*Submitted to the Annals of Statistics*
arXiv: 2009.02609

# ISOTONIC REGRESSION WITH UNKNOWN PERMUTATIONS: STATISTICS, COMPUTATION, AND ADAPTATION

BY ASHWIN PANANJADY[1,*] AND RICHARD J. SAMWORTH[2,†]

[1]*Georgia Institute of Technology, 755 Ferst Dr NW, Atlanta 30318, USA*
*ashwinpm@gatech.edu*

[2]*Statistical Laboratory, Wilberforce Road, Cambridge, CB3 0WB, United Kingdom*
*r.samworth@statslab.cam.ac.uk*

*Dedicated to the memory of Matthew Brennan*

Motivated by models for multiway comparison data, we consider the problem of estimating a coordinate-wise isotonic function on the domain $[0,1]^d$ from noisy observations collected on a uniform lattice, but where the design points have been permuted along each dimension. While the univariate and bivariate versions of this problem have received significant attention, our focus is on the multivariate case $d \geq 3$. We study both the minimax risk of estimation (in empirical $L_2$ loss) and the fundamental limits of adaptation (quantified by the adaptivity index) to a family of piecewise constant functions. We provide a computationally efficient Mirsky partition estimator that is minimax optimal while also achieving the smallest adaptivity index possible for polynomial time procedures. Thus, from a worst-case perspective and in sharp contrast to the bivariate case, the latent permutations in the model do not introduce significant computational difficulties over and above vanilla isotonic regression. On the other hand, the fundamental limits of adaptation are significantly different with and without unknown permutations: Assuming a hardness conjecture from average-case complexity theory, a statistical-computational gap manifests in the former case. In a complementary direction, we show that natural modifications of existing estimators fail to satisfy at least one of the desiderata of optimal worst-case statistical performance, computational efficiency, and fast adaptation. Along the way to showing our results, we improve adaptation results in the special case $d = 2$ and establish some properties of estimators for vanilla isotonic regression, both of which may be of independent interest.

**1. Introduction.** Consider the problem of estimating a degree $d$, real-valued tensor $\theta^* \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, whose entries are observed with noise. As in many problems in high-dimensional statistics, this tensor estimation problem requires a prohibitively large number of observations to solve without the imposition of further structure, and consequently, many structural constraints have been placed in particular applications of tensor estimation. For instance, low "rank" structure is common in chemistry and neuroscience applications (Andersen and Bro, 2003; Möcks, 1988), blockwise constant structure is common in applications to clustering and classification of relational data (Zhou et al., 2007), sparsity is commonly used in data mining applications (Kolda et al., 2005), and variants and combinations of such assumptions have also appeared in other contexts (Zhou et al., 2015). In this paper, we study

*Supported in part by a research fellowship from the Simons Institute for the Theory of Computing.
†Supported by EPSRC Fellowship EP/P031447/1 and EPSRC Programme Grant EP/N031938/1.
*MSC2020 subject classifications:* 62G05.
*Keywords and phrases:* permutation-based model; statistical-computational gap; adaptive estimation; shape-constrained estimation.

a flexible, nonparametric structural assumption that generalizes parametric assumptions in applications of tensor models to discrete choice data.

Suppose we are interested in modeling ordinal data, which arises in applications ranging from information retrieval (Dwork et al., 2001) and assortment optimization (Kök et al., 2015) to recommender systems (Baltrunas et al., 2010) and crowdsourcing (Chen et al., 2013); in a generic such problem, we have $n_1$ "items", subsets of which are evaluated using a multiway comparison. In particular, each datum takes the form of a tuple containing $d$ of these items, and a single item that is chosen from the tuple as the "winner" of this comparison. Such data can be represented using a stochastic model of choice: For each tuple $A$ and each item $i \in A$, suppose that $i$ wins the comparison with probability $p(i, A)$. The winner of each comparison is then modeled as a random variable; equivalently, the overall statistical model is described by a $d$-dimensional mean tensor $\theta^* \in \mathbb{R}^{n_1 \times \cdots \times n_1}$, where $\theta^*(i_1, \ldots, i_d) = p(i_1, (i_1, \ldots, i_d))$, and our data consist of noisy observations of entries of this tensor. Imposing sensible constraints on the tensor $\theta^*$ in these applications goes back to classical, axiomatic work on the subject due to Luce (1959) and Plackett (1975). A natural and flexible assumption is given by *simple scalability* (Krantz, 1965; McFadden, 1981; Tversky, 1972): this states that each of the $n_1$ items can be associated with some scalar utility (item $i$ with utility $u_i$), and that the comparison probability is given by

$$\theta^*(i_1, \ldots, i_d) = f(u_{i_1}, \ldots, u_{i_d}), \tag{1}$$

where $f$ is a non-decreasing function of its first argument and a coordinate-wise non-increasing function of the remaining arguments. Operationally, an item should not have a lower chance of being chosen as a winner if—all else remaining equal—its utility were to be increased.

There are many models that satisfy the nonparametric simple scalability assumption, in particular, *parametric* assumptions in which a specific form of the function $f$ is posited. The simplest parameterization is given by $f(u_1, \ldots, u_d) = u_1 / \sum_{j=1}^{d} u_j$, which dates back to Luce (1959). A logarithmic transformation of Luce's parameterization leads to the multinomial logit (MNL) model, which has seen tremendous popularity in applications ranging from transportation (Ben-Akiva et al., 1985) to marketing (Chandukala et al., 2008). See, e.g., McFadden (1974) for a classical but comprehensive introduction to this class of models. However the parametric assumptions of the MNL model have been called into question by a line of work showing that greater flexibility in modeling can lead to improved results in many applications (see, e.g., Farias et al. (2013) and references therein).

The simple scalability (SS) assumption has also been extensively explored when $d = 2$, i.e., in the pairwise comparison case. In this case, the nonparametric SS assumption is equivalent to *strong stochastic transitivity*, and a long line of work (Marschak and Davidson, 1957; McLaughlin and Luce, 1965; Fishburn, 1973) has studied its empirical properties. In particular, the parametric MNL model specialized to this case corresponds to the popular Bradley–Terry–Luce model (Bradley and Terry, 1952; Luce, 1959), and nonparametric models are known to be significantly more robust to misspecification than their parametric counterparts in common applications (Marschak and Davidson, 1957; Ballinger and Wilcox, 1997).

Let us return now to the SS assumption in the general case, and state an equivalent formulation in terms of structure on the tensor $\theta^*$. For two vectors of equal dimension, let $x \preceq y$ denote that $x - y \leq 0$ entrywise, and let $\pi$ denote any permutation of $[n_1]$ that orders the utilities in the sense that $u_{\pi(1)} \leq \cdots \leq u_{\pi(n_1)}$. The monotonicity of the function $f$ in the SS assumption (1) ensures that whenever $(i_1, \ldots, i_d) \preceq (i'_1, \ldots, i'_d)$, we have

$$\begin{aligned} \theta^*\big(\pi(i_1), \pi^{-1}(i_2), \ldots, \pi^{-1}(i_d)\big) &= f\big(u_{\pi(i_1)}, u_{\pi^{-1}(i_2)}, \ldots, u_{\pi^{-1}(i_d)}\big) \\ &\leq f\big(u_{\pi(i'_1)}, u_{\pi^{-1}(i'_2)}, \ldots, u_{\pi^{-1}(i'_d)}\big) \\ &= \theta^*\big(\pi(i'_1), \pi^{-1}(i'_2), \ldots, \pi^{-1}(i'_d)\big). \end{aligned} \tag{2}$$

Crucially, since the utilities themselves are latent, the permutation $\pi$ is *unknown*—indeed, it represents the ranking that must be estimated from our data—and so $\theta^*$ is a coordinate-wise isotonic tensor with unknown permutations. In the multiway comparison problem, this tensor represents the stochastic model underlying our data, and accurate knowledge of these probabilities is useful, for instance, in informing pricing and revenue management decisions in assortment optimization applications (Kök et al., 2015).

While multiway comparisons form our primary motivation, the flexibility afforded by non-parametric models with latent permutations has also been noticed and exploited in other applications. For instance, in psychometric item-response theory, the Mokken model—which corresponds to imposing structure of the form (2) when $d = 2$—is known to be significantly more robust to misspecification that the parametric Rasch model; see van Schuur (2003) for an introduction and survey. In crowd-labeling, the permutation-based model (Shah et al., 2021) has seen empirical success in applications where the parametric Dawid–Skene model (Dawid and Skene, 1979) imposes stringent assumptions. Besides these, there are also several other examples of tensor estimation problems in which parametric structure is frequently assumed; for example, in click modeling (Craswell et al., 2008) and random hypergraph models (Ghoshdastidar and Dukkipati, 2017; Angelini et al., 2015). Similarly to before, nonparametric structure has the potential to generalize and lend flexibility to these parametric models.

It is worth noting that in many of the aforementioned applications, the underlying objects can be clustered into near identical sets. For example, there is evidence that such "indifference sets" of items exist in crowdsourcing (see Shah et al. (2019a, Figure 1) for an illuminating example) and peer review applications involving comparison data (Nguyen et al., 2014); clustering is often used in the application of psychometric evaluation methods (Hardouin and Mesbah, 2004), and many models for communities in hypergraphs posit the existence of such clusters of nodes (Abbe and Montanari, 2013; Ghoshdastidar and Dukkipati, 2017). For a precise mathematical definition of indifference sets and how they induce further structure in the tensor $\theta^*$, see Section 2. Whenever such additional structure exists, it is conceivable that estimation can be performed in a more sample-efficient manner; we will precisely quantify such a phenomenon in our exposition in Section 3.

Using these applications as motivation, our goal in this paper is to study the tensor estimation problem under the nonparametric structural assumptions (2) of monotonicity constraints and unknown permutations.

1.1. *Related work.* Regression problems with unknown permutations were classically studied in applications to record-linkage (DeGroot and Goel, 1980), and similar models have witnessed recent interest driven by other modern applications in machine learning and signal processing; see, e.g., Collier and Dalalyan (2016); Unnikrishnan et al. (2018); Pananjady et al. (2017a,b); Hsu et al. (2017); Abid and Zou (2018); Behr and Munk (2017) for theoretical results and applications. We focus our discussion on the sub-class of such problems involving monotonic shape-constraints and (vector/matrix/tensor) estimation. When $d = 1$, the assumption (2) corresponds to the "uncoupled" or "shuffled" univariate isotonic regression problem (Carpentier and Schlueter, 2016). Here, an estimator based on Wasserstein deconvolution is known to attain the minimax rate $\log\log n/\log n$ in (normalized) squared $\ell_2$-error for estimation of the underlying (sorted) vector of length $n$ (Rigollet and Weed, 2019). In a recent paper, Balabdaoui et al. (2020) considered a closely related problem, with a focus on isolating the effect of the noise distribution on the deconvolution procedure. A multivariate version of this problem (estimating multiple isotonic functions under a common unknown permutation of coordinates) has also been studied under the moniker of "statistical seriation", and has been shown to have applications to archaeology and metagenomics (Flammarion et al., 2019; Ma et al., 2021+).

The case $d = 2$ has also seen a long line of work in the mathematical statistics community in the context of estimation from pairwise comparisons, wherein the monotonicity assumption (2) corresponds to strong stochastic transitivity, or SST for short (e.g., Chatterjee, 2015; Shah et al., 2017; Chatterjee and Mukherjee, 2019; Shah et al., 2019a; Mao et al., 2020). Relatives of this model have also appeared in the context of prediction in graphon estimation (Chan and Airoldi, 2014; Airoldi et al., 2013) and calibration in crowd-labeling (Shah et al., 2021; Mao et al., 2020). The minimax rate (in normalized, squared Frobenius error) of estimating an $(n^{1/2} \times n^{1/2})$ SST matrix is known to be of the order $n^{-1/2}$ up to a polylogarithmic factor, but many computationally efficient algorithms (Chatterjee, 2015; Shah et al., 2017; Chatterjee and Mukherjee, 2019; Shah et al., 2019a) achieved only the rate $n^{-1/4}$. Recent progress has shown more sophisticated (but still efficient) procedures with improved rates: an algorithm with rate $n^{-3/8}$ was given by Mao et al. (2018), and in recent work, Liu and Moitra (2020) show that a rate $n^{-5/12}$ can be achieved. However, it is still not known whether the minimax rate is attainable by an efficient algorithm. The case of estimating rectangular matrices has also been studied, and the fundamental limits are known to be sensitive to the aspect ratio of the problem (Mao et al., 2020). Interesting adaptation properties are also known in this case, both to parametric structure (Chatterjee and Mukherjee, 2019), and to indifference sets (Shah et al., 2019a).

To the best of our knowledge, analogs of these results have not been explored in the multivariate setting $d \geq 3$, although a significant body of literature has studied parametric models for choice data in this case (see, e.g., Negahban et al. (2018) and references therein).

1.2. *Overview of contributions.* We begin by considering the minimax risk of estimating bounded tensors satisfying assumption (2), and show in Proposition 1 that when $d \geq 2$, it is dominated by the risk of estimating the underlying *ordered* coordinate-wise isotonic tensor. In other words, the latent permutations do not significantly influence the statistical difficulty of the problem. We also study the fundamental limits of estimating tensors having indifference set structure, and this allows us to assess the ability of an estimator to adapt to such structure via its *adaptivity index* (to be defined precisely in equation (3)). We establish two surprising phenomena in this context: First, we show in Proposition 3 that the fundamental limits of estimating these objects preclude a parametric rate, in sharp contrast to the case without unknown permutations. Second, we prove in Theorem 1 that the adaptivity index exhibits a statistical-computational gap under the assumption of a widely-believed conjecture in average-case complexity. In particular, we show that the adaptivity index of any polynomial time computable estimator must grow at least polynomially in $n$, assuming the hypergraph planted clique conjecture (Brennan and Bresler, 2020). Our results also have interesting consequences for the isotonic regression problem without unknown permutations (see Proposition 2 and Corollary 1).

Having established these fundamental limits, we then turn to our main methodological contribution. We propose and analyze—in Theorem 2—an estimator based on Mirsky's partitioning algorithm (Mirsky, 1971) that estimates the underlying tensor (a) at the minimax rate (up to poly-logarithmic factors) for each $d \geq 3$ whenever this tensor has bounded entries, and (b) with the best possible adaptivity index for polynomial time procedures for all $d \geq 2$. The first of these findings is particularly surprising because it shows that the case $d \geq 3$ of this problem is distinctly different from the bivariate case, in that the minimax risk is achievable with a computationally efficient algorithm. This is in spite of the fact that there are more permutations to estimate as the dimension increases, which, at least in principle, ought to make the problem more difficult both statistically and computationally.

In addition to its favorable risk properties, the Mirsky partition estimator also has several other advantages: it is computable in time sub-quadratic in the size of the input, and its *computational complexity* also adapts to underlying indifference set structure. In particular, when

there are a fixed number of indifference sets, the estimator has almost linear computational complexity with high probability. When specialized to $d = 2$, this estimator exhibits significantly better adaptation properties to indifference set structure than known estimators that were designed specifically for this purpose; see Section 3.5 and Appendix A of the supplementary material for statements and discussions of these results.

To complement our upper bounds on the Mirsky partition estimator, we also show, somewhat surprisingly, that many other estimators proposed in the literature (Chatterjee and Mukherjee, 2019; Shah et al., 2019a, 2017), and natural variants thereof, suffer from an extremely large adaptivity index. In particular, they are unable to attain the polynomial time optimal adaptivity index (given by the fundamental limit established by Theorem 1) for any $d \geq 4$. This is in spite of the fact that some of these estimators are minimax optimal for estimation over the class of bounded tensors (see Propositions 4 and 2) for all $d \geq 3$. Thus, we see that simultaneously achieving good worst-case risk properties while remaining computationally efficient and adaptive to structure is a challenging requirement, thereby providing further evidence of the value of the Mirsky partitioning estimator.

1.3. *Organization.* The rest of this paper is organized as follows. In Section 2, we introduce formally the estimation problem at hand. Section 3 contains full statements and discussions of our main results, and Section 4 contains some concluding remarks. Proofs of the main results are given in the supplementary material (Pananjady and Samworth, 2021+). The supplementary material also contains appendices that may be of independent interest. In particular, some of our results on adaptation in the special cases $d = 2, 3$ are postponed to Appendix A, and Appendix B collects some properties of the vanilla isotonic regression estimator that may be of independent interest.

**2. Background and problem formulation.** Let $\mathfrak{S}_k$ denote the set of all permutations on the set $[k] := \{1, \ldots, k\}$. We interpret $\mathbb{R}^{n_1 \times \cdots \times n_d}$ as the set of all real-valued, tensors of dimension $n_1 \times \cdots \times n_d$. For a set of positive integers $i_j \in [n_j], j \in [d]$, we use $T(i_1, \ldots, i_d)$ to index entry $i_1, \ldots, i_d$ of a tensor $T \in \mathbb{R}^{n_1 \times \cdots \times n_d}$.

The set of all real-valued, coordinate-wise isotonic functions on the set $[0,1]^d$ is denoted by

$$\mathcal{F}_d := \left\{ f : [0,1]^d \to \mathbb{R} : f(x_1, x_2, \ldots, x_d) \leq f(x'_1, x'_2, \ldots, x'_d) \text{ when } x_j \leq x'_j \text{ for } j \in [d] \right\}.$$

Let $n_j$ denote the number of observations along dimension $j$, with the total number of observations given by $n := \prod_{j=1}^d n_j$. For $n_1, \ldots, n_d \in \mathbb{N}$, let $\mathbb{L}_{d,n_1,\ldots,n_d} := \prod_{j=1}^d [n_j]$ denote the $d$-dimensional lattice. With this notation, we assume access to a tensor of observations $Y \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, where

$$Y(i_1, \ldots, i_d) = f^*\left(\frac{\pi_1^*(i_1)}{n_1}, \frac{\pi_2^*(i_2)}{n_2}, \ldots, \frac{\pi_d^*(i_d)}{n_d}\right) + \epsilon(i_1, \ldots, i_d) \text{ for each } i_j \in [n_j],\ j \in [d].$$

Here, the function $f^* \in \mathcal{F}_d$ is unknown, and for each $j \in [d]$, we also have an unknown permutation $\pi_j^* \in \mathfrak{S}_{n_j}$. The tensor $\epsilon \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ represents noise in the observation process, and we assume that its entries are given by independent standard normal random variables[1]. Denote the noiseless observations on the lattice by

$$\theta^*(i_1, \ldots, i_d) := f^*\left(\frac{\pi_1^*(i_1)}{n_1}, \frac{\pi_2^*(i_2)}{n_2}, \ldots, \frac{\pi_d^*(i_d)}{n_d}\right) \quad \text{for each } i_j \in [n_j],\ j \in [d];$$

[1]We study the canonical Gaussian setting for convenience, but our results extend straightforwardly to sub-Gaussian noise distributions.

this is a generalization of[2] the nonparametric structure that was posited in equation (2).

It is also convenient to define the set of tensors that can be formed by permuting evaluations of a coordinate-wise monotone function on the lattice by the permutations $(\pi_1, \ldots, \pi_d)$. Denote this set by

$$\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d}; \pi_1, \ldots, \pi_d) := \Bigg\{ \theta \in \mathbb{R}^{n_1 \times \cdots \times n_d} : \exists f \in \mathcal{F}_d \text{ such that } \forall i_j \in [n_j],\ j \in [d], \\ \theta(i_1, \ldots, i_d) = f\left( \frac{\pi_1(i_1)}{n_1}, \ldots, \frac{\pi_d(i_d)}{n_d} \right) \Bigg\}.$$

We use the shorthand $\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})$ to denote this set when the permutations are all the identity. Also define the set

$$\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n_1,\ldots,n_d}) := \bigcup_{\pi_1 \in \mathfrak{S}_{n_1}} \cdots \bigcup_{\pi_d \in \mathfrak{S}_{n_d}} \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d}; \pi_1, \ldots, \pi_d)$$

of tensors that can be formed by permuting evaluations of any coordinate-wise monotone function.

For a collection of permutations $\{\pi_j \in \mathfrak{S}_{n_j}\}_{j=1}^d$ and a tensor $T \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, we let $T\{\pi_1, \ldots, \pi_d\}$ denote the tensor $T$ viewed along permutation $\pi_j$ on dimension $j$. Specifically, we have

$$T\{\pi_1, \ldots, \pi_d\}(i_1, \ldots, i_d) = T(\pi_1(i_1), \ldots, \pi_d(i_d)) \quad \text{for each} \quad i_j \in [n_j],\ j \in [d].$$

With this notation, note the inclusion $\theta^*\{(\pi_1^*)^{-1}, \ldots, (\pi_d^*)^{-1}\} \in \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})$. However, since we do not know the permutations $\pi_1^*, \ldots, \pi_d^*$ a priori, we may only assume that $\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n_1,\ldots,n_d})$, and our goal is to denoise our observations and produce an estimate of $\theta^*$. We study the empirical $L_2$ risk of any such estimate $\widehat{\theta} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, given by

$$\mathcal{R}_n(\widehat{\theta}, \theta^*) := \mathbb{E}\left[ \ell_n^2(\widehat{\theta}, \theta^*) \right], \text{ where}$$

$$\ell_n^2(\theta_1, \theta_2) := \frac{1}{n} \sum_{j=1}^{d} \sum_{i_j=1}^{n_j} \left( \theta_1(i_1, \ldots, i_d) - \theta_2(i_1, \ldots, i_d) \right)^2.$$

Note that the expectation is taken over both the noise $\epsilon$ and any randomness used to compute the estimate $\widehat{\theta}$. In the case where $\widehat{\theta} \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n_1,\ldots,n_d})$, we also produce a function estimate $\widehat{f} \in \mathcal{F}_d$ and permutation estimates $\widehat{\pi}_j \in \mathfrak{S}_{n_j}$ for $j \in [d]$, with

$$\widehat{\theta}(i_1, \ldots, i_d) := \widehat{f}\left( \frac{\widehat{\pi}_1(i_1)}{n_1}, \frac{\widehat{\pi}_2(i_2)}{n_2}, \ldots, \frac{\widehat{\pi}_d(i_d)}{n_d} \right) \quad \text{for each } i_j \in [n_j],\ j \in [d].$$

Note that in general, the resulting estimates $\widehat{f}, \widehat{\pi}_1, \ldots, \widehat{\pi}_d$ need not be unique, but this identifiability issue will not concern us since we are only interested in the tensor $\widehat{\theta}$ as an estimate of the tensor $\theta^*$.

As alluded to in the introduction, it is common in multiway comparisons for there to be *indifference sets* of items that all behave identically. These sets are easiest to describe in the space of functions. For each $j \in [d]$ and $s_j \in [n_j]$, let $I_1^j, \ldots, I_{s_j}^j$ denote a set of $s_j$ disjoint intervals such that $[0,1] = \cup_{\ell=1}^{s_j} I_\ell^j$. Suppose that for each $\ell$, the length of the interval $I_\ell^j$ exceeds $1/n_j$, so that we are assured that the intersection of $I_\ell^j$ with the set $\frac{1}{n_j}\{1, \ldots, n_j\}$ is nonempty. With a slight abuse of terminology, we also refer to this intersection as an interval, and

[2]Note that unlike in equation (2), we now allow for a different unknown permutation along each dimension for greater flexibility.

let the tuple $\mathbf{k}^j = (k_1^j, \dots, k_{s_j}^j)$ denote the cardinalities of these intervals, with $\sum_{\ell=1}^{s_j} k_\ell^j = n_j$. Let $\mathbf{K}_{s_j}$ denote the set of all such tuples, and define $k_{\max}^j := \max_{\ell \in [s_j]} \mathbf{k}_\ell^j$. Collect $\{\mathbf{k}^j\}_{j=1}^d$ in a tuple $\mathbb{k} = (\mathbf{k}^1, \dots, \mathbf{k}^d)$, and the $d$ values $\{s_j\}_{j=1}^d$ in a tuple $\mathbf{s} = (s_1, \dots, s_d)$. Let $\mathbb{K}_{\mathbf{s}}$ denote the set of all such tuples $\mathbb{k}$, and note that the possible values of $\mathbf{s}$ range over the lattice $\mathbb{L}_{d,n_1,\dots,n_d}$. Finally, let $k^* := \min_{j \in [d]} k_{\max}^j$. See Figure 1 for an illustration when $d = 2$.

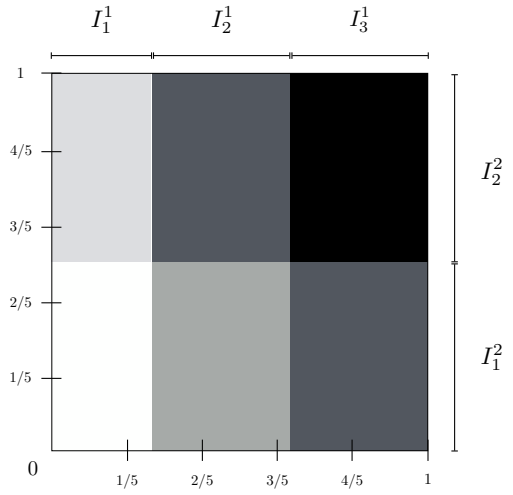


FIG 1. *An illustration of a block-wise constant isotonic function on* $[0,1]^2$*, with lighter colors indicating smaller values. We observe this function on a* $5 \times 5$*, equally spaced grid (i.e.,* $n_1 = 5$*). The number of indifference sets along the two dimensions satisfies* $s_1 = 3$ *and* $s_2 = 2$*, and their size tuples satisfy* $\mathbf{k}^1 = (1,2,2)$ *and* $\mathbf{k}^2 = (2,3)$*. As a result, we have* $k^* = 2$*.*

.

If, for each $j \in [d]$, dimension $j$ of the domain is partitioned into the intervals $I_1^j, \dots, I_{s_j}^j$, then the set $[0,1]^d$ is partitioned into $s := \prod_{j=1}^d s_j$ hyper-rectangles. Here, the hyper-rectangular partition is a Cartesian product of univariate partitions, i.e., each hyper-rectangle takes the form $\prod_{j=1}^d I_{\ell_j}^j$ for some sequence of indices $\ell_j \in [s_j], j \in [d]$. We refer to the intersection of a hyper-rectangle with the lattice $\mathbb{L}_{d,n_1,\dots,n_d}$ also as a hyper-rectangle, and note that $\mathbb{k}$ fully specifies such a hyper-rectangular partition. Denote by $\mathcal{M}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n_1,\dots,n_d})$ the set of all $\theta \in \mathcal{M}(\mathbb{L}_{d,n_1,\dots,n_d})$ that are piecewise constant on a hyper-rectangular partition specified by $\mathbb{k}$—we have chosen to be explicit about the tuple $\mathbf{s}$ in our notation for clarity. Let $\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n_1,\dots,n_d})$ denote the set of all coordinate-wise permuted versions of $\theta \in \mathcal{M}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n_1,\dots,n_d})$.

For the rest of this paper, we operate in the *uniform, or balanced* case $2 \le n_1 = \dots = n_d = n^{1/d}$, which is motivated by the comparison setting introduced in Section 1. We use the shorthand $\mathbb{L}_{d,n}$ to represent the uniform lattice and $\mathbb{R}_{d,n}$ to represent balanced tensors. We continue to use the notation $n_j$ in some contexts since this simplifies our exposition, and also continue to accommodate distinct permutations $\pi_1^*, \dots, \pi_d^*$ and cardinalities of indifference sets $s_1, \dots, s_d$ along the different dimensions for flexibility.

Let $\widehat{\Theta}$ denote the set of all estimators of $\theta^*$, i.e. the set of all measurable functions (of the observation tensor $Y$) taking values in $\mathbb{R}_{d,n}$. Denote the minimax risk over the class of tensors in the set $\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n})$ by

$$\mathfrak{M}_{d,n}(\mathbb{k},\mathbf{s}) := \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*).$$

Note that $\mathfrak{M}_{d,n}(\mathbb{k}, \mathbf{s})$ measures the smallest possible risk achievable with a priori knowledge of the inclusion $\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$. On the other hand, we are interested in estimators that *adapt* to hyper-rectangular structure without knowing of its existence in advance. One way to measure the extent of adaptation of an estimator $\widehat{\theta}$ is in terms of its *adaptivity index* to indifference set sizes $\mathbb{k}$, defined as

$$\mathfrak{A}^{\mathbb{k},\mathbf{s}}(\widehat{\theta}) := \frac{\sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*)}{\mathfrak{M}_{d,n}(\mathbb{k}, \mathbf{s})}.$$

A large value of this index indicates that the estimator $\widehat{\theta}$ is unable to adapt satisfactorily to the set $\mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$, since a much lower risk is achievable when the inclusion $\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ is known in advance. The *global* adaptivity index of $\widehat{\theta}$ is then given by

$$\mathfrak{A}(\widehat{\theta}) := \max_{\mathbf{s} \in \mathbb{L}_{d,n}} \max_{\mathbb{k} \in \mathbb{K}_{\mathbf{s}}} \mathfrak{A}^{\mathbb{k},\mathbf{s}}(\widehat{\theta}). \tag{3}$$

We note that similar definitions of an adaptivity index or factor have appeared in the literature; our definition most closely resembles the index defined by Shah et al. (2019a), but similar concepts go back at least to Lepski and co-authors (Lepski, 1991; Lepski and Spokoiny, 1997).

Finally, for a tensor $X \in \mathbb{R}_{d,n}$ and closed set $\mathcal{C} \subseteq \mathbb{R}_{d,n}$, it is useful to define the $L_2$-projection of $X$ onto $\mathcal{C}$ by

$$\widehat{\theta}_{\mathsf{LSE}}(\mathcal{C}, X) \in \operatorname*{argmin}_{\theta \in \mathcal{C}} \ell_n^2(X, \theta). \tag{4}$$

In our exposition to follow, the set $\mathcal{C}$ will either be compact or a finite union of closed, convex sets, and so the projection is guaranteed to exist. When $\mathcal{C}$ is closed and convex, the projection is additionally unique.

*General notation.* For a (semi-)normed space $(\mathcal{F}, \|\cdot\|)$ and positive scalar $\delta$, let $N(\delta; \mathcal{F}, \|\cdot\|)$ denote its $\delta$-covering number, i.e., the minimum cardinality of any set $U \subseteq \mathcal{F}$ such that $\inf_{u \in U} \|x - u\| \leq \delta$ for all $x \in \mathcal{F}$. Let $\mathbb{B}_\infty(t)$ and $\mathbb{B}_2(t)$ denote the $\ell_\infty$ and $\ell_2$ closed balls of radius $t$ in $\mathbb{R}_{d,n}$, respectively. Let $\mathbb{I}\{\cdot\}$ denote the indicator function, and denote by $\mathbb{1}_{d,n} \in \mathbb{R}_{d,n}$ the all-ones tensor. For two sequences of non-negative reals $\{f_n\}_{n \geq 1}$ and $\{g_n\}_{n \geq 1}$, we use $f_n \lesssim g_n$ to indicate that there is a universal positive constant $C$ such that $f_n \leq C g_n$ for all $n \geq 1$. The relation $f_n \gtrsim g_n$ indicates that $g_n \lesssim f_n$, and we say that $f_n \asymp g_n$ if both $f_n \lesssim g_n$ and $f_n \gtrsim g_n$ hold simultaneously. We also use standard order notation $f_n = \mathcal{O}(g_n)$ to indicate that $f_n \lesssim g_n$ and $f_n = \widetilde{\mathcal{O}}(g_n)$ to indicate that $f_n \lesssim g_n \log^c n$, for a universal constant $c > 0$. We say that $f_n = \Omega(g_n)$ (resp. $f_n = \widetilde{\Omega}(g_n)$) if $g_n = \mathcal{O}(f_n)$ (resp. $g_n = \widetilde{\mathcal{O}}(f_n)$). The notation $f_n = o(g_n)$ is used when $\lim_{n \to \infty} f_n / g_n = 0$, and $f_n = \omega(g_n)$ when $g_n = o(f_n)$. Throughout, we use $c, C$ to denote universal positive constants, and their values may change from line to line. Finally, we use the symbols $c_d, C_d$ to denote $d$-dependent constants; once again, their values will typically be different in each instantiation. All logarithms are to the natural base unless otherwise stated. We denote by $\mathcal{N}(\mu, \sigma^2)$ a normal distribution with mean $\mu$ and variance $\sigma^2$. We use $\mathsf{Ber}(p)$ to denote a Bernoulli distribution with success probability $p$, and denote by $\mathsf{Bin}(n, p)$ a binomial distribution with $n$ trials and success probability $p$. We let $\mathsf{Hyp}(n, N, K)$ denote a hypergeometric distribution with $n$ trials, a universe of size $N$, and $K$ defectives[3]. Finally, we denote the total variation distance between two distributions $\mu$ and $\nu$ by $d_{\mathsf{TV}}(\mu, \nu)$.

[3]Recall that a hypergeometric random variable is formed as follows: Suppose that there is a universe of $N$ items containing $K$ defective items. Then $\mathsf{Hyp}(n, N, K)$ is the distribution of the number of defective items in a collection of $n$ items drawn uniformly at random, without replacement from the universe.

**3. Main results.** We begin by characterizing the fundamental limits of estimation and adaptation, and then turn to developing an estimator that achieves these limits. Finally, we analyze variants of existing estimators from this point of view.

3.1. *Fundamental limits of estimation.* In this subsection, our focus is on the fundamental limits of estimation over various parameter spaces without imposing any computational constraints on our procedures. We begin by characterizing the minimax risk over the class of bounded, coordinate-wise isotonic tensors with unknown permutations.

PROPOSITION 1. *There is a universal positive constant $C$ such that for each $d \geq 2$,*

$$c_d \cdot n^{-1/d} \leq \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}, \theta^*) \leq C \cdot n^{-1/d} \log^2 n, \tag{5}$$

*where $c_d > 0$ depends on $d$ alone.*

The lower bound on the minimax risk in equation (5) follows immediately from known results on estimating bounded monotone functions on the lattice without unknown permutations (Han et al., 2019; Deng and Zhang, 2020). These results show that one can take $c_d \asymp (d-1)^{-(d-1)}$, but the dependence of this constant on $d$ can likely be improved.

The upper bound is our main contribution to the proposition, and is achieved by the bounded least squares estimator

$$\widehat{\theta}_{\mathsf{BLSE}} := \widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}\big(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1), Y\big). \tag{6}$$

In fact, the risk of $\widehat{\theta}_{\mathsf{BLSE}}$ can be expressed as a sum of two terms:

$$\mathcal{R}_n(\widehat{\theta}_{\mathsf{BLSE}}, \theta^*) \leq C\big(n^{-1/d} \log^2 n + n^{-(1-1/d)} \log n\big). \tag{7}$$

The first term corresponds to the error of estimating the unknown isotonic function, and the second to the price paid for having unknown permutations. Such a characterization was known in the case $d = 2$ (Shah et al., 2017; Mao et al., 2020), and our result shows that a similar decomposition holds even for larger $d$. Note that for all $d \geq 2$, the first term of equation (7) dominates the bound, and this is what leads to Proposition 1.

Although the bounded LSE (6) achieves the worst case risk (5), we may use its analysis as a vehicle for obtaining risk bounds for the vanilla least squares estimator without imposing any boundedness constraints. This results in the following proposition.

PROPOSITION 2. *There is a universal positive constant $C$ such that for each $d \geq 2$:*
*(a) The least squares estimator over the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ has worst case risk bounded as*

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y), \theta^*\big) \leq C n^{-1/d} \log^{5/2} n. \tag{8a}$$

*(b) The isotonic least squares estimator over $\mathcal{M}(\mathbb{L}_{d,n})$ has worst case risk bounded as*

$$\sup_{\theta^* \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}), Y), \theta^*\big) \leq C n^{-1/d} \log^{5/2} n. \tag{8b}$$

Part (a) of Proposition 2 deals with the LSE computed over the entire set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$, and appears to be new even when $d = 2$; to the best of our knowledge, prior work (Shah et al., 2017; Mao et al., 2020) has only considered the bounded LSE $\widehat{\theta}_{\mathsf{BLSE}}$ (6). Part (b) of Proposition 2, on the other hand, provides a risk for the vanilla isotonic least squares estimator when estimating functions in the set $\mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$. This estimator has a long

history in both the statistics and computer science communities (Robertson and Wright, 1975; Dykstra and Robertson, 1982; Stout, 2015; Kyng et al., 2015; Chatterjee et al., 2018; Han et al., 2019); unlike the other estimators considered so far, the isotonic LSE is the solution to a convex optimization problem and can be computed in time polynomial in $n$. Bounds on the worst case risk of this estimator are also known: results for $d = 1$ are classical (see, e.g., Brunk (1955); Nemirovski et al. (1985); Zhang (2002)); when $d = 2$, risk bounds were derived by Chatterjee et al. (2018); and the general case $d \geq 2$ was considered by Han et al. (2019). Proposition 2(b) improves the logarithmic factor in the latter two papers from $\log^4 n$ to $\log^{5/2} n$, and is obtained via a different proof technique involving a truncation argument.

Two other comments are worth making. First, it should be noted that there are other estimators for tensors in the set $\mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$ besides the isotonic LSE. The block-isotonic estimator of Deng and Zhang (2020), first proposed by Fokianos et al. (2020), enjoys a risk bound of the order $C_d \cdot n^{-1/d}$ for all $d \geq 2$, where $C_d > 0$ is a $d$-dependent constant. This eliminates the logarithmic factor entirely, and matches the minimax lower bound up to a $d$-dependent constant. In addition, the block-isotonic estimator also enjoys significantly better adaptation properties than the isotonic LSE. On the other hand, the best known algorithm to compute the block-isotonic estimator takes time $\mathcal{O}(n^3)$, while the isotonic LSE can be computed in time $\widetilde{\mathcal{O}}(n^{3/2})$ (Kyng et al., 2015).

Second, we note that when the design is random in the setting without unknown permutations Han (2021+, Theorem 3.6) improves, at the expense of a $d$-dependent constant, the logarithmic factors in the risk bounds of prior work (Han et al., 2019). His proof techniques are based on the concentration of empirical processes on upper and lower sets of $[0,1]^d$, and do not apply to the lattice setting considered here. On the other hand, our proof works on the event on which the LSE is suitably bounded, and is not immediately applicable to the random design setting. Both of these techniques should be viewed as particular ways of establishing the optimality of global empirical risk minimization procedures even when the entropy integral for the corresponding function class diverges; this runs contrary to previous heuristic beliefs about the suboptimality of these procedures (see, e.g., Birgé and Massart (1993), van de Geer (2000, pp. 121–122), Kim and Samworth (2016), Rakhlin et al. (2017), and Han (2021+) for further discussion).

Let us now turn to establishing the fundamental limits of estimation over the class $\mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$. The following proposition characterizes the minimax risk $\mathfrak{M}_{d,n}(\mathbb{k},\mathbf{s})$. Recall that $s = \prod_{j=1}^d s_j$ and $k^* = \min_{j\in[d]} \max_{\ell\in[s_j]} k^j_\ell$.

PROPOSITION 3. *There is a pair of universal positive constants $(c, C)$ such that for each $d \geq 1$, $\mathbf{s} \in \mathbb{L}_{d,n}$, and $\mathbb{k} \in \mathbb{K}_{\mathbf{s}}$, the minimax risk $\mathfrak{M}_{d,n}(\mathbb{k},\mathbf{s})$ satisfies*

$$\frac{c}{n} \cdot \Big(s + (n_1 - k^*)\Big) \leq \inf_{\widehat{\theta}\in\widehat{\Theta}} \sup_{\theta^*\in\mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta},\theta^*) \leq \frac{C}{n} \cdot \Big(s + (n_1 - k^*)\log n\Big). \tag{9}$$

A few comments are in order. As before, the risk can be decomposed into two terms: the first term represents the *parametric* rate of estimating a tensor with $s$ constant pieces, and the second term is the price paid for unknown permutations. When the parameter space is also bounded in $\ell_\infty$-norm, such a decomposition does not occur transparently even in the special case $d = 2$ (Shah et al., 2019a). Also note that when $s = \mathcal{O}(1)$ and $n_1 - k^* = \omega(1)$, the second term of the bound (9) dominates and the minimax risk is no longer of the parametric form $s/n$. This is in sharp contrast to isotonic regression without unknown permutations, where there are estimators that achieve the parametric risk up to poly-logarithmic factors (Deng and Zhang, 2020). Thus, the fundamental adaptation behavior that we expect changes significantly in the presence of unknown permutations.

Second, note that when $s_j = n_1$ for all $j \in [d]$, we have $\mathcal{M}_{\mathsf{perm}}^{\Bbbk,\mathbf{s}}(\mathbb{L}_{d,n}) = \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$, in which case the result above shows that consistent estimation is impossible over the set of all isotonic tensors with unknown permutations. This does *not* contradict Proposition 1, since Proposition 3 computes the minimax risk over isotonic tensors without imposing boundedness constraints.

Finally, we note that Proposition 3 yields the following corollary in the setting where we do not have unknown permutations. With a slight abuse of notation, we let

$$\mathcal{M}^s(\mathbb{L}_{d,n}) := \bigcup_{\mathbf{s}\,:\,\prod_{j=1}^d s_j = s} \; \bigcup_{\Bbbk \in \mathbb{K}_{\mathbf{s}}} \mathcal{M}^{\Bbbk,\mathbf{s}}(\mathbb{L}_{d,n})$$

denote the set of all coordinate-wise monotone tensors that are piecewise constant on a $d$-dimensional partition having $s$ pieces.

COROLLARY 1. *There is a pair of universal positive constants $(c, C)$ such that for each $d \geq 1$, the following statements hold.*
*(a) For each $\mathbf{s} \in \mathbb{L}_{d,n}$ and $\Bbbk \in \mathbb{K}_{\mathbf{s}}$, we have*

$$c \cdot \frac{s}{n} \leq \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}^{\Bbbk,\mathbf{s}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*) \leq C \cdot \frac{s}{n}. \tag{10a}$$

*(b) For each $s \in [n]$, we have*

$$c \cdot \frac{s}{n} \leq \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}^{s}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*) \leq C \cdot \frac{s \log n}{n}. \tag{10b}$$

Let us interpret this corollary in the context of known results. When $d = 1$ and there are no permutations, Bellec and Tsybakov (2015) established minimax lower bounds of order $s/n$ and upper bounds of the order $s \log n/n$ for estimating $s$-piece monotone functions, and the bound (10b) recovers this result. The problem of estimating a univariate isotonic vector with $s$ pieces was also considered by Gao et al. (2020), who showed a rate-optimal characterization of the minimax risk that exhibits an iterated logarithmic factor in the sample size whenever $s \geq 3$. When $d \geq 2$, however, the results of Corollary 1 are new to the best of our knowledge.

The fundamental limits of estimation over the class $\mathcal{M}_{\mathsf{perm}}^{\Bbbk,\mathbf{s}}(\mathbb{L}_{d,n})$ in Proposition 3 will allow us to assess the adaptivity index of particular estimators. Before we do that, however, we establish a baseline for adaptation by proving a lower bound on the adaptivity index of polynomial time estimators.

3.2. *Lower bounds on polynomial time adaptation.* We now turn to our average-case reduction showing that any computationally efficient estimator cannot have a small adaptivity index. Our primitive is the hypergraph planted clique conjecture $\mathsf{HPC}_D$, which is a hypergraph extension of the planted clique conjecture. Let us introduce the testing, or "detection", version of this conjecture. Denote the set of $D$-uniform hypergraphs on the vertex set $[N]$ (hypergraphs in which each hyperedge is incident on $D$ vertices) by $\mathbb{H}_{D,N}$. Define, via their generative models, the random hypergraphs:

1. $\mathcal{G}_D(N, p)$: Generate each hyperedge independently with probability $p$, and
2. $\mathcal{G}_D(N, p; K)$: Choose $K \geq D$ vertices uniformly at random and form a clique, adding all $\binom{K}{D}$ possible hyperedges between them. Add each remaining hyperedge independently with probability $p$.

Given an instantiation of a random hypergraph $G \in \mathbb{H}_{D,N}$, the testing problem is to distinguish the hypotheses $H_0 : G \sim \mathcal{G}_D(N,p)$ and $H_1 : G \sim \mathcal{G}_D(N,p;K)$. The error of any test $\psi_N : \mathbb{H}_{D,N} \mapsto \{0,1\}$ is given by

$$\mathcal{E}(\psi_N) := \frac{1}{2}\,\mathbb{E}_{H_0}\left[\psi_N(G)\right] + \frac{1}{2}\,\mathbb{E}_{H_1}\left[1-\psi_N(G)\right]. \tag{11}$$

CONJECTURE 1 ($\mathsf{HPC}_D$ conjecture). *Suppose that $p = 1/2$, and that $D \geq 2$ is a fixed integer. If*

$$\limsup_{N\to\infty} \frac{\log K}{\log \sqrt{N}} < 1,$$

*then for any sequence of tests $\{\psi_N\}_{N\geq 1}$ such that $\psi_N$ is computable in time polynomial in $N^D$, we have*

$$\liminf_{N\to\infty} \mathcal{E}(\psi_N) \geq 1/2.$$

Note that when $D = 2$, Conjecture 1 is equivalent to the planted clique conjecture, which is a widely believed conjecture in average-case complexity (Jerrum, 1992; Feige and Krauthgamer, 2003; Barak et al., 2019). The $\mathsf{HPC}_3$ conjecture was used by Zhang and Xia (2018) to show statistical-computational gaps for third order tensor completion; their evidence for the validity of this conjecture was based on the threshold at which the natural spectral method for the problem fails. In a recent paper on the general case $D \geq 3$, Luo and Zhang (2020) showed that MCMC algorithms and methods based on low-degree polynomials—see Hopkins (2018); Kunisky et al. (2019) and the references therein for an introduction to such methods, which comprise a large family of popular algorithms—also fail at this threshold. In concurrent work to that of Luo and Zhang, Brennan and Bresler (2020) showed that the planted clique conjecture with "secret leakage" can be reduced to $\mathsf{HPC}_D$. Recall our definition of the adaptivity index (3); the $\mathsf{HPC}_D$ conjecture implies the following computational lower bound.

THEOREM 1. *Suppose that Conjecture 1 holds, and that $d \geq 2$ is a fixed integer. Then for any sequence of estimators $\{\widehat{\theta}_n\}_{n\geq 1}$ such that $\widehat{\theta}_n$ is computable in time polynomial in $n$, we have*

$$\liminf_{n\to\infty} \frac{\log \mathfrak{A}(\widehat{\theta}_n)}{\log n^{\frac{1}{2}\left(1-1/d\right)}} \geq 1. \tag{12}$$

Assuming Conjecture 1, Theorem 1 thus posits that the adaptivity index of any computationally efficient estimator must grow at least at rate $n^{\frac{1}{2}\left(1-\frac{1}{d}\right)}$, up to sub-polynomial factors in $n$. In particular, this precludes the existence of efficient estimators with adaptivity index bounded poly-logarithmically in $n$. Contrast this with the case of isotonic regression without unknown permutations, where the block-isotonic estimator has adaptivity index[4] of the order $\mathcal{O}(\log^d n)$ (Deng and Zhang, 2020). This demonstrates yet another salient difference in adaptation behavior with and without unknown permutations.

Finally, while Theorem 1 is novel for all $d \geq 3$, we note that when $d = 2$, Shah et al. (2019a) established a computational lower bound for the case where the noise distribution

[4] Deng and Zhang (2020) consider the more general case where the hyper-rectangular partition need not be consistent with the Cartesian product of one-dimensional partitions, but the adaptivity index claimed here can be obtained as a straightforward corollary of their results.

is Bernoulli and the indifference sets are identical along all the dimensions. On the other hand, Theorem 1 applies in the case where the indifference sets induced by the univariate partitions may be different along the different dimensions, and also to the case of Gaussian noise. The latter, technical reduction is accomplished via the machinery of Gaussian rejection kernels introduced by Brennan et al. (2018). This device shares many similarities with other reduction "gadgets" used in earlier arguments (e.g., Berthet and Rigollet (2013); Ma and Wu (2015); Wang et al. (2016)).

We have thus established both the fundamental limits of estimation without computational considerations (5), and a lower bound on the adaptivity index of polynomial time estimators (12). Next, we show that a simple, efficient estimator simultaneously attains both lower bounds for all $d \geq 3$.

3.3. *Achieving the fundamental limits in polynomial time.* We begin with notation that will be useful in defining our estimator. We say that a tuple $\mathsf{bl} = (S_1, \ldots, S_L)$ is a *one-dimensional ordered partition* of the set $[n_1]$ of size $L$ if the sets $S_1, \ldots, S_L \subseteq [n_1]$ are non-empty and pairwise disjoint, with $[n_1] = \bigcup_{\ell=1}^{L} S_\ell$. Note that any such one-dimensional ordered partition induces a partial order, which we denote by $\prec$, on the set $[n_1]$; to be specific, the induced partial order is such that for $a, b \in [n_1]$, we write $a \prec b$ if $a \in S_\ell$ and $b \in S_{\ell'}$ with $\ell < \ell'$. Furthermore, each $S_\ell$, $\ell = 1, \ldots, L$ is an *antichain* of this partial order[5]. As a concrete example, suppose that $n_1 = 6$; then the one-dimensional ordered partition $\mathsf{bl} = (\{2,4,6\}, \{1,5\}, \{3\})$ induces a partial order on $[6]$ with the set of binary relations

$$\{2 \prec 1,\ 2 \prec 5,\ 2 \prec 3,\ 4 \prec 1,\ 4 \prec 5,\ 4 \prec 3,\ 6 \prec 1,\ 6 \prec 5,\ 6 \prec 3,\ 1 \prec 3,\ 5 \prec 3\}.$$

The antichains of this partial order are indeed $\{2,4,6\}$, $\{1,5\}$, and $\{3\}$.

Denote the set of all one-dimensional ordered partitions of size $L$ by $\mathfrak{P}_L$, and let $\mathfrak{P} := \bigcup_{L=1}^{n_1} \mathfrak{P}_L$. Note that any one-dimensional ordered partition of size $L$ induces a map $\sigma_{\mathsf{bl}} : [n_1] \to [L]$, where $\sigma_{\mathsf{bl}}(i)$ is the index $\ell$ of the set $S_\ell \ni i$. In the example above, we have $\sigma_{\mathsf{bl}}(1) = \sigma_{\mathsf{bl}}(5) = 2$ and $\sigma_{\mathsf{bl}}(3) = 3$. Now given $d$ ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d \in \mathfrak{P}$, define

$$\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) := \Bigg\{ \theta \in \mathbb{R}_{d,n} : \ \exists f \in \mathcal{F}_d \ \text{ such that } \forall i_j \in [n_j],\ j \in [d], \\ \theta(i_1, \ldots, i_d) = f\left( \frac{\sigma_{\mathsf{bl}_1}(i_1)}{n_1}, \ldots, \frac{\sigma_{\mathsf{bl}_d}(i_d)}{n_d} \right) \Bigg\}.$$

In other words, the set[6] $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ represents all tensors that are piecewise constant on the hyper-rectangles[7] $\prod_{j=1}^{d} \mathsf{bl}_j$, while also being coordinate-wise isotonic on the partial orders specified by $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$. We refer to any such hyper-rectangular partition of the lattice $\mathbb{L}_{d,n}$ that can be written in the form $\prod_{j=1}^{d} \mathsf{bl}_j$ as a *$d$-dimensional ordered partition*.

Our estimator computes various statistics of the observation tensor $Y$, and we require some more terminology to define these precisely. For each $j \in [d]$, define the vector $\widehat{\tau}_j \in \mathbb{R}^{n_j}$ of "scores", whose $k$-th entry is given by

$$\widehat{\tau}_j(k) := \sum_{i_1, \ldots, i_d \in [n_1]} Y(i_1, \ldots, i_d) \cdot \mathbb{I}\{i_j = k\}. \tag{13a}$$

[5]Recall that a subset of a partially ordered set is an antichain if no two elements in the subset are comparable with each other in the partial order.

[6]Note that we have abused notation slightly in defining the sets $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ and $\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)$ similarly to each other. The reader should be able to disambiguate the two from context, depending on whether the arguments are ordered partitions or permutations.

[7]Note that $\prod_{j=1}^{d} \mathsf{bl}_j = \{\prod_{j=1}^{d} S_j \mid S_j \in \mathsf{bl}_j, j \in [d]\}$.

The score vector $\widehat{\tau}_j$ provides noisy information about the permutation $\pi_j^*$. In order to see this clearly, it is helpful to specialize to the noiseless case $Y = \theta^*$, in which case we obtain the population scores

$$\tau_j^*(k) := \sum_{i_1,\ldots,i_d \in [n_1]} \theta^*(i_1,\ldots,i_d) \cdot \mathbb{I}\{i_j = k\}. \tag{13b}$$

One can verify that the entries of the vector $\tau_j^*$ are increasing when viewed along permutation $\pi_j^*$, i.e., that $\tau_j^*(\pi_j^*(1)) \le \cdots \le \tau_j^*(\pi_j^*(n_j))$.

For each pair $k, \ell \in [n_j]$, also define the pairwise statistics

$$\widehat{\Delta}_j^{\mathsf{sum}}(k,\ell) := \widehat{\tau}_j(\ell) - \widehat{\tau}_j(k) \quad \text{and} \tag{14a}$$

$$\widehat{\Delta}_j^{\max}(k,\ell) := \max_{(i_q)_{q\neq j} \in \prod_{q\neq j}[n_q]} \left\{ Y(i_1,\ldots,i_{j-1},\ell,i_{j+1},\ldots,i_d) - Y(i_1,\ldots,i_{j-1},k,i_{j+1},\ldots,i_d) \right\}. \tag{14b}$$

Given that the scores provide noisy information about the unknown permutation, the statistic $\widehat{\Delta}_j^{\mathsf{sum}}(k,\ell)$ provides noisy information about the event $\{\pi_j^*(k) < \pi_j^*(\ell)\}$, i.e., a large positive value of $\widehat{\Delta}_j^{\mathsf{sum}}(k,\ell)$ provides evidence that $\pi_j^*(k) < \pi_j^*(\ell)$ and a large negative value indicates otherwise. Now clearly, the scores are not the sole carriers of information about the unknown permutations; for instance, the statistic $\widehat{\Delta}_j^{\max}(k,\ell)$ measures the maximum difference between *individual* entries and a large, positive value of this statistic once again indicates that $\pi_j^*(k) < \pi_j^*(\ell)$. The statistics (14) thus allow us to distinguish pairs of indices, and our algorithm is based on precisely this observation. Finally, recall that similarly to before, one may define an antichain of a directed acyclic graph: for any pair of nodes in the antichain, there is no directed path in the graph going from one node to the other.

Having set up the necessary notation, we are now ready to describe the algorithm formally.

---

**Algorithm: Mirsky partition estimator**

I. (Partition estimation): For each $j \in [d]$, perform the following steps:

a. Create a directed graph $G_j'$ with vertex set $[n_j]$ and add the edge $u \to v$ if either

$$\widehat{\Delta}_j^{\mathsf{sum}}(u,v) > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} \quad \text{or} \quad \widehat{\Delta}_j^{\max}(u,v) > 8\sqrt{\log n}. \tag{15a}$$

If $G_j'$ has cycles, then prune the graph and only keep the edges corresponding to the first condition above, i.e.,

$$u \to v \quad \text{if and only if} \quad \widehat{\Delta}_j^{\mathsf{sum}}(u,v) > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}. \tag{15b}$$

Let $G_j$ denote the pruned graph.

b. Compute a one-dimensional ordered partition $\widehat{\mathsf{bl}}_j$ as the minimal partition of the vertices of $G_j$ into disjoint antichains, via Mirsky's algorithm (Mirsky, 1971).

II. (Piecewise constant isotonic regression): Project the observations on the set of isotonic functions that are consistent with the blocking obtained in step I to obtain

$$\widehat{\theta}_{\mathsf{MP}} = \operatorname*{argmin}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d)} \ell_n^2(Y,\theta).$$

---

Some discussion of the pruning step is in order. Note that at the end of step Ia, the graph $G_j$ is guaranteed to have no cycles, since the pruning step is based exclusively on the score

vector $\widehat{\tau}_j$. The purpose of the pruning step is precisely to accomplish this, and there are other heuristics that might also work. For instance, if the graph $G'_j$ consists of disjoint cycles, then the pruning step can instead proceed by pruning each cycle individually. As will be make clear in the proof, the probability that the graph $G'_j$ is pruned is vanishingly small, and so the exact mechanics of the pruning step are not crucial to the algorithm.

Owing to its acyclic structure, the vertices of graph $G_j$ can always be decomposed as the union of disjoint antichains, since a directed acyclic graph defines a partial order on its vertices in the natural way. The presence of an edge $u \to v$ indicates that $u \succ v$ in the partial order, and the acyclic nature of the graph ensures that there are no inconsistencies.

Let us now describe the intuition behind the estimator as a whole. On each dimension $j$, we produce a partial order on the set $[n_j]$. We employ the statistics (14) in order to determine such a partial order, with two indices placed in the same block if they cannot be distinguished based on these statistics. This partitioning step serves a dual purpose: first, it discourages us from committing to orderings over indices when our observations on these indices look similar, and second, it serves to cluster indices that belong to the same indifference set, since the statistics (14) computed on pairs of indices lying in the same indifference set are likely to have small magnitudes. Once we have determined the partial order via Mirsky's algorithm, we project our observations onto isotonic tensors that are piecewise constant on the $d$-dimensional partition specified by the individual partial orders. We note that the Mirsky partition estimator presented here derives some inspiration from existing estimators. For instance, the idea of associating a partial order with the indices has appeared before (Pananjady et al., 2020; Mao et al., 2020), and variants of the pairwise statistics (14) have been used in prior work on permutation estimation (Flammarion et al., 2019; Mao et al., 2020). However, to the best of our knowledge, no existing estimator computes a partition of the indices into antichains: a natural idea that significantly simplifies both the algorithm—speeding it up considerably when there are a small number of indifference sets (see the following paragraph for a discussion)—and its analysis.

We now turn to a discussion of the computational complexity of this estimator. Suppose that we compute the score vectors $\widehat{\tau}_j, j \in [d]$ first, which takes $\mathcal{O}(dn)$ operations. Now for each $j \in [d]$, step I of the algorithm can be computed in time $\mathcal{O}(n_j^2)$, since it takes $\mathcal{O}(n_j^2)$ operations to form the graph $G_j$, and Mirsky's algorithm (Mirsky, 1971) for the computation of a "dual Dilworth" decomposition into antichains runs in time $\mathcal{O}(n_j^2)$. Thus, the total computational complexity of step I is given by $\mathcal{O}(d \cdot n_1^2)$. Step II of the algorithm involves an isotonic projection onto a partially ordered set. As we establish in Lemma 8 in the supplementary material, such a projection can be computed by first averaging the entries of $Y$ on the hyper-rectangular blocks formed by the $d$-dimensional ordered partition $\prod_{j=1}^{d} \widehat{\mathsf{bl}}_j$, and then performing multivariate isotonic regression on the result. The first operation takes linear time $\mathcal{O}(n)$, and the second operation is a weighted isotonic regression problem that can be computed in time $\widetilde{\mathcal{O}}(B^{3/2})$ if there are $B$ blocks in the $d$-dimensional ordered partition (Kyng et al., 2015). Now clearly, $B \leq n$, so that step II of the Mirsky partition estimator has worst-case complexity $\widetilde{\mathcal{O}}(n^{3/2})$. Thus, the overall estimator (from start to finish) has worst-case complexity $\widetilde{\mathcal{O}}(n^{3/2})$. Furthermore, we show in Lemma 4 that if $\theta^* \in \mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n})$, then $B \leq s$ with high probability, and on this event, step II only takes time $\mathcal{O}(n) + \widetilde{\mathcal{O}}(s^{3/2})$. When $s$ is small, the overall complexity of the Mirsky partition procedure is therefore dominated by that of computing the scores, and given by $\mathcal{O}(dn)$ with high probability. Thus, the computational complexity also adapts to underlying structure.

Having discussed its algorithmic properties, let us now turn to the risk bounds enjoyed by the Mirsky partition estimator. Recall, once again, the notation $k^* = \min_{j \in [d]} k^j_{\max}$.

THEOREM 2. *There is a universal positive constant $C$ such that for all $d \geq 2$:*
*(a) We have the worst-case risk bound*

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}_{\mathsf{MP}}, \theta^*) \leq C \left\{ n^{-1/d} \log^{5/2} n + d^2 n^{-\frac{1}{2}(1-1/d)} \log n \right\}. \tag{16}$$

*(b) We have the adaptive bound*

$$\sup_{\theta^* \in \mathcal{M}^{\mathbf{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}_{\mathsf{MP}}, \theta^*) \leq \frac{C}{n} \left\{ s + d^2(n_1 - k^*) \cdot n^{\frac{1}{2}\left(1-\frac{1}{d}\right)} \right\} \log n. \tag{17a}$$

*Consequently, the estimator $\widehat{\theta}_{\mathsf{MP}}$ has adaptivity index bounded as*

$$\mathfrak{A}(\widehat{\theta}_{\mathsf{MP}}) \leq C d^2 \cdot n^{\frac{1}{2}\left(1-\frac{1}{d}\right)} \log n. \tag{17b}$$

When taken together, the two parts of Theorem 2 characterize both the risk and adaptation behaviors of the Mirsky partition estimator $\widehat{\theta}_{\mathsf{MP}}$. Let us discuss some particular consequences of these results, starting with part (a) of the theorem. When $d = 2$, we see that the second term of equation (16) dominates the bound, leading to a risk of order $n^{-1/4}$. Comparing with the minimax lower bound (5), we see that this is sub-optimal by a factor $n^{1/4}$. There are other estimators that attain strictly better rates (Mao et al., 2020; Liu and Moitra, 2020), but to the best of our knowledge, it is not yet known whether the minimax lower bound (5) can be attained by an estimator that is computable in polynomial time. On the other hand, for $d \geq 3$, the first term of equation (16) dominates, and we achieve the lower bound on the minimax risk (5) up to a poly-logarithmic factor. Thus, the case $d \geq 3$ of this problem is distinctly different from the bivariate case: The minimax risk is achievable with a computationally efficient algorithm in spite of the fact that there are more permutations to estimate in higher dimensions. This surprising behavior can be reconciled with prevailing intuition by two high-level observations. First, as $d$ grows, the isotonic function becomes much harder to estimate, so we are able to tolerate more sub-optimality in estimating the permutations. Second, in higher dimensional problems, a single permutation perturbs large blocks of the tensor, and this allows us to obtain more information about it than when $d = 2$. Both of these observations are made quantitative and precise in the proof.

As a side note, we believe that the logarithmic factor in the bound (16) can be improved; one way to do so is to use other isotonic regression estimators (like the bounded LSE) in step II of our algorithm. But since our notion of adaptation requires an estimator that performs well even when the signal is unbounded, we have used the vanilla isotonic LSE in step II.

Turning our attention now to part (b) of the theorem, notice that we achieve the lower bound (12) on the adaptivity index of polynomial time procedures up to a sub-polynomial factor in $n$. Such a result was not known, to the best of our knowledge, for any $d \geq 3$. Even when $d = 2$, the Count-Randomize-Least-Squares (CRL) estimator of Shah et al. (2019a) was shown to have adaptivity index bounded by $\widetilde{\mathcal{O}}(n^{1/4})$ over a sub-class of *bounded* bivariate isotonic matrices with unknown permutations that are also piecewise constant on two-dimensional ordered partitions $\mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{2,n}) \cap \mathbb{B}_\infty(1)$. As we show in Proposition 5 presented in Appendix A of the supplementary material, the Mirsky partition estimator is also adaptive in this case, and attains an adaptivity index that significantly improves upon the best bound known for the CRL estimator in terms of the logarithmic factor. In particular, the adaptivity index $n^{1/4} \log^8 n$ for the CRL estimator[8] is improved to $n^{1/4} \log^{5/4} n$ for

[8]To be clear, this is the best known upper bound on the adaptivity index of the CRL estimator due to Shah et al. (2019a). These results, in turn, rely on the adaptation properties of the bivariate isotonic least squares estimator (Chatterjee et al., 2018), and it is not clear if they can be improved substantially.

the Mirsky partition estimator $\widehat{\theta}_{\mathsf{MP}}$ and further to $n^{1/4}\log n$ for a bounded variant (see Remark 1). Appendix A of the supplementary material also establishes some other adaptation properties for a variant of the CRL estimator in low dimensions. An even starker difference between the adaptation properties of the CRL and Mirsky partition estimators is evident in higher dimensions. We show in Theorem 3 to follow that for higher dimensional problems with $d \geq 4$, the CRL estimator has strictly sub-optimal adaptivity index. Thus, in an overall sense, the Mirsky partition estimator is better equipped to adapt to indifference set structure than the CRL estimator.

Let us also briefly comment on the proof of part (b) of the theorem, which has several components that are novel to the best of our knowledge. We begin by employing a decomposition of the error of the estimator in terms of the sum of estimation and approximation errors; while there are also compelling aspects to our bound on the estimation error, let us showcase some interesting components involved in bounding the approximation error. The first key insight is a structural result (given as Lemma 8 in Appendix B of the supplementary material) that allows us to write step II of the algorithm as a composition of two simpler steps. Besides having algorithmic consequences (alluded to in our discussion of the running time of the Mirsky partition estimator), Lemma 8 allows us to write the approximation error as a sum of two terms corresponding to the two simpler steps of this composition. In bounding these terms, we make repeated use of a second key component: Mirsky's algorithm groups the indices into clusters of disjoint antichains, so our bound on the approximation error incurred on any single block of the partition makes critical use of the condition (15a) used to accomplish this clustering. Our final key component, which is absent from proofs in the literature to the best of our knowledge, is to handle the approximation error on unbounded mean tensors $\theta^*$, which is crucial to establishing that the bound (17a) holds in expectation—this is, in turn, necessary to provide a bound on the adaptivity index. This component requires us to leverage the pruning condition (15b) of the algorithm in conjunction with some careful conditioning arguments.

Taking both parts of Theorem 2 together, then, we have produced a computationally efficient estimator that is both worst-case optimal when $d \geq 3$ and optimally adaptive among the class of computationally efficient estimators. Let us now turn to other natural estimators for this problem, and assess their worst-case risk, computation, and adaptation properties.

3.4. *Adaptation properties of existing estimators.* Arguably, the most natural estimator for this problem is the global least squares estimator $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y)$, which corresponds to the maximum likelihood estimator in our setting with Gaussian errors. The worst-case risk behavior of the LSE over the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{n,d}) \cap \mathbb{B}_\infty(1)$ was already discussed in Proposition 2(a): It attains the minimax lower bound (5) up to a poly-logarithmic factor. However, computing such an estimator is NP-hard in the worst-case even when $d = 2$, since the notoriously difficult max-clique instance can be straightforwardly reduced to the corresponding quadratic assignment optimization problem (see, e.g., Pitsoulis and Pardalos (2001) for reductions of this type).

Another class of procedures consists of two-step estimators that first estimate the unknown permutations defining the model, and then the underlying isotonic function. Estimators of this form abound in prior work (Chatterjee and Mukherjee, 2019; Shah et al., 2019a; Pananjady et al., 2020; Mao et al., 2020; Liu and Moitra, 2020). We unify such estimators under Definition 1 to follow, but first, let us consider a particular instance of such an estimator in which the permutation-estimation step is given by a multidimensional extension of the Borda or Copeland count. A close relative of such an estimator has been analyzed when $d = 2$ (Chatterjee and Mukherjee, 2019).

**Algorithm: Borda count estimator**

I. (Permutation estimation): Recall the score vectors $\widehat{\tau}_1, \ldots, \widehat{\tau}_d$ from (13a). Let $\widehat{\pi}_j^{\mathsf{BC}}$ be any permutation along which the entries of $\widehat{\tau}_j$ are non-decreasing; i.e.,

$$\widehat{\tau}_j\big(\widehat{\pi}_j^{\mathsf{BC}}(k)\big) \leq \widehat{\tau}_j\big(\widehat{\pi}_j^{\mathsf{BC}}(\ell)\big) \text{ for all } 1 \leq k \leq \ell \leq n_j.$$

II. (Isotonic regression): Project the observations onto the class of isotonic tensors that are consistent with the permutations obtained in step I to obtain

$$\widehat{\theta}_{\mathsf{BC}} := \operatorname*{argmin}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1^{\mathsf{BC}}, \ldots, \widehat{\pi}_d^{\mathsf{BC}})} \ell_n^2(Y, \theta).$$

The rationale behind the estimator is simple: If we were given the true permutations $(\pi_1^*, \ldots, \pi_d^*)$, then performing isotonic regression on the permuted observations $Y\{(\pi_1^*)^{-1}, \ldots, (\pi_d^*)^{-1}\}$ would be the most natural thing to do. Thus, a natural idea is to *plug-in* permutation estimates $(\widehat{\pi}_1^{\mathsf{BC}}, \ldots, \widehat{\pi}_d^{\mathsf{BC}})$ of the true permutations. The computational complexity of this estimator is dominated by the isotonic regression step, and is thus given by $\widetilde{\mathcal{O}}(n^{3/2})$ (Kyng et al., 2015). The following proposition provides an upper bound on the worst-case risk of this estimator over bounded tensors in the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$.

PROPOSITION 4. *There is a universal positive constant $C$ such that for each $d \geq 2$, we have*

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}_{\mathsf{BC}}, \theta^*) \leq C \cdot \left( n^{-1/d} \log^{5/2} n + d^2 n^{-\frac{1}{2}(1-1/d)} \right). \tag{18}$$

A few comments are in order. First, note that a variant of this estimator has been analyzed previously in the case $d = 2$, but with the bounded isotonic LSE in step II instead of the (unbounded) isotonic LSE (Chatterjee and Mukherjee, 2019). When $d = 2$, the second term of equation (18) dominates the bound and Proposition 4 establishes the rate $n^{-1/4}$, without the logarithmic factor present in Chatterjee and Mukherjee (2019).

Second, note that when $d \geq 3$, the first term of equation (18) dominates the bound, and comparing this bound with the minimax lower bound (5), we see that the Borda count estimator is minimax optimal up to a poly-logarithmic factor for all $d \geq 3$. In this respect, it resembles both the full least squares estimator $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y)$ and the Mirsky partition estimator $\widehat{\theta}_{\mathsf{MP}}$.

Unlike the Mirsky partition estimator, however, both the global LSE and the Borda count estimator are unable to adapt optimally to indifference sets. This is a consequence of a more general result that we state after the following definition.

DEFINITION 1 (Permutation-projection based estimator). *We say that an estimator $\widehat{\theta}$ is permutation-projection based if it can be written as either*

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1, \ldots, \widehat{\pi}_d)} \ell_n^2(Y, \theta) \qquad \text{or} \qquad \widehat{\theta} = \operatorname*{argmin}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1, \ldots, \widehat{\pi}_d) \cap \mathbb{B}_\infty(1)} \ell_n^2(Y, \theta)$$

*for a tuple of permutations $(\widehat{\pi}_1, \ldots, \widehat{\pi}_d)$. These permutations may be chosen in a data-dependent fashion.*

The bounded LSE (6), the global LSE, and the Borda count estimator are permutation-projection based, as is the CRL estimator of Shah et al. (2019a). The Mirsky partition estimator, on the other hand, is not. The following theorem proves a lower bound on the adaptivity index of any permutation-projection based estimator.

THEOREM 3. *For each $d \geq 4$, there is a pair of constants $(c_d, C_d)$ that depend only on the dimension $d$ such that for each $n \geq C_d$ and any permutation-projection based estimator $\widehat{\theta}$, we have*

$$\mathfrak{A}(\widehat{\theta}) \geq c_d \cdot n^{1-2/d}.$$

For each $d \geq 4$, we have $n^{1-2/d} \gg n^{\frac{1}{2}(1-1/d)}$, so by comparing Theorem 3 with Theorem 1, we see that no permutation-projection based estimator can attain the smallest adaptivity index possible for polynomial time algorithms. In fact, even the global LSE, which is not computable in polynomial time to the best of our knowledge, falls short of the polynomial time benchmark of Theorem 1.

On the other hand, when $d = 2$, we note once again that Shah et al. (2019a) leveraged the favorable adaptation properties of the bivariate isotonic LSE (Chatterjee et al., 2018) to show that their CRL estimator has the optimal adaptivity index for polynomial time algorithms over the class $\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{2,n}) \cap \mathbb{B}_\infty(1)$. They also showed that the bounded LSE (6) does not adapt optimally in this case. In higher dimensions, however, even the isotonic LSE—which must be employed within any permutation-projection based estimator—has poor adaptation properties (Han et al., 2019), and this leads to our lower bound in Theorem 3.

The case $d = 3$ represents a transition between these two extremes, where the isotonic LSE adapts sub-optimally, but a good enough adaptivity index is still achievable owing to the lower bound of Theorem 1. Indeed, we show in Proposition 7 in Appendix A of the supplementary material that a variant of the CRL estimator also attains the polynomial time optimal adaptivity index for this case. Consequently, a result as strong as Theorem 3—valid for all permutation-projection based estimators—cannot hold when $d = 3$.

3.5. *Adaptation of the Mirsky partition estimator in the bounded case.* The careful reader would have noticed that our results on adaptation hold for *unbounded* signals, and as such, do not recover our minimax results in the bounded case (see the discussion following Proposition 3). This raises the natural question of whether one can show adaptation results for signals that are piecewise constant on hyper-rectangles but also uniformly bounded.

In order to answer this question, let us first define the adaptivity index over a hierarchy of bounded sets $\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$. Begin by defining the minimax risk

$$\overline{\mathfrak{M}}_{d,n}(\mathbb{k}, \mathbf{s}) := \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}, \theta^*).$$

Now for an estimator $\widehat{\theta} \in \widehat{\Theta}$, let

$$\overline{\mathfrak{A}}^{\mathbb{k},\mathbf{s}}(\widehat{\theta}) := \frac{\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}, \theta^*)}{\overline{\mathfrak{M}}_{d,n}(\mathbb{k}, \mathbf{s})} \quad \text{and} \tag{19a}$$

$$\overline{\mathfrak{A}}(\widehat{\theta}) := \max_{\mathbf{s} \in \mathbb{L}_{d,n}} \max_{\mathbb{k} \in \mathbb{K}_{\mathbf{s}}} \overline{\mathfrak{A}}^{\mathbb{k},\mathbf{s}}(\widehat{\theta}). \tag{19b}$$

With these definitions set up, we are now ready to state our main result of this section: an adaptation result for the Mirsky partition estimator for bounded, two-dimensional signals.

PROPOSITION 5. *Let $d = 2$. There is a universal positive constant $C$ such that the Mirsky partition estimator satisfies*

$$\sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}_{\mathsf{MP}}, \theta^*) \leq \frac{C}{n} \cdot (n_1 - k^* + 1) \cdot n^{1/4} \log^{5/4} n.$$

*Consequently*[9]*, we have*

$$\overline{\mathfrak{A}}(\widehat{\theta}_{\mathsf{MP}}) \leq C \cdot n^{1/4} \log^{5/4} n.$$

Let us begin by comparing[10] Proposition 5 to the results of Shah et al. (2019a). Assuming the planted clique conjecture, Shah et al. (2019a, Theorem 3) show a lower bound on the (bounded) adaptivity index of any polynomial time procedure. Proposition 5 shows that the Mirsky partition estimator matches this bound up to a poly-logarithmic factor, thereby achieving the smallest adaptivity index achievable for any polynomial time procedure. Comparing Proposition 5 with Shah et al. (2019a, Theorem 2), we also see that in the bounded case, the Mirsky partition estimator significantly improves the logarithmic factor in the best-known upper bound, from $\log^8 n$ (for their CRL estimator), to $\log^{5/4} n$. In fact, the following remark shows that an even smaller adaptivity index can be achieved.

REMARK 1. *If step II of the Mirsky partition estimator is changed to a projection onto the bounded set $\mathcal{M}(\mathbb{L}_{d,n}; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) \cap \mathbb{B}_\infty(1)$, then it can be shown by repeating the steps of our proof of Proposition 5 and using metric entropy bounds from the proof of Proposition 1 that the resulting estimator $\widehat{\theta}^{\mathsf{bd}}_{\mathsf{MP}}$ satisfies*

$$\sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}^{\mathsf{bd}}_{\mathsf{MP}}, \theta^*) \leq \frac{C}{n} \cdot (n_1 - k^* + 1) \cdot n^{1/4} \log n,$$

*leading to the adaptivity index*

$$\overline{\mathfrak{A}}(\widehat{\theta}^{\mathsf{bd}}_{\mathsf{MP}}) \leq C \cdot n^{1/4} \log n.$$

Having established results when $d = 2$, let us now turn to a discussion of the general case $d \geq 3$. By straightforward modifications to our arguments used to prove Proposition 5, it is possible to prove a general upper bound of the form

$$\sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}_{\mathsf{MP}}, \theta^*) \leq \frac{C}{n} \Big\{ \min\big(s, n^{1-1/d} \log^{3/2} n\big) + d^2 (n_1 - k^*) \cdot n^{\frac{1}{2}\left(1 - \frac{1}{d}\right)} \Big\} \log n. \tag{20}$$

However, in order to turn the guarantee (20) into a bound on the adaptivity index $\overline{\mathfrak{A}}(\widehat{\theta}_{\mathsf{MP}})$, we would require a corresponding minimax lower bound over the set $\mathcal{M}^{\mathbb{k},\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$. Now such a lower bound appears to be particularly challenging to obtain even in the case *without unknown permutations*, and is likely to exhibit an intricate dependence on the pair $(\mathbb{k}, \mathbf{s})$ over and above just the number of pieces $s$. Obtaining a sharp guarantee on this minimax risk—and subsequently, providing a sharper analysis of the Mirsky partition estimator to attempt to match this guarantee—are both interesting open problems that we discuss in more detail below.

[9] The reason for this consequence is an existing minimax lower bound (Shah et al., 2019a), and is made clear in the proof.

[10] When making this comparison, note the differences between our notation and theirs: we consider $n_1 \times n_1$ matrices with $n = n_1^2$, while Shah et al. (2019a) work with $n \times n$ matrices.

**4. Discussion.** We considered the problem of estimating a multivariate isotonic regression function on the lattice from noisy observations that were also permuted along each coordinate, and established several results. In this section, we summarize these results, and discuss some related and open questions.

*Summary of results.* First, we showed that unlike in the bivariate case, computationally efficient estimators are able to achieve the minimax lower bound for estimation of bounded tensors in this class. Second, when the tensor is also structured, in that it is piecewise constant on a $d$-dimensional partition with a small number of blocks, we showed that the fundamental limits of adaptation are still nonparametric. Third, by appealing to the hypergraph planted clique conjecture, we also argued that the adaptivity index of polynomial time estimators is significantly poorer than that of their inefficient counterparts. The second and third phenomena are both significantly different from the case without unknown permutations. Fourth, we introduced a novel Mirsky partition estimator that was simultaneously optimal both in worst-case risk and adaptation, while being computable in sub-quadratic time. This procedure also enjoys better adaptation properties than existing estimators when $d = 2$ (see Appendix A of the supplementary material), and its computational complexity adapts to structure in the underlying tensor. Our results for the Mirsky partition estimator are particularly surprising given that a large class of natural estimators does not exhibit fast adaptation in the multivariate case. Finally, we also established risk bounds and structural properties (see Appendix B of the supplementary material) for natural isotonic regression estimators without unknown permutations.

*Dependence on signal strength.* Let us briefly comment on a particular facet of our results that was not emphasized in Section 3.1: The dependence of the derived rates on the signal-to-noise ratio of the problem. In particular, suppose that the true signal $\theta^* \in \mathbb{B}_\infty(r)$ for some positive scalar $r$. Then how do our results in Propositions 1 and 2 change? By carefully repeating the steps in the respective proofs, it can be shown that for all $d \geq 2$, the BLSE over the class of isotonic functions with unknown permutations achieves the worst-case risk bound

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r)} \mathcal{R}_n(\widehat{\theta}_{\mathsf{BLSE}}, \theta^*) \leq C \cdot (rn^{-1/d} \log^2 n + n^{-(1-1/d)} \log n).$$

On the other hand, the LSE (without boundedness constraints) is shown by our techniques (for all $d \geq 2$) to achieve the risk bounds

(21a)
$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y), \theta^*\big) \leq C \cdot (\sqrt{\log n} + r) \cdot (rn^{-1/d} \log^2 n + n^{-(1-1/d)} \log n)$$

and

(21b)
$$\sup_{\theta^* \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}), Y), \theta^*\big) \leq C \cdot \Big\{(\sqrt{\log n} + r) \cdot rn^{-1/d} \log^2 n + n^{-1}\Big\}$$

over the sets defined with and without unknown permutations, respectively.

It is instructive to compare the latter bound (21b) on the vanilla isotonic regression estimator with the one that can be derived from the proof of Han et al. (2019, Theorem 1). There, the authors show the worst case bound

(22)
$$\sup_{\theta^* \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}), Y), \theta^*\big) \leq C \cdot (rn^{-1/d} \log^4 n + n^{-2/d} \log^8 n).$$

Comparing the bounds (21b) and (22) in terms of their dependence on $r$, we see that our bound (21b) is sharper in the regime $r \to 0$, since the error floor is much smaller: $n^{-1} \ll n^{-2/d}\log^8 n$. On the other hand, the bound (22) is better when $r$ is very large, i.e., growing with $n$. Since both bounds are on the same estimator, one can combine them to obtain the guarantee

$$\sup_{\theta^* \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}), Y), \theta^*\big) \\ \leq C \cdot \Big(rn^{-1/d}\log^{5/2} n\Big) \cdot \min\Big\{\sqrt{\log n} + r, \log^2 n\Big\} + Cn^{-1},$$

which inherits the favorable properties of both bounds.

*Open questions.* Our work raises many interesting questions from both the modeling and theoretical standpoints. From a modeling perspective, the isotonic regression model with unknown permutations should be viewed as just a particular nonparametric model for tensor data. There are many ways one may extend these models. For instance, taking a linear combination of $k > 1$ tensors in the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ directly generalizes the class of nonnegative tensors of (canonical polyadic) rank $k$. Studying such models would parallel a similar investigation that was conducted in the case $d = 2$ for matrix estimation (Shah et al., 2019b). It would also be interesting to incorporate latent permutations within other multidimensional nonparametric function estimation tasks that are not shape constrained; a similar study has been carried out in the case $d = 2$ in the context of graphon estimation (Gao et al., 2015). In the case $d \geq 3$, the analogous application would be in modeling hypergraphs in a flexible manner, going beyond existing models involving planted partitions (Abbe and Montanari, 2013; Ghoshdastidar and Dukkipati, 2017).

Methodological and theoretical questions also abound. First, note that in typical applications, $n_1$ will be very large, and we will only observe a subset of entries chosen at random. Indeed, when $d = 2$, Mao et al. (2020) showed that the fundamental limits of the problem exhibit an intricate dependence on the probability of observing each entry and the dimensions of the tensor. What are the analogs of these results when $d \geq 3$? The second question concerns adaptation. Our focus on indifference sets to define structure in the tensor was motivated by the application to multiway comparisons, but other structures are also interesting to study. For instance, what does a characterization of adaptation look like when there is simply a partition into hyper-rectangles—not necessarily Cartesian products of one-dimensional partitions—on which the tensor is piecewise constant? Such structure has been extensively studied in the isotonic regression literature (Chatterjee et al., 2018; Han et al., 2019; Deng and Zhang, 2020). What about cases where $\theta^*$ is a nonnegative tensor of rank 1? It would be worth studying spectral methods for tensor estimation for this problem, especially in the latter case. Finally, an interesting open question is whether the block-isotonic regression estimator of Fokianos et al. (2020) can be employed in conjunction with permutation estimation to yield an estimator that is minimax optimal as well as adaptive. For instance, we could replace step II in the Borda count estimator with the block-isotonic regression estimator (Fokianos et al., 2020; Deng and Zhang, 2020), and call this estimator $\widehat{\theta}_{\mathsf{block}}$. Note that $\widehat{\theta}_{\mathsf{block}}$ is *not* permutation-projection based, so it is possible that it achieves the optimal adaptivity index for polynomial time algorithms while remaining minimax optimal. On the other hand, the best existing algorithms for the block-isotonic estimator require time $\mathcal{O}(n^3)$, as opposed to our estimation procedure that runs in time $\widetilde{\mathcal{O}}(n^{3/2})$ in the worst-case, and faster if the problem is structured. From a technical perspective, understanding the behavior of the estimator $\widehat{\theta}_{\mathsf{block}}$ in our setting is intricately related to the oracle properties of the block-isotonic regression estimator around permuted versions of isotonic tensors.

Finally, let us discuss in more detail two independently interesting questions in shape-constrained estimation. The first was raised in the context of adaptation in Section 3.5. Can we obtain a characterization of the minimax risk of estimation over the set $\mathcal{M}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$ *as a function of* the pair $(\mathbb{k}, \mathbf{s})$? In spite of multiple investigations of related issues (Chatterjee et al., 2015; Bellec and Tsybakov, 2015; Chatterjee et al., 2018; Gao et al., 2020), this question is complementary and does not seem to have been addressed (or even asked) in the literature. While a complete answer to this question would make significant progress towards characterizing adaptation in the bounded case (with unknown permutations), the question is one of independent interest even in the case of univariate isotonic regression, as witnessed by the following examples. First, suppose that the partition into $s$ pieces induces blocks of equal sizes, i.e., $k_1 = \cdots = k_s = n/s$. For this pair $(\mathbb{k}, \mathbf{s})$, existing results (see, e.g., Bellec and Tsybakov (2015)) show that we have

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}, \theta^*) \gtrsim \frac{1}{n} \min\bigl(s, n^{1/3}\bigr), \tag{23}$$

and it can be shown that this bound is matched by the block-wise isotonic estimator that first averages the observations within each block and then performs isotonic regression on the result. Indeed, the vanilla isotonic regression estimator (without averaging within blocks) also achieves the same rate up to a logarithmic factor (Chatterjee et al., 2015). On the other hand, consider the second case in which the pair $(\mathbb{k}, \mathbf{s})$ satisfies $k_1 = \cdots = k_{s-1} = 1$ and $k_s = n - (s-1)$. By treating the first $s-1$ entries of the problem as standard isotonic regression and setting the last $n-(s-1)$ entries (deterministically) to 1, one can establish the minimax lower bound

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}, \theta^*) \gtrsim \frac{(s-1)^{1/3} + 1}{n}. \tag{24}$$

Once again, this bound can be achieved by the block-wise isotonic regression estimator. Comparing the bounds (23) and (24), we see that the minimax risk in this case must depend on the values of the block sizes $k_1, \ldots, k_s$, and not just the number of blocks $s$. Characterizing the risk as a function of $\mathbb{k}$ and $\mathbf{s}$ (especially in the general multivariate case) is thus likely to be a challenging problem.

The second question is about measuring adaptation with respect to a larger class of structured tensors. Note that we considered isotonic tensors with piecewise constant structure on a hyper-rectangular partition that was formed by indifference sets along different dimensions, i.e., a Cartesian product of univariate partitions. While our focus on this type of structure was motivated by the application to multi-way comparisons—in which each block of the univariate partition represents a set of items among which we are indifferent—a more general type of structure has been studied in the isotonic regression literature (without unknown permutations), in which we have a general hyper-rectangular partition that is not necessarily a Cartesian product of univariate partitions (Chatterjee et al., 2018; Han et al., 2019; Deng and Zhang, 2020). An interesting open question is to derive analogs of Proposition 3 and Theorems 2 and 3 under this more general notion of structure. The key difference in such a bound is that while the tuple $\mathbb{k} = (\mathbf{k}_1, \ldots, \mathbf{k}_d)$ of indifference set sizes (and the associated functionals $s$ and $k^*$) suffice to characterize structure in the tensor in the setting of the current paper, a different set of quantities would be needed to measure complexity in this more general class of tensors. Note that one can always refine a general hyper-rectangular partition into a Cartesian product of univariate partitions, though this can increase the number of hyper-rectangular pieces exponentially in the dimension and yield suboptimal rates.

**Acknowledgments.** We thank the anonymous referees for their constructive feedback, which improved the scope and presentation of the paper.

## SUPPLEMENTARY MATERIAL

**Supplement to: Isotonic regression with unknown permutations: Statistics, computation, and adaptation:**
(doi: COMPLETED BY THE TYPESETTER; .pdf). In the supplementary document (Pananjady and Samworth, 2021+), we provide proofs of all our main results stated in the body of the paper, and also some auxiliary results of independent interest.

## REFERENCES


Abbe, E. and Montanari, A. (2013). Conditional random fields, planted constraint satisfaction and entropy concentration, *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques*, Springer, pp. 332–346.

Abid, A. and Zou, J. (2018). A stochastic expectation-maximization approach to shuffled linear regression, *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, pp. 470–477.

Airoldi, E. M., Costa, T. B. and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation, *Advances in Neural Information Processing Systems*, pp. 692–700.

Andersen, C. M. and Bro, R. (2003). Practical aspects of PARAFAC modeling of fluorescence excitation-emission data, *Journal of Chemometrics: A Journal of the Chemometrics Society* **17**(4): 200–215.

Angelini, M. C., Caltagirone, F., Krzakala, F. and Zdeborová, L. (2015). Spectral detection on sparse hypergraphs, *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, pp. 66–73.

Balabdaoui, F., Doss, C. R. and Durot, C. (2020). Unlinked monotone regression, *arXiv preprint arXiv:2007.00830* .

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

Baltrunas, L., Makcinskas, T. and Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering, *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, ACM, New York, USA, pp. 119–126.

Barak, B., Hopkins, S., Kelner, J., Kothari, P. K., Moitra, A. and Potechin, A. (2019). A nearly tight sum-of-squares lower bound for the planted clique problem, *SIAM Journal on Computing* **48**(2): 687–735.

Behr, M. and Munk, A. (2017). Minimax estimation in linear models with unknown finite alphabet design, *arXiv preprint arXiv:1711.04145* .

Bellec, P. C. and Tsybakov, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation., *The Journal of Machine Learning Research* **16**: 1879–1892.

Ben-Akiva, M. E., Lerman, S. R. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, Vol. 9, MIT press.

Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension, *The Annals of Statistics* **41**(4): 1780–1815.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators, *Probability Theory and Related Fields* **97**(1-2): 113–150.

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

Brennan, M. and Bresler, G. (2020). Reducibility and statistical-computational gaps from secret leakage, *in* J. Abernethy and S. Agarwal (eds), *Proceedings of 33rd Conference on Learning Theory*, Vol. 125 of *Proceedings of Machine Learning Research*, PMLR, pp. 648–847.

Brennan, M., Bresler, G. and Huleihel, W. (2018). Reducibility and computational lower bounds for problems with planted sparse structure, *in* S. Bubeck, V. Perchet and P. Rigollet (eds), *Proceedings of the 31st Conference On Learning Theory*, Vol. 75 of *Proceedings of Machine Learning Research*, PMLR, pp. 48–166.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters, *Ann. Math. Statist.* **26**(4): 607–616.

Carpentier, A. and Schlueter, T. (2016). Learning relationships between data obtained independently, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 658–666.

Chan, S. and Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models, *International Conference on Machine Learning*, pp. 208–216.

Chandukala, S. R., Kim, J., Otter, T. and Allenby, G. M. (2008). *Choice Models in Marketing: Economic Assumptions, Challenges and Trends*, Now Publishers Inc.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding, *The Annals of Statistics* **43**(1): 177–214.

Chatterjee, S., Guntuboyina, A. and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems, *The Annals of Statistics* **43**(4): 1774–1800.

Chatterjee, S., Guntuboyina, A. and Sen, B. (2018). On matrix estimation under monotonicity constraints, *Bernoulli* **24**(2): 1072–1100.

Chatterjee, S. and Mukherjee, S. (2019). Estimation in tournaments and graphs under monotonicity constraints, *IEEE Transactions on Information Theory* **65**(6): 3525–3539.

Chen, X., Bennett, P. N., Collins-Thompson, K. and Horvitz, E. (2013). Pairwise ranking aggregation in a crowd-sourced setting, *Proceedings of the sixth ACM international conference on Web search and data mining*, ACM, pp. 193–202.

Collier, O. and Dalalyan, A. S. (2016). Minimax rates in permutation estimation for feature matching, *The Journal of Machine Learning Research* **17**(1): 162–192.

Craswell, N., Zoeter, O., Taylor, M. and Ramsey, B. (2008). An experimental comparison of click position-bias models, *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 87–94.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1): 20–28.

DeGroot, M. H. and Goel, P. K. (1980). Estimation of the correlation coefficient from a broken random sample, *The Annals of Statistics* **8**(2): 264–278.

Deng, H. and Zhang, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs, *The Annals of Statistics* **48**(6): 3672 – 3698.

Dwork, C., Kumar, R., Naor, M. and Sivakumar, D. (2001). Rank aggregation methods for the web, *Proceedings of the 10th International Conference on World Wide Web*, ACM, pp. 613–622.

Dykstra, R. L. and Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables, *The Annals of Statistics* **10**(3): 708–716.

Farias, V. F., Jagabathula, S. and Shah, D. (2013). A nonparametric approach to modeling choice with limited data, *Management Science* **59**(2): 305–322.

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

Fishburn, P. C. (1973). Binary choice probabilities: on the varieties of stochastic transitivity, *Journal of Mathematical Psychology* **10**(4): 327–352.

Flammarion, N., Mao, C. and Rigollet, P. (2019). Optimal rates of statistical seriation, *Bernoulli* **25**(1): 623–653.

Fokianos, K., Leucht, A. and Neumann, M. H. (2020). On integrated $l^1$ convergence rate of an isotonic regression estimator for multivariate observations, *IEEE Transactions on Information Theory* **66**(10): 6389–6402.

Gao, C., Han, F. and Zhang, C.-H. (2020). On estimation of isotonic piecewise constant signals, *The Annals of Statistics* **48**(2): 629–654.

Gao, C., Lu, Y. and Zhou, H. H. (2015). Rate-optimal graphon estimation, *The Annals of Statistics* **43**(6): 2624–2652.

Ghoshdastidar, D. and Dukkipati, A. (2017). Consistency of spectral hypergraph partitioning under planted partition model, *The Annals of Statistics* **45**(1): 289–315.

Han, Q. (2021+). Set structured global empirical risk minimizers are rate optimal in general dimensions, *The Annals of Statistics, to appear* .

Han, Q., Wang, T., Chatterjee, S. and Samworth, R. J. (2019). Isotonic regression in general dimensions, *The Annals of Statistics* **47**(5): 2440–2471.

Hardouin, J.-B. and Mesbah, M. (2004). Clustering binary variables in subscales using an extended Rasch model and Akaike information criterion, *Communications in Statistics-Theory and Methods* **33**(6): 1277–1294.

Hopkins, S. B. K. (2018). *Statistical inference and the sum of squares method*, PhD thesis, Cornell University.

Hsu, D. J., Shi, K. and Sun, X. (2017). Linear regression without correspondence, *Advances in Neural Information Processing Systems*, pp. 1531–1540.

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

Kim, A. K. H. and Samworth, R. J. (2016). Global rates of convergence in log-concave density estimation, *The Annals of Statistics* **44**(6): 2756–2779.

Kök, A. G., Fisher, M. L. and Vaidyanathan, R. (2015). Assortment planning: Review of literature and industry practice, *in* N. Agrawal and S. A. Smith (eds), *Retail Supply Chain Management: Quantitative Models and Empirical Studies*, Springer US, Boston, MA, pp. 175–236.

Kolda, T. G., Bader, B. W. and Kenny, J. P. (2005). Higher-order web link analysis using multilinear algebra, *Fifth IEEE International Conference on Data Mining (ICDM'05)*, IEEE, pp. 8–pp.

Krantz, D. H. (1965). *The Scaling of Small and Large Color Differences.*, PhD thesis, University of Pennsylvania.

Kunisky, D., Wein, A. S. and Bandeira, A. S. (2019). Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio, *arXiv preprint arXiv:1907.11636* .

Kyng, R., Rao, A. and Sachdeva, S. (2015). Fast, provable algorithms for isotonic regression in all $\ell_p$-norms, *Advances in Neural Information Processing Systems*, pp. 2719–2727.

Lepski, O. (1991). On a problem of adaptive estimation in Gaussian white noise, *Theory of Probability & Its Applications* **35**(3): 454–466.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation, *The Annals of Statistics* **25**(6): 2512–2546.

Liu, A. and Moitra, A. (2020). Better algorithms for estimating non-parametric models in crowd-sourcing and rank aggregation, *in* J. Abernethy and S. Agarwal (eds), *Proceedings of 33rd Conference on Learning Theory*, Vol. 125 of *Proceedings of Machine Learning Research*, PMLR, pp. 2780–2829.

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

Luo, Y. and Zhang, A. R. (2020). Tensor clustering with planted structures: Statistical optimality and computational limits, *arXiv preprint arXiv:2005.10743v2* .

Ma, R., Cai, T. T. and Li, H. (2021+). Optimal estimation of bacterial growth rates based on a permuted monotone matrix, *Biometrika, to appear* .

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

Mao, C., Pananjady, A. and Wainwright, M. J. (2018). Breaking the $1/\sqrt{n}$ barrier: Faster rates for permutation-based models in polynomial time, *in* S. Bubeck, V. Perchet and P. Rigollet (eds), *Proceedings of the 31st Conference On Learning Theory*, Vol. 75 of *Proceedings of Machine Learning Research*, PMLR, pp. 2037–2042.

Mao, C., Pananjady, A. and Wainwright, M. J. (2020). Towards optimal estimation of bivariate isotonic matrices with unknown permutations, *The Annals of Statistics* **48**(6): 3183 – 3205.

Marschak, J. and Davidson, D. (1957). Experimental tests of stochastic decision theory, *Technical report*, Cowles Foundation for Research in Economics, Yale University.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, *in* P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, pp. 105–142.

McFadden, D. (1981). Econometric models of probabilistic choice, *in* C. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT press, pp. 198–272.

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

Mirsky, L. (1971). A dual of Dilworth's decomposition theorem, *The American Mathematical Monthly* **78**(8): 876–877.

Möcks, J. (1988). Decomposing event-related potentials: A new topographic components model, *Biological Psychology* **26**(1-3): 199–215.

Negahban, S., Oh, S., Thekumparampil, K. K. and Xu, J. (2018). Learning from comparisons and choices, *The Journal of Machine Learning Research* **19**(1): 1478–1572.

Nemirovski, A. S., Polyak, B. T. and Tsybakov, A. B. (1985). Convergence rate of nonparametric estimates of maximum-likelihood type, *Problemy peredachi informatsii* **21**(4): 17–33.

Nguyen, A., Piech, C., Huang, J. and Guibas, L. (2014). Codewebs: Scalable homework search for massive open online programming courses, *Proceedings of the 23rd International Conference on World Wide Web*, pp. 491–502.

Pananjady, A., Mao, C., Muthukumar, V., Wainwright, M. J. and Courtade, T. A. (2020). Worst-case versus average-case design for estimation from partial pairwise comparisons, *The Annals of Statistics* **48**(2): 1072–1097.

Pananjady, A. and Samworth, R. J. (2021+). Supplement to Isotonic regression with unknown permutations: Statistics, computation, and adaptation, *The Annals of Statistics, to appear* .

Pananjady, A., Wainwright, M. J. and Courtade, T. A. (2017a). Denoising linear models with permuted data, *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 446–450.

Pananjady, A., Wainwright, M. J. and Courtade, T. A. (2017b). Linear regression with shuffled data: Statistical and computational limits of permutation recovery, *IEEE Transactions on Information Theory* **64**(5): 3286–3300.

Pitsoulis, L. and Pardalos, P. M. (2001). Quadratic assignment problem, *in* C. A. Floudas and P. M. Pardalos (eds), *Encyclopedia of Optimization*, Springer US, Boston, MA, pp. 2075–2107.

Plackett, R. L. (1975). The analysis of permutations, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **24**(2): 193–202.

Rakhlin, A., Sridharan, K. and Tsybakov, A. B. (2017). Empirical entropy, minimax regret and minimax risk, *Bernoulli* **23**(2): 789–824.

Rigollet, P. and Weed, J. (2019). Uncoupled isotonic regression via minimum Wasserstein deconvolution, *Information and Inference: A Journal of the IMA* **8**(4): 691–717.

Robertson, T. and Wright, F. (1975). Consistency in generalized isotonic regression, *The Annals of Statistics* **3**(2): 350–362.

Shah, N. B., Balakrishnan, S., Guntuboyina, A. and Wainwright, M. J. (2017). Stochastically transitive models for pairwise comparisons: Statistical and computational issues, *IEEE Transactions on Information Theory* **63**(2): 934–959.

Shah, N. B., Balakrishnan, S. and Wainwright, M. J. (2019a). Feeling the Bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons, *IEEE Transactions on Information Theory* **65**(8): 4854–4874.

Shah, N. B., Balakrishnan, S. and Wainwright, M. J. (2019b). Low permutation-rank matrices: Structural properties and noisy completion, *Journal of Machine Learning Research* **20**(101): 1–43.

Shah, N. B., Balakrishnan, S. and Wainwright, M. J. (2021). A permutation-based model for crowd labeling: Optimal estimation and robustness, *IEEE Transactions on Information Theory* **67**(6): 4162–4184.

Stout, Q. F. (2015). Isotonic regression for multiple independent variables, *Algorithmica* **71**(2): 450–470.

Tversky, A. (1972). Elimination by aspects: A theory of choice., *Psychological Review* **79**(4): 281.

Unnikrishnan, J., Haghighatshoar, S. and Vetterli, M. (2018). Unlabeled sensing with random linear measurements, *IEEE Transactions on Information Theory* **64**(5): 3237–3253.

van de Geer, S. A. (2000). *Applications of Empirical Process Theory*, Vol. 91, Cambridge University Press.

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory, *Political Analysis* **11**(2): 139–163.

Wang, T., Berthet, Q. and Samworth, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components, *The Annals of Statistics* **44**(5): 1896–1930.

Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits, *IEEE Transactions on Information Theory* **64**(11): 7311–7338.

Zhang, C.-H. (2002). Risk bounds in isotonic regression, *The Annals of Statistics* **30**(2): 528–555.

Zhou, D., Huang, J. and Schölkopf, B. (2007). Learning with hypergraphs: Clustering, classification, and embedding, *Advances in Neural Information Processing Systems*, pp. 1601–1608.

Zhou, J., Bhattacharya, A., Herring, A. H. and Dunson, D. B. (2015). Bayesian factorizations of big sparse tensors, *Journal of the American Statistical Association* **110**(512): 1562–1576.