Multi-object tracking with representations of the symmetric group

Risi Kondor, Andrew Howard and Tony Jebara
{risi,ahoward,jebara}@cs.columbia.edu
Computer Science Department, Columbia University,
1214 Amsterdam Ave., New York, NY10027, USA

Abstract

We present an efficient algorithm for approximately maintaining and updating a distribution over permutations matching tracks to real world objects. The algorithm hinges on two insights from the theory of harmonic analysis on noncommutative groups. The first is that most of the information in the distribution over permutations is captured by certain "low frequency" Fourier components. The second is that Bayesian updates of these components can be efficiently realized by extensions of Clausen's FFT for the symmetric group.

1 Introduction

Multi-object trackers associate tracks $r_1, r_2, \ldots, r_n$ with real world objects $o_1, o_2, \ldots, o_n$ called targets. When the targets are well separated and good quality observations are available, following which track corresponds to which target is relatively easy. However, when two objects come very close, are occluded, or observations are not available, the association between tracks and objects becomes uncertain. This is what is referred to as the data association problem in multi-object tracking.

Most tracking systems in practical use today are probabilistic, in the sense that at time t, each $r_i(t)$ stands for a probability distribution describing our belief of where the corresponding object is likely to be in space. It is then natural to address the data association problem in a similarly probabiblistic fashion, by maintaining a probability distribution over the $n!$ possible matchings between objects and tracks. Equivalently, we regard this as a probability distribution $p(\sigma)$ over permutations of $o_1, o_2, \ldots, o_n$. Missing observations, occlusion, random noise, etc. tend to smear out $p(\sigma)$, while positively observing some target at a particular track makes the distribution more concentrated again. The central challenge in maintaining p is the factorial blow-up in the number of permutations.

Multi-object tracking has many applications in vision, security, control, observational problems in the natural sciences and even in sports [1][16]. The method we propose is particularly well suited to problems where location information is relatively abundant, but identity information is scarce. A typical example is air traffic control, where aircraft are more or less continuously visible on radar, but in the absence of transponders, their identity is only revelaled when the pilot reports by radio. A similar application is tracking animals released into the wild. Information about the animals might include sightings, or indirect evidence for their presence, such as killed prey, but only a small fraction of these observerations reveal the identity of the specific animal. The task is to extrapolate to the rest of the observations and reconstruct the paths of individual animals. The data association problem is also critical in sensor networks, such as "wired" office buildings. Observations come from security cameras, motion detectors, or even light switches, but identifying an individual is only possible when he/she passes by specific and potentially more intrusive sensors, such as RFID transmitters, or an identity card reader.

Sensor networks are the motivating application for recent papers on the information-form data association filter [12][14], which is perhaps closest in spirit to our current work. In particular, the combination of identity observations and a continuous noise process operating in the background is similar to our framework. However, in a certain sense the two approaches are complementary: our experiments indicate that the data association filter performs better when the association distribution is concentrated on a small set of permutations, while our method, based on smoothness arguments, performs better when the distribution is more vague. The data association filter scales very favorably with n, but its representational power is limited to an n^2 dimensional space. We can choose from a whole spectrum of representations at different levels of complexity, but richer representations scale more sharply with n and limit us to about $n \leq 30$. It is possible that in the future the two approaches can be fused, combining their respective strengths.

A continuum of problems

In general, what approximation is most appropriate for a given type of data association problem depends on a number of factors. When there is good reason to believe that the distribution over permutations is concentrated in the vicinity of a single permutation, it is best to keeep track of the likely permutations individually. This type of multiple hypothesis or particle filter type model is very simple and likely to perform better than any true probabilistic model provided that some process continually constrains the number of likely hypotheses to a very small number (e.g., [9]). In this regime there is little use for statistics or machine learning in the classical sense. However, as soon as a general mixing process sets in between the permutations, as in the examples investigated in the experiments section, because of the very strong mixing property of the symmetric group, the number of particles required blows up.

Another consideration is the form in which observations are made, and in which information is to be extracted from the distribution. In many cases these are both related to marginals of the distribution of the type $p\,[\sigma(i)=j]$, i.e., the probability of object i being at track j. It is then tempting to approach the data association problem from the point of view of maintaining an $n \times n$ matrix of these marginal probabilities. Many approaches, including, in some sense, [12] are based on this idea. In fact, in the most basic case of maintaining just two Fourier components, our method also performs something very similar. The difficulty here is that the marginal probabilities are not independent, and its very difficult to maintain them in a coherent state or read off a probability distribution from the marginals directly.

Finally, there are considerations of what it means for a probability distribution over permutations to be smooth, and whether the association distribution is smooth in any sense. Our spectral method, extending ideas from real harmonic analysis, implicitly makes very specific assumptions about the true distribution. As we show, certain natural processes, such as random mixing of targets (diffusion) fit naturally into this picture. Other features of real world problems, such as "hard" observations of object i being at track j with probability close to 1, blatantly violate the smoothness assumption. Just as in classical Fourier analysis, this gives rise to ringing and may lead to artifacts. We empirically found that in many cases this does not hurt performance becuase the form we extract information in is in the form of marginals, which averages over such fluctuations.

We do not claim to have all the answers to these questions, but our empirical results do seem to point us in a certain direction. In general, because of the factorial blow-up in the number of permutations, maintaining a distributions over permutations is a hard problem, and there is no one satisfactory solution at this time.

To the best of our knowledge, our approach based on non-commutative harmonic analysis is novel, certainly in the data association field. Abstract harmonic analysis itself is a well-developed branch of mathematics, but it has seldom been applied to solving real world problems. P. Diaconis wrote a very influential book on the subject sketching out numerous possible application [4], but did not address tracking.

The other inspiration for our work is the recent development of fast transforms for non-commutative groups by Clausen, Rockmore, Maslen and others [8][7]. Without these computational methods and our extensions of them tailored to multi-object tracking, spectral methods could not scale up to the number of particles typically encountered in practical applications. A software package implementing these algorithms, together with a more technical description than these pages allow has been made available on the first author's web pages.

2 In search of the canonical projection

A permutation of n objects is a bijective map $\sigma\colon \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$. In the data association problem $\sigma(i)=j$ signifies that target i is associated with track j. Defining the product of two permutations σ_1 and σ_2 as $(\sigma_2\sigma_1)(i)=\sigma_2(\sigma_1(i))$, permutations form a noncommutative group, called the symmetric group of degree n, and denoted $\mathbb{S}_n$.

A probability distribution over permutations is represented by an $n!$-dimensional vector with components $p_\sigma = p(\sigma)$, $\sigma \in \mathbb{S}_n$. For n greater than about 6 or 7, maintaining and updating such a large vector is prohibitively expensive. In this paper we project p from $\mathbb{R}^{n!}$ to an appropriately chosen subspace W. The subspace W has to satisfy a number of criteria:

1. It should be possible to reconstruct p from its projection p_W with minimal loss of information.
2. The projection should be symmetric in the sense that each component p_σ must suffer the same loss when projected to W.

3. It should be possible to apply observation and noise updates directly to p_W, without reconstructing the full distribution p.

The canonical choice of W is suggested by considerations from representation theory and noncommutative harmonic analysis. For background material in these subjects see [13] and [4].

2.1 Harmonic Analysis on Finite Groups

A (complex) representation of a finite group G is a matrix valued function $\rho\colon G \to \mathbb{C}^{d_\rho \times d_\rho}$ which obeys

$$\rho(g_2)\,\rho(g_1) = \rho(g_2 g_1) \qquad \forall\, g_1, g_2 \in G.$$

We say that d_ρ is the dimensionality of the representation. Two representations ρ_1 and ρ_2 of G are called equivalent if there is an invertible matrix T such that

$$T^{-1}\rho_1(g)\,T = \rho_2(g) \qquad \forall\, g \in G.$$

A representation ρ is said to be reducible if for some invertible matrix T it splits in the form

$$T^{-1}\rho(g)\,T = \left(\begin{array}{c|c} \rho_1(g) & 0 \\ \hline 0 & \rho_2(g) \end{array}\right) \qquad \forall\, g \in G$$

into a direct some of smaller representations ρ_1 and ρ_2. If no such matrix exists, ρ is called irreducible. Any complex-valued representation of a finite group G splits into a direct sum of irreducible representations in a unique way, up to equivalence of representations. One of the central problems in representation theory is to find a complete set of inequivalent irreducibles for G. Such a set we denote by $\mathcal{R}$.

Representation theory allows us to extend the concept of Fourier transformation to any finite group. The Fourier transform (FT) of a function $f\colon G \to \mathbb{C}$ is defined as the set of matrices

$$\widehat{f}(\rho) = \sum_{g \in G} f(g)\,\rho(g) \qquad \rho \in \mathcal{R}. \tag{1}$$

Similarly to the ordinary Fourier transform, $\mathfrak{F}\colon f \mapsto \widehat{f}$ is an invertible linear mapping, and it is unitary with respect to the norms

$$\|\, f\,\|^2 = \frac{1}{|\,G\,|} \sum_{g \in G} |\, f(g)\,|^2$$

and

$$\|\, \widehat{f}\,\|^2 = \frac{1}{|\,G\,|} \sum_{\rho \in \mathcal{R}} d_\rho \,\|\, \widehat{f}(\rho)\,\|^2_{\mathrm{Frob}}. \tag{2}$$

The inverse transform is given by

$$f(g) = \frac{1}{|\,G\,|} \sum_{\rho \in \mathcal{R}} d_\rho \,\mathrm{tr}\left[\widehat{f}(\rho)\,\rho(g)\right] \qquad g \in G,$$

where in the square brackets we have matrix multiplication. By unitarity, the map $\Phi_\rho\colon f \mapsto \widehat{f}(\rho)$ can be regarded as a projection of f onto a d_ρ^2-dimensional subspace V_ρ. Our assertion is that just as in classical Fourier analysis, the so-called isotypal components $\Phi_\rho(f)$ capture variation in f at different levels of "complexity" (c.f. [4][10]). To see this more clearly, we look more specifically at the case of the symmetric group.

2.2 The Spectrum of the Cayley Graph

Apart from the identity, the most elementary members of $\mathbb{S}_n$ are the so-called transpositions $\{[i,j]\}$, which are permutations that exchange i and j and leave the rest of $\{1, 2, \ldots, n\}$ fixed. To represent $\mathbb{S}_n$ as a graph, it is natural to regard two permutations σ and σ' as neighbors if they differ by only a transposition. This is called the Cayley graph, and its Laplacian is

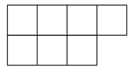
$$\Delta_{\sigma_1,\sigma_2} = \begin{cases} -1 & \begin{array}{l} \text{if } \sigma_1 = [i,j] \cdot \sigma_2 \\ \text{for some } 1 \le i < j \le n, \end{array} \\ n(n-1)/2 & \text{if } \ \sigma_1 = \sigma_2, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

According to general spectral theory, the orthonormal eigenvectors $v_1, v_2, \ldots, v_n$ of Δ ordered by their eigenvalues $\alpha_1 \le \alpha_2 \le \ldots \le \alpha_n$ correspond to increasingly "complex" functions on the Cayley graph [3][2][15]. Distributions corresponding to low complexity functions are smooth in the sense that on average a single transposition does not change $p(\sigma)$ much. Since noise enters p in the form of pairs or small groups of targets getting mixed up at a time, it is a reasonable assumption that p is smooth in this sense. The connection to the Fourier transform is given by the following proposition.

Proposition 1 *Each isotypal subspace V_ρ is spanned by eigenvectors of Δ with the same eigenvalue α_ρ.*

Proposition 1 tells us that the appropriate subspace to project p to is $W_{\alpha_{\mathrm{thresh}}} = \bigoplus_{\alpha_\rho \le \alpha_{\mathrm{thresh}}} V_\rho$ for some threshold α_{thresh}. In other words, the principled way of easing the computational burden of the data association problem is to work in Fourier space and maintain only the first few "low frequency" matrices of the Fourier transform of p. A low pass filter of this type does not affect the normalization of p, but might break $0 \le p(\sigma) \le 1$. In practice this usually does not cause problems in data association. However, it is important to bear in mind that what we denote by p below is only an approximation to a true distribution.

To find out which are the low frequency components, we need some more specific facts from the representation theory of $\mathbb{S}_n$[5][11]. First of all, the irreducibles are usually indexed by Young diagrams, such as



.

The total number of boxes must be n, and each row can contain at most as many boxes as the row above it. Young diagrams may be specified by the vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ giving the number of boxes in each row. The irreducible of shape (n) is the trivial representation $\rho_{(n)}(\sigma) \equiv 1$. The irreducible of shape $(1, 1, \ldots, 1)$ is the so-called alternating representation $\rho_{(1,1,\ldots,1)}(\sigma) = \mathrm{sgn}(\sigma)$, where $\mathrm{sgn}(\sigma) = 1$ if σ is an even permutation, and $\mathrm{sgn}(\sigma) = -1$ if it is odd. The rest of the irreducibles are multi-dimensional representations and are considerably more complicated.

While in general representation theory takes on its simplest form when working over $\mathbb{C}$, the symmetric group is special in that it does have a complete system of inequivalent irreducible (over $\mathbb{C}$) representations, which consist entirely of real matrices. This allows us to frame our approach to the data association problem in terms of real matrices and vector spaces.

The eigenvalues corresponding to each Young diagram are computed using the following proposition.

Proposition 2 *The eigenvalue of Δ corresponding to the isotypal indexed by λ is*

$$\alpha_\lambda = \binom{n}{2} \left(1 - \frac{\mathrm{tr}\,[\rho_\lambda([1,2])]}{d_\lambda} \right). \tag{4}$$

As for any Laplacian, the lowest eigenvalue of Δ is 0, and corresponds to the constant function on $\mathbb{S}_n$, i.e., $\alpha_{(n)} = 0$. The next lowest one corresponds to the $(n-1, 1)$ irreducible, which is $n-1$ dimensional. Following that we have $\alpha_{(n-2,2)}$ and $\alpha_{(n-2,1,1)}$, which correspond to $n(n-3)/2$ and $(n-1)(n-2)/2$ dimensional representations, respectively. After these come $\alpha_{(n-3,3)}$, $\alpha_{(n-3,2,1)}$ and $\alpha_{(n-3,1,1,1)}$ with dimensionalities $n(n-1)(n-5)/6$, $n(n-2)(n-4)/3$ and $\binom{n-1}{3}$. Beyond these are matrices with $O(n^8)$ elements, which are typically too large for practical applications. In the following sections D will denote the dimensionality of the largest $\widehat{p}(\rho)$ matrix that we maintain.

3 Tracking targets

A probabilistic data association algorithm must have at least the following three components:

1. A noise model describing how the distribution p degrades in the absence of observations.

2. A rule for updating p when one or more targets are observed at specific tracks.

3. An inference procedure for recovering the most likely assignment of targets to tracks.

We now describe the implementation of each of these components in our spectral framework.

3.1 Noise model

We model the uncertainty seeping into p by assuming that at each time t with some small probability β the targets corresponding to tracks i and j get exchanged, and this probability β is independent of i and j. If previously the correct assignment was σ, then the new assignment will be $[i, j]\,\sigma$. Considering all possible exchanges, the corresponding transition matrix is $I - \beta\Delta$, where Δ is again the Laplacian (3). Since this noise process is assumed to operate in the background in a continuous manner, we take the limiting case of infinitesimally small time intervals, leading to the update equation $p_{t'} = \exp\left(-\beta\,(t' - t)\,\Delta\right) p_t$, where

$$e^{-\beta(t'-t)\Delta} := \lim_{k\to\infty} \left(I - \frac{\beta\,(t'-t)}{k}\,\Delta \right)^k \tag{5}$$

is the matrix exponential. This, however, is just a diffusion process on the Cayley graph [6]. In particular, by Proposition 1 the Fourier space updates of p take on the particularly simple form

$$\widehat{p}_{t'}(\rho_\lambda) = e^{-\beta\alpha_\lambda(t'-t)}\; \widehat{p}_t(\rho_\lambda).$$

In our bandwidth-limited Fourier space scheme this update equation is not only very efficient to implement, but it is also exact: diffusion does not take us outside the subspace of distributions restricted to W.

3.2 Observations

The simplest type of observations consist of receiving information that with probability π, target o_i is at track r_j:

$$p\,(\,O_{i\to j} \,|\, \sigma\,) = \begin{cases} \pi & \text{if } \ \sigma(i) = j, \\ (1-\pi)/(n-1) & \text{if } \ \sigma(i) \neq j. \end{cases} \tag{6}$$

Fixing j, $p\,(\,O_{i\to j} \,|\, \sigma\,)$ is to be regarded as a probability distribution over which target i is observed at track j. The posterior over permutations is given by Bayes' rule:

$$p\,(\,\sigma \,|\, O_{i\to j}\,) = \frac{p\,(\,O_{i\to j} \,|\, \sigma\,)\ p(\sigma)}{\sum_{\sigma'\in\mathbb{S}_n} p\,(\,O_{i\to j} \,|\, \sigma'\,)\ p(\sigma')}\,, \tag{7}$$

giving rise to the update rule

$$p'(\sigma) = \begin{cases} \frac{1}{Z}\,\pi \cdot p(\sigma) & \text{for } \ \sigma(i) = j, \\ \frac{1}{Z}\,\frac{1-\pi}{n-1}\,p(\sigma) & \text{for } \ \sigma(i) \neq j, \end{cases} \tag{8}$$

where Z is the normalizer $Z = \pi \cdot p([\sigma(i) = j]) + \frac{1-\pi}{n-1} \cdot (1 - p([\sigma(i) \neq j]))$. In summary, observation updates hinge on (a) marginalizing p over sets of the form $C_{i\to j} = \{\, \sigma \in \mathbb{S}_n \mid \sigma(i) = j \,\}$ and (b) rescaling the projections of p onto subspaces corresponding to $C_{i\to j}$ and $C_{i\to j'},\ j' \neq j$.

A naive implementation of this update requiring explicit enumeration of group elements would run in time $O(n!)$. The next section will present a much faster algorithm. The observation updates couple the different Fourier components. This means when restricted to W the update step can only be approximate: an exact update would push some energy into some of the "higher frequency" Fourier components which we do not maintain in our bandwidth-limited representation of p. This is the only source of approximation error in our framework.

3.3 Inference

The third component to our data association system is a procedure for reading off the predicted association betweeen targets and tracks. In the simplest case this consists of returning the marginal probabilities $p([\sigma(i) = j])$ for a fixed object i as the track variable j ranges over $1, 2, \ldots, n$, or for a fixed track j as i ranges over $1, 2, \ldots, n$.

A more chellenging task is to find the optimal association $\tilde{\sigma} = \arg\max_{\sigma\in\mathbb{S}_n} p(\sigma)$. Again, the techniques presented in the next section lead to a search algorithm which can beat the apparent $O(n!)$ complexity of this operation.

4 Efficient computation

The spectral theory of the previous sections would be of little practical use if any of the required computations scaled with $n!$. One of the original contributions of this paper is to propose efficient computational procedures to perform the update and inference steps.

Both of these are based on Clausen's celebrated fast Fourier transform (FFT) for $\mathbb{S}_n$, which reduces the complexity of computing a complete Fourier transform from $n!^2$ to $n!\,(n+1)\,n\,(n-1)/3$ [8]. Unfortunately, space restrictions prevent us from giving more than just a sketch of Clausen's FFT and our extensions to it. Instead, the reader is referred to `http://www.cs.columbia.edu/~risi/SnOB` for a C++ implementation together with complete documentation describing the technical details.

Similarly to classical FFTs, Clausen's algorithm is based on decomposing the Fourier transform into a collection of Fourier transforms on subgroups. For

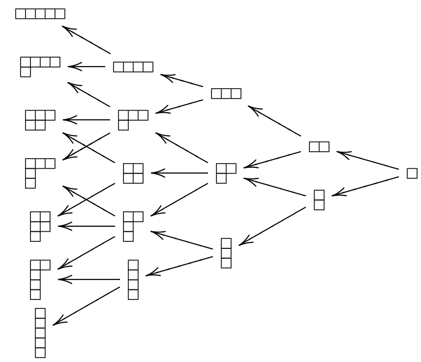


Figure 1: The Bratelli diagram for $\mathbb{S}_5$. The diagram shows that on restriction to $\mathbb{S}_{n-1}$, each irreducible representation ρ_λ of $\mathbb{S}_n$ decomposes into a sum of representations ρ_{λ^-} of $\mathbb{S}_{n-1}$, where each λ^- is derived from λ by removing a single square from its Young diagram.

$\mathbb{S}_n$ these subgroups are $\mathbb{S}_{n-1}, \mathbb{S}_{n-2}, \ldots, \mathbb{S}_1$, where $\mathbb{S}_k$ is understood to the be the subgroup of permutations permuting $1, 2, \ldots, k$ but leaving $k+1, \ldots, n$ fixed. To describe the decomposition, we introduce what we call contiguous cycles, denoted $[\![p, q]\!]$, which are permutations of the form

$$[\![p, q]\!]\,(i) = \begin{cases} i+1 & \text{if } \ p \leq i \leq q-1 \\ p & \text{if } \ i = q \\ i & \text{otherwise} \end{cases}$$

for $1 \leq p \leq q \leq n$. It is easy to see that the cosets $\{[\![j, n]\!]\,\mathbb{S}_{n-1}\}_{1\leq i\leq n}$ partition $\mathbb{S}_n$, in other words, any $\sigma \in \mathbb{S}_n$ can be written uniquely as $\sigma = [\![j, n]\!]\,\sigma'$ for some $\sigma' \in \mathbb{S}_{n-1}$ and some $1 \leq j \leq n$. This allows us to factor (1) in the form

$$\widehat{f}(\rho_\lambda) = \sum_{j=1}^{n} \rho_\lambda([\![j, n]\!]) \sum_{\sigma'\in\mathbb{S}_{n-1}} \rho_\lambda(\sigma')\, f_j(\sigma')\,, \qquad (9)$$

where $f_j(\sigma') = (f([\![j, n]\!]\,\sigma')$ are functions on $\mathbb{S}_{n-1}$. Clausen's FFT hinges on the observation that for certain specific choices of $\mathcal{R}$ (such as Young's orthogonal representation) the $\rho_\lambda([\![j, n]\!])$ can be written as a product of up to n very sparse matrices, and at the same time $\rho_\lambda(\sigma')$ decomposes into a direct sum of $\rho_{\lambda^-}(\sigma')$ matrices, where ρ_{λ^-} are now representations of $\mathbb{S}_{n-1}$. Plugging the latter into (9) gives the desired decomposition of $\widehat{f}$ into the smaller $\widehat{f}_i$ Fourier transforms:

$$\widehat{f}(\rho_\lambda) = \sum_{j=1}^{n} \rho([\![j, n]\!]) \bigoplus_{\lambda^-} \widehat{f}_j(\rho_{\lambda^-}). \qquad (10)$$

Exactly which partitions of n–1 the index λ^- runs over is revealed by the so called Bratelli diagram (Figure 1). Recursively applying (10) down to $\mathbb{S}_1$ gives the

```
find_largest_in_subgroup(σS_k){
  if k = 1 {
    if ‖p(σ)‖² > ‖p(σ̃)‖² { σ̃ ← σ; }
  }
  else {
    order {η_1, η_2, ..., η_k} so that ‖p(ση_1 S_{k-1})‖² ≥
              ‖p(ση_2 S_{k-1})‖² ≥ ... ≥ ‖p(ση_k S_{k-1})‖²;
    i ← 1;
    while ‖p(ση_i S_{k-1})‖² > ‖p(σ̃)‖² {
      find_largest_in_subgroup(ση_i S_{k-1});
      i ← i + 1;
    }
}}
```

Figure 2: Fast algorithm for finding the maximum probability permutation. The recursive procedure is called with the argument $e\mathbb{S}_n$, and $\tilde{\sigma}$ is initialized to any permutation. For $\sigma \in \mathbb{S}_n$, $\| p(\sigma\mathbb{S}_k) \|^2$ stands for $\frac{1}{k!}\sum_{\tau \in \mathbb{S}_k} \| p(\sigma\tau) \|^2$, which can be efficiently computed by partial Fourier transform methods, particularly when $\widehat{p}$ is sparse. Finally, $\eta_i \equiv [\![i,k]\!]$.

full FFT. The inverse transform is based on similar expressions for the $\widehat{f}_j$ in terms of components of $\widehat{f}$.

We generalize Clausen's FFT by modifying (10) to

$$\widehat{f}(\lambda) = \sum_{j=1}^{n} \rho([\![j,n]\!]) \Big[\bigoplus_{\lambda^-} \widehat{f}_{i\to j}(\rho_{\lambda^-}) \Big] \rho([\![i,n]\!]^{-1}),$$

where $1 \leq i \leq n$, and $f_{i\to j}(\sigma') = f([\![j,n]\!]\,\sigma'\,[\![i,n]\!]^{-1})$, leading to what we call "twisted" FFTs. The connection to the data association problem is that the two-sided coset $[\![j,n]\!]\,\mathbb{S}_{n-1}\,[\![i,n]\!]^{-1}$ to which $f_{i\to j}$ corresponds is exactly the set $C_{i\to j}$ introduced in 3.2. In particular, since $\rho_{(n-1)}$ is the trivial representation of $\mathbb{S}_{n-1}$, the Fourier component $\widehat{p}_{i\to j}(\rho_{n-1})$ is equal to the marginal $p([\sigma(i)=j])$.

Our Bayesian update step can then be performed by transforming to the $\widehat{p}_{i\to j}(\rho_{\lambda^-})$ matrices from the original $\widehat{p}(\rho_\lambda)$, reading off $\widehat{p}_{i\to j}(\rho_{(n-1)})$, rescaling the $\widehat{p}_{i\to j}$ according to (8), and then transforming back to $\widehat{p}$. For band-limited functions all these computations can be restricted to the subtree of the Bratelli diagram leading to those Fourier components $\widehat{p}(\rho_\lambda)$ which we actually maintain, and the total complexity of the update step becomes $O(D^2 n)$. This algorithmic viewpoint of information passing back and forth between the $\widehat{p}(\rho_\lambda)$ matrices and the $\widehat{p}_{i\to j}(\rho_{\lambda^-})$ matrices also sheds light on the nature of the coupling between the Fourier components: this arises because in the Bratelli diagram each partition of $n-1$ is linked to multiple partitions of n.

As regards the inference step, our proposed solution is a greedy search over the cosets that the FFT successively decomposes $\mathbb{S}_n$ into (Figure 2). The algorithm is based on the fact that the total energy $\| p(\sigma\mathbb{S}_k) \|^2$ in any $\mathbb{S}_k$-coset is the sum of the total energies of the constituent $\mathbb{S}_{k-1}$-cosets, and that by unitarity and (2) this energy is fast to compute. We cannot give a theoretical complexity guarantee for this algorithm, but in practical applications it is expected to be much faster than $O(n!)$.

4.1 Performance

In summary, our computational strategy is based on carefully avoiding complete Fourier transforms or any other $O(n!)$ computations. This makes our framework viable up to about $n = 30$. To indicate performance, on a 2.6 GHz PC, a single observation update takes 59ms for $n = 15$ when W is spanned by the first 4 Fourier components, and 3s when maintaining the first 7 components. For $n = 30$ and 4 components an update takes approximately 3s. Recall from the end of Section 2 that the third and fourth Fourier components have $O(n^4)$ scalar components. When storing and updating this many variables is no longer feasible, we are reduced to the just the first two components with 1 and $(n-1)^2$ scalar entries, respectively. At this point we lose the advantage in terms of detail of representation over other, simpler algorithms. Our spectral method is most likely to be of use in the $8 \leq n \leq 30$ range.

5 Higher order observations

One of the advantages of our algebraic approach is that it provides a consistent framework for incorporating more complicated observation and noise models. We consider the following.

1. $O_{(i_1,i_2,\ldots,i_k)\to(j_1,j_2,\ldots,j_k)}$ observations. These correspond to simultaneously observing o_{i_1} at r_{j_1}, o_{i_2} at r_{j_2}, etc.. A special case is when the entire permutation is observed.

2. $O_{\{i_1,i_2,\ldots,i_k\}\to\{j_1,j_2,\ldots,j_k\}}$ observations are like the above, except that there is no information about how the o_{i_a}'s and the r_{j_b}'s match up amongst themselves. As an example, we might localize a group of people in a room, but not be able to differentiate between them.

3. Diffusion localized to a subset of tracks. This becomes necessary when several tracks $r_{j_1}, r_{j_2}, \ldots, r_{j_k}$ approach each other, and hence it becomes more likely that corresponding targets get exchanged.

Let $\tau_{i_1,i_2,\ldots,i_k}$ denote any permutation such that $\tau_{i_1,i_2,\ldots,i_k}(i_a) = n+1-a$ for $a = 1,2,\ldots,k$. The coset $C_{(i_1,i_2,\ldots,i_k)\to(j_1,j_2,\ldots,j_k)} = \tau^{-1}_{j_1,j_2,\ldots,j_k}\mathbb{S}_{n-k}\,\tau_{i_1,i_2,\ldots,i_k}$ is

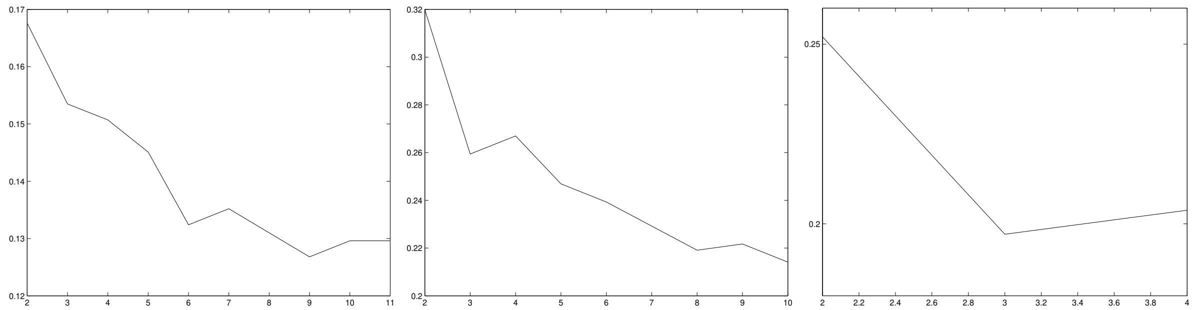


Figure 3: Tracking error on the air traffic control dataset for $n = \{6, 10, 15\}$ as a function of the number of Fourier components maintained ($p_{\mathrm{obs}} = 0.05, 0.1$ and 0.3)

then exactly the set of permutations corresponding the type of observation in case 1 above. The observation is implemented by transforming to the corresponding $\widehat{p}_{(i_1,i_2,\ldots,i_k)\to(j_1,j_2,\ldots,j_k)}$ partial Fourier transform, updating as in Section 4, and then transforming back.

Similarly, the second type of observation corresponds to transforming to the $\tau^{-1}_{j_1,j_2,\ldots,j_k}(\mathbb{S}_{n-k}\times\mathbb{S}_k)\,\tau_{i_1,i_2,\ldots,i_k}$ coset. For small k this is accomplished by transforming to all $k!$ possible $C_{(i'_1,i'_2,\ldots,i'_k)\to(j_1,j_2,\ldots,j_k)}$ cosets, where $i'_1, i'_2, \ldots, i'_k$ is a permutation of $i_1, i_2, \ldots, i_k$.

Finally, localized diffusion in the general case hinges on the convolution theorem (see Appendix). For the special case of diffusing over pairs of tracks (j_1, j_2), as used in our experiments, this reduces to $\widehat{p}(\rho_\lambda)$ being updated to

$$\left(\frac{1+e^{-\beta(t'-t)}}{2}I + \frac{1-e^{-\beta(t'-t)}}{2}\rho_\lambda([j_1,j_2])\right)\cdot\widehat{p}(\rho_\lambda).$$

6 Experiments

We performed experiments on a dataset of (x, y) radar positions of aircraft within an approximately 30 mile diameter area around New York's John F. Kennedy International Airport on a given day in May, 2006, updated at few second intervals. The data is available in streaming format from `http://www4.passur.com/jfk.html`. The presence of three major airports in the area (JFK, La Guardia and Newark), together with specific local procedures, such as the routing of aircraft over the Hudson river for noise abatement, and simultaneous approaches at closely spaced parallel runways at JFK, result in aircraft frequently crossing paths in the (x, y) plane, making the data particulary interesting for our purposes.

The number of simultaneously tracked aircraft was $n = 6, 10$ and 15, and whenever an aircraft's tracking was terminated either because it has landed or because it has left the area, its track was reassigned to a new aircraft. Our method could be extended to tracking varying numbers of objects, but that is beyond the scope of the present paper. The number 15 is significant becuase it is large enough to make storing a full distribution over permutations impossible ($15! \approx 1.3\cdot 10^{12}$), yet with up to 4 Fourier components our algorithm can still run on an entire day's worth of data in a matter of hours. The original data comes pe-labeled with track/target assignment information. We introduce uncertainty by at each timestep t for each pair (j_1, j_2) of tracks swapping their assignment with probability $p_{\mathrm{mix}}\exp\big(-\parallel \boldsymbol{x}_{j_1}(t) - \boldsymbol{x}_{j_2}(t)\parallel^2 / \big(2s^2\big)\big)$, where $\boldsymbol{x}_j(t)$ is the position of track j at time t ($s = 0.1$ and $p_{\mathrm{mix}} = 0.1$). This models noisy data, where the assignment breaks down when targets approach each other.

The algorithm has access to track locations from which it can model this shifting distribution over assignments using the pairwise diffusion updates discussed in Section 5. In addition, with probability p_{obs} for each t and each j, the algorithm is provided with a first order observation of the form $O_{i\to j}$, with $\pi = 1$. The task of the algorithm is to predict the target currently assigned to track j by computing $\arg\max_i p([\sigma(i) = j])$. The evaluation measure is the fraction of incorrect predictions.

Experimental results are promising (Figure 3) and clearly show that performance generally improves with the number of irreducibles. It is difficult to compare our alogrithm directly with other association methods, since most of them do not accept probabilistic observation and diffusion updates in the form supplied by the data. However, the best performance we could achieve with the information-form filter [12] for $n = 6$ was 0.194 error and for $n = 10$ was 0.2885, which is significantly worse than our own results of 0.127 and 0.215, respectively. Time limitations prevented us from more extensive comparisons, but our preliminary results indicate that on the flight tracking task the two approaches often perform similarly, but the information form filter is better suited to domains where the distribution is highly peaked, whereas our method performs better when there is more uncertainty and the distribution is smooth. We expect that our spectral method will give the largest performance gain for problems where complicated distributions need to be approximated accurately.

7 Conclusions

In this paper we argued that just as it is natural to approximate smooth periodic functions on $\mathbb{R}^n$ by their first few Fourier components, it is similarly natural to approximate smooth association distributions over the intractably large $\mathbb{S}_n$ by their first few Fourier matrices. As these matrices grow very rapidly, there is a severe tradeoff between accuracy and speed in how many components we maintain. We presented efficient computational methods based on Clausen's FFT for updating these matrices. The critical $O_{i\to j}$ observation step runs in $O(D^2 n)$, where D is the dimensionality of the largest maintained Fourier matrix (which has D^2 elements). In a realistic aircraft tracking scenario with $n = 15$ targets the algorithm can easily run in real time. The source of approximation error in our framework is the Fourier cutoff (band limit) and the the Bayesian identity observation updates, which cut across isotypics.

While we developed this machinery in the context of the Bayesian data association problem in multi-object tracking, clearly it has broader applicability to a wide range of problems involving distributions over the symmetric group. The experiments are in the exploratory stage, but show that in certain domains the new algorithm beats Schmutisch et. al.'s information form data association filter, which is considered state of the art.

Acknowledgments

We would like to thank to anonymous reviewers for correcting several minor errors. This work was supported in part by National Science Foundation grants IIS-0347499, CCR-0312690 and IIS-0093302.

References

[1] Yaakov Bar-Shalom and Thomas E. Formann. *Tracking and Data Association.* Academic Press, 1988.

[2] Mikhail Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems NIPS'2001*, volume 14. MIT Press, 2002.

[3] Fan R. K. Chung. *Spectral Graph Theory.* Number 92 in Regional Conference Series in Mathematics. American Mathematical Society, 1997.

[4] Persi Diaconis. *Group Representation in Probability and Statistics*, volume 11 of *IMS Lecture Series*. Institute of Mathematical Statistics, 1988.

[5] Gordon James and Adalbert Kerber. *The Representation Theory of the Symmetric Group.* Addison-Wesley, 1981.

[6] Risi Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the ICML*, 2002.

[7] D. Maslen and D. Rockmore. Generalized FFTs — a survey of some recent results. In *Groups and Computation II*, volume 28 of *DIMACS Ser. Discrete MAth. Theor. Comput. Sci.*, pages 183–287. AMS, Providence, RI, 1997.

[8] M.Clausen. Fast generalized Fourier transforms. *Theor. Comput. Sci.*, pages 55–63, 1989.

[9] D.B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Aerospace and Electronic Systems*, 1979.

[10] Daniel N. Rockmore. Some applications of generalized FFTs. In *Proceedings of the DIMACS Workshop on Groups and Computation*, pages 329–369, 1995.

[11] Bruce E. Sagan. *The Symmetric Group.* Graduate Texts in Mathematics. Springer, 2001.

[12] Brad Schumitsch, Sebastian Thrun, Gary R. Bradski, and Kunle Olukotun. The information-form data association filter. In *NIPS*, 2005.

[13] Jean-Pierre Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Texts in Mathamatics.* Springer-Verlag, 1977.

[14] Jaewon Shin, Nelso Lee, Sebastian Thrun, and Leonidas Gubias. Lazy inference on object identities in wireless sensor networks. Technical report, Stanford AI laboratory, submitted for publication, 2006.

[15] Alexander J. Smola and Risi Kondor. Kernels and regularization on graphs. In Manfred Warmuth and Bernhard Schökopf, editors, *Proceedings of the Conference on Learning Theory and Kernels Workshop*, 2003.

[16] Lawrence Stone, Carl Barlow, and Thomas Corwin. *Bayesian Multiple Target Tracking.* Artech House, 2000.