [n] \setminus \{m\}} A_{mj} \psi'(\xi_j)(\theta_j^{(t,m)} - \theta_j^*),$$

where ξ_j is a scalar between $\theta_m^* - \theta_j^*$ and $\theta_m^{(t,m)} - \theta_j^{(t,m)}$. Notice that $\psi'(x) = \frac{e^x}{(1+e^x)^2} \le \frac{1}{4}$ for any $c \in \mathbb{R}$, thus by Cauchy-Schwartz inequality we have

$$|\sum_{j \in [n] \setminus \{m\}} A_{mj} \psi'(\xi_j)(\theta_j^{(t,m)} - \theta_j^*)| \le \frac{1}{4} \sqrt{n_{\max}} \|\theta^{(t,m)} - \theta^*\|_2. \tag{25}$$

Also, since $\eta \le \frac{1}{\rho + n_{\max}}$,

$$1 - \eta\rho - \eta \sum_{j \in [n] \setminus \{m\}} A_{mj} \psi'(\xi_j) \ge 1 - \eta\rho - \eta \frac{n_{\max}}{4} \ge 0.$$

Therefore,

$$0 \le 1 - \eta\rho - \eta \sum_{j \in [n] \setminus \{m\}} A_{mj} \psi'(\xi_j) \le 1 - \eta n_{\min} \min_{j \in \mathcal{N}(m)} \psi'(\xi_j).$$

Since ξ_j is a scalar between $\theta_m^* - \theta_j^*$ and $\theta_m^{(t,m)} - \theta_j^{(t,m)}$, we have

$$\begin{aligned}\max_{j \in \mathcal{N}(m)} |\xi_j| &\le \max_{j \in \mathcal{N}(m)} |\theta_m^* - \theta_j^*| + \max_{j \in \mathcal{N}(m)} |\theta_m^* - \theta_j^* - (\theta_m^{(t,m)} - \theta_j^{(t,m)})| \\ &\le \kappa_E(\theta^*) + 2\|\theta^{(t,m)} - \theta^*\|_\infty \le \kappa_E(\theta^*) + \epsilon\end{aligned}$$

as long as

$$\|\theta^{(t,m)} - \theta^*\|_\infty \le f_c + f_d \le \epsilon/2. \tag{26}$$

Let $\epsilon = 0.2$, then $e^\epsilon \le 5/4$ and

$$\psi'(\xi_j) = \frac{e^{\xi_j}}{(1 + e^{\xi_j})^2} = \frac{e^{-|\xi_j|}}{(1 + e^{-|\xi_j|})^2} \ge \frac{e^{-|\xi_j|}}{4} \ge \frac{1}{4e^{\epsilon + \kappa_E(\theta^*)}} \ge \frac{1}{5e^{\kappa_E(\theta^*)}}.$$

By triangle inequality we get

$$\begin{aligned}|\theta_m^{(t+1,m)} - \theta_m^*| &\le \left(1 - \frac{\eta n_{\min}}{10 e^{\kappa_E(\theta^*)}} \right) |\theta_m^{(t,m)} - \theta_m^*| + \rho\eta \|\theta^*\|_\infty + \frac{\eta\sqrt{n_{\max}}}{4} \|\theta^{(t,m)} - \theta^*\|_2 \\ &\le f_b - \frac{\eta n_{\min}}{10 e^{\kappa_E(\theta^*)}} f_b + \eta\rho\kappa(\theta^*) + \eta \frac{\sqrt{n_{\max}}}{4} (f_a + f_c) \le f_b\end{aligned} \tag{27}$$

as long as

$$\frac{n_{\min}}{10e^{\kappa_E(\theta^*)}} f_b \geq \rho\kappa(\theta^*) + \frac{\sqrt{n_{\max}}}{4}(f_a + f_c).$$

By assumption, $f_a = C_a \frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)} \left[\sqrt{\frac{n_{\max}(n+r)}{L}} + \rho\kappa(\theta^*)\sqrt{n}\right]$, $f_c = C_c \frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)} \sqrt{\frac{n_{\max}(\log n + r)}{L}}$ with $C_a \gg \max\{C_c, 1\}$, so

$$f_a \gg f_c, \text{ and } \frac{\sqrt{n_{\max}}}{4} f_a \gg \frac{n_{\max}}{\lambda_2(\mathcal{L}_A)} \left[\sqrt{\frac{n+r}{L}} + \sqrt{\frac{n}{n_{\max}}}\rho\kappa(\theta^*)\right] \geq \rho\kappa(\theta^*).$$

Therefore, a sufficient condition for $|\theta_m^{(t+1,m)} - \theta_m^*| \leq f_b$ is

$$\frac{n_{\min}}{10e^{\kappa_E(\theta^*)}} f_b \geq \frac{3\sqrt{n_{\max}}}{4} f_a,$$

which is satisfied by our assumption. □

Lemma 14. *Suppose* (13) *holds with* $f_c = C_c \frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)} \sqrt{\frac{n_{\max}(\log n + r)}{L}}$ *for some sufficiently large constant* C_c, $f_d = f_b + f_c$, *then as long as the step size satisfies* $0 < \eta \leq \frac{1}{\rho + n_{\max}}$, *with probability at least* $1 - O(\kappa^{-2} e^{-3\kappa_E} n^{-10})$ *we have*

$$\max_{m \in [n]} \|\theta^{(t+1,m)} - \theta^{(t)}\|_2 \leq f_c.$$

Proof. By the update rules, we have

$$\begin{aligned}
\theta^{(t+1)} - \theta^{(t+1,m)} =& \theta^{(t)} - \eta\nabla\ell_\rho(\theta^{(t)}) - \left[\theta^{(t,m)} - \eta\nabla\ell_\rho^{(m)}(\theta^{(t,m)})\right] \\
=& \theta^{(t)} - \eta\nabla\ell_\rho(\theta^{(t)}) - \left[\theta^{(t,m)} - \eta\nabla\ell_\rho(\theta^{(t,m)})\right] \\
& - \eta\left[\nabla\ell_\rho(\theta^{(t,m)}) - \nabla\ell_\rho^{(m)}(\theta^{(t,m)})\right] \\
=& v_1 - v_2,
\end{aligned}$$

where

$$v_1 = \left[I_n - \eta\int_0^1 \nabla^2\ell_\rho(\theta(\tau))\mathrm{d}\tau\right](\theta^{(t)} - \theta^{(t,m)}), \quad v_2 = \eta\left[\nabla\ell_\rho(\theta^{(t,m)}) - \nabla\ell_\rho^{(m)}(\theta^{(t,m)})\right]$$

Now following the same arguments towards (22), as long as $\eta \leq \frac{1}{\rho + n_{\max}}$, we can get

$$\|v_1\|_2 \leq (1 - \frac{\eta\lambda_2(\mathcal{L}_A)}{10e^{\kappa_E(\theta^*)}})\|\theta^{(t)} - \theta^{(t,m)}\|_2.$$

For v_2, we know that

$$\begin{aligned}
\frac{1}{\eta}v_2 =& \sum_{i\in[n]\setminus\{m\}} \left\{A_{mi}\left[\psi(\theta_i^{(t,m)} - \theta_m^{(t,m)}) - \bar{y}_{im}\right] - A_{mi}\left[\psi(\theta_i^{(t,m)} - \theta_m^{(t,m)}) - \psi(\theta_i^* - \theta_m^*)\right]\right\}(\boldsymbol{e}_i - \boldsymbol{e}_m) \\
=& \sum_{i\in[n]\setminus\{m\}} A_{mi}\left[\psi(\theta_i^* - \theta_m^*) - \bar{y}_{im}\right](\boldsymbol{e}_i - \boldsymbol{e}_m).
\end{aligned}$$

By the form of the derivatives and Lemma 16, we know that with probability at least $1 - O(n^{\kappa^{-2}e^{-3\kappa_E}n^{-10}})$,

$$\begin{aligned}
\|\frac{1}{\eta}v_2\|_2^2 =& \left[\sum_{i\in[n]\setminus\{m\}} A_{im}\left(\bar{y}_{im} - \psi\left(\theta_i^* - \theta_m^*\right)\right)\right]^2 + \sum_{i\in[n]\setminus\{m\}} A_{im}\left(\bar{y}_{im} - \psi\left(\theta_i^* - \theta_m^*\right)\right)^2 \\
\lesssim& \frac{n_{\max}(\log n + r)}{L} + \frac{\log n + n_{\max} + r}{L} \lesssim \frac{n_{\max}(\log n + r)}{L}.
\end{aligned} \tag{28}$$

Therefore, we have

$$\begin{aligned}\|\theta^{(t+1)}-\theta^{(t+1,m)}\|_2 &\le \|v_1\|_2+\|v_2\|_2\\ &\le (1-\frac{\eta\lambda_2(\mathcal{L}_A)}{10e^{\kappa_E(\theta^*)}})\|\theta^{(t)}-\theta^{(t,m)}\|_2+C\eta\sqrt{\frac{n_{\max}(\log n+r)}{L}}\\ &\le (1-\frac{\eta\lambda_2(\mathcal{L}_A)}{10e^{\kappa_E(\theta^*)}})f_c+C\eta\sqrt{\frac{n_{\max}(\log n+r)}{L}}\le f_c,\end{aligned} \tag{29}$$

where the last inequality is due to the fact that C_c is a sufficiently large constant by our assumption and

$$\frac{\eta\lambda_2(\mathcal{L}_A)}{30e^{\kappa_E(\theta^*)}}f_c\ge C\eta\sqrt{\frac{n_{\max}(\log n+r)}{L}}\Leftarrow f_c=C_c\frac{e^{\kappa_E(\theta^*)}}{\lambda_2(\mathcal{L}_A)}\sqrt{\frac{n_{\max}(\log n+r)}{L}}. \tag{30}$$

□

Lemma 15. *Suppose* (13) *holds and* $f_d\ge f_b+f_c$*, then with probability at least* $1-O(\kappa^{-2}e^{-3\kappa_E}n^{-10})$ *we have*

$$\|\theta^{(t+1)}-\theta^*\|_\infty\le f_d.$$

Proof. By Lemma 13 and Lemma 14 we have

$$\begin{aligned}|\theta_m^{(t+1)}-\theta_m^*| &\le |\theta_m^{(t+1)}-\theta_m^{(t+1,m)}|+|\theta_m^{(t+1,m)}-\theta_m^*|\\ &\le \|\theta^{(t+1)}-\theta^{(t+1,m)}\|_2+|\theta_m^{(t+1,m)}-\theta_m^*|\le f_c+f_b\le f_d,\end{aligned}$$

since $f_d\ge f_b+f_c$ by our assumption. □

Lemma 16. *With probability at least* $1-O(\kappa^{-2}e^{-3\kappa_E}n^{-10})$ *it holds that*

$$\begin{aligned}\max_{i\in[n]}\left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right]^2 &\le C\frac{(\log n+r)\cdot n_{\max}}{L},\\ \sum_{i=1}^n\left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right]^2 &\le C\frac{(n+r)\cdot n_{\max}}{L}.\\ \max_{i\in[n]}\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]^2 &\le C\frac{\log n+n_{\max}+r}{L},\end{aligned} \tag{31}$$

where $r=\log\kappa+\kappa_E$.

Proof. To prove the first inequality, notice that by Hoeffding's inequality we have

$$\mathbb{P}\left(\sum_{j<i}A_{ij}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\ge\sqrt{\frac{8n_{\max}(\log n+r)}{L}}\right)\le 2\exp(-\frac{2L}{n_{\max}}\cdot\frac{8n_{\max}(\log n+r)}{L})=2\kappa^{-2}e^{-3\kappa_E}n^{-12},$$

where $r=\log\kappa+\kappa_E$. By union bound we know that on an event B with probability at least $1-\kappa^{-2}e^{-3\kappa_E}n^{-10}$ we have $\forall i\in[n]$, $\left[\sum_{j<i}A_{ij}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right]^2<\frac{8n_{\max}(\log n+r)}{L}$.

Next we prove the second inequality. Consider the unit ball $\mathcal{S}=\{v\in\mathbb{R}^n:\sum_{i\in[n]}v_i^2=1\}$ in $\mathbb{R}^n$. By Lemma 5.2 of Vershynin (2011), we can pick a subset $\mathcal{U}\subset\mathcal{S}$ so that $\log|\mathcal{U}|\le cn$ and for any $v\in\mathcal{S}$, there exists a vector $u\in\mathcal{U}$ such that $\|u-v\|_2\le\frac{1}{2}$. For a given $v\in\mathcal{S}$, pick $u\in\mathcal{U}$ such that $\|u-v\|_2\le\frac{1}{2}$ and we have

$$\begin{aligned}\sum_{i=1}^n v_i\left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right] &= \sum_{i=1}^n u_i\left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right]+\sum_{i=1}^n(v_i-u_i)\left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right]\\ &\le \sum_{i=1}^n u_i\left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right]+\frac{1}{2}\sqrt{\sum_{i=1}^n\left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij}-\psi(\theta_i^*-\theta_j^*)]\right]^2}.\end{aligned}$$

Taking maximum over v and the left hand side can achieve $\sqrt{\sum_{i=1}^n \left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]\right]^2}$, thus we have

$$\begin{aligned}\sqrt{\sum_{i=1}^n \left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]\right]^2} &\le 2\max_{u\in\mathcal{U}} \sum_{i=1}^n u_i \left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]\right] \\ &= 2\max_{u\in\mathcal{U}} \sum_{i<j} A_{ij}(u_i - u_j)\left[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)\right].\end{aligned}$$

To apply Hoeffding's inequality and union bound, we should account for $|\mathcal{U}| \le e^{cn}$, so we can get that with probability at least $1 - \kappa^{-2}e^{-3\kappa_E}n^{-10}$ it holds that

$$\begin{aligned}\sum_{i=1}^n \left[\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]\right]^2 &\le C_1\frac{1}{L}\left[(\log n + n + r)\sum_{i<j} A_{ij}(u_i - u_j)^2\right] \\ &\le C_1\frac{(\log n + n + r)\lambda_{\max}(\mathcal{L}_A)}{L},\end{aligned}$$

where $r = \log\kappa + \kappa_E$. Since $\lambda_{\max}(\mathcal{L}_A) \le 2n_{\max}$, the second inequality is proved.

Next we prove the third inequality. For each $i \in [n]$, let $\mathcal{V}_i := \{v \in \mathbb{R}^{n-1} : \sum_{j\ne i} A_{ij}v_j^2 \le 1\}$. By Lemma 5.2 of Vershynin (2011), we can pick a subset $\mathcal{U}_i \subset \mathcal{V}_i$ so that $\log|\mathcal{U}_i| \le 2\sum_{j\ne i} A_{ij}$ and for any $v \in \mathcal{V}_i$, there exists a vector $u \in \mathcal{U}_i$ such that $\|u - v\|_2 \le \frac{1}{2}$. For a given $v \in \mathcal{V}_i$, pick $u \in \mathcal{U}_i$ such that $\|u - v\|_2 \le \frac{1}{2}$ and we have

$$\begin{aligned}\sum_{j\in\mathcal{N}(i)} v_{ij}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)] &= \sum_{j\in\mathcal{N}(i)} u_{ij}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)] + \sum_{j\in\mathcal{N}(i)} (v_{ij} - u_{ij})[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)] \\ &\le \sum_{j\in\mathcal{N}(i)} u_{ij}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)] + \frac{1}{2}\sqrt{\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]^2}\end{aligned}$$

Taking maximum over v and the left hand side can achieve $\sqrt{\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]^2}$, thus we have

$$\sqrt{\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]^2} \le 2\max_{u\in\mathcal{U}_i}\sum_{j\in\mathcal{N}(i)} u_{ij}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)].$$

Therefore,

$$\sqrt{\max_{i\in[n]}\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]^2} \le 2\max_{i\in[n]}\max_{u\in\mathcal{U}_i}\sum_{j\in\mathcal{N}(i)} u_{ij}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)].$$

Now a straightforward application of Hoeffding's inequality and union bound gives

$$\max_{i\in[n]}\sum_{j\in\mathcal{N}(i)}[\bar{y}_{ij} - \psi(\theta_i^* - \theta_j^*)]^2 \le C\frac{1}{L}\left[\log n + n_{\max} + \kappa_E + \log\kappa\right]$$

with probability at least $1 - O(\kappa^{-2}e^{-3\kappa_E}n^{-10})$. □

Corollary 2 (Erdös-Rényi graph)**.** Suppose that the comparison graph comes from an Erdös-Rényi graph $ER(n,p)$. Assume that $\mathbf{1}_n^\top\boldsymbol{\theta}^* = 0$, $\kappa \le n$, $\kappa_E \le \log n$, $L \le n^8 e^{4\kappa_E}\max\{1,\kappa\}$, $np > C_1\log n$, and $L \ge C_2\ \max\{1,\kappa\}e^{4\kappa_E}n/\log^2 n$ for some sufficiently large constants $C_1, C_2 > 0$. Set $\rho = c_\rho/\kappa\sqrt{n_{max}/L}$. Then $\mathbf{1}_n^\top\hat{\boldsymbol{\theta}}_\rho = 0$, and with probability at least $1 - O(n^{-4})$, it holds that

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_\infty &\lesssim e^{2\kappa_E}\sqrt{\frac{1}{np^2L}} + e^{\kappa_E}\sqrt{\frac{\log n}{npL}}, \\ \|\hat{\boldsymbol{\theta}}_\rho - \boldsymbol{\theta}^*\|_2 &\lesssim e^{\kappa_E}\sqrt{\frac{1}{pL}}.\end{aligned} \tag{32}$$

Proof of Corollary 2. For an $ER(n,p)$ graph $\mathcal{G}$ with $p \geq c\frac{\log n}{n}$ for some larege $c > 0$, it holds with probability at least $1 - O(n^{-10})$ that $\mathcal{G}$ is connected, and

$$\frac{1}{2}np \leq n_{\min} \leq n_{\max} \leq 2np,$$

and

$$\lambda_2\left(\mathcal{L}_A\right) = \min_{u \neq 0: \mathbf{1}_n^\top u = 0} \frac{u^\top \mathcal{L}_A u}{\|u\|^2} \geq \frac{np}{2}.$$

The proof can be seen in either Chen et al. (2019) or Chen et al. (2020). Thus, by a union bound, we can replace the corresponding quantities in upper bounds in Theorem 1 and get the high probability bounds in Corollary 2. □

Proposition 3 (Path graph)**.** Suppose the comparison graph is a path graph $([n], E)$ with $E = \{(i, i+1)\}_{i \in [n-1]}$ and for each i, item i and item $i+1$ are compared $L_{i,i+1}$ times such that $\min_i L_{i,i+1} > ce^{2\kappa_E} n \log n$ for some universal constant c, then with probability at least $1 - n^{-4}$, the vanilla MLE $\hat{\boldsymbol{\theta}}_0$ satisfies

$$\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_\infty \lesssim \sqrt{\sum_{i=1}^{n-1} \frac{\exp(2|\theta_i^* - \theta_{i+1}^*|)\log n}{L_{i,i+1}}}. \tag{33}$$

In particular, when $L_{i,i+1} = L$ for all $i \in [n-1]$, we have

$$\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_\infty \lesssim e^{\kappa_E}\sqrt{\frac{n \log n}{L}}, \|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \lesssim e^{\kappa_E} n\sqrt{\frac{\log n}{L}} \tag{34}$$

Proof of Proposition 3. Consider a path graph with edge set $E = \{(i, i+1) : i \in [n-1]\}$. Let M_{ij} be the number of wins of item i against item j. For the ease of notations, in this proof we use $\hat{\theta}$ to denote the vanilla MLE $\hat{\boldsymbol{\theta}}_0$. We know that

$$\nabla \ell\left(\theta\right)_i = \sum_{j \in \mathcal{N}(i)} \left(\frac{M_{ij}}{M_{ij} + M_{ji}} - \frac{\exp(\theta_i - \theta_j)}{1 + \exp(\theta_i - \theta_j)}\right).$$

Thus the vanilla MLE solving $\nabla \ell(\theta) = 0$ is given by

$$\hat{\theta}_{i+1} - \hat{\theta}_i = \log M_{i+1,i} - \log M_{i,i+1} := \log R_{i+1,i},$$

where $M_{ij} := \#\{i \text{ beats } j\}$ and $R_{ij} := \frac{M_{ij}}{M_{ji}}$.

Let $\hat{\theta}_1 = 0$ and we can get

$$\hat{\theta}_1 = 0,\ \hat{\theta}_{i+1} = \sum_{j=1}^{i} (\log R_{j+1,j} - \log r_{j+1,j}),\ i = 1, \cdots, n-1,$$

where $\log R_{j+1,j} := \log M_{j+1,j} - \log M_{j,j+1}$, $\log r_{j+1,j} = \theta_{j+1}^* - \theta_j^*$. Shifting θ^* to make $\theta_1^* = 0$ and we have

$$|\hat{\theta}_{i+1} - \theta_{i+1}^*| = \left|\sum_{j=1}^{i} (\log R_{j+1,j} - \log r_{j+1,j})\right| \tag{35}$$

Let $F_{ij} := \frac{M_{ij}}{L_{ij}}$ with $L_{ij} = M_{ij} + M_{ji}$ and we have

$$\log R_{ij} - \log r_{ij} = \log \frac{1 - F_{ji}}{F_{ji}} - \log \frac{1 - p_{ji}}{p_{ji}}.$$

Now using the Talyor expansion of $f(x) = \log(\frac{1}{x} - 1)$ at p_{ji}, we can get

$$\log R_{ij} - \log r_{ij} = -v_{ij}(F_{ij} - p_{ij}) + \frac{1}{2}\frac{1 - 2z_{ji}}{z_{ji}^2(1 - z_{ji})^2}(F_{ij} - p_{ij})^2,$$

where $v_{ij} := [p_{ij}(1-p_{ij})]^{-1} = 2 + r_{ij} + r_{ji} = v_{ji} \le 4\exp(|\theta_i^* - \theta_j^*|)$, and z_{ji} is a number between p_{ji} and F_{ji}. We know that $v_{ij}(F_{ij} - p_{ij})$ is a subgaussian-$\frac{v_{ij}^2}{4T_{ij}}$ variable. In particular, for the path graph, by a standard sub-Gaussian tail bound on the first term, we can show that with probability at least $1 - n^{-9}$, it holds for $i = 1, \cdots, n-1$ simultaneously and some constant $C > 0$ that

$$|\sum_{j=1}^{i} v_{j,j+1}(F_{j,j+1} - p_{j,j+1})| \le C\sqrt{\sum_j \frac{v_{j,j+1}^2 \log n}{L_{j,j+1}}}.$$

Using Chernoff's method we can show that if $\min_j L_{ij} > 25e^{2\kappa_E}\log n$ then with probability at least $1 - n^{-9}$,

$$\max_j |F_{ij} - p_{ij}| \le C\sqrt{\frac{\log n}{L_{ij} v_{ij}}}.$$

Since $e^{\kappa_E} = \max_{(i,j)\in E} \frac{p_{ji}}{p_{ij}} = \frac{1}{\frac{1}{p_{ij}p_{ji}} p_{ij}^2} = \frac{1}{v_{ij}p_{ij}^2}$, the last inequality and the condition on L_{ij} imply that $|F_{ij} - p_{ij}| \le \min\{p_{ij}, p_{ji}\}/5$ simultaneously, and consequently $z_{ji} \in [0.8p_{ji}, 1.2p_{ji}]$. Therefore, the second term can be bounded as

$$\frac{1}{2}\frac{1-2z_{ji}}{z_{ji}^2(1-z_{ji})^2}(F_{ij} - p_{ij})^2 \le c\frac{1}{p_{ij}^2 p_{ji}^2}(F_{ij} - p_{ij})^2 \le c\frac{v_{ij}\log n}{L_{ij}}.$$

In particular, for the path graph, the summation of the second term is controlled by $\sum_j \frac{v_{j,j+1}\log n}{L_{j,j+1}}$. Thus, when

$$\min_j L_{j,j+1} > ce^{\kappa_E} n \log n,$$

for some sufficiently large constant c, the summation of the second term is negligible compared to the bound on the summation of the first term $\sqrt{\sum_j \frac{v_{j,j+1}^2 \log n}{L_{j,j+1}}}$. Therefore, the error (35) can be bounded as

$$|\hat{\theta}_{i+1} - \theta_{i+1}^*| \le C\sqrt{\sum_{j=1}^{i} \frac{\exp(2|\theta_j^* - \theta_{j+1}^*|)\log n}{L_{j,j+1}}},$$

which implies that $\|\hat{\theta} - \theta^*\|_\infty \le C\sqrt{\sum_{i=1}^{n-1} \frac{\exp(2|\theta_i^* - \theta_{i+1}^*|)\log n}{L_{i,i+1}}}$.

In the special case $L_{i,i+1} = L$ for all i, we have $\|\hat{\theta} - \theta^*\|_\infty \le C\sqrt{\frac{\exp(2\kappa_E) n \log n}{L}}$ and $\|\hat{\theta} - \theta^*\|_2 \le n\sqrt{\frac{\exp(2\kappa_E)\log n}{L}}$. □

In Shah (2016) they prove the lower bound $e^{-\kappa}\frac{n}{\sqrt{L}}$ for ℓ_2 error under path graph, while in our paper we prove the lower bound $e^{-\kappa}\sqrt{\frac{n}{L}}$ for ℓ_∞ error under path graph. Thus the upper bound above on the closed-form MLE achieves the minimax lower bound up to a $\sqrt{\log n}$ and $\exp(2\kappa_E)$ factor.

Proposition 4 (General tree graph)**.** Suppose the graph is a tree graph $([n], E)$ where item i and j are compared L_{ij} times such that $\min_{i,j} L_{ij} > ce^{2\kappa_E} n \log n$ for some universal constant c. Then with probability at least $1 - n^{-4}$, the vanilla MLE $\hat{\boldsymbol{\theta}}_0$ satisfies

$$\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_\infty \lesssim \sqrt{\max_{i_1,i_2} \sum_{(i,j)\in \mathrm{path}(i_1,i_2)} \frac{\exp(2|\theta_i^* - \theta_j^*|)\log n}{L_{ij}}}. \tag{36}$$

In particular, when $L_{i,j} = L$, we have

$$\|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_\infty \lesssim e^{\kappa_E}\sqrt{\frac{d\log n}{L}}, \|\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \lesssim e^{\kappa_E}\sqrt{\frac{dn\log n}{L}} \tag{37}$$

Proof of Proposition 4. Suppose $\mathcal{G} = ([n], E)$ is a tree graph and let d be the diameter of the graph, i.e., $d = \max_{i,j\in[n]} |\mathrm{path}(i,j)|$ where $\mathrm{path}(i,j) := \{(i,i_1),(i_2,i_3),\cdots,(i_m,j)\} \subseteq E$ is the shortest path from i to j. For instance, for complete graph and star graph, $d = 1$; for path graph, $d = n-1$; for a complete binary tree with depth or height h, $d = 2h$.

The key property of a tree graph is that it has three equivalent definitions: 1. it is a maximal loop-free graph; 2. it is a minimal connected graph; 3. it is a simple graph with $|E| = |V| - 1$. By the first definition, it can be seen that for any two nodes i and j, there is exactly one path from i to j, otherwise there will be a loop. Using this property, for a fixed item i_0, the MLE equation

$$\nabla \ell(\theta)_i = \sum_{j \in \mathcal{N}(i)} \left(\frac{M_{ij}}{M_{ij} + M_{ji}} - \frac{\exp(\theta_i - \theta_j)}{1 + \exp(\theta_i - \theta_j)} \right) = 0, \ \forall i \in [n]$$

can be solved by

$$\hat{\theta}_{i_0} = 0, \quad \hat{\theta}_l = \sum_{(i,j) \in \mathrm{path}(i_0, l)} (\log R_{ji} - \log r_{ji}), \ l \neq i_0.$$

Note that, the choice of i_0 is not crucial for the ℓ_∞ bound, as such shifting would at most lead to an inflation on the ℓ_∞ error by 2. In fact, if we let $\hat{\boldsymbol{\epsilon}}$ with $\hat{\epsilon}_{i_0} = 0$ be the entry-wise error when we choose $\hat{\theta}_{i_0} = \theta_{i_0} = 0$, then after shifting $\hat{\boldsymbol{\theta}}$ to get $\tilde{\boldsymbol{\theta}}$ with $\tilde{\theta}_{i_1} = \theta_{i_1} = 0$ for some $i_1 \neq i_0$ we will get $\|\tilde{\boldsymbol{\epsilon}}\|_\infty \leq \|\hat{\boldsymbol{\epsilon}}\|_\infty + \|\hat{\epsilon}_{i_1} \mathbf{1}_n\|_\infty \leq 2\|\hat{\boldsymbol{\epsilon}}\|_\infty$.

Following the same argument in the proof of Proposition 3, we can show that when $\min_{(i,j) \in E} L_{i,j} > e^{2\kappa_E} n \log n$, we have

$$\|\hat{\theta} - \theta^*\|_\infty \lesssim \sqrt{\max_{i_1, i_2} \sum_{(i,j) \in \mathrm{path}(i_1, i_2)} \frac{\exp(2|\theta_i^* - \theta_j^*|) \log n}{L_{ij}}}.$$

Again, the overall ℓ_∞ error is determined by the extreme values of $|\theta_i^* - \theta_j^*|$ and $1/L_{ij}$.

To simplify the bound, we can consider the uniform sampling scheme under which $L_{ij} = L$ for all $(i,j) \in E$. By the definition $d = \max_{i,j \in [n]} |\mathrm{path}(i,j)|$, we have

$$\|\hat{\theta} - \theta^*\|_\infty \lesssim \sqrt{\frac{\exp(2\kappa_E) d \log n}{L}}, \tag{38}$$

and as a corollary,

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{\exp(2\kappa_E) dn \log n}{L}}. \tag{39}$$

In particular, for a path graph, $d = n - 1$ and the two bounds become the same bounds in Proposition 3.

For a star graph, $d = 1$ and the bounds become

$$\|\hat{\theta} - \theta^*\|_\infty \lesssim \sqrt{\frac{\exp(2\kappa_E) \log n}{L}}, \quad \|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{\exp(2\kappa_E) n \log n}{L}},$$

which are both sharp up to a $\log n$ factor according to the lower bounds in Shah et al. (2016) and our work. In addition, for the star graph, one can actually show with much simpler arguments that the condition on L can be relaxed to $L > ce^{2\kappa_E} \log n$, because in a star graph essentially we are just comparing $n - 1$ pairs separately. □

A.3 PROOF IN SECTION 3

Let $F(t) = \frac{1}{1+e^{-t}}$, then the Bradley-Terry model can be written as $p_{ij} = F[(w_i^* - w_j^*)/\sigma]$ with $\sigma = 1$. We have $\max_{t\in[0,2\kappa/\sigma]} F'(t) = F'(0) = 1/4$, so ζ in Lemma 21 satisfies

$$\zeta = ce^{\frac{2\kappa}{\sigma}}.$$

Moreover, we denote $\mathcal{W} := \{\boldsymbol{\theta}^* \mid \|\boldsymbol{\theta}^*\|_\infty \le B\}$. Since $\mathbf{1}_n^\top \boldsymbol{\theta}^* = 0$, we have $B \asymp \kappa := \max_{i,j} |\theta_i^* - \theta_j^*|$ (in general, there is no more information, but for some special cases, e.g., when entries of $\boldsymbol{\theta}^*$ are equal-spaced, we have $\kappa = 2B$). In what follows in this section, we still use quantity σ for generality, but keep in mind that in our setting $\sigma = 1$. We still use B to differ it from κ.

A.3.1 Background and Required Results

Before writing our proof of Theorem 5 we note down specific results from (Shah et al., 2016) and other sources that we will use repeatedly.

Definition 17 (Pairwise $(\delta, \beta)-$packing set from (Shah et al., 2016))**.** Let $\theta \in \mathbb{R}^n$ be a parameter to be estimated, as indexed over a class of probability distributions $\mathcal{P} := \{\mathbb{P}_\theta \mid \theta \in \mathcal{W}\}$, and let $\rho : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ be a pseudo-metric. Suppose there exist a finite set of M vectors $\left\{\theta^1, \dots, \theta^M\right\}$ such that the following conditions hold:

$$\min_{\substack{j,k\in[M]\\ j\neq k}} \rho\left(\theta^j, \theta^k\right) \ge \delta \text{ and } \frac{1}{\binom{M}{2}} \sum_{\substack{j,k\in[M]\\ j\neq k}} D_{\mathrm{KL}}\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right) \le \beta$$

Then we refer to $\left\{\theta^1, \dots, \theta^M\right\}$ as $(\delta, \beta)-$packing set.

Lemma 18 (Fano minimax lower bound)**.** *Suppose that we can construct a $(\delta, \beta)-$packing set with cardinality M, then the minimax risk is lower bounded as:*

$$\inf_{\widehat{\theta}} \sup_{\theta^*\in\mathcal{W}} \mathbb{E}\left[\rho\left(\widehat{\theta}, \theta^*\right)\right] \ge \frac{\delta}{2}\left(1 - \frac{\beta + \log 2}{\log M}\right)$$

Proof. See (Yu, 1997, Lemma 3) for details. □

Lemma 19 (Equivalence of $\|\cdot\|_\infty$ and $\|\cdot\|_2$ norms)**.** *Given any vector $\theta \in \mathbb{R}^n$, with $n \in \mathbb{N}$ fixed, the following inequalities holds:*

$$\frac{1}{\sqrt{n}} \|\theta\|_2 \le \|\theta\|_\infty \le \|\theta\|_2$$

Proof. The result is standard e.g. see (Wendland, 2018, Proposition 2.10) for a more general version and proof. For the sake of completeness we provide a direct proof of the equivalent statement $\|\theta\|_\infty \le \|\theta\|_2 \le \sqrt{n}\,\|\theta\|_\infty$ as follows:

$$\|\theta\|_\infty = \max_{i\in[n]} |\theta_i| = \sqrt{\max_{i\in[n]} |\theta_i|^2} \le \sqrt{\sum_i^n {\theta_i}^2} \le \|\theta\|_2$$

$$\|\theta\|_2 = \sqrt{\sum_i^n {\theta_i}^2} \le \sqrt{\sum_i^n \max_{i\in[n]} |\theta_i|^2} = \sqrt{n \max_{i\in[n]} |\theta_i|^2} = \sqrt{n}\,\|\theta\|_\infty$$

This proves the lower and upper inequalities respectively, as required. □

Remark 3. We note that the above inequalities are tight i.e. have optimal constants. In the case of the upper bound consider $\theta = \mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$. So $\|\theta\|_\infty = 1$ and $\|\theta\|_2 = \sqrt{n} = \sqrt{n}\,\|\theta\|_\infty$ showing the tightness of the upper bound. In the case of the lower bound, consider $\theta = \mathbf{e}_1$ i.e. WLOG the first standard basis vector in $\mathbb{R}^n$. We then have that $\|\theta\|_\infty = \|\theta\|_2 = 1$, which shows tightness in the lower bound.

Lemma 20 (Lemma 7 in (Shah et al., 2016)). *For any $\alpha \in \left(0, \frac{1}{4}\right)$, there exists a set of $M(\alpha)$ binary vectors $\left\{z^1, \ldots, z^{M(\alpha)}\right\} \subset \{0,1\}^n$ such that*

$$\alpha n \leq \left\|z^j - z^k\right\|_2^2 \leq n \quad \textit{for all } j \neq k \in [M(\alpha)], \textit{ and} \tag{40}$$

$$\langle e_1, z^j \rangle = 0 \quad \textit{for all } j \in [M(\alpha)] \tag{41}$$

Lemma 21 (Lemma 8 in (Shah et al., 2016)). *For any pair of quality score vectors θ^j and θ^k, and for*

$$\zeta := \frac{\max_{x \in [0, 2B/\sigma]} F'(x)}{F(2B/\sigma)(1 - F(2B/\sigma))}$$

we have

$$D_{\mathrm{KL}}\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right) \leq \frac{N_{comp}\zeta}{\sigma^2}\left(\theta^j - \theta^k\right)^\top L\left(\theta^j - \theta^k\right) =: \frac{N_{comp}\zeta}{\sigma^2}\left\|\theta^j - \theta^k\right\|_L^2$$

Lemma 22 (Lemma 14 in (Shah et al., 2016)). *The Laplacian matrix $\tilde{\mathcal{L}}_{\mathbf{A}}$ satisfies the trace constraints:*

$$\mathrm{tr}\left(\tilde{\mathcal{L}}_{\mathbf{A}}\right) = 2 \tag{42}$$

$$\mathrm{tr}\left(\tilde{\mathcal{L}}_{\mathbf{A}}^{\dagger}\right) \geq \frac{n^2}{4} \tag{43}$$

Proof. See (Shah et al., 2016, Lemma 14) for details. □

The challenge - constructing a suitable pairwise packing set that meets Definition 17. The main tool to use here is the Varshamov-Gilbert Lemma

A.3.2 Sketch of Lower Bound Proof

In brief, we seek a minimax lower bound proof of the same form as (Shah et al., 2016, Theorem 2(a)), except in our case we choose our packing set norm to the ℓ_∞ norm, rather than the ℓ_2^2 in (Shah et al., 2016, Theorem 2(a)). We leverage their construction directly in two main ways. First, we use the slightly modified version of Fano's Lemma that enables the (δ, β)-packing set to be constructed for the ℓ_∞ norm, not the ℓ_∞^2 norm, consistent with our high probability upper bound per Lemma 18. Second, we can switch out their use of the (δ, β)-packing set in the ℓ_2 norm to the (δ', β')-packing set in the ℓ_∞ norm. This is done by using the topological equivalence of norms in finite dimensions per Lemma 19, which is shown to be tight in the dimension per Remark 3. For the sake of clarity, we use much of the same wording as the proof from (Shah et al., 2016, Appendix B), for the convenience of the reader.

A.3.3 Lower Bound Proof - Part I

Our proof follows directly the approach taken from (Shah et al., 2016, Section B.1). The normalized Laplacian $\tilde{\mathcal{L}}_{\mathbf{A}}$ of the comparison graph is symmetric and positive-semidefinite. We can thus decompose this via diagonalization as $L = U^\top \Lambda U$ where $U \in \mathbb{R}^{n \times n}$ is an orthonormal matrix, and Λ is a diagonal matrix of nonnegative eigenvalues $\Lambda_{jj} = \lambda_j(L)$ for each $j \in [n]$. Similar to (Shah et al., 2016, Section B.1) we first prove that the minimax risk is lower bounded by $c\sigma^2 \frac{n^2}{N_{\text{comp}}}$.

Fix scalars $\alpha \in (0, \frac{1}{4})$ and $\delta > 0$, with values to be specified later. Obtain set of vectors on the Boolean Hypercube $\{0,1\}^n$ i.e. $\left\{z^1, \ldots, z^{M(\alpha)}\right\}$ given by Lemma 20, where $M(\alpha)$ is set to be

$$M(\alpha) := \left\lfloor \exp\left\{\frac{n}{2}(\log 2 + 2\alpha \log 2\alpha + (1 - 2\alpha)\log(1 - 2\alpha))\right\}\right\rfloor. \tag{44}$$

Define another set of vectors of the same cardinality $\left\{\theta^j \,\middle|\, j \in [M(\alpha)]\right\}$ via $\theta^j := \frac{\delta}{\sqrt{n}} U^\top P z^j$, where P is a permutation matrix. The permutation matrix P has the constraint that it keeps the first coordinate constant i.e. $P_{11} = 1$. By construction for each $j \neq k$ we have that

$$\left\|\theta^j-\theta^k\right\|_\infty \overset{(i)}{\geq} \frac{1}{\sqrt{n}}\left\|\theta^j-\theta^k\right\|_2 \overset{(ii)}{=} \frac{1}{\sqrt{n}}\left(\frac{\delta}{\sqrt{n}}\left\|z^j-z^k\right\|_2\right) \overset{(iii)}{\geq} \delta\sqrt{\frac{\alpha}{n}} \tag{45}$$

Here (i) follows from Lemma 19. Additionally (ii) follows since $\theta^j := \frac{\delta}{\sqrt{n}}U^\top P z^j$. In the case of the final inequality (iii), we have $\frac{\delta^2}{n^2}\left\|z^j-z^k\right\|_2^2 \geq \frac{\alpha n\delta^2}{n^2} = \frac{\alpha\delta^2}{n}$. Since the set $\left\{z^1,\ldots,z^{M(\alpha)}\right\}$ are binary vectors with a minimum Hamming distance at least αn using Equation (40). Consider, any distinct $j,k \in [M(\alpha)]$, then for some subset $\{i_1,\ldots,i_r\} \subseteq \{2,\ldots,n\}$ with $\alpha n \leq r \leq n$ it must follow that

$$\left\|\theta^j-\theta^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 = \frac{\delta^2}{n}\left\|U^\top P z^j - U^\top P z^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 = \frac{\delta^2}{n}\left\|z^j-z^k\right\|_\Lambda^2 = \frac{\delta^2}{n}\sum_{m=1}^{r}\lambda_{i_m}(\tilde{\mathcal{L}}_{\mathbf{A}})$$

The last part follows since Λ is a diagonal matrix of non-negative eigenvalues with $\Lambda_{ii} = \lambda_j(L)$. Now for given $\{a_2,\ldots,a_n\}$ such that $\alpha n \leq \sum_{i=2}^n a_i \leq n$ we have that

$$\frac{1}{\binom{M(\alpha)}{2}}\sum_{j\neq k}\left\|\theta^j-\theta^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 = \frac{\delta^2}{n}\sum_{i=2}^{n} a_i\lambda_i(\tilde{\mathcal{L}}_{\mathbf{A}})$$

The permutation matrix P is chosen such that the last $n-1$ coordinates are permuted to have $a_1 \geq \ldots \geq a_n$ and keep the nth coordinate fixed. By this particular choice, and using the fact that $\operatorname{tr}(\tilde{\mathcal{L}}_{\mathbf{A}}) = 2$ we have that:

$$\frac{1}{\binom{M(\alpha)}{2}}\sum_{j\neq k}\left\|\theta^j-\theta^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 = \frac{\delta^2}{n}\frac{n}{n-1}\operatorname{tr}(\tilde{\mathcal{L}}_{\mathbf{A}}) \leq \frac{2\delta^2}{n}\operatorname{tr}(\tilde{\mathcal{L}}_{\mathbf{A}}) = \frac{4\delta^2}{n}$$

Now by the choice of P above, we have that for every choice of $j \in [M(\alpha)]$

$$\langle\tilde{\mathcal{L}}_{\mathbf{A}},\theta^j\rangle = \frac{\delta}{\sqrt{n}}\mathbf{e}_1^\top P z^j = \mathbf{e}_1^\top z^j = 0$$

where the last equality follows from Equation (41). Now the basic condition needs to be verified i.e. did the θ^j we chose satisfy the boundedness constraint, to ensure that $\theta^j \in \mathcal{W}_B$? Setting $\delta^2 = 0.01\frac{\sigma^2 n^2}{4N_{\text{comp}}\zeta}$, it indeed follows that $\left\|\theta^j\right\|_\infty \leq \frac{\delta}{\sqrt{n}}\left\|z^j\right\|_2 \overset{(i)}{\leq} \delta \overset{(ii)}{\leq} B$. Here (i) follows since $z^j \in \{0,1\}^n$. Furthermore (ii) follows from our choice of δ and our assumption that $N_{\text{comp}} \geq \frac{c\sigma^2\operatorname{tr}(\tilde{\mathcal{L}}_{\mathbf{A}}^\dagger)}{\zeta B^2}$ with $c = 0.002$, where Lemma 22 guarantees that $N_{\text{comp}} \geq \frac{c\sigma^2 n^2}{4\zeta B^2}$. We have thus verified that each vector θ_j also satisfies the boundedness constraint $\left\|\theta^j\right\|_\infty \leq B$, which is required for membership in $\mathcal{W}_B$. Finally by Lemma 21 we have that:

$$D_{\text{KL}}\left(\mathbf{P}_{\theta^j}\|\mathbf{P}_{\theta^k}\right) \leq \frac{N_{\text{comp}}\zeta}{\sigma^2}\frac{4\delta^2}{n} = 0.01n \tag{46}$$

To summarize, we have now constructed a (δ',β')-packing set with respect to the norm $\rho\left(\theta^j,\theta^k\right) := \left\|\theta^j-\theta^k\right\|_\infty$, where $\delta' = \delta\sqrt{\frac{\alpha}{n}}$ from Equation (45), and $\beta' = 0.01d$ from Equation (46).

Finally we have by substituting (δ',β') into the pairwise Fano's lower bound (Lemma 18) that:

$$\sup_{\theta^*\in\mathcal{W}_B}\mathbb{E}\left[\left\|\widetilde{\theta}-\theta^*\right\|_\infty\right] \geq \delta\sqrt{\frac{\alpha}{n}}\left(1-\frac{0.01n+\log 2}{\log M(\alpha)}\right) = c\sigma\sqrt{\frac{n}{\zeta N_{\text{comp}}}}\left(1-\frac{0.01n+\log 2}{\log M(\alpha)}\right)$$

which yields the claim, after appropriate substitution of δ and setting $\alpha = 0.01$.

For the case of $n \leq 9$, consider the set of the three n-length vectors $z^1 = (0,\ldots,-1)$, $z^2 = (0,\ldots,0)$ and $z^3 = (0,\ldots,0)$. Construct the packing set $\left(\theta^1,\theta^2,\theta^3\right)$ from these three vectors $\left(z^1,z^2,z^3\right)$ as done above for the case of $n > 9$. From the calculations made for the general case above, we have for all pairs $\min_{j\neq k}\left\|\theta^j-\theta^k\right\|_\infty^2 \geq \frac{1}{9}\min_{j\neq k}\left\|\theta^j-\theta^k\right\|_2^2 \geq \frac{\delta^2}{81}$ and $\max_{j,k}\left\|\theta^j-\theta^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 \leq 4\delta^2$, and as a result $\max_{j,k} D_{\text{KL}}\left(\mathbf{P}_{\theta^j}\|\mathbf{P}_{\theta^k}\right) \leq \frac{4N_{\text{comp}}\zeta\delta^2}{\sigma^2}$. Choosing $\delta^2 = \frac{\sigma^2\log 2}{8N_{\text{comp}}\zeta}$ and applying the pairwise Fano's lower bound (Lemma 18) yields the claim.

For the other case, the lower bound in terms of $\lambda_i(\tilde{\mathcal{L}}_{\mathbf{A}})$, the argument is similar, and we end up with an extra factor of $\frac{1}{n'}$.

A.3.4 Lower Bound Proof - Part II

Given an integer $n' \in \{2, \dots, n\}$, and constants $\alpha \in (0, \frac{1}{4})$, $\delta > 0$, define the integer:

$$M'(\alpha) := \left\lfloor \exp\left\{ \frac{n'}{2} (\log 2 + 2\alpha \log 2\alpha + (1-2\alpha)\log(1-2\alpha)) \right\} \right\rfloor \tag{47}$$

Applying Lemma 20 using n' as the dimension results in a subset $\left\{ z^1, \dots, z^{M'(\alpha)} \right\}$ of the Boolean hypercube $\{0,1\}^{n'}$, with specified properties. We then define a finite set of size $M'(\alpha)$, of n-length vectors $\left\{ \widetilde{\theta}^1, \dots, \widetilde{\theta}^{M'(\alpha)} \right\}$ using:

$$\widetilde{\theta}^j = \left[0 \left(z^j\right)^\top 0 \cdots 0 \right]^\top \qquad \text{for each } j \in [M(\alpha)]$$

For each $j \in [M(\alpha)]$, let us define $\theta^j := \frac{\delta}{\sqrt{n'}} U^\top \sqrt{\Lambda^\dagger} \tilde{\theta}^j$. For the first standard basis vector $\mathbf{e}_1 \in \mathbb{R}^n$, we then have that $\langle \mathbf{1}_n, \theta^j \rangle = \frac{\delta}{\sqrt{n'}} \tilde{\mathcal{L}}_{\mathbf{A}}^\top U^\top \sqrt{\Lambda^\dagger} \widetilde{\theta}^j = 0$. Here the main fact used is $\tilde{\mathcal{L}}_{\mathbf{A}} \mathbf{1}_n = 0$. Additionally we have that for any $j \neq k$, we have that:

$$\left\|\theta^j - \theta^k\right\|_\infty^2 \geq \frac{1}{n} \left\|\theta^j - \theta^k\right\|_2^2 = \frac{1}{n} \frac{\delta^2}{n'} \left(\widetilde{\theta}^j - \widetilde{\theta}^k\right)^\top \Lambda^\dagger \left(\widetilde{\theta}^j - \widetilde{\theta}^k\right) \geq \frac{1}{n} \frac{\delta^2}{n'} \sum_{i=\lceil (1-\alpha)n' \rceil}^{n'} \frac{1}{\lambda_i} \tag{48}$$

Now, setting $\delta^2 = 0.01 \frac{\sigma^2 n'}{N_{\text{comp}} \zeta}$ results in:

$$\left\|\theta^j\right\|_\infty \leq \frac{\delta}{\sqrt{n'}} \left\| \sqrt{\Lambda^\dagger} \tilde{\theta}^j \right\|_2 \overset{(i)}{\leq} \frac{\delta}{\sqrt{n'}} \sqrt{\operatorname{tr}\left(\Lambda^\dagger\right)} \overset{(ii)}{=} \frac{\delta}{\sqrt{n'}} \sqrt{\operatorname{tr}\left(L^\dagger\right)} \overset{(iii)}{\leq} B \tag{49}$$

where inequality (i) follows from the fact that z^j has entries in $\{0,1\}$; step (ii) follows because the matrices $\sqrt{\Lambda^\dagger}$ and $\sqrt{\tilde{\mathcal{L}}_{\mathbf{A}}^\dagger}$ have the same eigenvalues; and inequality (iii) follows from our choice of δ and our assumption $N_{\text{comp}} \geq \frac{c\sigma^2 \operatorname{tr}\left(\tilde{\mathcal{L}}_{\mathbf{A}}^\dagger\right)}{\zeta B^2}$ on the sample size with $c = 0.01$. We have thus verified that each vector θ^j also satisfies the boundedness constraint $\left\|\theta^j\right\|_\infty \leq B$, as required for membership in $\mathcal{W}_B$. Furthermore, for any pair of distinct vectors in this set, we have:

$$\left\|\theta^j - \theta^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 = \frac{\delta^2}{n'} \left\|z^j - z^k\right\|_2^2 \leq \delta^2$$

By Lemma 21 we have that:

$$D_{\text{KL}}\left(\mathbf{P}_{\theta^j} \| \mathbf{P}_{\theta^k}\right) \leq \frac{N_{\text{comp}} \zeta}{\sigma^2} \left\|\theta^j - \theta^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 = 0.01 n' \tag{50}$$

Finally we have by substituting (δ', β') into the pairwise Fano's lower bound (Lemma 18) that:

$$\sup_{\theta^* \in \mathcal{W}_B} \mathbb{E}\left[\left\| \widetilde{\theta} - \theta^* \right\|_\infty \right] \geq \frac{\frac{\delta}{\sqrt{n'd}} \sqrt{\sum_{i=\lceil (1-\alpha)n' \rceil}^{n'} \frac{1}{\lambda_i}}}{2} \left(1 - \frac{0.01 n' + \log 2}{\log M'(\alpha)} \right)$$

Substituting our choice of δ and setting $\alpha = 0.01$ proves the claim for $n' > 9$.

For the case of $n' \leq 9$. Consider the packing set of the three n'-length vectors $\theta^1 = \delta U \sqrt{\Lambda^\dagger}\,(0, 1, \dots, 0)$, $\theta^2 = -\theta^1$ and $\theta^3 = (0, \dots, 0)$. Then we have for all pairs $\min_{j \neq k} \left\|\theta^j - \theta^k\right\|_\infty^2 \geq \frac{1}{9} \min_{j \neq k} \left\|\theta^j - \theta^k\right\|_2^2 \geq \frac{\delta^2}{9\lambda_2(L)}$ and $\max_{j,k} \left\|\theta^j - \theta^k\right\|_{\tilde{\mathcal{L}}_{\mathbf{A}}}^2 \leq 4\delta^2$, and as a result $\max_{j,k} D_{\text{KL}}\left(\mathbf{P}_{\theta^j} \| \mathbf{P}_{\theta^k}\right) \leq \frac{4 N_{\text{comp}} \zeta \delta^2}{\sigma^2}$. Choosing $\delta^2 = \frac{\sigma^2 \log 2}{8 N_{\text{comp}} \zeta}$ and applying the pairwise Fano's lower bound (Lemma 18) yields the claim.

A.4 ADDITIONAL EXPERIMENTS

In this section, we show some additional results of experiments. A one-sentence summary of this section is, we demonstrate our discussions in Section 2.2 and Section 2.1 by real cases that our upper bound is tighter than those of Yan et al. (2012) and Shah et al. (2016) for general comparison graphs and maximal performance gap κ, except for some extremely sparse and sense comparison graphs and small κ.

There are two things we want to note for numerical experiments. First, since all upper bounds are high probability bounds in an asymptotic sense, i.e., they contain some constant factors that are ignorable only for large n, so the exact relative tightness of those bounds is hard to show numerically, because in experiments we can only check for certain moderate-magnitude values of n.

Another thing that should be noted is the factor $e^{c\kappa}$ or $e^{c\kappa_E}$. Again since we are working with a moderate n, this exponential term could be dominating even for a small κ, and bounds with different leading constants c in the exponents could show huge difference in magnitude. Thus the tightness of a method is significantly affected by this exponential factor. Because of this, our upper bound is definitely tighter than the ℓ_2 bound in Shah et al. (2016) for general κ. Moreover, as we show in Example 7, for some special design, κ_E can be much smaller than κ and makes our bound tighter than those with e^{κ}. But for constant κ and sufficiently large n this exponential factor is ignorable, and the ℓ_2 upper bound in Shah et al. (2016), acting as a trivial ℓ_∞ upper bound will be tighter than ours under some sparse or irregular comparison graphs, as we discussed in Section 2.2. We will see the same results of comparison in experiments below.

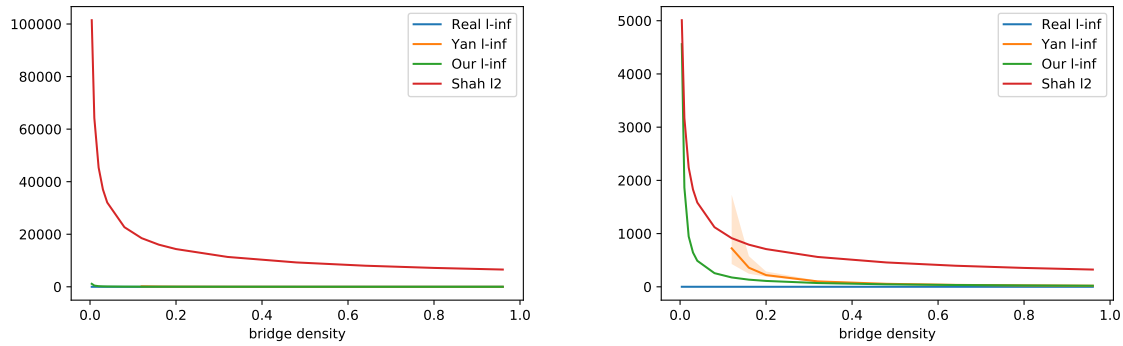

Figure 3: Comparison under the Barbell graph with random bridge edges. Shah-l2 shows the ℓ_2 high probability upper bound in Shah et al. (2016). In the left panel Shah-l2 uses the factor $e^{4\kappa}$, while in the right panel it uses factor $e^{2\kappa}$. Some points near the left end of the orange line are cut because $\min_{i,j} n_{ij} = 0$ and the corresponding upper bound is ∞. Each point on lines is an average of 20 trials. Upper and lower 0.05 quantiles are presented by colored areas.

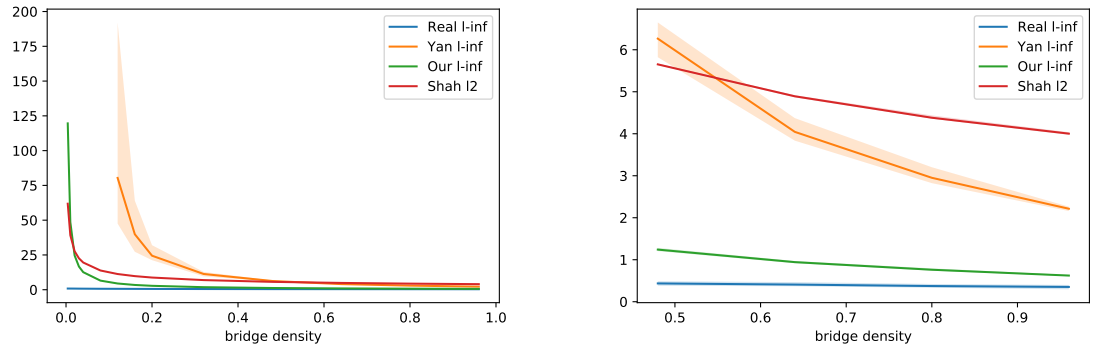

Figure 4: Comparison under the Barbell graph with random bridge edges, with $e^{c\kappa}$ replaced by some constant factor.

Additional results for Example 8. Fig. 3 shows a comparison under the Barbell graph with random bridges of the real ℓ_∞ error of the MLE and the 3 upper bounds. Here $n_1 = n_2 = 50$ and $n = 100$. The left panel shows the dominance of $e^{4\kappa}$ in

Shah's ℓ_2 bound. Since in our equal-spaced $\boldsymbol{\theta}^*$ setting, $\kappa = 2B$, so we tighten Shah's ℓ_2 bound for a "fair" comparison in the right panel. It doesn't change the relative order for most bridge density $|E_l|/(n_1 n_2)$, but suggests that Shah l2 will be lower than our l-inf for very small bridge density.

To illustrate the relative order of bounds for (very) large n and small κ, we replace the exponential factors by some common appropriate constant and compare the 3 bounds in Figure 4. It confirms our assertion at the beginning of this section.

Additional results for Example 7. Figure 5 shows a comparison under the island graph of the real ℓ_∞ error of the MLE, and the upper bounds of Shah and ours. The bound for Yan et al. (2012) is not shown because $\min_{i,j} n_{ij} = 0$. The results again confirm our assertion.

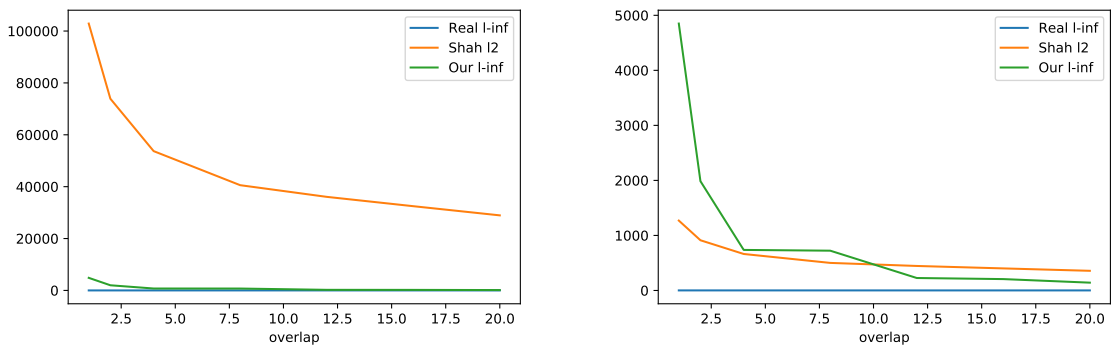

Figure 5: Comparison under the island graph. Shah-l2 shows the ℓ_2 high probability upper bound in Shah et al. (2016). In the left panel Shah-l2 uses the factor $e^{4\kappa}$, while in the right panel it uses factor $e^{2\kappa}$. Each point on lines is an average of 20 trials. Upper and lower 0.05 quantiles are presented by colored areas.

A.4.1 Cases where κ_E is loose

Consider a path graph of n nodes and edge set $\{(i, i+1) : i \in [n-1]\}$. Assume $\theta_i^* = \theta_1^* + (i-1)\delta$, then $\kappa_E = \delta$ and $\kappa = (n-1)\delta$. In this case, a factor of e^{κ_E} gives tighter control than e^{κ}. However, if we add one edge $(1, n)$ into the graph, κ_E becomes $(n-1)\delta$ and our upper bound will increase a lot, which is counter-intuitive because the newly-added 1 out of n edges should not affect the estimation accuracy too much. In other words, the bound gets looser after the new edge is added.

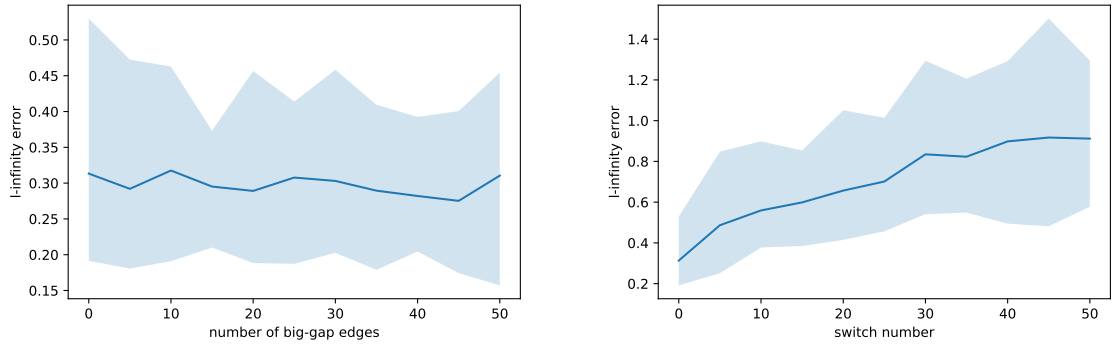

Figure 6: Experiment on the path graph. Left: the ℓ_∞ error of the regularized MLE when adding new edges with big performance gaps to the path graph. Right: the ℓ_∞ error of the regularized MLE when switch some pairs of the performance parameter $\{\theta_i^*\}$ so that there are more big-gap edges while keeping the algebraic connectivity. The results show that when the proportion of big-gap edges is small, the estimation error would not be affected a lot. Both experiments are under $\kappa \approx 6.9$, equal-gap θ^*, $n = 200$ and $L = 5000$ and based on 40 trials for each hyperparameter, with 0.05 and 0.95 quantiles shown by the shaded area.

The reason is that κ_E itself is not enough to provide tight control across all comparison graphs. For instance, to avoid such

loose cases, one also needs to take into account the proportion of edges with big performance gaps compared to edges with small performance gaps. Figure 6 shows some numerical results on the impact of big-gap edges. In the left panel, we add edges to the top-right and bottom-left corner of the adjacency matrix so that their performance gaps are big. In the right panel, we switch some pairs of parameters $\{\theta_i^*\}$ so that there are more big-gap edges while keeping the algebraic connectivity. The first switch will switch $\theta_1^*, \theta_{[n/2]+1}^*$. The i-th switch will switch the pair $\theta_{2i-1}^*, \theta_{[n/2]+2i-1}^*$. As an example, taking $n = 8$, $\theta_1^* = 0$, and $\delta = 1$, 2 switches will make parameters change as

$$(1, 2, 3, 4, 5, 6, 7, 8) \longrightarrow (5, 2, 7, 4, 1, 6, 3, 8). \tag{51}$$

In the left panel, as new big-gap edges come in, the algebraic connectivity of the graph gets larger as well, keeping estimation errors in the same level. In the right panel, as we keep the algebraic connectivity constant, the impact of the proportion of big-gap edges is shown more clearly.

Another thing we need to point out is that for cases like path graphs, even if we ignore the factor of κ, bounding estimation error itself is hard due to the poor connectivity of the graph, as is argued in Shah et al. (2016) (their upper bound for both path and cycle graphs is not optimal). As we have shown in Section 2.2, even for d-regular graphs with relatively small d, the algebraic connectivity is not big enough and our upper bound cannot match the lower bound. But as the comparison graph gets denser and more regular, κ_E will get closer to κ, making the difference between e^{κ_E} and e^{κ} not as dramatic, although in finite sample phase the difference can still be big because it is an exponential factor.

The last thing we want is that the estimation itself (without providing a theoretical tight upper bound for the estimation error) under the path graph is hard: one need a huge L to make accurate estimation when n is big, as is shown in Figure 7.

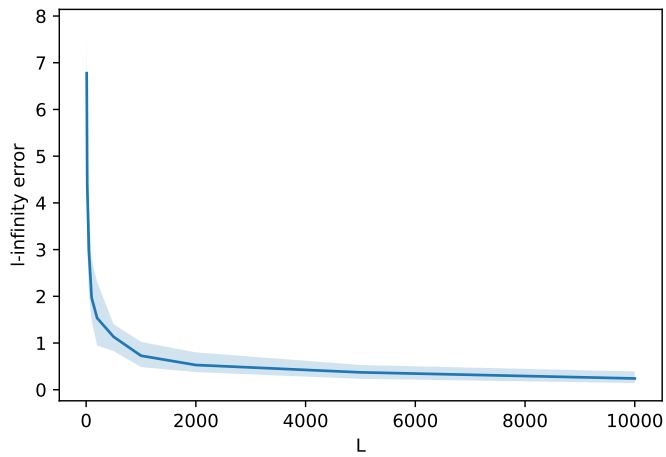

Figure 7: Experiment on the path graph, under $\kappa \approx 6.9$, equal-gap θ^*, $n = 200$ and based on 40 trials for each hyperparameter L, with 0.05 and 0.95 quantiles shown by the shaded area.

A.5 OTHER SUPPORTING RESULTS

Lemma 23. *Given a d-Cayley graph, where $(i,j) \in E$ if and only if $i - j \equiv k (\mathrm{mod}\ n)$ with $-d \le k \le d$, $k \neq 0$. It satisfies $\lambda_2(\mathcal{L}_{\mathbf{A}}) \asymp d^3/n^2$.*

Proof. By definition, a d-Cayley graph is a $2d$-regular graph, and it's well known that (see, e.g. Brouwer and Haemers, 2012) the spectra of its adjacency matrix $\mathbf{A}$ is given by

$$\lambda_j(\mathbf{A}) = \sum_{k=1}^{d} (\zeta_j^k + \zeta_{-j}^k), \quad \zeta_j := \cos\frac{2\pi j}{n} + \sqrt{-1}\sin\frac{2\pi j}{n},$$

for $j = 1, \cdots, n$. Thus, $\lambda_2(\mathcal{L}_{\mathbf{A}}) = 2d - 2\sum_{k=1}^{d}\cos(\frac{2\pi k}{n})$. Since $\cos(\frac{2\pi k}{n}) = 1 - 2\sin^2\frac{\pi}{n}k$ and for $k \le d < 0.5n$, $\sin\frac{\pi}{n}k \in (0.5\frac{\pi}{n}k, \frac{\pi}{n}k)$, we have

$$\lambda_2(\mathcal{L}_{\mathbf{A}}) = c' 4\pi^2/n^2 \sum_{k=1}^{d} k^2 = cd^3/n^2,$$

for $c \in (2\pi^2/3, 4\pi^2/3)$. □

Lemma 4. (Subadditivity) Let I_1, I_2, I_3 be three subsets of $[n]$ such that $\cup_{j=1}^3 I_j = [n]$ and, for each $j \neq k$, $I_j \not\subseteq I_k$ and for $i = 1, 2$, $I_i \cap I_3 \neq \emptyset$. For any vector $\boldsymbol{\theta} \in \mathbb{R}^n$, let $\boldsymbol{\theta}_{(j)} \in \mathbb{R}^{|I_j|}$ be the sub-vector of $\boldsymbol{\theta}$ consisting of the entries indexed by I_j, $j = 1, 2, 3$. Let $\boldsymbol{\theta}^*$ be the vector of preference scores in the BTL model over n items and $\hat{\boldsymbol{\theta}}_{(j)}$ be the MLE of $\boldsymbol{\theta}^*_{(j)}$ for the BTL model involving only items in I_j, $j = 1, 2, 3$, assuming that the sub-graphs induced by the I_j's are connected and allowing for a different number of comparisons in each sub-graph. Finally, let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^n$ be the ensemble MLE based on $\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}, \hat{\boldsymbol{\theta}}_{(3)}$. Then

$$\frac{1}{4} d_\infty(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \le d_\infty(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}^*_{(1)}) + d_\infty(\hat{\boldsymbol{\theta}}_{(2)}, \boldsymbol{\theta}^*_{(2)}) + d_\infty(\hat{\boldsymbol{\theta}}_{(3)}, \boldsymbol{\theta}^*_{(3)}), \tag{52}$$

where $d_\infty(\mathbf{v}_1, \mathbf{v}_2) := \|(\mathbf{v}_1 - \mathbf{1}^\top \mathrm{avg}(\mathbf{v}_1)) - (\mathbf{v}_2 - \mathbf{1}^\top \mathrm{avg}(\mathbf{v}_2))\|_\infty$, where $\mathrm{avg}(\mathbf{x}) := \frac{1}{n}\mathbf{1}_n^\top \mathbf{x}$.

Proof of Lemma 6. First, for each $i = 1, 2, 3$, let's shift $\hat{\boldsymbol{\theta}}_{(i)}$, $\boldsymbol{\theta}^*_{(i)}$ by some constant vector $a\mathbf{1}_{|I_i|}$ so that

$$\mathrm{avg}(\hat{\boldsymbol{\theta}}_{(i)}) = \frac{1}{|I_i|}\mathbf{1}_{|I_i|}^\top \hat{\boldsymbol{\theta}}_{(i)} = 0, \quad \mathrm{avg}(\boldsymbol{\theta}^*_{(i)}) = 0,$$

and hence

$$d_\infty(\hat{\boldsymbol{\theta}}_{(i)}, \boldsymbol{\theta}^*_{(i)}) = \|\hat{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta}^*_{(i)}\|_\infty. \tag{53}$$

Next, for each $i = 1, 2, 3$, let $\tilde{\boldsymbol{\theta}}_{(i)} \in \mathbb{R}^n$ be the augmented version of $\hat{\boldsymbol{\theta}}_{(i)} \in \mathbb{R}^{|I_i|}$, given by

$$\tilde{\boldsymbol{\theta}}_{(i)}(j) = \begin{cases} \hat{\boldsymbol{\theta}}_{(i)}(j), & j \in I_i, \\ 0, & j \notin I_i, \end{cases}$$

where $\mathbf{v}(j)$ refers to the j-th entry of vector $\mathbf{v}$. Similarly, we define

$$\tilde{\boldsymbol{\theta}}^*_{(i)}(j) = \begin{cases} \boldsymbol{\theta}^*_{(i)}(j), & j \in I_i, \\ 0, & j \notin I_i, \end{cases}$$

Now let's define the ensemble MLE $\hat{\boldsymbol{\theta}}$ and show the subadditivity. The idea is to first fix $\tilde{\boldsymbol{\theta}}_{(1)}$, and then shift nonzero entries in $\tilde{\boldsymbol{\theta}}_{(2)}$ and $\tilde{\boldsymbol{\theta}}_{(3)}$ to comply with the difference in common entries of I_1, I_3 and I_2, I_3.

Let $S_1 = I_1$, $S_2 = I_2 \setminus I_1$ (note that we don't put any constraint on $I_1 \cap I_2$), $S_3 = I_3 \setminus (I_1 \cup I_2)$. We allow S_3 to be $\emptyset$, but by the assumption that $I_j \not\subseteq I_k$, we have $S_2 \neq \emptyset$. Since $\cup_{j=1}^3 I_j = [n]$ and $I_i \cap I_3 \neq \emptyset$ for $i = 1, 2$, we have $\cup_{j=1}^3 S_j = [n]$

and $S_i \cap S_k = \emptyset$ for any $i \neq k$. Pick $t_1 \in I_1 \cap I_3$, $t_2 \in I_2 \cap I_3$, and let $\delta_3 = \tilde{\boldsymbol{\theta}}_{(1)}(t_1) - \tilde{\boldsymbol{\theta}}_{(3)}(t_1)$, $\delta_2 = \tilde{\boldsymbol{\theta}}_{(3)}(t_2) - \tilde{\boldsymbol{\theta}}_{(2)}(t_2)$. Moreover, define $\check{\boldsymbol{\theta}}_{(3)}$, $\check{\boldsymbol{\theta}}_{(2)}$ by

$$\check{\boldsymbol{\theta}}_{(3)}(j) = \begin{cases} \tilde{\boldsymbol{\theta}}_{(3)}(j) + \delta_3, & j \in S_3, \\ 0, & j \notin S_3, \end{cases}, \quad \check{\boldsymbol{\theta}}_{(2)}(j) = \begin{cases} \tilde{\boldsymbol{\theta}}_{(2)}(j) + \delta_3 + \delta_2, & j \in S_2, \\ 0, & j \notin S_2, \end{cases}.$$

Let

$$\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_{(1)} + \check{\boldsymbol{\theta}}_{(2)} + \check{\boldsymbol{\theta}}_{(3)}, \tag{54}$$

then it can be seen that

$$\hat{\boldsymbol{\theta}}(j) = \begin{cases} \tilde{\boldsymbol{\theta}}_{(1)}(j), & j \in S_1, \\ \tilde{\boldsymbol{\theta}}_{(2)}(j) + \delta_3 + \delta_2, & j \in S_2, \\ \tilde{\boldsymbol{\theta}}_{(3)}(j) + \delta_3, & j \in S_3, \end{cases}.$$

To analyze the error, we first notice that for any $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$ and $a \in \mathbb{R}$,

$$d_\infty(\mathbf{u}, \mathbf{v}) = d_\infty(\mathbf{u}, \mathbf{v} + a\mathbf{1}_n) \leq \|\mathbf{u} - (\mathbf{v} + a\mathbf{1}_n)\|_\infty + \frac{1}{n}\sum_{i=1}^n |u_i - (v_i + a)| \leq 2\|\mathbf{u} - (\mathbf{v} + a\mathbf{1}_n)\|_\infty.$$

Let $\delta_3^* = \tilde{\boldsymbol{\theta}}^*_{(1)}(t_1) - \tilde{\boldsymbol{\theta}}^*_{(3)}(t_1)$, $\delta_2^* = \tilde{\boldsymbol{\theta}}^*_{(3)}(t_2) - \tilde{\boldsymbol{\theta}}^*_{(2)}(t_2)$. Define a new true parameter $\check{\boldsymbol{\theta}}^*$ by shifting $\boldsymbol{\theta}^*$ via

$$\check{\boldsymbol{\theta}}^*(j) = \begin{cases} \tilde{\boldsymbol{\theta}}^*_{(1)}(j), & j \in S_1, \\ \tilde{\boldsymbol{\theta}}^*_{(2)}(j) + \delta_3^* + \delta_2^*, & j \in S_2, \\ \tilde{\boldsymbol{\theta}}^*_{(3)}(j) + \delta_3^*, & j \in S_3, \end{cases}.$$

It can be verified that $\check{\boldsymbol{\theta}}^* = \boldsymbol{\theta}^* + (\check{\boldsymbol{\theta}}^*(t_1) - \tilde{\boldsymbol{\theta}}^*(t_1))\mathbf{1}_n$. Therefore,

$$d_\infty(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \leq 2\|\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}^*\|_\infty.$$

For $j \in S_1$,

$$|\hat{\boldsymbol{\theta}}(j) - \check{\boldsymbol{\theta}}^*(j)| = |\tilde{\boldsymbol{\theta}}_{(1)}(j) - \tilde{\boldsymbol{\theta}}^*_{(1)}(j)| \leq \|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}\|_\infty.$$

For $j \in S_3$,

$$|\hat{\boldsymbol{\theta}}(j) - \check{\boldsymbol{\theta}}^*(j)| = |\tilde{\boldsymbol{\theta}}_{(3)}(j) - \tilde{\boldsymbol{\theta}}^*_{(3)}(j)| + |\delta_3 - \delta_3^*| \leq 2\|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}\|_\infty + \|\hat{\boldsymbol{\theta}}_{(3)} - \boldsymbol{\theta}^*_{(3)}\|_\infty.$$

For $j \in S_2$.

$$\begin{aligned} |\hat{\boldsymbol{\theta}}(j) - \check{\boldsymbol{\theta}}^*(j)| &= |\tilde{\boldsymbol{\theta}}_{(2)}(j) - \tilde{\boldsymbol{\theta}}^*_{(2)}(j)| + |\delta_3 - \delta_3^*| + |\delta_2 - \delta_2^*| \\ &\leq \|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}\|_\infty + 2\|\hat{\boldsymbol{\theta}}_{(2)} - \boldsymbol{\theta}^*_{(2)}\|_\infty + 2\|\hat{\boldsymbol{\theta}}_{(3)} - \boldsymbol{\theta}^*_{(3)}\|_\infty. \end{aligned}$$

Therefore by definition of $\|\cdot\|_\infty$ and Eq. (53),

$$\begin{aligned} \|\hat{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}}^*\|_\infty &\leq 2(\|\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}^*_{(1)}\|_\infty + \|\hat{\boldsymbol{\theta}}_{(2)} - \boldsymbol{\theta}^*_{(2)}\|_\infty + \|\hat{\boldsymbol{\theta}}_{(3)} - \boldsymbol{\theta}^*_{(3)}\|_\infty) \\ &= 2(d_\infty(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}^*_{(1)}) + d_\infty(\hat{\boldsymbol{\theta}}_{(2)}, \boldsymbol{\theta}^*_{(2)}) + d_\infty(\hat{\boldsymbol{\theta}}_{(3)}, \boldsymbol{\theta}^*_{(3)})), \end{aligned}$$

and we have

$$\frac{1}{4} d_\infty(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \leq d_\infty(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}^*_{(1)}) + d_\infty(\hat{\boldsymbol{\theta}}_{(2)}, \boldsymbol{\theta}^*_{(2)}) + d_\infty(\hat{\boldsymbol{\theta}}_{(3)}, \boldsymbol{\theta}^*_{(3)}).$$

□

A.6 SPECIAL CASES OF PAIRWISE GRAPH TOPOLOGIES

By Theorem 1, for the estimator $\hat{\boldsymbol{\theta}}_\rho$ to be consistent, L needs to be sufficiently large. We can check some common types of comparison graph topologies and see in what order the necessary sample complexity $N_{\text{comp}} = |E|L$ needs to be, to achieve consistency. To simplify results, we assume $e^{2\kappa_E} \lesssim \log n$. Spectral properties of graphs listed here can be found in well-known textbooks (Brouwer and Haemers, 2012). Shah et al. (2016) provide analogous comparisons. Per Section 2.2 we include results for the d-Caley graphs, and expander graphs here. For reader convenience other results from Section 2.2 are also noted below.

Complete graph: In this case, $\lambda_2(\mathcal{L}_{\mathbf{A}}) = n_{\max} = n_{\min} = n - 1$. Thus, we need $e^{2\kappa_E} \log n/n = o(L)$. Hence $L = \Omega(1)$ and $N_{\text{comp}} = \Omega(n^2)$.

Expander graph: If the comparison graph is a d-regular expander graph with edge expansion (or Cheeger number) coefficient ϕ, then $\lambda_2(\mathcal{L}_{\mathbf{A}}) \geq \phi^2/(2d)$ (Alon et al., 2008), $n_{\max} = n_{\min} = d$. Here ϕ is defined by $\phi := \min_{|S|} e(S,T)/|S|$ where $\{S,T\}$ is a partition of the vertex set and $|S| \leq |T|$. We need $d^2 e^{2\kappa_E}/\phi^4 \cdot (e^{2\kappa_E} n \vee d \log n) = o(L)$, so $N_{\text{comp}} = \Omega(nd^3 e^{2\kappa_E}/\phi^4 \cdot (e^{2\kappa_E} n \vee d \log n))$.

Complete bipartite graph: If the comparison graph has two partitioned sets of size m_1 and m_2 such that $m_1 \leq m_2$, then $\lambda_2(\mathcal{L}_{\mathbf{A}}) = m_1$, $n_{\max} = m_2$, $n_{\min} = m_1$. We need $e^{2\kappa_E} m_2/m_1^2 \cdot [(e^{2\kappa_E} n m_2/m_1^2) \vee \log n] = o(L)$. When $m_1 = \Omega(n)$, we have $N_{\text{comp}} = \Omega(n^2)$.

d-Cayley graph: $(i,j) \in E$ if and only if $i - j \equiv k (\text{mod } n)$ with $-d \leq k \leq d$, $k \neq 0$. It is a $2d$-regular graph, and $\lambda_2(\mathcal{L}_{\mathbf{A}}) \asymp d^3/n^2$ (see Appendix A.5 in the supplement), $n_{\max} = n_{\min} = 2d$. Thus, we need $e^{2\kappa_E} n^4/d^6 \cdot (e^{2\kappa_E} n \vee 2d \log n) = o(L)$, so $N_{\text{comp}} = \Omega(e^{2\kappa_E} n^5/d^5 \cdot (e^{2\kappa_E} n \vee 2d \log n))$ for $d = o(n)$ and $N_{\text{comp}} = \Omega(n^2)$ for $d = \Omega(n)$. When $d = 1$, the d-Cayley graph is a cycle graph.

Path or Cycle graph: When comparisons occur based on a path or cycle comparison graph, then $\lambda_2(\mathcal{L}_{\mathbf{A}}) = 2 - 2\cos(c\pi/n) \approx c^2\pi^2/n^2$, $n_{\max} = 2$, $n_{\min} = c$ with $c = 1$ for path and $c = 2$ for cycle. Thus, we need $e^{4\kappa_E} n^5 = o(L)$, so $N_{\text{comp}} = \Omega(e^{4\kappa_E} n^6)$.

Star graph: For a start graph on n node one node has degree $n - 1$ and the remaining have degree 1. In this case, $\lambda_2(\mathcal{L}_{\mathbf{A}}) = n_{\min} = 1$, $n_{\max} = n$. We need $e^{4\kappa_E} n^3 = o(L)$, hence $N_{\text{comp}} = e^{4\kappa_E} n^4$.

Barbell graph: It contains two size-$n/2$ complete sub-graphs connected by 1 edge, so $\lambda_2(\mathcal{L}_{\mathbf{A}}) \asymp 1/n$, $n_{\max} = n/2$, $n_{\min} = n/2 - 1$. We need $e^{2\kappa_E} n^3 \log n = o(L)$, and $N_{\text{comp}} = e^{2\kappa_E} n^5 \log n$.

A.7 UPPER BOUND FOR UNREGULARIZED/VANILLA MLE

A.7.1 Main theorem

The unregularized or vanilla MLE is defined as

$$\hat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}\in\mathbb{R}^n\,:\,\mathbf{1}_n^\top\boldsymbol{\theta}=0}{\arg\min}\ \ell(\boldsymbol{\theta};\mathbf{y}), \tag{55}$$

where $\ell(\boldsymbol{\theta};\mathbf{y})$ is the negative log-likelihood, given by

$$\ell(\boldsymbol{\theta};\mathbf{y}) := -\sum_{1\le i<j\le n} A_{ij}\left\{\bar{y}_{ij}\log\psi(\theta_i-\theta_j)+(1-\bar{y}_{ij})\log[1-\psi(\theta_i-\theta_j)]\right\}, \tag{56}$$

and $t\in\mathbb{R}\mapsto\psi(t)=1/[1+e^{-t}]$ the sigmoid function. To make the expressions of results simpler, we consider the parameter range $\kappa\le n$ and $\kappa_E\le\log n$ as discussed in the comments after Theorem 1.

Theorem 24 (Vanilla MLE)**.** *Assume the BT model with parameter $\boldsymbol{\theta}^*=(\theta_1^*,\dots,\theta_n^*)^\top$ such that $\mathbf{1}_n^\top\boldsymbol{\theta}^*=0$ and a comparison graph $\mathcal{G}=\mathcal{G}([n],E)$ with adjacency matrix $\mathbf{A}$, algebraic connectivity $\lambda_2(\mathcal{L}_{\mathbf{A}})$ and maximum and minimum degrees $n_{\max}$ and $n_{\min}$, respectively. Suppose that each pair of items $(i,j)\in E$ are compared L times. Let $\kappa=\max_{i,j}|\theta_i^*-\theta_j^*|$ and $\kappa_E=\max_{(i,j)\in E}|\theta_i^*-\theta_j^*|$. Assume that $\mathcal{G}$ is connected, or equivalently, $\lambda_2(\mathcal{L}_A)>0$. In addition, assume that 1. $\lambda_2(\mathcal{L}_A)^2L>Ce^{2\kappa_E}\max\{n_{\max}\log n, e^{2\kappa_E}nn_{\max}^2/n_{\min}^2\}$ for some large constant $C>0$, and 2. $\lambda_2(\mathcal{L}_{\mathbf{A}})\ge 2e^{2\kappa_E}n_{\max}/n_{\min}$. Then, with probability at least $1-O(n^{-5})$, the unregularized MLE $\hat{\boldsymbol{\theta}}$ from (55) satisfies*

$$\|\hat{\theta}-\theta^*\|_\infty \lesssim e^{\kappa_E}\sqrt{\frac{n_{\max}\log n}{Ln_{\min}^2}}+e^{2\kappa_E}\sqrt{\frac{n_{\max}^3}{L\lambda_2(\mathcal{L}_A)^2n_{\min}^2}}\left[1+\frac{e^{\kappa_E}\sqrt{\log n}}{n_{\min}}+\frac{e^{2\kappa_E}\sqrt{n_{\max}n}}{\lambda_2(\mathcal{L}_A)n_{\min}}\right], \tag{57}$$

and

$$\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2 \lesssim \frac{e^{\kappa_E}}{\lambda_2(\mathcal{L}_{\mathbf{A}})}\sqrt{\frac{n_{\max}n}{L}}, \tag{58}$$

provided that $L\lesssim n^5$, $\kappa<n$, $\kappa_E\le\log n$, and the right hand side of Equation (57) is smaller than a sufficiently small constant $C>0$.

Remark 4. The expression of the ℓ_∞ bound looks messy, but in the $ER(n,p)$ case it reduces to the same form as the upper bound of the regularized MLE in Corollary 2. Moreover, for vanilla MLE we require an additional pure topological assumption $\lambda_2(\mathcal{L}_{\mathbf{A}})\ge 2e^{2\kappa_E}n_{\max}/n_{\min}$, which is stringent in some sense as it exclude many graphs with small $n_{\min}$. But for such graphs with high degree heterogeneity $n_{\max}/n_{\min}$, it's reasonable that we need some regularity in our objective function. We believe that this condition can be weakened, and the ℓ_∞ upper bound can be improved or tightened by improving the proof techniques, which can be a good future direction for researchers.

To prove Theorem 24, we need two lemmas, Lemma 25 anbd 26.

Lemma 25. *Under the setting of Theorem 24, it holds with probability at least $1-O(n^{-5})$ that*

$$\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_\infty\le 5. \tag{59}$$

Following Chen et al. (2020), we decompose the full negative loglikelihood function as

$$\ell_n(\boldsymbol{\theta})=\ell_n^{(-m)}(\boldsymbol{\theta}_{-m})+\ell_n^{(m)}(\boldsymbol{\theta}_{-m}), \tag{60}$$

where $\theta_m\in\mathbb{R}$ is the m-th entry of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{-m}\in\mathbb{R}^{n-1}$ is the subvector containing the rest of entries, and the two functions are given by

$$\ell_n^{(-m)}(\boldsymbol{\theta}_{-m})=\sum_{1\le i<j\le n:i,j\ne m}A_{ij}\left[\bar{y}_{ij}\log\frac{1}{\psi(\theta_i-\theta_j)}+(1-\bar{y}_{ij})\log\frac{1}{1-\psi(\theta_i-\theta_j)}\right],$$

$$\ell_n^{(m)}(\theta_m|\boldsymbol{\theta}_{-m})=\sum_{j\in[n]\setminus\{m\}}A_{mj}\left[\bar{y}_{mj}\log\frac{1}{\psi(\theta_m-\theta_j)}+(1-\bar{y}_{mj})\log\frac{1}{1-\psi(\theta_m-\theta_j)}\right].$$

Let $H^{(-m)} := \nabla^2 \ell_n^{(-m)}(\ell_{-m})$, and

$$\boldsymbol{\theta}_{-m}^{(m)} = \underset{\boldsymbol{\theta}_{-m}:\|\boldsymbol{\theta}_{-m}-\boldsymbol{\theta}_{-m}^*\|_\infty \le 5}{\operatorname{argmin}} \ell_n^{(-m)}\left(\boldsymbol{\theta}_{-m}\right).$$

Lemma 26. *Under the setting of Theorem 24, it holds with probability at least $1 - O(n^{-9})$ that*

$$\max_{m\in[n]} \|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \boldsymbol{1}_{n-1}\|_2^2 \le C \frac{e^{2\kappa_E} n_{\max} n}{L\lambda_2(\mathcal{L}_A)^2} \tag{61}$$

for some constant $C > 0$, where $a_m = \operatorname{avg}\left(\theta_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^*\right) := \frac{1}{n-1}\mathbf{1}_{n-1}^\top \left(\theta_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^*\right)$.

Proof of Theorem 24. Again we define the leave-one-out negative log-likelihood as

$$\begin{aligned}\ell_n^{(m)}(\theta) = &\sum_{1\le i<j\le n: i,j\ne m} A_{ij}\left[\bar{y}_{ij}\log\frac{1}{\psi\left(\theta_i-\theta_j\right)} + (1-\bar{y}_{ij})\log\frac{1}{1-\psi\left(\theta_i-\theta_j\right)}\right] \\ &+ \sum_{i\in[n]\setminus\{m\}} A_{mj}\left[\psi\left(\theta_i^*-\theta_m^*\right)\log\frac{1}{\psi\left(\theta_i-\theta_m\right)} + \psi\left(\theta_m^*-\theta_i^*\right)\log\frac{1}{\psi\left(\theta_m-\theta_i\right)}\right],\end{aligned} \tag{62}$$

For the ℓ_2 bound, notice that by Taylor expansion

$$\ell_n(\hat{\theta}) = \ell_n(\theta^*) + (\hat{\theta}-\theta^*)^T\nabla\ell_n(\theta^*) + \frac{1}{2}(\hat{\theta}-\theta^*)^T H(\xi)(\hat{\theta}-\theta^*),$$

where ξ is a convex combination of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$. By Lemma 25, we have $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \le 5$ and hence $\|\xi - \boldsymbol{\theta}^*\|_\infty \le 5$. By Lemma 10, we have $\frac{1}{2}(\hat{\theta}-\theta^*)^T H(\xi)(\hat{\theta}-\theta^*) \ge ce^{-\kappa_E}\lambda_2\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$ for some constant $c > 0$. By the fact that $\ell_n(\boldsymbol{\theta}^*) \ge \ell_n(\hat{\boldsymbol{\theta}})$, Cauchy-Schwartz inequality, and Lemma 9, we have

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \lesssim \frac{e^{\kappa_E}}{\lambda_2}\|\nabla\ell(\boldsymbol{\theta}^*)\|_2 \le \frac{e^{\kappa_E}}{\lambda_2}\sqrt{\frac{n_{\max} n}{L}}.$$

For the ℓ_∞ bound, the proof can be sketched as following steps.

0. By Lemma 25, $\|\hat{\theta} - \theta^*\|_\infty \le 5$ with probability at least $1 - O(n^{-5})$.
1. By Lemma 26, it holds with probability exceeding $1 - O(n^{-9})$ that,

$$\max_{m\in[n]} \|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbf{1}_{n-1}\|_2^2 \le C \frac{e^{2\kappa_E} n_{\max} n}{L\lambda_2(\mathcal{L}_A)^2}$$

   where $a_m = \operatorname{avg}\left(\theta_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^*\right) := \frac{1}{n-1}\mathbf{1}_{n-1}^\top \left(\theta_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^*\right)$.
2. Show that on the same event,

$$\begin{aligned}\max_{m\in[n]} \|\theta_{-m}^{(m)} - \widehat{\theta}_{-m} - a_m \mathbf{1}_{n-1}\|_2^2 \le &C_1 \frac{\max_{m\in[n]}\sum_{i\in\mathcal{N}(m)}\left(\bar{y}_{mi} - \psi\left(\theta_m^* - \theta_i^*\right)\right)^2}{\lambda_2(\mathcal{L}_A)^2 e^{-2\kappa_E}} \\ &+ C_1 \frac{e^{2\kappa_E} n_{\max}}{\lambda_2(\mathcal{L}_A)^2}\|\widehat{\theta} - \theta^*\|_\infty^2.\end{aligned} \tag{63}$$

   Following the same arguments towards Equation (75), we can get

$$\|\boldsymbol{\theta}_{-m}^{(m)} - \hat{\boldsymbol{\theta}}_{-m} - \bar{a}_m \mathbf{1}_{n-1}\|_2^2 \le \frac{e^{2\kappa_E}}{c^2\lambda_2(\mathcal{L}_A)^2}\|\nabla\ell_n^{(-m)}(\hat{\boldsymbol{\theta}}_{-m})\|_2^2.$$

   By Equation (60), for each $i \in [n] \setminus \{m\}$, we have

$$\frac{\partial}{\partial\boldsymbol{\theta}_i}\ell_n^{(-m)}\left(\boldsymbol{\theta}_{-m}\right) = \frac{\partial}{\partial\theta_i}\ell_n(\boldsymbol{\theta}) - \frac{\partial}{\partial\theta_i}\ell_n^{(m)}\left(\theta_m \mid \boldsymbol{\theta}_{-m}\right),$$

Using the fact that $\nabla \ell_n(\hat{\boldsymbol{\theta}}) = 0$, we get

$$\frac{\partial}{\partial \theta_i} \ell_n^{(-m)}\left(\boldsymbol{\theta}_{-m}\right)_{|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\frac{\partial}{\partial \theta_i} \ell_n^{(m)}\left(\theta_m \mid \boldsymbol{\theta}_{-m}\right)|_{\theta=\widehat{\boldsymbol{\theta}}} = -A_{mi}\left[\bar{y}_{mi} - \psi(\widehat{\theta}_m - \widehat{\theta}_i)\right].$$

Therefore, we have

$$\begin{aligned}
\|\nabla \ell_n^{(-m)}(\widehat{\boldsymbol{\theta}}_{-m})\|_2^2 =& \sum_{i\in[n]\setminus\{m\}} A_{mi}\left[\bar{y}_{mi} - \psi(\widehat{\theta}_m - \widehat{\theta}_i)\right]^2 \\
\leq& 2\sum_{i\in[n]\setminus\{m\}} A_{mi}\left(\bar{y}_{mi} - \psi\left(\theta_m^* - \theta_i^*\right)\right)^2 \\
&+ 2\sum_{i\in[n]\setminus\{m\}} A_{mi}\left[\psi\left(\theta_m^* - \theta_i^*\right) - \psi(\widehat{\theta}_m - \widehat{\theta}_i)\right]^2 \\
\leq& 2\sum_{i\in[n]\setminus\{m\}} A_{mi}\left(\bar{y}_{mi} - \psi\left(\theta_m^* - \theta_i^*\right)\right)^2 + 2\|\widehat{\theta} - \theta^*\|_\infty^2 \sum_{i\in[n]\setminus\{m\}} A_{mi} \\
\leq& 2\sum_{i\in[n]\setminus\{m\}} A_{mi}\left(\bar{y}_{mi} - \psi\left(\theta_m^* - \theta_i^*\right)\right)^2 + 4n_{\max}\|\widehat{\theta} - \theta^*\|_\infty^2.
\end{aligned}$$

Now use the fact that $\mathbf{1}_n^\top \boldsymbol{\theta}^* = \mathbf{1}_n^\top \hat{\boldsymbol{\theta}} = 0$, we have

$$\|a_m \mathbf{1}_{n-1} - \bar{a}_m \mathbf{1}_{n-1}\|_2^2 = (n-1)[\mathrm{avg}(\widehat{\boldsymbol{\theta}}_{-m} - \boldsymbol{\theta}_{-m}^*)]^2 = \frac{(\hat{\boldsymbol{\theta}}_m - \theta_m^*)^2}{n-1} \leq \frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty^2}{n-1}.$$

These results, together with the fact that $\lambda_2(\mathcal{L}_{\mathbf{A}}) \leq 2n_{\max} \leq 2n$, give Equation (63).

3. Show that on the same event,

$$\begin{aligned}
\|\widehat{\theta} - \theta^*\|_\infty \cdot C_4 e^{-\kappa_E} n_{\min} \leq& \max_{m\in[n]} \Big| \sum_{i\in\mathcal{N}(m)} \left(\bar{y}_{mi} - \psi\left(\theta_m^* - \theta_i^*\right)\right) \Big| \\
&+ \sqrt{n_{\max}} \max_{m\in[n]} \|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbf{1}_{n-1}\|_2 \\
&+ \sqrt{n_{\max}} \max_{m\in[n]} \|\theta_{-m}^{(m)} - \widehat{\theta}_{-m} - a_m \mathbf{1}_{n-1}\|_2.
\end{aligned} \tag{64}$$

First, define two univariate functions as some proxy of gradient and hessian:

$$\begin{aligned}
g^{(m)}\left(\theta_m \mid \boldsymbol{\theta}_{-m}\right) &= \frac{\partial}{\partial \theta_m} \ell_n^{(m)}\left(\theta_m \mid \boldsymbol{\theta}_{-m}\right) = -\sum_{i\in[n]\setminus\{m\}} A_{mi}\left(\bar{y}_{mi} - \psi\left(\theta_m - \theta_i\right)\right) \\
h^{(m)}\left(\theta_m \mid \boldsymbol{\theta}_{-m}\right) &= \frac{\partial^2}{\partial \theta_m^2} \ell_n^{(m)}\left(\theta_m \mid \boldsymbol{\theta}_{-m}\right) = \sum_{i\in[n]\setminus\{m\}} A_{mi}\psi\left(\theta_m - \theta_i\right)\psi\left(\theta_i - \theta_m\right).
\end{aligned}$$

By the definition of $\hat{\boldsymbol{\theta}}$ and the shift invariance of ℓ_n, we have $\ell_n(\hat{\boldsymbol{\theta}}) \leq \ell_n(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \mathbb{R}^n$, thus

$$\ell_n^{(m)}(\theta_m^* \mid \hat{\boldsymbol{\theta}}_{-m}) + \ell_n^{(-m)}(\hat{\boldsymbol{\theta}}_{-m}) \geq \ell_n(\hat{\boldsymbol{\theta}}).$$

This implies

$$\begin{aligned}
\ell_n^{(m)}(\theta_m^* \mid \hat{\theta}_{-m}) &\geq \ell_n^{(m)}(\hat{\theta}_m \mid \hat{\theta}_{-m}) \\
&= \ell_n^{(m)}(\theta_m^* \mid \hat{\boldsymbol{\theta}}_{-m}) + (\hat{\theta}_m - \theta_m^*) g^{(m)}(\theta_m^* \mid \hat{\boldsymbol{\theta}}_{-m}) + \frac{1}{2}(\hat{\theta}_m - \theta_m^*)^2 h^{(m)}(\xi \mid \hat{\boldsymbol{\theta}}_{-m}),
\end{aligned}$$

where ξ is a convex combination of θ_m^* and $\hat{\theta}_m$. By Lemma 25, we have $|\xi - \theta_m^*| \leq |\hat{\theta}_m - \theta_m^*| \leq 5$. Thus for any $i \neq m$ it holds that $|\xi - \hat{\theta}_i| \leq |\xi - \theta_m^*| + |\theta_m^* - \theta_i^*| + \left|\widehat{\theta}_i - \theta_i^*\right| \leq 10 + \kappa$. By definition of $h^{(m)}$, we have $\frac{1}{2}h^{(m)}(\xi|\hat{\boldsymbol{\theta}}_{-m}) \geq c_2 e^{-\kappa_E} n_{\min}$ for some constant $c_2 > 0$. Therefore, we get

$$(\hat{\theta}_m - \theta_m^*)^2 \leq \frac{e^{2\kappa_E}}{(c_2 n_{\min})^2} |g^{(m)}(\theta_m^* \mid \hat{\boldsymbol{\theta}}_{-m})|^2. \tag{65}$$

To bound $|g^{(m)}(\theta_m^* \mid \hat{\boldsymbol{\theta}}_{-m})|$, we decompose it as

$$
\begin{aligned}
|g^{(m)}(\theta_m^* \mid \hat{\theta}_{-m})| &= \left| \sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta_m^* - \widehat{\theta}_i)) \right| \\
&\leq \left| \sum_{i\in[n]\setminus\{m\}} A_{mi}(\bar{y}_{mi} - \psi(\theta_m^* - \theta_i^*)) \right| && (66)\\
&\quad + \left| \sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta_m^* - \theta_i^*) - \psi(\theta_m^* - \theta_i^{(m)} + a_m)) \right| && (67)\\
&\quad + \left| \sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta_m^* - \theta_i^{(m)} + a_m) - \psi(\theta_m^* - \widehat{\theta}_i)) \right|. && (68)
\end{aligned}
$$

By Cauchy-Schwartz inequality, we can bound (67) and (68) by

$$
\begin{aligned}
\left| \sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta_m^* - \theta_i^*) - \psi(\theta_m^* - \theta_i^{(m)} + a_m)) \right|^2 &\leq n_{\max} \|\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1}\|_2^2, \\
\left| \sum_{i\in[n]\setminus\{m\}} A_{mi}(\psi(\theta_m^* - \theta_i^{(m)} + a_m) - \psi(\theta_m^* - \widehat{\theta}_i)) \right|^2 &\leq n_{\max} \|\boldsymbol{\theta}_{-m}^{(m)} - \hat{\boldsymbol{\theta}}_{-m} - a_m \mathbf{1}_{n-1}\|_2^2.
\end{aligned}
$$

Plugging these bounds into Equation (65) and taking maximum over $m \in [n]$ give the desired bound (64).

4. Plug (64) back into (63) and get

$$
\begin{aligned}
\max_{m\in[n]} \|\theta_{-m}^{(m)} - \widehat{\theta}_{-m} - a_m \mathbf{1}_{n-1}\|_2^2 \lesssim{}& \frac{e^{2\kappa_E}}{\lambda_2(\mathcal{L}_A)^2} \max_m \sum_{i\in\mathcal{N}(m)} (\bar{y}_{mi} - \psi_{mi}^*)^2 \\
&+ \frac{e^{4\kappa_E}}{\lambda_2(\mathcal{L}_A)^2} \frac{n_{\max}}{n_{\min}^2} \max_{m\in[n]} |\sum_{i\in\mathcal{N}(m)} (\bar{y}_{mi} - \psi_{mi}^*)|^2 \\
&+ \frac{e^{4\kappa_E}}{\lambda_2(\mathcal{L}_A)^2} \frac{n_{\max}^2}{n_{\min}^2} \max_{m\in[n]} \|\theta_{-m}^{(m)} - \theta_{-m}^* - a_m \mathbf{1}_{n-1}\|_2^2 \\
&+ \frac{e^{4\kappa_E}}{\lambda_2(\mathcal{L}_A)^2} \frac{n_{\max}^2}{n_{\min}^2} \max_{m\in[n]} \|\theta_{-m}^{(m)} - \hat{\theta}_{-m} - a_m \mathbf{1}_{n-1}\|_2^2.
\end{aligned}
$$

as we assume $\frac{e^{2\kappa_E}}{\lambda_2(\mathcal{L}_A)} \frac{n_{\max}}{n_{\min}} \leq \frac{1}{2}$, we have

$$
\begin{aligned}
\max_{m\in[n]} \|\theta_{-m}^{(m)} - \widehat{\theta}_{-m} - a_m \mathbf{1}_{n-1}\|_2^2 \lesssim{}& \frac{e^{2\kappa_E}(\log n + n_{\max})}{\lambda_2(\mathcal{L}_A)^2 L} + \frac{e^{4\kappa_E}}{\lambda_2(\mathcal{L}_A)^2} \frac{n_{\max}^2}{n_{\min}^2} \frac{\log n}{L} + \\
& \frac{e^{4\kappa_E}}{\lambda_2(\mathcal{L}_A)^2} \frac{n_{\max}^2}{n_{\min}^2} \frac{e^{2\kappa_E} n_{\max} n}{\lambda_2(\mathcal{L}_A)^2 L}.
\end{aligned} \tag{69}
$$

5. Plug (69) back into (64) and we can get

$$
\begin{aligned}
\|\widehat{\theta} - \theta^*\|_\infty^2 \lesssim{}& \frac{e^{2\kappa_E}}{n_{\min}^2} \frac{n_{\max} \log n}{L} + \frac{e^{2\kappa_E} n_{\max}^2}{n_{\min}^2 n} \frac{e^{2\kappa_E} n_{\max} n}{\lambda_2(\mathcal{L}_A)^2 L} \\
&+ \frac{e^{2\kappa_E} n_{\max}}{n_{\min}^2 \lambda_2(\mathcal{L}_A)^2 L} \left[e^{2\kappa_E}(\log n + n_{\max}) + \frac{e^{4\kappa_E} n_{\max}^2 \log n}{n_{\min}^2} + \frac{e^{6\kappa_E} n_{\max}^3 n}{\lambda_2(\mathcal{L}_A)^2 n_{\min}^2} \right]
\end{aligned} \tag{70}
$$

One term can be reduced and the inequality becomes

$$
\|\hat{\theta} - \theta^*\|_\infty \lesssim e^{\kappa_E} \sqrt{\frac{n_{\max} \log n}{L n_{\min}^2}} + e^{2\kappa_E} \sqrt{\frac{n_{\max}^3}{L \lambda_2(\mathcal{L}_A)^2 n_{\min}^2}} \left[1 + e^{\kappa_E} \sqrt{\frac{\log n}{n_{\min}^2}} + e^{2\kappa_E} \sqrt{\frac{n_{\max} n}{\lambda_2(\mathcal{L}_A)^2 n_{\min}^2}} \right]. \tag{71}
$$

□

A.7.2 Proof of Lemmas

Proof of Lemma 25. We will use a gradient descent sequence defined by

$$\theta^{(t+1)} = \theta^{(t)} - \eta[\nabla \ell_n(\theta^{(t)}) + \rho\theta^{(t)}].$$

The leave-one-out negative log-likelihood is defined as

$$\begin{aligned}\ell_n^{(m)}(\theta) = &\sum_{1\le i<j\le n: i,j\neq m} A_{ij}\left[\bar{y}_{ij}\log\frac{1}{\psi(\theta_i-\theta_j)} + (1-\bar{y}_{ij})\log\frac{1}{1-\psi(\theta_i-\theta_j)}\right]\\ &+ \sum_{i\in[n]\setminus\{m\}} A_{mi}\left[\psi(\theta_i^*-\theta_m^*)\log\frac{1}{\psi(\theta_i-\theta_m)} + \psi(\theta_m^*-\theta_i^*)\log\frac{1}{\psi(\theta_m-\theta_i)}\right],\end{aligned}$$

so the leave-one-out gradient descent sequence is defined as

$$\theta^{(t+1,m)} = \theta^{(t,m)} - \eta[\nabla \ell_n^{(m)}(\theta^{(t,m)}) + \rho\theta^{(t,m)}].$$

We initialize both sequences by $\theta^{(0)} = \theta^{(0,m)} = \theta^*$ and set $\rho = \frac{1}{\kappa}\sqrt{\frac{n_{\max}}{L}}$ and step size $\eta = \frac{1}{\lambda + n_{\max}}$. We will show that under the assumption $\lambda_2(\mathcal{L}_A)^2 L > Ce^{2\kappa_E}\max\{n_{\max}\log n, e^{2\kappa_E} n n_{\max}^2/n_{\min}^2\}$ for some large constant $C > 0$, we have

$$\begin{aligned}\max_{m\in[n]}\|\theta^{(t,m)} - \theta^{(t)}\|_2 \le& f_1 := C_1\frac{e^{\kappa_E}}{\lambda_2}\sqrt{\frac{n_{\max}\log n}{L}} \le 1\\ \|\theta^{(t)} - \theta^*\|_2 \le& f_2 := C_2\frac{e^{\kappa_E}}{\lambda_2}\sqrt{\frac{n_{\max} n}{L}} \le \sqrt{\frac{n}{\log n}}\\ \max_{m\in[n]}|\theta_m^{(t,m)} - \theta_m^*| \le& f_3 := C_3\frac{e^{2\kappa_E}}{\lambda_2}\frac{n_{\max}}{n_{\min}}\sqrt{\frac{n}{L}} \le 1.\end{aligned} \tag{72}$$

A useful fact given that (72) holds is that

$$\|\theta^{(t,m)} - \theta^*\|_\infty \le f_1 + f_2. \tag{73}$$

We again have the Taylor expansion

$$\theta^{(t+1)} - \theta^{(t+1,m)} = [(1-\eta\rho)I_n - \eta H(\xi)](\theta^{(t)} - \theta^{(t,m)}) - \eta[\nabla\ell_n(\theta^{(t,m)}) - \nabla\ell_n^{(m)}(\theta^{(t,m)})].$$

Now by the fact that $\lambda_{\min,\perp}(H(\xi)) \ge c_0 e^{-\kappa}\lambda_2$, we have

$$\|\left((1-\eta\rho)I_n - \eta H(\xi)\right)(\theta^{(t)} - \theta^{(t,m)})\|_2 \le (1-\eta\rho - c_1\eta\lambda_2)\|\theta^{(t)} - \theta^{(t,m)}\|_2$$

for some constant $c_1 > 0$ and the other term can be bounded as

$$\begin{aligned}&\|\nabla\ell_n(\theta^{(t,m)}) - \nabla\ell_n^{(m)}(\theta^{(t,m)})\|_2^2\\ =&\left[\sum_{j\in[n]\setminus\{m\}} A_{jm}\left(\bar{y}_{jm} - \psi\left(\theta_j^* - \theta_m^*\right)\right)\right]^2 + \sum_{j\in[n]\setminus\{m\}} A_{jm}\left(\bar{y}_{jm} - \psi\left(\theta_j^* - \theta_m^*\right)\right)^2\\ \le& C_1\frac{1}{L}n_{\max}\log n + C_1\frac{1}{L}(\log n + n_{\max}).\end{aligned}$$

Therefore, for

$$\|\theta^{(t+1)} - \theta^{(t+1,m)}\|_2 \le (1-c_1\eta\lambda_2)f_1 + \eta\sqrt{2C_1\frac{1}{L}n_{\max}\log n} \le f_1$$

to hold, we need $f_1 > C\frac{e^{\kappa_E}}{\lambda_2}\sqrt{\frac{n_{\max}\log n}{L}}$ for some sufficiently large positive constant $C > 0$.

Next, we bound $\|\theta^{(t+1)} - \theta^*\|$. By Tarlor expansion,

$$\theta^{(t+1)} - \theta^* = \left((1-\eta\lambda)I_n - \eta H(\xi)\right)(\theta^{(t)} - \theta^*) - \eta\lambda\theta^* - \eta\nabla\ell_n(\theta^*).$$

Equation (41) becomes

$$((1-\eta\lambda)I_n - \eta H(\xi))\left(\theta^{(t)} - \theta^*\right) \le (1-\eta\lambda - c_2\eta\lambda_2)\left\|\theta^{(t)} - \theta^*\right\|_2,$$

and equation (42) becomes

$$\left\|\nabla \ell_n\left(\theta^*\right)\right\|_2^2 = \sum_{i=1}^{n}\left[\sum_{j\in[n]\setminus\{i\}} A_{ij}\left(\bar{y}_{ij} - \psi\left(\theta_i^* - \theta_j^*\right)\right)\right]^2 \le C_2\frac{nn_{\max}}{L},$$

for some constants $c_2, C_2 > 0$. Therefore,

$$\|\theta^{(t+1)} - \theta^*\|_2 \le (1-c_2\eta\lambda_2)f_2 + \eta\sqrt{C_2\frac{nn_{\max}}{L}} + \eta\lambda\|\theta^*\|_2.$$

For $\|\theta^{(t+1)} - \theta^*\| \le f_2$ to hold, we need $\frac{\lambda_2}{e^{\kappa E}} f_2 > C\sqrt{\frac{n_{\max}n}{L}}$ for some sufficiently large constant C, which is guaranteed by the definition of f_2.

Next, we bound $|\theta_m^{(t+1,m)} - \theta_m^*|$. Note that by the definition of the gradient descent

$$\theta_m^{(t+1,m)} - \theta_m^* = \left[1 - \eta\lambda - \eta\sum_{j\in[n]\setminus\{m\}} A_{mj}\psi'(\xi_j)\right](\theta_m^{(t,m)} - \theta_m^*) - \lambda\eta\theta_m^* + \eta\sum_{j\in[n]\setminus\{m\}} A_{mj}\psi'(\xi_j)(\theta_j^{(t,m)} - \theta_j^*)$$

where ξ_j is a scalar between $\theta_m^* - \theta_j^*$ and $\theta_m^{(t,m)} - \theta_j^{(t,m)}$. Since $\|\theta^{(t,m)} - \theta^*\|_\infty \le 3$, we have $|\xi_j - \theta_m^* + \theta_j^*| \le |\theta_m^* - \theta_j^* - \theta_m^{(t,m)} + \theta_j^{(t,m)}| \le 6$ and $\|\xi\|_\infty$ is bounded. Therefore,

$$\sum_{j\in[n]\setminus\{m\}} A_{mj}\psi'(\xi_j) \ge c_3 \min_{i\in[n]} n_i$$

and

$$\begin{aligned}|\sum_{j\in\mathcal{N}(m)} A_{mj}\psi'(\xi_j)(\theta_j^{(t,m)} - \theta_j^*)| &\le \sqrt{\sum_{j\in\mathcal{N}(m)}[A_{mj}\psi'(\xi_j)]^2}\sqrt{\sum_{j\in\mathcal{N}(m)}(\theta_j^{(t,m)} - \theta_j^*)^2} \\ &\le c_4\sqrt{n_{\max}}(f_1 + f_2).\end{aligned}$$

for some constant $c_3, c_4 > 0$. Thus we have

$$|\theta_m^{(t+1,m)} - \theta_m^*| \le (1 - c_3\eta n_{\min}) + \eta\sqrt{n_{\max}}(f_1 + f_2) + \lambda\eta|\theta_m^*|.$$

For $|\theta_m^{(t+1,m)} - \theta_m^*| \le f_3$ to hold, we need $\frac{n_{\min}}{e^{\kappa E}} f_3 > C\sqrt{n_{\max}} f_2$ for some sufficiently large constant $C > 0$, which is ensured by the definition of f_2, f_3.

As the last step, we again use the fact that $\ell_\rho(\cdot)$ is ρ-strongly convex and $(\rho + n_{\max})$-smooth (see definition in the paragraph before Equation (11)), so by Theorem 3.10 in Bubeck (2015), we have

$$\|\theta^{(t)} - \hat{\theta}_\rho\|_2 \le (1 - \frac{\rho}{\rho + n_{\max}})^t\|\theta^* - \hat{\theta}_\rho\|_2. \tag{74}$$

By a union bound, Equation (72) holds for all $t \le T$ with probability at leaast $1 - O(Tn^{-10})$. Triangle inequality implies that

$$\|\hat{\theta}_\rho - \theta^*\|_\infty \le \|\theta^{(T)} - \hat{\theta}_\rho\|_2 + \|\theta^{(T)} - \theta^*\|_\infty \le (1 - \frac{\rho}{\rho + n_{\max}})^T\sqrt{n}\|\hat{\theta}_\rho - \theta^*\|_\infty + 2$$

Take $T = n^5$ and remember that $L \lesssim n^5$. If $\rho > n_{\max}$, then $(1 - \frac{\rho}{\rho + n_{\max}})^T\sqrt{n} \le 2^{-n^5}\sqrt{n} \le 1/2$. Otherwise, since

$$(1 - \frac{\rho}{\rho + n_{\max}})^T\sqrt{n} \le \exp\left(-\frac{T\rho}{\rho + n_{\max}}\right)\sqrt{n},$$

using the fact that $\kappa < n$, we have

$$(1 - \frac{\rho}{\rho + n_{\max}})^T\sqrt{n} \le \exp\left(-\frac{T}{c\kappa}\sqrt{\frac{1}{n_{\max}L}}\right)\kappa\sqrt{n} \le ce^{-n}n^{3/2} \le \frac{1}{2}.$$

In conclusion, we have $\|\hat{\theta}_\rho - \theta^*\|_\infty \le \frac{1}{2}\|\hat{\theta}_\rho - \theta^*\|_\infty + 2$, thus $\|\hat{\theta}_\rho - \theta^*\|_\infty \le 4$, with probability at least $1 - O(n^{-5})$. □

Proof of Lemma 26. By definition, $\boldsymbol{\theta}_{-m}^{(m)}$ is a constrained MLE on a subset of the data, thus by Taylor expansion, for ξ given by a convex combination of $\boldsymbol{\theta}_{-m}^*$ and $\boldsymbol{\theta}_{-m}^{(m)}$, we have

$$\begin{aligned}\ell_n^{(-m)}\left(\boldsymbol{\theta}_{-m}^*\right) \geq & \ell_n^{(-m)}(\boldsymbol{\theta}_{-m}^{(m)}) \\ = & \ell_n^{(-m)}\left(\boldsymbol{\theta}_{-m}^*\right) + (\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1})^T \nabla \ell_n^{(-m)}\left(\boldsymbol{\theta}_{-m}^*\right) \\ & + \frac{1}{2}(\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1})^T H^{(-m)}(\xi)(\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1}),\end{aligned}$$

where we use the invariant property of $\ell_n^{(-m)}(\boldsymbol{\theta}_{-m})$, i.e., $\ell_n^{(-m)}(\boldsymbol{\theta}_{-m}) = \ell_n^{(-m)}(\boldsymbol{\theta}_{-m} + c\mathbf{1}_{n-1})$. By the fact that $\|\xi - \boldsymbol{\theta}_{-m}^*\|_\infty \leq \|\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^*\|_\infty \leq 5$ and Lemma 10, we have

$$(\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1})^T H^{(-m)}(\xi)(\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1}) \geq c e^{-\kappa_E} \lambda_2(\mathcal{L}_{A_{-m}}) \|\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1}\|_2^2.$$

Applying Cauchy-Schwartz inequality to the expansion and we can get

$$\|\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1}\|_2^2 \leq \frac{e^{2\kappa_E}}{c^2 \lambda_2(\mathcal{L}_{A_{-m}})^2} \|\nabla \ell_n^{(-m)}(\boldsymbol{\theta}_m^*)\|_2^2.$$

Where A_{-m} is the adjacency matrix of the comparison graph with node m excluded. By the interlacing property of the eigenvalue sequences of Laplacians of graph and its induced subgraph (see, e.g. Brouwer and Haemers, 2012, Proposition 3.2.1), we have $\lambda_2(\mathcal{L}_{A_{-m}}) \geq \lambda_2(\mathcal{L}_A)$, thus

$$\|\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1}\|_2^2 \leq \frac{e^{2\kappa_E}}{c^2 \lambda_2(\mathcal{L}_A)^2} \|\nabla \ell_n^{(-m)}(\boldsymbol{\theta}_m^*)\|_2^2. \tag{75}$$

Now by Lemma 9, it holds with probability at least $1 - O(n^{-10})$ that

$$\|\boldsymbol{\theta}_{-m}^{(m)} - \boldsymbol{\theta}_{-m}^* - a_m \mathbf{1}_{n-1}\|_2^2 \leq C \frac{e^{2\kappa_E} n_{\max} n}{L \lambda_2(\mathcal{L}_A)^2},$$

and the conclusion is guaranteed by a union bound. □