Fourier-Information Duality in the Identity Management Problem

Xiaoye Jiang[1], Jonathan Huang[2], and Leonidas Guibas[1]

[1] Stanford University, Stanford, CA, 94305, USA
[2] Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Abstract. We compare two recently proposed approaches for representing probability distributions over the space of permutations in the context of multi-target tracking. We show that these two representations, the Fourier approximation and the information form approximation can both be viewed as low dimensional projections of a true distribution, but with respect to different metrics. We identify the strengths and weaknesses of each approximation, and propose an algorithm for converting between the two forms, allowing for a *hybrid* approach that draws on the strengths of both representations. We show experimental evidence that there are situations where hybrid algorithms are favorable.

1 Introduction

In this paper we consider the *identity management problem* which arises in a number of multi-target tracking scenarios in computer vision and robotics. Typical multi-target tracking systems maintain tracks of n people and the identity of the person corresponding to each track. A successful tracking system must reason in the face of noisy *evidence events*, in which an identity may be partially revealed to be at a particular track, as well as *mixing events*, in which identities can be confused when tracks cross paths.

To handle this uncertainty algorithmically, identity management is formalized mathematically as a filtering problem for identity-to-track associations, in which one must maintain a distribution over permutations. Since the space of permutations scales factorially in the number of tracked objects, n, however, it is not tractable to explicitly represent distributions over permutations for nontrivial n. Moreover, typical compact representations, such as graphical models, are not effective due to the mutual exclusivity constraints associated with permutations.

To efficiently represent and reason with such distributions, researchers have turned to a number of compact approximate representations. There are two competing methodologies in the identity management literature which have garnered the most attention in the last decade: the *Fourier theoretic* approach [6, 7, 11], and the *information theoretic* approach [14, 17]. Cosmetically, both methods seem similar in spirit — the Fourier theoretic approach represents distributions over possible associations by maintaining marginal probabilities involving small subsets of objects, while the information theoretic approach represents similar terms, but working with unnormalized log-probabilities.

Despite progress made on both approaches over the last several years, there has been little work in unifying or even comparing the two approaches. In this paper we compare the Fourier and information approaches, drawing parallels between the two methods, and contrasting their strengths and weaknesses. The main contributions of our work is as follows:

1. Among the many parallels between the two representations, we identify an interesting duality between the two types of events (mixing and evidence) that must be processed during identity management. [6] showed that mixing events can be handled within the Fourier representation without increasing representational complexity, while evidence events always increase the representation complexity. We show that the opposite is true for the information form representation — that while evidence events can be handled without increased complexity, mixing events cannot be handled exactly without increasing representation complexity. We also make a connection between the two representations by viewing them as parametric representation of projected distributions with different metrics.
2. We explore the problem of converting between the Fourier and information theoretic representations and show that the conversion problem is #P-hard, but that due to recent advances in permanent approximation theory, approximate conversion is possible in polynomial time.
3. Using our algorithm for converting between the two forms, we propose a hybrid method that draws on the strengths of both representations and show experimental evidence that there are situations where hybrid algorithms are favorable.

2 Probabilistic Identity Management

In identity management, we are interested in maintaining a distribution over possible permutations which assign n identities to n tracks maintained by an internal tracker. We denote permutations as σ, where $\sigma(k)$ is the track belonging to the kth identity. Over time, the distribution over permutations in identity management is subject to change due to two causes: *mixing events* and *observation events*. In a mixing event, a subset of people can walk too closely together, leading to confusion about the identity-to-track associations for their respective tracks. This confusion is balanced by observation events, in which, for example, the color of an individual's clothing is captured by a sensor, giving information about his or her identity.

Uncertainty over permutations in identity management can be modeled with a hidden Markov model, where the joint probability of a sequence of latent permutations $(\sigma^{(1)}\dots,\sigma^{(T)})$ and observed data $(z^{(1)},\dots,z^{(T)})$ factors as :

$$h(\sigma^{(1)},\dots,\sigma^{(T)},z^{(1)},\dots,z^{(T)}) = h(z^{(1)}|\sigma^{(1)})\cdot\prod_{t=1}^{T} h(z^{(t)}|\sigma^{(t)})\cdot h(\sigma^{(t)}|\sigma^{(t-1)}).$$

We will refer to $h(\sigma^{(t)}|\sigma^{(t-1)})$ as the *mixing model*, which captures, for example, that tracks i and j swapped identities with some probability. We refer to $h(z^{(t)}|\sigma^{(t)})$ as the *observation model*, which captures, for example, the probability of observing a green blob given that Alice was at Track 1.

2.1 Inference Operations

There are two fundamental probabilistic inference operations that we focus on. The first is the *prediction/rollup* operation, which, given the distribution at time t, $h(\sigma^{(t)}|z^{(1)},\dots,z^{(t)})$, and a mixing event, computes the distribution at the following timestep by multiplying by the mixing model and marginalizing over the permutation at the previous timestep:

$$h(\sigma^{(t+1)}|z^{1},\ldots,z^{(t)}) = \sum_{\pi\in S_n} h(\sigma^{(t+1)}|\sigma^{(t)}=\pi)\cdot h(\sigma^{(t)}=\pi|z^{(1)},\ldots,z^{(t)}).$$

The second is the *conditioning* operation, which, given a new observation $z^{(t+1)}$, performs a Bayesian update to compute the posterior distribution:

$$h(\sigma^{(t+1)}|z^{1},\ldots,z^{(t+1)}) \propto \ell(z^{(t+1)}|\sigma^{(t+1)})\cdot h(\sigma^{(t+1)}|z^{1},\ldots,z^{(t)}).$$

For explicit representations of the distribution $h(\sigma^{(t)})$, inference is intractable for all but very small n with running time complexities of $O((n!)^2)$ and $O(n!)$ for prediction/rollup and conditioning respectively. In this paper, we discuss two methods which have been proposed in recent years for compactly representing distributions over permutations and how these two inference operations can be performed efficiently with respect to each representation.

3 Two Dueling Representations

In this section we introduce the Fourier and information representations for distributions over permutations. In the simplest case, both representations maintain coefficients corresponding to the event that a single track j is associated with a single identity k, for all (track,identity) pairs j,k. Additionally, in both representations, one can also formulate generalizations which maintain coefficients corresponding to joint events that small subsets of identities map to small subsets of tracks. However, we show that with respect to the Fourier representation, the prediction/rollup step of inference is ‘easy’ in the sense that it can be performed efficiently and exactly, while the conditioning step of inference is ‘difficult’ since it can only be performed approximately. With respect to the information form representation, the roles are reversed, with prediction/rollup ‘difficult’ and conditioning ‘easy’.

3.1 Fourier Domain Representation

The identity management problem was first introduced by Shin et al. [16], who proposed a representation based on collapsing the factorial sized distribution over permutations to just its *first-order marginals*, the n^2 marginal probabilities of the form:

$$H_{jk} = h(\sigma:\sigma(k)=j) = \sum_{\sigma\in S_n\,:\,\sigma(k)=j} h(\sigma).$$

The first-order marginals can be represented in a doubly stochastic matrix[3] (called a *belief matrix* in [16]). As an example, the matrix

$$H = \left[\begin{array}{c|ccc} & \text{Alice} & \text{Bob} & \text{Charlie} \\ \hline \text{Track 1} & 1/4 & 1/2 & 1/4 \\ \text{Track 2} & 3/8 & 3/8 & 1/4 \\ \text{Track 3} & 3/8 & 1/8 & 1/2 \end{array}\right].$$

[3] A doubly stochastic matrix has rows and columns which sum to 1. In the identity management setting, it reflects the constraint that every identity must map to *some* track, and that there is *some* identity on every track.

By simply representing these first-order terms, it is already possible to make useful predictions. For example, we can predict the track at which the identity Alice is currently located, or predict the identity currently located at track 2.

The first-order marginal probabilities can be generalized to higher-order marginals which maintain, for example, the probability that a pair of tracks is jointly associated with a pair of identities. For example, we might be interested in the *second-order* probability that Alice and Bob are jointly in Tracks 1 and 2, respectively.

The reason for referring to these simple matrix-of-marginal type representations as 'Fourier' representations is due to the mathematical theory of generalized Fourier transforms for the symmetric group (see [3, 13, 15]). Just like the Fourier transform of a function on the real line can be separated into low and high frequency terms, a function over the symmetric group (the group of permutations) can be separated into low-order effects and higher-order effects. We remark that the Fourier coefficients of [6, 11] do not literally take the form of marginal probabilities but instead can be thought of as a set of coefficients which can be used to uniquely reconstruct the marginals. Loosely speaking, low-order marginal probabilities of a distribution can always be reconstructed using a subset of 'low-frequency' terms of its Fourier transform. Varying the maximum represented frequency yields a principled way for trading between accuracy and speed of inference.

Matrices of marginals can be viewed as a *compact summary* of a distribution over permutations, but they can additionally be viewed as an *approximation* to that distribution by applying the inverse Fourier transform to a truncated Fourier expansion. Given the first-order marginals H of a distribution, the approximate distribution is:

$$h(\sigma) = \frac{n-1}{n!}\mathrm{Tr}(H^T M_\sigma) - \frac{n-2}{n!}$$

where M_σ is the first-order *permutation matrix* associated with σ.[4] The above equation can be generalized to higher-order Fourier representations allowing for successively better approximations to the original distribution h.

3.2 Information Form Representation

Instead of representing the marginal probability that an identity k will be associated with track j, in the information form representation, one maintains a 'score' Ω_{jk} for each identity-track pair (k, j). The probability of a joint assignment of identities to tracks is parameterized as:

$$h(\sigma) = \frac{1}{Z_\Omega}\exp\left(\sum_{k=1}^{n} \Omega_{\sigma(k),k}\right) = \exp\Big(\mathrm{Tr}(\Omega^T M_\sigma)\Big),$$

where M_σ is the first-order *permutation matrix* associated with σ and Z_Ω is the normalizing constant. We observe that if we add a constant to every entry within a single row or single column of Ω, the distribution parameterized by Ω does not change. The

[4] Given a $\sigma \in S_n$, the permutation matrix associated with σ is defined as the $n \times n$ matrix M, with entries $M_{jk} = 1$ if $j = \sigma(k)$, 0 otherwise. This (first-order) permutation matrix can easily be generalized to higher-order permutation matrices whose nonzero entries represent assignments of tuples of identities $(k_1, \ldots, k_m)$ to tuples $(j_1, \ldots, j_m)$ of tracks.

Inference Operation	**Fourier (First Order)**		**Information Form (First Order)**	
	Accuracy	Complexity	Accuracy	Complexity
Prediction/Rollup	Exact	$\mathcal{O}(n)$	Approximate	$\mathcal{O}(n)$
Conditioning	Approximate	$\mathcal{O}(n^3)$	Exact	$\mathcal{O}(n)$
Normalization	Exact	$\mathcal{O}(n^2)$	Approximate	$\mathcal{O}(n^4 \log n)$
Maximization	Exact	$\mathcal{O}(n^3)$	Exact	$\mathcal{O}(n^3)$

Table 1. We compare common inference operations for the Fourier and information forms assuming the simplest case using a first-order representation, pairwise mixing, and first-order observations.

entries of Ω are referred to as the *information coefficients* of the distribution P. Note that multiple settings of the information coefficient matrix Ω can correspond to the same distribution. For example, adding a constant c to any row or column of Ω does not change the distribution parameterized by Ω.

As with Fourier coefficients, it is possible to consider generalizations of the information form to higher order terms. For example, we can maintain a nonzero 'score' $\Omega'_{(j_1,j_2),(k_1,k_2)}$ where (j_1, j_2) denote a pair of tracks and (k_1,k_2) denote a pair of identities. Thus, in the information domain, the probability over permutations is parameterized as:

$$h(\sigma) = \frac{1}{Z_{\Omega'}} \exp\left(\sum_{k_1=1}^{n} \sum_{k_2 \neq k_1} \Omega_{(\sigma(k_1),\sigma(k_2)),(k_1,k_2)} \right) = \exp\left(\mathrm{Tr}(\Omega^T M_\sigma) \right),$$

where Ω' is a *second-order information coefficient matrix*.

3.3 Comparing the Two Representations

We now compare and contrast the two representations. Of particular interest are the probabilistic inference operations which are common in identity management. The challenge is how to perform these probabilistic operations using either the Fourier or information forms, exactly or approximately, in polynomial time. For simplicity, we will restrict our focus to first order representations for both the Fourier and information domains. Additionally, we assume that mixing only occurs between a pair of tracks i and j at any given time, leading to the following simple mixing model in which one draws a permutation $\pi \sim m_{ij}(\pi)$, where:

$$m_{ij}(\pi) = \begin{cases} p & \text{if } \pi = id \\ 1-p & \text{if } \pi = (i,j) \\ 0 & \text{otherwise} \end{cases},$$

and sets $\sigma^{(t+1)} \leftarrow \pi \cdot \sigma_t$ (where $\cdot$ represents the composition of two permutations).

We also assume the simple observation model (employed in [6, 11]) which assumes that we get observations z of the form: 'track j is color r'. The probability of seeing color r at track j given an identity-to-track association σ is

$$\ell(\sigma) = \mathrm{Prob}(\text{track } j \text{ is color } r|\sigma) = \alpha_{\sigma^{-1}(j),r},$$

where $\sum_r \alpha_{\sigma^{-1}(j),r} = 1$. The likelihood model parameters α can be constructed based on prior knowledge of color profiles of the moving targets [6].

For a tabular summary of the inference operations considered in this section, we refer the reader to Table 1.

Prediction/Rollup. In general, the pairwise mixing models considered in this paper can be thought of as a special case of random walk transitions over a group, which assume that $\sigma^{(t+1)}$ is generated from $\sigma^{(t)}$ by drawing a random permutation $\pi^{(t)}$ from some distribution $m^{(t)}$ and setting $\sigma^{(t+1)} = \pi^{(t)}\sigma^{(t)}$. The permutation $\pi^{(t)}$ represents a random identity permutation that might occur among tracks when they get close to each other (what we call a *mixing event*).

The motivation behind the random walk transition model is that it allows us to write the prediction/rollup operation as a *convolution* of distributions, and as a result the familiar *convolution theorem* of Fourier analysis holds. Below we state the convolution theorem for the special case of first order Fourier representations, but a more general statement can be found in, for example, [6].

Proposition 1 (Convolution theorem). *Let $M^{(t)}$ be the first order matrix of marginals for the distribution $m^{(t)}$ and $H^{(t)}$ be the first order matrix for $h(\sigma^{(t)}|z^{(1)},\dots,z^{(t)})$. The first order matrix for the distribution after the prediction step, $h(\sigma^{(t+1)}|z^{(1)},\dots,z^{(t)})$ is:*

$$H^{(t+1)} = M^{(t)} \cdot H^{(t)},$$

where the operation on the right side is matrix multiplication.

Prediction/rollup in the Fourier domain is *exact* in the sense that first order marginals for timestep $t+1$ can be computed exactly from first order marginals at timestep t. In contrast the same operation cannot be performed exactly with respect to information form coefficients and in particular, we argue that, if the distribution $h(\sigma^{(t)})$ can be represented with first order information form coefficients, then under pairwise mixing, second order information form coefficients are necessary and sufficient for representing the distribution $h(\sigma^{(t+1)})$.

Proposition 2. *Let $\Omega^{(t)}$ be the first order information coefficient matrix for the distribution $h(\sigma^{(t)}|z^{(1)},\dots,z^{(t)})$. There exists a second order information coefficient matrix $\Omega^{(t+1)}$ which exactly parameterizes the distribution obtained by the prediction/rollup step $h(\sigma^{(t+1)}|z^{(1)},\dots,z^{(t)})$ in the information domain.*

Proof. Given information coefficients Ω which parametrize $h(\sigma^{(t)}|z^{(1)},\dots,z^{(t)})$, we argue that there exists an $n(n-1)$-by-$n(n-1)$ 2nd order information coefficient matrix Ω' which exactly parametrize $h(\sigma^{(t+1)}|z^{(1)},\dots,z^{(t)})$. To see this, suppose that track k_1 and k_2 mixed up, then the distribution after the rollup operation evaluated on σ would be proportional to

$$p\exp\left(\mathrm{Tr}(\Omega^T M_\sigma)\right) + (1-p)\exp\left(\mathrm{Tr}(\Omega^T M_{\pi\sigma})\right).$$

In such a expression, any entries in Ω that does not lie in row k_1 or k_2 is still an additive term in the logarithmic space for characterizing the posterior.

For entries that lie in either row k_1 or k_2, we need to form $n(n-1)$ 2nd order information coefficients $\Omega_{(\sigma(k_1),\sigma(k_2)),(k_1,k_2)}$. With those coefficients, we can represent the posterior distribution evaluated on σ using logarithmic likelihoods $\Omega_{(\sigma(k_1),\sigma(k_2)),(k_1,k_2)}$ together with $\Omega_{\sigma(k),k}$, where $k \neq k_1, k_2$.

It turns out that the above logarithmic likelihoods can be combined together into a second order information matrix. This is because the representation theory applies to the logarithmic space of the information form representation. □

Instead of increasing the size of the representation at each timestep, a sensible approximation is to compute a projection of $h(\sigma^{(t+1)})$ to the space of distributions which can be represented in first-order information form. Schumitsch et al. [14] proposed the following update:

$$\Omega^{(t+1)} = \log\left(M^{(t)} \cdot \exp(\Omega^{(t)})\right)$$

which they showed worked well in practice. The exponential and logarithmic functions in the formula refer to elementwise operations rather than matrixwise operations.

Conditioning. In contrast with the ease of prediction/rollup operations, conditioning a distribution in the Fourier domain is more complex and increases the size of the representation.

Proposition 3 (Kronecker conditioning [6]). *Let $H^{(t+1)}$ be the first order matrix of marginals for the distribution $h(\sigma^{(t+1)}|z^{(1)},\ldots,z^{(t)})$, then there exists a second order matrix of marginals which exactly parametrize the distribution obtained by the conditioning step $h(\sigma^{(t+1)}|z^{(1)},\ldots,z^{(t+1)})$ in the Fourier domain.*

Using information coefficients, however, conditioning can be performed exactly, and takes a particularly simple and efficient form (that of a local addition) which does *not* increase the representation complexity.

Proposition 4 (Schumitsch et al. [14]). *If $h(\sigma) \propto \exp\left(Tr(\Omega^T M_\sigma)\right)$, then the update is of the form*

$$\Omega_{jk} \leftarrow \Omega_{jk} + \log \alpha_{k,r}.$$

where $k = \sigma^{-1}(j)$. The complexity of this update is $\mathcal{O}(n)$.

Normalization and Maximization. Normalization is a major inference operation and appears, for example, as a subroutine of the conditioning and marginalization operations, i.e., computing $\sum_\sigma \ell(\cdot|\sigma)h(\sigma|\cdots)$ or $\sum_\sigma h(\sigma)$. In the Fourier domain, normalization is 'free' since the *zeroth-order* marginal is exactly the normalization constant $Z = \sum_\sigma h(\sigma)$. Thus with respect to the irreducible Fourier coefficients of [6, 11], normalization can be performed by dividing all Fourier coefficients by the lowest-frequency coefficient. Alternatively, if the matrix of marginals, H, is represented explicitly, the normalization constant Z is simply the sum across any row or column of H. One can then normalize by scaling every entry of H by Z.

It may be somewhat surprising to realize that the normalization problem is provably hard in the information domain since the probability of a joint assignment may at first glance seem to factorize as:

$$h(\sigma) \propto \prod_{k=1}^{n} w_{k,\sigma(k)} = \exp(\sum_k \Omega_{\sigma(k),k}),$$

which would allow one to factor the normalization problem into tractable pieces. However, due to mutual exclusivity constraints which disallow identities from mapping to the same track, probabilistic independence is not present. Instead, the normalization constant, $Z = \sum_{\sigma \in S_n} \prod_k W_{k,\sigma(k)}$, is exactly the matrix permanent of $W = \exp(\Omega)$, whose computation is #P-complete (even for binary matrices). We have:

Proposition 5. *Computing the normalization constant of the information form parameterization is #P-complete.*

We remark that despite the dramatic differences with respect to normalization, computing the permutation which is assigned the maximum probability under h (instead of summing over h) reduces to the same problem for both the Fourier and information forms due to the fact that the exponential is a monotonic function. In the first-order case, for example, one must compute $\arg\max_\sigma \mathrm{Tr}\left(H^T M_\sigma\right)$ (see Equation 3.1), which can be efficiently solved using either linear programming or a number of other combinatorial algorithms.

Both Forms are Low-Dimensional Projections. Since the Fourier transform is linear and orthogonal [3], the Fourier approximation of a distribution h over permutations can be thought of as an ℓ_2 projection of h onto a low-frequency Fourier basis V which can be interpreted as affine marginal constraints. This projection is associated with the following *Pythagorean theorem*, which says that if g is any function lying in the span of V, then $\|g-h\|_{\ell_2}^2 = \|g-h'\|_{\ell_2}^2 + \|h'-h\|_{\ell_2}^2$, where h' is the Fourier projection of h onto the span of V.

The information form representation can be thought of, on the other hand, as an information projection of h to the same low-frequency Fourier subspace V using the *KL-divergence* metric. Recall that the KL-divergence, also known as the *relative entropy* is defined as $D(q||h) = \sum_\sigma q(\sigma) \log \frac{q(\sigma)}{h(\sigma)}$. Given a doubly stochastic matrix H of first order marginals, the information projection (IP) can be formulated as follows:

$$\begin{array}{lll} (IP)\ \min_q & \sum_\sigma q(\sigma) \log \dfrac{q(\sigma)}{h(\sigma)} \\ \text{s.t.} & \sum_\sigma q(\sigma) M_\sigma = H \\ & q(\sigma) \geq 0, \forall \sigma \end{array} \qquad \begin{array}{lll} (ME) & \min_q & \sum_\sigma q(\sigma) \log q(\sigma) \\ & \text{s.t.} & \sum_\sigma q(\sigma) M_\sigma = H \\ & & q(\sigma) \geq 0, \forall \sigma \end{array}$$

In the special case, where the distribution h to be projected is uniform, i.e., we have no prior knowledge, then the information projection problem becomes the maximum entropy (ME) problem. The objective in (ME) coincides with the maximum entropy principle in Bayesian probability, where the information entropy of a distribution q over S_n is $H[q] = -\sum_\sigma q(\sigma) \log q(\sigma)$. The maximum entropy distribution can be thought of as the least biased distribution encoding some given information (about low-order marginals in our case). We remark that the normalization constraint $\sum_\sigma q(\sigma) = 1$ is implicitly contained in the first constraint, $\sum_\sigma q(\sigma) M_\sigma = H$.

The following result (see proof of Proposition 7) shows that the solution to maximum entropy problem *must* be parametrizable as an information form distribution:

Proposition 6. *The solution to (IP) is guaranteed to take the form $h(\sigma)\exp\left(Tr(\Omega^T M_\sigma)\right)$ while the solution to (ME) is guaranteed to take the form $q(\sigma) \propto \exp\left(Tr(\Omega^T M_\sigma)\right)$. The* Pythagorean theorem *holds: if g is any function that satisfies the marginal constraints, then $D(g||h) = D(g||h') + D(h'||h)$, where h' is the information projection of h.*

3.4 Discussion

As we have shown, both the Fourier and information forms can be thought of as methods for approximating distributions over permutations via a low-dimensional projection. However, we have also argued that each method has their own respective advantages and disadvantages with respect to the two inference operations of prediction/rollup and conditioning. While prediction/rollup updates, which increase the information entropy of the maintained distribution, can be performed exactly with respect to a Fourier representation, conditioning updates, which typically decrease the entropy, can be performed exactly with respect to an information form representation. As a result, Fourier representations are typically much more suitable for modeling problems with high uncertainty, while information form representations are more suitable for problems with low uncertainty. In Section 7, we will validate these claims with experiments.

4 Representation Conversion

In this section we show a natural method for converting between the two representations. Since the two representations do not describe the same space of functions, conversion can only be approximate. We show in particular that much like the normalization problem which we discussed in the previous section, converting between the two representations requires solving the *matrix permanent problem*.

4.1 From Information Coefficients to Fourier Coefficients

We first consider the problem of estimating low-order marginals from information coefficients. Given the information coefficients Ω, we can compute the first order marginal probability that identity k maps to j, H_{jk}, by conditioning on $\sigma(k) = j$, then normalizing. Note that the posterior after conditioning can also be written in information form and that the normalization operation corresponds to taking the permanent of the information matrix of the posterior distribution. We have:

$$H_{jk} = \sum_{\sigma:\sigma(k)=j} h(\sigma) = \frac{\exp(\Omega_{jk})\mathrm{perm}(\exp(\hat{\Omega}_{jk}))}{\mathrm{perm}(\exp(\Omega))}.$$

Here $\hat{\Omega}_{jk}$ denotes the $n-1$ by $n-1$ submatrix of Ω with the j'th row and k'th column removed. The matrix $\exp(\hat{\Omega}_{jk})$ denotes component-wise exponentials rather than matrix exponentials. We therefore conclude that to convert from information coefficients to Fourier coefficients, one must compute matrix permanents.

4.2 From Fourier Coefficients to Information Coefficients

We now discuss the opposite conversion from Fourier coefficients to Information coefficients, for which we take the maximum entropy approach described in the previous section (problem (ME)). Given, say, the first-order marginal probabilities, we are interested in computing the maximum entropy distribution consistent with the given marginals, which we argued can be parameterized in information form. We now turn to the problem of algorithmically optimizing the entropy with respect to low-order constraints. Our approach is to solve the dual problem [1]:

Proposition 7. *The dual problem of* (ME) *is:*

$$\max_{Y} \; Tr\Big(Y^T H\Big) - \sum_{\sigma} \exp\Big(Tr(Y^T M_\sigma) - 1\Big),$$
$$s.t. \; Y \leq 0.$$

Proof. The Lagrangian for (ME) is given by:

$$\sum_{\sigma} q(\sigma)\log q(\sigma) - \sum_{\sigma} s_\sigma q(\sigma) - \mathrm{Tr}\Big(Y^T(\sum_{\sigma} q(\sigma)M_\sigma - H)\Big),$$

where s_σ and Y are dual variables associated with the constraint $q(\sigma) \geq 0$, and $\sum_\sigma q(\sigma)M_\sigma = Q$. The KKT conditions tell us that for (ME):

$$1 + \log q(\sigma) - s_\sigma - \mathrm{Tr}\Big(Y^T M_\sigma\Big) = 0.$$

Assuming all $q(\sigma) > 0$, which gives us $s_\sigma = 0$ becuase of the dual complementary condition, we have

$$q(\sigma) = \exp\Big(\mathrm{Tr}(Y^T M_\sigma) - 1\Big) = \exp\Big(\Omega^T M_\sigma\Big)$$

where $\Omega = Y - 1/n$. Thus implies that the distribution q is completely characterized by n^2 information coefficients Ω. So the dual objective of (ME) is therefore

$$\mathrm{Tr}\Big(Y^T H\Big) - \sum_{\sigma} \exp\Big(\mathrm{Tr}(Y^T M_\sigma) - 1\Big).$$ □

Gradient Based Optimization for the Maximum Entropy Problem. We now give a simple gradient descent algorithm to find the solution of the dual problem. Note that the gradient of the objective function is given by the matrix

$$G(Y) = H_{jk} - \sum_{\sigma} \exp\Big(\mathrm{Tr}(Y^T M_\sigma)\Big) - 1)(M_\sigma)_{jk} = H_{jk} - \exp(Y_{jk} - 1)\mathrm{perm}(\exp(\hat{Y}_{jk})).$$

Thus we can have a simple gradient descent algorithm, where at each iteration we find an optimal step length α such that the objective function values is improved, i.e.,

$$\mathrm{Tr}\Big((Y + \alpha G(Y)^T H\Big) - \sum_{\sigma} \exp\Big(\mathrm{Tr}((Y + \alpha G(Y))^T M_\sigma) - 1\Big) > \mathrm{Tr}\Big(Y^T H\Big) - \sum_{\sigma} \exp\Big(\mathrm{Tr}(Y^T M_\sigma) - 1\Big),$$

while the feasibility $Y + \alpha G(Y) \leq 0$ is still maintained. We note that the estimation of the gradient involves estimating the matrix permanent which we now discuss.

The pseudocode for the algorithm is given below.

4.3 Computation of the Matrix Permanent

We have shown that the problems of converting between the two above representations both require one to solve the matrix permanent problem, one of the prototypically #*P*-complete problems (even when all of the entries are binary [9]). The fastest known general exact algorithm is due to Ryser [12] based on the inclusion-exclusion formula.

Algorithm 1 Computing Information Coefficients Ω from Marginals Q

```
$Y \Leftarrow 0$
while $\|G(Y)\| \geq \varepsilon$ do
    Find an optimal step length $\alpha$
    $Y \Leftarrow Y + \alpha G(Y)$
end while
$\Omega \Leftarrow Y$
```

In some special cases, polynomial time algorithms exist for estimating the matrix permanent (e.g., for planar graphs [10]), but we have not found any such special cases to be applicable for general identity management problems.

When the entries of the matrix are non-negative, which is true in our setting there is an FPRAS (fully polynomial-time randomized approximation scheme) for approximating the permanent in probabilistic polynomial time [8, 9].

Finally, the fastest approximation that we are aware of is based on the Bethe free energy approximation [5, 18, 19] which frames the permanent problem as an inference problem in a graphical model, which can then be solved using loopy belief propagation.

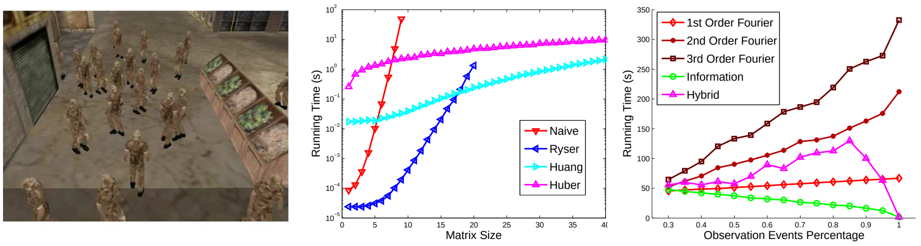
5 A Hybrid Approach for Identity Management

Using the conversion algorithms presented in the previous section, we now present a *hybrid identity management approach* in which we switch between the Fourier and information form domains depending on which domain is more convenient for certain inference operations. There are several issues that one must consider in designing a scheme for switching between the two domains. In this section, we present three simple switching strategies which we compare experimentally in Section 7.

We have argued that to handle mixing events, it is better to use a Fourier representation and that to handle evidence events, it is better to use the information form representation. A simple switching strategy (which we call the *myopic switching* scheme) thus *always* switches to either the Fourier or information form domain depending on whether it must perform a prediction/rollup operation or a conditioning operation.

In a similar spirit, we can also consider a *smoothness based switching* scheme, in which we switch based on the diffuseness of the distribution. In our implementation, we consider a heuristic in which we switch to a Fourier representation whenever the first-order matrix of marginals is within ε of a uniform matrix with respect to the Frobenius norm. Similarly, we switch to the information form representation whenever the first-order matrix comes within ε of some delta distribution.

What both the myopic and smoothness based approaches ignore, however, is the computational cost of switching between representations. To minimize this switching cost, we finally propose the *lagged block switching* scheme in which switching is only allowed to happen every k timeslices, where k is a parameter set by the user. In lagged block switching, we allow the identity management algorithm to lag the incoming data by k timesteps and therefore it can look ahead to see whether there are more mixing events or evidence events in the next k timesteps. As with myopic switching, the algorithm switches to Fourier if there are a majority of mixing events, and switches to information form if there are a majority of evidence events. After potentially switching, the algorithm processes the next k timesteps sequentially.



(a) (b) (c)

Fig. 1. (a) A view of the simulated data. (b) Running time comparison of different approaches in computing matrix permanent. (c) Comparing the running time of the three approaches.

6 An Adaptive Approach for Identity management

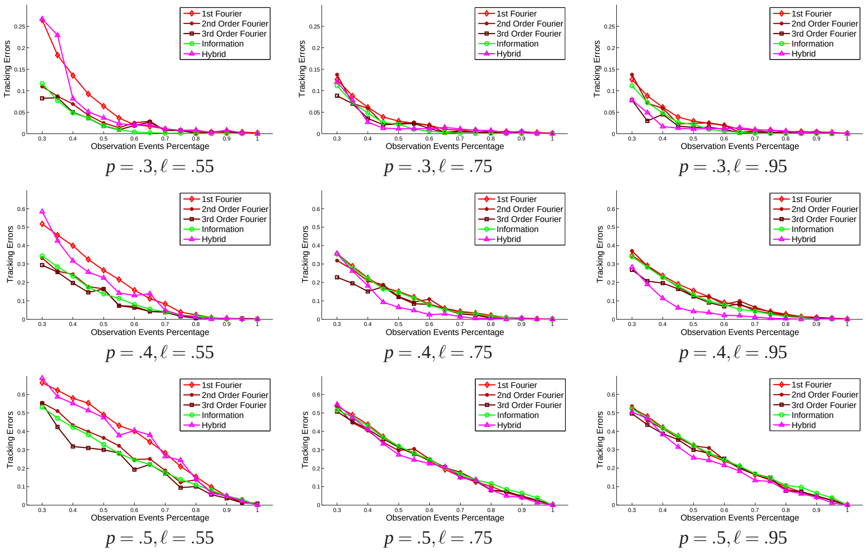
There are two extremal cases in the identity management problem: if we are completely uncertain about the assignment of target identities to tracks, then we have a uniform distribution over permutations, this smooth distribution can be represented compactly with Fourier coefficients; at the limit when we know the location of every identity, our distribution becomes very peaked, and we can use information coefficients to represent such a distribution compactly. In a real tracking scenario, we can pull highly certain or uncertain groups of targets out of a global Fourier or information representation and represent them separately, so that the problem breaks up into independent subproblems [7]. We now propose a method based on exploiting probabilistic independence of distributions over permutations, which can achieve significantly improved scalability.

Due to the mutual exclusivity constraints associated with permutations, we say the distribution $h(\sigma)$ has a independence factorization if there exists a subset X of identities and a subset Y of tracks, and also their corresponding complement subsets $\bar{X}$ and $\bar{Y}$, such that $h(\sigma)$ can be factorized into a product of two distributions over all mappings between X and Y and all mappings between $\bar{X}$ and $\bar{Y}$.

It turns out that whenever probabilistic independence holds, then both first order Fourier coefficients and information coefficients can be rendered block diagonal under an appropriate reordering of the rows and columns [7]. Since X and Y are unknown, our task is to find permutations of the rows and columns of the first order Fourier or information coefficients to obtain a block diagonal matrix. Viewing such a matrix as a set of edge weights on a bipartite graph between identities and tracks, we can approach the detection step as a biclustering problem with an extra balance constraint forcing $|X| = |Y|$. In practice, we use a cubic time SVD-based technique presented in [20] which finds bipartite graph partitions optimizing the normalized cut measure modified to satisfy the balance constraint. We note that such bipartite graph partitioning problems can be approached using either the ℓ_2 metric [20] or KL-divergence metric [2].

7 Experiments

In this section, we perform several experiments to compare the Fourier approach, information approach and the proposed hybrid approach. We use the Delta3D game engine to generate simulated crowds of up to 50 moving targets which walk around in an outdoor market [4]; Figure 1-(a) depicts a snapshot view of the simulated crowd. Such a simulation approach allows us to obtain accurate ground truth for large crowds than



$p = .3, \ell = .55$ $p = .3, \ell = .75$ $p = .3, \ell = .95$

$p = .4, \ell = .55$ $p = .4, \ell = .75$ $p = .4, \ell = .95$

$p = .5, \ell = .55$ $p = .5, \ell = .75$ $p = .5, \ell = .95$

Fig. 2. Comparing tracking accuracy of the three approaches with different parameters.

would be feasible in a typical physical testbed. The data contains interesting movement patterns and we can extract mixing and observation events directly from the data. We log a mixing event whenever two targets get within some distance of each other and an observation event whenever one target is separated from all the other targets for some distance. The percentages of observation events can be controlled by adjusting those distance parameters. We measure tracking errors using the fraction of mislabeled target identities over the tracks.

We first run an experiment for testing the running time performance of different algorithms in estimating the matrix permanent. As shown in Figure 1-(b), we generate random matrices and compare the running time of the four approaches: the naive method which sums up all products of matrix elements that lie in different rows and columns, the fastest known exact algorithm by Ryser [12], the Monte Carlo sampling algorithm by Huber et al. [8], and the loopy belief propagation algorithm by Huang et al. [5]. The naive approach has a super-exponential complexity and the Ryser's formula has an exponential complexity, thus, they scale poorly as the matrix size grows; On the other hand, the two randomized approximate algorithms have much better running time performance than the exact algorithms. In the hybrid algorithm, we use Monte Carlo sampling algorithms [8]. for estimating the matrix permanent.

In our experiments, we can control two sets of parameters which determine the tracking quality, one is the swapping probability — if we can keep track of who is who when two targets mix with high probability during the prediction/rollup operation, we can achieve better tracking performance; the other is the likelihood function, if the likelihood for observing the identity of a target is high, then conditioning step can resolve the ambiguities better. We set up nine cases to explore the tracking accuracy with different swapping probability and likelihood function parameters. As depicted in Figure 2, the probability p characterizing confusions of the mixing events grows larger

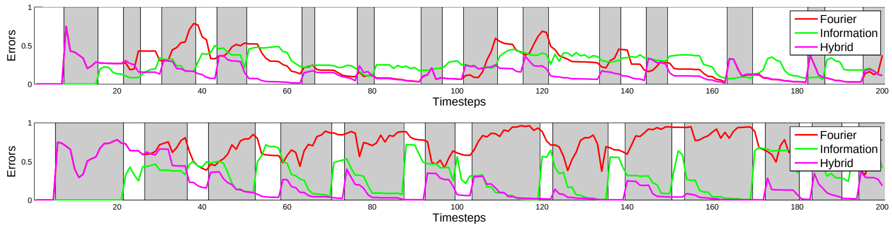
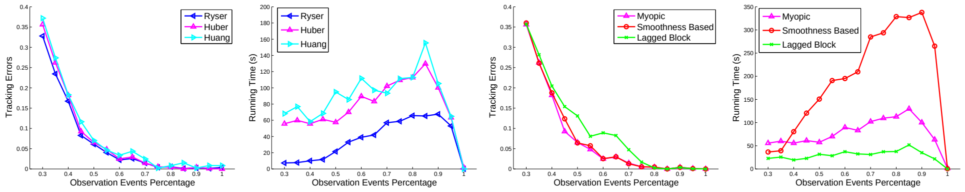


Fig. 3. Compare the errors in distribution of the three approaches. The white intervals denote the rollup steps and the grey intervals denote the conditioning steps.



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

Fig. 4. (a,b) Tracking accuracy and running time of the hybrid approach with different algorithms for estimating matrix permanent. (c,d) Tracking accuracy and running time of the hybrid approach with different rules for switching.

from left to right, and the likelihood ℓ for observing target identity grows larger from top to bottom. We can get better tracking accuracy if p is small or ℓ is large.

From Figure 2, we can see that the information approach outperforms the Fourier approach in most cases, while the Fourier approach can beat the information approach only slightly in some cases, e.g., the case $p = .5, \ell = .75$ where the mixings are quite confusing. The tracking accuracy can be improved if we incorporate high order Fourier coefficients. We can achieve better performances in lots of cases if we use the hybrid approach, whose tracking accuracy are comparable to the 2nd order or even 3rd order Fourier approach. The running time for those approaches are shown in Figure 1-(c). In general, the Fourier approach has a fundamental trade-off between tracking complexity in terms of the number of coefficients used and the tracking accuracy: we can improve tracking accuracy by using more coefficients. The hybrid approach makes a good balance which can improve tracking accuracy when there are observation events that confirm the target identities (large ℓ) with moderate running time. We can see that the running time for the hybrid approach is strictly less than the second order Fourier approach. This is because the complexity for the conditioning step in the Fourier domain is very expensive if we use high order Fourier coefficients.

We also compare the errors of approximating the distribution over permutations of the three approaches (see Figure 3). It turns out that the Fourier approach decrease (increase) the errors during the rollup (conditioning) steps, while the information approach decrease (increase) the errors during the conditioning (rollup) steps. However, if we use the hybrid approach, we can always keep the errors at a much lower level.

We also compare the tracking accuracy and runing time of the hybrid approach by varing the algorithms for estimating the matrix permanent, as well as varying the strategies for switching between two domains. Specifically, we compare Ryser, Huber, and the LBP algorithms for estimating matrix permanents. The tracking accuracy of those approaches does not differ too much. However, the two approximation algorithms (by

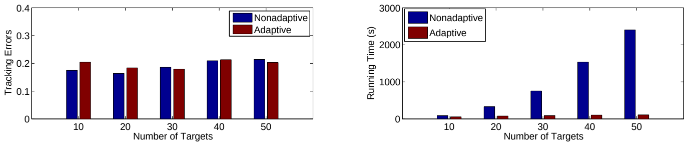


Fig. 5. Tracking accuracy and running time of the adaptive approach compared with the nonadaptive approach.

Huber and the loopy belief propagation method) have longer running times for the small scale experiments because it takes longer time to converge in solving the maximum entropy problem (see Figure 4-(a,b)). We also evaluate our three different switching strategies for the hybrid approach. Compared with the smoothness based switching strategy which switches 38 times, the lagged block strategy switches between two domains 91 times among the 1000 timesteps while can not improve the tracking accuracy too much and take very long running time. The myopic strategy suffers a little on the tracking accuracy while the running time can be improved because there are only 20 times of switchings (see Figure 4-(c,d)).

We finally evaluate the performance of the adaptive approach. As depicted in Figure 5, the tracking accuracy for the adaptive approach is comparable to the nonadaptive approach, while the running time can always be controled using the adaptive approach. In particular, the tracking accuracy for the adaptive approach is often worse than the nonadaptive approach when the number of targets is small because it is usually difficult to factorize the problem in those cases. When the number of targets is larger, however, the benefit of adaptive approach becomes more evident in both tracking accuracy and complexity.

8 Conclusion

In this paper we compare the computational advantages and disadvantages of two popular distributional representations for the identity management problem. We show that the two approaches are complementary - the Fourier representation is closed under prediction operations and is thus better suited for handling problems with high uncertainty while the information form representation is closed under conditioning operations and is better suited for handling problems in which a lot of observations are available. As our experiments show, using a combination of both approaches seems to often be the best approach. While converting between the two representations is a #P-hard problem in general, we show that with some of the modern permanent approximation algorithms, conversion is tractable and yields surprisingly good performance in practice.

We have focused primarily on the first-order versions of both the Fourier and information form approximations. It would be interesting to develop high order analysis with the hybrid approach.

9 Acknowledgement

The authors would like to thank Prof. Yinyu Ye for the helpful discussion. Thanks also to Kyle Heath for providing experimental data. Leonidas Guibas and Xiaoye Jiang wish to acknowledge the support of ARO grants W911NF-10-1-0037 and W911NF-07-2-0027, as well as NSF grant CCF 1011228 and a gift from the Google Corporation. Leonidas Guibas and Jonathan Huang acknowledge the support of grant ONR MURI N000140710747.

References

1. S. Agrawal, Z. Wang, and Y. Ye. Parimutuel betting on permutations. In *the Workshop on Internet and Network Economics (WINE)*, pages 126–137, 2008.
2. I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98, 2003.
3. P. Diaconis. *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics, 1988.
4. K. Heath and L. J. Guibas. Multi-person tracking from sparse 3d trajectories in a camera sensor network. In *Proceedings of IEEE ICDSC*, 2008.
5. B. Huang and T. Jebara. Approximating the permanent with belief propagation. *Computing Research Repository*, 2009.
6. J. Huang, C. Guestrin, and L. J. Guibas. Fourier theoretic probabilistic inference over permutations. *Journal of Machine Learning Reserach (JMLR)*, 10:997–1070, 2009.
7. J. Huang, C. Guestrin, X. Jiang, and L. J. Guibas. Exploiting probabilistic independence for permutations. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.
8. M. Huber and J. Law. Fast approximation of the permanent for very dense problems. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 681–689, 2008.
9. M. Jerrum, A. Sinclair, and E. Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries. In *ACM Symposium on Theory of Computing*, pages 712–721, 2001.
10. P. W. Kasteleyn. The statistics of dimers on a lattice. i. the number of dimer arrangements on a quadratic lattice. *Physica*, page 12091225, 1961.
11. R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.
12. H. Ryser. *Combinatorial Mathematics - The Carus Mathematical Monographs Series*. the Mathematical Association of America, 1963.
13. B. Sagan. *The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions*. Springer-Verlage, 2001.
14. B. Schumitsch, S. Thrun, G. Bradski, and K. Olukotun. The information-form data association filter. In *Proceedings of the Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2005. MIT Press.
15. J.-P. Serre. *Linear Representation of Finite Groups*. Springer-Verlag, 1977.
16. J. Shin, L. J. Guibas, and F. Zhao. A distributed algorithm for managing multi-target identities in wireless ad-hoc sensor networks. In *Proceedings of the International Conference on Information Processing in Sensor Networks (IPSN)*, pages 223–238, 2003.
17. J. Shin, N. Lee, S. Thrun, and L. J. Guibas. Lazy inference on object identities in wireless sensor networks. In *Proceeings of the International Conference on Information Processing in Sensor Networks (IPSN)*, 2005.
18. P. Vontobel. The bethe permanent of a non-negative matrix. In *Proceedings of the Allerton Conference on Communications, Control, and Computing*, 2010.
19. Y. Watanabe and M. Chertkov. Belief propagation and loop calculus for the permanent of a non-negative matrix. *Journal of Physics A: Mathematical and Theoretical*, 43(24):242002, 2010.
20. H. Zha, X. He, C. Ding, H. Simon, and M. Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 25–32, 2001.