Feeling the Bern: Adaptive Estimators for Bernoulli Probabilities of Pairwise Comparisons

Nihar B. Shah
Dept. of EECS
UC Berkeley
nihar@eecs.berkeley.edu

Sivaraman Balakrishnan
Dept. of Statistics
CMU
siva@stat.cmu.edu

Martin J. Wainwright
Dept. of EECS and Statistics
UC Berkeley
wainwrig@berkeley.edu

Abstract

We study methods for aggregating pairwise comparison data in order to estimate outcome probabilities for future comparisons among a collection of n items. Working within a flexible framework that imposes only a form of strong stochastic transitivity (SST), we introduce an adaptivity index defined by the indifference sets of the pairwise comparison probabilities. In addition to measuring the usual worst-case risk of an estimator, this adaptivity index also captures the extent to which the estimator adapts to instance-specific difficulty relative to an oracle estimator. We prove three main results that involve this adaptivity index and different algorithms. First, we propose a three-step estimator termed Count-Randomize-Least squares (CRL), and show that it has adaptivity index upper bounded as $\sqrt{n}$ up to logarithmic factors. We then show that that conditional on the hardness of planted clique, no computationally efficient estimator can achieve an adaptivity index smaller than $\sqrt{n}$. Second, we show that a regularized least squares estimator can achieve a poly-logarithmic adaptivity index, thereby demonstrating a $\sqrt{n}$-gap between optimal and computationally achievable adaptivity. Finally, we prove that the standard least squares estimator, which is known to be optimally adaptive in several closely related problems, fails to adapt in the context of estimating pairwise probabilities.

1 Introduction

There is an extensive literature on modeling and analyzing data in the form of pairwise comparisons between items, with much of the earliest literature focusing on applications in voting, social choice theory, and tournaments. The advent of new internet-scale applications, particularly search engine ranking [RKJ08], online gaming [HMG07], and crowdsourcing [SBB$^+$15], has renewed interest in ranking problems, particularly in the statistical and computational challenges that arise from the aggregation of large data sets of paired comparisons.

The problem of aggregating pairwise comparisons, which may be inconsistent and/or noisy, presents a number of core challenges, including: (i) how to produce a consensus ranking from the paired comparisons [BM08, RGLA15, SW15]; (ii) how to estimate a notional "quality" for each of the underlying objects [NOS12, HOX14, SBB$^+$15]; and (iii) how to estimate the probability of the outcomes of subsequent comparisons [Cha14, SBGW15]. In this paper, we focus on the third task—that is, the problem of estimating the probability that one object is preferred to another. Accurate knowledge of such pairwise comparison probabilities is useful in various applications, including (in operations research) estimating the probability of a customer picking one product over another, or (in sports bookmaking and tournament design) estimating the probability of one team beating another.

In more detail, given a set of n items $\{1, \ldots, n\}$, the paired comparison probabilities can be described by an $(n \times n)$ matrix M^* in which the $(i, j)^{\text{th}}$ entry corresponds to the probability that item i beats item j. From this perspective, problem of estimating the comparison probabilities amounts to estimating the unknown matrix M^*. In practice, one expects that the pairwise comparison probabilities exhibit some form of structure, and in this paper, in line with some past work on the problem, we assume that the entries of the matrix M^* satisfy the strong stochastic transitivity (SST) constraint. It is important to note that the SST constraint is considerably weaker than standard parametric assumptions that are often made in the literature—for instance, that the entries of M^* follow a Bradley-Terry-Luce [BT52, Luc59] or Thurstone [Thu27] model. The SST constraint is quite flexible and models satisfying this constraint often provide excellent fits to paired comparison data in a variety of applications. There is also a substantial body of empirical work that validates the SST assumption—for instance, see the papers [ML65, DM59, BW97] in the psychology and economics literatures.

On the theoretical front, some past work [Cha14, SBGW15] has studied the problem of estimating SST matrices in the Frobenius norm. These works focus exclusively on the *global* minimax error, meaning that the performance of any estimator is assessed in a worst-case sense globally over the entire SST class. It is well-known that the criterion of global minimax can lead to a poor understanding of an estimator, especially in situations where the intrinsic difficulty of the estimation task is highly variable over the parameter space (see, for instance, the discussion and references in Donoho et al. [DJKP95]). In such situations, it can be fruitful to benchmark the risk of an estimator against that of a so-called oracle estimator that is provided with side-information about the local structure of the parameter space. Such a benchmark can be used to show that a given estimator is *adaptive*, in the sense that even though it is not given side-information about the problem instance, it is able to achieve lower risk for "easier" problems (e.g., see the papers [Can06, Kol11, CL11] for results of this type).[1] In this paper, we study the problem-specific difficulty of estimating a pairwise comparison matrix M^* by introducing an adaptivity index that involves the size of the indifference sets in the matrix M^*. These indifference sets, which arise in many relevant applications, correspond to subsets of items that are all equally desirable.

In addition, our work makes contributions to a growing body of work (e.g., [BR13, MW15, Wai14]) that studies the notion of a computationally-constrained statistical risk.

[1] The term "adaptivity" in this paper derives its meaning from the literature on statistics, and refers to the property of an estimator of automatically adapting its performance to the complexity of the problem. It should not be confused with the notion of "adaptive sampling" used in the context of sequential design or adaptive learning, which refers to the ability to obtain samples one at a time in a sequential and data-dependent manner.

In more detail, we make the following contributions in this paper:

- We show that the risk of estimating a pairwise comparison probability matrix M^* depends strongly on the size of its largest indifference set. This fact motivates us to define an adaptivity index that benchmarks the performance of an estimator relative to that of an oracle estimator that is given additional side information about the size of the indifference sets in M^*. By definition, an estimator with lower values of this index is said to exhibit better adaptivity, and the oracle estimator has an adaptivity index of 1.
- We characterize the fundamental limits of adaptivity, in particular by proposing a regularized least squares estimator with a carefully chosen regularization function. With a suitable choice of regularization parameter, we prove that this estimator achieves an $\widetilde{\mathcal{O}}(1)$ adaptivity index, which matches the best possible up to poly-logarithmic factors.
- We then show that conditional on the planted clique hardness conjecture, the adaptivity index achieved by any polynomial-time algorithm must be lower bounded as $\widetilde{\Omega}(\sqrt{n})$. This result exhibits an interesting gap between the adaptivity of polynomial-time versus statistically optimal estimators.
- We propose a computationally-efficient three-step "Count–Randomize–Least squares" (CRL) estimator for estimation of SST matrices, and show that its adaptivity index is upper bounded as $\widetilde{\mathcal{O}}(\sqrt{n})$. Due to the aforementioned lower bound, the CRL estimator has the best possible adaptivity among all possible computationally efficient estimators.
- Finally, we investigate the adaptivity of the standard (unregularized) least squares estimator. This estimator is found to have good, or even optimal adaptivity in several related problems, and is also minimax-optimal for the problem of estimating SST matrices. We prove that surprisingly, the adaptivity of the least squares estimator for estimating SST matrices is of the order $\widetilde{\Theta}(n)$, which is as bad as a constant estimator that is independent of the data.

The remainder of this paper is organized as follows. We begin in Section 2 with background on the problem. Section 3 is devoted to the statement of our main results, as well as discussion of their consequences. In Section 4, we provide the proofs of our main results, with the more technical details deferred to appendices. Finally, Section 5 presents concluding remarks.

2 Background and problem setting

In this section, we provide background and a more precise problem statement.

2.1 Estimation from pairwise comparisons

Given a collection of n items, suppose that we arrange the paired comparison probabilities in a matrix $M^* \in [0,1]^{n \times n}$, where M^*_{ij} is the probability that item i is preferred to item j in a paired comparison. Accordingly, the upper and lower halves of M^* are related by the shifted-skew-symmetry condition $M^*_{ji} = 1 - M^*_{ij}$ for all $i, j \in [n]$, where we assume that $M^*_{ii} = 0.5$ for all $i \in [n]$

for concreteness. In other words, the shifted matrix $M^* - \frac{1}{2}11^T$ is skew-symmetric. Here we have adopted the standard shorthand $[n] := \{1, 2, \ldots, n\}$.

Suppose that we observe a random matrix $Y \in \{0,1\}^{n \times n}$ with (upper-triangular) independent Bernoulli entries, in particular, with

$$\mathbb{P}[Y_{ij} = 1] = M^*_{ij} \qquad \text{for every } 1 \leq i \leq j \leq n, \tag{1}$$

and $Y_{ji} = 1 - Y_{ij}$ except on the diagonal. We take the diagonal entries Y_{ii} to be $\{0,1\}$ with equal probability, for every $i \in [n]$. The focus of this work is not to evaluate the effects of the choice of the pairs compared, but to understand the effects of the noise models. Consequently, we restrict attention to the case of a single observation per pair, but keeping mind in that one may extend the result to other observation models via techniques similar to those proposed in our past work [SBB$^+$15, SBGW15]. Based on observing Y, our goal in this paper is to recover an accurate estimate, in the squared Frobenius norm, of the full matrix M^*.

We consider matrices M^* that satisfy the constraint of *strong stochastic transitivity* (SST), which reflects the natural transitivity of any complete ordering. Formally, suppose that the set of all items $[n]$ is endowed with a complete ordering π^*. We use the notation $\pi^*(i) \succ \pi^*(j)$ to indicate that item i is preferred to item j in the total ordering π^*. We say that the M^* satisfies the SST condition with respect to the permutation π^*—or is π^*-SST for short—if

$$M^*_{ik} \geq M^*_{jk} \qquad \text{for every triple } (i, j, k) \text{ such that } \pi^*(i) \succ \pi^*(j). \tag{2}$$

The intuition underlying this constraint is as follows: since i dominates j in the true underlying order, when we make noisy comparisons, the probability that i is preferred to k should be at least as large as the probability that j is preferred to k. The class of all SST matrices is given by

$$\mathbb{C}_{\text{SST}} := \Big\{ M \in [0,1]^{n \times n} \;\mid\; M_{ba} = 1 - M_{ab} \;\forall\, (a,b) \text{ and } \exists\, \pi \text{ such that } M \text{ is } \pi\text{-SST} \Big\}. \tag{3}$$

The goal of this paper[2] is to design estimators that can estimate the true underlying matrix $M^* \in \mathbb{C}_{\text{SST}}$ from the observed matrix Y.

2.2 Indifference sets

We now turn to the notion of *indifference sets*, which allows for a finer-grained characterization of the difficulty of estimating a particular matrix. Suppose that the set $[n]$ of all items is partitioned into the union of s disjoint sets $\{P_i\}_{i=1}^s$ of sizes $\mathbf{k} = (k_1, \ldots, k_s)$ such that $\sum_{i=1}^s k_i = n$. For reasons to be clarified in a moment, we term each of these sets as an *indifference set*. We write $i \sim i'$ to mean that the pair i and i' belong to the same index set, and we say that a matrix $M^* \in \mathbb{R}^{n \times n}$ *respects the indifference set partition* $\{P_i\}_{i=1}^s$ if

$$M^*_{ij} = M^*_{i'j'} \quad \text{for all quadruples } (i, j, i', j') \text{ such that } i \sim i' \text{ and } j \sim j'. \tag{4}$$

[2] We note that an accurate estimate of M^* leads to an accurate estimate of the underlying permutation as well [SBGW15].

For instance, in the special case of a two-contiguous-block partition, the matrix M^* must have a (2×2) block structure, with all entries equaling $1/2$ in the two diagonal blocks, all entries equaling $\alpha \in [0,1]$ in the upper right block, and equaling $(1-\alpha)$ in the lower left block. Intuitively, matrices with this type of block structure should be easier to estimate.

Indifference sets arise in various applications of ranking: for instance, in buying cars, frugal customers may be indifferent between high-priced cars; or in ranking news items, people from a certain country may be indifferent to the domestic news from other countries. Block structures of this type are also studied in other types of matrix estimation problems, in which contexts they have been termed communities, blocks, or level sets, depending on the application under consideration. For instance, see the papers [AS15, MW15, CGS15] as well as references therein for more discussion in such structures.

Given the number of partitions s and their size vector $\mathbf{k} = (k_1, \ldots, k_s)$, we let $\mathbb{C}_{\text{SST}}(s, \mathbf{k})$ denote the subset of $\mathbb{C}_{\text{SST}}$ comprising all SST matrices that respect some indifference set partition $\{P_i\}_{i=1}^s$ of sizes $\mathbf{k}$. The size of the largest indifference set $k_{\max} := \|\mathbf{k}\|_\infty = \max\limits_{i \in \{1,\ldots,s\}} k_i$ plays an important role in our analysis. We also use the notation $\mathbb{C}_{\text{SST}}(k_{\max})$ to denote all SST matrices that have at least one indifference set of size at least $k_{\max}$, that is,

$$\mathbb{C}_{\text{SST}}(k_{\max}) := \bigcup_{\|\mathbf{k}\|_\infty \geq k_{\max}} \mathbb{C}_{\text{SST}}(s, \mathbf{k}), .$$

Finally, with a minor abuse of notation, for any matrix $M \in \mathbb{C}_{\text{SST}}$, we let $k_{\max}(M)$ denote the size of the largest indifference set in M.

2.3 An oracle estimator and the adaptivity index

We begin by defining a benchmark based on the performance of the best estimator that has side-information that $M^* \in \mathbb{C}_{\text{SST}}(s, \mathbf{k})$, along with the values of $(s, \mathbf{k})$. We evaluate any such estimator $\widetilde{M}(s, \mathbf{k})$ based on its mean-squared Frobenius error

$$\mathbb{E}[|\!|\!|\widetilde{M}(s, \mathbf{k}) - M^*|\!|\!|_{\text{F}}^2] = \mathbb{E}\Big[\sum_{i,j=1}^{n} \big(\widetilde{M}_{ij}(s, \mathbf{k}) - M_{ij}^*\big)^2\Big], \tag{5}$$

where the expectation is taken with respect to the random matrix $Y \in \{0,1\}^{n \times n}$ of noisy comparisons. With this notation, the $(s, \mathbf{k})$-*oracle risk* is given by

$$R_n(s, \mathbf{k}) := \inf_{\widetilde{M}(s,\mathbf{k})} \sup_{M^* \in \mathbb{C}_{\text{SST}}(s,\mathbf{k})} \mathbb{E}[|\!|\!|\widetilde{M}(s, \mathbf{k}) - M^*|\!|\!|_{\text{F}}^2], \tag{6}$$

where the infimum is taken over all measurable functions $\widetilde{M}(s, \mathbf{k})$ of the data Y.

For a given estimator $\widehat{M}$ that *does not know* the values of $(s, \mathbf{k})$, we can then compare its performance to this benchmark via the $(s, \mathbf{k})$-*adaptivity index*

$$\alpha_n(\widehat{M}; s, \mathbf{k}) := \frac{\sup\limits_{M^* \in \mathbb{C}_{\text{SST}}(s,\mathbf{k})} \mathbb{E}[|\!|\!|\widehat{M} - M^*|\!|\!|_{\text{F}}^2]}{R_n(s, \mathbf{k})}. \tag{7a}$$

The *global adaptivity index* $\alpha_n(\widehat{M})$ of an estimator $\widehat{M}$ is then given by

$$\alpha_n(\widehat{M}) := \max_{s,\mathbf{k}:\|\mathbf{k}\|_\infty<n} \alpha_n(\widehat{M}; s, \mathbf{k}). \tag{7b}$$

In this definition, we restrict the maximum to the interval $\|\mathbf{k}\|_\infty < n$ since in the (degenerate) case of $\|\mathbf{k}\|_\infty = n$, the only valid matrix M^* is the all-half matrix and hence the estimator with the knowledge of the parameters trivially incurs an error of zero.

Given these definitions, the goal is to construct estimators that are computable in polynomial time, and possess a low adaptivity index. Finally, we note that an estimator with a low adaptivity index also achieves a good worst-case risk: any estimator $\widehat{M}$ with global adaptivity index $\alpha_n(\widehat{M}) \leq \gamma$ is minimax-optimal within a factor γ.

3 Main results

In this section, we present the main results of this paper on both statistical and computational aspects of the adaptivity index. We begin with an auxiliary result on the risk of the oracle estimator which is useful for our subsequent analysis.

3.1 Risk of the oracle estimator

Recall from Section 2.3 that the oracle estimator has access to additional side information on the values of the number s and the sizes $\mathbf{k} = (k_1, \ldots, k_s)$ of the indifference sets of the true underlying matrix M^*. The oracle estimator is defined as the estimator that incurs the lowest possible risk (6) among all such estimators.

The following result provides tight bounds on the risk of the oracle estimator.

Proposition 1. *There are positive universal constants c_ℓ and c_u such that the $(s, \mathbf{k})$-oracle risk (6) is sandwiched as*

$$c_\ell(n - k_{\max}) \leq R_n(s, \mathbf{k}) \leq c_u(n - k_{\max} + 1)(\log n)^2. \tag{8}$$

Proposition 1 provides a characterization minimax risk of estimation under various subclasses of $\mathbb{C}_{\text{SST}}$. Remarkably, the minimax risk depends on only the size $k_{\max} := \|\mathbf{k}\|_\infty$ of the largest indifference set: given this value, it is not affected by the number of indifference sets s nor their sizes $\mathbf{k}$. This property is in sharp contrast to known results [CGS15] for the related problem of bivariate isotonic regression, in which the number s of indifference sets does play a strong role.

Note that when $k_{\max} < n$, we have $\frac{1}{2}(n - k_{\max} + 1) \leq (n - k_{\max})$, and consequently the lower bound in (8) can be replaced by $\frac{c_\ell}{2}(n - k_{\max} + 1)$.

3.2 Fundamental limits on adaptivity

Proposition 1 provides a sharp characterization of the denominator in the adaptivity index (7a). In this section, we investigate the fundamental limits of adaptivity by studying the numerator but disregarding computational constraints. The main result of this section is to show that a suitably regularized form of least-squares estimation has optimal adaptivity up to logarithmic factors.

More precisely, recall that $k_{\max}(M)$ denotes the size of the largest indifference set in the matrix M. Given the observed matrix Y, consider the M-estimator

$$\widehat{M}_{\text{REG}} \in \arg\min_{M \in \mathbb{C}_{\text{SST}}} \Big(|\!|\!| M - Y |\!|\!|_{\text{F}}^2 - k_{\max}(M) (\log n)^3 \Big). \tag{9}$$

Here the inclusion of term $-k_{\max}(M)$, along with its logarithmic weight, serves to "reward" the estimator for returning a matrix with a relatively large maximum indifference set. As our later analysis in Section 3.5 will clarify, the inclusion of this term is essential: the unregularized form of least-squares has very poor adaptivity properties.

The following theorem provides an upper bound on the estimation error and the adaptivity of the estimator $\widehat{M}_{\text{REG}}$.

Theorem 1. *There are universal constants c_u and c_u' such that for every $M^* \in \mathbb{C}_{SST}$, the regularized least squares estimator* (9) *has squared Frobenius error at most*

$$\frac{1}{n^2} |\!|\!| M^* - \widehat{M}_{REG} |\!|\!|_F^2 \leq c_u \frac{n - k_{\max}(M^*) + 1}{n^2} (\log n)^3, \tag{10a}$$

with probability at least $1 - e^{-\frac{1}{2}(\log n)^2}$. Consequently, its adaptivity index is upper bounded as

$$\alpha_n(\widehat{M}_{REG}) \leq c_u' (\log n)^3. \tag{10b}$$

Since the adaptivity index of any estimator is at least 1 by definition, we conclude that the regularized least squares estimator $\widehat{M}_{\text{REG}}$ is optimal up to logarithmic factors.

The reader may notice that the optimization problem (9) defining the regularized least squares estimator $\widehat{M}_{\text{REG}}$ is non-trivial to solve; it involves both a nonconvex regularizer, as well as a nonconvex constraint set. We shed light on the intrinsic complexity of computing this estimator in Section 3.4, where we investigate the adaptivity index achievable by estimators that are computable in polynomial time.

3.3 Adaptivity of the CRL estimator

In this section, we propose a polynomial-time computable estimator termed the *Count-Randomize-Least-Squares (CRL)* estimator, and prove an upper bound on its adaptivity index. In order to define the CRL estimator, we requre some additional notation. For any permutation π on n items, let $\mathbb{C}_{\text{SST}}(\pi) \subseteq \mathbb{C}_{\text{SST}}$ denote the set of all SST matrices that are faithful to the permutation π—that is

$$\mathbb{C}_{\text{SST}}(\pi) := \big\{ M \in [0,1]^{n \times n} \;\mid\; M_{ba} = 1 - M_{ab} \; \forall\, (a,b),\; M_{ik} \geq M_{jk} \; \forall\, i,j,k \in [n] \text{ s.t. } \pi(i) > \pi(j) \big\}. \tag{11}$$

One can verify that the sets $\{\mathbb{C}_{\text{SST}}(\pi)\}$ for all permutations π on n items form a partition of the SST class $\mathbb{C}_{\text{SST}}$.

The CRL estimator acts on the observed matrix Y and outputs an estimate $\widehat{M}_{\text{CRL}} \in \mathbb{C}_{\text{SST}}$ via a three step procedure:
Step 1 (Count): For each $i \in [n]$, compute the total number $N_i = \sum_{j=1}^{n} Y_{ij}$ of pairwise comparisons that it wins. Order the n items in terms of $\{N_i\}_{i=1}^{n}$, with ties broken arbitrarily.
Step 2 (Randomize): Find the largest subset of items S such that $|N_i - N_j| \leq \sqrt{n}\log n$ for all $i, j \in S$. Using the order computed in Step 1, permute this (contiguous) subset of items uniformly at random within the subset. Denote the resulting permutation as π_{CRL}.
Step 3 (Least squares): Compute the least squares estimate assuming that the permutation π_{CRL} is the true permutation of the items:

$$\widehat{M}_{\text{CRL}} \in \underset{M \in \mathbb{C}_{\text{SST}}(\pi_{\text{CRL}})}{\arg\min} \; |||Y - M|||_{\text{F}}^2. \tag{12}$$

The optimization problem (12) corresponds to a projection onto the polytope of bi-isotone matrices contained within the hypercube $[0,1]^n$, along with skew symmetry constraints. Problems of the form (12) have been studied in past work [BDPR84, RWDR88, Cha14, KRS15], and the estimator $\widehat{M}_{\text{CRL}}$ is indeed computable in polynomial time. By construction, it is agnostic to the values of $(s, \mathbf{k})$.

To provide intuition for the second step of randomization, it serves to discard "non-robust" information from the order computed in Step 1. Any such information corresponds to noise due to the Bernoulli sampling process, as opposed to structural information about the matrix. If we do not perform this second step—effectively retaining considerable bias from Step 1—then then isotonic regression procedure in Step 3 may amplify it, leading to a poorly performing estimator. To clarify our choice of threshold $T = \sqrt{n}\log(n)$, the factor $\sqrt{n}$ corresponds to the standard deviation of a typical win count N_i (as a sum of Bernoulli variables), whereas the $\log n$ serves to control fluctuations in a union bound.

The following theorem provides an upper bound on the adaptivity index achieved by the CRL estimator.

Theorem 2. *There are universal constants c_u and c_u' such that for every $M^* \in \mathbb{C}_{SST}$, the CRL estimator $\widehat{M}_{CRL}$ has squared Frobenius norm error*

$$\frac{1}{n^2}|||\widehat{M}_{CRL} - M^*|||_F^2 \leq c_u \frac{n - k_{\max}(M^*) + 1}{n^{3/2}} (\log n)^8, \tag{13a}$$

with probability at least $1 - n^{-20}$. Consequently, its adaptivity index is upper bounded as

$$\alpha_n(\widehat{M}_{CRL}) \leq c_u' \sqrt{n} (\log n)^8. \tag{13b}$$

It is worth noting that equation (13a) in yields an upper bound on the minimax risk of the CRL estimator—namely

$$\sup_{M^* \in \mathbb{C}_{\text{SST}}} \frac{1}{n^2}\mathbb{E}[|||\widehat{M}_{\text{CRL}} - M^*|||_{\text{F}}^2] \leq c_u \frac{(\log n)^8}{\sqrt{n}},$$

with this worst-case achieved when $k_{\max}(M^*) = 1$. Up to logarithmic factors, this bound matches the best known upper bound on the minimax rate of polynomial-time estimators [SBGW15, Theorem 2].

3.4 A lower bound on adaptivity for polynomial-time algorithms

By comparing the guarantee (13b) for the CRL estimator with the corresponding guarantee (10b) for the regularized least-squares estimator, we see that (apart from log factors and constants), their adaptivity indices differ by a factor of $\sqrt{n}$. Given this polynomial gap, it is natural to wonder whether our analysis of the CRL estimator might be improved, or if not, whether there is another polynomial-time estimator with a lower adaptivity index than the CRL estimator. In this section, we answer *both of these questions in the negative,* at least conditionally on a certain well-known conjecture in average case complexity theory.

More precisely, we prove a lower bound that relies on the average-case hardness of the planted clique problem [Jer92, Kuč95]. The use of this conjecture as a hardness assumption is widespread in the literature [JP00, AAK$^+$07, Dug14], and there is now substantial evidence in the literature supporting the conjecture [Jer92, FK03, MPW15, DM15]. It has also been used as a tool in proving hardness results for sparse PCA and related matrix recovery problems [BR13, MW15].

In informal terms, the planted clique conjecture asserts that it is hard to detect the presence of a planted clique in an Erdős-Rényi random graph. In order to state it more precisely, let $\mathcal{G}(n, \kappa)$ be a random graph on n vertices constructed in one of the following two ways:

H_0: Every edge is included in $\mathcal{G}(n, \kappa)$ independently with probability $\frac{1}{2}$.

H_1: Every edge is included in $\mathcal{G}(n, \kappa)$ independently with probability $\frac{1}{2}$. In addition, a set of κ vertices is chosen uniformly at random and all edges with both endpoints in the chosen set are added to $\mathcal{G}$.

The planted clique conjecture then asserts that when $\kappa = o(\sqrt{n})$, then there is no polynomial-time algorithm that can correctly distinguish between H_0 and H_1 with an error probability that is strictly bounded below 1/2.

Using this conjectured hardness as a building block, we have the following result:

Theorem 3. *Suppose that the planted clique conjecture holds. Then there is a universal constant* $c_\ell > 0$ *such that for any polynomial-time computable estimator* $\widehat{M}$*, its adaptivity index is lower bounded as*

$$\alpha_n(\widehat{M}) \geq c_\ell \sqrt{n} (\log n)^{-3}.$$

Together, the upper and lower bounds of Theorems 2 and 3 imply that the estimator $\widehat{M}_{\text{CRL}}$ achieves the optimal adaptivity index (up to logarithmic factors) among all computationally efficient estimators.

3.5 Negative results for the least squares estimator

In this section, we study the adaptivity of the (unregularized) least squares estimator given by

$$\widehat{M}_{LS} \in \arg\min_{M \in \mathbb{C}_{\text{SST}}} |\!|\!| Y - M |\!|\!|_{\text{F}}^2. \tag{14}$$

Least squares estimators of this type are known to possess very good adaptivity in various other problems of shape-constrained estimation; for instance, see the papers [C$^+$11, CGS13, CL15, CGS15, Bel16] and references therein for various examples of such phenomena. From our own past work [SBGW15], the estimator (14) is known to be minimax optimal for estimating SST matrices.

Given this context, the following theorem provides a surprising result—namely, that the least-squares estimator (14) has remarkably poor adaptivity:

Theorem 4. *There is a universal constant $c_\ell > 0$ such that the adaptivity index of the least squares estimate* (14) *is lower bounded as*

$$\alpha_n(\widehat{M}_{LS}) \geq c_\ell \, n \, (\log n)^{-2}. \tag{15}$$

In order to understand why the lower bound (15) is very strong, consider the trivial estimator M_0 that simply *ignores the data*, and returns the constant matrix $M_0 = \frac{1}{2} 1 1^T$. It can be verified that we have

$$|\!|\!| M^* - \frac{1}{2} 1 1^T |\!|\!|_{\text{F}}^2 \leq 3n(n - k_{\max}(M^*) + 1)$$

for every $M^* \in \mathbb{C}_{\text{SST}}$. Thus, for this trival estimator M_0, we have $\alpha_n(M_0) \leq c_u n$. Comparing to the lower bound (15), we see that apart from logarithmic factors, the adaptivity of the least squares estimator is no better than that of the trivial estimator M_0.

4 Proofs

In this section, we present the proofs of our results. We note in passing that our proofs additionally lead to some auxiliary results that may be of independent interest. These auxiliary results pertain to the problem of bivariate isotonic regression—that is, estimating M^* when the underlying permutation is known—an important problem in the field of shape-constrained estimation [RWDR88, THT11, Cha14]. Prior works restrict attention to the expected error and assume that the underlying permutation is correctly specified; our results provide exponential tail bounds and also address settings when the permutation is misspecified.

A few comments about assumptions and notation are in order. In all of our proofs, so as to avoid degeneracies, we assume that the number of items n is greater than a universal constant. (The cases when n is smaller than some universal constant all follow by adjusting the pre-factors in front of our results suitably.) For any matrix M, we use $k_{\max}(M)$ to denote the size of the largest indifference set in M, and we define $k^* = k_{\max}(M^*)$. The notation c, c_1, c_u, c_ℓ etc. all denote positive universal constants. For any two square matrices A and B of the same size, we let $\langle\!\langle A,\ B \rangle\!\rangle = \text{trace}(A^T B)$ denote their trace inner product. For an $(n \times n)$ matrix M and any permutation π on n items, we let $\pi(M)$ denote an $(n \times n)$ matrix obtained by permuting the rows and columns of M by π. For a given class $\mathbb{C}$ of matrices, metric ρ and tolerance $\epsilon > 0$, we use $N(\epsilon, \mathbb{C}, \rho)$ to denote the ϵ covering number of the class $\mathbb{C}$ in the metric ρ. The metric entropy is given by the logarithm of the covering number—namely $\log N(\epsilon, \mathbb{C}, \rho)$.

It is also convenient to introduce a linearized form of the observation model (1). Observe that we can write the observation matrix Y in a linearized fashion as

$$Y = M^* + W, \tag{16a}$$

where $W \in [-1, 1]^{n \times n}$ is a random matrix with independent zero-mean entries for every $i > j$, and and $W_{ji} = -W_{ij}$ for every $i < j$. For $i > j$, its entries follow the distribution

$$W_{ij} \sim \begin{cases} 1 - M^*_{ij} & \text{with probability } M^*_{ij} \\ -M^*_{ij} & \text{with probability } 1 - M^*_{ij}. \end{cases} \tag{16b}$$

In summary, all entries of the matrix W above the main diagonal are independent, zero-mean, and uniformly bounded by 1 in absolute value. This fact plays an important role in several parts of our proofs.

4.1 A general upper bound on regularized M-estimators

In this section, we prove a general upper bound that applies to a relatively broad class of regularized M-estimators for SST matrices. Given a matrix Y generated from the model (16a), consider an estimator of the form

$$\widehat{M} \in \arg\min_{M \in \mathbb{C}} \Big\{ |\!|\!| Y - M |\!|\!|_{\text{F}}^2 + \lambda(M) \Big\}. \tag{17}$$

Here $\lambda : [0, 1]^{n \times n} \to \mathbb{R}_+$ is a regularization function to be specified by the user, and $\mathbb{C}$ is some subset of the class $\mathbb{C}_{\text{SST}}$ of SST matrices. Our goal is to derive a high-probability bound on the Frobenius norm error $|\!|\!| \widehat{M} - M^* |\!|\!|_{\text{F}}$. As is well-known from theory on M-estimators [vdG00, BBM05, Kol06], doing so requires studying the empirical process in a localized sense.

In order to do so, it is convenient to consider sets of the form

$$\mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C}) := \{\alpha(M - M^*) \mid M \in \mathbb{C}, \alpha \in [0, 1],\ b |\!|\!| \alpha(M - M^*) |\!|\!|_{\text{F}} \le bt\},$$

where $t \in [0, n]$, and $b \in \{0, 1\}$. The binary flag b controls whether or not the set is localized around M^*, and the radius t controls the extent to which the set is localized.

In the analysis to follow, we assume that for each $\epsilon \geq n^{-8}$, the ϵ-metric entropy of $\mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C})$ satisfies an upper bound of the form

$$\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C}), \|\!| \cdot \|\!|_{\text{F}}) \leq \frac{t^{2b}(g(M^*))^2}{\epsilon^2} + (h(M^*))^2, \tag{18a}$$

where $g : \mathbb{R}^{n \times n} \mapsto \mathbb{R}_+$ and $h : \mathbb{R}^{n \times n} \mapsto \mathbb{R}_+$ are some functions. In the sequel, we provide concrete examples of sets $\mathbb{C}$ and functions (g, h) for which a bound of this form holds.

Given (g, h, λ), we can then define a critical radius $\delta_n \geq 0$ as

$$\delta_n^2 = c\big((g(M^*) \log n)^{1+b} + (h(M^*))^2 + \lambda(M^*) + n^{-7}\big), \tag{18b}$$

where $c > 0$ is a universal constant. The following result guarantees that the Frobenius norm can be controlled by the square of this critical radius:

Lemma 1. *For any set $\mathbb{C}$ satisfying the metric entropy bound* (18a)*, the Frobenius norm of the M-estimator* (17) *can be controlled as*

$$\mathbb{P}\Big[\|\!|\widehat{M} - M^*\|\!|_F^2 > u\delta_n^2\Big] \leq e^{-u\delta_n^2} \qquad \textit{for all } u \geq 1, \tag{19}$$

where δ_n is the critical radius (18b).

The significance of this claim is that it reduces the problem of controlling the error in the M-estimator to bounding the metric entropy (as in equation (18a)), and then computing the critical radius (18b). The remainder of this section is devoted to the proof of this claim.

4.1.1 Proof of Lemma 1

Define the difference $\widehat{\Delta} = \widehat{M} - M^*$ between M^* and the optimal solution $\widehat{M}$ to the constrained least-squares problem. Since $\widehat{M}$ is optimal and M^* is feasible, we have

$$\|\!|Y - \widehat{M}\|\!|_{\text{F}}^2 + \lambda(\widehat{M}) \leq \|\!|Y - M^*\|\!|_{\text{F}}^2 + \lambda(M^*).$$

Following some algebra, and using the assumed non-negativity condition $\lambda(\cdot) \geq 0$, we arrive at the basic inequality

$$\frac{1}{2}\|\!|\widehat{\Delta}\|\!|_{\text{F}}^2 \leq \langle\!\langle \widehat{\Delta},\, W \rangle\!\rangle + \lambda(M^*),$$

where $W \in [0,1]^{n \times n}$ is the noise matrix in the linearized observation model (16a), and $\langle\!\langle \widehat{\Delta},\, W \rangle\!\rangle$ denotes the trace inner product between $\widehat{\Delta}$ and W.

Now define the supremum $Z(t) := \sup\limits_{D \in \mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C})} \langle\!\langle D,\, W \rangle\!\rangle$. With this definition, we find that the error matrix $\widehat{\Delta}$ satisfies the inequality

$$\frac{1}{2}\|\!|\widehat{\Delta}\|\!|_{\text{F}}^2 \leq \langle\!\langle \widehat{\Delta},\, W \rangle\!\rangle + \lambda(M^*) \;\leq\; Z\big(\|\!|\widehat{\Delta}\|\!|_{\text{F}}\big) + \lambda(M^*). \tag{20}$$

Thus, in order to obtain a high probability bound, we need to understand the behavior of the random quantity $Z(t)$.

By definition, the set $\mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C})$ is star-shaped, meaning that $\alpha D \in \mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C})$ for every $\alpha \in [0,1]$ and every $D \in \mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C})$. Using this star-shaped property, it is straightforward to verify that $Z(t)$ grows at most linearly with t, ensuring that there is a non-empty set of scalars $t > 0$ satisfying the critical inequality:

$$\mathbb{E}[Z(t)] + \lambda(M^*) \leq \frac{t^2}{2}. \tag{21}$$

Our interest is in an upper bound on the smallest (strictly) positive solution δ_n to the critical inequality (21). Moreover, our goal is to show that for every $t \geq \delta_n$, we have $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}^2 \leq ct\delta_n$ with probability at least $1 - c_1 e^{-c_2 t\delta_n}$.

Define the "bad" event

$$\mathcal{A}_t := \big\{\exists \Delta \in \mathbb{C}_{\text{DIFF}}(M^*, t) \mid |\!|\!|\Delta|\!|\!|_{\text{F}} \geq \sqrt{t\delta_n} \quad \text{and} \quad \langle\!\langle \Delta,\ W \rangle\!\rangle + \lambda(M^*) \geq 2|\!|\!|\Delta|\!|\!|_{\text{F}}\sqrt{t\delta_n}\big\}. \tag{22}$$

Using the star-shaped property of $\mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C})$ and the fact that $\lambda(\cdot) \geq 0$, it follows by a rescaling argument that

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_n) + \lambda(M^*) \geq 2\delta_n\sqrt{t\delta_n}] \qquad \text{for all } t \geq \delta_n.$$

The entries of W lie in $[-1, 1]$, have a mean of zero, are i.i.d. on and above the diagonal, and satisfy skew-symmetry. Moreover, the function $W \mapsto Z(u)$ is convex and Lipschitz with parameter u. Consequently, by Ledoux's concentration theorem [Led01, Theorem 5.9], we have

$$\mathbb{P}\big[Z(\delta_n) \geq \mathbb{E}[Z(\delta_n)] + \sqrt{t\delta_n}\delta_n\big] \leq e^{-c_1 t\delta_n} \qquad \text{for all } t \geq \delta_n,$$

for some universal constant c_1. By the definition of δ_n, we have $\mathbb{E}[Z(\delta_n)] + \lambda(M^*) \leq \delta_n^2 \leq \delta_n\sqrt{t\delta_n}$ for any $t \geq \delta_n$, and consequently

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z(\delta_n) + \lambda(M^*) \geq 2\delta_n\sqrt{t\delta_n}] \ \leq \ e^{-c_1 t\delta_n} \quad \text{for all } t \geq \delta_n.$$

Consequently, either $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq \sqrt{t\delta_n}$, or we have $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} > \sqrt{t\delta_n}$. In the latter case, conditioning on the complement $\mathcal{A}_t^c$, the basic inequality (20) implies that $\frac{1}{2}|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}^2 \leq 2|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}\sqrt{t\delta_n}$. Putting together the pieces yields that

$$|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq 4\sqrt{t\delta_n},$$

with probability at least $1 - e^{-c_1 t\delta_n}$ for every $t \geq \delta_n$. Substituting $u = \frac{t}{\delta_n}$, we get

$$\mathbb{P}\Big(|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}}^2 > c_2 u\delta_n^2\Big) \leq e^{-c_1 u\delta_n^2}, \tag{23}$$

for every $u \geq 1$.

In order to determine a feasible δ_n satisfying the critical inequality (21), we need to bound the expectation $\mathbb{E}[Z(\delta_n)]$. To this end, we introduce an auxiliary lemma:

Lemma 2. *There is a universal constant c such that for any set $\mathbb{C}$ satisfying the metric entropy bound* (18a)*, we have*

$$\mathbb{E}[Z(t)] \leq c\Big\{t^b g(M^*) \log n + t\, h(M^*) + n^{-7}\Big\} \qquad \textit{for all } t \geq 0. \tag{24}$$

See Section 4.1.2 for the proof of this claim.

Using Lemma 2, we see that the critical inequality (21) is satisfied for

$$\delta_n = c_0\Big\{\big(g(M^*)\log n\big)^{\frac{1}{2}(b+1)} + h(M^*) + \sqrt{\lambda(M^*)} + n^{-\frac{7}{2}}\Big\},$$

for a positive universal constant c_0. With this choice, our claim follows from the tail bound (23), absorbing the constants c_1 and c_2 into c_0.

It remains to prove Lemma 2.

4.1.2 Proof of Lemma 2

By the truncated form of Dudley's entropy inequality, we have

$$\begin{aligned}\mathbb{E}[Z(t)] &\leq c \inf_{\delta \in [0,n]} \Big\{n\delta + \int_{\frac{\delta}{2}}^{t} \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C}), |\!|\!|.|\!|\!|_{\text{F}})}d\epsilon\Big\} \\ &\leq c\,\Big\{2n^{-7} + \int_{n^{-8}}^{t} \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*, t, b, \mathbb{C}), |\!|\!|.|\!|\!|_{\text{F}})}d\epsilon\Big\},\end{aligned} \tag{25}$$

where the second step follows by setting $\delta = 2n^{-8}$. Combining our assumed upper bound (18a) on the metric entropy with the earlier inequality (25) yields

$$\mathbb{E}[Z(t)] \leq c\big\{2n^{-7} + t^b g(M^*)\log(nt) + th(M^*)\big\} \;\leq\; 2c\big\{2n^{-7} + t^b g(M^*)\log n + th(M^*)\big\},$$

where the final step uses the upper bound $t \leq n$. We have thus established the claimed bound (24).

4.2 Proof of Proposition 1

We are now equipped to prove bounds on the risk incurred by the oracle estimator from equation (6).

4.2.1 Upper bound

Let $k^* = \|\mathbf{k}\|_\infty$ denote the size of the largest indifference set in M^*, and recall that the oracle estimator knows the value of k^*. For our upper bound, we use Lemma 1 from the previous section with

$$\mathbb{C} = \mathbb{C}_{\text{SST}}(k^*), \qquad \lambda(M) = 0, \quad \text{and} \quad b = 0.$$

With these choices, the estimator (17) for which Lemma 1 provides guarantees is equivalent to the oracle estimator (6). We then have

$$\mathbb{C}_{\text{DIFF}}(M^*, t, \mathbb{C}_{\text{SST}}(k^*)) = \Big\{\alpha(M - M^*) \mid M \in \mathbb{C}, \alpha \in [0,1]\Big\}.$$

In order to apply the result of Lemma 1, we need to compute the metric entropy of the set $\mathbb{C}_{\text{DIFF}}$. For ease of exposition, we further define the set

$$\widetilde{\mathbb{C}}_{\text{SST}}(k) := \{\alpha M \mid M \in \mathbb{C}_{\text{SST}}(k), \, \alpha \in [0,1]\}.$$

Since $M^* \in \mathbb{C}_{\text{SST}}(k^*)$, the metric entropy of $\mathbb{C}_{\text{DIFF}}$ is at most twice the metric entropy of $\widetilde{\mathbb{C}}_{\text{SST}}(k^*)$. The following lemma provides an upper bound on the metric entropy of the set $\widetilde{\mathbb{C}}_{\text{SST}}(k)$:

Lemma 3. *For every $\epsilon > 0$ and every integer $k \in [n]$, the metric entropy is bounded as*

$$\log N(\epsilon, \widetilde{\mathbb{C}}_{SST}(k), |\!|\!|.|\!|\!|_F) \leq c\frac{(n-k+1)^2}{\epsilon^2}\Big(\log\frac{n}{\epsilon}\Big)^2 + c(n-k+1)\log n, \tag{26}$$

where $c > 0$ is a universal constant.

With this lemma, we are now equipped to prove the upper bound in Proposition 1. The bound (26) implies that

$$\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*, t, \mathbb{C}_{\text{SST}}(k^*)), |\!|\!|.|\!|\!|_{\text{F}}) \leq c'\frac{(n-k^*+1)^2}{\epsilon^2}(\log n)^2 + c'(n-k^*+1)\log n,$$

for all $\epsilon \geq n^{-8}$. Consequently, a bound of the form (18a) holds with $g(M^*) = \sqrt{c'}(n-k^*+1)\log n$ and $h(M^*) = \sqrt{c'(n-k^*+1)\log n}$. Applying Lemma 1 with $u = 1$ yields

$$\mathbb{P}\big(|\!|\!|\widetilde{M}(s, \mathbf{k}) - M^*|\!|\!|_{\text{F}}^2 > c(n-k^*+1)(\log n)^2\big) \leq e^{-(n-k^*+1)(\log n)^2},$$

where $c > 0$ is a universal constant. Integrating this tail bound (and using the fact that the Frobenius norm is bounded as $|\!|\!|\widetilde{M}(s, \mathbf{k}) - M^*|\!|\!|_{\text{F}} \leq n$) gives the claimed result.

4.2.2 Lower bound

We now turn to proving the lower bound in Proposition 1. By re-ordering as necesseary, we may assume without loss of generality that $k_1 \geq \cdots \geq k_s$, so that $k_{\max} = k_1$. The proof relies on the following technical preliminary that establishes a lower bound on the minimax rates of estimation when there are two indifference sets.

Lemma 4. *If there are $s = 2$ indifference sets (say, of sizes $k_1 \geq k_2$), then any estimator $\widehat{M}$ has error lower bounded as*

$$\sup_{M^* \in \mathbb{C}_{SST}(2,(k_1,k_2))} \frac{1}{n^2}\mathbb{E}[|\!|\!|\widehat{M} - M^*|\!|\!|_F^2] \geq c_\ell \frac{n-k_1}{n^2}. \tag{27}$$

See Section 4.2.4 for the proof of this claim.

Let us now complete the proof of the lower bound in Proposition 1. We split the analysis into two cases depending on the size of the largest indifference set.

Case I: First, suppose that $k_1 > \frac{n}{3}$. We then observe that $\mathbb{C}_{\text{SST}}(2, (k_1, n - k_1))$ is a subset of $\mathbb{C}_{\text{SST}}(\mathbf{k})$: indeed, every matrix in $\mathbb{C}_{\text{SST}}(2, (k_1, n - k_1))$ can be seen as a matrix in $\mathbb{C}_{\text{SST}}(\mathbf{k})$ which has identical values in entries corresponding to all items not in the largest indifference set. Since the induced set $\mathbb{C}_{\text{SST}}(2, (k_1, n - k_1))$ is a subset of $\mathbb{C}_{\text{SST}}(\mathbf{k})$, the lower bound for estimating a matrix in $\mathbb{C}_{\text{SST}}(\mathbf{k})$ is at least as large as the lower bound for estimating a matrix in the class $\mathbb{C}_{\text{SST}}(2, (k_1, n - k_1))$. Now applying Lemma 4 to the set $\mathbb{C}_{\text{SST}}(2, (k_1, n - k_1))$ yields a lower bound of $c_\ell \frac{\min\{n-k_1, k_1\}}{n^2}$. Since $k_1 > \frac{n}{3}$, we have $k_1 \geq \frac{n-k_1}{2}$. As a result, we get a lower bound of $\frac{c_\ell}{2} \frac{n-k_1}{n^2}$.

Case II: Alternatively, suppose that $k_1 \leq \frac{n}{3}$. In this case, we claim that there exists a value $u \in [n/3, 2n/3]$ such that $\mathbb{C}_{\text{SST}}(2, (u, n - u))$ is a subset of the set $\mathbb{C}_{\text{SST}}(\mathbf{k})$ with $k_1 \leq \frac{n}{3}$. Observe that for any collection of sets with sizes $\mathbf{k}$ with $k_1 \leq \frac{n}{3}$, there is a grouping of sets into two groups, both of size between $n/3$ and $2n/3$. This is true since the largest set is of size at most $n/3$. Denoting the size of either of these groups as u, we have established our earlier claim.

As in the previous case, we can now apply Lemma 4 to the subset $\mathbb{C}_{\text{SST}}(2, (u, n - u))$ to obtain a lower bound of $c_\ell \frac{1}{3n} \geq c_\ell \frac{n-k_1}{3n^2}$.

4.2.3 Proof of Lemma 3

In order to upper bound the metric entropy of $\widetilde{\mathbb{C}}_{\text{SST}}(k)$, we first separate out the contributions of the permutation and the bivariate monotonicity conditions. Let $\widetilde{\mathbb{C}}_{\text{SST}}(\text{id})(k)$ denote the subset of matrices in $\widetilde{\mathbb{C}}_{\text{SST}}(k)$ that are faithful to the identity permutation. With this notation, the ϵ-metric entropy of $\widetilde{\mathbb{C}}_{\text{SST}}(k^*)$ is upper bounded by the sum of two parts:

(a) the ϵ-metric entropy of the set $\widetilde{\mathbb{C}}_{\text{SST}}(\text{id})(k)$; and

(b) the logarithm of the number of distinct permutations of the n items.

Due to the presence of an indifference set of size at least k, the quantity in (b) is upper bounded by $\log(\frac{n!}{k!}) \leq (n - k) \log n$.

We now upper bound the ϵ-metric entropy of the set $\widetilde{\mathbb{C}}_{\text{SST}}(\text{id})(k)$. We do so by partitioning the n^2 positions in the matrix, computing the ϵ-metric entropy of each partition separately, and then adding up these metric entropies. More precisely, letting $S_k \subseteq [n]$ denote some set of k items that belong to the same indifference set, let us partition the entries of each matrix into four sub-matrices as follows:

(i) The $(k \times k)$ sub-matrix comprising entries (i, j) where both $i \in S_k$ and $j \in S_k$;

(ii) the $(k \times (n - k))$ sub-matrix comprising entries (i, j) where $i \in S_k$ and $j \in [n] \backslash S_k$;

(iii) $((n - k) \times k)$ sub-matrix comprising entries (i, j) where $i \in [n] \backslash S_k$ and $j \in S_k$; and

(iv) the $((n-k)\times(n-k))$ sub-matrix comprising entries (i,j) where both $i \in [n]\backslash S_k$ and $j \in [n]\backslash S_k$.

By construction, the metric entropy of $\widetilde{\mathbb{C}}_{\text{SST}}(\text{id})(k)$ is at most the sum of the metric entropies of these sub-matrices.

The set of sub-matrices in (i) comprises only constant matrices, and hence its metric entropy is at most $\log \frac{n}{\epsilon}$. Any sub-matrix from set (ii) has constant-valued columns, and so the metric entropy of this set is upper bounded by $(n-k)\log\frac{n}{\epsilon}$. An identical bound holds for the set of sub-matrices in (iii). Finally, the set of sub-matrices in (iv) are all contained in the set of all $((n-k)\times(n-k))$ SST matrices. The metric entropy of the SST class is analyzed in Theorem 1 of our past work [SBGW15], where we showed that the metric entropy of this set is at most $2\left(\frac{n-k}{\epsilon}\right)^2\left(\log\frac{n-k}{\epsilon}\right)^2 + (n-k)\log n$. Summing up each of these metric entropies, some algebraic manipulations yield the claimed result.

4.2.4 Proof of Lemma 4

For the first part of the proof, we assume k_2 is greater than a universal constant. (See the analysis of Case 2 below for how to handle small values of k_2.) Under this condition, the Gilbert-Varshamov bound [Gil52, Var57] guarantees the existence of a binary code $\mathbb{B}$ of length k_2, minimum Hamming distance c_0k_2, and number of code words $\text{card}(\mathbb{B}) = T = 2^{ck_2}$. (As usual, the quantities c and c_0 are positive numerical constants.)

We now construct a set of T matrices contained within the set $\mathbb{C}_{\text{SST}}(2,(k_1,k_2))$, whose constituents have a one-to-one correspondence with the T codewords of the binary code constructed above. Let items $S=\{1,\ldots,k_1\}$ correspond to the first indifference set, so that the complementary set $S^c := \{k_1+1,\ldots,n\}$ indexes the second indifference set.

Fix some $\delta \in (0,\frac{1}{3}]$, whose precise value is to be specified later. Define the base matrix $M(\mathbf{0})$ with entries

$$M_{ij}(\mathbf{0}) = \begin{cases} \frac{1}{2} & \text{if } i,j\in S \text{ or } i,j \in S^c \\ \frac{1}{2}+\delta & \text{if } i \in S \text{ and } j \in S^c \\ \frac{1}{2}-\delta & \text{if } i \in S^c \text{ and } j\in S. \end{cases}$$

For any other codeword $\mathbf{z}\in\mathbb{B}$, the matrix $M(\mathbf{z})$ is defined by starting with the base matrix $M(\mathbf{0})$, and then swapping row/column i with row/column (k_1+i) if and only if $\mathbf{z}_i = 1$. For instance, if the codeword is $\mathbf{z} = [1\ 1\ 0\ \cdots\ 0]$, then the new ordering in the matrix $M(\mathbf{z})$ is given by $(k_1+1),(k_1+2),3,\ldots,k_1,1,2,(k_1+3),\ldots,n$, which is obtained by swapping the first two items of the two indifference sets.

We have thus constructed a set of T matrices that are contained within the set $\mathbb{C}_{\text{SST}}(2,(k_1,k_2))$. We now evaluate certain properties of these matrices which will allow us prove the claimed lower bound. Consider any two matrices M_1 and M_2 in this set. Since any two codewords in our binary code have a Hamming distance at least c_0k_2, we have from the aforementioned construction:

$$c_1k_2n\delta^2 \leq |||M_1-M_2|||_{\text{F}}^2 \leq 2k_2n\delta^2,$$

for a constant $c_1 \in (0, 1)$.

Let $\mathbb{P}_{M_1}$ and $\mathbb{P}_{M_2}$ correspond to the distributions of the random matrix Y based on Bernoulli sampling (1) from the matrices M_1 and M_2, respectively. Since $\delta \in (0, \frac{1}{3}]$, all entries of the matrices M_1 and M_2 lie in the interval $[1/3, 2/3]$. Under this boundedness condition, the KL divergence may be sandwiched by the Frobenius norm up to constant factors. Applying this result in the current setting yields

$$c_2 k_2 n \delta^2 \leq D_{\mathrm{KL}}(\mathbb{P}_{M_1} \| \mathbb{P}_{M_2}) \leq c_3 k_2 n \delta^2,$$

again for positive universal constants c_2 and c_3. An application of Fano's inequality to this set gives that the error incurred by any estimator $\widehat{M}$ is lower bounded as

$$\sup_{M^* \in \mathbb{C}_{\mathrm{SST}}(2,(k_1,k_2))} \mathbb{E}[\|\!|\widehat{M} - M^*|\!\|_{\mathrm{F}}^2] \geq \frac{c_1 k_2 n \delta^2}{2} \left(1 - \frac{c_3 k_2 n \delta^2 + \log 2}{c k_2}\right). \tag{28}$$

From this point, we split the remainder of the analysis into two cases.

Case 1: First suppose that k_2 is larger than some suitably large (but still universal) constant. In this case, we may set $\delta^2 = \frac{c''}{n}$ for a small enough universal constant c'', and the Fano bound (28) then implies that

$$\sup_{M^* \in \mathbb{C}_{\mathrm{SST}}(2,(k_1,k_2))} \mathbb{E}[\|\!|\widehat{M} - M^*|\!\|_{\mathrm{F}}^2] \geq c' k_2,$$

for some universal constant $c' > 0$. Since $k_2 = n - k_1$, this completes the proof the claimed lower bound (27) in this case.

Case 2: Otherwise, the parameter k_2 is smaller than the universal constant in the above part of the proof. In this case, the claimed lower bound (27) on $\mathbb{E}[\|\!|\widehat{M} - M^*|\!\|_{\mathrm{F}}^2]$ is just a constant, and we can handle this case with a different argument. In particular, suppose that the estimator is given partition forming the two indifference sets, and only needs to estimate the parameter δ. For this purpose, the sufficient statistics of the observation matrix Y are those entries of the observation matrix that correspond to matches between two items of different indifference sets; note that there are $k_1 k_2$ such entries in total. From standard bounds on estimation of a single Bernoulli probability, any estimator $\hat{\delta}$ of δ must have mean-squared error lower bounded as $\mathbb{E}[(\delta - \hat{\delta})^2] \geq \frac{c}{k_1 k_2}$. Finally, observe that the error in estimating the matrix M^* in the squared Frobenius norm is at least $2k_1 k_2$ times the error in estimating the parameter δ. We have thus established the claimed lower bound of a constant.

4.3 Proof of Theorem 1

We now prove the upper bound (10a) for the regularized least squares estimator (9). Note that it has the equivalent representation

$$\widehat{M}_{\mathrm{REG}} \in \arg\min_{M \in \mathbb{C}_{\mathrm{SST}}} \Big\{ \|\!|Y - M|\!\|_{\mathrm{F}}^2 + (n - k_{\max}(M) + 1)(\log n)^3 \Big\}. \tag{29}$$

Defining $k^* := k_{\max}(M^*)$, it is also convenient to consider the family of estimators

$$\widehat{M}_k \in \arg\min_{M \in \mathbb{C}_{\text{SST}}(k) \cup \mathbb{C}_{\text{SST}}(k^*)} \Big\{ ⦀Y - M⦀_{\text{F}}^2 + (n - k_{\max}(M) + 1)(\log n)^3 \Big\}, \tag{30}$$

where k ranges over $[n]$. Note that these estimators cannot be computed in practice (since the value of k^* is unknown), but they are convenient for our analysis, in particular because $\widehat{M}_{\text{REG}} = \widehat{M}_k$ for some value $k \in [n]$.

We first show that there exists a universal constant $c_0 > 0$ such that

$$\mathbb{P}\Big[⦀\widehat{M}_k - M^*⦀_{\text{F}}^2 > c_0(n - k^* + 1)(\log n)^3\Big] \leq e^{-(\log n)^2} \tag{31}$$

for each fixed $k \in [n]$. Since $\widehat{M}_{\text{REG}} = \widehat{M}_k$ for some k, we then have

$$\begin{aligned} \mathbb{P}\Big[⦀\widehat{M}_{\text{REG}} - M^*⦀_{\text{F}}^2 > c_0(n - k^* + 1)(\log n)^3\Big] &\leq \mathbb{P}\Big[\max_{k \in [n]} ⦀\widehat{M}_k - M^*⦀_{\text{F}}^2 > c_0(n - k^* + 1)(\log n)^3\Big] \\ &\overset{(i)}{\leq} ne^{-(\log n)^2} \;\leq\; e^{-\frac{1}{2}(\log n)^2}. \end{aligned}$$

where step (i) follows from the union bound. We have thus established the claimed tail bound (10a).

In order to prove the bound (10b) on the adaptivity index, we first integrate the tail bound (10a). Since all entries of M^* and $\widehat{M}_{\text{REG}}$ all lie in $[0,1]$, we have $⦀M^* - \widehat{M}_{\text{REG}}⦀_{\text{F}}^2 \leq n^2$, and so this integration step yields an analogous bound on the expected error:

$$\mathbb{E}[⦀M^* - \widehat{M}_{\text{REG}}⦀_{\text{F}}^2] \leq c_u(n - k_{\max}(M^*) + 1)(\log n)^3.$$

Coupled with the lower bound on the risk of the oracle estimator established in Proposition 1, we obtain the claimed bound (10b) on the adaptivity index of $\widehat{M}_{\text{REG}}$.

It remains to prove the tail bound (31). We proceed via a two step argument: first we use the general upper bound given by Lemma 1 to derive a weaker version of the required bound; and second, we then refine this weaker bound so as to obtain the bound (31).

Establishing a weaker bound: Beginning with the first step, let us apply Lemma 1 with the choices

$$b = 0, \quad \mathbb{C} = \mathbb{C}_{\text{SST}}(k) \cup \mathbb{C}_{\text{SST}}(k^*), \quad \text{and} \quad \lambda(M) = (n - k_{\max}(M) + 1)(\log n)^3.$$

With these choices, the $\mathbb{C}_{\text{DIFF}}(M^*, t)$ in the statement of Lemma 1 takes the form

$$\mathbb{C}_{\text{DIFF}}(M^*, t) \subseteq \{\alpha(M - M^*) \mid \alpha \in [0,1], M \in \mathbb{C}_{\text{SST}}(k) \cup \mathbb{C}_{\text{SST}}(k^*)\}.$$

Lemma 3 implies that

$$\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*, t), ⦀.⦀_{\text{F}}) \leq c\frac{(n - \min\{k, k^*\} + 1)^2}{\epsilon^2}(\log n)^2 + c(n - \min\{k, k^*\} + 1)\log n$$

for all $\epsilon \geq n^{-8}$. Applying Lemma 1 with $u = 1$ then yields

$$\mathbb{P}\Big(⦀\widehat{M}_k - M^*⦀_{\text{F}}^2 > c(n - \min\{k, k^*\} + 1)(\log n)^2\Big) \leq e^{-c(\log n)^2}. \tag{32}$$

Note that this bound is weaker than the desired bound (31), since $\min\{k, k^*\} \leq k^*$. Thus, our next step is to refine it.

Refining the bound (32)**:** Before proceeding with the proof, we must take care of one subtlety. Recall that the set $\mathbb{C}_{\text{SST}}(k^*)$ consists of all matrices in $\mathbb{C}_{\text{SST}}$ that have an indifference set containing at least (but not necessarily exactly) k^* items. If $k \geq k^*$, then the bound (32) is equivalent to the bound (31). Otherwise, we evaluate the estimator $\widehat{M}_k$ for the choices $k = 1, \ldots, k^* - 1$ (in this particular order). For any $k \in \{1, \ldots, k^* - 1\}$ under consideration, suppose $k_{\max}(\widehat{M}_k) = k' < k$. Then the estimate under consideration is either also an optimal estimator for the case of $M_{k'}$, or it is suboptimal for the aggregate estimation problem (29). In the former case, the error incurred by this estimate is already handled in the analysis of $M_{k'}$, and in the latter case, it is irrelevant. Consequently, it suffices to evaluate the case when $k_{\max}(\widehat{M}_k) = k$.

Observe that the matrix $\widehat{M}_k$ is optimal for the optimization problem (30) and the matrix M^* lies in the feasible set. Consequently, we have the basic inequality:

$$|\!|\!| Y - \widehat{M}_k |\!|\!|_{\text{F}}^2 + (n - k + 1)(\log n)^3 \leq |\!|\!| Y - M^* |\!|\!|_{\text{F}}^2 + (n - k^* + 1)(\log n)^3.$$

Using the linearized form of the observation model (16a), some simple algebraic manipulations give

$$\frac{1}{2} |\!|\!| \widehat{M}_k - M^* |\!|\!|_{\text{F}}^2 \leq \langle\!\langle \widehat{M}_k - M^*, \; W \rangle\!\rangle - (n - k + 1)(\log n)^3 + (n - k^* + 1)(\log n)^3, \tag{33}$$

where W is the noise matrix (16b) in the linearized form of the model. The following lemma helps bound the first term on the right hand side of inequality (33). Consistent with the notation elsewhere in the paper, for any value of $t > 0$, let us define a set of matrices $\mathbb{C}_{\text{DIFF}}(M^*, t) \subseteq \mathbb{C}_{\text{SST}}$ as

$$\mathbb{C}_{\text{DIFF}}(M^*, t) := \{\alpha(M - M^*) \mid M \in \mathbb{C}_{\text{SST}}(k), \; \alpha \in [0, 1], \; |\!|\!| \alpha(M - M^*) |\!|\!|_{\text{F}} \leq t\}.$$

With this notation, we then have the following result:

Lemma 5. *For any $M^* \in \mathbb{C}_{SST}$, any fixed $k \in [n]$, and any $t > 0$, we have*

$$\sup_{D \in \mathbb{C}_{DIFF}(M^*, t)} \langle\!\langle D, \; W \rangle\!\rangle \leq ct\sqrt{(n - \min\{k, k^*\} + 1)} \log n + c(n - \min\{k, k^*\} + 1)(\log n)^2 \tag{34}$$

with probability at least $1 - e^{-(\log n)^2}$.

See Section 4.3.1 for the proof of this lemma.

From our weaker guarantee (32), we know that $|\!|\!| \widehat{M}_k - M^* |\!|\!|_{\text{F}} \leq c' \sqrt{(n - \min\{k, k^*\} + 1)(\log n)^2}$, with high probability. Consequently, the term $\langle\!\langle \widehat{M}_k - M^*, \; W \rangle\!\rangle$ is upper bounded by the quantity (34) for some value of $t \leq c' \sqrt{(n - \min\{k, k^*\} + 1)(\log n)^2}$, and hence

$$\langle\!\langle \widehat{M}_k - M^*, \; W \rangle\!\rangle \leq c''(n - \min\{k, k^*\} + 1)(\log n)^2,$$

with probability at least $1 - e^{-(\log n)^2}$. Applying this bound to the basic inequality (33) and performing some algebraic manipulations yields the claimed result (31).

4.3.1 Proof of Lemma 5

Consider the function $\zeta : [0,n] \to \mathbb{R}_+$ given by $\zeta(t) := \sup_{D \in \mathbb{C}_{\text{DIFF}}(M^*,t)} \langle\!\langle D,\ W \rangle\!\rangle$. In order to control the behavior of this function, we first bound the metric entropy of the set $\mathbb{C}_{\text{DIFF}}(M^*,t)$. Note that Lemma 3 ensures that

$$\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*,t), \|\!| \cdot |\!\|_{\text{F}}) \leq c\frac{(n - \min\{k,k^*\}+1)^2}{\epsilon^2}\big(\log\frac{n}{\epsilon}\big)^2 + c(n - \min\{k,k^*\}+1)\log n.$$

Based on this metric entropy bound, the truncated version of Dudley's entropy integral then guarantees that

$$\mathbb{E}[\zeta(t)] \leq c(n - \min\{k,k^*\}+1)(\log n)^2 + ct\sqrt{(n - \min\{k,k^*\}+1)\log n}.$$

It can be verified that the function $\zeta(t)$ is t-Lipschitz. Moreover, the random matrix W has entries (16b) that are independent on and above the diagonal, bounded by 1 in absolute value, and satisfy skew-symmetry. Consequently, Ledoux's concentration theorem [Led01, Theorem 5.9] guarantees that

$$\mathbb{P}\big[\zeta(t) \geq \mathbb{E}[\zeta(t)] + tv\big] \leq e^{-v^2} \qquad \text{for all } v \geq 0.$$

Combining the pieces, we find that

$$\mathbb{P}\Big[\zeta(t) \geq c(n - \min\{k,k^*\}+1)(\log n)^2 + ct\sqrt{(n - \min\{k,k^*\}+1)\log n} + tv\Big] \leq e^{-v^2},$$

valid for all $v \geq 0$. Setting $v = \sqrt{(n - \min\{k,k^*\}+1)}\log n$ yields the claimed result.

4.4 Proof of Theorem 2

We now prove the upper bound for the CRL estimator, as stated in Theorem 2. In order to simplify the presentation, we assume without loss of generality that the true permutation of the n items is the identity permutation id. Let $\pi_{\text{CRL}} = (\pi_1,\ldots,\pi_n)$ denote the permutation obtained at the end of the second step of the CRL estimator. The following lemma proves two useful properties of the outcomes of the first two steps.

Lemma 6. *With probability at least $1 - n^{-20}$, the permutation π_{CRL} obtained at the end of the first two steps of the estimator satisfies the following two properties:*

(a) $\max_{i\in[n]} \sum_{\ell=1}^n |M^*_{i\ell} - M^*_{\pi_{CRL}(i)\ell}| \leq \sqrt{n}(\log n)^2$, and

(b) the group of similar items obtained in the first step is of size at least $k^* = k_{\max}(M^*)$.

See Section 4.4.1 for the proof of this claim.

Given Lemma 6, let us complete the proof of the theorem. Let $\widehat{\Pi}$ denote the set of all permutations on n items which satisfy the two conditions (a) and (b) stated in Lemma 6. Given that every entry of M^* lies in the interval $[0,1]$, any permutation $\hat{\pi} \in \widehat{\Pi}$ satisfies

$$\|\!| M^* - \hat{\pi}(M^*) |\!\|_{\text{F}}^2 = \sum_{i\in[n]}\sum_{\ell\in[n]} (M^*_{i\ell} - M^*_{\hat{\pi}(i)\ell})^2 \ \leq\ \sum_{i\in[n]}\sum_{\ell\in[n]} \mid M^*_{i\ell} - M^*_{\hat{\pi}(i)\ell} \mid. \tag{35}$$

Now consider any item $i \in [n]$. Incorrectly estimating item i as lying in position $\hat{\pi}(i)$ contributes a non-zero error only if either item i or item $\hat{\pi}(i)$ lies in the $(n-k^*)$-sized set of items outside the largest indifference set. Consequently, there are at most $2(n-k^*)$ values of i in the sum (35) that make a non-zero contribution. Moreover, from property (a) of Lemma 6, each such item contributes at most $\sqrt{n}(\log n)^2$ to the error. As a consequence, we have the upper bound

$$|||M^* - \hat{\pi}(M^*)|||_{\text{F}}^2 \leq 2(n-k^*)\sqrt{n}(\log n)^2. \tag{36}$$

Let us now analyze the third step of the CRL estimator. The problem of bivariate isotonic regression refers to estimation of the matrix $M^* \in \mathbb{C}_{\text{SST}}$ when the true underlying permutation of the items *is known* a priori. In our case, the permutation is known only approximately, so that we need also to track the associated approximation error. In order to derive a tail bound on the error of bivariate isotonic regression, we call upon the general upper bound proved earlier in Lemma 1 with the choices $b = 1$, $\mathbb{C} = \mathbb{C}_{\text{SST}}(\text{id})$, and $\lambda = 0$. Now let

$$\mathbb{C}_{\text{DIFF}}(M^*, t) := \{\alpha(M - M^*) \mid M \in \mathbb{C}_{\text{SST}}(\text{id})\}.$$

The following lemma uses a result from the paper [CGS15] to derive an upper bound on the metric entropy of $\mathbb{C}_{\text{DIFF}}(M^*, t)$. For any matrix $M^* \in \mathbb{C}_{\text{SST}}$, let $s(M^*)$ denote the number of indifference sets in M^*.

Lemma 7. *For every $\epsilon > n^{-8}$ and $t \in (0, n]$, we have the metric entropy bound*

$$\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*, t), |||.|||_F) \leq c\frac{t^2(s(M^*))^2(\log n)^6}{\epsilon^2}.$$

where $c > 0$ is a universal constant.

With this bound on the metric entropy, an application of Lemma 1 with $u = \frac{(n-k^*+1)^2}{(s(M^*))^2}$ gives that for every $M^* \in \mathbb{C}_{\text{SST}}(\text{id})$, the least squares estimator $\widehat{M}_{\text{id}} \in \underset{M \in \mathbb{C}_{\text{SST}}(\text{id})}{\arg\min} |||M - Y|||_{\text{F}}^2$ incurs an error upper bounded as

$$|||\widehat{M}_{\text{id}} - M^*|||_{\text{F}}^2 \leq c(n-k^*+1)^2(\log n)^8,$$

with probability at least $1 - e^{-(n-k^*+1)^2(\log n)^8}$. Note that this application of Lemma 1 is valid since $s(M^*) \leq n - k^* + 1$ and hence $u \geq 1$. Furthermore, it follows from a corollary of Theorem 1 in the paper [SBGW15] that

$$|||\widehat{M}_{\text{id}} - M^*|||_{\text{F}}^2 \leq cn(\log n)^2,$$

with probability at least $1 - e^{-cn}$. Combining these upper bounds yields

$$|||\widehat{M}_{\text{id}} - M^*|||_{\text{F}}^2 \leq c\min\{(n-k^*+1)^2, n\}(\log n)^8 \overset{(i)}{\leq} c(n-k^*+1)\sqrt{n}(\log n)^8, \tag{37}$$

with probability at least $1 - e^{-c(n-k^*+1)^2(\log n)^8}$, where c is a positive universal constant. Inequality (i) makes use of the bound $\min\{u^2, v^2\} \leq uv$ for any two non-negative numbers u and v.

Let us put together the analysis of the approximation error (36) in the permutation obtained in the first two steps and the error (37) in estimating the matrix in the third step. To this end, consider any permutation $\hat{\pi} \in \widehat{\Pi}$. For clarity, we augment the notation of $\widehat{M}_{\text{CRL}}$ (defined in (12)) and use $\widehat{M}_{\text{CRL}}(Y, \hat{\pi})$ to represent the estimator $\widehat{M}_{\text{CRL}}$ under the permutation $\hat{\pi}$ for the observation matrix Y, that is,

$$\widehat{M}_{\text{CRL}}(Y, \hat{\pi}) := \underset{M \in \mathbb{C}_{\text{SST}}(\hat{\pi})}{\arg\min} \ \|M - Y\|_{\text{F}}^2.$$

Consider any matrix $M^* \in \mathbb{C}_{\text{SST}}(\text{id})$ under the identity permutation. We can then write

$$\begin{aligned} &\|\widehat{M}_{\text{CRL}}(M^* + W, \hat{\pi}) - M^*\|_{\text{F}}^2 \\ &= \|\widehat{M}_{\text{CRL}}(M^* + W, \hat{\pi}) - \widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi}) + \widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi}) - M^*\|_{\text{F}}^2 \\ &\leq 2\|\widehat{M}_{\text{CRL}}(M^* + W, \hat{\pi}) - \widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi})\|_{\text{F}}^2 + 2\|\widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi}) - M^*\|_{\text{F}}^2. \end{aligned} \tag{38}$$

We separately bound the two terms on the right hand side of equation (38). First observe that the least squares step of the estimator $\widehat{M}_{\text{CRL}}$ (for a given permutation $\hat{\pi}$ in its second argument) is a projection onto the convex set $\mathbb{C}_{\text{SST}}(\hat{\pi})$, and hence we have the deterministic bound

$$\|\widehat{M}_{\text{CRL}}(M^* + W, \hat{\pi}) - \widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi})\|_{\text{F}}^2 \leq \|M^* - \hat{\pi}(M^*)\|_{\text{F}}^2. \tag{39a}$$

In addition, we have

$$\|\widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi}) - M^*\|_{\text{F}}^2 \leq 2\|\widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi}) - \hat{\pi}(M^*)\|_{\text{F}}^2 + 2\|\hat{\pi}(M^*) - M^*\|_{\text{F}}^2. \tag{39b}$$

From our earlier bound (37), we have that for any *fixed* permutation $\hat{\pi} \in \widehat{\Pi}$, the least squares estimate satisfies

$$\|\widehat{M}_{\text{CRL}}(\hat{\pi}(M^*) + W, \hat{\pi}) - \hat{\pi}(M^*)\|_{\text{F}}^2 \leq c_u (n - k^* + 1)\sqrt{n}(\log n)^8, \tag{40}$$

with probability at least $1 - e^{-c(n-k^*+1)^2(\log n)^8}$.

In conjunction, the bounds (36), (38), (39a), (39b) and (40) imply that for any *fixed* $\hat{\pi} \in \widehat{\Pi}$,

$$\mathbb{P}\Big(\|\widehat{M}_{\text{CRL}}(M^* + W, \hat{\pi}) - M^*\|_{\text{F}}^2 \leq c_u (n - k^* + 1)\sqrt{n}(\log n)^8\Big) \geq 1 - e^{-c(n-k^*+1)^2(\log n)^8}. \tag{41}$$

Although we are guaranteed that $\pi_{\text{CRL}} \in \widehat{\Pi}$, we cannot apply the bound (41) directly to it, since π_{CRL} is a data-dependent quantity. In order to circumvent this issue, we need to obtain a uniform version of the bound (41), and we do so by applying the union bound over the data-dependent component of π_{CRL}.

In more detail, let us consider Steps 1 and 2 of the CRL algorithm as first obtaining a total ordering of the n items via a count of the number of pairwise victories, then converting it to a partial order by putting all items in the subset identified by Step 2 in an equivalence class, and then obtaining a total ordering by permuting the items in the equivalence class in a *data-independent* manner. Lemma 6 ensures that the size of this equivalence class is at least k^*. Consequently, the number of possible (data-dependent) partial orders obtained is at most $\frac{n!}{k^*!} \leq e^{(n-k^*)\log n}$. Taking a union bound over each of these $e^{(n-k^*)\log n}$ cases, we get that

$$\mathbb{P}\Big[|\!|\!|\widehat{M}_{\text{CRL}}(M^* + W, \pi_{\text{CRL}}) - M^*|\!|\!|_{\text{F}}^2 \leq c_u(n - k^* + 1)\sqrt{n}(\log n)^8 \mid \pi_{\text{CRL}} \in \widehat{\Pi}\Big] \quad \geq 1 - e^{-(\log n)^7}.$$

Recalling that Lemma 6 ensures that $\mathbb{P}\big[\pi_{\text{CRL}} \in \widehat{\Pi}\big] \geq 1 - n^{-20}$, we have established the claim.

It remains to prove the two auxiliary lemmas stated above.

4.4.1 Proof of Lemma 6

We first prove that for any fixed item $i \in [n]$, the inequality of part (a) holds with probability at least $1 - n^{-22}$. The claimed result then follows via a union bound over all items.

Consider any item $j > i$ such that

$$\sum_{\ell=1}^{n} M^*_{i\ell} - \sum_{\ell=1}^{n} M^*_{j\ell} > \sqrt{n}(\log n)^2. \tag{42}$$

An application of the Bernstein inequality then gives (see the proof of Theorem 1 in the paper [SW15] for details) that

$$\mathbb{P}\big(\sum_{\ell=1}^{n} Y_{j\ell} \geq \sum_{\ell=1}^{n} Y_{i\ell}\big) \leq \frac{1}{n^{23}}.$$

Likewise, for any item $j < i$ such that $\sum_{\ell=1}^{n} M^*_{j\ell} - \sum_{\ell=1}^{n} M^*_{i\ell} > \sqrt{n}(\log n)^2$, we have $\mathbb{P}\big(\sum_{\ell=1}^{n} Y_{i\ell} \geq \sum_{\ell=1}^{n} Y_{j\ell}\big) \leq \frac{1}{n^{23}}$.

Now consider any $j \geq i$. In order for item i to be located in position j in the total order given by the row sums, there must be at least $(j - i)$ items in the set $\{i + 1, \ldots, n\}$ whose row sums are at least as big as the sum of the i^{th} row of Y. In particular, there must be at least one item in the set $\{j, \ldots, n\}$ such that its row sum is as big as the sum of the i^{th} row of Y. It follows from our results above that under the condition (42), this event occurs with probability no more than $\frac{1}{n^{21}}$. Likewise when $j \leq i$, thereby proving the claim.

We now move to the condition of part (b). Observe that for any two items i and j in the same indifference set, we have that $M^*_{i\ell} = M^*_{j\ell}$ for every $\ell \in [n]$. An application of the Bernstein inequality now gives that

$$\mathbb{P}\big(\sum_{\ell=1}^{n} Y_{j\ell} - \sum_{\ell=1}^{n} Y_{i\ell} \geq \sqrt{n}\log n\big) \leq \frac{1}{n^{23}}.$$

A union bound over all pairs of items in the largest indifference set gives that all k^* items in the largest indifference set have their row sums differing from each other by at most $\sqrt{n}\log n$. Consequently, the group must be of at least this size.

4.4.2 Proof of Lemma 7

For the proof, it is be convenient to define a class $\mathbb{C}_{\text{SST}}(;$ [-1,1]) that is similar to the class $\mathbb{C}_{\text{SST}}(\text{id})$, but contains matrices with entries in $[-1,1]$:

$$\mathbb{C}_{\text{SST}}(t;[-1,1]) := \big\{M \in [-1,1]^{n\times n} \mid |||M|||_{\text{F}} \leq t,\ M_{k\ell} \geq M_{ij} \text{ whenever } k \leq i \text{ and } \ell \geq j\big\}.$$

We now call upon Theorem 3.3 of the paper [CGS15]. It provides the following upper bound on the metric entropy of bivariate isotonic matrices within a Frobenius ball:

$$\log N(\epsilon, \mathbb{C}_{\text{SST}}(t;[-1,1]), |||\cdot|||_{\text{F}}) \leq c_0 \frac{t^2(\log n)^4}{\epsilon^2}\big(\log\frac{t\log n}{\epsilon}\big)^2.$$

Substituting $\epsilon \geq n^{-8}$ and $t \leq n$ yields

$$\log N(\epsilon, \mathbb{C}_{\text{SST}}(t;[-1,1]), |||\cdot|||_{\text{F}}) \leq c\frac{t^2(\log n)^6}{\epsilon^2}. \tag{43}$$

We now use this result to derive an upper bound on the metric entropy of the set $\mathbb{C}_{\text{DIFF}}(M^*,t)$. Consider the following partition of the entries of any $(n\times n)$ matrix into $(s(M^*))^2$ submatrices. Submatrix $(i,j) \in [s(M^*)]\times[s(M^*)]$ in this partition is the $(k_i \times k_j)$ submatrix corresponding to the pairwise comparison probabilities between every item in the i^{th} indifference set with every item in the j^{th} indifference set in M^*. Such a partition ensures that each partitioned submatrix of M^* is a constant matrix. Consequently, for any $M \in \mathbb{C}_{\text{DIFF}}(M^*,t)$, each partitioned submatrix belongs to the set of matrices $\mathbb{C}_{\text{SST}}(t;[-1,1])$ (where we slightly abuse notation to ignore the size of the matrices as long as no dimension is greater than $(n\times n)$). The metric entropy of the set of matrices in $\mathbb{C}_{\text{DIFF}}(M^*,t)$ can now be upper bounded by the sum of the metric entropies of each set of submatrices. Consequently, we have

$$\begin{aligned}\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(M^*,t), |||\cdot|||_{\text{F}}) &\leq (s(M^*))^2\log N(\epsilon, \mathbb{C}_{\text{SST}}(t;[-1,1]), |||\cdot|||_{\text{F}})\\ &\leq c\frac{t^2(s(M^*))^2(\log n)^6}{\epsilon^2},\end{aligned}$$

where the final inequality follows from our earlier bound (43).

4.5 Proof of Theorem 3

We now turn to the proof of the lower bound for polynomial-time computable estimators, as stated in Theorem 3. We proceed via a reduction argument. Consider any estimator that has Frobenius norm error upper bounded as

$$\sup_{M^*\in\mathbb{C}_{\text{SST}}} \mathbb{E}[|||\widehat{M}-M^*|||_{\text{F}}^2] \leq c_u\sqrt{n}(n-k_{\max}(M^*)+1)(\log n)^{-1}. \tag{44}$$

We show that any such estimator defines a method that, with probability at least $1-\frac{1}{\sqrt{\log n}}$, is able to identify the presence or absence a planted clique with $\frac{\sqrt{n}}{\log\log n}$ vertices. This result, coupled with the upper bound on the risk of the oracle estimator established in Proposition 1 proves the claim of Theorem 3.

Our reduction from the bound (44) proceeds by identifying a subclass of $\mathbb{C}_{\text{SST}}$, and showing that any estimator satisfying the bound (44) on on this subclass can be used to identify a planted clique in an Erdős-Rényi random graph. Naturally, in order to leverage the planted clique conjecture, we need the planted clique to be of size $o(\sqrt{n})$.

Our construction involves a partition with $s = 3$ components, maximum indifference set size $k_{\max} = k_1 = n - 2k$, with the remaining two indifference sets of size $k_2 = k_3 = k$. We choose the parameter $k := \frac{\sqrt{n}}{\log \log n}$ so that any constant multiple of it will be within the hardness regime of planted clique (for sufficiently large values of n). Now let M_0^* be a matrix with all ones in the $(k \times k)$ sub-matrix in its top-right, zeros on the corresponding sub-matrix in the bottom-left and all other entries set equal to $\frac{1}{2}$. By construction, the matrix M_0^* belongs to the class $\mathbb{C}_{\text{SST}}(k_{\max})$ with $k_{\max} = n - 2k$.

For any permutation π on $\frac{n}{2}$ items and any $(n \times n)$ matrix M, define another $(n \times n)$ matrix $P_\pi(M)$ by applying the permutation π to:

- the first $\frac{n}{2}$ rows of M^*, and the last $\frac{n}{2}$ rows of M^*
- the first $\frac{n}{2}$ columns of M^*, and to the last $\frac{n}{2}$ columns of M^*.

We then define the set $\widetilde{\mathbb{C}}_{\text{SST}} := \Big\{ P_\pi(M_0^*) \mid \text{for all permutations } \pi \text{ on } [n/2] \Big\}$. By construction, it is a subset of $\mathbb{C}_{\text{SST}}(n - k_{\max})$.

For any estimator $\widehat{M}$ that satisfies the bound (44), we have

$$\sup_{M^* \in \widetilde{\mathbb{C}}_{\text{SST}} \cup \{\frac{1}{2} 1 1^T\}} \mathbb{E}[|\!|\!| \widehat{M} - M^* |\!|\!|_{\text{F}}^2] \leq ck \frac{\sqrt{n}}{\log n}.$$

On the other hand, Markov's inequality implies that

$$\mathbb{E}[|\!|\!| \widehat{M} - M^* |\!|\!|_{\text{F}}^2] > c \frac{k\sqrt{n}}{\sqrt{\log n}} \mathbb{P}\Big[|\!|\!| \widehat{M} - M^* |\!|\!|_{\text{F}}^2 \geq c \frac{k\sqrt{n}}{\sqrt{\log n}} \Big].$$

Combining the two bounds, we find that

$$\mathbb{P}\Big[|\!|\!| \widehat{M} - M^* |\!|\!|_{\text{F}}^2 < \frac{c\, k\sqrt{n}}{\sqrt{\log n}} \Big] \geq 1 - \frac{1}{\sqrt{\log n}}. \tag{45}$$

Consider the set of $(\frac{n}{2} \times \frac{n}{2})$ matrices comprising the top-right $(\frac{n}{2} \times \frac{n}{2})$ sub-matrix of every matrix in $\widetilde{\mathbb{C}}_{\text{SST}}$. We claim that this set is identical to the set of all possible matrices in the planted clique problem with $\frac{n}{2}$ vertices and a planted clique of size k. Indeed, the set contains the all-half matrix corresponding to the absence of a planted clique, and all symmetric matrices that have all entries equal to half except for a $(k \times k)$ all-ones submatrix corresponding to the planted clique.

Now consider the problem of testing the hypotheses of whether M^* is equal to the all-half matrix ("no planted clique") or if it lies in $\widetilde{\mathbb{C}}_{\text{SST}}$ ("planted clique"). Let us consider a decision rule that declares the absence of a planted clique if $|\!|\!| \widehat{M} - \frac{1}{2} 1 1^T |\!|\!|_{\text{F}}^2 \leq \frac{1}{16} k^2$, and the presence of a planted clique otherwise.

Null case: On one hand, if there is no planted clique ($M^* = \frac{1}{2}11^T$), then the bound (45) guarantees that

$$|\!|\!|\widehat{M} - \frac{1}{2}11^T|\!|\!|_{\text{F}}^2 < k\frac{\sqrt{n}}{\sqrt{\log n}} \tag{46}$$

with probability at least $1 - \frac{1}{\sqrt{\log n}}$. Recalling that $k = \frac{\sqrt{n}}{\log\log n}$, we find that our decision rule can detect the absence of the planted clique with probability at least $1 - \frac{1}{\sqrt{\log n}}$.

Case of planted clique: On the other hand, if there is a planted clique ($M^* \in \widetilde{\mathbb{C}}_{\text{SST}}$), then we have

$$|\!|\!|\widehat{M} - \frac{1}{2}11^T|\!|\!|_{\text{F}}^2 \geq \frac{1}{2}|\!|\!|M^* - \frac{1}{2}11^T|\!|\!|_{\text{F}}^2 - |\!|\!|\widehat{M} - M^*|\!|\!|_{\text{F}}^2 = \frac{1}{4}k^2 - |\!|\!|\widehat{M} - M^*|\!|\!|_{\text{F}}^2.$$

Thus, in this case, the bound (45) guarantees that

$$|\!|\!|\widehat{M} - \frac{1}{2}11^T|\!|\!|_{\text{F}}^2 \geq \frac{1}{4}k^2 - \frac{k\sqrt{n}}{\sqrt{\log n}},$$

with probability at least $1 - \frac{1}{\sqrt{\log n}}$. Since $k = \frac{\sqrt{n}}{\log\log n}$, our decision rule successfully detects the presence of a planted clique with probability at least $1 - \frac{1}{\sqrt{\log n}}$.

In summary, given the planted clique conjecture, our decision rule cannot be computed in polynomial time. Since it can be computed in polynomial-time given the estimator $\widehat{M}$, it must also be the case that $\widehat{M}$ cannot be computed in polynomial time, as claimed.

4.6 Proof of Theorem 4

We now prove lower bounds on the standard least-squares estimator. A central piece in our proof is the following lemma, which characterizes an interesting structural property of the least-squares estimator.

Lemma 8. *Let $M^* = \frac{1}{2}11^T$ and consider any matrix $Y \in \{0,1\}^{n \times n}$ satisfying the shifted-skew-symmetry condition. Then the least squares estimator $\widehat{M}_{LS}$ from equation* (14) *must satisfy the quadratic equation*

$$|\!|\!|Y - M^*|\!|\!|_F^2 = |\!|\!|Y - \widehat{M}_{LS}|\!|\!|_F^2 + |\!|\!|M^* - \widehat{M}_{LS}|\!|\!|_F^2.$$

See Section 4.6.1 for the proof of this claim.

Let us now complete the proof of Theorem 4 using Lemma 8. Our strategy is as follows: we first construct a "bad" matrix $\widetilde{M} \in \mathbb{C}_{\text{SST}}$ that is far from M^* but close to Y. We then use Lemma 8 to show that the least squares estimate $\widehat{M}_{LS}$ must also be far from M^*.

In the matrix Y, let item ℓ be an item that has won the maximum number of pairwise comparisons—that is $\ell \in \arg\max_{i \in [n]} \sum_{j=1}^{n} Y_{ji}$. Let S denote the set of all items that are beaten by item ℓ—that is, $S := \{j \in [n]\backslash\{\ell\} \mid Y_{\ell j} = 1\}$. Note that $\text{card}(S) \geq \frac{n-1}{2}$. Now define a matrix $\widetilde{M} \in \mathbb{C}_{\text{SST}}$ with entries $\widetilde{M}_{\ell,j} = 1 = 1 - \widetilde{M}_{j,\ell}$ for every $j \in S$, and all remaining entries equal to $\frac{1}{2}$. Some simple calculations then give

$$\|\!|Y - M^*|\!\|_{\text{F}}^2 = \|\!|Y - \widetilde{M}|\!\|_{\text{F}}^2 + \|\!|M^* - \widetilde{M}|\!\|_{\text{F}}^2, \qquad \text{and} \tag{47a}$$

$$\|\!|\widetilde{M} - M^*|\!\|_{\text{F}}^2 \geq \frac{n-1}{4}. \tag{47b}$$

Next we exploit the structural property of the least squares solution guaranteed by Lemma 8. Together with the conditions (47) and the fact that $\|\!|Y - \widehat{M}_{LS}|\!\|_{\text{F}}^2 \leq \|\!|Y - \widetilde{M}|\!\|_{\text{F}}^2$, some simple algebraic manipulations yield the lower bound

$$\|\!|M^* - \widehat{M}_{LS}|\!\|_{\text{F}}^2 \geq \frac{n-1}{4}. \tag{48}$$

This result holds for any arbitrary observation matrix Y, and consequently, holds with probability 1 when the observation matrix Y is drawn at random. For $k_{\max} = n - 1$, Proposition 1 yields an upper bound of $c(\log n)^2$ on the oracle risk. Combining this upper bound with the lower bound (48) yields the claimed lower bound on the adaptivity index of the least squares estimator.

4.6.1 Proof of Lemma 8

From our earlier construction of $\widetilde{M}$ in Section 4.6, we know that $\|\!|Y - \widehat{M}_{LS}|\!\|_{\text{F}} \leq \|\!|Y - \widetilde{M}|\!\|_{\text{F}} < \|\!|Y - M^*|\!\|_{\text{F}}$, which guarantees that $\widehat{M}_{LS} \neq M^*$. Consequently, we may consider the line

$$\mathbb{L}(M^*, \widehat{M}_{LS}) := \{\theta M^* + (1-\theta)\widehat{M}_{LS} \mid \theta \in \mathbb{R}\}$$

that passes through the two points M^* and $\widehat{M}_{LS}$. Given this line, consider the auxiliary estimator

$$\widehat{M}_1 := \arg\min_{M \in \mathbb{L}(M^*, \widehat{M}_{LS})} \|\!|Y - M|\!\|_{\text{F}}^2. \tag{49}$$

Since $\widehat{M}_1$ is the Euclidean projection of Y onto this line, it must satisfy the Pythagorean relation

$$\|\!|Y - M^*|\!\|_{\text{F}}^2 = \|\!|Y - \widehat{M}_1|\!\|_{\text{F}}^2 + \|\!|M^* - \widehat{M}_1|\!\|_{\text{F}}^2. \tag{50}$$

Let $\Pi_{[0,1]} : \mathbb{R}^{n \times n} \to [0,1]^{n \times n}$ denote the Euclidean projection of any $(n \times n)$ matrix onto the hypercube $[0,1]^{n \times n}$. This projection actually has a simple closed-form expression: it simply clips every entry of the matrix M to lie in the unit interval $[0,1]$. Since projection onto the convex set $[0,1]^{n \times n}$ is non-expansive, we must have

$$\|\!|Y - \widehat{M}_1|\!\|_{\text{F}}^2 \geq \|\!|\Pi_{[0,1]}(Y) - \Pi_{[0,1]}(\widehat{M}_1)|\!\|_{\text{F}}^2 = \|\!|Y - \Pi_{[0,1]}(\widehat{M}_1)|\!\|_{\text{F}}^2. \tag{51}$$

Here the final equation follows since $Y \in [0,1]^{n \times n}$, and hence $\Pi_{[0,1]}(Y) = Y$.

Furthermore, we claim that $\Pi_{[0,1]}(\widehat{M}_1) \in \mathbb{C}_{\text{SST}}$. In order to prove this claim, first recall that the matrix $\widehat{M}_1$ can be written as $\widehat{M}_1 = (1-\theta)(\widehat{M}_{LS} - \frac{1}{2}11^T) + \frac{1}{2}11^T$ for some $\theta \in \mathbb{R}$, and $\widehat{M}_{LS} \in \mathbb{C}_{\text{SST}}$. Consequently, if $\theta \leq 1$, then the rows/columns of the projected matrix $\Pi_{[0,1]}(\widehat{M}_1)$ obey the same monotonicity conditions as those of $\widehat{M}_{LS}$; conversely, if $\theta > 1$, the rows/columns obey an inverted set of monotonicity conditions, again specified by the rows/columns of $\widehat{M}_1$. Moreover, since the two matrices $\widehat{M}_{LS}$ and $\frac{1}{2}11^T$ satisfy shifted-skew-symmetry, so does the matrix $\widehat{M}_1$. One can further verify that any two real numbers $a \geq b$ must also satisfy the inequalities

$$\min(a,1) \geq \min(b,1), \quad \text{and} \quad \max(a,0) \geq \max(b,0).$$

If in addition, the pair (a,b) satisfy the constraint, $a + b = 1$, then we have $\max(\min(a,1),0) + \max(\min(b,1),0) = 1$. Using these elementary facts, it can be verified that the monotonicity and shifted-skew-symmetry conditions of any matrix are thus retained by the projection $\Pi_{[0,1]}$.

The arguments above imply that $\Pi_{[0,1]}(\widehat{M}_1) \in \mathbb{C}_{\text{SST}}$ and hence the matrix $\Pi_{[0,1]}(\widehat{M}_1)$ is feasible for the optimization problem (14). By the optimality of $\widehat{M}_{LS}$, we must have $|||Y - \Pi_{[0,1]}(\widehat{M}_1)|||_{\text{F}}^2 \geq |||Y - \widehat{M}_{LS}|||_{\text{F}}^2$. Coupled with the inequality (51), we find that

$$|||Y - \widehat{M}_1|||_{\text{F}}^2 \geq |||Y - \widehat{M}_{LS}|||_{\text{F}}^2.$$

On the other hand, since $\widehat{M}_{LS}$ is feasible for the optimization problem (49) and $\widehat{M}_1$ is the optimal solution, we must actually have

$$|||Y - \widehat{M}_1|||_{\text{F}}^2 = |||Y - \widehat{M}_{LS}|||_{\text{F}}^2,$$

so that $\widehat{M}_{LS}$ is also optimal for the optimization problem (49). However, the optimization problem (49) amounts to Euclidean projection on to a line, it must have a unique minimizer, which implies that $\widehat{M}_{LS} = \widehat{M}_1$. Substituting this condition in the Pythagorean relation (50) yields the claimed result.

5 Conclusions

We proposed the notion of an adaptivity index to measure the abilities of any estimator to automatically adapt to the intrinsic complexity of the problem. This notion helps to obtain a more nuanced evaluation of any estimator that is more informative than the classical notion of the worst-case error. We provided sharp characterizations of the optimal adaptivity that can be achieved in a statistical (information-theoretic) sense, and that can be achieved by computationally efficient estimators.

The logarithmic factors in our results arise from corresponding logarithmic factors in the metric entropy results of Gao and Wellner [GW07], and understanding their necessity is an open question. In statistical practice, we often desire estimators, that perform well in a variety of different senses. We believe that estimating SST matrices at the minimax-optimal rate in Frobenius norm, as studied in more detail in the paper [SBGW15], is also computationally difficult. We hope to formally establish this in future work. Finally, developing a broader understanding of fundamental limits imposed by computational considerations in statistical problems is an important avenue for continued investigation.

Acknowledgements: This work was partially supported by NSF grant CIF-31712-23800, ONR-MURI grant DOD 002888, and AFOSR grant FA9550-14-1-0016. The work of NBS was supported in part by a Microsoft Research PhD fellowship.

References

[AAK+07] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k-wise and almost k-wise independence. In *ACM STOC*, 2007.

[AS15] E. Abbe and C. Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in Neural Information Processing Systems*, pages 676–684, 2015.

[BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[BDPR84] G. Bril, R. Dykstra, C. Pillers, and T. Robertson. Algorithm as 206: isotonic regression in two independent variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):352–357, 1984.

[Bel16] P. C. Bellec. Adaptive confidence sets in shape restricted regression. *arXiv preprint arXiv:1601.05766*, 2016.

[BM08] M. Braverman and E. Mossel. Noisy sorting without resampling. In *Proc. ACM-SIAM symposium on Discrete algorithms*, pages 268–276, 2008.

[BR13] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, 2013.

[BT52] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, pages 324–345, 1952.

[BW97] T. P. Ballinger and N. T. Wilcox. Decisions, error and heterogeneity. *The Economic Journal*, 107(443):1090–1105, 1997.

[C+11] E. Cator et al. Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli*, 17(2):714–735, 2011.

[Can06] E. J. Candes. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15:257–325, 2006.

[CGS13] S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *arXiv preprint arXiv:1311.3765*, 2013.

[CGS15] S. Chatterjee, A. Guntuboyina, and B. Sen. On matrix estimation under monotonicity constraints. *arXiv:1506.03430*, 2015.

[Cha14] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2014.

[CL11] T. T. Cai and M. G. Low. A framework for estimation of convex functions. Technical report, Technical report, 2011.

[CL15] S. Chatterjee and J. Lafferty. Adaptive risk bounds in unimodal regression. *arXiv preprint arXiv:1512.02956*, 2015.

[DJKP95] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 301–369, 1995.

[DM59] D. Davidson and J. Marschak. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, pages 233–69, 1959.

[DM15] Y. Deshpande and A. Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. *arXiv:1502.06590*, 2015.

[Dug14] S. Dughmi. On the hardness of signaling. In *IEEE Foundations of Computer Science (FOCS)*, pages 354–363, 2014.

[FK03] U. Feige and R. Krauthgamer. The probable value of the lovász–schrijver relaxations for maximum independent set. *SIAM Journal on Computing*, 32(2):345–370, 2003.

[Gil52] E. N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31(3):504–522, 1952.

[GW07] F. Gao and J. A. Wellner. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764, 2007.

[HMG07] R. Herbrich, T. Minka, and T. Graepel. Trueskill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, 2007.

[HOX14] B. Hajek, S. Oh, and J. Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2014.

[Jer92] M. Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[JP00] A. Juels and M. Peinado. Hiding cliques for cryptographic security. *Designs, Codes and Cryptography*, 20(3):269–280, 2000.

[Kol06] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

[Kol11] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 38. Springer Science & Business Media, 2011.

[KRS15] R. Kyng, A. Rao, and S. Sachdeva. Fast, provable algorithms for isotonic regression in all l_p-norms. In *Advances in Neural Information Processing Systems*, pages 2701–2709, 2015.

[Kuč95] L. Kučera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2):193–212, 1995.

[Led01] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[Luc59] R. D. Luce. *Individual choice behavior: A theoretical analysis*. New York: Wiley, 1959.

[ML65] D. H. McLaughlin and R. D. Luce. Stochastic transitivity and cancellation of preferences between bitter-sweet solutions. *Psychonomic Science*, 2(1-12):89–90, 1965.

[MPW15] R. Meka, A. Potechin, and A. Wigderson. Sum-of-squares lower bounds for planted clique. *arXiv:1503.06447*, 2015.

[MW15] Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

[NOS12] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2012.

[RGLA15] A. Rajkumar, S. Ghoshal, L.-H. Lim, and S. Agarwal. Ranking from stochastic pairwise preferences: Recovering Condorcet winners and tournament solution sets at the top. In *International Conference on Machine Learning*, 2015.

[RKJ08] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *ACM conference on Information and knowledge management*, pages 43–52, 2008.

[RWDR88] T. Robertson, F. Wright, R. L. Dykstra, and T. Robertson. *Order restricted statistical inference*, volume 229. Wiley New York, 1988.

[SBB$^+$15] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Conference on Artificial Intelligence and Statistics*, pages 856–865, 2015.

[SBGW15] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint 1510.05610*, 2015.

[SW15] N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint 1512.08949*, 2015.

[THT11] R. J. Tibshirani, H. Hoefling, and R. Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011.

[Thu27] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273, 1927.

[Var57] R. Varshamov. Estimate of the number of signals in error correcting codes. In *Dokl. Akad. Nauk SSSR*, 1957.

[vdG00] S. van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[Wai14] M. J. Wainwright. Constrained forms of statistical minimax: Computation, communication and privacy. In *Proceedings of the International Congress of Mathematicians*, Seoul, Korea, 2014.