18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

# Efficient Information Planning in Gaussian MRFs

Georgios Papachristoudis
CSAIL, MIT
Cambridge, MA 02139
Email: geopapa@mit.edu

John W. Fisher III
CSAIL, MIT
Cambridge, MA 02139
Email: fisher@csail.mit.edu

***Abstract*—We are interested in information planning of structures represented by sparse graphical models where measurements correspond to a limited number of nodes. Choosing a set of measurements, which better describe spatiotemporal phenomena is a fundamental task whose optimal solution becomes intractable as the number of measurements grows. Krause et al. (2005) and Williams et al. (2007) have shown that by exploiting the submodular property of mutual information, a simple polynomial greedy selection algorithm comes with near-optimal guarantees. Most previous works assume *oracle value* models, where the value of a set of measurements is provided in constant time. However, the complexity of evaluating the reward of different measurement sets might be nontrivial in realistic settings. Here, we show that by taking advantage of sparsity in the measurement process, the complexity of information planning in Gaussian models is dramatically reduced. We additionally demonstrate that working with the information form reduces the computational load to the absolutely necessary computations. Lastly, we present an analysis of the computational complexity of different orders of selecting measurements known as *visit walks*, and suggest how this could help in forming a measurement schedule. We restrict ourselves to Gaussian Hidden Markov Models (HMMs), but the underlying analysis generalizes to general Markov Random Fields (MRFs).**

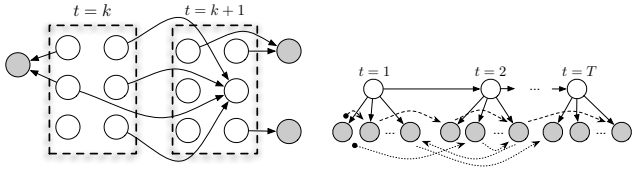
## I. Introduction

The idea of using information measures as a reward function for Bayesian experimental design, active inference being a special case, is a well-studied problem. Early analysis includes Lindley (1956) [1], which considers an alternative to the classic work of Blackwell (1950) on comparisons of experiments [2], and that of Bernardo (1979) which extends the concepts to a more principled setting [3]. The relations of such measures to risk have also been studied extensively. See [4] for a review and [5], [6] on relations to surrogate loss functions. Work in [7]–[10] considers their use in the setting of sequential inference subject to resource constraints on measurements.

In the sequential setting, complexity issues arise when planning multiple time-steps ahead. Specifically, the complexity of active planning methods is combinatorial in the number of sensing actions and exponential in the planning horizon. Krause et al. (2005) observed that a commonly utilized choice of information measure, mutual information (MI), is *submodular* [11] when the measurements are statistically independent conditioned on the quantity of interest [12]. As such, the results of Nemhauser et al. (1978) [13] apply and tractable greedy selection methods are guaranteed to be within a factor of the optimal (though, intractable) selection. The analysis of Williams et al. (2007) [14] provides guarantees for greedy selection for the more general case of inference in graphical models when measurements are divided into subsets with local constraints on subset selection and when the latent variable structure may not be fully specified a priori (e.g., inference in Markov chains for streaming data). It is important to emphasize that this analysis is about the basis for *planning* the sensing actions. Inference proceeds *after* having selected a plan. Consequently, a first step is to evaluate the reward of the prospective plan.

An important, often neglected, aspect of information-based approaches, however, is the computational cost of evaluating a given plan. While the bounds for greedy selection hold for *any* plan subject to the same constraints, one is free to reorder the sequence in which subsets are considered. Some reorderings have significantly higher information rewards than others. A simple example occurs in a Markov chain where at each node one may choose $k$ out of $N$ measurements. A naïve plan considers each node in order (greedily selecting $k$ out of $N$ available measurements at each node). Alternatively, one may consider nodes in random order selecting a single measurement (from those that have not already been selected), but ensuring each node is considered $k$ times. Evaluating the information reward of the naïve plan has significantly lower computational complexity than the random plan, but the random plan will often have significantly higher information reward. Thus, there is motivation to expend computational resources for exploring multiple plans subject to the same constraints. Furthermore, when exploring multiple plans, the plan with lowest reward provides the lowest upper bound on the optimal plan yielding a tighter performance guarantee as compared to the greedy plan with highest reward.

Here, we consider the computational complexity of evaluating information rewards for measurement selection in Gaussian models. In such models, complexity depends on the number of latent variables, the number of measurements to be explored and the visitation order. We show speedups up to a thousand times by taking advantage of sparsity in the measurement process without changing the outcome of the greedy algorithm. In addition, we demonstrate that by working with the information form of Gaussian, we can provide the sufficient statistics at every step with much reduced computation. We achieve that by deploying a variant of belief propagation that is more suitable for adaptive inference settings. The results of the later technique are exact for Markov chains, trees, and poly-trees. This analysis is particularly useful for large-scale models, since the evaluation of information rewards poses a




(a) (b)

Fig. 1. **Sparsity and different walks.** (a) Each measurement of $\mathsf{X}_k$ depends only on a few components of $\mathsf{X}_k$. Dashed rectangles represent vectors of variables at a point, $\mathsf{X}_k, \mathsf{X}_{k+1}$. (b) This structure can be collectively represented as an HMM. Two walks, that is, orders in which observation sets are visited, are visualized. A forward walk represented with dashed arrows, and another walk represented with dotted arrows. Arrows with a circle in one end denote the beginning of the walk. Different walks visit each observation set the same number of times, but in different orders.

major computational bottleneck. Additionally, we demonstrate empirically that both the information reward and evaluation complexity are largely decoupled and as such, exploration of low-complexity walks yields high information rewards and tighter upper bounds.

## II. Information Planning in Gaussian Models

For brevity and clarity of discussion, we restrict ourselves to HMMs. However, the results easily extend to trees. It also can generalized to general MRFs by using the method of Liu et al. (2012) as described in [15]. The underlying dynamical system of such models can be described by

$$\mathsf{X}_t = A_{t-1}\mathsf{X}_{t-1} + \mathsf{V}_{t-1} \tag{1}$$

$$\mathsf{Y}_t = C_t\mathsf{X}_t + \mathsf{W}_t, \tag{2}$$

where $\mathsf{X}_t, \mathsf{Y}_t, t \in \{1, \ldots, T\}$ are the latent and observed variables, respectively. In addition, $\mathsf{V}_t \sim \mathcal{N}(\mu_{v,t}, Q_t)$, $\mathsf{W}_t \sim \mathcal{N}(\mu_{w,t}, R_t)$ are the respective process and measurement noises, with $R_t$ being block-diagonal. Let $\mathsf{X} = \{\mathsf{X}_1, \ldots, \mathsf{X}_T\}$ denote the set of latent variables up to time $T$. Each $\mathsf{X}_t$ is a $d$-dimensional vector. For each $\mathsf{X}_t$, we define an *observation set*, denoted by $\mathcal{V}_t$, where $|\mathcal{V}_t| = N_t$ ($N_t$ comparable to $d$). Each measurement $\mathsf{Y}_{t,u}$ from set $\mathcal{V}_t$ is an $m$-dimensional vector. In Eq. (2), $\mathsf{Y}_t$ represents the set of all $N_t$ measurements of $\mathcal{V}_t$. Therefore, $C_t$ is a $N_t m \times d$ matrix, where consecutive $m$-row patches correspond to one measurement from $\mathcal{V}_t$. W.l.o.g., we assume $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset, \forall i \neq j$. As shown in Fig. 1a, a measurement depends at most on $q$ elements of $\mathsf{X}_t$. As a consequence, $C_t$ is a highly sparse matrix.

Our goal is to characterize approximate solutions to the following (generally intractable) combinatorial optimization problem:

$$\mathcal{O} \in \underset{\{S \mid\mid S \cap \mathcal{V}_t| \leq k_t, \forall t\}}{\arg\max} f(\mathcal{S}), \tag{3}$$

where $f(\mathcal{S})$ is a set function and $k_t$ are the selection constraints for the $t$-th set. We restrict ourselves to monotonic functions and thus it never "hurts" to obtain more measurements. Hence, the inequality constraint becomes tight. To give an indication of the hardness of problem (3), there are $\prod_t \binom{N_t}{k_t} = \binom{N}{k}^T$ feasible solutions, which is an extremely large number as $N, k, T$ grow.

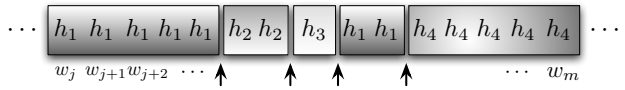



Fig. 2. **Composition of a walk.** A walk, $\boldsymbol{w} = \{w_1, \ldots, w_M\}$, represents the particular order observation sets are visited during measurement selection. A walk segment is a part of the walk that consists of elements from the same observation set. Here, we have two segments of length 5 corresponding to sets $\mathcal{V}_{h_1}, \mathcal{V}_{h_4}$, two segments of length 2 corresponding to sets $\mathcal{V}_{h_1}, \mathcal{V}_{h_2}$, and one of length one corresponding to set $\mathcal{V}_{h_3}$. Vertical arrows indicate *transition points*, which are points of transition between different observation sets.

### A. Greedy Selection

Greedy methods select elements sequentially conditioned on the previous selection. A *walk* $\boldsymbol{w} = \{w_1, \ldots, w_M\}$ denotes the particular order in which observation sets are visited. Greedy selection for a particular walk is defined as

$$g_j = \underset{u \in \mathcal{V}_{w_j} \setminus \mathcal{G}_{j-1}}{\arg\max} f(u \mid \mathcal{G}_{j-1}), \tag{4}$$

where $w_j$ is the observation set index corresponding to the $j$-th element of the walk and $\mathcal{G}_{j-1}$ is the greedy set obtained up to the previous iteration. The marginal increase in the reward (incremental reward) of incorporating a measurement $u$ in a given set $\mathcal{G}_{j-1}$ is denoted as $f(u \mid \mathcal{G}_{j-1}) = f(\mathcal{G}_{j-1} \cup \{j\}) - f(\mathcal{G}_{j-1})$. Essentially, at each greedy step we choose the measurement that maximizes the incremental reward based on past selections $\mathcal{G}_{j-1}$. Since we explore at most all $N$ measurements from a set and there are $kT$ steps of the algorithm, its overall complexity is $\mathcal{O}(kTN)$ as compared to $\mathcal{O}(\binom{N}{k}^T)$ of the optimal approach. Here, we use MI as the reward function, $f(\mathcal{S}) = I(\mathsf{X}; \mathsf{Y}_\mathcal{S})$, with the resulting incremental reward being $f(u \mid \mathcal{G}_{j-1}) = I(\mathsf{X}; \mathsf{Y}_u \mid \mathsf{Y}_{\mathcal{G}_{j-1}})$. A walk has length $M = \sum_{t=1}^T k_t$. A *forward walk* is a walk with non-decreasing order (referred as *naïve* walk in the previous section). W.l.o.g., we assume that each observation set has the same number of measurements $N_t = N$.

### B. Theoretical guarantees on greedy selection

A greedy solution achieves in the worst case half of the optimal reward [14]:

$$f(\mathcal{G}) \geq \frac{1}{2} f(\mathcal{O}).$$

The above bound, seen differently, serves also as an upper bound on the optimal reward which cannot exceed twice the reward of an arbitrary walk. Depending on the problem structure, different walks yield significantly different rewards. Of most interest are the walks of highest and lowest reward, since the first offers the best greedy solution, while the second gives the lowest bound on the optimal reward. Note that tighter on-line bounds are available with some computation [14].

### C. Gaussian HMMs

Gaussian models are appealing since information rewards can be expressed as a function of the covariance of the underlying process. In addition, obtaining values for the measurements does not have an effect in the selection of measurements, which makes planning completely independent of the actual measurement values. As stated, the incremental reward of a measurement $u$ is defined as

$$\begin{aligned} f(u \mid \mathcal{G}_{j-1}) &= I(\mathsf{X}; \mathsf{Y}_u \mid \mathsf{Y}_{\mathcal{G}_{j-1}}) \\ &= H(\mathsf{X} \mid \mathsf{Y}_{\mathcal{G}_{j-1}}) - H(\mathsf{X} \mid \mathsf{Y}_u, \mathsf{Y}_{\mathcal{G}_{j-1}}). \end{aligned} \tag{5}$$

In the Gaussian setting, entropy can be analytically expressed as

$$H(\mathsf{X}) = \frac{d}{2}(1 + \log 2\pi) + \frac{1}{2}\log|\Sigma_{\mathsf{X}}|.$$

Therefore, Eq. (5) becomes

$$f(u \mid \mathcal{G}_{j-1}) = \frac{1}{2}\log\frac{|\Sigma_{\mathsf{X}|\mathcal{G}_{j-1}}|}{|\Sigma_{\mathsf{X}|\{u\}\cup\mathcal{G}_{j-1}}|} = \frac{1}{2}\log\frac{|J_{\mathsf{X}|\{u\}\cup\mathcal{G}_{j-1}}|}{|J_{\mathsf{X}|\mathcal{G}_{j-1}}|}, \tag{6}$$

depending on whether we work with the moment (covariance) or information (precision) form.

In Gaussian HMMs, covariance updates are as follows:

$$\Sigma_{t|t-1} = A_{t-1}\Sigma_{t-1|t-1}A_{t-1}^T + Q_{t-1} \tag{7}$$

$$\Sigma_{t|t} = \Sigma_{t|t-1} - G_t C_t \Sigma_{t|t-1} \tag{8}$$

$$G_t = \Sigma_{t|t-1}C_t^T(C_t\Sigma_{t|t-1}C_t^T + R_t)^{-1},$$

where $\Sigma_{t|t-1} = \mathrm{cov}(\mathsf{X}_t \mid \mathsf{Y}_1, \ldots, \mathsf{Y}_{t-1}) = \mathrm{cov}(\mathsf{X}_t \mid \mathsf{Y}_{1:t-1})$, $\Sigma_{t|t} = \mathrm{cov}(\mathsf{X}_t \mid \mathsf{Y}_{1:t})$. Eqs. (7), (8) are referred to as the *propagation* and *update* steps, respectively. It is important to note that pursuant to propagation, the incremental reward depends only on the *local* update to the node entropy.[1] Consequently, updates to the local covariance (equivalently precision) matrix *fully* quantify the information reward with respect to the full set of latent nodes.

## III. Complexity Reduction Through Sparsity

For a given walk, three primary sources of computational complexity arise when evaluating the information reward: **(i) propagation**, in which the covariance at the next time point is computed; **(ii) exploration**, in which the information rewards of the remaining measurements of the current observation set are computed to find the best measurement in a greedy sense; and **(iii) updates** in which the covariance matrix of the hidden variable linked to the current observation set incorporates the selected measurement. The last two sources of complexity do not depend on the structure of the hidden graph and hence the underlying analysis extends straightforwardly to general Gaussian MRFs. Greedy algorithms proceed as follows; propagation, exploration and update: (i) propagate the uncertainty given the existing measurements to the next node in the walk, (ii) explore the available candidate measurements in the observation set of that node choosing the measurement with highest incremental reward, as dictated by Eq. (4), and lastly, (iii) update the uncertainty at that node after selecting the measurement. We defer the discussion of the propagation step to Sec. IV as it relies on the sparsity of $A$ and primarily applies to *forward* walks. Exploration and updates depend on the structure of $C$.

With slight abuse of notation, we denote the $m$-row portion of matrix $C_t$ corresponding to measurement $\mathsf{Y}_{t,u}$ as $C_t(u,:)$. After we select a measurement $g_j$ at step $j$ as dictated by Eq. (4), we need to update the uncertainty at the current node of the walk, $\mathsf{X}_{w_j}$. For notational consistency with the analysis later in the text, we will denote the greedily selected measurement $g_j$ by $u$ in the discussion below. From Eq. (8), we see that the complexity of the update step is $\mathcal{O}(md^2)$, where $d$ is the dimensionality of $\mathsf{X}_{w_j}$, since the computation is dominated by the term $C_{w_j}(u,:)\Sigma_{w_j|\mathcal{G}_{j-1}}$ for $m \ll d$. Updates at each iteration yield overall complexity of $\mathcal{O}(\sum_{t=1}^{T} k_t m d^2) \overset{k_t=k}{=} \mathcal{O}(Tkmd^2)$. Exploration of one measurement takes $\mathcal{O}(d^3)$ time as shown in Eq. (6), since it requires the computation of the determinant of a $d\times d$ matrix. The number of measurements that should be considered is $\mathcal{O}(kTN)$. Therefore, the overall complexity of exploration is $\mathcal{O}(TkNd^3)$, which makes it the dominant term in the computational load.

The above analysis is agnostic to the sparsity of $C$. Here, we show that evaluation complexity is dramatically reduced by taking advantage of this sparsity. Additionally, adopting the information form yields further significant efficiency. Let $I_c$ denote the indicator matrix of the non-zero elements of $C$. Computation of $I_c$ requires $\mathcal{O}(Nmd)$ time or $\mathcal{O}(TNmd)$, if time-varying. We further assume that the largest parent set for a measurement vector is of size $q$, where $q \ll d$.

### A. Reductions during updates

For an $m\times 1$ measurement $u$, denote $I_u$ as $I_u = \vee_{i=1}^{m} I_c(u_i,:)$, where $u_i$ is the $i^{\text{th}}$ element of measurement $u$. That is, $I_u$ is the $d \times 1$ indicator vector representing the nodes of latent graph $\mathsf{X}$ that generated measurement $u$. If latent variable $\mathsf{X}_i$ is linked to measurement $\mathsf{Y}_{t,u}$, $I_u(i) = 1$, while 0 otherwise. Since a measurement depends on at most $q$ latent variables, we have that $\sum_{i=1}^{d} I_u(i) \le q$. Making use of Eq. (8), the updated covariance $\Sigma'$ with $\Sigma$ as a prior is:

$$\Sigma' = \Sigma - \Sigma C(u,:)^T(C(u,:)\Sigma C(u,:)^T + R(u,u))^{-1}C(u,:)\Sigma. \tag{9}$$

By inverting (9), we obtain the updated precision matrix:

$$\begin{aligned} J' &= (\Sigma - \Sigma C(u,:)^T(C(u,:)\Sigma C(u,:)^T + R(u,u))^{-1}C(u,:)\Sigma)^{-1} \\ &= \Sigma^{-1} + C(u,:)^T R(u,u)^{-1}C(u,:) \\ &= J + C(u,:)^T R(u,u)^{-1}C(u,:), \end{aligned} \tag{10}$$

where we have made use of the Woodbury matrix identity.

We denote by $\hat{C}_u = R(u,u)^{-1/2}C(u,I_u)$ the $m \times q$ matrix that takes only into account the part of $C$ matrix that is relevant to the latent variables that generated measurement $u$. The matrix square root $R(u,u)^{-1/2}$ can be recovered in $\mathcal{O}(m^3)$

[1] Note that this property is a consequence of conditional independence and is not restricted to Gaussian models.

time. In addition, $\hat{C}_u$ is computed in $\mathcal{O}(m^2 q)$ time. Therefore, (10) can be rewritten as

$$J' = \begin{bmatrix} J(I_u, I_u) & J(I_u, \neg I_u) \\ J(\neg I_u, I_u) & J(\neg I_u, \neg I_u) \end{bmatrix} + \begin{bmatrix} \hat{C}_u^T \\ 0^T \end{bmatrix} \begin{bmatrix} \hat{C}_u & 0 \end{bmatrix}, \tag{11}$$

where $\neg I_u$ denotes all the latent variables that are not directly linked to measurement $u$. Eq. (11) can be written more concisely as

$$J'(I_u, I_u) = J(I_u, I_u) + \hat{C}_u^T \hat{C}_u,$$

since only the block of $J$ dictated by $I_u$ will be affected by $\hat{C}_u^T \hat{C}_u$.

The above calculation requires $\mathcal{O}(mq^2)$ steps. As such, if $m > q$, the complexity is dominated by $\mathcal{O}(m^3)$, while by $\mathcal{O}(mq^2)$, otherwise. Compare this to the complexity of the standard calculation, which is $\mathcal{O}(md^2)$ per update, resulting in a speedup on the order of $(\frac{d}{\max\{m,q\}})^2$ per update.

### *B. Reductions during exploration*

Greedy selection for Gaussian models simplifies to

$$\begin{aligned} g_j &= \underset{u \in \mathcal{V}_{w_j} \setminus \mathcal{G}_{j-1}}{\arg\max} \; I(\mathsf{X}; \mathsf{Y}_{w_j,u} \mid \mathsf{Y}_{\mathcal{G}_{j-1}}) \\ &= \underset{u \in \mathcal{V}_{w_j} \setminus \mathcal{G}_{j-1}}{\arg\max} \; I(\mathsf{X}_{w_j}; \mathsf{Y}_{w_j,u} \mid \mathsf{Y}_{\mathcal{G}_{j-1}}) \\ &= \underset{u \in \mathcal{V}_{w_j} \setminus \mathcal{G}_{j-1}}{\arg\max} \; \log \frac{|J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}|}{|J_{w_j|\mathcal{G}_{j-1}}|}, \end{aligned} \tag{12}$$

where $J_{w_j|\mathcal{G}_{j-1}}$ is the precision of $\mathsf{X}_{w_j}$ given observations $\mathsf{Y}_{\mathcal{G}_{j-1}}$. Similarly, $J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}$ is the precision after the incorporation of measurement $u$.

We can express $J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}$ as a function of $J_{w_j|\mathcal{G}_{j-1}}$ by using Eq. (11):

$$J_{w_j|\{u\}\cup\mathcal{G}_{j-1}} = J_{w_j|\mathcal{G}_{j-1}} + \begin{bmatrix} \hat{C}_{w_j,u}^T \\ 0^T \end{bmatrix} \begin{bmatrix} \hat{C}_{w_j,u} & 0 \end{bmatrix}.$$

As we observe in (12), we are only interested in ratios of determinants to select the next measurement. If we use the Matrix Determinant Lemma on $J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}$ we obtain

$$\begin{aligned} &|J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}| \\ &= |J_{w_j|\mathcal{G}_{j-1}}| \left| \mathbb{I}_{m\times m} + \begin{bmatrix} \hat{C}_{w_j,u} & 0 \end{bmatrix} J_{w_j|\mathcal{G}_{j-1}}^{-1} \begin{bmatrix} \hat{C}_{w_j,u}^T \\ 0^T \end{bmatrix} \right| \\ &= |J_{w_j|\mathcal{G}_{j-1}}| \left| \mathbb{I}_{m\times m} + \begin{bmatrix} \hat{C}_{w_j,u} & 0 \end{bmatrix} \Sigma_{w_j|\mathcal{G}_{j-1}} \begin{bmatrix} \hat{C}_{w_j,u}^T \\ 0^T \end{bmatrix} \right| \end{aligned}$$

Therefore,

$$\frac{|J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}|}{|J_{w_j|\mathcal{G}_{j-1}}|} = |\mathbb{I}_{m\times m} + \hat{C}_{w_j,u} \Sigma_{w_j|\mathcal{G}_{j-1}}(I_u, I_u) \hat{C}_{w_j,u}^T|. \tag{13}$$

The term inside the determinant can be recovered in $\mathcal{O}(m^2 q)$ or $\mathcal{O}(mq^2)$ time, if $m > q$ or $m < q$, respectively. In addition, evaluating the determinant of the above matrix is an $\mathcal{O}(m^3)$ operation. So, overall the complexity of calculating the ratios is $\mathcal{O}(m \max\{m, q\}^2)$ as compared to that of the standard calculation which is $\mathcal{O}(d^3)$.

TABLE I
SPEEDUPS ACHIEVED BY SPARSITY

| | Speedup | Arbitrary Walk | Forward Walk (see Sec. IV) |
|---|---|---|---|
| $m > q$ | propagation | – | $\mathcal{O}(d/p)$ |
| | exploration | $\mathcal{O}(N)$ | $\mathcal{O}((d/m)^3)$ |
| | update | $\mathcal{O}((d/m)^2)$ | $\mathcal{O}(d/q)$ |
| $m < q$ | exploration | $\mathcal{O}(N)$ | $\mathcal{O}((d/q)^3)$ |
| | update | $\mathcal{O}((d/q)^2)$ | $\mathcal{O}(d/q)$ |

Unfortunately, as Eq. (13) implies we need to know $\Sigma_{w_j|\mathcal{G}_{j-1}}$ at the begining of each greedy step for the consideration of each measurement's information content in set $\mathcal{V}_{w_j}$. As we showed before, it is much more beneficial to evaluate the precision rather than the covariance matrix in an update step. However, the covariance of $\mathsf{X}_{w_j}$ given measurements $\mathcal{G}_{j-1}$ can be recovered by inverting $J_{w_j|\mathcal{G}_{j-1}}$ from the previous step, in $\mathcal{O}(d^3)$ time. We need to do the inversion at every greedy step resulting in a total complexity of $\mathcal{O}(Tkd^3)$. After we obtain the covariance matrix, the evaluation of the reward of each measurement is accomplished in $\mathcal{O}(mq^2)$ time (assuming $m < q$), resulting in a total of $\mathcal{O}(TkNmq^2)$ for all iterations. Assuming that $m, q \ll d, N$, the overall complexity of exploration is dominated by $\mathcal{O}(Tkd^3)$ as opposed to $\mathcal{O}(TkNd^3)$ of the standard calculations. For example, if $d = N = 10^4$, the savings are on the order of $10^4$. We summarize the gains we obtain in the update and exploration steps by making use of sparsity in Table I.

## IV. FORWARD WALKS

Forward walks, walks with non-decreasing orders, are a special case as planning is accomplished solely with forward propagation. While the evaluation complexity of such walks is low, they tend to produce significantly lower information rewards. However, such walks are still of use in that they provide tighter upper bounds on the optimal solution. We assume that the dynamics matrix $A$ is sparse. In other words, each row has at most $p$ non-zero elements. The sparsity in the dynamics matrix $A$ affords a computational advantage for such walks. It is straightforward to see that propagation requires $\mathcal{O}(d^3)$ time. We can take advantage of sparsity, by storing the indicator matrix $I_a$, which contains the non-zero elements of each row of $A$. The above operation requires $\mathcal{O}(d^2)$ time, or $\mathcal{O}(Td^2)$ if it is a time-varying model. Assuming we know the covariance $\Sigma_{t-1|t-1}$, we can evaluate the elements of $\Sigma_{t|t-1}$ as follows:

$$\begin{aligned} \Sigma_1(i,:) &= A_{t-1}(i, I_a(i,:)) \Sigma_{t-1|t-1}(I_a(i,:),:), \forall i \\ \Sigma_2(:,\ell) &= \Sigma_1(:, I_a(\ell,:)) A_{t-1}(\ell, I_a(\ell,:))^T, \forall \ell \\ \Sigma_{t|t-1} &= \Sigma_2 + Q_{t-1}, \end{aligned}$$

where $\Sigma_1, \Sigma_2$ are temporary $d \times d$ matrices. The above evaluation completes in $\mathcal{O}(pd^2)$ time, where $p \ll d$ and so is much faster than $\mathcal{O}(d^3)$ of the standard calculation. Overall, we need only propagate at the transition points, which leads

to an overall complexity of $\mathcal{O}(Tpd^2)$. Interestingly, the worst-case complexity of a forward walk does not depend on the composition of the walk but rather the length of the chain.

Here, the information form provides no computational advantage since Kalman filtering is sufficient for uncertainty propagation. Furthermore, matrix inversion is unnecessary for measurement selection. From Eq. (12), if we denote by $\mathcal{N}(u)$ the (latent) neighbors of measurement $u$ we have that

$$\begin{aligned} g_j &= \underset{u \in \mathcal{V}_{w_j} \setminus \mathcal{G}_{j-1}}{\arg\max} \; I(\mathsf{X}_{w_j}; \mathsf{Y}_{w_j,u} \mid \mathsf{Y}_{\mathcal{G}_{j-1}}) \\ &\overset{(a)}{=} \underset{u \in \mathcal{V}_{w_j} \setminus \mathcal{G}_{j-1}}{\arg\max} \; I(\mathsf{X}_{w_j,\mathcal{N}(u)}; \mathsf{Y}_{w_j,u} \mid \mathsf{Y}_{\mathcal{G}_{j-1}}) \\ &= \underset{u \in \mathcal{V}_{w_j} \setminus \mathcal{G}_{j-1}}{\arg\max} \; \log \frac{|\Sigma_{w_j|\mathcal{G}_{j-1}}(I_u, I_u)|}{|\Sigma_{w_j|\{u\}\cup\mathcal{G}_{j-1}}(I_u, I_u)|}, \end{aligned}$$

where $(a)$ holds since conditioned on the neighbors of measurement $u$, the remaining hidden variables gain no information from $u$. If we know $\Sigma_{w_j|\mathcal{G}_{j-1}}$, $\Sigma_{w_j|\{u\}\cup\mathcal{G}_{j-1}}(I_u, I_u)$ can be evaluated from Eq. (8) by just focusing on the block that corresponds to the (latent) parents $I_u$. If $m > q$, the dominant calculation is the inversion of an $m \times m$ matrix, which is $\mathcal{O}(m^3)$. If $m < q$, the dominant operation is the evaluation of the determinant of a $q \times q$ matrix, which is $\mathcal{O}(q^3)$. The complexity of each exploration step using the standard approach is $\mathcal{O}(d^3)$. Lastly, if we take advantage of sparsity in the moment form, the complexity per update is $\mathcal{O}(mqd)$, while the standard update complexity is $\mathcal{O}(md^2)$. Table I summarizes these results. By way of example, if $m = p = 10, q = 5, d = 10^4$, we gain a speedup of order $10^9$ over the standard approach.

## V. Single Node Decompositions

This section discusses the savings we obtain in the propagation step by taking advantage of a variant of belief propagation that we present below. The update and exploration steps discussed in Sec. III assumed the knowledge of $J_{w_j|\mathcal{G}_{j-1}}$ (or $\Sigma_{w_j|\mathcal{G}_{j-1}}$) at each greedy step, which is obtained during propagation. While the analysis below is restricted to Gaussian HMMs, it extends to trees and general Gaussian MRFs straightforwardly.

The naïve way to do this is to first propagate forward the covariance up to the node corresponding to the first element of the walk ($w_1$) and greedily select the measurement from that node. In general, having selected a measurement, we need to propagate forward from the current node up to the maximum walk element encountered so far and then smooth back to the node corresponding to the next walk element. Each filtering step requires $\mathcal{O}(d^3)$ time assuming $k_t$ is comparatively smaller than $d$. Similarly, each smoothing step requires $\mathcal{O}(d^3)$ steps. Filtering is bounded by $\mathcal{O}(Td^3)$, while smoothing is bounded by $\mathcal{O}\left((T - w_{j+1})d^3\right)$. In the worst case, $w_j \neq w_{j+1}, \forall j$, so we need to use Kalman filtering and smoothing at every step of the greedy algorithm. Since, there are $kT$ steps (assuming $k_t = k$), the worst-case overall complexity is $\mathcal{O}(kT^2d^3)$. In the above approach, the farther the next walk element is from the node with the maximum index encountered so far, the greater the number of unnecessary computations. However, the increase in the total information reward only depends on the covariance update to the node corresponding to the next walk element

$$\begin{aligned} &I(\mathsf{X}; \mathsf{Y}_{w_{j+1},u} \mid \mathsf{Y}_{\mathcal{G}_j}) = I(\mathsf{X}_{w_{j+1}}; \mathsf{Y}_{w_{j+1},u} \mid \mathsf{Y}_{\mathcal{G}_j}) \\ &= \frac{1}{2} \log \frac{|\Sigma_{w_{j+1}|\mathcal{G}_j}|}{|\Sigma_{w_{j+1}|\{u\}\cup\mathcal{G}_j}|} = \frac{1}{2} \log \frac{|J_{w_{j+1}|\{u\}\cup\mathcal{G}_j}|}{|J_{w_{j+1}|\mathcal{G}_j}|}. \end{aligned}$$

Consequently, computations may be focused on the single node that corresponds to the next walk element $w_{j+1}$ after having properly updated its uncertainty from the previous step. Using the information form is preferrable since uncertainty propagation from the current node $w_j$ to the next walk element $w_{j+1}$ is accomplished via a modified version of the Gaussian belief propagation (GaBP) as we describe in the next section.

### A. Adaptive BP

It is well known that a Gaussian HMM may be described by node and edge potentials as

$$\begin{aligned} \varphi_t(x_t) &= \exp\left(-\frac{1}{2} x_t^T J_{t,t} x_t + x_t^T h_t\right) \\ \psi_{t,t+1}(x_t, x_{t+1}) &= \exp\left(-x_t^T J_{t,t+1} x_{t+1}\right), \text{ where} \end{aligned}$$

$$h_t = Q_{t-1}^{-1}\mu_{v,t-1} - A_t^T Q_t^{-1}\mu_{v,t} + C_t^T R_t^{-1}(y_t - \mu_{w,t}) \tag{14}$$

$$J_{t,t} = Q_{t-1}^{-1} + A_t^T Q_t^{-1} A_t + C_t^T R_t^{-1} C_t \tag{15}$$

$$J_{t,t+1} = -A_t^T Q_t^{-1}. \tag{16}$$

where node $\varphi_t$ and edge potentials $\psi_{t,t+1}$ have information parameters $(h_t, J_{t,t}), (0, J_{t,t+1})$, respectively. The "forward" and "backward" passes take the form of belief propagation messages [16], [17]:

Forward Pass

$$h_{t\to t+1} = -J_{t,t+1}^T (J_{t,t} + J_{t-1\to t})^{-1}(h_t + h_{t-1\to t}) \;, \forall t \tag{17}$$

$$J_{t\to t+1} = -J_{t,t+1}^T (J_{t,t} + J_{t-1\to t})^{-1} J_{t,t+1} \tag{18}$$

Backward Pass

$$h_{t\to t-1} = -J_{t-1,t} (J_{t,t} + J_{t+1\to t})^{-1}(h_t + h_{t+1\to t}) \;, \forall t \tag{19}$$

$$J_{t\to t-1} = -J_{t-1,t} (J_{t,t} + J_{t+1\to t})^{-1} J_{t-1,t}^T \tag{20}$$

where $J_{0\to 1} = 0, h_{0\to 1} = 0$, $J_{T+1\to T} = 0, h_{T+1\to T} = 0$.

Finally, the marginals $\mathsf{X}_t$ given all the measurements, $\mathsf{X}_t \mid \mathsf{Y}_{1:T} \sim \mathcal{N}^{-1}(h_{t|T}, J_{t|T})$, are obtained as:

$$\begin{aligned} h_{t|T} &= h_t + h_{t-1\to t} + h_{t+1\to t} \\ J_{t|T} &= J_{t,t} + J_{t-1\to t} + J_{t+1\to t}. \end{aligned}$$

*1) Precision updates with standard BP:* Assuming noise is independent across different measurements within an observation set, $R_t$ is an $N_t m \times N_t m$ block-diagonal matrix. Each block is of size $m \times m$. Incorporating an $m \times 1$ measurement $u$ from set $\mathcal{V}_t$, only affects $h_t, J_{t,t}$ according to Eqs. (14), (15), (16) as:

$$h'_t = h_t + C_t(u,:)^T R_t(u,u)^{-1}(y_t(u) - \mu_{w,t}(u))$$
$$J'_{t,t} = J_{t,t} + C_t(u,:)^T R_t(u,u)^{-1} C_t(u,:).$$

The above operation completes in $\mathcal{O}(m \max(m,q)^2)$ time. Since only the node potential of the current walk element $(h_{w_j}, J_{w_j,w_j})$ changes after the incorporation of measurement $g_j$ from observation set $\mathcal{V}_{w_j}$, we need only propagate messages forward from node $w_j$ to the end of the chain and backwards from node $w_j$ back to the beginning of the chain, which takes exactly $T-1$ steps. Each forward/backward pass results in $\mathcal{O}(d^3)$ complexity and hence the overall complexity is $\mathcal{O}(Td^3)$. In order to find the updated covariance of the next walk element $\Sigma_{w_{j+1}|\mathcal{G}_j}$ for use in Eq (13) (if we replace $j$ by $j+1$), we need to invert the precision matrix $J_{w_{j+1}|\mathcal{G}_j}$,

$$J_{w_{j+1}|\mathcal{G}_j} = J_{w_{j+1}} + J_{w_{j+1}-1 \to w_{j+1}} + J_{w_{j+1}+1 \to w_{j+1}},$$

which adds another $\mathcal{O}(d^3)$ complexity.

*2) Precision updates with adaptive BP:* As our reward computation only requires the covariance of the next walk element, we can further reduce the complexity by restricting updates to the nodes between the current and next walk element. Namely, if $w_j < w_{j+1}$, we only need to update the forward messages from $w_j$ to $w_{j+1}$. Similarly, if $w_j > w_{j+1}$, we only need to update the backward messages from $w_j$ to $w_{j+1}$. We call this modified version of belief propagation (BP), *adaptive BP*.

### *B. Description of the method*

At initialization ($\mathcal{G}_0$), we evaluate all node potentials assuming no measurements are available propagating messages along the entire chain in both directions. As a new measurement is selected ($g_j$) from set $\mathcal{V}_{w_j}$, we compute and update messages from $X_{w_j}$ to $X_{w_{j+1}}$. This results in correct node marginals along that path. Within same walk segment, ($w_j = w_{j+1}$), propagation is unnecessary and we need only update the node potential $(h_{w_j}, J_{w_j,w_j})$ Please, see Alg. 1 for details. Since relevant node potentials are correct at every iteration, and the message from $w_j - 1$ to $w_j$ (if $w_j < w_{j+1}$), or $w_j + 1$ to $w_j$ (if $w_j > w_{j+1}$) is correct, then the messages from $w_j$ to $w_{j+1}$ are guaranteed to be correct as well. Messages from $w_j - 1$ to $w_j$ (or $w_j + 1$ to $w_j$) are also guaranteed to be correct, since during the previous walk step, from $w_{j-1}$ to $w_j$, that message was either not included in the path from $w_{j-1}$ to $w_j$ and was unchanged or has been correctly updated (as part of the directed message schedule from $w_{j-1}$ to $w_j$). Please, see Fig. 3 for an example flow of the algorithm.

At every iteration, we need to update exactly $|w_{j+1} - w_j|$ messages. Therefore, the overall complexity is $\mathcal{O}\left(\sum_{j=1}^{M-1} |w_{j+1} - w_j| d^3\right)$. If we denote by $\bar{\ell} \triangleq \frac{1}{M-1} \sum_{j=1}^{M-1} |w_{j+1} - w_j|$, the average length of the path connecting two nodes of neighboring walk elements, the average-case complexity becomes $\mathcal{O}\left(kT\bar{\ell}d^3\right)$. Compare this with the complexity of the naïve approach which is $\mathcal{O}\left(kT^2d^3\right)$. If $\bar{\ell} \ll T$, we achieve a speedup on the order of $\mathcal{O}(T/\bar{\ell})$ compared to Kalman filtering/smoothing or standard BP. Small $\bar{\ell}$ implies that there are only short jumps in the walk. In other words, it is "cheaper" to obtain the marginal of the node of the next walk element when there is a small distance from the current node. On the other hand, nodes farther from the current walk element generally give higher information gain at the cost of more intensive computation. Thus, there is a trade-off between obtaining higher incremental information rewards and reducing the complexity of computations.

**Algorithm 1** ADAPTIVE BP

Initialization (no measurements)

**if** $j = 1$ **then**

$\{h_{t\to t+1}, J_{t\to t+1}\}_{t=1}^{T-1}$, $\{h_{t\to t-1}, J_{t\to t-1}\}_{t=2}^{T}$ from Eqs. (17), (18), (19), (20).

**end if**

Update node potential of the current walk element

$h_{w_j} = h_{w_j} + C_{w_j}(g_j,:)^T R_{w_j}(g_j,g_j)^{-1} Y_{w_j}(g_j) - \mu_{w,w_j}(g_j)$

$J_{w_j,w_j} = J_{w_j,w_j} + C_{w_j}(g_j,:)^T R_{w_j}(g_j,g_j)^{-1} C_{w_j}(g_j,:)$

Update messages between $X_{w_j}$ and $X_{w_{j+1}}$

**if** $w_j < w_{j+1}$ **then**

$(h_{t\to t+1}, J_{t\to t+1})$ from Eqs. (17), (18), $t = w_j, \ldots, w_{j+1} - 1$

**else if** $w_j > w_{j+1}$ **then**

$(h_{t\to t-1}, J_{t\to t-1})$ from (19), (20), $t = w_j, \ldots, w_{j+1} + 1$

**end if**

Evaluate the parameters of $p(X_{w_{j+1}} \mid \mathcal{G}_j)$

$h_{w_{j+1}|\mathcal{G}_j} = h_{w_{j+1}} + h_{w_{j+1}-1\to w_{j+1}} + h_{w_{j+1}+1\to w_{j+1}}$

$J_{w_{j+1}|\mathcal{G}_j} = J_{w_{j+1},w_{j+1}} + J_{w_{j+1}-1\to w_{j+1}} + J_{w_{j+1}+1\to w_{j+1}}$

## VI. COMPUTATION WALK TREE

The number of different walks for this class of problems is $\frac{(\sum_{t=1}^{T} k_t)!}{k_1!k_2!\cdots k_T!}$, and hence consideration of all walks is intractable. However, many walks share subsets of elements, suggesting redundant computations. For instance, if a partial walk is comprised of $\ell_1, \ldots, \ell_T$ measurements from sets $\mathcal{V}_1, \ldots, \mathcal{V}_T$, there are $\frac{(\sum_{t=1}^{T}(k_t - \ell_t))!}{(k_1-\ell_1)!(k_2-\ell_2)!\cdots(k_T-\ell_T)!}$ walks that start with this partial walk. Greedy selection to time $T$ is the same for all the different walks. We can create at most $T$ tree-like structures with roots $1, 2, \ldots, T$. Each path from a root to a leaf would represent a walk satisfying the constraints. If a partial walk has been evaluated starting from observation set $\mathcal{V}_t$ and ending in set $\mathcal{V}_{w_j}$, then we can evaluate any path of the subtree with root $w_j$ without repeating any of the computations of the path from root $t$ to node $w_j$, as shown in Fig. 4.

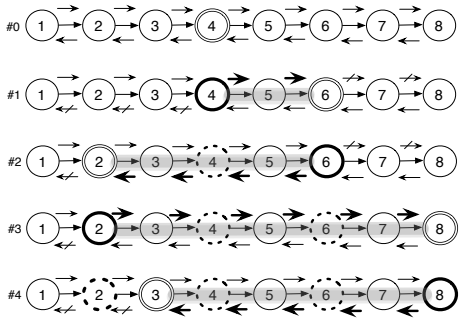



Fig. 3. **Adaptive BP.** Solid thick node represents the node of current walk element under consideration, while the node with double stroke the next one. Dashed nodes represent nodes whose measurements have been obtained in the past. The current node's potentials are updated after the incorporation of a measurement. Thick arrows represent the messages that are transmitted in the current iteration. Solid and strikethrough arrows represent correct messages and messages not being properly updated, respectively. Gray bands encompass all the nodes whose marginals can be correctly computed. During initialization (iter #0), all node potentials, forward and backward messages are computed.

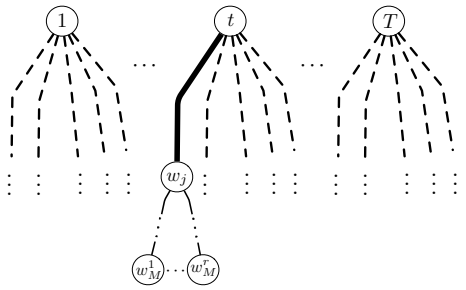



Fig. 4. **Computation Walk Tree.** Each path from a root to a leaf represents one of the walks that satisfy the same set of constraints. If a partial walk has been evaluated (e.g., one starting from set $\mathcal{V}_t$ and ending in set $\mathcal{V}_{w_j}$), there is no need to repeat the computations of this part for all the walks that start from set $\mathcal{V}_t$ and pass through set $\mathcal{V}_{w_j}$.

## VII. Experimental Results

We consider a synthetic tracking experiment. The primary goal of our experiments is to demonstrate the utility of the method from a computational perspective. We additionally observe that in some cases, information rewards - depending on the structure of the problem - may be decoupled from the complexity of walk. In such cases, exploration may be restricted to low-complexity walks while yielding high information rewards. The properties under which this condition arises remains an open question.

In our setup, there are three objects. Two of the objects move away from the third one. Each object has a 6-dimensional states, $p_x, p_y, p_z, v_x, v_y, v_z$ representing the positions and velocities along the three axes. We consider the following linear state-space model:

$$\begin{aligned} \mathsf{X}_t &= A_{t-1}\mathsf{X}_{t-1} + \mathsf{V}_{t-1}, \forall t \in \{1, \ldots, 20\} \\ Y_t &= C_t\mathsf{X}_t + \mathsf{W}_t \ , \end{aligned}$$

where $A_{t-1}$ captures linear dynamics, $\mathsf{V}_{t-1} \sim \mathcal{N}(0, Q_{t-1})$ is driving noise and $\mathsf{W}_t \sim \mathcal{N}(0, R_t)$ is measurement noise.

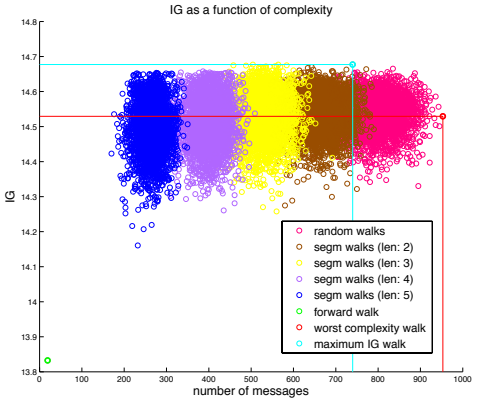



Fig. 5. **Empirical Analysis of Structured and Unstructured Walks.** The figure compares information rewards versus computation complexity (as quantified by number of messages) for walks of different minimum size segment (magenta:1, brown:2, yellow:3, purple:4, blue:5). As we see, the distribution of rewards is similar to all cases, but the evaluation complexity is significantly lower for walks with harder constraints on the minimum segment length.

Potential measurements are available for each hidden variable (position, velocity), which accounts to 18 measurements per time point (6 per object) of which we are may select six (at each time point). We consider five different types of walks with the following minimum sizes of a walk segment; $\ell_{\min} \in \{1, 2, 3, 4, 5\}$. By minimum size, we mean that for every set $\mathcal{V}_t$, there is a walk segment with size at least $\ell_{\min}$.[2] In Fig. 5 we compare rewards of different walks to their complexities. As expected, the forward walk (green circle) has lowest complexity and lowest reward. Interestingly, even though walks with larger minimum segment sizes (blue, purple) result in much lower complexities, they have comparable rewards to those of lower minimum segment sizes (brown, magenta) and higher complexity. Additionally, the walk with the maximum reward which belongs in one of the highest complexity clusters is not significantly higher than the maximum from lower complexity clusters.

We also examine speed up due to sparsity by considering 200 moving objects with different degrees of correlated motion. The hidden dimension in this case is $d = 1200$. We consider different observation sizes, constituting $\{10\%, 25\%, 50\%, 75\%, 100\%\}$ of the hidden dimension and different degrees of sparsity in the measurement model. Fig. 6 shows the efficiency gain as a function of sparsity and observation size when using the information form. Here, color indicates factor of speedup, the maximum being 1400.

Lastly, we examine the advantage of adaptive BP. We construct 10 Markov chains of varying length (from $T = 10$ to $T = 300$) and compare adaptive BP to the standard Kalman filtering and smoothing. In Fig. 7a, certain speedups are obtained converging to a single number for reasons explained in the caption. We also construct multiple walks with specified average distance between consecutive walk elements, (1–

[2]A *walk segment* is a subset of the walk where all elements are the same.

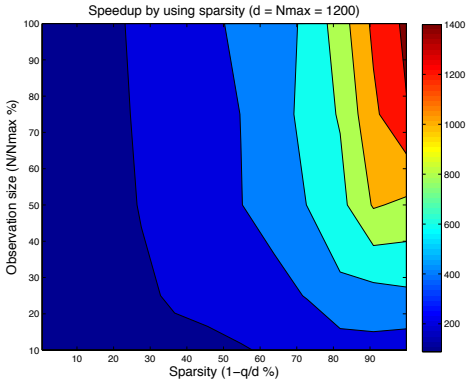
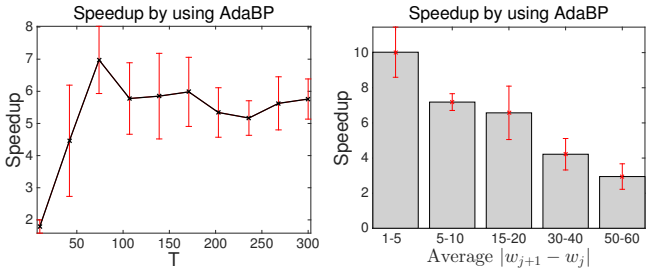



Fig. 6. **Speedup by taking advantage of sparsity.** The figure shows the speedup we gain by taking sparsity into account and working with the information form. We explore the speedup for different degrees of sparsity defined as $1 - q/d$ and different ratios of observation sizes $N$, to hidden dimension $d$. As expected, we see that the gains are more imminent as observation size and sparsity grows.




(a) (b)

Fig. 7. **Sparsity and different walks.** (a) Efficiency gains as the length of the chain $T$ increases. Gains stabilize around a single number due to the construction of the walks. Since we choose the same number of measurements from each observation set, and observation sets are sampled uniformly with probability $1/T$, the mean distance between two consecutive points is $T/3$. Kalman filtering and smoothing send $2T$ messages on average at every iteration, since we need to propagate to the end of the chain and then smooth back to the node that corresponds to the next walk element $w_{j+1}$. Therefore, the expected speedup converges to a number close to $\frac{2T}{T/3} = 6$. (b) As the average distance between consecutive walk elements decreases, the speedups increase.

5, 5–10, 15–20, 50–60) for a chain of length $T = 100$. As the average distance between consecutive walk elements decreases, the greater the advantage of Adaptive BP.

## VIII. Conclusion

We have considered the problem of efficient evaluation of information rewards in Gaussian HMMs. The analysis extends straightforwardly to trees and with some modifications to general MRFs. Naïve evaluation of such rewards is generally prohibitive for all but forward walks. However, in the case of sparse measurement matrices, a detailed analysis has shown that significant computational savings are available. Consequently, in such situations it is computationally feasible to explore multiple plans leading to improved information rewards and tighter upper bounds on the optimal reward. Furthermore, our experimental results reveal that in some cases the information reward can be decoupled from the complexity of the walk. As a result, exploration can be restricted to low-complexity walks while still yielding high information rewards that are guaranteed to be within a computable factor of the (intractable) optimal solution.

## Acknowledgements

The authors would like to thank the reviewers for their valuable comments and suggestions. This work was partially supported by the Army Research Office (ARO) Multidisciplinary Research Initiative (MURI) program (Award number W911NF-11-1-0391), NSF/DNDO Collaborative Research ARI-LA (Award ECCS-1348328), and by the Department of Energy (NA-22) Consortium for Verification Technology.

## References

[1] D. V. Lindley, "On a measure of the information provided by an experiment," *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 986–1005, December 1956.

[2] D. Blackwell, "Comparison of experiments," in *Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, Ed. Berkeley, CA: Statistical Laboratory of the University of California, Berkeley, August 1950, pp. 93–102.

[3] J. M. Bernardo, "Expected Information as Expected Utility," *The Annals of Statistics*, vol. 7, no. 3, pp. 686–690, May 1979.

[4] M. Basseville, "Distance Measures for Signal Processing and Pattern Recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, Dec 1989.

[5] P. L. Bartlett, M. J. Jordan, and J. D. Mcauliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, 2003.

[6] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *Annals of Statistics*, 2009.

[7] F. Zhao, J. Shin, and J. Reich, "Information-driven Dynamic Sensor Collaboration," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 61–72, March 2002.

[8] E. Ertin, J. W. Fisher III, and L. C. Potter, "Maximum Mutual Information Principle for Dynamic Sensor Query Problems," in *Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks (IPSN)*, February 2003, pp. 405–416.

[9] C. Kreucher, K. Kastella, and A. Hero, "Sensor management using an active sensing approach," *Signal Processing*, vol. 85, no. 3, pp. 607–624, 2005.

[10] J. L. Williams, J. W. Fisher III, and A. S. Willsky, "Approximate Dynamic Programming for Communication-Constrained Sensor Network Management," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 3995–4003, August 2007.

[11] S. Fujishige, *Submodular Functions and Optimization*, ser. Annals of discrete mathematics. Elsevier, 2005.

[12] A. Krause and C. Guestrin, "Near-optimal Nonmyopic Value of Information in Graphical Models," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2005.

[13] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An Analysis of Approximations for Maximizing Submodular Set Functions - I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[14] J. L. Williams, J. W. Fisher III, and A. S. Willsky, "Performance Guarantees for Information Theoretic Active Inference," in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, March 2007.

[15] Y. Liu, V. Chandrasekaran, A. Anandkumar, and A. S. Willsky, "Feedback Message Passing for Inference in Gaussian Graphical Models," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4135–4150, Aug 2012.

[16] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach," in *Proceedings of the American Association of Artificial Intelligence National Conference (AAAI)*, 1982, pp. 133–136.

[17] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, pp. 2173–2200, 2001.