Learning sparse mixtures of permutations from noisy information

Anindya De ANINDYAD@CIS.UPENN.EDU
University of Pennsylvania

Ryan O'Donnell ODONNELL@CS.CMU.EDU
Carnegie Mellon University

Rocco A. Servedio ROCCO@CS.COLUMBIA.EDU
Columbia University

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the problem of learning an unknown mixture of k permutations over n elements, given access to noisy samples drawn from the unknown mixture. We consider a range of different noise models, including natural variants of the "heat kernel" noise framework and the Mallows model. We give an algorithm which, for each of these noise models, learns the unknown mixture to high accuracy under mild assumptions and runs in $n^{O(\log k)}$ time. Our approach is based on a new procedure that recovers an unknown mixture of permutations from noisy higher-order marginals.

Keywords: mixture models, rankings, noise-tolerant learning

1. Introduction

One of the simplest distribution learning problems is that of learning an unknown distribution f with support size k over some domain $\mathcal{D}$. It is well known that given independent samples drawn from f, with sample and time complexity $O(k/\varepsilon^2)$ the standard empirical estimator produces a hypothesis $\widehat{f}$ such that the total variation distance $\|f - \widehat{f}\|_1$ is at most ε.

Despite the simplicity of the above scenario, it gives rise to a rich landscape of challenging problems once *noise* enters the picture. In more detail, suppose that the learning algorithm only receives *noisy samples*, i.e. each draw from f is independently corrupted by some type of noise before it is given to the learning algorithm. Is it still possible to efficiently recover the underlying sparse distribution f? As we now describe, this question captures several important and well-studied problems in learning theory. We shall refer to this general class of problems as "population recovery" problems, since they may be viewed as the problem of recovering the underlying "population" (along with the frequencies) of the k objects in $\mathcal{D}$ that comprise the support of the distribution f.

One well studied variant of this problem is the problem of *population recovery problem over the discrete cube*. Here the domain $\mathcal{D}$ is $\{0,1\}^n$, and the standard noise model is the *Bonami-Beckner noise operator* (see e.g. O'Donnell (2014)),which independently flips each coordinate with some fixed probability. Motivated by the simplicity and elegance of this problem and its connection to DNF learning, this problem has of late been extensively studied in the theoretical computer science community, see e.g. Dvir et al. (2012); Wigderson and Yehudayoff (2012); Moitra and Saks (2013); Lovett and Zhang (2015); De et al. (2016).

Beyond population recovery over the discrete cube, by varying the choice of the domain $\mathcal{D}$ and the type of noise in question, the above general "population recovery" formulation captures many other well-studied problems in machine learning and statistics:

1. When $\mathcal{D} = \mathbb{R}^n$ and the noise is given by (convolution with) a standard Gaussian, then the corresponding population recovery problem is the problem of learning a Gaussian mixture model with identity covariances, which is a problem of central interest in algorithmic statistics Dasgupta (1999); Regev and Vijayaraghavan (2017); Hopkins and Li (2018); Kothari et al. (2018).

2. When $\mathcal{D} = \mathbb{R}^2$ and the noise is given by the Bessel function, then the resulting problem is equivalent to the mathematical formulation of the so-called "diffraction limit" (see Chen and Moitra (2020)).

3. When $\mathcal{D} = \mathbb{S}_n$ (the symmetric group on n elements) and the noise is given by the so-called Mallows model, then the resulting problem is the same as the problem of learning a mixture of Mallows models[1] Mallows (1957). This problem has attracted significant attention in theoretical machine learning Awasthi et al. (2014); Liu and Moitra (2018); Braverman and Mossel (2008); Lu and Boutilier (2011) (see also Jiao and Vert (2018); Kondor and Barbosa (2010); Murphy and Martin (2003), which consider closely related settings).

This paper studies the population recovery problem on $\mathbb{S}_n$; our main contribution is a single unified efficient algorithm that succeeds for several different noise models. These noise models include the standard "heat kernel" noise model Diaconis (1988a); Kondor and Lafferty (2002); Kondor and Barbosa (2010); Jiao and Vert (2018) and the Ewens model Ewens (1972); Fligner and Verducci (1986); Diaconis and Hanlon (1992) (which is a natural variant of the Mallows model), among others.

The population recovery problem over $\mathbb{S}_n$: We now formally describe the population recovery problems over $\mathbb{S}_n$ that we consider. We model the noise by a family of distributions $\mathcal{K}_\theta$, where each distribution $\mathcal{K}_\theta$ is supported on $\mathbb{S}_n$. Here θ is a model parameter capturing the "noise rate" (we will have much more to say about this for each of the specific noise models we consider below). Given a fixed noise distribution $\mathcal{K}_\theta$, an instance of the population recovery problem is defined by

(a) k unknown weights $w_1, \ldots, w_k \geq 0$ such that $w_1 + \ldots + w_k = 1$; and

(b) k unknown permutations $\sigma_1, \ldots, \sigma_k \in \mathbb{S}_n$.

Let $f : \mathbb{S}_n \to \mathbb{R}^{\geq 0}$ denote the function (distribution) which is w_i at σ_i and 0 otherwise. The learner gets *noisy samples* from f where each sample is independently generated by first choosing a random permutation $\boldsymbol{\sigma} \sim f$; then independently drawing a random $\boldsymbol{\pi} \sim \mathcal{K}_\theta$; and finally, providing the learner with the permutation $\boldsymbol{\pi}\boldsymbol{\sigma} \in \mathbb{S}_n$. We write "$\mathcal{K}_\theta * f$" to denote the distribution over noisy samples described above, and the goal of the learner is to approximately recover f given independent noisy samples of the above sort. The reader may verify that the distribution defined by $\boldsymbol{\pi}\boldsymbol{\sigma}$ is precisely given by the group convolution $\mathcal{K}_\theta * f$ (and hence the notation).

1.1. Motivation

The population recovery framework is a basic one in the theory of distribution learning, and as we have argued earlier, the framework is able to capture a wide range of problems in unsupervised learning. Since the domain $\mathcal{D} = \mathbb{S}_n$ is one of the most fundamental and basic *non-commutative*

1. (with identical noise parameter)

domains, it is natural to study the population recovery problem over this domain. In particular, for the domain $\mathcal{D} = \{0,1\}^n$, much of the progress to date has been achieved using analytic methods, in particular methods of discrete Fourier analysis Dvir et al. (2012); Wigderson and Yehudayoff (2012); Moitra and Saks (2013); Lovett and Zhang (2015); De et al. (2016). Moving to the non-commutative domain $\mathbb{S}_n$ presents many new challenges, but as our results show, it is possible to adapt various methods of discrete Fourier analysis to the domain $\mathbb{S}_n$ using the representation theory of $\mathbb{S}_n$. This is a topic which has been studied in great depth in algebra and combinatorics (see e.g. James (2006); Méliot (2017)); we give a gentle introduction to this area in Appendix G.

The non-commutativity of $\mathbb{S}_n$ gives rise to some intriguing features of the population recovery problem over this domain (which are not present over the discrete cube $\{0,1\}^n$). In particular, for the Ewens noise model our upper and lower bounds (Theorem 4 and Theorem 5) together show that the sample complexity of population recovery at noise rate θ is inversely related to the *fractional part of* e^θ; hence for Ewens noise the complexity of the population recovery problem is not a monotonic function of the noise rate θ. This is in sharp contrast to the behavior of the population recovery problem over the hypercube.

Another motivation for studying the population recovery problem on $\mathbb{S}_n$ comes from its connections to the problem of learning mixture models of rankings. To elaborate on this connection, suppose that there are k subgroups in a population and for each $1 \leq i \leq k$,

1. the i^{th} subgroup has an unknown "central preference order" specifying a ranking over a fixed set of n items (equivalently, there is an unknown permutation $\sigma_i \in \mathbb{S}_n$ for the i-th subgroup);

2. the fraction of the i^{th} subgroup in the population is an unknown parameter $w_i \geq 0$.

Suppose the preference order of a random individual in the i^{th} subgroup is given by a *noisy version* of σ_i. Modeling the noise by the distribution $\mathcal{K}_\theta$, the random variable $\mathcal{K}_\theta * f$ (where f is the distribution putting weight w_i on the permutation σ_i) is the same as the preference order of a random individual from the entire population. Thus, the population recovery problem is now the problem of recovering the central preferences of the subgroups along with their weights in the population. This problem is known as the task of *learning mixtures of ranking models*, and is a well-studied problem in algorithmic machine learning Braverman and Mossel (2008); Awasthi et al. (2014); Liu and Moitra (2018); Chierichetti et al. (2015). Many different noise models (i.e., choices of $\mathcal{K}_\theta$) have been studied in the literature Mallows (1957); Fligner and Verducci (1986); Mukherjee (2016); Kondor and Lafferty (2002); Awasthi et al. (2014); Jiao and Vert (2018), including some of the noise models studied in the current paper. (We note that the noise model that has been most studied in this context in the machine learning community is is the Mallows noise model Braverman and Mossel (2008); Awasthi et al. (2014); Liu and Moitra (2018). As we discuss later, our current techniques are applicable only when the noise model $\mathcal{K}_\theta$ is a class function; since the Mallows model is not a class function, our technique does not yield an algorithmic result for mixtures of Mallows models.)

1.2. The noise models that we consider

We consider a range of different noise models, corresponding to different choices for the parametric family $\{\mathcal{K}_\theta\}$, and we give a single unified algorithm which, for each of the three noise models, can efficiently recover the population in the presence of that kind of noise. In this subsection we detail the three specific noise models that we will work with (though as we discuss later, our general mode of analysis could be applied to other noise models as well).

(A.) **Symmetric noise.** In the *symmetric noise* model, the parametric family of distributions over $\mathbb{S}_n$ is denoted $\{\mathcal{S}_{\overline{p}}\}_{\overline{p} \in \Delta^n}$. Given a vector $\overline{p} = (p_0, \ldots, p_n) \in \Delta^n$ (so each $p_i \geq 0$ and $\sum_{i=0}^n p_i = 1$), a draw of $\boldsymbol{\pi} \sim \mathcal{S}_{\overline{p}}$ is obtained as follows:

1. Choose $0 \leq \boldsymbol{j} \leq n$, where value j is chosen with probability p_j.
2. Choose a uniformly random subset $\boldsymbol{A} \subseteq [n]$ of size exactly $\boldsymbol{j}$. Draw $\boldsymbol{\pi}$ uniformly from $\mathbb{S}_{\boldsymbol{A}}$; in other words, $\boldsymbol{\pi}$ is a uniformly random permutation over the set $\mathcal{A}$ and is the identity permutation on elements in $[n] \setminus \boldsymbol{A}$. (We denote this uniform distribution over $\mathbb{S}_{\boldsymbol{A}}$ by $\mathbb{U}_{\boldsymbol{A}}$.)

Note that in this model, if the noise vector $\overline{p}$ has $p_n = 1$, then every draw from $\mathcal{S}_{\overline{p}} * f$ is a uniform random permutation and there is no useful information available to the learner.

In order to define the next two noise models that we consider, let us recall the notion of a *right-invariant* metric on $\mathbb{S}_n$. Such a metric $d(\cdot, \cdot)$ is one that satisfies $d(\sigma, \pi) = d(\sigma\tau, \pi\tau)$ for all $\sigma, \pi, \tau \in \mathbb{S}_n$. We note that a metric is right-invariant if and only if it is invariant under relabeling of the items $1, \ldots, n$, and that most metrics considered in the literature satisfy this condition (see Kumar and Vassilvitskii (2010); Diaconis (1988b) for discussions of this point). In this paper, for technical convenience we restrict our attention to the metric $d(\cdot, \cdot)$ being the *Cayley distance* over $\mathbb{S}_n$ (though see Section 1.6 for a discussion of how our methods and results could potentially be generalized to other right-invariant metrics):

Definition 1 *Let G be the undirected graph with vertex set $\mathbb{S}_n$ and an edge between permutations σ and π if there is a transposition τ such that $\sigma = \tau \cdot \pi$. The* Cayley distance *over $\mathbb{S}_n$ is the metric induced by this graph; in other words, $d(\pi, \sigma) = t$ where t is the smallest value such that there are transpositions $\tau_1, \ldots, \tau_t$ satisfying $\sigma = \tau_1 \cdots \tau_t \pi$.*

Now we are ready to define the next two parameterized families of noise distributions that we consider. We note that each of the noise distributions $\mathcal{K}$ considered below has the natural property that $\mathbf{Pr}_{\boldsymbol{\pi} \sim \mathcal{K}}[\boldsymbol{\pi} = \pi]$ decreases with $d(\pi, e)$ where e is the identity distribution.

(B.) **Heat kernel random walk under Cayley distance.** Let $\mathcal{L}$ be the Laplacian of the graph G from Definition 1. Given a "temperature" parameter $t \in \mathbb{R}^+$, the *heat kernel* is the $n! \times n!$ matrix $H_t = e^{-t\mathcal{L}}$. It is well known that H_t is the transition matrix of the random walk induced by choosing a Poisson-distributed time parameter $\mathbf{T} \sim \mathsf{Poi}(t)$ and then taking $\mathbf{T}$ steps of a uniform random walk in the graph G. With this motivation, we define the *heat kernel noise model* as follows: the parametric family of distributions is $\{\mathcal{H}_t\}_{t \in \mathbb{R}^+}$, where the probability weight that $\mathcal{H}_t$ assigns to permutation π is the probability that the above-described random walk, starting at the identity permutation $e \in \mathbb{S}_n$, reaches π. (Observe that higher temperature parameters t correspond to higher rates of noise. More precisely, it is well known that the mixing time of a uniform random walk on G is $\Theta(n \log n)$ steps, so if t grows larger than $n \log n$ then the distribution $\mathcal{H}_t$ converges rapidly to the uniform distribution on $\mathbb{S}_n$; see Diaconis and Shahshahani (1981) for detailed results along these lines.) We note that these probability distributions (or more precisely, the associated heat kernel H_t) have been previously studied in the context of learning rankings, see e.g. Kondor and Lafferty (2002); Kondor and Barbosa (2010); Jiao and Vert (2018). In some of this work, a different underlying distance measure was used over $\mathbb{S}_n$ rather than the Cayley distance; see our discussion of related work in Section 1.4.

(C.) **A Mallows-type model under Cayley distance: the Ewens model.** While the heat kernel noise model arises naturally from an analyst's perspective, a somewhat different model, called the *Mallows model*, has been more popular in the statistics and machine learning literature. The Mallows model is defined using the "Kendall τ-distance" $K(\cdot,\cdot)$ between permutations (defined in Section 1.4) rather than the Cayley distance $d(\cdot,\cdot)$; the Mallows model with parameter $\theta > 0$ assigns probability weight $e^{-\theta K(\pi,e)}/Z_K(\theta)$ to the permutation π, where $Z_k(\theta) = \sum_{\pi\in\mathbb{S}_n} e^{-\theta K(\pi,e)}$ is a normalizing constant. As proposed by Fligner and Verducci Fligner and Verducci (1986), it is natural to consider generalizations of the Mallows model in which other distance measures take the place of the Kendall τ-distance. The model which we consider is one in which the Cayley distance is used as the distance measure; so given $\theta > 0$, the noise distribution $\mathcal{E}_\theta$ which we consider assigns weight $e^{-\theta d(\pi,e)}/Z(\theta)$ to each permutation $\pi \in \mathbb{S}_n$, where $Z(\theta) = \sum_{\pi\in\mathbb{S}_n} e^{-\theta d(\pi,e)}$ is a normalizing constant. In fact, this noise model was already proposed in 1972 by W. Ewens in the context of population genetics Ewens (1972) and has been intensively studied in that field (according to Google Scholar, Ewens (1972) has been cited more than 2000 times). We observe that for the Ewens model $\mathcal{E}_\theta$, in contrast with the heat kernel noise model now *smaller* values of θ correspond to higher levels of noise, and that when $\theta = 0$ the distribution $\mathcal{E}_\theta$ is simply the uniform distribution over $\mathbb{S}_n$ and there is no useful information available to the learner.

1.3. Our results

We present a general algorithm which, for each of the noise models defined above, provably recovers the unknown permutations $\sigma_1,\ldots,\sigma_k$ and associated mixing weights $w_1,\ldots,w_k$ up to high accuracy (under a mild technical assumption, that no mixing weight w_i is too small). A notable feature of our results is that the sample and running time dependence is only *quasipolynomial* in the number of elements n and the number of permutations k; as we detail in Section 1.4 below, this is in contrast with recent results for similar problems in which the dependence on k is exponential.

Below we give detailed statements of the various specific results that follow from our algorithmic approach. The following notation and terminology will be used in these statements: for f a distribution over $\mathbb{S}_n$ (or any function from $\mathbb{S}_n$ to $\mathbb{R}$) we write $\mathrm{supp}(f)$ to denote the set of permutations $\sigma \in \mathbb{S}_n$ that have $f(\sigma) \neq 0$. For a given noise model $\mathcal{K}$, we write "$\mathcal{K} * f$" to denote the distribution over noisy samples that is provided to the learning algorithm as described earlier. Given two functions $f, g : \mathbb{S}_n \to \mathbb{R}$, we write "$\|f - g\|_1$" to denote $\sum_{\pi\in\mathbb{S}_n} |f(\pi) - g(\pi)|$, the ℓ_1 distance between f and g. If f and g are both distributions then we write $d_{\mathrm{TV}}(f,g)$ to denote the *total variation distance* between f and g, which is $\frac{1}{2}\|f-g\|_1$. Finally, if f is a distribution over $\mathbb{S}_n$ in which $f(\sigma) > \varepsilon$ for every σ such that $f(\sigma) > 0$, we say that f is *ε-heavy*.

Learning from noisy permutations: Positive and negative results. Our first algorithmic result is for the symmetric noise model (A) defined earlier. Theorem 2, stated below, gives an efficient algorithm as long as the vector $\overline{p}$ is "not too extreme" (i.e. not too biased towards putting almost all of its weight on values very close to n):

Theorem 2 (Algorithm for symmetric noise) *There is an algorithm with the following guarantee: Let f be an unknown ε-heavy distribution over $\mathbb{S}_n$ with $|\mathrm{supp}(f)| \le k$. Let $\overline{p} = (p_0,\ldots,p_n) \in$*

Δ^n *be such that*[2]

$$\sum_{j=0}^{n-\log k} p_j \geq \frac{1}{n^{O(\log k)}}.$$

Given $\overline{p}$*, the value of* $\varepsilon > 0$*, a confidence parameter* $\delta > 0$*, and access to random samples from* $\mathcal{S}_{\overline{p}} * f$*, the algorithm runs in time* $\mathrm{poly}(n^{\log k}, 1/\varepsilon, \log(1/\delta))$ *and with probability* $1 - \delta$ *outputs a distribution* $g : \mathbb{S}_n \to \mathbb{R}$ *such that* $d_{\mathrm{TV}}(f, g) \leq \varepsilon$*.*

Our second algorithmic result, which is similar in spirit to Theorem 2, is for the heat kernel noise model:

Theorem 3 (Algorithm for heat kernel noise) *There is an algorithm with the following guarantee: Let* f *be an unknown* ε*-heavy distribution over* $\mathbb{S}_n$ *with* $|\mathrm{supp}(f)| \leq k$*. Let* $t \in \mathbb{R}^+$ *be any value that is* $O(n \log n)$*. Given* t*, the value of* $\varepsilon > 0$*, a confidence parameter* $\delta > 0$*, and access to random samples from* $\mathcal{H}_t * f$*, the algorithm runs in time* $\mathrm{poly}(n^{\log k}, 1/\varepsilon, \log(1/\delta))$ *and with probability* $1 - \delta$ *outputs a distribution* $g : \mathbb{S}_n \to \mathbb{R}$ *such that* $d_{\mathrm{TV}}(f, g) \leq \varepsilon$*.*

Recalling that the uniform random walk on the Cayley graph of $\mathbb{S}_n$ mixes in $\Theta(n \log n)$ steps, we see that the algorithm of Theorem 3 is able to handle quite high levels of noise and still run quite efficiently (in quasi-polynomial time).

Our third positive result, for the Ewens model, displays an intriguing qualitative difference from Theorems 2 and 3. To state our result, let us define the function $\mathsf{dist} : \mathbb{R}^+ \times \mathbb{N} \to \mathbb{R}^+$ as follows:

$$\mathsf{dist}(\theta, \ell) := \min_{j \in \{1,\ldots,\ell\}} \left|e^\theta - j\right|,$$

so $\mathrm{dist}(\theta, \ell)$ measures the minimum distance between e^θ and any integer in $\{1, \ldots, \ell\}$. Theorem 4 gives an algorithm which can be quite efficient for the Ewens noise model if the noise parameter θ is such that $\mathrm{dist}(\theta, \log k)$ is not too small:

Theorem 4 (Algorithm for the Ewens model) *There is an algorithm with the following guarantee: Let* f *be an unknown* ε*-heavy distribution over* $\mathbb{S}_n$ *with* $|\mathrm{supp}(f)| \leq k$*. Given* $\theta > 0$*, the value of* $\varepsilon > 0$*, a confidence parameter* $\delta > 0$*, and access to random samples from* $\mathcal{E}_\theta * f$*, the algorithm runs in time* $\mathrm{poly}(n^{\log k}, 1/\varepsilon, \log(1/\delta), \mathrm{dist}(\theta, \log k)^{-\sqrt{\log k}})$ *and with probability* $1 - \delta$ *outputs a distribution* $g : \mathbb{S}_n \to \mathbb{R}$ *such that* $d_{\mathrm{TV}}(f, g) \leq \varepsilon$*.*

As alluded to earlier, as θ approaches 0 the difficulty of learning in the $\mathcal{E}_\theta$ noise model increases (and indeed learning becomes impossible at $\theta = 0$); since for small θ we have $\mathrm{dist}(\theta, \ell) \approx \theta$, this is accounted for by the $\mathrm{dist}(\theta, \log k)^{-\sqrt{\log k}}$ factor in our running time bound above. However, for larger values of θ the $\mathrm{dist}(\theta, \log k)^{-\sqrt{\log k}}$ dependence may strike the reader as an unnatural artifact of our analysis: is it really hard to learn when θ is very close to $\ln 2 \approx 0.63147$, easy when θ is very close to $\ln 2.5 \approx 0.91629$, and hard again when θ is very close to $\ln 3 \approx 1.09861$? Perhaps surprisingly, the answer is yes: it turns out that the $\mathrm{dist}(\cdot, \cdot)$ parameter captures a *fundamental barrier* to learning in the Ewens model. We establish this by proving the following lower bound for the Ewens model, which shows that a dependence on dist very similar to the one in Theorem 4 is in fact inherent in the problem:

2. Here and throughout the paper we write "log" to denote log base two and "ln" to denote natural logarithm.

Theorem 5 *Given $j \in \mathbb{N}$, there are infinitely many values of k and $m = m(k)$ such that the following holds for all $\eta, \theta > 0$ such that $|e^{\theta} - j| \leq \eta \leq 1/2$: Let A be any algorithm which, when given access to random samples from $\mathcal{E}_\theta * f$ where f is a distribution over $\mathbb{S}_m$ with $|\mathrm{supp}(f)| \leq k$, with probability at least 0.51 outputs a distribution h over $\mathbb{S}_m$ that has $d_{\mathrm{TV}}(f, h) \leq 0.99$. Then A must use $\eta^{-\Omega\left(\sqrt{\frac{\log k}{\log \log k}}\right)}$ samples.*

1.4. Relation to prior work

Population recovery on the discrete cube: As mentioned earlier, at a thematic level this paper is akin to to the rich body of work on the *population recovery problem* on the discrete cube Dvir et al. (2012); Wigderson and Yehudayoff (2012); Moitra and Saks (2013); Lovett and Zhang (2015); De et al. (2016). Over the discrete cube, the goal of population recovery is to recover an unknown distribution $f : \{0,1\}^n \to \mathbb{R}$ (with support size k), given samples from $T_\mu f$, where $T_\mu(\cdot)$ is the Bonami-Beckner operator with correlation μ (a random sample from $T_\mu f$ is generated by sampling $\boldsymbol{x} \sim f$ and flipping every bit independently with probability $(1-\mu)/2$). De et al. (2016) gave an algorithm to recover f with sample and time complexity $k^{\mathsf{poly}(1/\mu)}$, so the complexity of recovering f with their algorithm is $\mathrm{poly}(k)$ for any $\mu > 0$. In this paper, we obtain an analogue of this result over $\mathbb{S}_n$, but with a quasipolynomial dependence on the sparsity parameter k. A notable difference between our setting and $\{0,1\}^n$ is that for $\mathbb{S}_n$ there is no canonical choice of noise operator; rather, a number of different noise models appear in the literature (depending on the application). Our techniques are well suited to analyzing noise operators defined by class functions.

Mixture models of rankings: As noted earlier, another motivation for studying the population recovery problem over $\mathbb{S}_n$ comes from learning mixture models of rankings. Here the noise distribution models how the ordinal preferences of a homogenous population are distributed around a *central preference order*. Several noise models have been considered in this context including the Ewens model (and generalizations of it) Fligner and Verducci (1986); Murphy and Martin (2003); Mukherjee (2016); Diaconis and Hanlon (1992); Diaconis (1988a); Ewens (1972) and the heat kernel model Kondor and Lafferty (2002); Kondor and Barbosa (2010); Jiao and Vert (2018); the most popular choice is the Mallows model Mallows (1957). We recall that the Mallows model (with noise parameter θ, denoted by $\mathcal{M}_\theta$) is identical to the Ewens model $\mathcal{E}_\theta$ described earlier, except that the Cayley distance used in the description of $\mathcal{E}_\theta$ is replaced by the Kendall-τ distance. While this may seem like a minor difference, as we explain later the population recovery problem for the Ewens model $\mathcal{E}_\theta$ exhibits qualitatively different behavior from the Mallows model $\mathcal{M}_\theta$.

The problem of learning mixture of k-Mallows models has been quite popular in learning theory Braverman and Mossel (2008); Awasthi et al. (2014); Liu and Moitra (2018). Mao and Wu (2020); Liu and Moitra (2018) give algorithms for the population recovery problem with noise $\mathcal{M}_\theta$ with running time $\mathrm{poly}(n, 1/\theta) \cdot \exp(k)$[3]. Thus, in contrast to the models considered in our paper, the currently best known algorithms for the population recovery problem with the Mallows noise distribution have an exponential dependence on the sparsity parameter k.

It is a challenging open problem to extend the analysis in this paper to the population recovery problem with the Mallows noise distribution (see Section 1.6 for details). Here we note that there is a sense in which our results show that the Mallows and Ewens models are fundamentally incomparable. This is because, while the results of Liu and Moitra (2018) show that mixtures of Mallows

3. The algorithms of Mao and Wu (2020); Liu and Moitra (2018) work in the more general settings where different components can have different noise parameters.

models are identifiable whenever each $\theta_i \neq 1$, Theorem 5 shows that mixtures of Ewens models are information-theoretically not identifiable at various larger values of θ such as $\ln 3, \ln 4, \ldots$, even when all of the noise parameters are the same value θ (which is given to the algorithm).

Inference from marginal information: In Theorem 7, we give an efficient algorithm to recover a mixture of k permutations given (roughly speaking) all the $O(\log k)$-way marginals. Further, and crucially for us, getting the marginals to error $\varepsilon/n^{\log k}$ suffices to recover f up an ℓ_1 error ε. Motivated by compressive sensing, the broad question of recovering sparse distributions over $\mathbb{S}_n$ from marginals has also been studied in statistics Jagabathula and Shah (2011); Farias et al. (2012); Chatterjee (2015). A key conceptual innovation in this work is to exploit noisy information about higher order marginals to recover the underlying sparse distribution over $\mathbb{S}_n$. In contrast, most previous work either (i) only uses information about pairwise marginals, or (ii) assumes access to exact higher order marginal information for f. It is an interesting direction to explore how our techniques can be used in that line of work.

1.5. Our techniques

A key notion for our algorithmic approach is that of the *marginal* of a distribution f over $\mathbb{S}_n$:

Definition 6 *Fix $f : \mathbb{S}_n \to [0,1]$ to be some distribution over $\mathbb{S}_n$. Let $t \in \{1,\ldots,n\}$, let $\bar{i} = (i_1,\ldots,i_t)$ be a vector of t distinct elements of $\{1,\ldots,n\}$ and likewise $\bar{j} = (j_1,\ldots,j_t)$. We say the* $(\bar{i},\bar{j})$-marginal of f *is the probability*

$$\Pr_{\boldsymbol{\sigma}\sim f}[\boldsymbol{\sigma}(i_1) = j_1 \text{ and } \cdots \text{ and } \boldsymbol{\sigma}(i_t) = j_t]$$

that for all $\ell = 1,\ldots,t$, the i_ℓ-th element of a random $\boldsymbol{\sigma}$ drawn from f is j_ℓ. When $\bar{i}$ and $\bar{j}$ are of length t we refer to such a probability as a t-way marginal of f.

The first key ingredient of our approach for learning from noisy permutations is a reduction from the problem of learning f (the unknown distribution supported on k permutations $\sigma_1,\ldots,\sigma_k$) given access to samples from $\mathcal{K}*f$, to the problem of estimating t-way marginals (for a not-too-large value of t, roughly $\log k$). More precisely, in Section 2 we give an algorithm which, given the ability to efficiently estimate t-way marginals of f, efficiently computes a high-accuracy approximation for an unknown ε-heavy distribution f with support size at most k (see Theorem 7). This algorithm builds on ideas in the population recovery literature, suitably extended to the domain $\mathbb{S}_n$ rather than $\{0,1\}^n$.

With the above-described reduction in hand, in order to obtain a positive result for a specific noise model $\mathcal{K}$ the remaining task is to develop an algorithm A_{marginal} which, given access to noisy samples from $\mathcal{K} * f$, can reliably estimate the required marginals. In Section 3 we show that if the noise distribution $\mathcal{K}$ (a distribution over $\mathbb{S}_n$) is efficiently samplable, then given samples from $\mathcal{K} * f$, the time required to estimate the required marginals essentially depends on the minimum, over a certain set of matrices arising from the Fourier transform (over the symmetric group $\mathbb{S}_n$) of the noise distribution, of the minimum singular value of the matrix. (See Theorem 8 for a detailed statement.) At this point, we have reduced the algorithmic problem of obtaining a learning algorithm for a particular noise model to the analytic task of lower bounding the relevant singular values. We carry out the required analyses on a noise-model-by-noise-model basis in Sections C, D, and E. These analyses employ ideas and results from the representation theory of the symmetric group and its connections to enumerative combinatorics; we review the necessary background in Appendix G.

To establish our lower bound for the Ewens model, Theorem 5, we exhibit two distributions f_1 and f_2 over the symmetric group such that the distributions of noisy permutations $\mathcal{E}_\theta * f_1$ and $\mathcal{E}_\theta * f_2$ have very small statistical distance from each other. Not surprisingly, the inspiration for this construction also comes from the representation theory of the symmetric group; more precisely, the two above-mentioned distributions are obtained from the character (over the symmetric group) corresponding to a particular carefully chosen partition of $[n]$. A crucial ingredient in the proof is the fact that characters of the symmetric group are rational-valued functions, and hence any character can be split into a positive part and a negative part; details are given in Appendix F.

Finally, we note that whereas some of the earlier results in the literature (such as Awasthi et al. (2014); Braverman and Mossel (2008)) only use 2-way or 3-way marginals of the samples, our approach uses $(\log k)$-way marginals. This is not an artifact of our approach but rather is inherent in the problem we consider (learning mixtures of k permutations); this is because it is possible to construct two distributions f_1 and f_2 over permutations, with disjoint supports each of size at most k, such that all $t = \tilde{\Theta}(\log k)$-way marginals of f_1 and f_2 are identical. (This is an easy consequence of the result of Kuperberg et al. (2017) showing the existence of small t-wise permutation families.) Thus, using $(\log k)$-way marginals is essentially necessary to recover mixtures of k permutations. Indeed, the early results such as Awasthi et al. (2014); Braverman and Mossel (2008) seek only to recover a single hidden permutation or a mixture of two permutations.

1.6. Discussion and future work

In this paper we consider three particular noise models — symmetric noise, heat kernel noise, and Ewens noise — and give an efficient algorithm for these noise models. Looking beyond these specific noise models, though, our approach provides a general framework for obtaining algorithms for learning mixtures of noisy permutations. Indeed, for essentially any efficiently samplable noise distribution $\mathcal{K}$, given access to samples from $\mathcal{K} * f$ our approach reduces the algorithmic problem of learning f to the analytic problem of lower bounding the minimum singular values of matrices arising from the Fourier transform of $\mathcal{K}$ (see Theorem 8). We believe that this technique may be useful in a broader range of contexts, e.g. to obtain results analogous to ours for the original Mallows model or for other noise models.

As is made clear in Sections C, D, and E, the representation-theoretic analysis that we require for our noise models is facilitated by the fact that each of the noise distributions considered in those sections is a *class function* (in other words, the value of the distribution on a given input permutation depends only on the cycle structure of the permutation). There are other models, most prominently the Mallows model, for which the noise distribution is not a class function. Extending the kinds of analyses that we perform to other noise distributions which are not class functions is a technical challenge that we leave for future work.

Finally, another natural question is whether our framework can be adapted to handle mixture models in which each component has a different noise rate. Roughly speaking, the difficulty which arises is that in this more general setting it is no longer possible to express the samples from the mixture model as $\mathcal{K} * f$ (i.e., the noise process and the draw from f are no longer independent). Given this, we cannot use our Claim 17, which shows that samples from the mixture model can be used to efficiently yield certain representations of f which in turn yields marginals of f. Whether our results or approaches can extended to the setting in which different mixture components have different noise rates is an intriguing question for future work.

2. Algorithmic recovery of sparse functions

The main result of this section is the reduction alluded to in Section 1.5. In more detail, we give an algorithm which, given the ability to efficiently estimate t-way marginals, efficiently computes a high-accuracy approximation for an unknown ε-heavy distribution f with support size at most k:

Theorem 7 *Let f be an unknown ε-heavy distribution over $\mathbb{S}_n$ with $|\mathrm{supp}(f)| \leq k$. Suppose there is an algorithm A_{marginal} with the following property: given as input a value $\delta > 0$ and two vectors $\bar{i} = (i_1, \ldots, i_t)$ and $\bar{j} = (j_1, \ldots, j_t)$ each composed of t distinct elements of $\{1, \ldots, n\}$, algorithm A_{marginal} runs in time $T(\delta, t, k, n)$ and outputs an additively $\pm\delta$-accurate estimate of the $(\bar{i}, \bar{j})$-marginal of f (recall Definition 6). Then there is an algorithm A_{learn} with the following property: given the value of ε, algorithm A_{learn} runs in time $\mathrm{poly}(n/\varepsilon, n^{\log k}) \cdot T(\frac{\varepsilon}{2k^{O(\log k)}}, 2\log k, k^2, n)$ and returns a function $g : \mathbb{S}_n \to \mathbb{R}^+$ such that $\|f - g\|_1 \leq \varepsilon$.*

Because of space constraints the proof of Theorem 7 is given in Appendix A. Given Theorem 7, in order to obtain a positive result for a specific noise model $\mathcal{K}$ the remaining task is to develop an algorithm A_{marginal} which, given access to noisy samples from $\mathcal{K} * f$, can reliably estimate the required marginals. The algorithm is given in Section 3 (with its proof in Appendix B) and the detailed analyses establishing its efficiency for each of the noise models (by bounding minimum singular values of certain matrices arising from each specific noise distribution) are given in Sections C, D, and E. To the best of our knowledge, the algorithm A_{learn} of Theorem 7 has not appeared in earlier work, though, as we mention later, it is quite similar to the algorithm of Wigderson-Yehudayoff Wigderson and Yehudayoff (2012) for the population recovery problem over $\{0,1\}^n$. At a higher level, our algorithm uses a so-called "extend and prune" approach which can be traced back to early works in computational learning theory Kushilevitz and Mansour (1993); Goldreich and Levin (1989).

3. Computing limited way marginals from noisy samples

Recall that the noisy ranking learning problems we consider are of the following sort: There is a known noise distribution $\mathcal{K}$ supported on $\mathbb{S}_n$, and an unknown k-sparse ε-heavy distribution $f : \mathbb{S}_n \to [0,1]$. Each sample provided to the learning algorithm is generated by the following probabilistic process: independent draws of $\boldsymbol{\pi} \sim \mathcal{K}$ and $\boldsymbol{\sigma} \sim f$ are obtained, and the sample given to the learner is $(\boldsymbol{\pi}\boldsymbol{\sigma}) \in \mathbb{S}_n$. By the reduction established in Theorem 7, in order to give an algorithm that learns the distribution f in the presence of a particular kind of noise $\mathcal{K}$, it suffices to give an algorithm that can efficiently estimate t-way marginals given samples $\boldsymbol{\pi}\boldsymbol{\sigma} \sim \mathcal{K} * f$.

The main result of this section, Theorem 8, gives such an algorithm. Before stating the theorem we need some terminology and notation and we need to recall some necessary background from representation theory of the symmetric group (see Appendix G for a detailed overview of all of the required background).

First, let $\mathcal{K}$ be a distribution over $\mathbb{S}_n$ (which should be thought of as a noise distribution as described earlier). We say that $\mathcal{K}$ is *efficiently samplable* if there is a $\mathrm{poly}(n)$-time randomized algorithm which takes no input and, each time it is invoked, returns an independent draw of $\boldsymbol{\pi} \sim \mathcal{K}$.

Next, we recall that a *partition* λ of the natural number n (written "$\lambda \vdash n$") is a vector of natural numbers $(\lambda_1, \ldots, \lambda_k)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$ and $\lambda_1 + \ldots + \lambda_k = n$ (see Appendix G.2 for more detail). For two partitions λ and μ of n, we say that μ *dominates* λ, written $\mu \rhd \lambda$, if

$\sum_{j \le i} \mu_j \ge \sum_{j \le i} \lambda_j$ for all $i > 0$ (see Definition 41). Given any $\lambda \vdash n$, let $\mathsf{Up}(\lambda)$ denote the set of all partitions $\mu \vdash n$ such that $\mu \rhd \lambda$.

We recall that a *representation* of the symmetric group $\mathbb{S}_n$ is a group homomorphism from $\mathbb{S}_n$ to $\mathbb{C}^{m \times m}$ (see Appendix G). We further recall that for each partition $\lambda \vdash n$ there is a corresponding *irreducible* representation, denoted ρ_λ (see Appendix G.2). For a matrix M we write $\sigma_{\min}(M)$ to denote the smallest singular value of M. Given a partition $\lambda \vdash n$ we define the value $\sigma_{\min,\mathsf{Up}(\lambda),\mathcal{K}}$ to be

$$\sigma_{\min,\mathsf{Up}(\lambda),\mathcal{K}} := \min_{\mu \in \mathsf{Up}(\lambda)} \sigma_{\min}(\widehat{\mathcal{K}}(\rho_\mu)), \tag{1}$$

the smallest singular value across all Fourier coefficients of the noise distribution of irreducible representations corresponding to partitions that dominate λ. (We recall that the Fourier coefficients of functions over the symmetric group, and indeed over any finite group, are matrices; see Appendix G.2.)

Finally, for $0 \le \ell \le n-1$ we define the partition $\lambda_{\mathsf{hook},\ell} \vdash n$ to be

$$\lambda_{\mathsf{hook},\ell} := (n-\ell, 1, \ldots, 1).$$

Now we can state the main result of this section:

Theorem 8 *Let $\mathcal{K}$ be an efficiently samplable distribution over $\mathbb{S}_n$. Let f be an unknown distribution over $\mathbb{S}_n$. There is an algorithm A_{marginal} with the following properties: A_{marginal} receives as input a parameter $\delta > 0$, a confidence parameter $\tau > 0$, a pair of ℓ-tuples $\bar{i} = (i_1, \ldots, i_\ell) \in [n]^\ell$, $\bar{j} = (j_1, \ldots, j_\ell) \in [n]^\ell$ each composed of ℓ distinct elements, and has access to random samples from $\mathcal{K} * f$. Algorithm A_{marginal} runs in time $\mathsf{poly}(\binom{n}{\ell}, \delta^{-1}, \sigma^{-1}_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{K}}, \log(1/\tau))$ and outputs a value $\kappa_{\bar{i},\bar{j}}$ which with probability at least $1-\tau$ is a $\pm\delta$-accurate estimate of the $(\bar{i},\bar{j})$-marginal of f.*

Because of space constraints we give the proof of Theorem 8 in Appendix B.

3.1. Efficient samplability of our noise distributions

In order to apply Theorem 8 to a particular noise distribution $\mathcal{K}$ we need to confirm that $\mathcal{K}$ is efficiently samplable; we now do this for each of the three noise models that we consider. It is immediate from the definition that it is straightforward (given $\overline{p}$) to efficiently generate a random $\boldsymbol{\sigma}$ drawn from the symmetric noise distribution $\mathcal{S}_{\overline{p}}$, and the same is true for the heat kernel noise distribution $\mathcal{H}_t$.

For the Ewens model $\mathcal{E}_\theta$, the characterization $\mathbf{Pr}_{\boldsymbol{\sigma} \sim \mathcal{E}_\theta}[\boldsymbol{\sigma} = \pi] = e^{-\theta d(\pi,e)}/Z(\theta)$ given earlier does not directly yield an efficient sampling algorithm, since it may be hard to compute or approximate the normalizing factor $Z(\theta) = \sum_{\pi \in \mathbb{S}_n} e^{-\theta d(\pi,e)}$. Instead, we recall (see e.g. Section 2.1 of Diaconis and Saloff-Coste (1998)) that the Metropolis algorithm can be used to efficiently perform a random walk on $\mathbb{S}_n$ whose unique stationary distribution is the Ewens distribution $\mathcal{E}_\theta$. (Each step of the random walk can be carried out efficiently because it is computationally easy to compute the Cayley distance between two permutations: if π is the permutation that brings σ to τ, then the Cayley distance $d(\sigma,\tau)$ is $n - \mathrm{cycles}(\pi)$ where $\mathrm{cycles}(\pi)$ is the number of cycles in π.) It is known (see e.g. Theorem 2 of Diaconis and Hanlon (1992)) that this random walk has rapid convergence, and consequently it is indeed possible to sample efficiently from $\mathcal{E}_\theta$ (up to an exponentially small statistical distance which can be ignored in our applications since our algorithms use a sub-exponential number of samples).

4. Representations of heat kernel, symmetric, and Ewens model noise

In this section we record lower bounds on the smallest singular value for the relevant matrices corresponding to "symmetric noise" $\mathcal{S}_{\overline{p}}$ on $\mathbb{S}_n$; to "heat kernel noise" $\mathcal{H}_t$ at temperature parameter t; and to "Ewens model noise" $\mathcal{E}_\theta$ with parameter θ. Proofs of these lower bounds are given in Appendix C, Appendix D, and Appendix E respectively.

Theorem 9 (Symmetric noise) *Let $\ell \in \{1,\dots,n\}$ and let $\overline{p} = (p_0,\dots,p_n) \in \Delta^n$ (i.e. $\overline{p}$ is a non-negative vector whose entries sum to 1) which is such that*

$$\sum_{j=0}^{n-\ell} p_j \geq \kappa.$$

Then (recalling Equation (1)) we have that

$$\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{S}_{\overline{p}}} \geq \frac{\kappa}{n^\ell}. \tag{2}$$

Theorem 10 (Heat kernel noise) *Let $t \geq 1$ and let $\ell \in \{1,\dots,cn\}$ for some suitably small universal constant $c > 0$. Then (recalling Equation (1)) we have that*

$$\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{H}_t} \geq \frac{1}{2} \cdot e^{-O(\ell t)/n}. \tag{3}$$

Theorem 11 (Ewens model noise) *Let $\theta > 0$, let $\ell \in \{1,\dots,n\}$, and let $\eta := \mathrm{dist}(\theta,\ell) = \min_{j\in\{1,\dots,\ell\}} \left|e^\theta - j\right|$. Then (recalling Equation (1)) we have that*

$$\sigma_{\min,\mathsf{Up}(\mu_{\mathsf{hook},\ell}),\mathcal{E}_\theta} \geq (2n)^{-\ell}\eta^{2\sqrt{\ell}}. \tag{4}$$

5. Our positive results for noisy rankings: Putting the pieces together

In this brief section we put all the pieces together to obtain our main positive results, Theorems 2, 3 and 4, for the symmetric, heat kernel, and Ewens noise models respectively.

Symmetric noise. Under the assumptions of Theorem 2 (that $\sum_{j=0}^{n-\log k} p_j \geq \frac{1}{n^{O(\log k)}}$), taking $\ell = \log k$ in Theorem 9, we have that $\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\log k}),\mathcal{S}_{\overline{p}}} \geq \frac{1}{n^{O(\log k)}}$. Since (as discussed in Section 3.1) $\mathcal{S}_{\overline{p}}$ is efficiently samplable given $\overline{p}$, by Theorem 8 in time $\mathrm{poly}(n^{\log k}, 1/\delta, \log(1/\tau))$ with probability $1-\tau$ it is possible to obtain $\pm\delta$-accurate estimates of all of the $(\log k)$-way marginals of f. Setting $\delta = \frac{\varepsilon}{2k^{O(\log k)}}$ and applying Theorem 7, we get Theorem 2.

Heat kernel noise. First observe that we may assume that the temperature parameter t is at least 1 (since otherwise it is easy to artificially add noise to achieve $t = 1$). Under the assumptions of Theorem 3 (that $t = O(n\log n)$), taking $\ell = \log k$ in Theorem 10, we have that $\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\log k}),\mathcal{H}_t} \geq \frac{1}{n^{O(\log k)}}$. Theorem 3 follows as in the previous paragraph (this time using the efficient samplability of $\mathcal{H}_t$ given t).

Ewens noise. Under the assumptions of Theorem 4, taking $\ell = \log k$ in Theorem 11 we get that $\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\log k}),\mathcal{E}_\theta} \geq \frac{1}{n^{O(\log k)}} \cdot \mathrm{dist}(\theta,\log k)^{2\sqrt{\log k}}$. Theorem 4 follows as in the previous paragraph (this time using the efficient samplability of $\mathcal{E}_\theta$ given θ).

Acknowledgments

We thank Mike Saks for allowing us to include his proof of Claim 13 here. We also thank Vic Reiner and Yuval Roichman for answering several questions about representation theory. A.D. is grateful to Aravindan Vijayaraghavan for many useful discussions about ranking models. The work of A.D. was done while the author was at Northwestern University. Supported by NSF grant CCF-1910534, CCF-1926872 and CCF-2045128. R.O'D. was supported by NSF grants CCF-1618679 and CCF-1717606. R. A. S. was supported by the Simons Collaboration on Algorithms and Geometry, NSF CCF-1563155, CCF-1814873, and IIS-1838154. This material is based upon work supported by the National Science Foundation under grant numbers listed above. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

P. Awasthi, A. Blum, O. Sheffet, and A. Vijayaraghavan. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2014.

M. Braverman and E. Mossel. Noisy sorting without resampling. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 268–276, 2008.

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

S. Chen and A. Moitra. Algorithmic Foundations for the Diffraction Limit. *arXiv:2004.07659*, 2020.

F. Chierichetti, A. Dasgupta, R. Kumar, and S. Lattanzi. On learning mixture models for permutations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 85–92, 2015.

K. P. Choi. On the medians of gamma distributions and an equation of Ramanujan. *Proc. Amer. Math. Soc.*, 121:245–251, 1994.

C. Curtis and I. Reiner. *Representation theory of finite groups and associative algebras*, volume 356. American Mathematical Society, 1966.

S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

A. De, M. Saks, and S. Tang. Noisy population recovery in polynomial time. In *2016 Foundations of Computer Science*, pages 675–684. IEEE, 2016.

P. Diaconis. Group representations in probability and statistics. *Lecture Notes-Monograph Series*, 11:i–192, 1988a.

Persi Diaconis. *Chapter 6: Metrics on Groups, and Their Statistical Uses*, volume Volume 11 of *Lecture Notes–Monograph Series*, pages 102–130. Institute of Mathematical Statistics, 1988b.

Persi Diaconis and Phil Hanlon. Eigen Analysis for Some Examples of the Metropolis Algorithm. *Contemporary Mathematics*, 138:99–117, 1992.

Persi Diaconis and Laurent Saloff-Coste. What do we know about the metropolis algorithm? *J. Comput. Syst. Sci.*, 57(1):20–36, 1998.

Persi Diaconis and Mehrdad Shahshahani. Generating a Random Permutation with Random Transpositions. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 57:159–179, 1981.

Z. Dvir, A. Rao, A. Wigderson, and A. Yehudayoff. Restriction access. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 19–33, 2012.

W. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3: 87–112, 1972.

Vivek F Farias, Srikanth Jagabathula, and Devavrat Shah. Sparse choice models. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–28. IEEE, 2012.

M. Fligner and J. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.

O. Goldreich and L. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the Twenty-First Annual Symposium on Theory of Computing*, pages 25–32, 1989.

B. Green and A. Wigderson. Lecture notes for the 22nd McGill Invitational Workshop on Computational Complexity. 2010.

S. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

Srikanth Jagabathula and Devavrat Shah. Inferring rankings using constrained sensing. *IEEE Transactions on Information Theory*, 57(11):7288–7306, 2011.

Gordon Douglas James. *The representation theory of the symmetric groups*, volume 682. Springer, 2006.

Y. Jiao and J. Vert. The Kendall and Mallows kernels for permutations. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1755–1769, 2018.

R. Kondor and M. Barbosa. Ranking with Kernels in Fourier space. In *COLT 2010*, pages 451–463, 2010.

R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Machine Learning, Proceedings of the 19th International Conference (ICML 2002)*, 2002.

P. Kothari, J. Steinhardt, and D. Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *WWW*, pages 571–580, 2010.

Greg Kuperberg, Shachar Lovett, and Ron Peled. Probabilistic existence of regular combinatorial structures. *Geometric and Functional Analysis*, 27(4):919–972, 2017.

E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. on Computing*, 22(6):1331–1348, 1993.

A. Liu and A. Moitra. Efficiently Learning Mixtures of Mallows Models. In *Proceedings of FOCS, 2018*, 2018.

Shachar Lovett and Jiapeng Zhang. Improved noisy population recovery, and reverse Bonami-Beckner inequality for sparse functions. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 137–142, 2015.

T. Lu and C. Boutilier. Learning Mallows models with pairwise preferences. In *Proceedings of the 28th ICML*, pages 145–152, 2011.

C. Mallows. Non-null ranking models. I. *Biometrika*, 44(1/2):114–130, 1957.

Cheng Mao and Yihong Wu. Learning mixtures of permutations: Groups of pairwise comparisons and combinatorial method of moments, 2020.

P. Méliot. *Representation theory of symmetric groups*. Chapman and Hall/CRC, 2017.

Ankur Moitra and Michael Saks. A polynomial time algorithm for lossy population recovery. In *2013 Foundations of Computer Science*, pages 110–116. IEEE, 2013.

S. Mukherjee. Estimation in exponential families on permutations. *The Annals of Statistics*, 44(2): 853–875, 2016.

T. Murphy and D. Martin. Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, 41(3-4):645–655, 2003.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

O. Regev and A. Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.

M. Saks. Personal communication, 2018.

Richard P. Stanley. *Enumerative Combinatorics: Volume 2*. Cambridge University Press, 1999.

G. W. Stewart. On the Perturbation of Pseudo-Inverses, Projections and Linear Least Squares Problems. *SIAM Review*, 19(4):634–662, 1977.

Avi Wigderson and Amir Yehudayoff. Population Recovery and Partial Identification. In *53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 390–399, 2012.

Appendix A. Proof of Theorem 7

We recall the statement of Theorem 7:

Theorem 12 (Restatement of Theorem 7) *Let f be an unknown ε-heavy distribution over $\mathbb{S}_n$ with $|\mathrm{supp}(f)| \leq k$. Suppose there is an algorithm A_{marginal} with the following property: given as input a value $\delta > 0$ and two vectors $\bar{i} = (i_1, \ldots, i_t)$ and $\bar{j} = (j_1, \ldots, j_t)$ each composed of t distinct elements of $\{1, \ldots, n\}$, algorithm A_{marginal} runs in time $T(\delta, t, k, n)$ and outputs an additively $\pm\delta$-accurate estimate of the $(\bar{i}, \bar{j})$-marginal of f (recall Definition 6). Then there is an algorithm A_{learn} with the following property: given the value of ε, algorithm A_{learn} runs in time $\mathrm{poly}(n/\varepsilon, n^{\log k}) \cdot T(\frac{\varepsilon}{2k^{O(\log k)}}, 2\log k, k^2, n)$ and returns a function $g : \mathbb{S}_n \to \mathbb{R}^+$ such that $\|f - g\|_1 \leq \varepsilon$.*

A.1. A useful structural result

The following structural result on functions from $\mathbb{S}_n$ to $\mathbb{R}$ with small support will be useful for us:

Theorem 13 (Small-support functions are correlated with juntas) *Fix $1 \leq \ell \leq n$ and let $g : [n]^\ell \to \mathbb{R}$ be such that $\|g\|_1 = 1$ and $|\mathrm{supp}(g)| \leq k$. There is a subset $U \subseteq [n]$ and a list of values $\alpha_1, \ldots, \alpha_{|U|} \in [n]$ such that $|U| \leq \log k$ and*

$$\left| \sum_{x \in [n]^\ell} g(x) \cdot \mathbf{1}[x_i = \alpha_i \text{ for all } i \in U] \right| \geq k^{-O(\log k)}. \tag{5}$$

Theorem 13 is reminiscent of analogous structural results for functions over $\{0,1\}^\ell$ which are implicit in the work of Wigderson and Yehudayoff (2012) (specifically, Theorem 1.5 of that work), and indeed Theorem 13 can be proved by following the techniques of Wigderson and Yehudayoff (2012). Michael Saks Saks (2018) has communicated to us an alternative, and arguably simpler, argument for the relevant structural result over $\{0,1\}^\ell$; here we follow that alternative argument (extending it in the essentially obvious way to the domain $[n]^\ell$ rather than $\{0,1\}^\ell$).

Proof Let the support of g be $S \subseteq [n]^\ell$. Note that since $|S| \leq k$, there must exist some set of $k' := \min\{k, \ell\}$ coordinates such that any two elements of S differ in at least one of those coordinates. Without loss of generality, we assume that this set is the first k' coordinates $\{1, \ldots, k'\}$.

We prove Theorem 13 by analyzing an iterative process that iterates over the coordinates $1, \ldots, k'$. At the beginning of the process, we initialize a set $\mathrm{Coord}_{\mathsf{live}}$ of "live coordinates" to be $[k']$, initialize a set Constr of constraints to be initially empty, and initialize a set $S_{\mathsf{live}} \subseteq [n]^\ell$ of "live support elements" to be the entire support S of g. We will see that the iterative process maintains the following invariants:

(I1) The coordinates in $\mathrm{Coord}_{\mathsf{live}}$ are sufficient to distinguish between the elements in S_{live}, i.e. any two distinct strings in S_{live} have distinct projections onto the coordinates in $\mathrm{Coord}_{\mathsf{live}}$;

(I2) The only elements of S that satisfy all the constraints in Constr are the elements of S_{live}.

Before presenting the iterative process we need to define some pertinent quantities. For each coordinate $j \in \mathrm{Coord}_{\mathsf{live}}$ and each index $\alpha \in [n]$, we define

$$\mathsf{Wt}(j, \alpha) := \sum_{x \in S_{\mathsf{live}} : x_j = \alpha} |g(x)|,$$

the *weight* under g of the live support elements x that have $x_j = \alpha$, and we define

$$\mathsf{Num}(j, \alpha) := |\{x \in S_{\mathsf{live}} : x_j = \alpha\}|,$$

the number of live support elements x that have $x_j = \alpha$ (note that $\mathsf{Num}(j, \alpha)$ has nothing to do with the specific values of the weights assigned by g). It will also be useful to have notation for fractional versions of each of these quantities, so we define

$$\mathsf{FracWt}(j, \alpha) := \frac{\mathsf{Wt}(j, \alpha)}{\sum_{x \in S_{\mathsf{live}}} |g(x)|}. \qquad \text{and} \qquad \mathsf{Frac}(j, \alpha) := \frac{\mathsf{Num}(j, \alpha)}{|S_{\mathsf{live}}|}$$

Note that for any $j \in \mathrm{Coord}_{\mathsf{live}}$ we have that $\sum_\alpha \mathsf{Num}(j, \alpha) = |S_{\mathsf{live}}|$, or equivalently $\sum_\alpha \mathsf{Frac}(j, \alpha) = 1$.

For each coordinate $j \in \mathrm{Coord}_{\mathsf{live}}$, we write $\mathsf{MAJ}(j)$ to denote the element $\beta \in [n]$ which is such that $\mathsf{Num}(j, \beta) \geq \mathsf{Num}(j, \alpha)$ for all $\alpha \in [n]$ (we break ties arbitrarily). Finally, we let $\mathsf{FracWtMaj}(j) = \mathsf{FracWt}(j, \mathsf{MAJ}(j))$.

Now we are ready to present the iterative process:

1. If every $j \in \mathrm{Coord}_{\mathsf{live}}$ has $\mathsf{FracWtMaj}(j) > 1 - \frac{1}{10k'}$[4], then halt the process. Otherwise, let j be any element of $\mathrm{Coord}_{\mathsf{live}}$ for which $\mathsf{FracWtMaj}(j) \leq 1 - \frac{1}{10k'}$.
2. For this coordinate j, choose $\alpha \in [n]$ which maximizes the ratio $\frac{\mathsf{FracWt}(j,\alpha)}{\mathsf{Frac}(j,\alpha)}$ (or equivalently, maximizes $\frac{\mathsf{FracWt}(j,\alpha)}{\mathsf{Num}(j,\alpha)}$) subject to $\mathsf{Frac}(j, \alpha) \neq 0$ and $\alpha \neq \mathsf{MAJ}(j)$.
3. Add the constraint $x_j = \alpha$ to Constr, remove j from $\mathrm{Coord}_{\mathsf{live}}$, and remove all x such that $x_j \neq \alpha$ from S_{live}. Go to Step 1.

When the iterative process ends, suppose that the set Constr is $\{x_{j_1} = \alpha_1, \ldots, x_{j_\ell} = \alpha_\ell\}$. Then we claim that Equation (5) holds for $U = \{j_1, \ldots, j_\ell\}$.

To argue this, we first observe that both invariants (I1) and (I2) are clearly maintained by each round of the iterative process. We next observe that each time a pair (j, α) is processed in Step 3, it holds that $\mathsf{Frac}(j, \alpha) \leq \frac{1}{2}$, and hence each round shrinks S_{live} by a factor of at least 2. Thus, after $\log k$ steps, the set S_{live} must be of size at most 1 and hence the process must halt. (Note that the claimed bound $|U| \leq \log k$ follows from the fact that the process runs for at most $\log k$ stages.)

Next, note that when the process halts, by a union bound over the at most k' coordinates in $\mathrm{Coord}_{\mathsf{live}}$ it holds that

$$\sum_{x \in S_{\mathsf{live}} : x_j = \mathsf{MAJ}(j) \text{ for all } j \in \mathrm{Coord}_{\mathsf{live}}} |g(x)| \geq \frac{9}{10} \cdot \sum_{x \in S_{\mathsf{live}}} |g(x)|.$$

On the other hand, by the first invariant (I1), the cardinality of the set $\{x \in S_{\mathsf{live}} : x_j = \mathsf{MAJ}(j) \text{ for all } j \in \mathrm{Coord}_{\mathsf{live}}\}$ is precisely 1. This immediately implies that almost all of the weight of g, across elements of S_{live}, is on a single element; more precisely, that

$$\left| \sum_{x \in S_{\mathsf{live}}} g(x) \right| \geq \frac{4}{5} \cdot \sum_{x \in S_{\mathsf{live}}} |g(x)|,$$

4. Note that this means almost all of the weight under g of the live support elements is on elements that all agree with the majority value on coordinate j. Note further that if $\mathrm{Coord}_{\mathsf{live}}$ is empty then this condition trivially holds.

from which it follows that

$$\left|\sum_{x\in[n]^\ell} g(x)\cdot \mathbf{1}[x_i=\alpha_i \text{ for all } i\in U]\right| \geq \frac{4}{5}\cdot \sum_{x\in S_{\mathsf{live}}} |g(x)|. \tag{6}$$

So to establish Equation (5), it remains only to establish a lower bound on $\sum_{x\in S_{\mathsf{live}}} |g(x)|$ when the process terminates. To do this, let us suppose that the process runs for T steps where in the t^{th} step the coordinate chosen is j_t. Now, at any stage t, we have

$$\frac{\sum_{\beta\in[n]:\beta\neq \mathsf{MAJ}(j_t)} \mathsf{FracWt}(j_t,\beta)}{\sum_{\beta\in[n]:\beta\neq \mathsf{MAJ}(j_t)} \mathsf{Frac}(j_t,\beta)} \geq \frac{1}{10k'}.$$

(because the denominator is at most 1 and since the process does not terminate, the numerator is at least $\frac{1}{10k}$). As a result, we get that if the constraint chosen at time t is $x_{j_t}=\alpha_t$, then

$$\frac{\mathsf{FracWt}(j_t,\alpha_t)}{\mathsf{Frac}(j_t,\alpha_t)} \geq \frac{1}{10k'}. \tag{7}$$

By Equation (7), when the process halts we have

$$\sum_{x\in S_{\mathsf{live}}} |g(x)| = \prod_{t=1}^{T} \mathsf{FracWt}(j_t,\alpha_t) \geq \frac{1}{(10k')^T} \prod_{t=1}^{T} \mathsf{Frac}(j_t,\alpha_t).$$

But since at least one element remains, we have that $\prod_{t=1}^{T} \mathsf{Frac}(j_t,\alpha_t) \geq \frac{1}{k}$, and since $T\leq \log k$, we conclude (recalling that $k'\leq k$) that

$$\sum_{x\in S_{\mathsf{live}}} |g(x)| \geq k^{-O(\log k)}.$$

Combining with (6), this yields the claim. ■

A.2. Proof of Theorem 12

The idea of the proof is in the spirit of the algorithmic component of several recent works on population recovery Moitra and Saks (2013); Wigderson and Yehudayoff (2012); Lovett and Zhang (2015); De et al. (2016). Given any function $f:\mathbb{S}_n\to\mathbb{R}$ and any integer $i\in\{1,\ldots,n\}$, we define the function $f_i:[n]^i\to\mathbb{R}$ as follows:

$$f_i(x_1,\ldots,x_i) := \sum_{\sigma\in\mathbb{S}_n} f(\sigma)\cdot \mathbf{1}[\sigma(1)=x_1\wedge\ldots\wedge\sigma(i)=x_i]. \tag{8}$$

At a high level, the algorithm A_{learn} of Theorem 12 works in stages, by successively reconstructing $f_0,\ldots,f_n$. In each stage it uses the procedure described in the following claim, which says that high-accuracy approximations of the $(\log k)$-marginals *together with the support of* f_ℓ (or a not-too-large superset of it) suffices to reconstruct f_ℓ:

Claim 14 *Let f_ℓ be an unknown distribution over $[n]^\ell$ supported on a given set S of size k. There is an algorithm $A_{\mathrm{one-stage}}$ which has the following guarantee: The algorithm is given as input the set S, $\delta > 0$, and parameters $\beta_{J,y}$ (for every set $J \subseteq [\ell]$ of size at most $\log k$ and every $y \in [n]^J$) which satisfy*

$$\left| \beta_{J,y} - \sum_{x \in S} f(x) \cdot \mathbf{1}[x_i = y_i \text{ for all } i \in J] \right| \leq \delta.$$

$A_{\mathrm{one-stage}}$ runs in time $\mathsf{poly}(n, \ell^{\log k})$ and outputs a function $\tilde{f} : [n]^\ell \to [0,1]$ such that $\|f - \tilde{f}\|_1 \leq \delta \cdot k^{O(\log k)}$.

Proof We consider a linear program which has a variable s_x for each $x \in S$ (representing the probability that f puts on x) and is defined by the following constraints:

1. $s_x \geq 0$ and $\sum_{x \in S} s_x = 1$.
2. For each $J \subseteq [\ell]$ of size at most $\log k$ and each $y \in [n]^J$, include the constraint

$$\left| \beta_{J,y} - \sum_{x \in S} s_x \cdot \mathbf{1}[x_i = y_i \text{ for all } i \in J] \right| \leq \delta. \tag{9}$$

Algorithm $A_{\mathsf{one-stage}}$ sets up and solves the above linear program (this can clearly be done in time $\mathsf{poly}(n, \ell^{\log k})$). We observe that the linear program is feasible since by definition $s_x = f_\ell(x)$ is a feasible solution. To prove the claim it suffices to show that every feasible solution is ℓ_1-close to f_ℓ; so let $f^*(x)$ denote any other feasible solution to the linear program, and let η denote $\|f^* - f_\ell\|_1$. Define $h(x) = f^*(x) - f_\ell(x)$, so $\|h\|_1 = \eta$. By Theorem 13, we have that there is a subset $J \subseteq [\ell]$ of size at most $\log k$ and a $y \in [n]^\ell$ such that

$$\left| \sum_x h(x) \cdot \mathbf{1}[x_i = y_i \text{ for all } i \in J] \right| \geq \eta \cdot k^{-O(\log k)}. \tag{10}$$

On the other hand, since both $f_\ell(x)$ and $f^*(x)$ are feasible solutions to the linear program, by the triangle inequality it must be the case that

$$\left| \sum_x h(x) \cdot \mathbf{1}[x_i = y_i \text{ for all } i \in J] \right| \leq 2\delta. \tag{11}$$

Equations 10 and A.2 together give the desired upper bound on η, and the claim is proved. ■

Essentially the only remaining ingredient required to prove Theorem 12 is a procedure to find (a not-too-large superset of) the support of f. This is given by the following claim, which inductively uses the algorithm $A_{\mathrm{one-stage}}$ to successively construct suitable (approximations of) the support sets for $f_1, \ldots, f_n$.

Claim 15 *Under the assumptions of Theorem 12, there is an algorithm A_{support} with the following property: given as input a value $\delta > 0$, algorithm A_{support} runs in time $\mathrm{poly}(n/\varepsilon, n^{\log k}) \cdot T(\frac{\varepsilon}{2k^{O(\log k)}}, 2\log k, k^2, n)$ and for each $\ell = 1, \ldots, n$ outputs a set $S'_{(\ell)}$ of size at most k which contains the support of f_ℓ.*

Proof The algorithm A_{support} works inductively, where at the start of stage ℓ (in which it will construct the set $S'_{(\ell)}$) it is assumed to have a set $S'_{(\ell-1)}$ with $|S'_{(\ell-1)}| \leq k$ which contains the support of $f_{\ell-1}$. (Note that at the start of the first stage $\ell = 1$ this holds trivially since f_0 trivially has empty support).

Let us describe the execution of the ℓ-th stage of A_{support}. For $1 \leq \ell \leq n$, we define the set $S_{\mathsf{marg},\ell}$ as follows:

$$S_{\mathsf{marg},\ell} = \big\{t : \sum_{\sigma \in \mathbb{S}_n} f(\sigma) \cdot \mathbf{1}[\sigma(\ell) = t] > 0\big\}.$$

Observe that in time $\mathsf{poly}(n/\varepsilon) \cdot T(\frac{\varepsilon}{4}, 1, k, n)$, we can compute $f(\sigma) \cdot \mathbf{1}[\sigma(\ell) = t]$ up to error $\pm\varepsilon/4$ (denote this estimate by $\beta_{\ell,t}$) for all $1 \leq t \leq n$. Since f is ε-heavy, we have that

$$t \in S_{\mathsf{marg},\ell} \text{ implies } \beta_{\ell,t} \geq \frac{3\varepsilon}{4} \quad \text{and} \quad t \notin S_{\mathsf{marg},\ell} \text{ implies } \beta_{\ell,t} \leq \frac{\varepsilon}{4}.$$

Consequently, we can compute the set $S_{\mathsf{marg},\ell}$ in time $\mathsf{poly}(n/\varepsilon) \cdot T(\frac{\varepsilon}{4}, 1, k, n)$. The final observation is that the set $S^*_{(\ell)}$ (of cardinality at most k^2) obtained by appending each final ℓ-th character from $S_{\mathsf{marg},\ell}$ to each element of $S'_{(\ell-1)}$ must contain the support $S_{(\ell)}$ of f_ℓ. Set $\delta = \frac{\varepsilon}{2k^{O(\log k)}}$; by the assumption of Theorem 12, in time $T(\frac{\varepsilon}{2k^{O(\log k)}}, 2\log k, k^2, n)$ it is possible to obtain additively $\pm\delta$-accurate estimates of each of the $(2\log k)$-way marginals of f_ℓ. In the ℓ-th stage, algorithm A_{support} runs $A_{\mathrm{one-stage}}$ using $S^*_{(\ell)}$ and these estimates of the marginals; by Theorem 14, this takes time $\mathrm{poly}(n/\varepsilon, n^{\log k})$ and yields a function $\tilde{f}_\ell : [n]^\ell \to [0,1]$ such that $\|f_\ell - \tilde{f}_\ell\|_1 \leq \frac{\delta}{2k^{O(\log k)}} \cdot k^{O(\log k)} = \varepsilon/4$. Since by assumption f is ε-heavy, it follows that any element x in the support of $\tilde{f}_\ell$ such that $\tilde{f}_\ell(x) \leq \varepsilon/4$ must not be in the support of f_ℓ; so the algorithm removes all such elements x from $S^*_{(\ell)}$ to obtain the set $S'_{(\ell)}$. This resulting $S'_{(\ell)}$ is precisely the support of f_ℓ, and is clearly of size at most k. ■

Finally, the overall algorithm A_{learn} works by running A_{support} to get the set $S' = S'_{(n)}$ of size at most k which is the support of $f_n = f$, and then uses S' and the algorithm A_{marginal} from the assumptions of Theorem 12) to run algorithm $A_{\mathrm{one-stage}}$ and obtain the required ε-accurate approximator g of f. This concludes the proof of Theorem 12.

Appendix B. Proof of Theorem 8

We recall the statement of Theorem 8:

Theorem 16 (Restatement of Theorem 8) *Let $\mathcal{K}$ be an efficiently samplable distribution over $\mathbb{S}_n$. Let f be an unknown distribution over $\mathbb{S}_n$. There is an algorithm A_{marginal} with the following properties: A_{marginal} receives as input a parameter $\delta > 0$, a confidence parameter $\tau > 0$, a pair of ℓ-tuples $\bar{i} = (i_1, \ldots, i_\ell) \in [n]^\ell$, $\bar{j} = (j_1, \ldots, j_\ell) \in [n]^\ell$ each composed of ℓ distinct elements, and has access to random samples from $\mathcal{K} * f$. Algorithm A_{marginal} runs in time $\mathsf{poly}(\binom{n}{\ell}, \delta^{-1}, \sigma^{-1}_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{K}}, \log(1/\tau))$ and outputs a value $\kappa_{\bar{i},\bar{j}}$ which with probability at least $1 - \tau$ is a $\pm\delta$-accurate estimate of the $(\bar{i},\bar{j})$-marginal of f.*

We will use the following claim to prove Theorem 16:

Claim 17 *Let $\rho : \mathbb{S}_n \to \mathbb{C}^{m\times m}$ be any unitary representation of $\mathbb{S}_n$, let $\mathcal{K}$ be any efficiently samplable distribution over $\mathbb{S}_n$, and let $\sigma_{\min}$ denote the smallest singular value of $\widehat{\mathcal{K}}(\rho)$. Let f be an unknown distribution over $\mathbb{S}_n$. There is an algorithm which, given random samples from $\mathcal{K} * f$ and an error parameter $0 < \delta < 1$, runs in time $\mathsf{poly}(m, n, \sigma_{\min}^{-1}, \delta^{-1})$ and with high probability outputs a matrix $M_{f,\rho}$ such that $\|M_{f,\rho} - \widehat{f}(\rho)\| \leq \delta$.*

Proof Let $\eta_1, \eta_2 > 0$ denote two error parameters that will be fixed later. Since f is a distribution, the Fourier coefficient $\widehat{f}(\rho)$ is equal to $\mathbf{E}_{\boldsymbol{\sigma}\sim f}[\rho(\boldsymbol{\sigma})]$. Consequently, since $\mathcal{K}$ is assumed to be efficiently samplable and the algorithm is given samples from $\mathcal{K} * f$, by sampling from $\mathcal{K}$ and from $\mathcal{K} * f$ it is straightforward to obtain matrices M_1, M_2 in time $\mathrm{poly}(m, n, \log(1/\tau))$ which with probability $1 - \tau$ satisfy

$$\|M_1 - \widehat{\mathcal{K}}(\rho)\|_2 \leq \eta_1 \text{ and } \|M_2 - \widehat{\mathcal{K} * f}(\rho)\|_2 \leq \eta_2.$$

Now we recall the following matrix perturbation inequality (see Theorem 2.2 of Stewart (1977)):

Lemma 18 *Let $A \in \mathbb{R}^{n\times n}$ be a non-singular matrix and further let $\Delta A \in \mathbb{R}^{n\times n}$ be such that $\|\Delta A\|_2 \cdot \|A^{-1}\|_2 < 1$. Then $A + \Delta A$ is non-singular. Further, if $\gamma = 1 - \|A^{-1}\|_2\|\Delta A\|_2$, then*

$$\|A^{-1} - (A + \Delta A)^{-1}\|_2 \leq \frac{\|A^{-1}\|_2^2 \|\Delta A\|_2}{\gamma}.$$

Let us now set the error parameters η_1 and η_2 as follows (recall that $\delta < 1$):

$$\eta_1 = \min\Big\{\frac{\delta \cdot \sigma_{\min}^2}{4}, \frac{\delta \cdot \sigma_{\min}}{4}\Big\} \text{ and } \eta_2 = \min\{\frac{\delta \cdot \sigma_{\min}}{4}, 1\}. \tag{12}$$

Applying Lemma 18 with $\widehat{\mathcal{K}}(\rho)$ in place of A and $M_1 - \widehat{\mathcal{K}}(\rho)$ in place of ΔA, using (12) (more precisely, the upper bound $\eta_1 \leq \delta \cdot \sigma_{\min}^2/4$ in the numerator and the upper bound $\eta_1 \leq \delta \cdot \sigma_{\min}/4$ in the denominator) we get that

$$\|M_1^{-1} - \widehat{\mathcal{K}}(\rho)^{-1}\|_2 \leq \frac{\|\widehat{\mathcal{K}}(\rho)^{-1}\|_2^2 \cdot \|M_1 - \widehat{\mathcal{K}}(\rho)\|_2}{1 - \|\widehat{\mathcal{K}}(\rho)^{-1}\|_2 \cdot \|M_1 - \widehat{\mathcal{K}}(\rho)\|_2} \leq \frac{\delta}{3}. \tag{13}$$

Now using $\widehat{\mathcal{K} * f}(\rho) = \widehat{\mathcal{K}}(\rho) \cdot \widehat{f}(\rho)$, we get

$$\begin{aligned}
\|M_1^{-1} \cdot M_2 - \widehat{f}(\rho)\|_2 &= \|M_1^{-1} \cdot M_2 - \widehat{\mathcal{K}}(\rho)^{-1} \cdot \widehat{\mathcal{K} * f}(\rho) \cdot\|_2 \\
&\leq \|M_1^{-1} \cdot M_2 - M_1^{-1} \cdot \widehat{\mathcal{K} * f}(\rho)\|_2 + \|M_1^{-1} \cdot \widehat{\mathcal{K} * f}(\rho) - \widehat{\mathcal{K}}(\rho)^{-1} \cdot \widehat{\mathcal{K} * f}(\rho)\|_2 \\
&\leq \|M_1^{-1}\|_2 \cdot \|M_2 - \widehat{\mathcal{K} * f}(\rho)\|_2 + \|M_1^{-1} - \widehat{\mathcal{K}}(\rho)^{-1}\|_2 \cdot \|\widehat{\mathcal{K} * f}(\rho)\|_2 \\
&\leq \|M_1^{-1}\|_2 \cdot \eta_2 + \|\widehat{\mathcal{K} * f}(\rho)\|_2 \cdot \frac{\delta}{3}. \qquad \text{(using (13))} \\
&\leq \eta_2 \left(\|\widehat{\mathcal{K}}(\rho)^{-1}\|_2 + \|M^{-1} - \widehat{\mathcal{K}}(\rho)^{-1}\|_2\right) + \|\widehat{\mathcal{K} * f}(\rho)\|_2 \cdot \frac{\delta}{3}. \qquad \text{(using (13))} \\
&\leq \sigma_{\min}^{-1} \cdot \eta_2 + \frac{\delta}{3} \cdot \eta_2 + \|\widehat{\mathcal{K} * f}(\rho)\|_2 \cdot \frac{\delta}{3}.
\end{aligned} \tag{14}$$

Next we use the following fact, which is an easy consequence of the triangle inequality and the assumption that ρ is unitary:

Fact 19 *Let $\rho : \mathbb{S}_n \to \mathbb{C}^{m\times m}$ be a unitary representation and let $g : \mathbb{S}_n \to \mathbb{R}^+$. Then we have that $\|\widehat{g}(\rho)\|_2 \le \|g\|_1$.*

Combining this fact with (14) and (12), since $\|\mathcal{K} * f\|_1 = 1$, we get that

$$\|M_1^{-1} \cdot M_2 - \widehat{f}(\rho)\|_2 \le \sigma_{\min}^{-1} \cdot \eta_2 + \frac{\delta}{3} \cdot \eta_2 + \frac{\delta}{3} \le \frac{\delta}{4} + \frac{\delta}{3} + \frac{\delta}{3} < \delta.$$

This concludes the proof of Claim 17. ■

With Theorem 17 in hand we are ready to prove Theorem 16:

Proof of Theorem 16. Let $\tau_{\lambda_{\mathsf{hook},\ell}}$ be the permutation representation corresponding to the partition $\lambda_{\mathsf{hook},\ell}$; for conciseness we subsequently write ρ for $\tau_{\lambda_{\mathsf{hook},\ell}}$. Definition 40 immediately gives that the dimension of ρ is $\binom{n}{\ell}$. Observe that ρ is a unitary representation. Let $\sigma_{\min}$ denote the smallest singular value of $\widehat{\mathcal{K}}(\rho)$; applying Theorem 17, we get an algorithm running in time $\mathrm{poly}(\binom{n}{\ell}, \sigma_{\min}^{-1}, \delta)$ which outputs a matrix $M_{f,\rho}$ such that $\|M_{f,\rho} - \widehat{f}(\rho)\| \le \delta$. Next, we observe that the Young tableaux corresponding to the partition $\lambda_{\mathsf{hook},\ell}$ (which, recalling Definition 40, index the rows and columns of $\rho(\cdot)$) correspond precisely to ordered t-tuples of distinct entries of $[n]$. If $\mathsf{Y}_{\lambda_{\mathsf{hook},\ell},i} = \bar{i}$ and $\mathsf{Y}_{\lambda_{\mathsf{hook},\ell},j} = \bar{j}$, then it follows that

$$\widehat{f}(\rho)(i,j) = \sum_{\sigma\in\mathbb{S}_n} f(\sigma) \cdot \mathbf{1}[f(i_1) = j_1 \text{ and } \cdots \text{ and } f(i_\ell) = j_\ell)],$$

which is the $(\bar{i},\bar{j})$-marginal of f as desired; so the output of the algorithm is $M_{f,\rho}(\bar{i},\bar{j})$.

To finish the correctness argument it remains only to argue that $\sigma_{\min}^{-1}$ is at most $\mathrm{poly}(\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell})}^{-1})$. To see that this is indeed, the case, we observe that by Theorem 42, the permutation representation $\tau_{\lambda_{\mathsf{hook},\ell}}$ block diagonalizes into a direct sum of irreducible representations ρ_μ where each μ belongs to $\mathsf{Up}(\lambda_{\mathsf{hook},\ell})$. This finishes the proof of Theorem 16.

Appendix C. Representations of symmetric noise

In this section we establish lower bounds on the smallest singular value for the relevant matrices corresponding to "symmetric noise" $\mathcal{S}_{\overline{p}}$ on $\mathbb{S}_n$. In more detail, the main result of this section is the following lower bound:

Theorem 20 (Symmetric noise - Restatement of Theorem 9) *Let $\ell \in \{1,\dots,n\}$ and let $\overline{p} = (p_0,\dots,p_n) \in \Delta^n$ (i.e. $\overline{p}$ is a non-negative vector whose entries sum to 1) which is such that*

$$\sum_{j=0}^{n-\ell} p_j \ge \kappa.$$

Then (recalling Equation (1)) we have that

$$\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{S}_{\overline{p}}} \ge \frac{\kappa}{n^\ell}. \tag{15}$$

C.1. Setup

To analyze the smallest singular value of $\widehat{\mathcal{S}_{\overline{p}}}(\rho_\mu)$ (as required by the definition of $\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{S}_{\overline{p}}}$), we start by observing that symmetric noise is a *class function* (meaning that it is invariant under conjugation, see Definition 36):

Claim 21 *For any vector $\overline{p} = (p_0, \ldots, p_n) \in \Delta^n$, the distribution $\mathcal{S}_{\overline{p}}$ (viewed as a function from $\mathbb{S}_n$ to $[0,1]$) is a class function (i.e. $\mathcal{S}_{\overline{p}}(\pi) = \mathcal{S}_{\overline{p}}(\tau\pi\tau^{-1})$ for every $\pi, \tau \in \mathbb{S}_n$).*

Proof For $0 \le j \le n$, let $\overline{e}_j$ denote the vector in $\mathbb{R}^{n+1}$ which has a 1 in the j-th position and a 0 in every other position. By linearity, to prove Claim 21 it suffices to prove that $\mathcal{S}_{\overline{e}_j}$ is invariant under conjugation for every j; to establish this, it suffices to show that $\mathcal{S}_{\overline{e}_j}$ is invariant under conjugation by any transposition τ. By symmetry, it suffices to consider the transposition $\tau = (1,2)$.

We observe that $\mathcal{S}_{\overline{e}_j}$ is a uniform average of $\mathbb{U}_A$ over all $\binom{n}{j}$ subsets A of $[n]$ of size exactly j. Now we consider two cases: the first is that $|A \cap \{1,2\}|$ is 0 or 2. In this case it is easy to see that $\mathbb{U}_A$ does not change under conjugation by the transposition $(1,2)$. The remaining case is that $|A \cap \{1,2\}| = 1$; in this case it is easy to see that conjugation by $(1,2)$ converts $\mathbb{U}_A$ into $\mathbb{U}_{A\Delta\{1,2\}}$. Since the collection of size-j sets A with $A \cap \{1,2\} = \{1\}$ are in 1-1 correspondence with the collection of size-j sets A with $A \cap \{1,2\} = \{2\}$, it follows that $\mathcal{S}_{\overline{e}_j}$ is invariant under conjugation by $\tau = (1,2)$, and the proof is complete. ∎

Before stating the next lemma we remind the reader that for partitions $\mu \vdash m, \lambda \vdash n$ where $m \le n$, we write $\mathrm{Paths}(\mu, \lambda)$ to denote the number of paths from μ to λ in Young's lattice (see Appendix G.2 and Theorem 45). We write Triv_j to denote the trivial partition (j) of j.

Lemma 22 *Let $\lambda \vdash n$ and let ρ_λ be the corresponding irreducible representation of $\mathbb{S}_n$. Given $\overline{p} = (p_0, \ldots, p_n) \in \Delta^n$, we have that*

$$\widehat{\mathcal{S}_{\overline{p}}}(\rho_\lambda) = \mathsf{c}(\overline{p}, \lambda) \cdot \mathsf{Id} \text{ where } \mathsf{c}(\overline{p}, \lambda) := \frac{\sum_{j=0}^n p_j \cdot \mathrm{Paths}(\mathsf{Triv}_j, \lambda)}{\mathsf{dim}(\rho_\lambda)}. \tag{16}$$

Proof By Claim 21, we have that $\mathcal{S}_{\overline{p}}$ is a class function, so we may apply Lemma 39 to conclude that

$$\widehat{\mathcal{S}_{\overline{p}}}(\rho_\lambda) = \mathsf{c}(\overline{p}, \lambda) \cdot \mathsf{Id},$$

where

$$\mathsf{c}(\overline{p}, \lambda) = \frac{1}{\mathsf{dim}(\rho_\lambda)} \cdot \left(\sum_{\sigma \in \mathbb{S}_n} \mathcal{S}_{\overline{p}}(\sigma) \cdot \chi_\lambda(\sigma) \right)$$

and χ_λ denotes the character of the irreducible representation ρ_λ. Thus it remains to show that $\sum_{\sigma \in \mathbb{S}_n} \mathcal{S}_{\overline{p}}(\sigma) \cdot \chi_\lambda(\sigma)$ is equal to the numerator of Equation (16). By definition of $\mathcal{S}_{\overline{p}}$, we have that

$$\sum_{\sigma \in \mathbb{S}_n} \mathcal{S}_{\overline{p}}(\sigma) \cdot \chi_\lambda(\sigma) \quad = \quad \sum_{0 \le j \le n} \overline{p}_j \mathop{\mathbf{E}}_{\mathcal{A}:|\mathcal{A}|=j} \mathop{\mathbf{E}}_{\boldsymbol{\sigma} \in \mathbb{U}_\mathcal{A}} \chi_\lambda(\boldsymbol{\sigma}). \tag{17}$$

We proceed to analyze $\mathbf{E}_{\boldsymbol{\sigma} \in \mathbb{U}_\mathcal{A}} \chi_\lambda(\boldsymbol{\sigma})$. Let $\rho_\lambda^\mathcal{A}$ denote the representation ρ_λ restricted to the subgroup $\mathbb{S}_\mathcal{A}$. By Theorem 45, the representation $\rho_\lambda^\mathcal{A}$ splits as follows:

$$\rho_\lambda^\mathcal{A} = \bigoplus_{\mu \vdash |\mathcal{A}|} \mathrm{Paths}(\mu, \lambda) \rho_\mu.$$

Thus, we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{\sigma}\in\mathbb{U}_{\mathcal{A}}}\chi_\lambda(\boldsymbol{\sigma}) = \sum_{\mu\vdash|\mathcal{A}|}\mathrm{Paths}(\mu,\lambda)\mathop{\mathbf{E}}_{\boldsymbol{\sigma}\in\mathbb{U}_{\mathcal{A}}}\chi_\mu(\boldsymbol{\sigma}) = \mathrm{Paths}(\mathsf{Triv}_{|\mathcal{A}|},\lambda).$$

The second equality follows from that fact that if μ is a non-trivial partition of $|\mathcal{A}|$ then $\mathbf{E}_{\boldsymbol{\sigma}\in\mathbb{U}_{\mathcal{A}}}\chi_\mu(\boldsymbol{\sigma}) = 0$, while if $\mu = \mathsf{Triv}_{|\mathcal{A}|}$ then $\mathbf{E}_{\boldsymbol{\sigma}\in\mathbb{U}_{\mathcal{A}}}\chi_\mu(\boldsymbol{\sigma}) = 1$. Plugging this into (17) we get that $\sum_{\sigma\in\mathbb{S}_n}\mathcal{S}_{\overline{p}}(\sigma)\cdot\chi_\lambda(\sigma) = \sum_{j=0}^n p_j\cdot\mathrm{Paths}(\mathsf{Triv}_j,\lambda)$, and the lemma is proved. ■

C.2. Proof of Theorem 20

We recall from Equation (1) that

$$\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{S}_{\overline{p}}} := \min_{\mu\in\mathsf{Up}(\lambda_{\mathsf{hook},\ell})}\sigma_{\min}(\widehat{\mathcal{S}_{\overline{p}}}(\rho_\mu)).$$

Fix any $\mu\in\mathsf{Up}(\lambda_{\mathsf{hook},\ell})$, so μ is a partition of n of the form $(n-\ell',\ell_2,\dots,\ell_r)$ where $\ell'\le\ell$. By Lemma 22 we have that the smallest singular value of $\widehat{\mathcal{S}_{\overline{p}}}(\rho_\mu)$ is

$$\mathsf{c}(\overline{p},\mu) := \frac{\sum_{j=0}^n p_j\cdot\mathrm{Paths}(\mathsf{Triv}_j,\mu)}{\mathsf{dim}(\rho_\mu)}. \tag{18}$$

To upper bound $\mathsf{dim}(\rho_\mu)$, we observe that

$$\dim(\rho_\mu)\le\dim(\tau_\mu) = \binom{n}{n-\ell',\ell_2,\dots,\ell_r}\le\frac{n!}{(n-\ell')!}\le n^{\ell'}\le n^{\ell},$$

where the first inequality is by Theorem 42. For the numerator, we observe that if $j\le n-\ell$ then there is at least one path in the Young lattice from Triv_j to μ, so under the assumptions of Theorem 20 the numerator of Equation (18) is at least κ. This proves the theorem.

Appendix D. Representations of heat kernel noise

In this section, analogous to Appendix C, we lower bound Equation (1) when the noise distribution $\mathcal{K}$ is $\mathcal{H}_t$, corresponding to "heat kernel noise" at temperature parameter t:

Theorem 23 (Heat kernel noise - Restatement of Theorem 10) *Let $t\ge1$ and let $\ell\in\{1,\dots,cn\}$ for some suitably small universal constant $c>0$. Then we have that*

$$\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{H}_t}\ge\frac{1}{2}\cdot e^{-O(\ell t)/n}. \tag{19}$$

D.1. Setup

Let $\mathrm{trans}:\mathbb{S}_n\to[0,1]$ be the following probability distribution over $\mathbb{S}_n$:

$$\mathrm{trans}(\pi) = \begin{cases}1/n & \text{if } \pi \text{ is the identity,}\\ 2/n^2 & \text{if } \pi \text{ is a transposition,}\\ 0 & \text{otherwise.}\end{cases}$$

Since $\mathrm{trans}(\pi)$ depends only on the cycle structure of π, the function $\mathrm{trans}(\cdot)$ is a class function. Fix any $\mu \in \mathsf{Up}(\lambda_{\mathsf{hook},\ell})$, so μ is a partition of n of the form $(\mu_1, \ldots, \mu_r)$ where $\mu_1 \geq n - \ell$. As in the proof of Lemma 22 we may apply Lemma 39 to conclude that

$$\widehat{\mathrm{trans}}(\rho_\mu) = \mathsf{c}_{\mathrm{trans},\mu} \cdot \mathsf{Id}$$

for some constant $\mathsf{c}_{\mathrm{trans},\mu}$. By Corollary 1 of Diaconis and Shahshahani Diaconis and Shahshahani (1981), we have that

$$\mathsf{c}_{\mathrm{trans},\mu} = \frac{1}{n} + \frac{n-1}{n} \cdot \frac{\chi_\mu(\tau)}{\dim(\rho_\mu)}, \tag{20}$$

where as before χ_μ denotes the character of the irreducible representation ρ_μ and τ is any transposition. Diaconis and Shahshahani (1981) further shows that for ρ_μ an irreducible representation of $\mathbb{S}_n$ with μ as above and τ any transposition, it holds that

$$\frac{\chi_\mu(\tau)}{\dim(\rho_\mu)} = \frac{1}{n(n-1)} \cdot \sum_{j=1}^{r} (\mu_j - j)(\mu_j - j + 1) - j(j-1). \tag{21}$$

In our setting we have

$$(21) \geq \frac{(n-\ell)(n-\ell-1)}{n(n-1)} + \frac{1}{n(n-1)} \sum_{j=2}^{r} (\mu_j - j)(\mu_j - j + 1) - j(j-1). \tag{22}$$

where the inequality holds because $\mu_1 \geq n - \ell$. Now, we observe that for each summand in Equation (22), we have

$$\begin{aligned}(\mu_j - j)(\mu_j - j + 1) - j(j-1) &= \mu_j^2 - \mu_j(2j-1) \\ &\geq -\mu_j(2j-1) \\ &\geq \frac{-\ell}{j-1} \cdot (2j-1) \geq -3\ell.\end{aligned}$$

The second inequality above holds because $\mu_2 + \cdots + \mu_j \leq \ell$ and the μ_j's are non-increasing, so $\mu_j \leq \frac{\ell}{j-1}$. Since $r - 1 \leq \ell$, this means that

$$(21) \geq \frac{(n-\ell)(n-\ell-1)}{n(n-1)} - \frac{3\ell^2}{n(n-1)} \geq 1 - \frac{O(\ell)}{n},$$

and recalling Equation (20) we get that

$$1 \geq \mathsf{c}_{\mathrm{trans},\mu} \geq 1 - \frac{O(\ell)}{n}. \tag{23}$$

D.2. Proof of Theorem 23

As in Appendix C we recall from Equation (1) that

$$\sigma_{\min,\mathsf{Up}(\lambda_{\mathsf{hook},\ell}),\mathcal{H}_t} := \min_{\mu \in \mathsf{Up}(\lambda_{\mathsf{hook},\ell})} \sigma_{\min}(\widehat{\mathcal{H}_t}(\rho_\mu)),$$

Fix any $\mu \in \mathsf{Up}(\lambda_{\mathsf{hook},\ell})$ (so μ is a partition of n of the form $(\mu_1, \dots, \mu_r)$ where $\mu_1 \geq n - \ell$). We recall that the function $\mathcal{H}_t : \mathbb{S}_n \to [0,1]$ is defined by

$$\mathcal{H}_t = \sum_{j=0}^{\infty} \mathbf{Pr}_{\mathbf{T} \sim \mathsf{Poi}(t)}[\mathbf{T} = j](\mathrm{trans})^j,$$

where "$(\mathrm{trans})^T$" denotes T-fold convolution of trans. Since convolution corresponds to multiplication of Fourier coefficients, this gives that

$$\widehat{\mathcal{H}_t}(\rho_\mu) = \mathsf{c}(t,\mu) \cdot \mathsf{Id}, \text{ where } \mathsf{c}(t,\mu) := \sum_{j=0}^{\infty} \mathbf{Pr}_{\mathbf{T} \sim \mathsf{Poi}(t)}[\mathbf{T} = j](\mathsf{c}_{\mathrm{trans},\mu})^j. \tag{24}$$

Recalling Choi (1994) that the median of the Poisson distribution $\mathsf{Poi}(t)$ is at most $t + 1/3$, we get that

$$\mathsf{c}(t,\mu) \geq \frac{1}{2} \cdot (\mathsf{c}_{\mathrm{trans},\mu})^{t+1/3} \geq \frac{1}{2} \cdot e^{-O(\ell t)/n},$$

(where the second inequality uses $\ell \leq cn$ and $t \geq 1$), and the theorem is proved.

Appendix E. Representations of Ewens model noise

In this section we lower bound Equation (1) when the noise distribution $\mathcal{K}$ is $\mathcal{E}_\theta$, corresponding to the Ewens noise model with parameter θ:

Theorem 24 (Ewens model noise - Restatement of Theorem 11) *Let $\theta > 0$, let $\ell \in \{1, \dots, n\}$, and let $\eta := \mathrm{dist}(\theta, \ell) = \min_{j \in \{1,\dots,\ell\}} \left|e^\theta - j\right|$. Then (recalling Equation (1)) we have that*

$$\sigma_{\min, \mathsf{Up}(\mu_{\mathsf{hook},\ell}), \mathcal{E}_\theta} \geq (2n)^{-\ell} \eta^{2\sqrt{\ell}}. \tag{25}$$

Similar to the previous two sections, Theorem 24 follows immediately from the following lower bound on singular values of certain irreducible representations:

Lemma 25 *Let μ be a partition of n of the form $(\mu_1, \dots, \mu_r)$ where $\mu_1 \geq n - \ell$. Let $\theta > 0$ and let $\eta := \mathrm{dist}(\theta, \ell) = \min_{j \in \{1,\dots,\ell\}} \left|e^\theta - j\right|$. Then we have that*

$$\widehat{\mathcal{E}_\theta}(\rho_\mu) = c_{\mu,\theta} \cdot \mathsf{Id} \;\; \textit{where} \;\; |c_{\mu,\theta}| \geq (2n)^{-\ell} \eta^{2\sqrt{\ell}}.$$

To prove Lemma 25, we will need the notions of *content* and *hook length* for boxes in a Young diagram:

Definition 26 *Let μ be a partition $\mu \vdash n$. The* hook length *of a box u in the Young diagram for μ, denoted by $h(u)$, is the sum*

(# of boxes to the right of u in its row) + (# of boxes below u in its column) + 1 (for u itself).

The content *$c(u)$ of a box u is $c(u) := j - i$, where j is its column number (from the left, starting with column 1) and i is its row number (from the top, starting with row 1).*

7	5	2	1
4	1		
3	1		
1			

0	1	2	3
−1	0		
−2	−1		
−3			

Figure 1: On the left is a Young diagram in which each box has been labeled with its hook length; on the right is a Young diagram in which each box has been labeled with its content.

The left portion of Figure 1 depicts a Young diagram annotated with the hook lengths of each of its boxes. The right portion of Figure 1 depicts the same Young diagram annotated with the contents of each of its boxes.

We will need the following technical result to prove Lemma 25:

Lemma 27 *Let $\mu \vdash n$ and let χ_μ be the corresponding character in $\mathbb{S}_n$. For any $q \in \mathbb{R}$,*

$$\frac{1}{n!} \sum_{\sigma \in \mathbb{S}_n} \chi_\mu(\sigma) \cdot q^{\text{cycles}(\sigma)} = \prod_{u \in \mu} \frac{q + c(u)}{h(u)},$$

where the subscript "$u \in \mu$" means that u ranges over all the boxes in the Young diagram corresponding to μ.

Proof The above identity is given as Exercise 7.50 in Stanley's book Stanley (1999). For the sake of completeness, we provide the proof here.

For any $\bar{t} = (t_1, \ldots, t_n)$, we define the polynomial

$$a_{\bar{t}}(x_1, \ldots, x_n) := \det \begin{bmatrix} x_1^{t_1} & x_2^{t_1} & x_3^{t_1} & \ldots & x_n^{t_1} \\ x_1^{t_2} & x_2^{t_2} & x_3^{t_2} & \ldots & x_n^{t_2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_1^{t_n} & x_2^{t_n} & x_3^{t_n} & \ldots & x_n^{t_n} \end{bmatrix}.$$

Given any partition $\mu \vdash n$, we now define the Schur polynomial $s_\mu(x_1, \ldots, x_n)$ as follows: Define $\bar{t}_\mu = (\mu_1 + n - 1, \ldots, \mu_n + 0)$ and $\bar{t}_0 = (n-1, \ldots, 0)$. Then,

$$s_\mu(x_1, \ldots, x_n) := \frac{a_{\bar{t}_\mu}(x_1, \ldots, x_n)}{a_{\bar{t}_0}(x_1, \ldots, x_n)}.$$

The denominator is just the Vandermonde determinant of the variables $(x_1, \ldots, x_n)$. As the polynomial $a_{\bar{t}_\mu}(x_1, \ldots, x_n)$ is alternating, it follows that $s_\mu(x_1, \ldots, x_n)$ is a polynomial (as opposed to a rational function) and further, it is symmetric.

The following is a fundamental fact connecting Schur polynomials and cycles: For any $0 \leq k \leq n$,

$$s_\mu(\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{n-k}) = \sum_{\sigma \in \mathbb{S}_n} \frac{1}{n!} \cdot \chi_\mu(\sigma) \cdot k^{\text{cycles}(\sigma)} \tag{26}$$

(see equation 7.78 in Stanley (1999)). On the other hand, there are known explicit formulas for evaluations of the Schur polynomial at specific inputs. In particular, Corollary 7.21.4 of Stanley (1999) states that

$$s_\mu(\underbrace{1,\dots,1}_{k},\underbrace{0,\dots,0}_{n-k}) = \prod_{u\in\mu}\frac{k+c(u)}{h(u)}. \tag{27}$$

Combining (26) and (27), we get that for any $0 \le k \le n$, we have

$$\frac{1}{n!}\sum_{\sigma\in\mathbb{S}_n}\chi_\mu(\sigma)\cdot k^{\mathrm{cycles}(\sigma)} = \prod_{u\in\mu}\frac{k+c(u)}{h(u)}.$$

However, note that both the left and the right hand sides can be seen as polynomials of degree at most n in the variable k. Since they agree at $n+1$ values $k = 0,\dots,n$, they must be identical as formal functions. This concludes the proof. ■

Proof of Lemma 25. Recall that the distribution $\mathcal{E}_\theta$ over $\mathbb{S}_n$ is defined by $\mathcal{E}_\theta(\pi) = e^{-\theta d(\pi,e)}/Z(\theta)$, where $Z(\theta) = \sum_{\pi\in\mathbb{S}_n} e^{-\theta d(\pi,e)}$ is a normalizing constant. Since the Cayley distance $d(\sigma,\tau)$ is equal to $n - \mathrm{cycles}(\sigma^{-1}\tau)$, where $\mathrm{cycles}(\pi)$ is the number of cycles in π, we have that

$$\mathcal{E}_\theta(\pi) = \frac{e^{\theta\cdot\mathrm{cycles}(\pi)}}{C}, \text{ where } C = \sum_{\pi\in\mathbb{S}_n} e^{\theta\cdot\mathrm{cycles}(\pi)}.$$

Since the $\mathrm{cycles}(\cdot)$ function is a class function so is $\mathcal{E}_\theta$, so we can apply Lemma 39 and we get that $\widehat{\mathcal{E}_\theta}(\rho_\mu) = c_{\mu,\theta}\cdot \mathsf{Id}$, where

$$c_{\mu,\theta} = \frac{\sum_{\sigma\in\mathbb{S}_n}\mathcal{E}_\theta(\sigma)\cdot\chi_\mu(\sigma)}{\mathsf{dim}(\rho_\mu)} = \frac{\sum_{\sigma\in\mathbb{S}_n} e^{\theta\cdot\mathrm{cycles}(\sigma)}\cdot\chi_\mu(\sigma)}{\mathsf{dim}(\rho_\mu)\cdot(\sum_{\sigma\in\mathbb{S}_n} e^{\theta\cdot\mathrm{cycles}(\sigma)})} = \frac{\sum_{\sigma\in\mathbb{S}_n} q^{\mathrm{cycles}(\sigma)}\cdot\chi_\mu(\sigma)}{\mathsf{dim}(\rho_\mu)\cdot(\sum_{\sigma\in\mathbb{S}_n} q^{\mathrm{cycles}(\sigma)})},$$

where $q := e^\theta$. We re-express the numerator by applying Lemma 27 to get

$$\sum_{\sigma\in\mathbb{S}_n} q^{\mathrm{cycles}(\sigma)}\cdot\chi_\mu(\sigma) = n!\cdot\prod_{u\in\mu}\frac{q+c(u)}{h(u)}. \tag{28}$$

To analyze the denominator of $c_{\mu,\theta}$, applying Lemma 27 to the trivial partition $\mathsf{Triv}_n = (n)$ of n (the character of which is identically 1), we get that

$$\sum_{\sigma\in\mathbb{S}_n} q^{\mathrm{cycles}(\sigma)} = n!\cdot\prod_{u\in\mathsf{Triv}_n}\frac{q+c(u)}{h(u)} = q(q+1)\cdots(q+n-1). \tag{29}$$

For the rest of the denominator, we recall the following well-known fact about the dimension of irreducible representations of the symmetric group:

Fact 28 (Hook length formula, see e.g. Theorem 3.41 of Méliot (2017)) *For $\mu\vdash n$, $\mathsf{dim}(\rho_\mu) = \frac{n!}{\prod_{u\in\mu}h(u)}$.*

Combining (28), (29) and Fact 28, we get

$$c_{\mu,\theta} = \frac{\prod_{u\in\mu}(q+c(u))}{q(q+1)\cdots(q+n-1)}. \tag{30}$$

Let $\mathcal{A}$ denote the set consisting of the cells of the Young diagram of μ which are not in the first row. Since $n-\mu_1 = \ell'$ for some $\ell' \le \ell$, the above expression simplifies to

$$c_{\mu,\theta} = \frac{\prod_{u\in\mathcal{A}}(q+c(u))}{(q+n-\ell')\cdots(q+n-1)}. \tag{31}$$

To bound this ratio, first observe that both the numerator and denominator are ℓ'-way products. There are two possibilities now:

1. **Case 1:** $q \ge \ell+1$**.** In this case we observe that each cell $u \in \mathcal{A}$ satisfies $c(u) \ge -\ell' \ge -\ell$. Thus $c_{\mu,\theta}$ can be expressed as a product of ℓ' many fractions, each of which is at least $\frac{q-\ell}{q+n-1} \ge \frac{1}{\ell+n}$. This implies that

$$c_{\mu,\theta} \ge \left(\frac{1}{n+\ell}\right)^{\ell'} \ge (2n)^{-\ell}.$$

2. **Case 2:** $q \le \ell$**.** In this case, the denominator of Equation (31) is at most $(2n)^{\ell}$. To lower bound the numerator, observe that for every cell u of $\mathcal{A}$, the value of $c(u)$ is an integer in $\{-\ell,\dots,\ell\}$. Let j_0 and j_1 denote the two values in $\{-\ell,\dots,\ell\}$ for which $|q-j|$ achieves its smallest value η and its next smallest value (note that these values are equal if $\eta = 1/2$). Next, we observe that at most $\sqrt{\ell}$ many cells of $\mathcal{A}$ have content equal to any given fixed integer value. Since j_0 and j_1 are the only possible values of $j \in \{-\ell,\dots,\ell\}$ for which $|q+j| < 1$, it follows that

$$\prod_{u\in\mathcal{A}} |(q+c(u))| \ge \left(\prod_{u\in\mathcal{A}:c(u)=j_0} |(q+c(u))|\right)\cdot\left(\prod_{u\in\mathcal{A}:c(u)=j_1} |(q+c(u))|\right) \ge \eta^{2\sqrt{\ell}}.$$

This finishes the proof of Lemma 25.

Appendix F. Lower bound for Ewens models

Recall that because of the $\mathrm{poly}(\mathrm{dist}(\theta,\log k)^{-\sqrt{\log k}})$ dependence in Theorem 4, the algorithm of that theorem is inefficient if e^{θ} is very close to an integer. In this section we prove Theorem 5, which establishes that *any algorithm* for learning in the presence of Ewens noise *must* be inefficient if e^{θ} is very close to an integer.

F.1. A key technical result

The following lemma is at the heart of our lower bound. It shows that if e^{θ} is close to an integer, then any partition μ of $n \ge m$ which extends a particular partition λ_{sq} of m must be such that the Fourier coefficient $\widehat{\mathcal{E}}_{\theta}(\rho_{\mu})$ of Ewens noise has small singular values.

Lemma 29 *Let λ_{sq} denote the partition $(t,\dots,t)$ of $m = t(t+j)$ whose Young tableau is a rectangle with $t+j$ rows and t columns. Let $\theta > 0$ be such that $\left|e^\theta - j\right| \le \eta$ where $\eta \le 1/2$. Let $n \ge m$, $\mu \vdash n$ and $\lambda_{\mathsf{sq}} \Uparrow \mu$ (recall Definition 43). Then*

$$\widehat{\mathcal{E}}_\theta(\rho_\mu) = c_{\mu,\theta} \cdot \mathsf{Id}, \quad \text{where } c_{\mu,\theta} \le \eta^t.$$

Here ρ_μ denotes the irreducible representation of $\mathbb{S}_n$ corresponding to the partition μ.

Proof Let $\mu = (\mu_1,\dots,\mu_r)$. By Lemma 25, we have that

$$\widehat{\mathcal{E}}_\theta(\rho_\mu) = c_{\mu,\theta} \cdot \mathsf{Id},$$

where Equation (31) gives the precise value of $c_{\mu,\theta}$ as

$$c_{\mu,\theta} = \frac{\prod_{u\in\mathcal{A}}(q+c(u))}{\prod_{u\in\mathcal{B}}(q+c(u))}, \quad \text{where} \quad q = e^\theta. \tag{32}$$

Here $\mathcal{A}$ denotes the set of cells of the Young diagram of μ which are not in the first row and $\mathcal{B}$ denotes the rightmost $n-\mu_1$ many cells in the Young diagram of the trivial partition $\mathsf{Triv}_n = (n)$. Note that in this lemma, we are trying to upper bound Equation (32) whereas Lemma 25 was about lower bounding this quantity.

To upper bound Equation (32), we first observe that there is an obvious bijection $\Phi : \mathcal{A} \to \mathcal{B}$ such that if $\Phi(u) = v$, then $c(v) > |c(u)| > 0$.

Next, let $\mathcal{A}_{-j} \subset \mathcal{A}$ be $\mathcal{A} := \{(r,s) : s - r = j \text{ and } (r,s) \in \mathcal{A}\}$. Since $\lambda_{\mathsf{sq}} \Uparrow \mu$, it follows that $|\mathcal{A}_{-j}| \ge t$. As a result, we can upper bound $c_{\mu,\theta}$ as follows:

$$\begin{aligned} c_{\mu,\theta} = \frac{\prod_{u\in\mathcal{A}}(q+c(u))}{\prod_{u\in\mathcal{B}}(q+c(u))} &= \prod_{u\in\mathcal{A}} \frac{q+c(u)}{q+c(\Phi(u))} = \left(\prod_{u\in\mathcal{A}_{-j}} \frac{q+c(u)}{q+c(\Phi(u))}\right)\left(\prod_{u\in\mathcal{A}\setminus\mathcal{A}_{-j}} \frac{q+c(u)}{q+c(\Phi(u))}\right) \\ &\le \prod_{u\in\mathcal{A}_{-j}} q + c(u) \qquad \text{(using } c(\Phi(u)) > |c(u)| > 0 \text{ and } q > 0) \\ &\le \eta^t. \end{aligned}$$

■

F.2. Proof of Theorem 5

Theorem 5 is an immediate consequence of the following result. It shows that if e^θ is close to an integer j, then it may be statistically impossible to learn a distribution f supported on k rankings without using many samples from $\mathcal{E}_\theta * f$:

Theorem 30 *Given $j \in \mathbb{N}$, there are infinitely many values of k and $m = m(k) \approx \frac{\log k}{\log\log k}$ such that the following holds: there are two distributions f_1, f_2 over $\mathbb{S}_m$ with the following properties:*

1. *$d_{\mathrm{TV}}(f_1, f_2) = 1$ (i.e. the distributions f_1 and f_2 have disjoint support);*
2. *$|\mathrm{supp}(f_1)|, |\mathrm{supp}(f_2)| \le k$;*
3. *For any $\theta > 0$ such that $|e^\theta - j| \le \eta \le 1/2$, we have that $d_{\mathrm{TV}}(\mathcal{E}_\theta * f_1, \mathcal{E}_\theta * f_2) \le 2 \cdot \eta^{\Theta\left(\sqrt{\frac{\log k}{\log\log k}}\right)}$.*

Proof Let $t \geq j$ be any integer, let $m = t(t+j)$, and let $k = m!$. We first construct the two distributions f_1, f_2 over $\mathbb{S}_m$ and argue that properties (1) and (2) hold.

Let $\lambda_{\mathsf{sq}} \vdash m$ be the partition whose Young tableau is a rectangle with $t+j$ rows and t columns. Let us consider the character $\chi_{\mathsf{sq}} : \mathbb{S}_m \to \mathbb{Q}$ corresponding to the partition λ_{sq}. By Fact 46 we have that χ_{sq} is rational valued, and by Theorem 38 we have that $\sum_{\sigma \in \mathbb{S}_n} \chi_{\mathsf{sq}}(\sigma) = 0$. Thus, we have that

$$\sum_{\sigma \in \mathbb{S}_n} |\chi_{\mathsf{sq}}(\sigma)| \cdot \mathbf{1}_{\chi_{\mathsf{sq}}(\sigma) > 0} = \sum_{\sigma \in \mathbb{S}_n} |\chi_{\mathsf{sq}}(\sigma)| \cdot \mathbf{1}_{\chi_{\mathsf{sq}}(\sigma) < 0} =: C_{\mathsf{sq}} \tag{33}$$

for some C_{sq} (which is nonzero again by Theorem 38). We now define distributions f_1 and f_2 over $\mathbb{S}_m$ as

$$f_1(\sigma) = \begin{cases} \frac{1}{C_{\mathsf{sq}}} \cdot \chi_{\mathsf{sq}}(\sigma) & \text{if } \chi_{\mathsf{sq}}(\sigma) > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad f_2(\sigma) = \begin{cases} \frac{-1}{C_{\mathsf{sq}}} \cdot \chi_{\mathsf{sq}}(\sigma) & \text{if } \chi_{\mathsf{sq}}(\sigma) < 0 \\ 0 & \text{otherwise.} \end{cases}$$

From their definitions and Equation (33) it is immediate that f_1 and f_2 are distributions over $\mathbb{S}_m$ which have disjoint support. Since $|\mathbb{S}_m| = k$, this gives items 1 and 2 of the theorem.

To prove the third item, observe (recalling the comment immediately after Definition 37) that the function $g : \mathbb{S}_m \to \mathbb{C}$, defined as $g(\sigma) := f_1(\sigma) - f_2(\sigma) = \frac{1}{C_{\mathsf{sq}}} \cdot \chi_{\mathsf{sq}}(\sigma)$, is a class function. Choose any partition $\lambda \vdash m$ and the corresponding irreducible representation ρ_λ of $\mathbb{S}_m$. By applying Lemma 39, we have that

$$\widehat{g}(\rho_\lambda) = c_\lambda \cdot \mathsf{Id} \quad \text{where} \quad c_\lambda = \frac{\sum_{\sigma \in \mathbb{S}_m} g(\sigma) \cdot \chi_\lambda(\sigma)}{\dim(\rho_\lambda)}. \tag{34}$$

We analyze the multiplier c_λ by noting that

$$\begin{aligned} c_\lambda = \frac{\sum_{\sigma \in \mathbb{S}_m} g(\sigma) \cdot \chi_\lambda(\sigma)}{\dim(\rho_\lambda)} &= \frac{\sum_{\sigma \in \mathbb{S}_m} \chi_{\mathsf{sq}}(\sigma) \cdot \chi_\lambda(\sigma)}{\dim(\rho_\lambda) \cdot C_{\mathsf{sq}}} \\ &= \frac{m! \cdot \mathbf{1}[\lambda = \lambda_{\mathsf{sq}}]}{\dim(\rho_\lambda) \cdot C_{\mathsf{sq}}} \text{ using Theorem 38.} \end{aligned} \tag{35}$$

Thus, we have

$$\begin{aligned}
\|\mathcal{E}_\theta * f_1 - \mathcal{E}_\theta * f_2\|_1 &= \sum_{\sigma \in \mathbb{S}_m} |\mathcal{E}_\theta * f_1(\sigma) - \mathcal{E}_\theta * f_2(\sigma)| \\
&= \sum_{\sigma \in \mathbb{S}_m} |\mathcal{E}_\theta * g(\sigma)| && \text{(linearity and } g = f_1 - f_2\text{)} \\
&= \frac{1}{m!} \sum_{\sigma \in \mathbb{S}_m} \left| \sum_{\mu \vdash m} \dim(\rho_\mu) \mathsf{Tr}[\widehat{\mathcal{E}_\theta * g}(\rho_\mu) \rho_\mu(\sigma^{-1})] \right| \\
& && \text{(Definition 35, inverse Fourier transform of } \mathcal{E}_\theta * g\text{)} \\
&= \frac{1}{m!} \sum_{\sigma \in \mathbb{S}_m} \left| \sum_{\mu \vdash m} \dim(\rho_\mu) \mathsf{Tr}[\widehat{\mathcal{E}}_\theta(\rho_\mu) \widehat{g}(\rho_\mu) \rho_\mu(\sigma^{-1})] \right| && \text{(convolution identity)} \\
&= \frac{1}{\dim(\rho_{\lambda_{\mathsf{sq}}}) \cdot C_{\mathsf{sq}}} \sum_{\sigma \in \mathbb{S}_m} \left| \dim(\rho_{\lambda_{\mathsf{sq}}}) \mathsf{Tr}[\widehat{\mathcal{E}}_\theta(\rho_{\lambda_{\mathsf{sq}}}) \rho_{\lambda_{\mathsf{sq}}}(\sigma^{-1})] \right| \\
& && \text{(Equations 34 and 35)} \\
&= \frac{1}{C_{\mathsf{sq}}} \sum_{\sigma \in \mathbb{S}_m} \left| \mathsf{Tr}[\widehat{\mathcal{E}}_\theta(\rho_{\lambda_{\mathsf{sq}}}) \rho_{\lambda_{\mathsf{sq}}}(\sigma^{-1})] \right| && (36)
\end{aligned}$$

To deal with $\widehat{\mathcal{E}}_\theta(\rho_{\lambda_{\mathsf{sq}}})$, we apply Lemma 29. In particular, by setting $n = m$ and $\mu = \lambda_{\mathsf{sq}}$ in Lemma 29, we get that

$$\widehat{\mathcal{E}}_\theta(\rho_{\lambda_{\mathsf{sq}}}) = c_{\lambda_{\mathsf{sq}},\theta} \cdot \mathsf{Id},$$

where $|c_{\lambda_{\mathsf{sq}},\theta}| \leq \eta^t$, and we thus get that

$$\|\mathcal{E}_\theta * f_1 - \mathcal{E}_\theta * f_2\|_1 \leq \frac{\eta^t}{C_{\mathsf{sq}}} \cdot \sum_{\sigma \in \mathbb{S}_m} \left| \mathsf{Tr}[\rho_{\lambda_{\mathsf{sq}}}(\sigma^{-1})] \right| = \frac{\eta^t}{C_{\mathsf{sq}}} \cdot \sum_{\sigma \in \mathbb{S}_m} \left| \chi_{\mathsf{sq}}(\sigma^{-1}) \right|. \tag{37}$$

Finally, recalling that

$$C_{\mathsf{sq}} = \frac{\sum_{\sigma \in \mathbb{S}_n} |\chi_{\mathsf{sq}}(\sigma)|}{2},$$

we get that the RHS of Equation (37) is $2\eta^t$. Recalling that $t \geq \sqrt{m/2}$, the theorem is proved. ■

Appendix G. Basics of representation theory over the symmetric group

Representation theory of the symmetric group $\mathbb{S}_n$ is at the technical core of this paper. In this appendix we briefly review the definitions and results that we require, starting first with general groups and then specializing to $\mathbb{S}_n$ as necessary. See Curtis and Reiner Curtis and Reiner (1966) (or many other sources) for an extensive reference on representation theory of finite groups and James James (2006) or Méliot Méliot (2017) for an extensive reference on representation theory of $\mathbb{S}_n$.

G.1. General groups

We start by recalling the definition of a representation:

Definition 31 *For any group G, a* representation $\rho : G \to \mathbb{C}^{m\times m}$ *is a group homomorphism, i.e. a function from G to $\mathbb{C}^{m\times m}$ that satisfies $\rho(g)\cdot\rho(h) = \rho(g\cdot h)$ for all $g, h \in G$. The* dimension *of such a representation ρ is m.*

In this paper, unless otherwise mentioned, all representations ρ are *unitary* – in other words, for every $g \in G$, $\rho(g)$ is a unitary matrix. Over finite groups, any representation can be made unitary by applying a similarity transformation; by this we mean that if ρ is a representation, then there is an invertible matrix Z such that the new map $\tilde{\rho}$ defined as $\tilde{\rho}(g) = Z^{-1}\cdot\rho(g)\cdot Z$ is a unitary representation. (The reader should verify that as long as Z is invertible, the map $\tilde{\rho}$ is always a representation if ρ is a representation.) Two such representations ρ and $\tilde{\rho}$ are said to be *equivalent*.

Next we recall the notion of an irreducible representation:

Definition 32 *A representation $\rho : G \to \mathbb{C}^{m\times m}$ is said to be* reducible *if there exists a proper subspace V of $\mathbb{C}^m$ such that $\rho(g)\cdot V \subseteq V$ for all $g \in G$. If there is no such proper subspace V, then ρ is said to be* irreducible*.*

It is well known that any finite group has only finitely many irreducible representations, up to the above notion of equivalence, and that every representation of a finite group G can be written as a direct sum of irreducible representations:

Theorem 33 (Maschke's theorem, see e.g. Theorem 1.3 of Méliot (2017)) *For G a finite group, there is a finite set of distinct irreducible representations $\{\rho_1, \ldots, \rho_r\}$ such that for any representation $\rho : G \to \mathbb{C}^{m\times m}$, there is a invertible transformation $Z \in \mathbb{C}^{m\times m}$ such that $Z^{-1}\rho Z$ is block diagonal where each block is one of $\{\rho_1, \ldots, \rho_r\}$. In other words, $Z^{-1}\rho Z$ is equal to the direct sum $\oplus_{\ell=1}^{M}\mu_\ell$ where each μ_ℓ is an element of $\{\rho_1, \ldots, \rho_r\}$.*

We remind the reader that elements g, h in a group G are said to be *conjugates* if there is an element $t \in G$ such that $tgt^{-1} = h$. Define $\mathsf{Cl}(g)$, the *conjugacy class* of g, to be $\{h : h \text{ is conjugate to } g\}$; it is easy to see that the different conjugacy classes form a partition of G.

We recall some very standard facts about irreducible representations:

Theorem 34 (see e.g. Theorem 2.3.1 of Green and Wigderson (2010)) *Let G be a finite group and let $\{\rho_1, \ldots, \rho_r\}$ be the set of its irreducible representations, where $\rho_i : G \to \mathbb{C}^{d_i\times d_i}$. Then*

1. $\sum_{i=1}^{r} d_i^2 = |G|$.
2. *The number of conjugacy classes is equal to r, the number of distinct irreducible representations.*
3. *For $1 \le s, t \le d_i$, let $\rho_{i,s,t} : G \to \mathbb{C}$ be the (s,t) entry of $\rho_i(g)$. Then, for $1 \le i_1, i_2 \le r$, $1 \le s_1, t_1 \le d_{i_1}$ and $1 \le s_2, t_2 \le d_{i_2}$*

$$\mathop{\mathbf{E}}_{g\in G}[\rho_{i_1,s_1,t_1}(g)\cdot\overline{\rho_{i_2,s_2,t_2}(g)}] = \begin{cases} \frac{1}{d_{i_1}} & \text{if } i_1 = i_2,\ s_1 = s_2 \text{ and } t_1 = t_2 \\ 0 & \text{otherwise} \end{cases}$$

4. The representations $\rho_1, \ldots, \rho_r$ are unitary.

A restatement of (3) above is that the functions $\{\rho_{i,s,t}(\cdot)\}$ are orthogonal. Combining this with $\sum_{i=1}^{r} d_i^2 = |G|$ (given by (1)), we get that the functions $\{\rho_{i,s,t}\}_{1 \le i \le r, 1 \le s,t \le d_i}$ form an orthogonal basis for $\mathbb{C}^G$.

With an orthonormal basis for the set of complex-valued functions on G in hand (in other words, a basis for the group algebra $\mathbb{C}[G]$), we are ready to define the *Fourier transform* of a function $f : G \to \mathbb{C}$:

Definition 35 *Let G be a finite group with irreducible representations given by $\{\rho_1, \ldots, \rho_r\}$ and let $f : G \to \mathbb{C}$. The* Fourier transform *of f is given by matrices $\widehat{f}(\rho_1), \ldots, \widehat{f}(\rho_r)$, where*

$$\widehat{f}(\rho_i) = \sum_{g \in G} f(g) \cdot \rho_i(g).$$

The inverse transform is given by

$$f(g) = \frac{1}{|G|} \sum_{i=1}^{r} \dim(\rho_i) \mathsf{Tr}[\widehat{f}(\rho_i)\rho_i(g^{-1})].$$

Parseval's identity states that for any f as above, we have

$$\sum_{i=1}^{r} \|\widehat{f}(\rho_i)\|_F^2 = |G| \cdot \sum_{g \in G} |f(g)|^2. \tag{38}$$

We next recall the definition of characters and class functions for a group G.

Definition 36 *Given a finite group G, a function $f : G \to \mathbb{C}$ is said to be a* class function of G *if $f(g)$ only depends on the conjugacy class of g, i.e. $f(g) = f(hgh^{-1})$ for every $h \in G$.*

Definition 37 *The* character $\chi_\rho : G \to \mathbb{C}$ *corresponding to a representation $\rho : G \to \mathbb{C}^{m \times m}$ is given by $\chi_\rho(g) := \mathsf{Tr}(\rho(g))$.*

We observe that $\chi_\rho(\cdot)$ is a class function of G, and that if ρ and $\tilde{\rho}$ are unitarily equivalent, then $\chi_\rho(\cdot) = \chi_{\tilde{\rho}}(\cdot)$. We recall some standard facts about characters and class functions:

Theorem 38 *Let G be a finite group and let $\{\rho_1, \ldots, \rho_r\}$ be its set of irreducible representations. Let $\chi_{\rho_1}, \ldots, \chi_{\rho_r}$ be the corresponding characters. Then we have:*

1. *[Schur's lemma] $\mathbf{E}_{g \in G}[\chi_{\rho_i}(g) \cdot \overline{\chi_{\rho_j}(g)}] = \delta_{i,j}$.*
2. *The functions $\{\chi_{\rho_i}(\cdot)\}_{1 \le i \le r}$ forms an orthonormal basis for all class functions of G.*

The following (standard) claim shows that the Fourier transform of any class function f is a diagonal matrix (in fact, a scalar multiple of the identity matrix):

Lemma 39 *Let $f : G \to \mathbb{C}$ be a class function and let $\rho : G \to \mathbb{C}^{m \times m}$ be an irreducible representation of G. Then $\widehat{f}(\rho) = c \cdot \mathsf{Id}$ where $c = \frac{\sum_{g \in G} f(g)\chi_\rho(g)}{m}$ and Id is the identity matrix.*

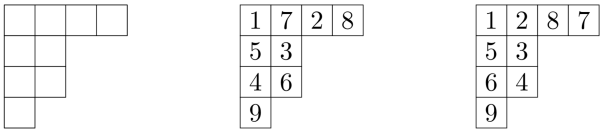

Figure 2: On the left is the Young diagram for the partition $\lambda = (4, 2, 2, 1)$. The middle and right present two equivalent Young tableaus for $(\{1, 7, 2, 8\}, \{5, 3\}, \{4, 6\}, \{9\})$.

Proof Choose any $h \in G$, and observe that

$$\begin{aligned}
\rho(h) \cdot \widehat{f}(\rho) &= \rho(h) \cdot \big(\sum_{g \in G} f(g)\rho(g) \big) \\
&= \rho(h) \cdot \big(\sum_{g \in G} f(h^{-1}gh)\rho(h^{-1}gh) \big) = \rho(h) \cdot \big(\sum_{g \in G} f(g)\rho(h^{-1}gh) \big) \\
&= \rho(h) \cdot \rho(h^{-1}) \cdot \big(\sum_{g \in G} f(g)\rho(g) \big) \cdot \rho(h) = \widehat{f}(\rho) \cdot \rho(h).
\end{aligned}$$

As a consequence of Schur's lemma, we have that if a matrix A is such that $A \cdot \rho(h) = \rho(h) \cdot A$ for all $h \in G$, then $A = c \cdot \mathsf{Id}$. Thus, we get that $\widehat{f}(\rho) = c \cdot \mathsf{Id}$. The lemma follows by taking trace on both sides. ■

G.2. Representation theory of the symmetric group

Representation theory of the symmetric group has many applications to algebra, combinatorics and statistical physics and has been intensively studied (as mentioned earlier, see e.g. James (2006); Méliot (2017) for detailed treatments). Below we only recall a few basics which we will need.

The first notion we require is that of a *Young diagram.* Consider a partition $\lambda = (\lambda_1, \ldots, \lambda_k)$ of n where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$ and $\lambda_1 + \ldots + \lambda_k = n$. We indicate that λ is such a partition by writing "$\lambda \vdash n$." The Young diagram corresponding to such a partition λ is a two-dimensional left-justified array of empty cells in which the i^{th} row has λ_i cells. See the left portion of Figure 2 for an example of a Young diagram. A *Young tableau* corresponding to a partition λ is obtained by filling in the n cells of the Young diagram with the elements of $[n]$, using each element exactly once, where the ordering within rows of the Young diagram is irrelevant.

For each partition $\lambda = (\lambda_1, \ldots, \lambda_k)$ of n, there is an associated representation, denoted τ_λ, which we now define. Let $N_\lambda = \binom{n}{\lambda_1, \ldots, \lambda_k}$ be the number of Young tableaus corresponding to partition λ, and let $\mathsf{Y}_{\lambda,1}, \ldots, \mathsf{Y}_{\lambda,N_\lambda}$ be an enumeration of these tableaus in some order.

Definition 40 *The* permutation representation τ_λ corresponding to λ *is defined as follows: For each $g \in \mathbb{S}_n$, $\tau_\lambda(g)$ is the $N_\lambda \times N_\lambda$ matrix (where we view rows and columns as indexed by Young tableaus corresponding to λ) which has $\tau_\lambda(g)(i, j) = 1$ iff $\mathsf{Y}_{\lambda,i}$ maps to $\mathsf{Y}_{\lambda,j}$ under the action of g.*

It is easy to check that $\tau_\lambda : \mathbb{S}_n \to \mathbb{C}^{N_\lambda \times N_\lambda}$ as defined above is indeed a representation. In fact, since the range of τ_λ is always a permutation matrix, τ_λ is also a unitary representation.

It turns that for $\lambda \neq (n)$, the permutation representation τ_λ is not an irreducible representation. However, it also turns out that all of the irreducible representations of $\mathbb{S}_n$ can be obtained from the permutation representations. To explain this, we need to define a partial order over partitions of n:

Definition 41 *For two partitions λ and μ of n, we say that λ* dominates *μ, written $\lambda \rhd \mu$, if $\sum_{j \leq i} \lambda_j \geq \sum_{j \leq i} \mu_j$ for all $i > 0$. The partial order defined by $\rhd$ is said to be the* dominance order *over the partitions (equivalently, Young diagrams) of n.*

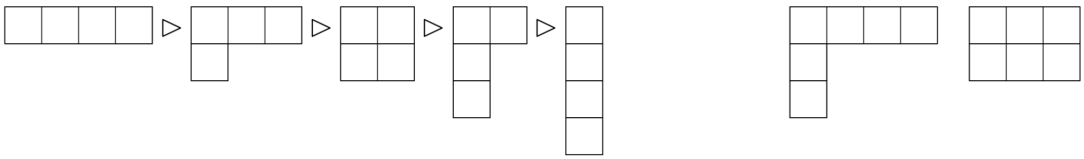

Figure 3: The left part of the picture depicts the dominance order across the partitions of 4; it happens to be the case that the dominance order is a total order across the partitions of 4. This is not true in general; as depicted on the right, the two partitions $(4,1,1)$ and $(3,3)$ of 6 are incomparable under the dominance order.

The next result explains how the irreducible representations of $\mathbb{S}_n$ can be obtained from the representations $\{\tau_\lambda\}_{\lambda \vdash n}$:

Theorem 42 (James submodule theorem, see e.g. Theorem 3.34 of Méliot (2017)) *The irreducible representations of $\mathbb{S}_n$ are in one-to-one correspondence with the partitions $\lambda \vdash n$; we denote the irreducible representation corresponding to λ by ρ_λ. In particular, when $\lambda = (n)$, then ρ_λ is the trivial irreducible representation (which maps each $g \in G$ to 1). Moreover, each permutation representation τ_λ is a direct sum of irreducible representations corresponding to partitions which dominate λ, i.e.*

$$\tau_\lambda = \bigoplus_{\mu \rhd \lambda} \bigoplus_{\ell=1}^{K_{\lambda,\mu}} \rho_\mu.$$

Here the $K_{\lambda,\mu}$'s are non-negative integers, known as the Kostka numbers*, which are such that $K_{\lambda,\lambda} = 1$.*

G.2.1. RESTRICTIONS OF IRREDUCIBLE REPRESENTATIONS

Fix $\lambda \vdash n$ and consider the irreducible representation ρ_λ of $\mathbb{S}_n$. For any $m \leq n$, $\mathbb{S}_m$ can be viewed as the subgroup of $\mathbb{S}_n$ where elements $\{m+1, \ldots, n\}$ are fixed. Hence ρ_λ can also be viewed as a representation of $\mathbb{S}_n$; this representation of $\mathbb{S}_m$ is written ρ_λ^m and is called the *restriction* of ρ_λ to $\mathbb{S}_m$. Note that ρ_λ^m may not be an *irreducible* representation of $\mathbb{S}_m$. By Theorem 33, we have that ρ_λ^m is equivalent to some direct sum

$$\bigoplus_{\mu \vdash m} M_{\lambda,\mu} \rho_\mu,$$

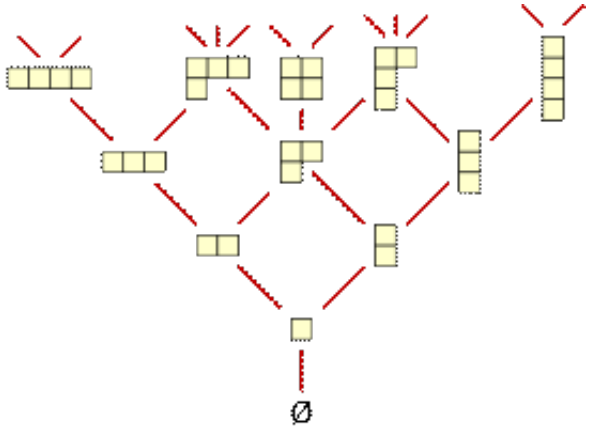

Figure 4: The first five levels of Young's lattice.

in which there are $M_{\lambda,\mu}$ many copies of p_μ, for some non-negative integers $M_{\lambda,\mu}$. These integers are given by the so-called "branching rule" on Young's lattice, which we now describe.

Definition 43 Young's lattice *is the partially ordered set of Young diagrams in which the partial order is given by inclusion in the following sense: given partitions μ and λ, we write "$\mu \uparrow \lambda$" if λ can be obtained by adding one box to μ (in such a way that λ is a valid partition, of course). If there are partitions $\mu^1, \ldots, \mu^r$ such that $\mu^1 \uparrow \mu^2 \uparrow \cdots \uparrow \mu^r$, we write "$\mu^1 \Uparrow \mu^r$."*

It is convenient to draw Young's lattice in such a way that the n-th level contains all and only the Young diagrams with n boxes. The diagram in Figure 4 depicts the first five levels of Young's lattice.

The next result, known as the "branching rule," states that for $\lambda \vdash n$, ρ_λ splits into a direct sum of ρ_μ over all $\mu \uparrow \lambda$ when ρ_λ is restricted to $\mathbb{S}_{n-1}$:

Lemma 44 (Branching rule) *Let λ be a partition of n and let ρ_λ be the corresponding irreducible representation of $\mathbb{S}_n$. Then ρ_λ^{n-1}, the restriction of ρ_λ to $\mathbb{S}_{n-1}$, is equivalent to*

$$\bigoplus_{\mu \vdash n-1 \,:\, \mu \uparrow \lambda} \rho_\mu.$$

By applying Lemma 44 inductively we get a complete description of how ρ_λ splits when it is restricted to any $\mathbb{S}_m$, $m < n$:

Theorem 45 *Let $\lambda \vdash n$ and let ρ_λ be the corresponding irreducible representation of $\mathbb{S}_n$. For $m < n$ we have that ρ_λ^m, the restriction of ρ_λ to $\mathbb{S}_m$, is equivalent to*

$$\bigoplus_{\mu \vdash m} \mathrm{Paths}(\mu, \lambda) \rho_\mu,$$

where $\mathrm{Paths}(\mu, \lambda)$ denotes the number of paths in Young's lattice from μ to λ.

Irreducible characters of the symmetric group. Finally, we recall the following fundamental fact (which is a consequence, e.g., of the Murnaghan-Nakayama rule) which we will use:

Fact 46 *[see e.g. Theorem 3.10 in Méliot (2017)] Let $\chi : \mathbb{S}_m \to \mathbb{C}$ be a character of $\mathbb{S}_m$. Then in fact χ is $\mathbb{Q}$-valued.*