Exploiting Probabilistic Independence for Permutations

Jonathan Huang, Carlos Guestrin
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Xiaoye Jiang, Leonidas Guibas
Stanford University
Stanford, California 94305

Abstract

Permutations are ubiquitous in many real world problems, such as voting, rankings and data association. Representing uncertainty over permutations is challenging, since there are $n!$ possibilities. Recent Fourier-based approaches can be used to provide a compact representation over low-frequency components of the distribution. Though polynomial, the complexity of these representations grows very rapidly, especially if we want to maintain reasonable estimates for peaked distributions. In this paper, we first characterize the notion of probabilistic independence for distributions over permutations. We then present a method for factoring distributions into independent components in the Fourier domain, and use our algorithms to decompose large problems into much smaller ones. We demonstrate that our method provides very significant improvements in terms of running time, on real tracking data.

1 Introduction

The need to reason about permutations arises in a broad variety of applications, such as information retrieval, webpage ranking, preference elicitation, and multiobject tracking (Huang et al., 2007; Lebanon and Mao). Yet exact solutions remain hopelessly intractable due to the fact that there are $n!$ permutations and that compact representations of uncertainty, such as graphical models, are ineffective due to the mutual exclusivity constraints that are associated with permutations.

A recent strand of research in the machine learning community (Huang et al., 2007; Kondor et al., 2007), however, has shown that maintaining the "low frequency" Fourier coefficients of a distribution over the symmetric group S_n (the group of permutations of n objects) offers a promising new approach for approximate inference over previous methods. Low frequency Fourier coefficients capture intuitive marginals and (Huang et al., 2007; Kondor et al., 2007) have developed a collection of general and efficient approximate inference operations, like marginalization and conditioning, which can be performed completely in the Fourier domain. Unfortunately, the current approach suffers from two shortcomings in scalability and accuracy:

- While low frequency Fourier coefficients provide a principled approximation to the underlying distribution and only require storing polynomially many numbers, the polynomials can grow quite fast for practical applications.
- Bandlimited approximations which discard high frequencies are most effective with diffuse distributions since smooth functions tend to be well approximated by linear combinations of low frequency basis functions, but are less effective at approximating highly peaked distributions.

In a sense, the two shortcomings listed above are at odds with each other since we can always achieve better approximations to sharp functions by maintaining higher frequency Fourier coefficients. But an interesting observation is that when the distribution is sharp, it often makes more sense to break up the problem into smaller parts and to reason about disjoint subsets of objects independently of each other.

Consider the *identity management* problem, for example, that arises in multiobject tracking, where one must maintain a belief over the joint one-to-one assignment of n tracks to n identities (Alice is at Track 1, Bob is at Track 2, etc.). If we are completely uncertain about the assignment of people to tracks, and have a uniform distribution over permutation, this smooth distribution can be represented with

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

only one parameter in the Fourier domain. At the limit when we know the location of every identity, our distribution becomes very peaked, and we need to maintain $n!$ Fourier coefficients. In this peaked setting, however, there is no reason to track all n identities jointly, and we can break up the problem into n subproblems. In this paper, we propose a principled method based on exploiting probabilistic independence which overcomes both issues and show that in practice, we can indeed "get the best of both worlds".

The main contributions of our paper are:

- We characterize the constraints on the Fourier coefficients of a distribution over permutations implied by probabilistic independence, and present two simple algorithms, *Join* and *Split*, which operate entirely in the Fourier domain for combining factors to form a joint distribution and factoring a distribution, respectively. Our algorithms are fully general in the sense that they work for any distribution over permutations.
- We identity the number of Fourier terms which are required in our Join/Split algorithms to achieve a desired number of Fourier terms in the result, and analyze the behaviour of our algorithms in near-independent situations in which a distribution does not fully factor.
- We discuss a method for detecting probabilistic independence using only Fourier coefficients.
- Finally, we use our algorithms to adaptively decompose large identity management problems into much smaller ones, improving previous methods both in scalability and approximation quality.

2 Probability distributions on permutations

We begin with a general discussion of distributions on permutations and their marginals. Any distribution $h(\sigma)$, defined over the symmetric group can be viewed as a joint distribution over the n variables $\sigma = (\sigma_1, \ldots, \sigma_n)$ (where $\sigma_i \in \{1, \ldots, n\}$), subject to *mutual exclusivity constraints* which ensure that objects i and j never map to the same thing ($h(\sigma_i = \sigma_j) = 0$ whenever $i \neq j$). Since there are $n!$ permutations, it is infeasible to represent entire distributions and one can only hope to maintain compact summary statistics instead. Perhaps the most common way to summarize distributions on S_n is to use the *first-order summary*, which stores a marginal distribution over singleton variables (for example, it might store the marginal probability of Alice being in Track 1), and thus requires storing a matrix of only $O(n^2)$ numbers. For example, if $P(\sigma)$ is specified by the following table,

$\sigma([ABC])$	[123]	[213]	[132]	[321]	[231]	[312]
$P(\sigma)$	1/3	1/6	1/3	0	1/6	0

then the first-order matrix of marginal probabilities associated with $P(\sigma)$ is given by:

	Alice	Bob	Cathy
Track 1	2/3	1/6	1/6
Track 2	1/3	1/3	1/3
Track 3	0	1/2	1/2

However, there are more complex marginals that can often provide important information about a distribution on permutations. In this paper, we will be interested in querying the *s^{th}-order marginals*, which are marginal probabilities of s-tuples. The second-order marginals, for example, take the form $P(\sigma : \sigma(k, \ell) = (i, j))$, and in identity management, might jointly capture the joint probability of Alice being in Track 1 and of Bob being in Track 2. While the first-order marginals require $O(n^2)$ storage, second-order marginals require $O(n^4)$ storage. As we discuss in Section 4, low order marginal probabilities correspond, in a certain sense, to the low frequency Fourier coefficients and thus the bandlimited approximations used in (Huang et al., 2007; Kondor et al., 2007) can be thought of as methods for maintaining low-order marginal probabilities.

3 First-order independence conditions

While band-limiting our representation can decrease the storage cost from $O(n!)$ to some polynomial in n, maintaining the s^{th}-order marginals requires, in the worst-case, $O(n^{2s})$ space. Thus, for small n we can maintain higher order coefficients (larger s), but this representation quickly becomes intractable as n becomes large. Over the next sections, we will show how probabilistic independence is manifested in the Fourier coefficients of a distribution, and how, by exploiting this independence, we can break our distribution into smaller subgroups, allowing higher-order coefficients to be maintained. We begin with a simple condition on the matrix of first-order marginal probabilities implied by independence.

Definition 1. Consider any subset $X \subset \{1, \ldots, n\}$ and its complement $\bar{X} \subset \{1, \ldots, n\}$. X and $\bar{X}$ are independent under a distribution $h(\sigma)$ if $h(\sigma)$ factors as the following product of distributions over X and $\bar{X}$: $h(\sigma) = f(\sigma_X) \cdot g(\sigma_{\bar{X}})$

If $X = \{1, \ldots, p\}$, for example, $h(\sigma) = f(\sigma_1, \ldots, \sigma_p)g(\sigma_{p+1}, \ldots, \sigma_n)$. We will refer to X and $\bar{X}$ as *cliques* since the variables of X and $\bar{X}$ form disjoint cliques in the graphical model representation of the above independence relation. In this section, we

discuss a simple first-order criterion for independence on the symmetric group and show that it naturally leads us to study functions over product groups of the form $S_p \times S_q$, where $p + q = n$.

Due to the mutual exclusivity constraints associated with permutations, a necessary (but insufficient) condition for a distribution on permutations h to factor into a product of factors over X and $\bar{X}$, respectively, is that there must exist a subset $Y \subset \{1, \ldots, n\}$ of the same size as X such that, with probability 1, elements of X map to Y and elements of $\bar{X}$ map to $\bar{Y}$. We will refer to the above condition as the *first-order independence criterion*. Intuitively, a distribution can only factor into independent parts if the set $\{1, \ldots, n\}$ can be partitioned into disjoint subsets of objects which do not interact with one another. See Figures 1(a) and 1(c) for example.

Lemma 2 (first-order independence criterion)**.** *If σ_X and $\sigma_{\bar{X}}$ are independent under the distribution $h(\sigma)$ (i.e., $h(\sigma) = f(\sigma_X) \cdot g(\sigma_{\bar{X}})$), then there exists a subset $Y \subset \{1, \ldots, n\}$ with $|Y| = |X|$ such that $h(\sigma) = 0$ unless $\sigma_X \subset Y$.*

To see why the first-order independence criterion is an insufficient indicator of independence, consider the simple example of a distribution on S_4 which always maps the set $X = \{1, 2\}$ to $Y = \{1, 2\}$ and the set $\bar{X} = \{3, 4\}$ to $\bar{Y} = \{3, 4\}$, but is constrained to map 1 to 1 whenever 3 maps to 3. In this case, the 1^{st}-order marginals exhibit independence, but the distribution is not independent when we examine the higher order components. Despite its insufficiency however, the first-order independence plays a crucial role for us in several ways. As we discuss later, it can serve as a first pass at detecting independence as it reduces the detection problem into a clustering-like problem. But on a somewhat more theoretical level, it also suggests that we should be thinking about groups of the form $S_p \times S_q \subset S_n$, where $|X| = |Y| = p$ and $|\bar{X}| = |\bar{Y}| = n - p = q$, which we will later consider in order to derive our Join and Split algorithms.

4 Fourier domain inference

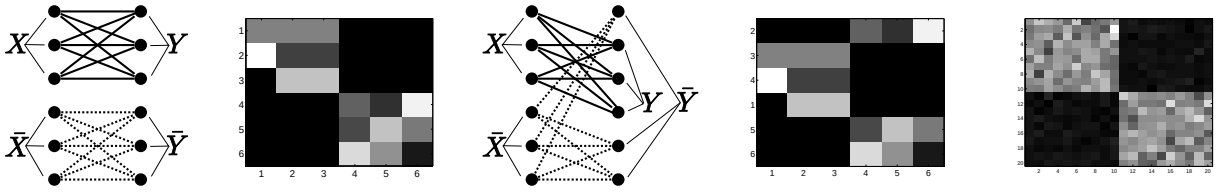
While the first-order condition described in Section 3 is fairly intuitive, understanding probabilistic independence at higher order marginals is considerably more complicated. In this paper, we discuss probabilistic independence at these higher order marginals, and despite the somewhat intimidating math, the algorithms we provide in this section (Section 5.1) are quite simple. We begin by providing a brief overview of several necessary concepts from Fourier analysis over the symmetric group and motivate the idea that "low-frequency" Fourier coefficients of a distribution on the symmetric group can be used to construct low-order marginal probabilities. See (Huang et al., 2008; Diaconis, 1988) for details.

Marginals and partitions. One of the key insights behind the Fourier-based methods, is that the first-order summaries can be constructed using a "low-frequency" subset of the Fourier coefficients of a distribution over S_n, and moreover, the appropriate generalization to "higher-frequency" coefficients allows one to capture more complicated marginals. From the Fourier theoretic view, the first-order marginals are lower frequency than the second-order marginals, which are, in turn, lower frequency than the third-order marginals, and so on. In the remainder of the paper, we will identify each type of marginal with some unique *partition* of n (which is defined to be an unordered tuple of positive integers $\lambda = (\lambda_1, \ldots, \lambda_\ell)$, which summing to n). In particular, we will say that the s^{th}-order marginal probabilities are of type $\lambda = (n - s, 1, \ldots, 1)$, where there are s trailing 1's. Thus $\lambda = (n - 1, 1)$ refers to the first-order marginals, while $(n - 2, 1, 1)$ refers to the second-order marginals. General partitions (not of the form $\lambda = (n - s, 1, \ldots, 1)$) can in fact also be thought of as marginals. For example, the partition $\lambda = (n - 2, 2)$ refers to marginals of unordered pairs: $P(\sigma : \sigma(\{k, \ell\}) \rightarrow \{i, j\})$, e.g., the probability that Alice and Bob occupy tracks 1 and 2, in either order. For simplicity, however, we will focus our discussion on the marginals of type $\lambda = (n - s, 1, \ldots, 1)$, but our results generalize to other partitions.

Fourier basis functions on groups. As it turns out, each partition will be associated with its own set of Fourier basis functions on the symmetric group. To make the connection more precise, we will first discuss Fourier basis functions on general finite groups. We start by defining a special class of functions on a group, called *representations*, which form a superset of the Fourier basis functions.

Definition 3. A *representation* of a group G is a map ρ from G to invertible $d_\rho \times d_\rho$ matrices such that for all $\sigma_1, \sigma_2 \in G$, $\rho(\sigma_1 \sigma_2) = \rho(\sigma_1) \cdot \rho(\sigma_2)$, where $\sigma_1 \sigma_2$ refers to the composition of σ_1 and σ_2. We refer to d_ρ as the *degree* of the representation.

The requirement that $\rho(\sigma_1 \sigma_2) = \rho(\sigma_1) \cdot \rho(\sigma_2)$ is analogous to the familiar property from discrete Fourier transforms that $e^{i(\theta_1 + \theta_2)} = e^{i\theta_1} \cdot e^{i\theta_2}$. For noncommutative groups, the difference is that representations are matrix-valued, and so a $d \times d$ representation can also be thought of as a collection of d^2 functions at once. The simplest example of a representation is the function $\rho_{(n)} : S_n \rightarrow \mathbb{R}^{1 \times 1}$ which maps every permutation to 1. As a less trivial example, we define the *first-order permutation representation* of S_n to be the degree n representation, $\tau_{(n-1,1)}$, which maps a permutation σ to its corresponding permutation



(a) (b) (c) (d) (e)

Figure 1: Examples of first-order independence where two groups (of three people each) interact within their own groups but not with the other group. In (a) and (c), we show how the identities and tracks can be partitioned into disjoint subsets, X, $\bar{X}$, Y, and $\bar{Y}$. (b) and (d), show an example of what the corresponding first-order marginals would look like. In practice, we expect first-order independence to only hold approximately, as in (e).

matrix given by: $[\tau_{(n-1,1)}(\sigma)]_{ij} = \mathbb{1}\{\sigma(j) = i\}$.

Though the d^2 functions encoded by a representation are, in general, neither linearly independent nor complete, there always exists a unique finite set of representations, $\rho_1, \ldots, \rho_m$ which *do* form a complete orthogonal basis for functions on G. These distinguished representations are known as the *irreducibles* of G, and one of the main results of representation theory, which we will use later, is that *any* representation can, in a sense, be "built" using only irreducibles.

Theorem 4 (Maschke (Diaconis, 1988))**.** *Given a group representation ρ, there exists a $d \times d$ orthogonal matrix C and multiplicities z_λ such that for any $\sigma \in G$: $C \cdot \rho(\sigma) \cdot C^T = \bigoplus_k \bigoplus_{\ell=1}^{z_k} \rho_k(\sigma)$, where $\bigoplus$ denotes the matrix direct sum operation, and ρ_k are irreducible representations.*

We can now simply define the Fourier transform of a function f to be the collection of dot products of f against each irreducible:

Definition 5. Let f be any real-valued function on a group G and let ρ_k be any irreducible representation on G. The *Fourier Transform of f at ρ_k* is defined to be: $\hat{f}_{\rho_k} = \sum_\sigma f(\sigma)\rho_k(\sigma)$. (Note that $\hat{f}_{\rho_k}$ is a $d_{\rho_k} \times d_{\rho_k}$ matrix).

Fourier transforms on S_n. While the problem of determining the irreducibles of a given group can be a highly nontrivial problem in general, the irreducibles of S_n are well understood. On S_n, the set of irreducibles is indexed by the partitions, λ, of n, reflecting the intuition that each irreducible gives information corresponding to some unique type of marginal. On S_3, for example, there are three irreducibles, $\rho_{(3)}$, $\rho_{(2,1)}$, and $\rho_{(1,1,1)}$. However, marginals and Fourier coefficients are not *exactly* the same — we cannot construct marginals of type λ directly from $\hat{f}_\lambda$. In addition to $\hat{f}_\lambda$, it is necessary to also know $\hat{f}_\mu$ at partitions μ which are at a "lower frequency" than λ, in order to obtain marginals of type λ. But what does it mean for an irreducible to be low frequency? To answer this question, we define the *dominance ordering* on partitions.

Definition 6. Let λ, μ be partitions of n. Then $\lambda \trianglerighteq \mu$ (we say λ *dominates* μ), if for each i, $\sum_{k=1}^{i} \lambda_k \geq \sum_{k=1}^{i} \mu_k$.

The dominance ordering imposes a partial (rather than linear) order on partitions. For example $(4,2) \trianglerighteq (3,2,1)$ since $4 \geq 3$, $4+2 \geq 3+2$, and $4+2+0 \geq 3+2+1$. However, $(3,3)$ and $(4,1,1)$ are incomparable since $3 \leq 4$, but $3+3 \geq 4+1$. The matrix of marginals of type λ^{MIN} for a distribution $f(\sigma)$ can always be reconstructed using $\hat{f}_{\lambda^{MIN}}$ and a collection of Fourier coefficients which are "lower frequency" (or higher in the dominance ordering) than λ^{MIN}.

Theorem 7 (See (Huang et al., 2008; Diaconis, 1988))**.** *Marginals of type λ^{MIN} of a distribution $f(\sigma)$ can be reconstructed using only the set of Fourier coefficients $\{\hat{f}_\lambda : \lambda \trianglerighteq \lambda^{MIN}\}$.*

For example, reconstructing the first order marginals requires knowledge of Fourier coefficients corresponding to (n) and $(n-1,1)$, while second order ordered marginals require Fourier coefficients at (n), $(n-1,1)$, $(n-2,2)$, and $(n-2,1,1)$.

Efficient inference in the Fourier domain. The Fourier-theoretic approach is appealing because it offers a unified and principled framework for approximations by allowing one to *bandlimit*, or discard high frequency Fourier terms. Additionally, Fourier coefficients lend themselves to natural reformulations of standard inference algorithms. Huang et al. (Huang et al., 2007) present general algorithms for performing inference in the Fourier domain for hidden Markov models. In a nutshell, they show that the prediction/rollup step of the Forward algorithm can be written as a convolution and can be performed in the Fourier domain as a pointwise product of coefficients. On the other hand, conditioning can be written as a pointwise product and can therefore be performed in the Fourier domain as a (generalized) convolution of Fourier coefficients.

5 Probabilistic independence in the Fourier domain

In this section, we return to probabilistic independence and generalize the results of Section 3 to hold for higher-order Fourier terms. Since we maintain the Fourier coefficients of the distribution instead of the actual distribution, there are several technical challenges associated with (1) splitting a distribution

into independent factors, (2) joining independent factors to form a joint distribution, and (3) detecting independent subsets in the Fourier domain. To solve these problems, we will now relate the Fourier coefficients of the joint distribution h with the Fourier coefficients of the factors f and g. We first focus on the special case when $X = Y = \{1, \ldots, p\}$ (we show how to deal with general X and Y when we discuss the detection step). Thus, we assume that:

$$h(\sigma) = \begin{cases} f(\sigma_p)g(\sigma_q) & \text{if } \sigma_p \subset \{1,\ldots,p\}, \text{ (and } \\ & \quad \sigma_q \subset \{p+1,\ldots,n\}) \\ 0 & \text{otherwise} \end{cases}, \tag{5.1}$$

where $\sigma_p = (\sigma_1, \ldots, \sigma_p)$, $\sigma_q = (\sigma_{p+1}, \ldots, \sigma_n)$, with $p + q = n$.

While the composite permutation $\sigma = [\sigma_p\ \sigma_q]$ can be seen as an element of S_n, it can additionally be viewed as an element of the subgroup $S_p \times S_q$ whenever $h(\sigma) \neq 0$, where $S_p \times S_q$ is the subgroup of permutations which map the set $\{1, \ldots, p\}$ into $\{1, \ldots, p\}$ (and thus also map $\{p+1, \ldots, n\}$ into $\{p+1, \ldots, n\}$). Evaluated at an element $\sigma \in S_p \times S_q$, any irreducible, ρ_λ of S_n, can also be viewed as a representation of the group $S_p \times S_q$. As a representation of $S_p \times S_q$, however, ρ_λ is not necessarily reducible, but it can be related to the irreducibles of $S_p \times S_q$ using Maschke's theorem (Theorem 4), as we show in the following.

But what are the irreducibles of $S_p \times S_q$? We use a standard representation theoretic result that the set of irreducibles of a direct product of two groups $H \times K$ is exactly the set of all pairwise tensor products of irreducibles of H and K. Thus the set of irreducibles of $S_p \times S_q$ is the set: $\{\rho_\mu \otimes \rho_\nu\}$, where μ and ν range over partitions of p and q, respectively. A representation ρ_λ, of S_n, when evaluated at a permutation σ which lies in the subgroup $S_p \times S_q$, therefore has the following decomposition by Maschke's Theorem (Theorem 4):

$$L^\lambda_{\mu\nu} \cdot \rho_\lambda(\sigma) \cdot {L^\lambda_{\mu\nu}}^T = \bigoplus_{\mu,\nu} \bigoplus_{\ell=1}^{c^\lambda_{\mu,\nu}} \rho_\mu(\sigma_p) \otimes \rho_\nu(\sigma_q). \tag{5.2}$$

The coupling matrix $L^\lambda_{\mu\nu}$, along with the multiplicities $c^\lambda_{\mu,\nu}$ are assumed to be precomputed (see (Sagan, 2001; Huang et al., 2008)). The following Proposition gives the desired relation between the Fourier coefficients of the joint and the Fourier coefficients of the factors.

Proposition 8. *Given Fourier coefficients of two independent factors f and g, the Fourier coefficient matrices of the joint distribution h, are:*

$$\widehat{h}_\lambda = \left[\widehat{f \cdot g}\right]_\lambda = {L^\lambda_{\mu\nu}}^T \cdot \bigoplus_{\mu,\nu} \bigoplus_{\ell=1}^{c^\lambda_{\mu,\nu}} \left(\widehat{f}_\mu \otimes \widehat{g}_\nu\right) \cdot L^\lambda_{\mu\nu}. \tag{5.3}$$

We remark that the independence assumption in Prop. 8 is *necessary* and Eqn. 5.3 does *not* hold for arbitrary functions even if they are zero outside of $S_p \times S_q$. As promised, Eqn. 5.3 characterizes the form of the Fourier matrices of the joint distribution at all frequencies. Recalling that (Lemma 2) that the first-order marginals are constrained to be block diagonal, we see (ignoring the change of basis) that Eqn. 5.3 in fact imposes block diagonal structure on the Fourier matrices at all orders. Additionally, we see that each nonzero block has Kronecker structure at higher orders and that the coefficients of the joint are redundant in the sense that information at lower frequencies of the factors f and g are duplicated to multiple higher frequencies of h. As it turns out, the multiplicities, $c^\lambda_{\mu,\nu}$, are equivalent to what mathematicians have studied in different contexts as *Littlewood-Richardson (LR)* coefficients. The LR coefficients tell us which crossterms contribute to the joint. For example, it can be shown that that first-order terms corresponding to the partition $(n-1, 1)$ can be reconstructed using only three terms, $(p) \otimes (q)$, $(p-1, 1) \otimes (q)$, and $(p) \otimes (q-1, 1)$. Computing the LR coefficients has been shown, in general, to be a $\#P$-complete problem (Narayanan, 2006). For low-order Fourier terms (corresponding to partitions with only a few rows), however, the *Littlewood-Richardson rule* ((Sagan, 2001)) computes the LR coefficients in reasonable time.Due to space constraints, we refer the reader to (Sagan, 2001) for a discussion of the Littlewood-Richardson rule. While the LR coefficients have been studied in various mathematical contexts,this paper provides, to the best of our knowledge, the first connection to probabilistic independence.

5.1 Algorithms and Approximation

We now discuss algorithms for merging independent factors to form a joint (*Join*), and for extracting independent factors from a joint (*Split*) based on our Fourier domain factorization (Proposition 8). There are two problems that one encounters in practice; first, it is impossible to maintain a complete set of Fourier coefficients, and second, it is rare for distributions to factor completely. We present novel theoretical results in this section addressing both issues and show that our algorithms behave reasonably in bandlimited and near-independent (rather than fully-independent) settings.

Join. The simplest operation of the two is the *Join* algorithm — which is a straightforward implementation of Equation 5.3. Given the Fourier transforms $\hat{f}$ and $\hat{g}$, the Fourier transform of the joint, $\hat{h}$, can be constructed by forming the direct sum of appropriate tensor product terms $\hat{f}_\mu \otimes \hat{g}_\nu$, and conjugating by the precomputed coupling matrix $L^\lambda_{\mu\nu}$. The complexity of the Join operation is dominated by the cost of matrix multiplication ($O(d^3_\rho)$ for each representation ρ), and is therefore no more expensive than the convolution operations from (Huang et al., 2007).

One might worry that we would require maintaining high-frequency terms of the independent factors in order to construct low frequency terms of the joint. We show, using the Littlewood-Richardson rule, that this is not the case when we maintain s^{th}-order marginals. For any integer s such that $0 \leq s < n$, define the following partitions:

$$\lambda^{MIN} = (n-s, \underbrace{1, \ldots, 1}_{s \text{ times}}), \qquad \mu^{MIN} = (p-k, \underbrace{1, \ldots, 1}_{k \text{ times}}),$$

$$\nu^{MIN} = (q-\ell, \underbrace{1, \ldots, 1}_{\ell \text{ times}}),$$

where $k = \min(s, p-1)$ and $\ell = \min(s, q-1)$. We have the following guarantee.

Theorem 9. *Given marginals of type μ^{MIN} for f and of type ν^{MIN} for g,* Join *returns Fourier coefficients of the joint distribution h which can reconstruct marginals of type λ^{MIN}* exactly.

Theorem 9 formalizes the intuitive idea that it is possible, using the Join algorithm, to exactly construct s^{th}-order marginals of the joint distribution using only the s^{th}-order marginals of each independent factor. The proof of Theorem 9 is given in the appendix. A more general principle holds for other partitions which do not take the form $\lambda^{MIN} = (n-s, 1, \ldots, 1)$,but we will focus on the simpler and more intuitive case of s^{th}-order marginals.

Split. Given the Fourier transform of the joint, $\hat{h}$, we wish to formulate an algorithm which computes the Fourier coefficients of the factors, $\hat{f}$ and $\hat{g}$, assuming that the sets $X = \{1, \ldots, p\}$ and $\bar{X} = \{p+1, \ldots, n\}$ are independent under h. One can imagine "inverting" the Join algorithm by computing $L^{\lambda}_{\mu\nu} \cdot \hat{h}_{\lambda} \cdot {L^{\lambda}_{\mu\nu}}^{T}$ and reading off the $\hat{f}_{\mu}$ and $\hat{g}_{\nu}$ from the resulting matrix, $\bigoplus_{\mu,\nu} \bigoplus_{\ell=1}^{c^{\lambda}_{\mu\nu}} \hat{f}_{\mu} \otimes \hat{g}_{\nu}$. The difficulty is that the matrices $\hat{f}_{\mu} \otimes \hat{g}_{\nu}$, in general, only determine $\hat{f}_{\mu}$ and $\hat{g}_{\nu}$ up to a scaling factor, and in the approximate case when X and $\bar{X}$ are only "nearly" independent, the appropriate blocks of the matrix $L^{\lambda}_{\mu\nu} \cdot \hat{h}_{\lambda} \cdot {L^{\lambda}_{\mu\nu}}^{T}$ do not take the form $A \otimes B$.

Happily though, we are in fact able to always construct coefficients of $\hat{f}$ and $\hat{g}$ using *only* blocks of the form $\hat{f}_{\mu} \otimes 1$, or $1 \otimes \hat{g}_{\nu}$, allowing us to literally read off the matrices for $\hat{f}_{\mu}$ and $\hat{g}_{\nu}$.

Theorem 10. *For any $\mu \trianglerighteq \mu^{MIN}$, there exists a block of $L^{\lambda}_{\mu\nu} \cdot \hat{h}_{\lambda} \cdot {L^{\lambda}_{\mu\nu}}^{T}$ for some $\lambda \trianglerighteq \lambda^{MIN}$ which is identically equal to $\hat{f}_{\mu}$.*

Likewise, for any $\nu \trianglerighteq \nu^{MIN}$, there exists a block of $L^{\lambda}_{\mu\nu} \cdot \hat{h}_{\lambda} \cdot {L^{\lambda}_{\mu\nu}}^{T}$ for some $\lambda \trianglerighteq \lambda^{MIN}$ which is identically equal to $\hat{g}_{\nu}$. See Algorithm 1 for pseudocode for the Split algorithm and the appendix for more details. As a corollary, we obtain a converse to Theorem 9 which says that given the s^{th}-order marginals of the joint, we will be able to recover the s^{th}-order marginals of the factors.

Algorithm 1: Pseudocode for the *Split* algorithm.

```
SPLIT
foreach partition μ of p such that μ ⊵ μ^MIN do
    λ ← (μ_1 + n − p, μ_2, ...) ;
    f̂_μ ← (μ,(q))-block of the matrix L^λ_μν · ĥ_λ · (L^λ_μν)^T ;
foreach partition ν of q such that ν ⊵ ν^MIN do
    λ ← (ν_1 + n − q, ν_2, ...) ;
    ĝ_ν ← ((p),ν)-block of the matrix L^λ_μν · ĥ_λ · (L^λ_μν)^T ;
Normalize f̂ and ĝ;
```

Corollary 11. *Given marginals of type λ^{MIN} of the joint h,* Split *returns Fourier coefficients of the factors f and g which can be used to exactly reconstruct marginals of type μ^{MIN} and ν^{MIN}, respectively.*

Near-independence. Although exploiting independence can significantly reduce computation, it is rare for full independence to hold in practice (Figure 1(e), for example). Consider calling the Split algorithm on a distribution which does *not* factor into distributions on S_p and S_q. Ideally, one would, in this case, hope to obtain the Fourier transform of the appropriate marginal distributions of $(1, \ldots, p)$ and $(p+1, \ldots, n)$. We now show that this is *almost* the case and that we *do* recover exact marginals. However, due to the fact that the Split algorithm effectively ignores mass outside of the subgroup $S_p \times S_q$, we are only able to accurately reconstruct marginals that could have been computed using mass concentrated in $S_p \times S_q$. For example, if $n = 6$ and $p = 3$, then the result of Split can reconstruct $P(\sigma : \sigma(1,2) = (2,3))$ but not $P(\sigma : \sigma(1,2) = (2,5))$ since the permutation $(2,5) \notin S_3 \times S_3 \subset S_6$.

Theorem 12. *Given marginals of type λ^{MIN} for an arbitrary distribution h, the output of Split can be used to reconstruct the subset of marginals of type μ^{MIN} and ν^{MIN} which can be computed using only elements of $S_p \times S_q$, for $(1, \ldots, p)$ and $(p+1, \ldots, n)$ respectively,*

Corollary 13. *Whenever first-order independence conditions hold for a distribution h, the output of Split can be used to exactly reconstruct* all marginals *of h.*

When first-order independence does not hold, the resulting coefficients do not correspond to a properly normalized distribution, and in particular, $\hat{f}_{(p)}$ is the amount of mass assigned to elements of $S_p \times S_q$ (instead of 1). However, since $\hat{f}$ still corresponds to a positive function, one can easily normalize $\hat{f}$ by dividing all coefficients by the zeroth-order Fourier coefficient, $\hat{f}_{(p)}$, without requiring a projection to the marginal polytope as in (Huang et al., 2007). Thus when a distribution is near first-order independent, we recover approximate marginals.

Detecting and measuring independence. We now discard the assumption that $X = Y = \{1, \ldots, p\}$ and deal with the problem of explicitly finding sets X and Y such that $h(\sigma_X \subset Y) = 1$ and $h(\sigma_{\bar{X}} \subset \bar{Y}) = 1$ as in the first-order independence criterion. We begin with the simple observation that, if we *knew* the sets X and Y, then the first-order matrix of marginals would be rendered block diagonal under an appropriate reordering of the rows and columns (Figure 1(b)). Since X and Y are unknown, our task is to find permutations of the rows and columns of the first-order matrix of marginals (Figure 1(d)) to obtain a block diagonal matrix. Viewing the matrix of first-order marginals as a set of edge weights on a bipartite graph between tracks and identities, we approach the detection step as a *biclustering* problem (in which one simultaneously clusters the tracks and identities) with an extra *balance* constraint forcing $|X| = |Y|$. In our experiments, we use a cubic time SVD-based technique presented in (Zha et al., 2001) which finds bipartite graph partitions optimizing the normalized cut measure modified to satisfy the balance constraint.

Assuming now that we have obtained the sets X and Y via the above clustering step, we can call the Split algorithm by first renaming the tracks and identities so that $X = Y = \{1, \ldots, p\}$. Suppose that, to achieve this reordering, we must permute the X (people) using a permutation π_1 and the Y (tracks) using π_2. The *Shift Theorem* can be applied to reorder the Fourier coefficients according to these new labels, and we can then apply Algorithm 1 unchanged.

Proposition 14 (Shift Theorem (Diaconis, 1988))**.** *Given $f : S_n \to \mathbb{R}$, define $f' : S_n \to \mathbb{R}$ by $f'(\sigma) = f(\pi_1 \sigma \pi_2)$ for some fixed $\pi_1, \pi_2 \in S_n$. The Fourier transforms of f and f' are related as: $\widehat{f'}_\lambda = \rho_\lambda(\pi_1) \cdot \widehat{f}_\lambda \cdot \rho_\lambda(\pi_2)$.*

We have focused on detecting independence in the first-order sense. As discussed in Section 3, first-order independence is necessary, but insufficient for higher order independence. However, as we showed in Section 5.1, we can approximately recover marginal probabilities when a distribution is near first-order independent. Furthermore, our biclustering approach can also be viewed as a first pass for proposing candidate splits. Once this factoring is performed, we can measure its effect on higher orders, e.g., using the Plancherel Theorem (Diaconis, 1988) to measure the distance between the original coefficients and the factored result, and decide whether or not to retain the partition.

6 Example: an adaptive solution for identity management

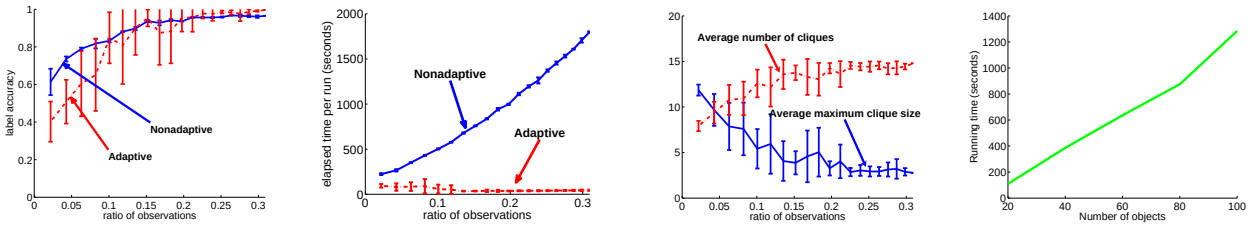
As an application, we apply our algorithms in the identity management setting. In both (Huang et al., 2007) and (Kondor et al., 2007), one reasons *jointly* over assignments of all n tracks to all n identities. In realistic settings however, we believe that it is often sufficient to only reason over small cliques of tracks at a time. Thus instead of maintaining Fourier coefficients over all of S_n, we search for independent cliques and *adaptively* split the distribution into factors over smaller cliques whenever possible.

In our adaptive approach, we maintain a collection of disjoint cliques over the tracks and identities. After conditioning on any observation, we attempt to split. We also force splits whenever cliques grow to be too large to handle. Upon splitting, we allow the representational size to grow to higher orders — thus for very large n, we might only maintain first-order coefficients, but for smaller sized cliques, we might choose to represent higher-order coefficients. Finally, whenever mixing events occur between tracks belonging to distinct cliques, we merge the cliques using our Join algorithm and perform a mixing on the newly formed joint distribution.

7 Experiments

We evaluted our adaptive identity management algorithm on a biotracking dataset from (Khan et al., 2006). In their data, there are 20 ants (Fig. 7) moving in an enclosed area. The data is interesting for our purposes since it is a relatively large n compared to many multiobject tracking datasets with interesting movement patterns and plenty of mixing events (which we log whenever ants walk within some distance of each other). At each timestep, we allow each ant to 'reveal' its identity with some probability (in our experiments, ranging from $p_{obs} = .005$ to $p_{obs} = .05$ per timeframe), and our task is to jointly label all tracks with identities for all timeframes. Due to the fact that our experimental setup is quite different from (Khan et al., 2006) and that we do not consider positional uncertainty, we are not able to compare with their results. Instead, we compare with the nonadaptive algorithm from (Huang et al., 2007). We measure accuracy using the fraction of correctly labeled tracks over the entire sequence (note that the accuracy of random guessing is $1/n = 5\%$ in expectation). As a splitting criterion, we decide to to split if, after clustering, the sum over all off-block elements fall below a certain threshold ϵ (in all experiments, we fixed $\epsilon = 1/(2n)$).

In Figures 2(a) and 2(b), we compare the performance of an adaptive approach against the nonadaptive algorithm from (Huang et al., 2007) as we vary the ratio of observations to mixing events. Figure 2(a) shows that the two algorithms perform similarly in accuracy, with the nonadaptive approach faring slightly better with fewer observations (due to more diffuse distributions) and slightly worse with more observations (due to the



(a) Accuracy comparison (b) Running times (c) Cliquesizes (d) Running times vs. n

Figure 2: Experimental results on biotracking data

fact that the adaptive approach can represent higher-order Fourier terms). The real advantage of our adaptive approach is shown in Figure 2(b) which plots a running time comparison. Since the conditioning step is the complexity bottleneck of performing inference in the Fourier domain, the running time scales according to the proportion observations. However, since the adaptive algorithm typically conditions smaller cliques on average (especially with more observations), we see that it is a far more scalable algorithm. In Figure 2(c), we plot the average number of cliques and sizes of cliques which were formed in the same experiment. As expected, we see that the cliques get smaller and more numerous as the number of observations grows.

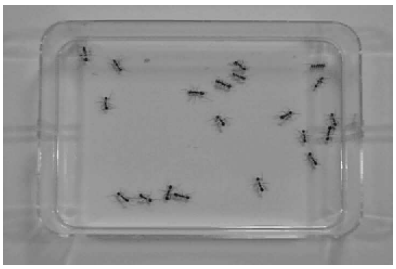


Figure 3: Sample image from biotracking data

Finally, we simulated larger tracking problems by taking m different segments of the ant data and tracking $m \cdot n$ ants at the same time allowing for ants to 'teleport' to other segments with some probability. Figure 2(d) shows a comparison of average running time for these larger problems. Note that at such sizes, we can no longer feasibly run the original nonadaptive algorithms from Huang et al. (2007); Kondor et al. (2007).

8 Conclusion

A pervasive technique in machine learning for making large problems tractable is to exploit independence structures for decomposing large problems into much smaller ones. It is the structure of (conditional) independence, for example, which has made Bayes net and Markov random field representations so powerful. In this paper, we have contributed to the existing collection of efficient Fourier-theoretic inference operations by presenting a formulation of probabilistic independence for permutations based on the Littlewood-Richardson decomposition. While such decompositions have been used in mathematics, we are the first to use them in the context of probabilistic independence and to consider their bandlimiting properties. Combined with the bandlimited inference algorithms from (Huang et al., 2007; Kondor et al., 2007), we believe that our algorithms will contribute to making these Fourier methods highly scalable and practical.

Finally we view our contributions as a first step towards understanding and exploiting more intermediate notions which lie somewhere between full independence and fully connected, such as conditional or context-specific independence which have proven themselves to be indispensible in the fields of machine learning and AI.

Acknowledgements

This work is supported by the ONR under MURI N000140710747, the ARL under grants W911NF-06-1-0275, W911NF-07-2-0027, the NSF under grants CCF-0634803, DGE-0333420, EEEC-540865, NeTS-NOSS 0626151, TF-0634803, and by the Pennsylvania Infrastructure Technology Alliance (PITA). Carlos Guestrin was also supported in part by an Alfred P. Sloan Fellowship.

References

J. Huang, C. Guestrin, and L. Guibas. Efficient inference for distributions on permutations. In *NIPS 2007*, 2007.

G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. In *NIPS 2007*.

R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *AISTATS*, 2007.

J. Huang, C. Guestrin, and L. Guibas. Inference for distributions over the permutation group. Technical Report CMU-ML-08-108, Machine Learning Department, Carnegie Mellon University, May 2008.

P. Diaconis. *Group Representations in Probability and Statistics*. IMS Lecture Notes, 1988.

B. Sagan. *The Symmetric Group*. Springer, 2001.

H. Narayanan. On the complexity of computing kostka numbers and littlewood-richardson coefficients. *J. Algebraic Comb.*, 24(3):347–354, 2006.

H. Zha, X. He, C. Ding, H. Simon, and M. Gu. Bipartite graph partitioning and data clustering. In *CIKM '01*, pages 25–32, New York, NY, USA, 2001. ACM.

Z. Khan, T. Balch, and F. Dellaert. Mcmc data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. *IEEE PAMI*, 28(12), 2006.