Efficient Data Association and Uncertainty Quantification for Multi-Object Tracking

David S. Hayden, Sue Zheng, John W. Fisher III
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02142
{dshayden, szheng, fisher}@csail.mit.edu

Abstract

Robust data association is critical for analysis of long-term motion trajectories in complex scenes. In its absence, trajectory precision suffers due to periods of kinematic ambiguity degrading the quality of follow-on analysis. Common optimization-based approaches often neglect uncertainty quantification arising from these events. Consequently, we propose the Joint Posterior Tracker (JPT), a Bayesian multi-object tracking algorithm that robustly reasons over the posterior of associations and trajectories. Novel, permutation-based proposals are crafted for exploration of posterior modes that correspond to plausible association hypotheses. JPT exhibits more accurate uncertainty representation of data associations with superior performance on standard metrics when compared to existing baselines. We also show the utility of JPT applied to automatic scheduling of user-in-the-loop annotations for improved trajectory quality.

1 Introduction

In multi-object tracking the trajectories of an unknown number of objects are estimated from noisy observations over time. Assigning observations to objects is known as the *data association problem*. The complexity of the multidimensional assignment formulation is factorial in the number of observations at each time and exponential in the number of timesteps making it NP-hard [8].

All approaches to multi-object tracking reason over data associations, but *few* represent uncertainty explicitly, much less make it available for subsequent tasks. Traditional applications in security [4], surveillance [23], sensor networks [30] and robotic localization [22] favor real-time performance. More recently, there is increased interest in the use tracking for follow-on decision making and analysis including the creation of gold-standard datasets [36], sports analytics [12], study of animal behavior [41] and cellular dynamics [2]. Such applications benefit significantly from *accurate* representations of uncertainty to inform subsequent analysis.

In response, we develop the Joint Posterior Tracker (JPT), a Bayesian multi-object tracker that emphasizes *joint* uncertainty over association hypotheses and object trajectories. We present model details in Section 2. In Section 3 we develop a novel Metropolis-Hastings inference procedure that, under some model parameterizations, generalizes the Extended-HMM proposals of [21]. JPT inference is exact, efficient and not limited by gating heuristics. In Section 4, we show that JPT explores posterior modes much more completely and efficiently with superior performance on standard multi-object tracking metrics on scientific and sports datasets as compared to a standard baseline. Finally, we show that JPT enables the automatic scheduling of a small number of disambiguations that facilitate rapid improvement in trajectory quality.

Preprint. Under review.

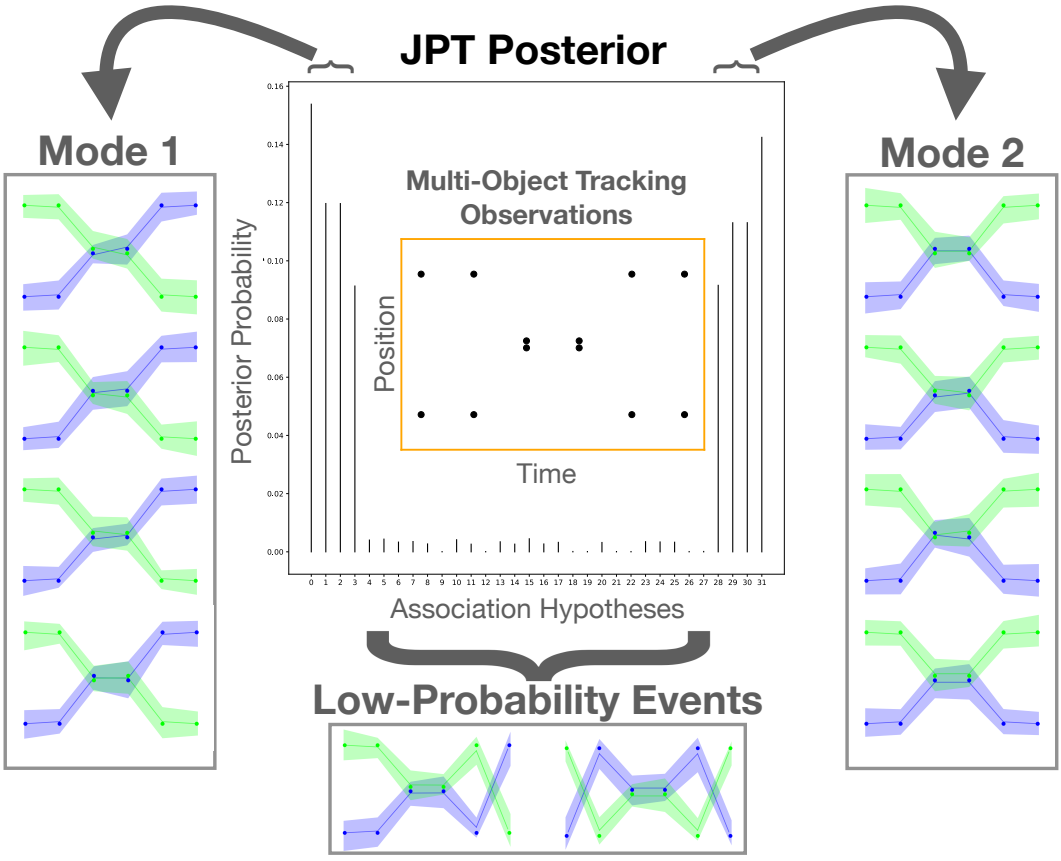


Figure 1: Multi-object tracking observations over time (orange box, middle) where two objects (green, blue) begin separated–briefly converge–then diverge. Such ambiguities are represented as distinct modes (left, right) which the Joint Posterior Tracker excels at exploring while avoiding low-probability events (bottom).

Figure 1 illustrates JPT reasoning over multimodal uncertainty. Two objects begin well separated, become kinematically ambiguous, then separate. Absent additional information, it is unclear whether or not their paths crossed. Evaluation of association events shows that the resulting JPT posterior accurately captures trajectory uncertainty via two modes–one for crossing, one for not. *Within* each mode, associations switch when objects are close but inferred trajectories remain similar whereas *between* modes, associations switch in ways that dramatically impact the inferred trajectories (and hence subsequent analysis).

Related Work Approaches to multi-object tracking can be distinguished in several ways. First, whether they process measurements one frame at a time (single-scan, [31, 17, 9]), multiple frames at a time (multi-scan, [18]) or all at once (batch, [10, 24, 37]). Second, whether they construct point estimates (as in optimization), [18, 14, 37] or entertain multiple solutions (sampling or variational methods, [31, 35]). Third, whether or not they employ gating heuristics that restrict possible hypotheses [18, 24]. JPT is a batch, sampling-based tracker with no gating heuristics that reasons over the joint distribution of an unknown number of objects, their trajectories and the association of objects to observations.

Many recent approaches to multi-object tracking focus on sophisticated appearance, motion or shape modeling in an optimization-based framework [14, 19, 18, 39, 10]. While providing a single point-estimate, they forego representing uncertainty in assignments, and do not permit recovery from errors. Regardless of the quality of the appearance model, errors are certain to occur in complex scenes.

Monte-Carlo approaches represent uncertainty via sampled realizations. These include [31], [17], [9], but each are single-scan, filtering-based approaches, that do not incorporate future information.

	JPT	MCMCDA [24]	Var [35]	BP [38]
Uncertainty	✓	✓	✓	✓
Exact Posterior	✓	✓		
General MOT	✓	✓	✓	

Table 1: Multi-Object Trackers capable of quantifying uncertainty in data association. JPT (Joint Posterior Tracker): our method; MCMCDA (Markov Chain Monte Carlo Data Association); Var (Variational Tracker); BP (Belief Propagation Tracker).

Table 1 summarizes related works that represent some degree of uncertainty, such as marginal uncertainty using belief propagation [38] and approximate uncertainty using variational methods [35]. The former treats a fixed number of objects, while the latter samples from a variational approximation rather than the true posterior.

Markov-Chain Monte Carlo Data Association (MCMCDA) and its variants [24, 4, 13] is most closely related to JPT. MCMCDA employs sampling and can be run online or in batch. It represents posterior uncertainty in data association, but does not incorporate information from the future (it performs filtering as opposed to smoothing or joint inference, even when running in batch). Furthermore, MCMCDA precomputes data structures that heavily rely on gating heuristics. This precludes uncertainty representation of some association hypotheses. Nevertheless, it is the only tracker the authors are aware of that can, in principle, represent posterior uncertainty, samples from its posterior exactly (modulo gating heuristics) and treats the general multi-object tracking problem. As such, we use batch MCMCDA as a baseline for comparison to JPT.

2 Bayesian Multi-Object Tracking

The multi-object tracking problem is to partition a set of observations across time into collections of objects such that every observation must be assigned and no two objects can claim the same observation. In its most general formulation, there can be clutter (false-positives), missing detections, unknown number of objects, and arbitrary object arrival and departure times. Multi-object tracking can be formulated in several ways; we discuss this further in Appendix A and define the multidimensional assignment formulation as it is most closely related to JPT.

JPT defines a joint distribution on trajectories and assignments whereas MCMCDA defines a posterior on assignments alone. Both operate in batch, but a key difference is that MCMCDA *only* considers past information when estimating trajectories via filtering. In contrast, JPT samples from a joint distribution over associations and trajectories that accounts for past and future information. Hence, there is a distribution over trajectories for any set of associations.

To reason over a joint distribution on trajectories and associations, we must define a generative model for observations $y = \{y_1, \dots, y_T\}$ over all times $1, \dots, T$ where $y_t = \{y_{tn}\}_{n=1}^{N_t}$. Vector-valued observation y_{tn} is the n^{th} observation at time t and has dimension D_y. N_t is the total number of observations at time t.

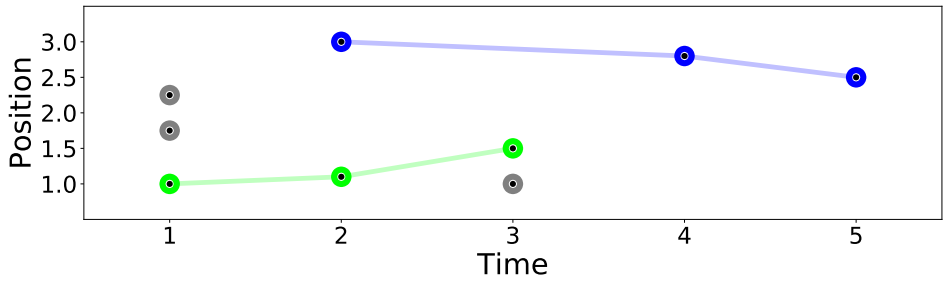
Associations z define a partitioning of y into objects and clutter and trajectories x are the latent states of objects over all times. For clarity of exposition since our goal is accurate representation of posterior uncertainty, we only model object locations. One can include a shape or appearance model without modification of *any* equation in this work. Next, we define the latent representation and generative model for JPT. Throughout, Figure 2 can be used to ground definitions in a toy example.

2.1 Event Counts $p(M)$

JPT explicitly models clutter (false-positive) and missing detections, as well as arbitrary arrival and departure times for an unknown number of objects. At each time t, counts of new object arrivals a_t, clutter observations f_t, existing object detections d_t and existing object departures λ_t are modeled as

$$a_t \sim \text{Pois}(a_t \mid \lambda_b) \quad f_t \sim \text{Pois}(f_t \mid \lambda_f) \quad d_t \sim \text{Bin}(d_t \mid e_{t-1}, p_d) \quad \lambda_t \sim \text{Bin}(\lambda_t \mid d_t, p_\lambda) \tag{1}$$

where $e_0 = d_0 = 0$ and $e_t = e_{t-1} + a_t - \lambda_t$ are counts of existing objects (those that arrived at some time $t' \leq t$ and have not yet departed). Prior parameters λ_b, λ_f are the new object arrival and



Time	x_t	z_t	a_t	f_t	d_t	λ_t
1	1.0, $\emptyset$	1, 0, 0	1	2	0	0
2	1.3, 3.0	1, 2	1	0	1	0
3	1.5	0, 1	0	1	1	1
4	$\emptyset$, 2.7	2	0	0	1	0
5	$\emptyset$, 2.5	2	0	0	1	1

Figure 2: JPT's Latent Representation. (**Top**): Observations y (circles) are colored by their association (green/blue for either of two objects, gray for clutter) and connected by sampled trajectories. (**Bottom**): Trajectories x, associations z and counts $M = \{a_t, f_t, d_t, \lambda_t\}_{t=1}^T$. Objects are observed at distinct arrival and departure times. Trajectories are length-T, padded by $\emptyset$ before arrival and after departure. We marginalize over states with missing detections (e.g. blue at $t = 3$).

false alarm rates and p_d, p_λ are the detection and departure probabilities for existing objects. Every object is assumed to be observed at least twice: when it arrives and when it departs. Denote the set of all event counts as $M = \{M_1, \ldots, M_T\}$ where $M_t = \{a_t, f_t, d_t, \lambda_t\}$. From Eqn. 1, the generative model for latent counts M is,

$$p(M) = \prod_{t=1}^{T} p(M_t \mid M_{t-1}) \quad = \prod_{t=1}^{T} p(a_t)\, p(f_t)\, p(d_t \mid e_{t-1})\, p(\lambda_t \mid d_t). \tag{2}$$

2.2 Associations $p(z \mid M)$

JPT represents the association of each observation to an integer-labeled object or clutter. Let the association of observation y_{tn} be the latent random variable $z_{tn} \in \mathbb{Z}^+ \cup \{0\}$ where $z_{tn} = k > 0$ if y_{tn} is associated to target k at time t and $z_{tn} = 0$ if y_{tn} is associated to clutter. Define an association hypothesis as the set of all associations $z = \{z_1, \ldots, z_T\}$ for $z_t = \{z_{tn}\}_{n=1}^{N_t}$.

Conditioned on event counts M, association hypotheses z have a uniform prior over the space of possible associations subject to constraints that enforce that all observations are either associated to an object or clutter (satisfied by definition of z_{tn}), that an object claim at most one observation at each time t (first constraint in Eqn. 3) and that associations be consistent with event counts M (remaining constraints in Eqn. 3):

$$p(z \mid M) \propto 1 \text{ if } \begin{cases} |\{n : z_{tn} = k\}| \leq 1 \quad \forall k > 0, \forall t \\ f_t = |\{n : z_{tn} = 0\}| \quad \forall t \\ a_t = |\{k > 0 : z_{tn} = k \text{ and } z_{t'n} \neq k \text{ for all } t' < t\}| \quad \forall t \\ \lambda_t = |\{k > 0 : z_{tn} = k \text{ and } z_{t'n} \neq k \text{ for all } t' > t\}| \quad \forall t \\ a_t + d_t = |\{n : z_{tn} > 0\}| \quad \forall t \end{cases} \tag{3}$$

Association hypotheses that do not satisfy these constraints have zero probability. We note that the space of possible associations is exponential in time T and factorial in the number of observations N_t

at each time t [26]. Reasoning over this large space is *the* fundamental challenge in data association. Doing so while satisfying these constraints makes inference difficult, as we discuss in Section 3.

2.3 Dynamics, Observations $p(x \mid z)\, p(y \mid x, z)$

Denote the trajectory of object $k > 0$ at time t by the latent random variable $x_{tk} \in \mathbb{R}^{D_x}$ for D_x the dimension of the latent states. Denote all trajectories as $x = \{x_1, \ldots, x_T\}$ where $x_t = \{x_{tk}\}_{k=1}^{K(z)}$ for $K(z)$ the number of objects in association hypothesis z. Every object has a length-T latent trajectory, but is represented by $x_{tk} = \emptyset$ for any time before its arrival or after its departure. Let $t' = t - 1$ and define the dynamics model for objects $1, \ldots, K(z)$ as:

$$p(x \mid z) = \prod_{t=1}^{T} p(x_t \mid x_{1:t-1}, z) = \prod_{t=1}^{T} \prod_{k=1}^{K(z)} p(x_{tk} \mid x_{t'k}) \tag{4}$$

Define the observation model as,

$$p(y \mid x, z) = \prod_{t=1}^{T} p(y_t \mid z_t, x_t) = \prod_{t=1}^{T} \prod_{n=1}^{N_t} p(y_{tn} \mid z_{tn}, x_t) \tag{5}$$

We now specialize Eqns. 4, 5 to a linear Gaussian system, as is common in tracking. For the dynamics,

$$p(x_{tk} \mid x_{t'k}) = \begin{cases} \mathrm{N}\left(x_{tk} \mid F x_{t'k}, Q\right) & \text{if } x_{t'k} \neq \emptyset \\ \mathrm{N}\left(x_{tk} \mid \mu_0, \Sigma_0\right) & \text{o.w.} \end{cases} \tag{6}$$

The first line is a linear Gaussian system with system model F and noise covariance Q. The second line specifies a shared prior on trajectories with prior parameters μ_0, Σ_0 that are typically set to be broad over the observation space. JPT marginalizes over missing detections (i.e., times when an object has already arrived but has no association). This can be computed in closed-form for linear Gaussian dynamics.

For the observation model,

$$p(y_{tn} \mid z_{tn} = k, x_t) = \begin{cases} \mathrm{N}\left(y_{tn} \mid H x_{tk}, R\right) & \text{if } k > 0 \\ \mathrm{N}\left(y_{tn} \mid \mu_{\text{FP}}, \Sigma_{\text{FP}}\right) & \text{o.w.} \end{cases} \tag{7}$$

The first line is a linear Gaussian system with observation projection H and observation noise covariance R. The second line specifies the model for clutter detections with prior parameters $\mu_{\text{FP}}, \Sigma_{\text{FP}}$, which are also typically set to be broad.

2.4 Joint Distribution

Finally, the joint posterior over trajectories x, associations z and counts M given observations y is,

$$p(x, z, M \mid y) = \frac{1}{Z} p(M)\, p(z \mid M)\, p(x \mid z)\, p(y \mid x, z) \tag{8}$$

where each term on the RHS is respectively given by Equations 2, 3, 4, 5 and Z is an intractable normalization constant (owing to the exponential and factorial number of terms). We show how to draw samples from this non-trivial posterior using Metropolis-Hastings proposals in Section 3.

3 JPT Inference

Sampling from the posterior in Eqn. 8 is complicated by the constraints in Eqn. 3 and the exponential in T and factorial in N_t scaling of possible association hypotheses. With no analytic form and computationally infeasible enumeration of all hypotheses, we turn to Metropolis-Hastings [15].

The Metropolis-Hastings (MH) algorithm enables sampling from intractable distributions by constructing a Markov chain whose unique stationary distribution is the desired distribution. Samples from this chain converge in distribution to the desired distribution, regardless of starting state. MH constructs transition distributions q^* that maintain detailed balance,

$$\frac{p(x', z', M' \mid y)}{p(x, z, M \mid y)} = \frac{q^*(x', z', M' \mid x, z, M, y)}{q^*(x, z, M \mid x', z', M', y)} \tag{9}$$

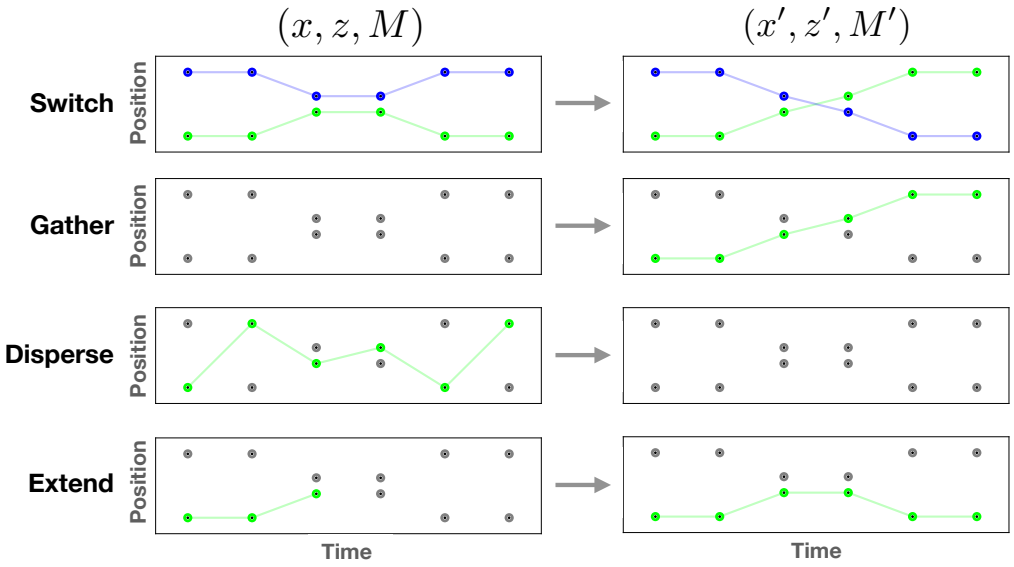


Figure 3: Examples of each JPT proposal. Left column is the input state (x, z, M) and right column is the output state (x', z', M'). Black points are observations y, encircled in the color of their association (green or blue for objects; grey for clutter). Example trajectories x are visualized as colored lines. Switch proposals can reason over many objects, but are shown here for two.

resulting in a chain where Equation 8 is a stationary distribution. MH accepts a proposed sample (x', z', M') from an arbitrary proposal distribution q with probability $\min(1, R)$,

$$R = \frac{p(x', z', M' \mid y)}{p(x, z, M \mid y)} \frac{q(x, z, M \mid x', z', M', y)}{q(x', z', M' \mid x, z, M, y)}. \tag{10}$$

where the normalizers cancel. In Section 3.1, we design proposals that rapidly explore high-probability regions in the JPT posterior, hopping between different modes as demonstrated in Figure 1 and later quantified in Section 4. We then describe closed-form Gibbs sampling of joint trajectories conditioned on associations in Section 3.2.

3.1 Proposals

We design Metropolis-Hastings proposals that make large moves in the latent space (including mode hopping) by reasoning over permutations of the latent state over time. JPT proposals are data-dependent–they make use of the observations y and current state (x, z, M) in proposing next state (x', z', M'). Data-dependent proposals are complicated but in our case avoid random exploration and the use of gating heuristics while retaining tractability. Broadly, JPT proposals reason over assignment and trajectory permutations between existing objects (Switch 3.1.1), between a new object and clutter (Gather 3.1.2, Disperse 3.1.3) or between existing objects and clutter (Extend, 3.1.4). Of these, the Switch proposal is a novel generalization of [21] and contributes most to JPT's exploration of posterior modes; we thus focus on it more than the other proposals. Pictorial examples for each proposal transitioning from state (x, z, M) to state (x', z', M') are shown in Figure 3.

3.1.1 Switch Proposal

Switch proposals consider possible trajectory and associations permutations between existing objects, and are sampled according to JPT's dynamics and observation models (Eqns. 4, 5). They cause rapid exploration of different posterior modes such as the ones shown in Figure 1. Strikingly, the Switch proposal is in many cases automatically accepted ($R_{\text{switch}} = 1$).

Following Algorithm 1, the Switch proposal samples uniformly at random a subset $\mathcal{K}$ of existing objects $\{1, \ldots, K(z)\}$ such that $2 \leq |\mathcal{K}| \leq \bar{\mathcal{K}}$ (Line 2) for $\bar{\mathcal{K}}$ a maximum size, discussed below.

Let σ_t be a *valid permutation* on objects $\{1, \ldots, K(z)\}$ at time t. Valid permutations do not permute objects outside of $\mathcal{K}$: for all $k \notin \mathcal{K}, \sigma_t(k) = k$. With slight abuse of notation, let $\sigma_t(x_t)$ and $\sigma_t(z_t)$

Algorithm 1: Switch Proposal

Input : x, z, M, y
Output : x', z', M'

1 Let $x' = x,\ z' = z$
2 Sample object set $\mathcal{K} \subset \{1, \ldots, K(z)\}$ s.t. $|\mathcal{K}| \geq 2$
3 Define switch times $\tau = \{t : z_{tn} = k \text{ for any } k \in \mathcal{K}\}$
4 Set permutations σ_t as the identity permutation on $(1, \ldots, K(z))$ for any $t \notin \tau$
5 **for** $t \in \tau$ *in order* **do**
6 | Sample valid permutation $p(\sigma_t \mid \sigma_{1:t-1}) \propto p\left(\sigma_t(x_t) \mid \sigma_{1:t-1}(x_{1:t-1})\right)\ p\left(y_t \mid \sigma_t(x_t)\right)$
7 | Let $x'_t = \sigma_t(x_t),\ z'_t = \sigma_t(z_t)$
8 **end**
9 Compute counts M' from z'
10 **if** $M' = M$ *or* $\mathit{rand}(0, 1) < \min(1, R_{\mathit{switch}})$ return x', z', M'
11 **else** return x, z, M

respectively represent the trajectory values and associations permuted according to σ_t. So for time t, the trajectory value x_{tk} (possibly an uninstantiated value) and association (possibly none) of object k become the trajectory value and association of object $\sigma_t(k)$. Define $\sigma_{1:t}(x_{1:t})$ over times $1, \ldots, t$ as $x'_{1:t}$ where $x'_t = \sigma_t(x_t)$.

The Switch proposal only considers permutations at times when at least one object $k \in \mathcal{K}$ has been observed. Let τ be all such times (Line 3). For any time $t \notin \tau$, set σ_t as the identity permutation, $\sigma_t(k) = k$ (Line 4).

For increasing time $t \in \tau$, iteratively sample permutation σ_t conditioned on the previously-sampled permutations $\sigma_{1:t-1}$ with probability proportional to the product of the observation and dynamics models (Equations 4, 5) evaluated with the appropriate swaps in trajectory and association values imposed by permutations $\sigma_{1:t}$ (Line 6). There are $|\mathcal{K}|!$ possible values for σ_t at each time t, but we find $\bar{\mathcal{K}} = 7$ balances efficient computation and posterior exploration.

After sampling σ_t for all $t \in \tau$, we compute new counts M' from the permuted associations z' (Line 9) and the Hastings ratio (Line 10) between (x', z', M') and (x, z, M), noting that Switch proposals are their own reverse move. In Appendix B.1, we show that:

$$R_{\text{switch}} = \frac{\prod_{t=1}^{T} p(M'_t \mid M'_{t-1})}{\prod_{t=1}^{T} p(M_t \mid M_{t-1})}. \tag{11}$$

The Switch proposal is always accepted ($R_{\text{switch}} = 1$) whenever $M' = M$. This occurs in several situations: when the number of objects are known in advance, when objects are assumed never to depart, when there are no missing observations and when all $k \in \mathcal{K}$ are observed at $\max \tau$. In many scientific and sports analytics applications, it is common for subjects to never depart. When these conditions don't hold, the event counts and Hastings ratio are efficiently evaluated (linear in time T and parallelizable) by only considering terms where the counts M', M differ. Switch proposals have complexity $\mathcal{O}(|\mathcal{K}|!\, T)$ where the factorial dependence on $|\mathcal{K}|$ comes from Line 6 and the linear dependence on T comes from its enclosing for loop. In practice, we limit the subset size $|\mathcal{K}| \leq \bar{\mathcal{K}}$.

In Appendix E, we show that Switch proposals generalize the Extended HMM proposals of [21] by proposing a discretization that depends on the current latent state (in their nomenclature, JPT "pool states" are permutations of x, z). In their work, sampled discretizations (or pool states) cannot depend on the current latent state, else detailed balance is lost.

3.1.2 Gather Proposal

Following Algorithm 2, the Gather proposal considers the formation of a new object $k = 1 + K(z)$ (Line 2) from the set of clutter-associated observations $\{y_{tn} : z_{tn} = 0\}$. Its reverse move is Disperse (3.1.3). Let τ_0 be the set of times t with at least one clutter association (Line 3). For increasing $t \in \tau_0$, assignments to object k are iteratively sampled either among observations that are currently associated to clutter or, with probability $\delta = 0.01$, no clutter association to allow for missing observations. In the former, association $z'_{tn} = k$ is sampled among all clutter observations with probability proportional to Line 6 where $t' = t - 1$ and marginalization occurs between states with missing associations.

Algorithm 2: Gather Proposal

Input : x, z, M, y
Output : x', z', M'

1 Let $x' = x,\ z' = z$
2 Let $k = 1 + K(z)$
3 Define gather times $\tau_0 = \{t : z_{tn} = 0 \text{ for any } 1 \leq t \leq T\}$
4 **for** $t = \min \tau_0, \ldots, \max \tau_0$ **do**
5 | **if** *rand*$(0, 1) < \delta$ continue
6 | Sample $p(z'_{tn} = k) \propto p(y_{tn} | x_{t'k}, z'_{tn} = k)\ \mathbb{I}(z_{tn} = 0)$
7 | Sample $p(x'_{tk} \mid x'_{t'k}, z'_{tn} = k) \propto p(x'_{tk} \mid x_{t'k})\ p(y_{tn} \mid x'_t, z'_{tn} = k)$
8 **end**
9 Compute counts M' from z'
10 **if** *rand*$(0, 1) < \min(1, R_{gather})$ return x', z', M'
11 **else** return x, z, M

Conditioned on the sampled assignment $z'_{tn} = k$, a trajectory value x'_{tk} is sampled (Line 7); this is analytic in the linear Gaussian case. Sampling the association then the trajectory allows unambiguous evaluation of the reverse move in the Hastings ratio (Line 10) as derived in Appendix B.2. Gather always requires an accept/reject step and has complexity $\mathcal{O}(T \prod_{t=1}^{T} N_t)$, where the linear complexity in N_t is a consequence of Lines 6–7 and linear complexity in T comes from the enclosing for loop.

3.1.3 Disperse Proposal

Algorithm 3: Disperse Proposal

Input : x, z, M, y
Output : x', z', M'

1 Let $z' = z$
2 Sample $k \in \{1, \ldots, K(z)\}$
3 Set $z'_{tn} = 0$ for all t, n such that $z_{tn} = k$
4 Let $x' = x \setminus \{x_{tk}\}_{t=1}^{T}$
5 Compute counts M' from z'
6 **if** *rand*$(0, 1) < \min(1, R_{disperse})$ return x', z', M'
7 **else** return x, z, M

Following Algorithm 3, the Disperse proposal simply chooses an existing object at random (Line 2), removes all its associations by setting them to clutter (Line 3) and deletes the trajectory values for that object (Line 4). It is the reverse move for the Gather proposal. Hence, $R_{\text{disperse}} = R_{\text{gather}}^{-1}$ where R_{gather} is defined in Equation 20. As in the case for the Gather proposal, an accept/reject step is required. Disperse has constant complexity.

3.1.4 Extend Proposal

The Extend proposal is similar to the Gather proposal but rather than consider permutations between clutter associations and a new object, it considers permutations between clutter associations and an existing object. Effectively, this allows an existing object to resample associations.

Following Algorithm 4, randomly sample object k from existing objects (Line 2) and iterate over all times $t \in \tau_k$ with an association to clutter $z_{tn} = 0$ or to the current object $z_{tn} = k$ (Line 3). As in the Gather proposal, skip a resampling of assignments at time $t \in \tau_k$ with probability δ. Otherwise, sample an association then a trajectory value (Lines 6-7) with definitions as in Gather, except that it is possible for $z'_{tn} = z_{tn}$ for some times t (it resamples the same assignment it already had).

By automatically rejecting any Extend proposal that leaves a object with fewer than two observations, we can ensure that Extend proposals are always their own reverse move. As in Gather, the observation and dynamics terms cancel in the posterior ratio for all objects other than k, but an accept/reject step must still be computed, and is of similar form to the Gather proposal. Like Gather, Extend

Algorithm 4: Extend Proposal

Input : x, z, M, y
Output : x', z', M'

1 Let $x' = x,\ z' = z$
2 Sample $k \in \{1, \ldots, K(z)\}$
3 Define extend times $\tau_k = \{t : z_{tn} \in \{0, k\} \text{ for any } 1 \leq t \leq T\}$
4 **for** $t = \min \tau_k, \ldots, \max \tau_k$ **do**
5 | **if** *rand*$(0, 1) < \delta$ continue
6 | Sample $p(z'_{tn} = k) \propto p(y_{tn} | x_{t'k}, z'_{tn} = k)\ \mathbb{I}(z_{tn} \in \{0, k\})$
7 | Sample $p(x'_{tk} \mid x'_{t'k}, z'_{tn} = k) \propto p(x'_{tk} \mid x_{t'k})\ p(y_{tn} \mid x'_t, z'_{tn} = k)$
8 **end**
9 Compute counts M' from z'
10 **if** *rand*$(0, 1) < \min(1, R_{extend})$ return x', z', M'
11 **else** return x, z, M

has complexity $\mathcal{O}(T\ \prod_{t=1}^{T} N_t)$ where linear complexity in N_t comes from Lines 6–7 and linear complexity in T comes from the enclosing for loop.

3.2 Forward-Filtering, Backward Sampling

Joint sampling of trajectories from the full conditional,

$$p(x \mid z, M, y) = p(x \mid z, y) \tag{12}$$

constitutes a fifth MH proposal in the form of a Gibbs sampler where M is dropped due to independence. As discussed, jointly sampling $x \mid z, y$ differs from typical filter- and smoothing-based approaches. If there are no states with missing associations, then the full conditional on trajectories can be sampled as,

$$p(x \mid z, y) = \prod_{k=1}^{K} \prod_{t=1}^{T} \frac{p(x_{tk} \mid y^k_{1:t})\ p(x_{(t+1)k} \mid x_{tk})}{p(x_{(t+1)k} \mid y^k_{1:t})} \tag{13}$$

where $p(x_{tk} \mid y^k_{1:t})$ is the filter distribution of x_{tk} and $y^k_{1:t} = \{y_{t'n} : z_{t'n} = k \text{ and } t' \leq t\}$. Sampling from this posterior is similar to smoothing [29], except that the backwards pass draws samples. Inference can be done in parallel over objects and, in the linear Gaussian case, is in closed form with complexity linear in T. For other (possibly non-linear) dynamics or observation models, any procedure that leaves the joint distribution invariant may be used. Described in Appendix B.3, we marginalize over latent states at times when the object has no association.

4 Experiments

Our experiments demonstrate that JPT provides a superior representation of posterior uncertainty as compared to batch MCMCDA (4.1) by virtue of more thoroughly and efficiently exploring the configuration space. Following, we show that JPT outperforms MCMCDA and a recent, optimization-based tracker on large datasets (4.2). Lastly, we show that JPT facilitates targeted queries to an oracle (e.g., a noisy human annotator) yielding significant improvement in trajectory quality with fewer iterations (4.3).

In all experiments, JPT and MCMCDA are run for 5 replicates (Markov chains), each drawing 2000 samples and discarding half as burn-in. All approaches use the same linear Gaussian dynamics.

4.1 Representation of Posterior Uncertainty

Consider the K33 dataset shown in Figure 4 (Top). Three objects begin well-separated but become ambiguous after each of three confusion events (yellow shading). Observe that for k ambiguously proximate objects, there will be $k!$ possible outcomes. Figure 4 (Bottom) shows the $24 = 2!\ 2!\ 3!$ modes that a multi-object tracker would ideally explore in this dataset.

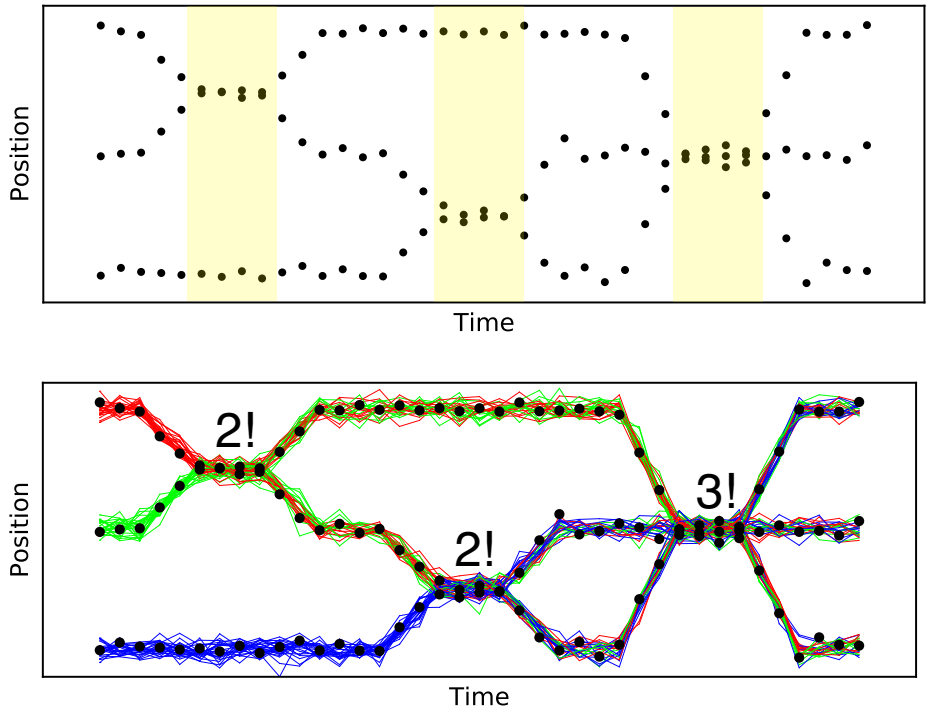


Figure 4: The K33 dataset. Observations over time (**Top**) with yellow shading for ambiguous regions and a joint trajectory sample from the $24 = 2!\ 2!\ 3!$ posterior modes (**Bottom**), each reflecting a possible outcome.

We investigate whether JPT and MCMCDA effectively explore the 24 modes of the K33 dataset by observing posterior trajectory variance for one Markov chain in Figure 5. JPT captures the uncertainty arising in each ambiguous region; for example, red and green cross in some posterior samples but not in others. In contrast, MCMCDA is *overconfident*: it fails to represent any uncertainty in either of the first two ambiguous regions, and is only partially successful in the final region. Not only is MCMCDA overconfident, it is wrong as will be shown by tracking metrics in (4.2).

We quantify how well uncertainty is captured by matching each posterior sample from JPT and MCMCDA to the nearest of the 24 likely outcomes. Details of this matching procedure are in Appendix C.3. Figure 6 shows histograms of modes matched by JPT and MCMCDA in different Markov chains. JPT represents every outcome within each Markov chain while MCMCDA captures at most 2 but usually 1 outcome in a single Markov chain. Noting that the ideal distribution over the 24 matched modes would be uniform, we compare total variation (L1) distance between that and the empirical distributions of matched modes for JPT and MCMCDA (Figure 7), plotted as a function of sample count. Observe that JPT's total variation is low but nonzero, implying imperfect mode exploration. This can be seen in Figure 5 (Top) where the red and green means are not perfectly balanced between the upper and middle paths after the first ambiguous region. Although not uniform in its mode exploration, JPT captures each outcome and is dramatically closer to the ideal uniform distribution than MCMCDA.

4.2 Performance on Real and Synthetic Data

We compare tracking performance of JPT, MCMCDA and a modern optimization-based tracker [18] (MHT) on three datasets: the K33 dataset (39 timesteps, 3 objects, 117 observations), a scientific dataset Marmoset (15k timesteps, 2 objects, 25k observations), and the sports dataset Soccer (1.5k timesteps, 22 objects, 12k observations). Metrics are computed over 200 evenly-spaced samples after burn-in for JPT and MCMCDA. Being deterministic, MHT only provides a point estimate. Details of each dataset are in the Appendix C.1.

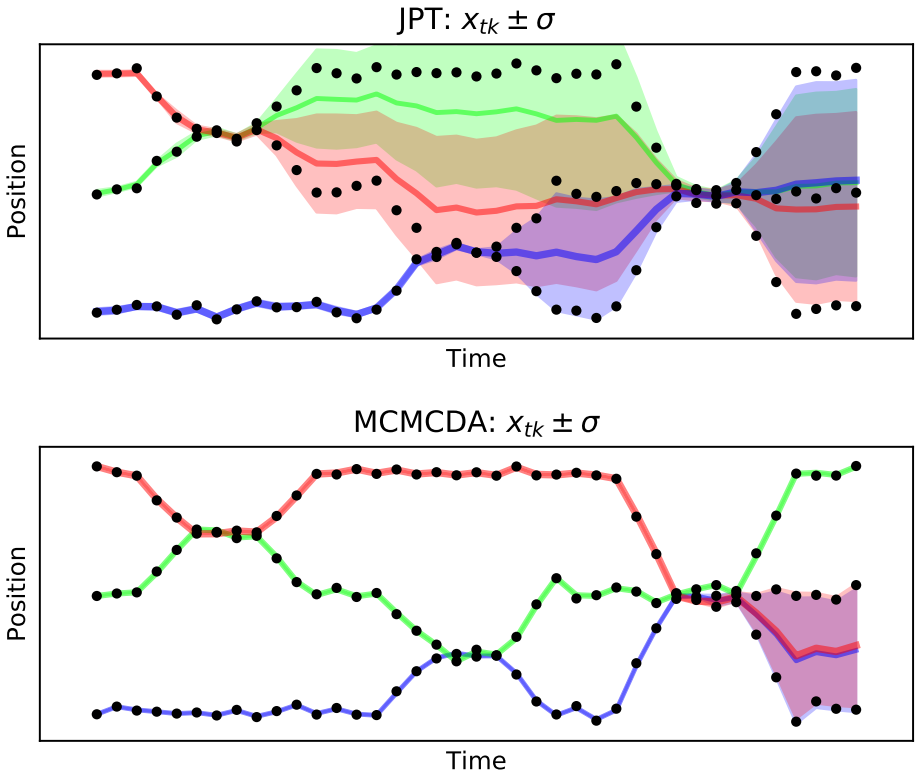


Figure 5: Posterior trajectory values $\pm$ one SD for JPT (**Top**) and MCMCDA (**Bottom**). JPT correctly captures the uncertainty from each ambiguous regions while MCMCDA fails to represent most ambiguities.

Figure 8 shows performance as evaluated by standard CLEAR MOT [6] metrics that account for identity switches (objects get confused), fragmentations (two inferred objects explain one actual object) and misses (an observation isn't correctly associated to an object).

Multi-object tracking accuracy (MOTA) is a summary statistic accounting for these events. JPT outperforms MCMCDA and MHT on all datasets and metrics with notably fewer identity switches, fragmentations and misses. Because MCMCDA uses gating heuristics, the reported performance is taken as the best-scoring MOTA from a grid search over parameter values for each heuristic. Details of the grid search are in Appendix C.2.

4.3 Uncertainty Reduction

We show JPT's accurate representation of posterior uncertainty facilitates follow-on analysis. Specifically, it provides a pathway to obtain high quality trajectories despite significant ambiguity in the original observation set. We augment the original dataset with disambiguations or *annotations* that have the potential to resolve modes in the posterior. Uncertainty representation allows us to identify *informative* annotations, thereby requiring few of them to arrive at quality trajectories.

Let there be L annotations $a = \{a_l\}_{l=1}^{L}$ where each indicates whether two observations $y_{t_1 n_1}, y_{t_2 n_2}$ belong to the same or different objects: $a_l(y_{t_1 n_1}, y_{t_2 n_2}) = 1$ or 0. Assume each annotation is correct with probability $p_a = 0.99$. Intuitively, *informative* annotations involve observations that flank an ambiguous region since the annotated value supports only a subset of modes in the posterior, thereby reducing uncertainty. We use sequential Bayesian experimental design (BED) for automated selection of informative observation pairs for L annotation rounds.

Sequential BED iteratively chooses observation pairs for annotation that yields the greatest information about the latent trajectories, as measured by mutual information (MI) [7]. MI quantifies the

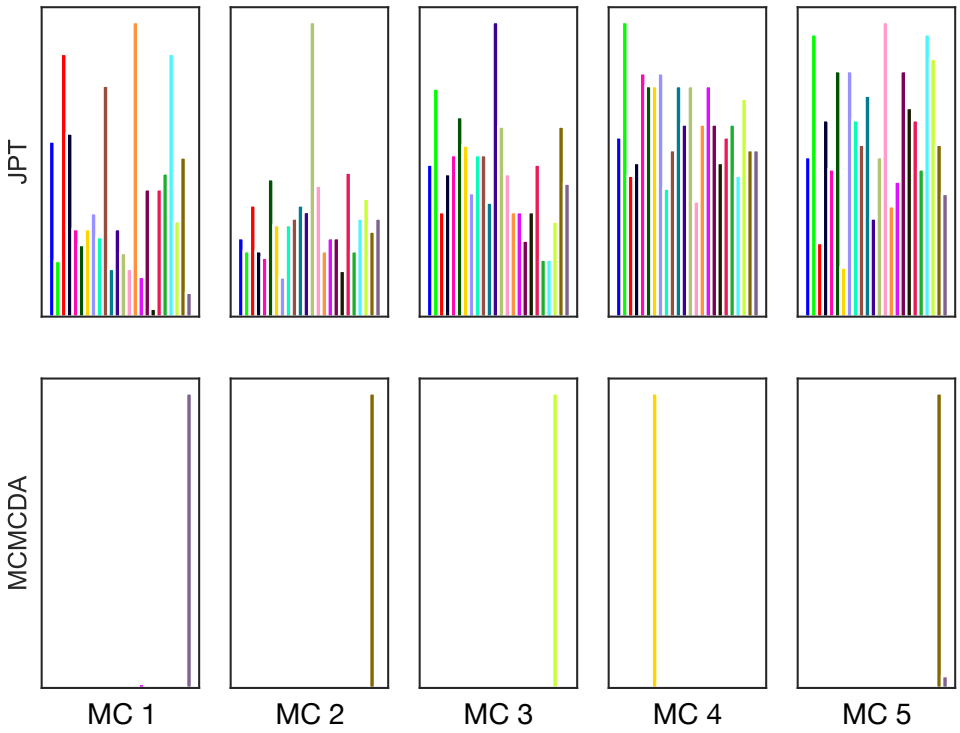


Figure 6: Histograms of the modes captured by JPT (**Top**) and MCMCDA (**Bottom**) in 5 Markov chains (MC). JPT explores all modes in each chain while MCMCDA gets stuck in one or two.

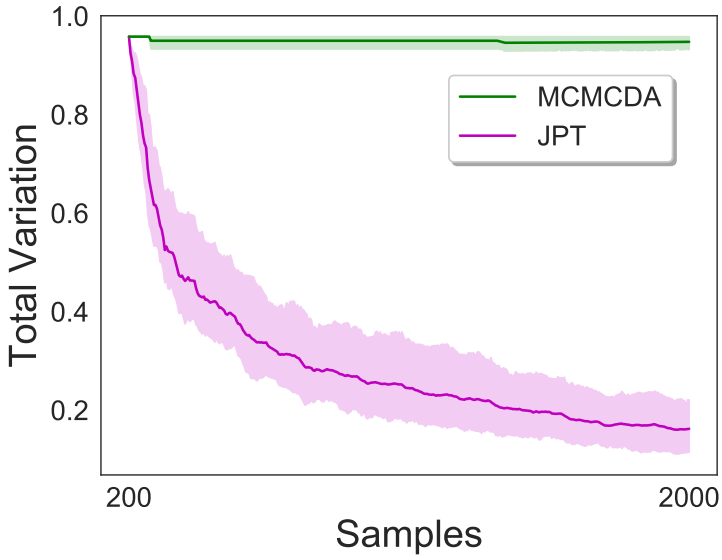


Figure 7: Total variation distance between the true distribution of modes on K33, and the histograms of matched modes for JPT and MCMCDA samples.

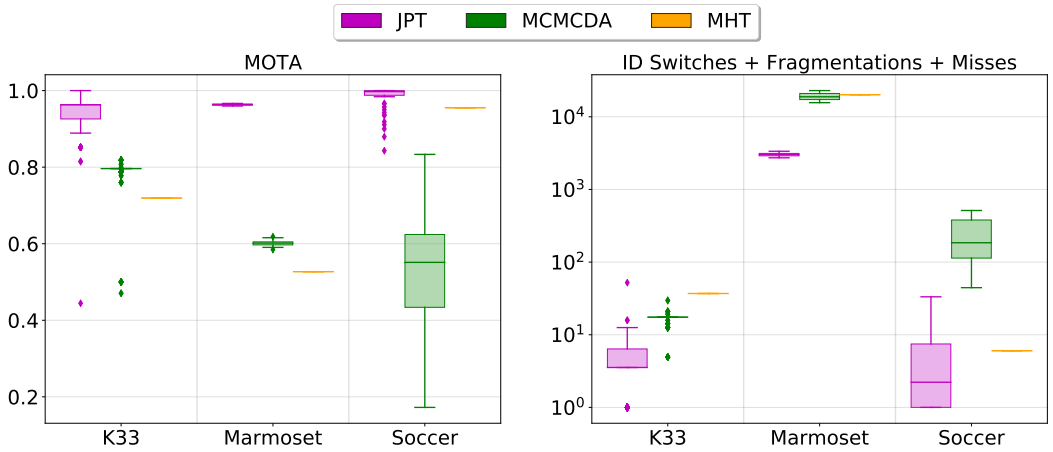


Figure 8: CLEAR MOT metrics for JPT, MCMCDA and MHT on datasets K33, Marmoset and Soccer. (**Left**), higher is better; (**Right**), lower is better.

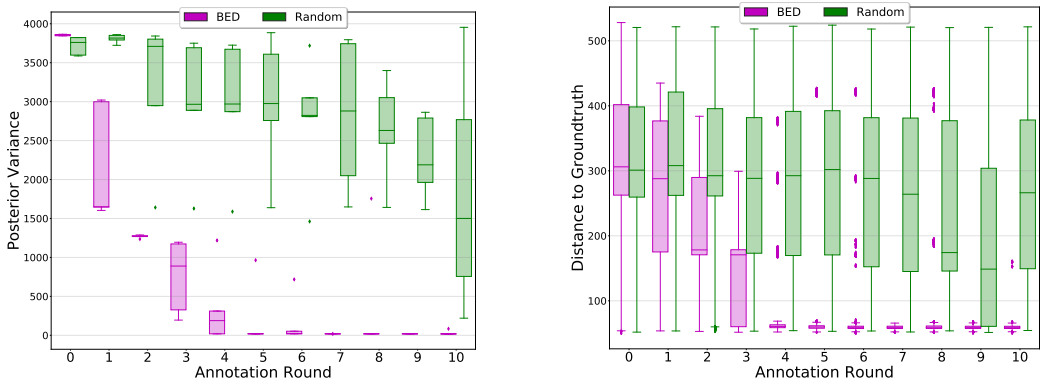


Figure 9: Successive rounds of automatically-scheduled annotations reduce posterior uncertainty (left) and improve trajectory quality (right) when planning with JPT's uncertainty representation (magenta) as compared to planning with no model of uncertainty (green).

expected reduction in posterior uncertainty provided by the annotation, and relies on an accurate representation of uncertainty. Details of the MI estimator and the sequential BED algorithm are in Appendix D.

We perform 5 replicated experiments on the K33 dataset, each with 10 rounds of annotation. The first round starts with no annotations; successive rounds add an annotation. We compare against a baseline that selects annotations at random (i.e., planning without a model of uncertainty).

Figure 9 (left) compares reduction in posterior trajectory uncertainty between BED annotations (magenta) and baseline annotations (green). Both posteriors begin with broad uncertainty, but BED rapidly reduces posterior uncertainty by picking informative annotations until it stabilizes at a low value by round 5. In contrast, the baseline chooses uninformative annotations that don't noticeably reduce uncertainty.

Figure 9 (right) plots distance to the groundtruth as a function of annotation round. As above, the informative annotations chosen by BED (magenta) rapidly improve the quality of posterior samples by reducing their distance to the groundtruth. In contrast, baseline planning yields little improvement in trajectory quality.

These results show that a small number of automatically-scheduled annotations enable rapid reduction in posterior uncertainty that correspond to improvements in track quality. Informative scheduling requires JPT's accurate representation of uncertainty.

5 Conclusion

We propose JPT, a Bayesian solution to the general multi-object tracking problem. We construct efficient inference to reason over permutations of associations and empirically demonstrate that JPT more effectively represents posterior uncertainty than baselines while outperforming them on standard tracking metrics. We then show that JPT's accurate representation of uncertainty enables automatic scheduling of informative disambiguations which rapidly drive down posterior uncertainty while improving trajectory quality.

Appendix

A Multidimensional Assignment Formulation

Multi-object tracking can be formulated in several common ways: as a set partitioning problem [3], a set packing problem [20, 40], a maximum-weight independent set problem [25] or a multidimensional assignment problem [28]. A review of these formulations is conducted by [11], who shows that the multidimensional assignment formulation is not limited to pairwise terms as the set-packing, network-flow solutions [27, 5] are. We next define the multidimensional assignment problem, as it is most similar to how JPT is formulated.

Consider $t = 1, \ldots, T$ timesteps with corresponding observation sets $y = \{y_1, \ldots, y_T\}$ where the time-t observation set $y_t = \{y_{tn}\}_{n=1}^{N_t}$ has N_t observations. Hence, y_{tn} is the n^{th} observation at time t. Let $I_t = \{0, 1, \ldots, N_t\}$ be an index set into y_t, where 0 indicates a false positive or missing detection. Define $\mathcal{P} = I_1 \times I_2 \times \cdots \times I_T$ as the set of paths through all index sets such that every path is length-T and has at least one non-zero index. Interpret a path with a single non-zero index as a false-positive. Interpret a path with two or more non-zero indices as an object. Define $\gamma(i_1, \ldots, i_T)$ as a fixed, real cost for path $(i_1, \ldots, i_T) \in \mathcal{P}$ where $i_t \in I_t$, and $B(i_1, \ldots, i_T)$ is a boolean variable signifying whether path $(i_1, \ldots, i_T)$ is included in a solution. Then the multidimensional assignment problem is to find the $B(i_1, \ldots, i_T)$ that minimizes:

$$
\begin{aligned}
\min \quad & \sum_{i_1=0}^{N_1} \sum_{i_2=0}^{N_2} \cdots \sum_{i_T=0}^{N_T} \gamma(i_1, \ldots, i_T)\, B(i_1, \ldots, i_T) \\
\text{subject to} \quad & \\
& \sum \sum_{I \setminus i_t} \cdots \sum B(i_1, \ldots, i_T) = 1 \qquad (\forall i_t = 1, \ldots, N_t, \forall t = 1, \ldots, T) \\
& B(i_1, \ldots, i_T) \in \{0, 1\}
\end{aligned}
\tag{14}
$$

The objective sums over the costs of all included paths in the solution. For each observation there is a constraint enforcing that it be claimed by exactly one path included in the solution (equivalently, that an observation is uniquely associated either to clutter or a distinct object). MHT [18] and JPDA [14] are deterministic solutions whereas MCMCDA [24] is a stochastic solution to the multidimensional assignment problem. The number of possible paths grow exponentially with T and factorially at each time with N_t. Solving this exactly is NP-hard [8, 26], forcing the above approaches to use gating heuristics such as a maximum distance between object locations and observations, a maximum distance between pairwise object locations, or a maximum number of consecutive missing detections.

We note that JPT represents a departure from the multidimensional assignment formulation because it does not assign a fixed cost to each association hypothesis. While a fixed cost could be constructed– such as by using smoothed state estimates or marginalizing out all trajectories–our focus is to explore and represent joint uncertainty in trajectories and data associations. JPT is not the only work to depart from a traditional multi-object tracking objective [1].

B JPT Metropolis-Hastings Inference

We derive the Hastings ratios for the Switch (B.1) and Gather (B.2) proposals discussed in the paper. We then treat trajectory inference when there are missing associations (B.3).

B.1 Switch Proposal Hastings Ratio

We show the derivation of the Switch proposal, recalling that σ_t is a permutation of objects $1, \ldots, K(z)$:

$$R_{\text{switch}} = \frac{p(x', z', M' \mid y)}{p(x, z, M \mid y)} \times \frac{q_{\text{switch}}(x, z, M \mid x', z', M', y)}{q_{\text{switch}}(x', z', M' \mid x, z, M, y)} \tag{15}$$

$$= \frac{p(z' \mid M') \prod_{t=1}^{T} \frac{1}{Z} p(M'_t \mid M'_{t-1}) p(x'_t \mid x'_{1:t-1}) p(y_t \mid x'_t, z'_t)}{p(z \mid M) \prod_{t=1}^{T} \frac{1}{Z} p(M_t \mid M_{t-1}) p(x_t \mid x_{1:t-1}) p(y_t \mid x_t, z_t)} \times \frac{\prod_{t=1}^{T} \frac{1}{Z_t} p(\sigma_t^{-1}(x'_t) \mid \sigma_{1:t-1}^{-1}(x'_{1:t-1}) p(y_t \mid \sigma_t^{-1}(x_t), \sigma_t^{-1}(z_t))}{\prod_{t=1}^{T} \frac{1}{Z_t} p(\sigma_t(x_t) \mid \sigma_{1:t-1}(x_{1:t-1}) p(y_t \mid \sigma_t(x_t), \sigma_t(z_t))} \tag{16}$$

$$= \frac{p(z' \mid M') \prod_{t=1}^{T} p(M'_t \mid M'_{t-1}) p(x'_t \mid x'_{1:t-1}) p(y_t \mid x'_t, z'_t)}{p(z \mid M) \prod_{t=1}^{T} p(M_t \mid M_{t-1}) p(x_t \mid x_{1:t-1}) p(y_t \mid x_t, z_t)} \times \frac{\prod_{t=1}^{T} p(x_t \mid x_{1:t-1}) p(y_t \mid x_t, z_t)}{\prod_{t=1}^{T} p(x'_t \mid x'_{1:t-1}) p(y_t \mid x'_t, z'_t)} \tag{17}$$

$$= \frac{\prod_{t=1}^{T} p(M'_t \mid M'_{t-1})}{\prod_{t=1}^{T} p(M_t \mid M_{t-1})} \tag{18}$$

where Equation 16 substitutes in the values for each term in the ratio, defining σ_t^{-1} as the inverse permutation of σ_t and Z_t as the normalizer for the sampled σ_t at time t (equal to 1 if $t \notin \tau$). Equation 17 substitutes $\sigma_t(x_t)$ for x'_t and $\sigma_t^{-1}(x'_t)$ for x_t (similarly for $\sigma_t(z_t)$). It also cancels common normalizers Z for the joint ratio and Z_t at each time t for the proposal ratio. Equation 18 cancels all terms related to the dynamics and observation models, and also cancels $p(z' \mid M')$ with $p(z \mid M)$ under the assumption that no object $k \in \mathcal{K}$ was rendered invalid by having fewer than two observations. That can easily be detected and automatically rejected or entirely avoided by defining valid permutations to require the first two observed times for any $k \in \mathcal{K}$ to not be permutable.

B.2 Gather Proposal Hastings Ratio

$$R_{\text{gather}} = \frac{p(x', z', M' \mid y)}{p(x, z, M \mid y)} \times \frac{q_{\text{disperse}}(x, z, M \mid x', z', M', y)}{q_{\text{gather}}(x', z', M' \mid x, z, M, y)} \tag{19}$$

$$= \frac{p(z' \mid M') \prod_{t=1}^{T} p(M'_t \mid M'_{t-1}) p(x'_t \mid x'_{1:t-1}) p(y_t \mid x'_t, z'_t)}{p(z \mid M) \prod_{t=1}^{T} p(M_t \mid M_{t-1}) p(x_t \mid x_{1:t-1}) p(y_t \mid x_t, z_t)} \times \frac{(K(z)+1)^{-1}}{\prod_{t \in \tau_0} \omega_t} \tag{20}$$

where $\omega_t = \delta$ if $z'_{tn} \neq k$ for any n else $\omega_t = \frac{1}{Z_t} p(y_{tn} \mid x_{t'k}, z'_{tn} = k)\, p(x'_{tk} \mid x'_{t'k}, z'_{tn} = k)(1-\delta)$. All dynamics and observation model terms cancel in the posterior ratio for objects other than k, but terms remain for observations that were previously clutter and are now associated to object k and counts $M' \neq M$.

B.3 Forward-Filtering, Backward Sampling with Missing Data

Recall from the main paper that if there are no missing observations (there is some n at every time t such that $z_{tn} = k$ for each $k \in \{1, \ldots, K(z)\}$, then the full conditional on trajectories can be sampled as,

$$p(x \mid z, y) = \prod_{t=1}^{T} p(x_t \mid x_{t+1:T}, y_{1:T}, z_{1:T}) \tag{21}$$

$$= \prod_{t=1}^{T} \frac{p(x_t \mid y_{1:t}, z_{1:t})\, p(x_{t+1} \mid x_t)}{p(x_{t+1} \mid y_{1:t}, z_{1:t})} \tag{22}$$

$$= \prod_{k=1}^{K(z)} \prod_{t=1}^{T} \frac{p(x_{tk} \mid y^k_{1:t})\, p(x_{(t+1)k} \mid x_{tk})}{p(x_{(t+1)k} \mid y^k_{1:t})} \tag{23}$$

where inference is performed independently for each object, $y^k_{1:t} = \{y_{t'n} : z_{t'n} = k \text{ and } t' \leq t\}$ and $p(x_{tk} \mid y^k_{1:t})$ is the marginal (filter) distribution of x_{tk}.

In the case of missing observations, we marginalize over the intervening latent states, realizing samples only at times where an object has an association. Thus, the distribution for x_{tk} under the joint (the numerator of Equation 23), assuming that the most recent previous association occurred at time $\overleftarrow{t} \leq t$ and most recent future association at time $\vec{t} > t$, is:

$$p(x_{tk} \mid y^k_{1:\overleftarrow{t}})\, p(x_{\vec{t}k} \mid x_{tk}) = \int p(x_{(\overleftarrow{t}+1:t)k} \mid y^k_{1:\overleftarrow{t}}) dx_{(\overleftarrow{t}+1:t-1)k} \int p(x_{(t+1:\vec{t})k} \mid x_{tk}) dx_{(t+1:\vec{t}-1k)}. \tag{24}$$

We emphasize that the first term in Equation 24 integrates over past missing states. If there is an association at time t (*i.e.* $\overleftarrow{t} = t$), then there is no integration to carry out in the first term and so it simplifies to $p(x_{tk} \mid y^k_{1:t})$. Similarly, the second term in Equation 24 integrates over future missing states. If $\vec{t} = t + 1$ then there is no integration to carry out in the second term and so it simplifies to $p(x_{(t+1)k} \mid x_{tk})$. Hence, when $\vec{t} = t + 1$ and $\overleftarrow{t} = t$, we recover the numerator of Equation 23.

C Experiment Details

We elaborate on experiments in the paper, starting with a description of each dataset (C.1), then detailing the grid search over gating heuristics we performed to give MCMCDA the best performance (C.2) and ending with details on how we compute distances between sets of associations, used in the uncertainty quantification and uncertainty reduction experiments of the paper (C.3).

C.1 Description of Datasets Used in Experiments

K33 is a synthetic dataset containing multiple ambiguous object crossing events. There are no clutter detections and all objects are detected at all times.

Marmoset contains two primates interacting in a laboratory environment over long periods of time where there are many total and partial occlusion events, as well as occasional clutter detections. Noisy observations are generated as the centroid of the detections from a trained Mask-RCNN neural network [16] and groundtruth accomplished by human annotation that correctly maintains object identities throughout the sequence. As a result, trackers must correctly re-identify objects that have been occluded to avoid getting penalized.

Soccer observations are the unassociated centers of players and referee. Groundtruth does not maintain the identities for objects that go out of frame; hence, re-identification after a total occlusion is not rewarded. Metrics are evaluated in chunks of 20 frames according to the protocol in [35].

C.2 MCMCDA Gating Heuristic Grid Search

The MCMCDA baseline contains two gating heuristics: thresholds on $\bar{v}$, the maximum L2 spatial and d, the maximum L1 temporal distances between two observations associated to the same object. Although they can be removed by setting the thresholds very high, this causes the inference procedures of MCMCDA to devolve into random exploration, severely damaging performance according to the CLEAR MOT metrics used in the paper. To make the comparison with JPT as competitive as possible, we performed a grid search over each gating threshold, as well as providing it with knowledge of the true number of objects or not, and restricted JPT/MCMCDA comparisons so that MCMCDA only used the parameters with best performance as measured by the CLEAR MOT multi-object tracking accuracy (MOTA) metric.

For MCMCDA on the K33 dataset, the best-performing spatial gating threshold was $\bar{v} = 6$ and the best temporal gating threshold was $d = 1$. For Marmoset and Soccer, the best-performing spatial gating threshold was $\bar{v} = 20$ and best-performing temporal gating threshold was $d = 6$. In all cases, MCMCDA performed best without knowledge of the true number of objects because this knowledge could limit its ability to explore by creating excess objects that it later destroyed. In some cases, it would also cause MCMCDA to be severely penalized by occlusion events that persisted for longer than its temporal gating threshold as it could either represent the object before or after the occlusion.

Increasing the gating thresholds to very large numbers caused random exploration of low-probability events in the MCMCDA posterior due to the way its inference is constructed. Specifically, MCMCDA precomputes a sparse graph of paths between observations that respect its gating thresholds. It then computes Metropolis-Hastings proposals that randomly sample from this graph on the assumption that the thresholds were set to encourage likely associations. Thus, MCMCDA inference has a fundamental limitation: either gating thresholds are set tight and some true association hypotheses are excluded, or they are set loose and random exploration occurs.

C.3 Computing Distances Between Association Hypotheses

To match an inferred set of trajectories to another set of trajectories, we begin with the Spatiotemporal Linear Combine (STLC) Distance of [32], which compared favorably in [34]. Briefly, STLC evaluates trajectories on both their L2 spatial and L1 temporal alignment; it supports uneven sampling rates and arbitrary trajectory start/end times. It is a similarity measure that ranges from $[0, 2]$, but we convert it to a cost by inverting the limits.

Given STLC as an object-to-object cost, we define a distance between multi-object tracking association hypotheses by using discrete optimal transport [33], where the cost matrix is filled with the STLC costs of each object pair between the two samples. Note that this supports arbitrary numbers of objects in each sample. This distance was then used in determining mode representation in posterior samples of JPT and MCMCDA in the Uncertainty Quantification experiments and again in the Uncertainty Reduction experiments, where we demonstrated that planned annotations rapidly reduce the distance of JPT samples to the groundtruth.

D Sequential Bayesian Experiment Design

Bayesian experiment design (BED) optimizes a utility function over the set of observation pairs for annotation, where each observation pair $(y_{t_1 n_1}, y_{t_2 n_2})$ is termed a *design*. The sequential form of BED allows annotation results from earlier rounds of BED to inform the selection of designs in subsequent rounds. Mutual information (MI) is a commonly used utility function for BED that quantifies the expected reduction in posterior uncertainty that results from annotation of a design. MI is especially suited for our task since we seek to reduce uncertainty in the trajectory posterior and has several appealing properties include invariance to reparameterization. We next describe the annotation model in detail.

Let $\kappa = (t, n)$ be the time and observation indices that uniquely identifies observation y_{tn}. A design then corresponds to a tuple of these index pairs $d = (\kappa_1, \kappa_2)$. We abuse notation and let $\kappa_1(d) = (t_1, n_1)$ indicate the first pair in design d such that $y_{\kappa_1(d)} = y_{t_1 n_1}$ and $z_{\kappa_1(d)} = z_{t_1 n_1}$ and likewise for $\kappa_2(d)$. Recall from the main text that the annotation indicates if two observations $y_{t_1 n_1}, y_{t_2 n_2}$ belong to the same or different objects – $a_l(y_{t_1 n_1}, y_{t_2 n_2}) = 1$ or 0 respectively – and is correct with probability $p_a = 0.99$. This event corresponds to whether the assignments $z_{t_1 n_1}, z_{t_2 n_2}$ share the same non-zero value – recall $z_{tn} = 0$ indicates clutter. After accounting for the annotation noise and design, we have the following annotation likelihood

$$p_d(a_l = 1 \mid x, y, z, M) = p_d(a_l = 1 \mid z_{\kappa_1(d)}, z_{\kappa_2(d)}) = \begin{cases} 0.99 & \text{if } z_{\kappa_1(d)} = z_{\kappa_2(d)} \text{ and } z_{\kappa_1(d)} > 0 \\ 0.01 & \text{o.w.} \end{cases} \tag{25}$$

When conditioned on just the two assignments $z_{\kappa_1(d)}, z_{\kappa_2(d)}$, the annotation is independent of the remaining variables in the model; this yields the first equality in Equation 25. This conditional taken in conjunction with the joint distribution $p(x, y, z, M)$ described in the paper yields a complete generative model that now includes annotations.

Mutual information between the annotation a_l and the latent trajectories x conditioned on the observations y and past annotations $D = \{a_{1:l-1}, d_{1:l-1}\}$ is given by,

$$I_d(a_l; x \mid y, D) = \mathbb{E}\left[\log \frac{p_d(a_l, x \mid y, D)}{p_d(a_l \mid y, D) p_d(x \mid y, D)}\right] \tag{26}$$

$$= \mathbb{E}\left[\log \frac{p_d(a_l \mid x, y, D)}{p_d(a_l \mid y, D)}\right] \tag{27}$$

$$= \mathbb{E}\left[-\log p_d(a_l \mid y, D)\right] - \mathbb{E}\left[-\log p_d(a_l \mid x, y, D)\right]. \tag{28}$$

We highlight that Equation 28 is the difference of entropies (a measure of uncertainty for random variables), illustrating how MI is a measure of the expected reduction in uncertainty.

We use a greedy approach to Sequential BED, wherein we select the highest MI design within each round of BED. While myopic, this approach avoids the complexity associated with searching for an optimal policy. Thus, at the $l^{\text{th}}-$round of sequential BED, we seek

$$d_l = \arg\max_d I_d(a_l; x \mid y, D). \tag{29}$$

We typically cannot evaluate MI in closed form and instead resort to Monte Carlo estimation using M samples drawn from the posterior $\{a_l^m, x^m, z^m\}_{m=1}^M \sim p(a_l, x, z \mid y, D)$:

$$\tilde{I}_d = \frac{1}{M} \sum_{m=1}^{M} \log \frac{p_d(a_l^m \mid x^m, y, D)}{p_d(a_l^m \mid y, D)}. \tag{30}$$

To evaluate the likelihoods in Equation 30, first we expand them as,

$$p_d(a_l^m \mid x^m, y, D) = \sum_z p_d(a_l^m \mid z, x^m, y, D)\, p(z \mid x^m, y, D) \tag{31}$$

$$= \sum_{z_{\kappa_1(d)}, z_{\kappa_2(d)}} p_d(a_l^m \mid z_{\kappa_1(d)}, z_{\kappa_2(d)})\, p(z_{\kappa_1(d)}, z_{\kappa_2(d)} \mid x^m, y, D). \tag{32}$$

$$p_d(a_l^m \mid y, D) = \sum_z p_d(a_l^m \mid z, y, D)\, p(z \mid y, D) \tag{33}$$

$$= \sum_{z_{\kappa_1(d)}, z_{\kappa_2(d)}} p_d(a_l^m \mid z_{\kappa_1(d)}, z_{\kappa_2(d)})\, p(z_{\kappa_1(d)}, z_{\kappa_2(d)} \mid y, D) \tag{34}$$

Equation 32 can be evaluated exactly because we can obtain $p(z_{\kappa_1(d)}, z_{\kappa_2(d)} \mid x^m, y, D)$ through enumeration of all pairwise assignments conditioned on the sampled trajectories x^m and observations y. Equation 34, on the other hand, requires $p_d(a_l^m \mid y, D)$ which is intractable, so we again use Monte Carlo estimation,

$$\hat{p}_d(a_l^m \mid y, D) = \frac{1}{M} \sum_{m'=1}^{M} p_d(a_l^m \mid z_{\kappa_1(d)}^{m'}, z_{\kappa_2(d)}^{m'}). \tag{35}$$

Our MI estimator is then,

$$\hat{I}_d = \frac{1}{M} \sum_{m=1}^{M} \log \frac{p_d(a_l^m \mid x^m, y, D)}{\hat{p}_d(a_l^m \mid y, D)} \tag{36}$$

where $p_d(a_l^m \mid x^m, y, D)$ is given in Equation 32 and $\hat{p}_d(a_l^m \mid y, D)$ is given in Equation 35.

E Switch Proposals Generalize Extended HMM Proposals

Switch proposals generalize the Extended HMM (EHMM) proposals of [21] by permitting discretizations that depend on the latent space. In brief, EHMM proposals compose an inference method that helps explore a posterior distribution by proposing a discretization of latent states (called "pool states") over time. A hidden Markov model is then defined over the pool states and a joint sample drawn using forward-filtering, backward sampling. Crucially, the discretization sampled by an EHMM proposal includes the current latent state, but *must not* otherwise depend on it. If it does, detailed balance is lost because calculating the reverse move probability would require a difficult integration over the latent space.

In contrast, Switch proposals sample from a discretization that depends on the current latent state while maintaining detailed balance. In the nomenclature of EHMM proposals, the "pool states" of JPT's Switch proposal are permutations of latent state x, z. JPT then samples from the generative model of an HMM that contains no future information; thus, no backwards pass is required. We note that the Switch proposal always contains the current state as represented by the identity permutation over all times.

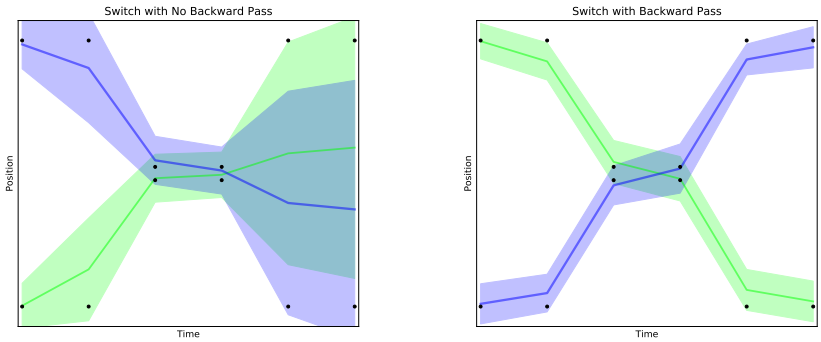


Figure 10: Observations y (black points) with two modes (blue, green objects crossed or not). Shading indicates marginal posterior trajectory variance. (**Left**): Switch statements that leave no future associations fixed have strong ability to explore modes because future associations do not force an outcome. (**Right**): Switch statements that leave future associations fixed (from the final timestep) will favor the modes supported by those fixed associations.

Switch proposals can easily be constructed to contain future information by restricting switch times τ so that some future associations from the sample set $\mathcal{K}$ remain fixed. Doing so causes a need for future information to propagate backward. We found that doing so without being careful about which associations to leave fixed impairs the ability of Switch proposals to explore different modes as future information encouraged the current sample to remain in the same mode. See Figure 10 for an example. Backward propagation *is* desirable when future information comes from annotations since they are intended to reduce posterior uncertainty. But, in the absence of annotations, future information in the form of restricted Switch times is not desirable.

References

[1] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *CVPR*, volume 2, page 7, 2011.

[2] A. Arasteh, B. V. Vahdat, and R. S. Yazdi. Multi-target tracking of human spermatozoa in phase-contrast microscopy image sequences using a hybrid dynamic bayesian network. *Scientific reports*, 8(1):1–19, 2018.

[3] E. Balas and M. W. Padberg. Set partitioning: A survey. *SIAM review*, 18(4):710–760, 1976.

[4] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR 2011*, pages 3457–3464. IEEE, 2011.

[5] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1806–1819, 2011.

[6] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[7] J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979.

[8] S. Blackman, S. S. Blackman, and R. Popoli. Design and analysis of modern tracking systems. 1999.

[9] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1820–1833, 2010.

[10] A. Butt and R. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1853, 2013.

[11] R. T. Collins. Multitarget data association with higher-order motion models. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1744–1751. IEEE, 2012.

[12] J. Ferryman. *PETS: Performance Evaluation of Tracking and Surveillance*, 2020.

[13] W. Ge and R. T. Collins. Multi-target data association by tracklets with unsupervised parameter estimation. In *BMVC*, volume 2. Citeseer, 2008.

[14] Hamid, S. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid. Joint probabilistic data association revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 3047–3055, 2015.

[15] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[17] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE transactions on pattern analysis and machine intelligence*, 27(11):1805–1819, 2005.

[18] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4696–4704, 2015.

[19] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[20] C. Morefield. Application of 0-1 integer programming to multitarget tracking problems. *IEEE Transactions on Automatic Control*, 22(3):302–312, 1977.

[21] R. M. Neal, M. J. Beal, and S. T. Roweis. Inferring state sequences for non-linear systems with embedded hidden markov models. In *Advances in neural information processing systems*, pages 401–408, 2004.

[22] J. Nieto, J. Guivant, E. Nebot, and S. Thrun. Real time data association for fastslam. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, volume 1, pages 412–418. IEEE, 2003.

[23] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011.

[24] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for multi-target tracking. *IEEE Transactions on Automatic Control*, 54(3):481–497, 2009.

[25] D. J. Papageorgiou and M. R. Salpukas. The maximum weight independent set problem for data association in multiple hypothesis tracking. In *Optimization and Cooperative Control Strategies*, pages 235–255. Springer, 2009.

[26] E. L. Pasiliao. Local neighborhoods for the multidimensional assignment problem. In *Dynamics of information systems*, pages 353–371. Springer, 2010.

[27] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1201–1208, 2011.

[28] A. B. Poore. Multidimensional assignment formulation of data association problems arising from multitarget and multisensor tracking. *Computational Optimization and Applications*, 3(1):27–57, 1994.

[29] H. E. Rauch, F. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

[30] S. Sastry, D. Culler, M. Howard, T. Roosta, B. Zhu, J. Taneja, S. Kim, S. Schaffert, J. Hui, P. Dutta, et al. Instrumenting wireless sensor networks for real-time surveillance. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 3128–3133. IEEE, 2006.

[31] A. V. Segal and I. Reid. Latent data association: Bayesian model selection for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2911, 2013.

[32] S. Shang, L. Chen, Z. Wei, C. S. Jensen, K. Zheng, and P. Kalnis. Trajectory similarity join in spatial networks. 2017.

[33] J. Solomon. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*, 2018.

[34] H. Su, S. Liu, B. Zheng, X. Zhou, and K. Zheng. A survey of trajectory distance measures and performance evaluation. *The VLDB Journal*, pages 1–30, 2019.

[35] R. D. Turner, S. Bottone, and B. Avasarala. A complete variational tracker. In *Advances in Neural Information Processing Systems*, pages 496–504, 2014.

[36] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling. *International Journal of Computer Vision*, 101(1):184–204, 2013.

[37] B. Wang, G. Wang, K. Luk Chan, and L. Wang. Tracklet association with online target-specific metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1234–1241, 2014.

[38] J. L. Williams and R. A. Lau. Data association by loopy belief propagation. In *2010 13th International Conference on Information Fusion*, pages 1–8. IEEE, 2010.

[39] A. R. Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *European Conference on Computer Vision*, pages 343–356. Springer, 2012.

[40] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[41] Y. Zhou, J. Sharma, Q. Ke, R. Landman, J. Yuan, H. Chen, D. S. Hayden, J. W. Fisher, M. Jiang, W. Menegas, et al. Atypical behaviour and connectivity in shank3-mutant macaques. *Nature*, 570(7761):326–331, 2019.