ON THE COMPLEXITY OF INFORMATION PLANNING IN GAUSSIAN MODELS

Georgios Papachristoudis and John W. Fisher III

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA, 02139
Email: geopapa@mit.edu, fisher@csail.mit.edu

ABSTRACT

We analyze the complexity of evaluating information rewards for measurement selection in sparse graphical models under the assumption that measurements are drawn from a limited number of nodes subject to a finite budget. Previous analyses [1, 2, 3] exploit the submodular property of conditional mutual information to demonstrate that greedy measurement selection come with near-optimal guarantees As noted in [4] typical formulations assume *oracle value* models. However, [1, 2, 5] allude to a more significant source of complexity, namely *computing* the measurement reward. Here, we focus on Gaussian models and show that by exploiting sparsity in the measurement model, the complexity of planning is substantially reduced. We also demonstrate that by utilizing the information form additional significant reductions in complexity may be realized.

***Index Terms*—** active learning, Gaussian HMMs, Kalman filtering and smoothing, experimental design, belief propagation

1. INTRODUCTION

There is a significant history on the use of information measure as a reward function for Bayesian experimental design, *e.g.*, [6, 7]. More recently [8, 9, 10, 11, 12] formulate such measures in the sequential inference setting subject to selection constraints where the complexity is combinatorial in the number of sensing actions and exponential in the planning horizon. As observed in [2], mutual information (MI), is *submodular* [13] when measurements are independent conditioned on the quantity of interest. Consequently, the results of [14] provide that tractable greedy selection guarantees information rewards to be within a factor of the optimal (though, intractable) selection.

The analysis of [3] provides more general guarantees for greedy selection in the case (applicable to probabilistic graphical models) where measurement subsets combined with subset selection constraints and for which the latent variable structure may not be fully specified à priori (*e.g.*, inference in Markov chains for streaming data). We stress that the analysis addresses *planning* sensing actions in contrast to *inference* (or data fusion). The latter follows the former as a consequence of a resulting choice of sensing actions. The distinction lies in the the difference between the *measurement* and its *value*. Planning considers the *expected* utility of a set of measurements while inference incorporates their values. Evaluation of information rewards across different measurement plans is a key component to any planning method.

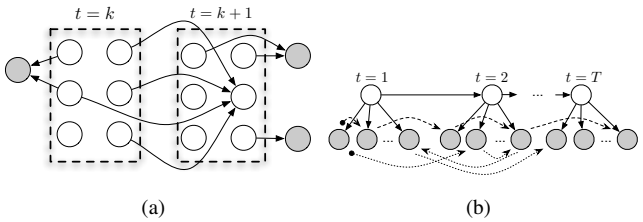


Fig. 1. **Sparsity and different walks.** (a) Measurements of X_k depend only on a few components of X_k. Dashed rectangles represent vectors of latent variables, X_k, X_{k+1}. (b) This structure can be represented as an HMM. Two walks are visualized. Arrows with a circle in one end denote the beginning of the walk. Different walks visit each observation set the same number of times, but in different orders.

An important, often neglected, aspect of information-based approaches, is the computational complexity of evaluating a given plan. In realistic settings, it can be a source of significant computational complexity. For Gaussian models, complexity primarily depends on the dimension of the latent variables, the number of measurements to be explored and the order in which different observation sets are visited. Here, we consider the computational complexity of evaluating the information rewards in active inference problems described by Markov chains, trees, and poly-trees. We show how to reduce the computational load substantially by taking advantage of sparsity in the measurement process and by utilizing the information form. This analysis allows one to trade off the computational complexity of exploring additional plans with the probability of achieving a higher information reward.

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 Royal Dutch Shell through the MIT Energy Initiative.

Consider the problem of selecting an optimal k-element subset (measurements for our purposes) from a ground set $\mathcal{V}$ of size N that maximizes some reward f. Due to combinatorial complexity, optimal solutions are intractable for even moderately large problems. However, [14] shows that if the function f is *submodular*, *i.e.,* it is monotone and has the property of "diminishing returns", then a greedy selection algorithm with $\mathcal{O}(kN)$ complexity achieves a reward that is a factor no worse than $(1 - 1/e)$ of the optimal reward. However, this result implicitly assumes an *oracle value* model, *i.e.*, that the reward function for any given subset can be computed in constant time. Specifically, the *oracle value* model assumes a universal "oracle" that provides the function value for any input set. Subsequent work [15], which generalizes [14] to matroidal structures, also assumes an oracle value model.

There have been a number of methods utilizing the preceding. As noted, [1, 2] consider information planning in the batch setting while [5] considers influential subset selection in social networks. The original bound is improved upon in [16]. Online resource allocation networks [17, 18], stochastic submodular maximization [19] , the study of the submodular welfare problem. [20], and additional extensions to submodular maximization [21, 22, 23] comprise just a small sampling of approaches and analyses which exploit the results of [14, 15]. All assume oracle value models.

However, as we discuss, evaluation of rewards presents a significant computational challenge. As such, [1] proposes truncation methods as an approximation for Gaussian models. Similarly, [2] notes that the evaluation of conditional entropies, intrinsic to greedy selection, can be prohibitive, while [5] acknowledge the complexity in evaluating the underlying influence function that guides the selection of the most influential nodes. We are not aware of previous results that exploit the structure of the latent graph to reduce the complexity of reward evaluations. Despite the inherent computational bottleneck, it is often overlooked. Here, we show for Gaussian models that exact computations are feasible by taking advantage of the graph structure and utilizing a variant of belief propagation (BP) [24] more suited for active learning settings.

2. INFORMATION PLANNING IN GAUSSIAN HMMS

For simplicity, we restrict ourselves to hidden Markov Models (HMMs). However, the results extend easily to trees and poly-trees. A Gaussian HMM can be described by the following dynamical system

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

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

where $\mathsf{X}_t, \mathsf{Y}_t$ are the respective latent and observed variables, and $\mathsf{V}_t \sim \mathcal{N}(0, Q_t)$, $\mathsf{W}_t \sim \mathcal{N}(0, R_t)$, are the respective process and measurement noise. R_t is assumed block-diagonal. Let $\mathsf{X} = \{\mathsf{X}_1, \ldots, \mathsf{X}_T\}$ be the set of latent variables up to time T. Each X_t is a d-dimensional vector. For each X_t, we define an *observation set*, $\mathcal{V}_t$, 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. Y_t represents all N_t measurements of set $\mathcal{V}_t$. As such, C_t is an $N_t m \times d$ matrix. As depicted in Fig. 1(a), we assume that a measurement depends at most on q elements of vector X_t. Consequently, C_t is a sparse matrix with non-zero row elements at locations corresponding to the q elements of X_t upon which a measurement depends. The assumption of sparsity is common in realistic settings (*e.g.*, in tracking applications, where noisy position measurements depend only upon a subset of the full kinematic state).

Constrained Selection over Measurement Subsets. We frame the problem as follows: given multiple subsets of measurements and constraints on the total number of measurements that can be selected from each subset (note that the constraint may differ for each subset), we wish to select the set of measurements which maximizes the expected information gain with respect to the full set of latent variables. This (generally intractable) combinatorial optimization problem is stated more formally as:

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

where $f(\cdot)$ is a set function representing the reward of a set and k_t are the selection constraints for the t-th set. We restrict ourselves to monotonic functions. A *walk* $\boldsymbol{w} = \{w_1, \ldots, w_M\}$ denotes the particular order in which observation sets are visited, which corresponds to a feasible solution of the above problem. That is, the walk defines an order that satisfies the selection constraints at all times as defined in Eq. (3). *Greedy* methods sequentially select elements that maximize the incremental value of the reward function conditioned on the previous selection. The incremental reward of a measurement u given set $\mathcal{S}$ is defined as $f(u \mid \mathcal{S}) \triangleq f(\{u\} \cup \mathcal{S}) - f(\mathcal{S})$. Greedy selection for a particular walk $\boldsymbol{w}$ 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 incremental reward of incorporating measurement u is $f(u \mid \mathcal{G}_{j-1}) = f(\{u\} \cup \mathcal{G}_{j-1}) - f(\mathcal{G}_{j-1})$. Here, we use MI as the reward function, $f(\mathcal{S}) = I(\mathsf{X}; \mathsf{Y}_\mathcal{S})$. The resulting incremental reward takes the form $f(u \mid \mathcal{G}_{j-1}) = I(\mathsf{X}; \mathsf{Y}_u \mid \mathsf{Y}_{\mathcal{G}_{j-1}})$. W.l.o.g., we assume that each observation set has the same size $N_t = N$ and identical selection constraints $k_k = k$.

To give an indication of the hardness of problem (3), there are $\prod_t \binom{N_t}{k_t} = (\frac{N!}{(N-k)!k!})^T$ feasible solutions (that satisfy the selection constraints), which is an extremely large number as N, k and T grow. On the contrary, the complexity of determining a greedy solution for a *given* walk is only $\mathcal{O}(\sum_t N_t k_t) = \mathcal{O}(kTN)$ assuming an oracle value model. In addition, there are $\binom{\sum_t k_t}{k_1,\ldots,k_t} = \frac{(kT)!}{k!^T}$ different walks. For every such walk, there are $\prod_t N_t!/(N_t - k_t)! =$

$(N!/(N-k)!)^T$ permutations of measurements that form feasible solutions. One such permutation (for a given walk) corresponds to the greedy solution as dictated by Eq. (4). Therefore, there are $\frac{(kT)!}{k!^T}$ greedy solutions, one for each different walk. Remarkably, Williams et al. [3] showed that the greedy solution as dictated by Eq. (4) results in a reward which is no worse than half of the optimal for *any* walk. While the performance guarantee holds for all feasible walks, it is the case that some walks may yield significantly better information rewards than others. Additionally, as our analysis shows, the complexity of evaluating information rewards cannot be done in constant time and varies considerably for different walks. This motivates developing efficient methods for evaluating different walks in order to obtain higher information rewards.

Gaussian HMMs. The information reward for Gaussian models are expressed as a function of the covariance matrix of the underlying process. In this case, the incremental reward of a measurement u is

$$f(u \mid \mathcal{G}_{j-1}) = H(\mathsf{X} \mid \mathsf{Y}_{\mathcal{G}_{j-1}}) - H(\mathsf{X} \mid \mathsf{Y}_u, \mathsf{Y}_{\mathcal{G}_{j-1}}). \quad (5)$$

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

Covariance updates in Gaussian HMMs are computed as:

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

$$\Sigma_{t|t} = \Sigma_{t|t-1} - G_t C_t \Sigma_{t|t-1} \quad (8)$$

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

where $\Sigma_{t|t-1} = \mathrm{cov}(\mathsf{X}_t \mid \mathsf{Y}_{1:t-1})$, $\Sigma_{t|t} = \mathrm{cov}(\mathsf{X}_t \mid \mathsf{Y}_{1:t})$. We refer to Eqs. (7), (8) 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 $\Sigma_{t|t}$. That is, *marginal* covariance (equivalently precision) updates are sufficient for quantifying the accumulated information reward with respect to the full set of latent variables.

3. COMPLEXITY REDUCTION VIA SPARSITY

There are two primary sources of computational complexity that arise when evaluating the information reward of a given walk: **(i) exploration**, in which the information rewards of the remaining measurements of the current observation set are computed; and **(ii) update**, in which the selected measurement is incorporated into the *marginal* covariance of the latent variable representing the current element of the walk. Greedy algorithms proceed as follows; (i) propagate uncertainty given the existing selections to the variable corresponding to the next walk element, (ii) evaluate the remaining candidate measurements in the observation set of the current walk element, choosing the one with maximum incremental reward, and lastly (iii) update the uncertainty at that variable after having selected the measurement. The complexity of the exploration and update steps depend on the structure of matrices C_t. We assume that the dimension of each measurement vector is much smaller than the latent dimension at all time points, *i.e.*, $m \ll d$. We denote the m-row portion of matrix C_t corresponding to measurement $\mathsf{Y}_{t,u}$ as $C_t(u,:)$. For notational consistency with the analysis later in the text, we will denote the greedily selected measurement g_j by u.

In Eq. (8), the complexity of the update step is $\mathcal{O}(md^2)$, as computation is dominated by $C_{w_j}(u,:)\Sigma_{w_j|\mathcal{G}_{j-1}}$. A total of kT updates yields overall update complexity of $\mathcal{O}(kTmd^2)$. Exploration of a single measurement as indicated by Eq. (6), takes $\mathcal{O}(d^3)$ time since it requires the computation of a $d \times d$ matrix determinant. Combined with the number of measurements that should be considered, which is $\mathcal{O}(kTN)$, the exploration component of greedy selection results in total complexity of $\mathcal{O}(kTNd^3)$. When $m \ll N, d$, exploration dominates the complexity of greedy selection with d^3 being the most significant term. However, the preceding ignores the sparsity of matrices C_t. Evaluation complexity can be reduced dramatically by taking advantage of this sparsity. Let I_c denote the indicator matrix of the non-zero elements of C_t. While the sparsity pattern defined by I_c may differ with t, we omit additional notation to account for that. Computation of I_c requires $\mathcal{O}(Nmd)$ time or $\mathcal{O}(TNmd)$, if time-varying.

Reductions during updates. For a measurement u of dimension m, I_u represents the nodes of latent graph X upon which u depends. If Σ' (equivalently J') represents the updated covariance (equivalently precision) after the incorporation of measurement u and Σ (equivalently J) the prior, Eq. (8) becomes:

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

$$J' = \Sigma'^{-1} = \underbrace{\Sigma^{-1}}_{J} + C(u,:)^T R(u,u)^{-1}C(u,:), \quad (9)$$

where we have made use of the Woodbury matrix identity. We denote by $\hat{C}_u$ the $m \times q$ matrix, $\hat{C}_u = R(u,u)^{-1/2}C(u,I_u)$. The matrix square root, $R(u,u)^{-1/2}$ and $\hat{C}_u$ can be recovered in $\mathcal{O}(m^3)$ and $\mathcal{O}(m^2 q)$ time, respectively. Therefore, Eq. (9) can be rewritten as

$$J' = J + \begin{bmatrix} \hat{C}_u^T \\ 0^T \end{bmatrix}\begin{bmatrix}\hat{C}_u & 0\end{bmatrix} \Rightarrow J'(I_u,I_u) = J(I_u,I_u) + \hat{C}_u^T\hat{C}_u, \quad (10)$$

since only the block of J dictated by I_u will be affected by $\hat{C}_u^T\hat{C}_u$. The above calculations are dominated by the term $\mathcal{O}(m\max\{m,q\}^2)$. Compare this to the complexity of the standard calculation, $\mathcal{O}(md^2)$, where $d \gg m, q$, translating to a speedup on the order of $\mathcal{O}((\frac{d}{\max\{m,q\}})^2)$.

Reductions during exploration. Greedy selection for Gaussian models simplifies to

$$g_j = \underset{u\in\mathcal{V}_{w_j}\setminus\mathcal{G}_{j-1}}{\arg\max} \underbrace{I(\mathsf{X}_{w_j}; \mathsf{Y}_{w_j,u} \mid \mathsf{Y}_{\mathcal{G}_{j-1}})}_{\log(|J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}|/|J_{w_j|\mathcal{G}_{j-1}}|)}, \quad (11)$$

where $J_{w_j|\mathcal{G}_{j-1}}$, $J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}$ is the precision of X_{w_j} before and after the incorporation of measurement u, respectively. By substituting $J' = J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}$ and $J = J_{w_j|\mathcal{G}_{j-1}}$ in Eq.

(10), we have:

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

As we observe in Eq. (11), the greedy step only requires ratios of determinants. Therefore, if we use the Matrix Determinant Lemma on $J_{w_j|\{u\}\cup\mathcal{G}_{j-1}}$ in Eq. (11), we obtain

$$g_j = \arg\max_{u\in\mathcal{V}_{w_j}} \log|\mathbb{I}_{m\times m} + \hat{C}_{w_j,u}\Sigma_{w_j|\mathcal{G}_{j-1}}(I_u, I_u)\hat{C}^T_{w_j,u}|. \tag{12}$$

The above computation takes $\mathcal{O}(m\max\{m,q\}^2)$ time and is orders of magnitude faster than the standard calculation with complexity $\mathcal{O}(d^3)$.

4. SINGLE NODE DECOMPOSITIONS

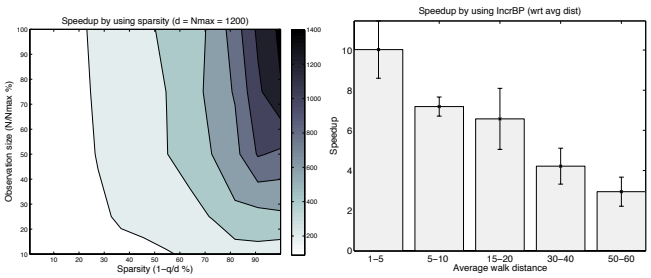
Upon incorporating a measurement in the current walk element $X_{w_{j-1}}$, one must propagate uncertainty $\Sigma_{w_j|\mathcal{G}_{j-1}}$ to the next element, X_{w_j}, as shown by Eq. (12). The standard approach is via Kalman filtering and smoothing with complexity linear in T, per greedy step. We obtain further complexity reductions utilizing the information form of belief propagation (BP) [24, 25] where at each step we propagate messages solely from the current to the next walk element. We refer to this variant of BP as *incremental BP*. The result is that only the latent variables which are necessary to compute information rewards are updated with cost relative to the distance between consecutive elements, rather than the width of the Markov chain, as shown in Fig. 2(b).

5. EXPERIMENTAL RESULTS

The primary goal of our experiment is to demonstrate the utility of the method from a computational perspective. We consider 200 moving objects with different degrees of correlated motion. Each object has a 6-dimensional state, $p_x, p_y, p_z, v_x, v_y, v_z$ representing the positions and velocities along the three axes. The problem is modeled with the following linear state-space model:

$$X_t = A_{t-1}X_{t-1} + V_{t-1}, \forall t \in \{1,\dots,20\}$$
$$Y_t = C_tX_t + W_t\ ,$$

where $X_t = [\ p^{1:200}_{t,(x,y,z)} \quad v^{1:200}_{t,(x,y,z)}\ ]^T$, A_{t-1} captures linear dynamics, $V_{t-1} \sim \mathcal{N}(0, Q_{t-1})$ is driving and $W_t \sim \mathcal{N}(0, R_t)$ measurement noise. Potential measurements are available for each latent variable (position, velocity), accounting to $N_t^{\max} = 1200$ measurements per time point (6 per object) of which we may select $k_t = 6$ (per point). The latent dimension is $d = 1200$. We consider different observation sizes, constituting $\{10\%, 25\%, 50\%, 75\%, 100\%\}$ of the latent dimension and different degrees of sparsity in the measurement model. Fig. 2(a) shows the gain as a function of sparsity and observation size when. Lastly, we examine the advantage of working with the information form of BP. We construct multiple walks with specified average distance between consecutive walk elements: (1–5, 5–10, 15–20, 50–60)



(a) (b)

Fig. 2. **Speedups from sparsity and single node decompositions.** (a) We explore the speedup for different degrees of sparsity defined as $1 - q/d$ and different ratios of observation size to latent dimension N/d. As expected, gains are more imminent as observation size and sparsity grow. (b) The figure shows how gains change as the average distance between consecutive walk elements increases. As expected, we see that when the average distance between consecutive walk elements is low, we gain significant speedups.

for a chain of length $T = 100$ and compare against standard Kalman filtering and smoothing. As the average distance between consecutive walk elements decreases, the computational advantage of incremental BP is much greater.

6. CONCLUSION

We have considered the problem of efficient evaluation of information rewards in Gaussian HMMs. We argue that the assumption of oracle value models to such problems is questionable due to the considerable computational costs of reward evaluations. We showed that the major computational bottleneck of the greedy process is in the exploration step, which is quartic on the observation size, N_t (assuming N_t and d are comparable). We further demonstrated that by using sparsity and proper manipulation of the covariance formulas, we can evaluate the exact reward while substantially reducing the computational cost from $\mathcal{O}(kTNd^3)$ to $\mathcal{O}(kTNm\max\{m,q\}^2)$, which is only linear in N – the largest parameter along with d.

7. REFERENCES

[1] C. Guestrin, A. Krause, and A. Singh, "Near-optimal sensor placements in gaussian processes," in *International Conference on Machine Learning (ICML)*, August 2005.

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

[3] 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.

[4] A. Krause and D. Golovin, "Submodular function maximization," in *Tractability: Practical Approaches to Hard Problems (to appear)*. Cambridge University Press, February 2014.

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the Spread of Influence Through a Social Network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2003, KDD '03, pp. 137–146, ACM.

[6] 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.

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

[8] K. J. Hintz and E. S. McVey, "Multi-process constrained estimation," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, no. 1, pp. 237–244, 1991.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

[15] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions–II," in *Polyhedral combinatorics*, M.L. Balinski and A.J. Hoffman, Eds., vol. 8 of *Mathematical programming*, pp. 73–87. Elsevier, 1978.

[16] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a Submodular Set Function Subject to a Matroid Constraint," in *Proceedings of the 12th International Conference on Integer Programming and Combinatorial Optimization*, Berlin, Heidelberg, 2007, IPCO '07, pp. 182–196, Springer-Verlag.

[17] M. J. Streeter and D. Golovin, "An Online Algorithm for Maximizing Submodular Functions," in *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, December 2008, pp. 1577–1584.

[18] M. Streeter, D. Golovin, and A. Krause, "Online Learning of Assignments," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 1794–1802.

[19] A. Asadpour, H. Nazerzadeh, and A. Saberi, "Stochastic Submodular Maximization," in *Proceedings of the 4th International Workshop on Internet and Network Economics, WINE*, December 2008, pp. 477–489.

[20] J. Vondrák, "Optimal Approximation for the Submodular Welfare Problem in the Value Oracle Model," in *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, New York, NY, USA, 2008, STOC '08, pp. 67–74, ACM.

[21] C. Chekuri, J. Vondrák, and R. Zenklusen, "Submodular Function Maximization via the Multilinear Relaxation and Contention Resolution Schemes," pp. 783–792, 2011.

[22] S. O. Gharan and J. Vondrák, "Submodular Maximization by Simulated Annealing," in *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, January 2011, pp. 1098–1116.

[23] Y. Filmus and J. Ward, "Monotone Submodular Maximization over a Matroid via Non-Oblivious Local Search," *SIAM J. Comput.*, vol. 43, no. 2, pp. 514–542, 2014.

[24] 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.

[25] 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.