



UNIVERSITEIT
VAN
AMSTERDAM

IAS technical report IAS-UVA-14-02

Exploiting Submodular Value Functions for Faster Dynamic Sensor Selection: Extended Version

Yash Satsangi¹, Shimon Whiteson¹, and Frans A. Oliehoek^{1,2}

¹Intelligent Systems Laboratory Amsterdam, University of Amsterdam, The Netherlands

²Dept of CS, University of Liverpool

A key challenge in the design of multi-sensor systems is the efficient allocation of scarce resources such as bandwidth, CPU cycles, and energy, leading to the *dynamic sensor selection* problem in which a subset of the available sensors must be selected at each timestep. While *partially observable Markov decision processes* (POMDPs) provide a natural decision-theoretic model for this problem, the computational cost of POMDP planning grows exponentially in the number of sensors, making it feasible only for small problems. We propose a new POMDP planning method that uses *greedy maximization* to greatly improve scalability in the number of sensors. We show that, under certain conditions, the value function of a dynamic sensor selection POMDP is *submodular* and use this result to bound the error introduced by performing greedy maximization. Experimental results on a real-world dataset from a multi-camera tracking system in a shopping mall show it achieves similar performance to existing methods but incurs only a fraction of the computational cost, leading to much better scalability in the number of cameras. *This paper is an extended version of [Satsangi et al., 2015] including all the proofs and further experimental details that were omitted in the shorter version.*

Keywords: Sensor selection, POMDPs.

IAS

intelligent autonomous systems

Contents

1	Introduction	1
2	Background	2
2.1	POMDPs	2
2.2	Dynamic Sensor Selection POMDPs	3
2.3	Point-Based Value Iteration	4
3	Greedy PBVI	4
4	Analysis: Bounds given Submodularity	5
5	Analysis: Submodularity under Belief Entropy	7
6	Analysis: Approximate Belief Entropy	11
7	Experiments	13
8	Related Work	16
9	Conclusions & Future Work	16

Intelligent Autonomous Systems
Informatics Institute, Faculty of Science
University of Amsterdam
Science Park 904, 1098 XH Amsterdam
The Netherlands
Tel (fax): +31 20 525 7463
<http://isla.science.uva.nl/>

Corresponding author:
Yash Satsangi
tel: +31 20 525 8516
y.satsangi@uva.nl

1 Introduction

Multi-sensor systems are becoming increasingly prevalent in a wide range of settings. For example, multi-camera systems are now routinely used for security, surveillance, and tracking. A key challenge in the design of such systems is the efficient allocation of scarce resources such as the bandwidth required to communicate the collected data to a central server, the CPU cycles required to process that data, and the energy costs of the entire system. This gives rise to the *dynamic sensor selection* problem [Spaan and Lima, 2009, Kreucher et al., 2005, Williams et al., 2007]: selecting, based on the system’s current uncertainty about its environment, K of the N available sensors to use at each timestep, where K is the maximum number of sensors allowed given the resource constraints.

When the state of the environment is static, a *myopic* approach that always selects the sensors that maximize the immediate expected reduction in uncertainty is typically sufficient. However, when that state changes over time, a non-myopic approach that reasons about the long-term effects of the sensor selection performed at each step can perform better. A natural decision-theoretic model for such an approach is the *partially observable Markov decision process* (POMDP) [Aström, 1965, Smallwood and Sondik, 1973, Kaelbling et al., 1998] in which actions specify different subsets of sensors.

In a typical POMDP, reducing uncertainty about the state is only a means to an end. For example, in a robot control task, the robot aims to determine its current location so it can more easily reach its goal. However, dynamic sensor selection is a type of *active perception* problem [Spaan, 2008, Spaan and Lima, 2009], which can be seen as a subclass of POMDPs in which reducing uncertainty is an end in itself. For example, a surveillance system’s goal is typically just to ascertain the state of its environment, not use that knowledge to achieve another goal. While perception is arguably always performed to aid decision-making, in an active perception problem that decision is made by another agent, e.g., a human, not modeled by the POMDP.

Although POMDPs are computationally expensive to solve, approximate methods such as point-based planners [Pineau et al., 2006, Araya et al., 2010] have made it practical to solve POMDPs with large state spaces. However, dynamic sensor selection poses a different challenge: as the number of sensors N grows, the size of the action space $\binom{N}{K}$ grows exponentially. Consequently, as the number of sensors grows, solving the POMDP even approximately quickly becomes infeasible with existing methods.

In this paper, we propose a new point-based planning method for dynamic sensor selection that scales much better with the number of sensors. The main idea is to replace maximization with *greedy maximization* [Nemhauser et al., 1978, Golovin and Krause, 2011, Krause and Golovin, 2014] in which a subset of sensors is constructed by iteratively adding the sensor that gives the largest marginal increase in value. Doing so avoids iterating over the entire action space, yielding enormous computational savings.

In addition, we present theoretical results bounding the error in the value functions computed by this method. Our core result is that, under certain conditions including *submodularity* [Krause and Golovin, 2014, Nemhauser et al., 1978], the value function computed using POMDP backups based on greedy maximization has bounded error. We also show that such conditions are met, or approximately met, if reward is defined using negative *belief entropy* or an approximation thereof. To our knowledge, these are the first results demonstrating the submodularity of value functions and bounding the error of greedy maximization in the full POMDP setting.

This is a corrected version of this paper. The original version contained a technical mistake in the proof of Lemma 5. We would like to thank Csaba Szepesvári for identifying this mistake.

Finally, we apply our method to a real-life dataset from a multi-camera tracking system with thirteen cameras installed in a shopping mall. Our empirical results demonstrate that our approach outperforms a myopic baseline and nearly matches the performance of existing point-based methods while incurring only a fraction of the computational cost.

2 Background

In this section, we provide background on POMDPs, dynamic sensor selection POMDPs, and point-based methods.

2.1 POMDPs

A POMDP is a tuple $\langle S, A, \Omega, T, O, R, b_0, \gamma, h \rangle$. At each timestep, the environment is in a state $s \in S$, the agent takes an action $a \in A$ and receives a reward whose expected value is $R(s, a)$, and the system transitions to a new state $s' \in S$ according to the transition function $T(s, a, s') = \Pr(s'|s, a)$. Then, the agent receives an observation $z \in \Omega$ according to the observation function $O(s', a, z) = \Pr(z|s', a)$. The agent can maintain a belief $b(s)$ using Bayes rule. Given $b(s)$ and $R(s, a)$, the *belief-based* reward, $\rho(b, a)$ is:

$$\rho(b, a) = \sum_s b(s)R(s, a). \quad (1)$$

A policy π specifies how the agent will act for each belief. The value $V_t^\pi(b)$ of π given t steps to go until the horizon h is given by the *Bellman equation*:

$$V_t^\pi(b) = \rho(b, a_\pi) + \gamma \sum_{z \in \Omega} \Pr(z|a_\pi, b)V_{t-1}^\pi(b^{z, a_\pi}). \quad (2)$$

The action-value function $Q_t^\pi(b, a)$ is the value of taking action a and following π thereafter:

$$Q_t^\pi(b, a) = \rho(b, a) + \gamma \sum_{z \in \Omega} \Pr(z|a, b)V_{t-1}^\pi(b^{z, a}). \quad (3)$$

The optimal value function $V_t^*(b)$ is given by the *Bellman optimality equation*:

$$\begin{aligned} V_t^*(b) &= \max_a Q_t^*(b, a) \\ &= \max_a [\rho(b, a) + \gamma \sum_{z \in \Omega} \Pr(z|a, b)V_{t-1}^*(b^{z, a})]. \end{aligned} \quad (4)$$

We can also define the *Bellman optimality operator* \mathfrak{B}^* :

$$(\mathfrak{B}^*V_{t-1})(b) = \max_a [\rho(b, a) + \gamma \sum_{z \in \Omega} \Pr(z|a, b)V_{t-1}(b^{z, a})], \quad (5)$$

and write (4) as: $V_t^*(b) = (\mathfrak{B}^*V_{t-1}^*)(b)$.

An important consequence of (1) is that V_t^* is *piecewise linear and convex* (PWLC). This property, which is exploited by most POMDP planners, allows V_t^* to be represented by a set of vectors: $\Gamma_t = \{\alpha_1, \alpha_2 \dots \alpha_m\}$, where each α -vector is an $|S|$ -dimensional hyperplane representing $V_t^*(b)$ in a bounded region of belief space. The value function can then be written as $V_t^*(b) = \max_{\alpha_i} \sum_s b(s)\alpha_i(s)$.

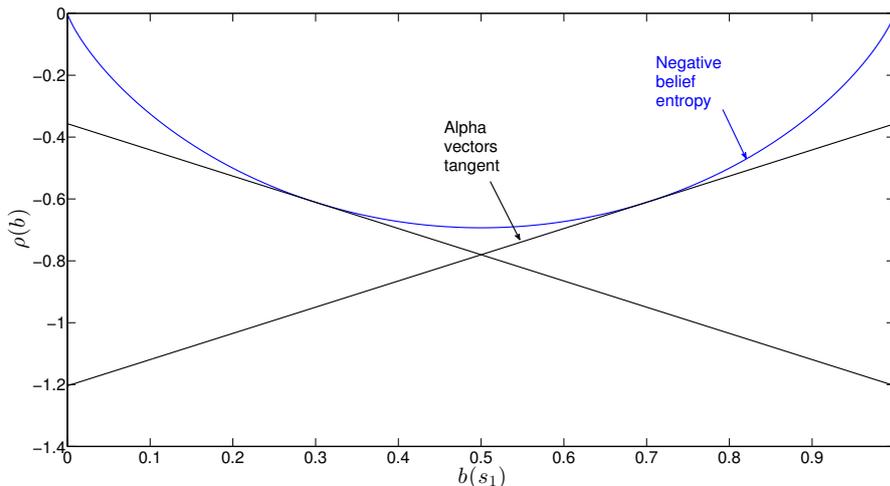


Figure 1: Tangents approximating negative belief entropy.

2.2 Dynamic Sensor Selection POMDPs

We model the dynamic sensor selection problem as a POMDP in which the agent must choose a subset of available sensors at each timestep. We assume that all selected sensors must be chosen simultaneously, i.e., it is not possible within a timestep to condition the choice of one sensor on the observation generated by another sensor. This corresponds to the common setting in which generating each sensor’s observation is time consuming, e.g., because it requires applying expensive computer vision algorithms, and thus all observations must be generated in parallel. Formally, a dynamic sensor selection POMDP has the following components:

- Actions $\mathbf{a} = \langle a_1 \dots a_N \rangle$ are modeled as vectors of N binary *action features*, each of which specifies whether a given sensor is selected or not (assuming N sensors). For each \mathbf{a} , we also define its set equivalent $\mathbf{a} = \{i : a_i = 1\}$, i.e., the set of indices of the selected sensors. Due to the resource constraints, the set of all actions $A = \{\mathbf{a} : |\mathbf{a}| \leq K\}$ contains only sensor subsets of size K or less. $A^+ = \{1, \dots, N\}$ indicates the set of all sensors.
- Observations $\mathbf{z} = \langle z_1 \dots z_N \rangle$ are modeled as vectors of N *observation features*, each of which specifies the sensor reading obtained by the given sensor. If sensor i is not selected, then $z_i = \emptyset$. The set equivalent of \mathbf{z} is $\mathfrak{z} = \{z_i : z_i \neq \emptyset\}$. To prevent ambiguity about which sensor generated which observation in \mathfrak{z} , we assume that, for all i and j , the domains of z_i and z_j share only \emptyset .
- The transition function $T(s', s) = \Pr(s'|s)$ is independent of \mathbf{a} because the agent’s role is purely observational.
- The belief-based reward $\rho(b)$ is also independent of \mathbf{a} and is typically some measure of the agent’s uncertainty. A natural choice is the negative entropy of the belief: $\rho(b) = -H_b(s) = -\sum_s p(s) \log(p(s))$. However, this definition destroys the PWLC property. Instead, we approximate $-H_b(s)$ using a set of vectors $\Gamma^\rho = \{\alpha_1^\rho, \dots, \alpha_m^\rho\}$, each of which is a tangent to $-H_b(s)$, as suggested by [Araya et al., 2010]. Figure 1 shows the tangents for an example Γ^ρ for a two-state POMDP. Because these tangents provide a PWLC approximation to belief entropy, the value function is also PWLC and can thus be computed using standard solvers.

2.3 Point-Based Value Iteration

Exact POMDP planners [Smallwood and Sondik, 1973, Monahan, 1982, Lovejoy, 1991, Kaelbling et al., 1998] compute the optimal Γ_t -sets for all possible belief points. However, this approach is intractable for all but small POMDPs. By contrast, *point-based value iteration* (PBVI) [Pineau et al., 2006] achieves much better scalability by computing the Γ_t -sets only for a set of sampled beliefs B , yielding an approximation of V_t^* .

At each iteration, PBVI computes Γ_t given Γ_{t-1} as follows. The first step is to generate intermediate $\Gamma_t^{\mathbf{z},\mathbf{a}}$ -sets for all $\mathbf{a} \in A$ and $\mathbf{z} \in \Omega$: $\Gamma_t^{\mathbf{z},\mathbf{a}} = \{\alpha^{\mathbf{z},\mathbf{a}} : \alpha \in \Gamma_{t-1}\}$, where

$$\alpha^{\mathbf{z},\mathbf{a}}(s) = \gamma \sum_{s' \in S} T(s, s') O(s', \mathbf{a}, \mathbf{z}) \alpha(s').$$

The next step is to use the intermediate sets to generate sets $\Gamma_t^{\mathbf{a}} = \{\alpha_{\mathbf{a},b} : b \in B\}$, where

$$\alpha_{\mathbf{a},b} = \arg \max_{\alpha \in \Gamma^{\rho}} \sum_s b(s) \alpha^{\rho}(s) + \sum_{\mathbf{z}} \arg \max_{\alpha^{\mathbf{z},\mathbf{a}} \in \Gamma_t^{\mathbf{z},\mathbf{a}}} \sum_s \alpha^{\mathbf{z},\mathbf{a}}(s) b(s).$$

The final step is to find the best vector for each $b \in B$ and thus generate Γ_t . To facilitate explanation of our algorithm in the following section, we describe this final step somewhat differently than Pineau et al. (2006). For each $b \in B$ and $\mathbf{a} \in A$ we must find the best $\alpha_{\mathbf{a},b} \in \Gamma_t^{\mathbf{a}}$:

$$\alpha_{\mathbf{a},b}^* = \arg \max_{\alpha_{\mathbf{a},b} \in \Gamma_t^{\mathbf{a}}} \sum_s \alpha_{\mathbf{a},b}(s) b(s), \quad (6)$$

and simultaneously record its value: $Q(b, \mathbf{a}) = \sum_s \alpha_{\mathbf{a},b}^*(s) b(s)$. Then, for each $b \in B$, we find the best vector across all actions: $\alpha_b = \alpha_{\mathbf{a}^*,b}^*$, where

$$\mathbf{a}^* = \arg \max_{\mathbf{a} \in A} Q(b, \mathbf{a}). \quad (7)$$

Finally, Γ_t is the union of these vectors: $\Gamma_t = \cup_{b \in B} \alpha_b$.

3 Greedy PBVI

The computational complexity of one iteration of PBVI is $O(|S||A||\Gamma_{t-1}||\Omega||B|)$ [Pineau et al., 2006]. While this is only linear in $|A|$, in our setting $|A| = \binom{N}{K}$. Thus, PBVI's complexity is $O(|S| \binom{N}{K} |\Gamma_{t-1}||\Omega||B|)$, leading to poor scalability in N , the number of sensors. In this section, we propose *greedy PBVI*, a new point-based POMDP planner for dynamic sensor selection whose complexity is only $O(|S||N||K||\Gamma_{t-1}||\Omega||B|)$, enabling much better scalability in N .

The main idea is to exploit *greedy maximization* [Nemhauser et al., 1978], an algorithm that operates on a set function $F : 2^X \rightarrow \mathbb{R}$. Algorithm 1 shows the argmax variant, which constructs a subset $Y \subseteq X$ of size K by iteratively adding elements of X to Y . At each iteration, it adds the element, e that maximally increases $F(Y)$.

Algorithm 1 greedy-argmax(F, X, K)

```

 $Y \leftarrow \emptyset$ 
for  $m = 1$  to  $K$  do
   $Y \leftarrow Y \cup \{\arg \max_{e \in X \setminus Y} F(Y \cup e)\}$ 
end for
return  $Y$ 

```

To exploit greedy maximization in PBVI, we need to replace an argmax over A with **greedy-argmax**. Our alternative description of PBVI above makes this straightforward: (7) contains such an

argmax , and $Q(b, \cdot)$ has been intentionally formulated to be a set function over A^+ . Thus, implementing greedy PBVI requires only replacing (7) with:

$$\mathbf{a}^* = \operatorname{greedy-argmax}(Q(b, \cdot), A^+, K). \quad (8)$$

Note that, since the point of greedy maximization is not to iterate over A , it is crucial that our implementation does not first compute $\alpha_{\mathbf{a}, b}^*$ and $Q(b, \mathbf{a})$ for all $\mathbf{a} \in A$, as this would already introduce an $|A| = \binom{N}{K}$ term into the complexity. Instead, $\alpha_{\mathbf{a}, b}^*$ and $Q(b, \mathbf{a})$ are computed on-the-fly only for the \mathbf{a} 's considered by $\operatorname{greedy-argmax}$. Since the complexity of $\operatorname{greedy-argmax}$ is only $O(|N||K|)$, this yields a complexity for greedy PBVI of only $O(|S||N||K||\Gamma_{t-1}||\Omega||B|)$. Note also that the $\alpha^{\mathbf{z}, a}$ that are generated can be cached because they are not specific to a given b and can thus be reused.

Greedy maximization can only be exploited by methods that perform explicit maximization to compute V_t^* . Exact methods that do not perform an explicit maximization and instead rely on pruning operators, e.g., [Cassandra et al., 1997, Kaelbling et al., 1998], cannot directly employ greedy maximization. Fortunately, point-based methods, which are fast and effective approximate solvers, are based on explicit maximization. These methods are thus an ideal starting point for our approach, and can be easily modified by using $\operatorname{greedy-argmax}$ instead of the regular argmax .

4 Analysis: Bounds given Submodularity

In this section, we present our core theoretical result, which shows that, under certain conditions, the most important of which is *submodularity*, the error in the value function computed by backups based on greedy maximization is bounded. Later sections discuss when reward based on negative belief entropy or an approximation thereof meets those conditions.

Submodularity is a property of set functions that corresponds to diminishing returns, i.e., adding an element to a set increases the value of the set function by a smaller or equal amount than adding that same element to a subset. In our notation, this is formalized as follows. The set function $Q_t^\pi(b, \mathbf{a})$ is submodular in \mathbf{a} , if for every $\mathbf{a}_M \subseteq \mathbf{a}_N \subseteq A^+$ and $a_e \in A^+ \setminus \mathbf{a}_N$,

$$\Delta_{Q_b}(a_e | \mathbf{a}_M) \geq \Delta_{Q_b}(a_e | \mathbf{a}_N), \quad (9)$$

where $\Delta_{Q_b}(a_e | \mathbf{a}) = Q_t^\pi(b, \mathbf{a} \cup \{a_e\}) - Q_t^\pi(b, \mathbf{a})$ is the *discrete derivative* of $Q_t^\pi(b, \mathbf{a})$. Equivalently, $Q_t^\pi(b, \mathbf{a})$ is submodular if for every $\mathbf{a}_M, \mathbf{a}_N \subseteq A^+$,

$$Q_t^\pi(b, \mathbf{a}_M \cap \mathbf{a}_N) + Q_t^\pi(b, \mathbf{a}_M \cup \mathbf{a}_N) \leq Q_t^\pi(b, \mathbf{a}_M) + Q_t^\pi(b, \mathbf{a}_N). \quad (10)$$

Submodularity is an important property because of the following result by Nemhauser et al. (1978):

Theorem 1. *Given any policy π , if $Q_t^\pi(b, \mathbf{a})$ is non-negative, monotone and submodular in \mathbf{a} , then for all b ,*

$$Q_t^\pi(b, \mathbf{a}^G) \geq (1 - e^{-1})Q_t^\pi(b, \mathbf{a}^*), \quad (11)$$

where $\mathbf{a}^G = \operatorname{greedy-argmax}(Q_t^\pi(b, \cdot), A^+, K)$ and $\mathbf{a}^* = \operatorname{arg max}_{\mathbf{a} \in A} Q_t^\pi(b, \mathbf{a})$.

However, Theorem 1 gives a bound only for a single application of $\operatorname{greedy-argmax}$, not for applying it within each backup, as greedy PBVI does. In this section, we establish such a bound. Let the *greedy Bellman operator* \mathfrak{B}^G be:

$$(\mathfrak{B}^G V_{t-1})(b) = \max_{\mathbf{a}}^G [\rho(b, \mathbf{a}) + \gamma \sum_{\mathbf{z} \in \Omega} \Pr(\mathbf{z} | \mathbf{a}, b) V_{t-1}(b^{\mathbf{z}, \mathbf{a}})],$$

where $\max_{\mathbf{a}}^G$ refers to greedy maximization. This immediately implies the following corollary to Theorem 1:

Corollary 1. *Given any policy π , if $Q_t^\pi(b, \mathbf{a})$ is non-negative, monotone, and submodular in \mathbf{a} , then for all b ,*

$$(\mathfrak{B}^G V_{t-1}^\pi)(b) \geq (1 - e^{-1})(\mathfrak{B}^* V_{t-1}^\pi)(b). \quad (12)$$

Proof. From Theorem 1 since $(\mathfrak{B}^G V_{t-1}^\pi)(b) = Q_t^\pi(b, \mathbf{a}^G)$ and $(\mathfrak{B}^* V_{t-1}^\pi)(b) = Q_t^\pi(b, \mathbf{a}^*)$. \square

In addition, we can prove that the error in the value function remains bounded after application of \mathfrak{B}^G .

Lemma 1. *Given a policy π , if for all b , $\rho(b) \geq 0$,*

$$V_t^\pi(b) \geq (1 - \epsilon)V_t^*(b), \quad (13)$$

and $Q_t^\pi(b, \mathbf{a})$ is non-negative, monotone, and submodular in \mathbf{a} , then, for $\epsilon \in [0, 1]$,

$$(\mathfrak{B}^G V_t^\pi)(b) \geq (1 - e^{-1})(1 - \epsilon)(\mathfrak{B}^G V_t^*)(b). \quad (14)$$

Proof. Starting from (13) and, for a given \mathbf{a} , on both sides adding $\gamma \geq 0$, taking the expectation over \mathbf{z} , and adding $\rho(b)$ (since $\rho(b) \geq 0$ and $\epsilon \leq 1$):

$$\rho(b) + \gamma \mathbb{E}_{\mathbf{z}|b, \mathbf{a}}[V_t^\pi(b^{\mathbf{z}, \mathbf{a}})] \geq (1 - \epsilon)(\rho(b) + \gamma \mathbb{E}_{\mathbf{z}|b, \mathbf{a}}[V_t^*(b^{\mathbf{z}, \mathbf{a}})]).$$

From the definition of Q_t^π (3), we thus have:

$$Q_{t+1}^\pi(b, \mathbf{a}) \geq (1 - \epsilon)Q_{t+1}^*(b, \mathbf{a}) \quad \forall \mathbf{a}. \quad (15)$$

From Theorem 1, we know

$$Q_{t+1}^\pi(b, \mathbf{a}_\pi^G) \geq (1 - e^{-1})Q_{t+1}^\pi(b, \mathbf{a}_\pi^*), \quad (16)$$

where $\mathbf{a}_\pi^G = \text{greedy-argmax}(Q_{t+1}^\pi(b, \cdot), A^+, K)$ and $\mathbf{a}_\pi^* = \arg \max_{\mathbf{a}} Q_{t+1}^\pi(b, \mathbf{a})$. Since $Q_{t+1}^\pi(b, \mathbf{a}_\pi^*) \geq Q_{t+1}^\pi(b, \mathbf{a})$ for any \mathbf{a} ,

$$Q_{t+1}^\pi(b, \mathbf{a}_\pi^G) \geq (1 - e^{-1})Q_{t+1}^\pi(b, \mathbf{a}_\pi^*), \quad (17)$$

where $\mathbf{a}_\pi^G = \text{greedy-argmax}(Q_{t+1}^\pi(b, \cdot), A^+, K)$. Finally, (15) implies that $Q_{t+1}^\pi(b, \mathbf{a}_\pi^G) \geq (1 - \epsilon)Q_{t+1}^*(b, \mathbf{a}_\pi^G)$, so:

$$\begin{aligned} Q_{t+1}^\pi(b, \mathbf{a}_\pi^G) &\geq (1 - e^{-1})(1 - \epsilon)Q_{t+1}^*(b, \mathbf{a}_\pi^G) \\ (\mathfrak{B}^G V_t^\pi)(b) &\geq (1 - e^{-1})(1 - \epsilon)(\mathfrak{B}^G V_t^*)(b). \end{aligned} \quad \square$$

Next, we define the *greedy Bellman equation*: $V_t^G(b) = (\mathfrak{B}^G V_{t-1}^G)(b)$, where $V_0^G = \rho(b)$. Note that V_t^G is the true value function obtained by greedy maximization, without any point-based approximations. Using Corollary 1 and Lemma 1, we can bound the error of V^G with respect to V^* .

Theorem 2. *If for all policies π , $Q_t^\pi(b, \mathbf{a})$ is non-negative, monotone and submodular in \mathbf{a} , then for all b ,*

$$V_t^G(b) \geq (1 - e^{-1})^{2t} V_t^*(b). \quad (18)$$

Proof. By induction on t . The base case, $t = 0$, holds because $V_0^G(b) = \rho(b) = V_0^*(b)$.

In the inductive step, for all b , we assume that

$$V_{t-1}^G(b) \geq (1 - e^{-1})^{2t-2} V_{t-1}^*(b), \quad (19)$$

and must show that

$$V_t^G(b) \geq (1 - e^{-1})^{2t} V_t^*(b). \quad (20)$$

Applying Lemma 1 with $V_t^\pi = V_{t-1}^G$ and $(1 - \epsilon) = (1 - e^{-1})^{2t-2}$ to (19):

$$\begin{aligned} (\mathfrak{B}^G V_{t-1}^G)(b) &\geq (1 - e^{-1})^{2t-2} (1 - e^{-1}) (\mathfrak{B}^G V_{t-1}^*)(b) \\ V_t^G(b) &\geq (1 - e^{-1})^{2t-1} (\mathfrak{B}^G V_{t-1}^*)(b). \end{aligned}$$

Now applying Corollary 1 with $V_{t-1}^\pi = V_{t-1}^*$:

$$\begin{aligned} V_t^G(b) &\geq (1 - e^{-1})^{2t-1} (1 - e^{-1}) (\mathfrak{B}^* V_{t-1}^*)(b) \\ V_t^G(b) &\geq (1 - e^{-1})^{2t} V_t^*(b). \end{aligned} \quad \square$$

5 Analysis: Submodularity under Belief Entropy

In this section, we show that, when using negative belief entropy as immediate belief-based reward, i.e., $\rho(b) = -H_b(s)$, then under certain conditions $Q_t^\pi(b, \mathbf{a})$ is submodular, non-negative and monotone, as required by Theorem 2. We start by observing that: $Q_t^\pi(b, \mathbf{a}) = \rho(b) + \sum_{k=1}^{t-1} G_k^\pi(b^t, \mathbf{a}^t)$, where $G_k^\pi(b^t, \mathbf{a}^t)$ is the expected immediate reward with k steps to go, conditioned on the belief and action with t steps to go and assuming policy π is followed after timestep t :

$$G_k^\pi(b^t, \mathbf{a}^t) = \gamma^{(h-k)} \sum_{\mathbf{z}^{t:k}} \Pr(\mathbf{z}^{t:k} | b^t, \mathbf{a}^t, \pi) (-H_{b^k}(s^k)).$$

where $\mathbf{z}^{t:k}$ is a vector of observations received in the interval from t steps to go to k steps to go, b^t is the belief at t steps to go, \mathbf{a}^t is the action taken at t steps to go, and $\rho(b^k) = -H_{b^k}(s^k)$, where s^k is the state at k steps to go.

Proving that $Q_t^\pi(b, \mathbf{a})$ is submodular in \mathbf{a} requires three steps. First, we show that $G_k^\pi(b^t, \mathbf{a}^t)$ equals the *conditional entropy* of b^k over s^k given $\mathbf{z}^{t:k}$. Second, we show that, under certain conditions, conditional entropy is a submodular set function. Third, we combine these two results to show that $Q_t^\pi(b, \mathbf{a})$ is submodular.

The *conditional entropy* [Cover and Thomas, 1991] of a distribution b over s given some observations \mathbf{z} is defined as: $H_b(s|\mathbf{z}) = -\sum_s \sum_{\mathbf{z}} \Pr(s, \mathbf{z}) \log(b(s|\mathbf{z}))$. Thus, conditional entropy is the expected entropy given \mathbf{z} has been observed but marginalizing across the values it can take on.

Lemma 2. *If $\rho(b) = -H_b(s)$, then the expected reward at each time step equals the negative discounted conditional entropy of b^k over s^k given $\mathbf{z}^{t:k}$:*

$$G_k^\pi(b^t, \mathbf{a}^t) = -\gamma^{(h-k)} (H_{b^k}(s^k | \mathbf{z}^{t:k})) \forall \pi. \quad (21)$$

Proof. To prove the above lemma, we take help of some additional notations and definitions, first we must elaborate on the definition of b^k :

$$b^k(s^k) \triangleq \Pr(s^k | b^t, \mathbf{a}^t, \pi, \mathbf{z}^{t:k}) = \frac{\Pr(\mathbf{z}^{t:k}, s^k | b^t, \mathbf{a}^t, \pi)}{\Pr(\mathbf{z}^{t:k} | b^t, \mathbf{a}^t, \pi)}. \quad (22)$$

For notational convenience, we also write this as:

$$b^k(s^k) \triangleq \frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})}. \quad (23)$$

The entropy of b^k is thus:

$$H_{b^k}(s^k) = \sum_{s^k} \frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})} \log\left(\frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})}\right), \quad (24)$$

and the conditional entropy of b^k over s^k given $\mathbf{z}^{t:k}$ is:

$$H_{b^k}(s^k|\mathbf{z}^{t:k}) = \sum_{s^k} \sum_{\mathbf{z}^{t:k}} \Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k) \log\left(\frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})}\right).$$

Then, by definition of $G_k^\pi(b^t, \mathbf{a}^t)$,

$$G_k^\pi(b^t, \mathbf{a}^t) = \gamma^{(h-k)} \left(- \sum_{\mathbf{z}^{t:k}} \Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}) H_{b^k}(s^k) \right)$$

By definition of entropy,

$$\begin{aligned} &= \gamma^{(h-k)} \sum_{\mathbf{z}^{t:k}} \Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}) \left[\sum_{s^k} \frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})} \log\left(\frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})}\right) \right] \\ &= \gamma^{(h-k)} \sum_{\mathbf{z}^{t:k}} \left[\sum_{s^k} \Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k) \log\left(\frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})}\right) \right] \\ &= \gamma^{(h-k)} \sum_{s^k} \left[\sum_{\mathbf{z}^{t:k}} \Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k) \log\left(\frac{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k}, s^k)}{\Pr_{b^t, \mathbf{a}^t}^\pi(\mathbf{z}^{t:k})}\right) \right] \end{aligned} \quad (25)$$

By definition of conditional entropy,

$$= \gamma^{(h-k)} (-H_{b^k}(s^k|\mathbf{z}^{t:k})).$$

□

Next, we identify the conditions under which $G_k^\pi(b^t, \mathbf{a}^t)$ is submodular in \mathbf{a}^t . We use the set equivalent \mathfrak{z} of \mathbf{z} since submodularity is a property of set functions. Thus:

$$G_k^\pi(b^t, \mathbf{a}^t) = \gamma^{(h-k)} (-H_{b^k}(s|\mathfrak{z}^{t:k})), \quad (26)$$

where $\mathfrak{z}^{t:k}$ is a set of observation features observed between t and k timesteps to go. The key condition required for submodularity of $G_k^\pi(b^t, \mathbf{a}^t)$ is *conditional independence* [Krause and Guestrin, 2007].

Definition 1. The observation set \mathfrak{z} is conditionally independent given s if any pair of observation features are conditionally independent given the state, i.e.,

$$\Pr(z_i, z_j|s) = \Pr(z_i|s) \Pr(z_j|s), \quad \forall z_i, z_j \in \mathfrak{z}. \quad (27)$$

Lemma 3. If \mathfrak{z} is conditionally independent given s then $-H(s|\mathfrak{z})$ is submodular in \mathfrak{z} , i.e., for any two observations \mathfrak{z}_M and \mathfrak{z}_N ,

$$H(s|\mathfrak{z}_M \cup \mathfrak{z}_N) + H(s|\mathfrak{z}_M \cap \mathfrak{z}_N) \geq H(s|\mathfrak{z}_M) + H(s|\mathfrak{z}_N). \quad (28)$$

Proof.

$$H(s|\mathfrak{z}_M \cup \mathfrak{z}_N) + H(s|\mathfrak{z}_M \cap \mathfrak{z}_N)$$

By chain rule of entropy,

$$= H(\mathfrak{z}_M \cup \mathfrak{z}_N, s) + H(\mathfrak{z}_M \cap \mathfrak{z}_N, s) - H(\mathfrak{z}_M \cup \mathfrak{z}_N) - H(\mathfrak{z}_M \cap \mathfrak{z}_N)$$

By chain rule of entropy,

$$= H(\mathfrak{z}_M \cup \mathfrak{z}_N|s) + 2H(s) + H(\mathfrak{z}_M \cap \mathfrak{z}_N|s) - H(\mathfrak{z}_M \cap \mathfrak{z}_N) - H(\mathfrak{z}_M \cup \mathfrak{z}_N)$$

Using conditional independence

$$= H(\mathfrak{z}_M|s) + H(\mathfrak{z}_N|s) + 2H(s) + H(\mathfrak{z}_M \cap \mathfrak{z}_N|s) - H(\mathfrak{z}_M \cap \mathfrak{z}_N) - H(\mathfrak{z}_M \cup \mathfrak{z}_N) \quad (29)$$

By chain rule of entropy,

$$= H(\mathfrak{z}_M, s) + H(\mathfrak{z}_N, s) + H(\mathfrak{z}_M \cap \mathfrak{z}_N|s) - H(\mathfrak{z}_M \cap \mathfrak{z}_N) - H(\mathfrak{z}_M \cup \mathfrak{z}_N)$$

By chain rule of entropy,

$$= H(s|\mathfrak{z}_M) + H(s|\mathfrak{z}_N) + H(\mathfrak{z}_M \cap \mathfrak{z}_N|s) + [H(\mathfrak{z}_M) + H(\mathfrak{z}_N) - H(\mathfrak{z}_M \cup \mathfrak{z}_N) - H(\mathfrak{z}_M \cap \mathfrak{z}_N)]$$

Since entropy is submodular & positive [Cover and Thomas, 1991]

$$= H(s|\mathfrak{z}_M) + H(s|\mathfrak{z}_N) + \text{a positive term}$$

Since the above expression is the right hand side of (28) plus a positive term, then the left hand side of (28) must be larger. \square

Lemma 4. *If $\mathfrak{z}^{t:k}$ is conditionally independent given s^k and $\rho(b) = -H_b(s)$, then $G_k^\pi(b^t, \mathbf{a}^t)$ is submodular in $\mathbf{a}^t \forall \pi$.*

Proof. Let \mathbf{a}_M^t and \mathbf{a}_N^t be two actions and $\mathfrak{z}_M^{t:k}$ and $\mathfrak{z}_N^{t:k}$ the observations they induce. Then, from Lemma 2,

$$G_k^\pi(b^t, \mathbf{a}_M^t) = \gamma^{(h-k)}(-H_{b^k}(s^k|\mathfrak{z}_M^{t:k})) \quad (30)$$

From Lemma 3,

$$\begin{aligned} H(s^k|\mathfrak{z}_M^{t:k} \cup \mathfrak{z}_N^{t:k}) + H(s^k|\mathfrak{z}_M^{t:k} \cap \mathfrak{z}_N^{t:k}) &\geq H(s^k|\mathfrak{z}_M^{t:k}) + H(s^k|\mathfrak{z}_N^{t:k}) \\ - H(s^k|\mathfrak{z}_M^{t:k} \cup \mathfrak{z}_N^{t:k}) - H(s^k|\mathfrak{z}_M^{t:k} \cap \mathfrak{z}_N^{t:k}) &\leq -H(s^k|\mathfrak{z}_M^{t:k}) - H(s^k|\mathfrak{z}_N^{t:k}) \\ G_k^\pi(b^t, \mathbf{a}_M^t \cup \mathbf{a}_N^t) + G_k^\pi(b^t, \mathbf{a}_N^t \cap \mathbf{a}_M^t) &\leq G_k^\pi(b^t, \mathbf{a}_M^t) + G_k^\pi(b^t, \mathbf{a}_N^t). \end{aligned} \quad (31)$$

\square

Now we can establish the submodularity of Q_t^π .

Theorem 3. *If $\mathfrak{z}^{t:k}$ is conditionally independent given s^k and $\rho(b) = -H_b(s)$, then $Q_t^\pi(b, \mathbf{a}) = \rho(b) + \sum_{k=1}^{t-1} G_k^\pi(b^t, \mathbf{a}^t)$ is submodular in \mathbf{a} , for all π .*

Proof. $\rho(b)$ is trivially submodular in \mathbf{a} because it is independent of \mathbf{a} . Furthermore, Lemma 4 shows that $G_k^\pi(b^t, \mathbf{a}^t)$ is submodular in \mathbf{a}^t . Since a positively weighted sum of submodular functions is also submodular [Krause and Golovin, 2014], this implies that $\sum_{k=1}^{t-1} G_k^\pi(b^t, \mathbf{a}^t)$ and thus $Q_t^\pi(b, \mathbf{a})$ are also submodular in \mathbf{a} . \square

While the conditional independence of \mathfrak{z}^k given s^k is easy to satisfy, the conditional independence of $\mathfrak{z}^{t:k}$, a whole sequence of observations, given s^k is more difficult. For $\mathfrak{z}^{t:k}$ to be conditionally independent given s^k , s^k must contain enough information to predict the past sequence of states $s^{t:k}$. One way to achieve this is by defining s^k such that it encodes all the information in the state history that is correlated with the observations $\mathfrak{z}^{t:k}$. Unfortunately,

this typically is not practical to do unless the transition function is deterministic and invertible. However, note that the conditions required by Theorem 3 are only sufficient, not necessary, conditions for the value function to be submodular. An important goal for future work is thus to identify weaker conditions for establishing submodularity of value functions based on belief entropy or other belief-based reward functions. As we show in Section 7, greedy PBVI performs well in practice even when these conditions do not hold, which suggests that establishing submodularity under weaker conditions may indeed be possible.

While Theorem 3 shows that $Q_t^G(b, \mathbf{a})$ is submodular, Theorem 2 also requires that it be monotone, which we now establish.

Lemma 5. *If V_t^π is convex over the belief space for all t , then $Q_t^\pi(b, \mathbf{a})$ is monotone in \mathbf{a} , i.e., for all b and $\mathbf{a}_M \subseteq \mathbf{a}_N$, $Q_t^\pi(b, \mathbf{a}_M) \leq Q_t^\pi(b, \mathbf{a}_N)$.*

Proof. By definition of $Q_t^\pi(b, \mathbf{a})$,

$$Q_t^\pi(b, \mathbf{a}_M) = [\rho(b) + \gamma \mathbb{E}_{\mathfrak{z}_M | b, \mathbf{a}_M} V_{t-1}^\pi(b^{\mathfrak{z}_M, \mathbf{a}_M})]. \quad (32)$$

Since $\rho(b)$ is independent of \mathbf{a}_M , we need only show that the second term is monotone in \mathbf{a} . Let $\mathbf{a}_P = \mathbf{a}_N \setminus \mathbf{a}_M$ and

$$F_b^\pi(\mathbf{a}_N) = \mathbb{E}_{\mathfrak{z}_N | b, \mathbf{a}_N} V_{t-1}^\pi(b^{\mathfrak{z}_N, \mathbf{a}_N}). \quad (33)$$

Since $\mathbf{a}_N = \{\mathbf{a}_M \cup \mathbf{a}_P\}$,

$$F_b^\pi(\mathbf{a}_N) = \mathbb{E}_{\{\mathfrak{z}_M, \mathfrak{z}_P\} | b, \{\mathbf{a}_M, \mathbf{a}_P\}} V_{t-1}^\pi(b^{\{\mathfrak{z}_M, \mathfrak{z}_P\}, \{\mathbf{a}_M, \mathbf{a}_P\}}). \quad (34)$$

Separating expectations,

$$F_b^\pi(\mathbf{a}_N) = \mathbb{E}_{\mathfrak{z}_M | b, \mathbf{a}_M} \mathbb{E}_{\mathfrak{z}_P | b, \mathbf{a}_P} V_{t-1}^\pi(b^{\{\mathfrak{z}_M, \mathfrak{z}_P\}, \{\mathbf{a}_M, \mathbf{a}_P\}}) \quad (35)$$

Applying Jensen's inequality, since V_{t-1}^π is convex,

$$F_b^\pi(\mathbf{a}_N) \geq \mathbb{E}_{\mathfrak{z}_M | b, \mathbf{a}_M} V_{t-1}^\pi(\mathbb{E}_{\mathfrak{z}_P | b, \mathbf{a}_P} b^{\mathfrak{z}_M, \mathfrak{z}_P, \mathbf{a}_M, \mathbf{a}_P}) \quad (36)$$

Since the expectation of the posterior is the prior,

$$\begin{aligned} F_b^\pi(\mathbf{a}_N) &\geq \mathbb{E}_{\mathfrak{z}_M | b, \mathbf{a}_M} V_{t-1}^\pi(b^{\mathfrak{z}_M, \mathbf{a}_M}) \\ F_b^\pi(\mathbf{a}_N) &\geq F_b^\pi(\mathbf{a}_M). \end{aligned} \quad (37)$$

Consequently, we have:

$$\begin{aligned} \rho(b) + \gamma^{(h-k)} F_b^\pi(\mathbf{a}_N) &\geq \rho(b) + \gamma^{(h-k)} F_b^\pi(\mathbf{a}_M) \\ Q_t^\pi(b, \mathbf{a}_N) &\geq Q_t^\pi(b, \mathbf{a}_M). \end{aligned} \quad (38)$$

□

Tying together our results so far:

Theorem 4. *If $\mathfrak{z}^{t:k}$ is conditionally independent given s^k , V_t^π is convex over the belief space for all t, π and $\rho(b) = -H_b(s)$, then for all b ,*

$$V_t^G(b) \geq (1 - e^{-1})^{2t} V_t^*(b). \quad (39)$$

Proof. Follows from Theorem 2, given $Q_t^G(b, \mathbf{a})$ is non-negative, monotone and submodular. For $\rho(b) = -H_b(s)$, it is easy to see that $Q_t^G(b, \mathbf{a})$ is non-negative, as entropy is always positive [Cover and Thomas, 1991]. Theorem 3 showed that $Q_t^G(b, \mathbf{a})$ is submodular if $\rho(b) = -H_b(s)$. The monotonicity of Q_t^G follows the fact that $-H_b(s)$ is convex [Cover and Thomas, 1991]: since Lemma ?? shows that \mathfrak{B}^G preserves convexity, V_t^G is convex if $\rho(b) = -H_b(s)$; Lemma 5 then shows that $Q_t^G(b, \mathbf{a})$ is monotone in \mathbf{a} . □

6 Analysis: Approximate Belief Entropy

While Theorem 4 bounds the error in $V_t^G(b)$, it does so only on the condition that $\rho(b) = -H_b(s)$. However, as discussed earlier, our definition of a dynamic sensor selection POMDP instead defines ρ using a set of vectors $\Gamma^\rho = \{\alpha_1^\rho, \dots, \alpha_m^\rho\}$, each of which is a tangent to $-H_b(s)$, as suggested by [Araya et al., 2010], in order to preserve the PWLC property. While this can interfere with the submodularity of $Q_t^\pi(b, \mathbf{a})$, in this section we show that the error generated by this approximation is still bounded in this case.

Let \tilde{V}_t^* denote the optimal value function when using a PWLC approximation to negative entropy for the belief-based reward, as in a dynamic sensor selection POMDP. Araya et al. (2010) showed that, if $\rho(b)$ verifies the α -Hölder condition [Gilbarg and Trudinger, 2001], a generalization of the Lipschitz condition, then the following relation holds between V_t^* and \tilde{V}_t^* :

$$\|V_t^* - \tilde{V}_t^*\|_\infty \leq \frac{C\delta_B^\alpha}{1-\gamma}, \quad (40)$$

where V_t^* is the optimal value function with $\rho(b) = -H_b(s)$, δ_B is a measure of how well belief entropy is approximated by the PWLC function, and C is a constant.

Let $\tilde{V}_t^G(b)$ be the value function computed by greedy PBVI for the dynamic sensor selection POMDP.

Lemma 6. *For all beliefs b , the error between $V_t^G(b)$ and $\tilde{V}_t^G(b)$ is bounded by $\frac{C\delta_B^\alpha}{1-\gamma}$. That is, $\|V_t^G - \tilde{V}_t^G\|_\infty \leq \frac{C\delta_B^\alpha}{1-\gamma}$.*

Proof. Follows exactly the strategy Araya et al. (2010) used to prove (40), which places no conditions on π and thus holds as long as \mathfrak{B}^G is a contraction mapping. Since for any policy the Bellman operator \mathfrak{B}^π defined as:

$$(\mathfrak{B}^\pi V_{t-1})(b) = [\rho(b, a_\pi) + \gamma \sum_{z \in \Omega} \Pr(z|a_\pi, b) V_{t-1}(b^{z, a_\pi})], \quad (41)$$

is a contraction mapping [Bertsekas, 2007], the bound holds for \tilde{V}_t^G . \square

Let $\eta = \frac{C\delta_B^\alpha}{1-\gamma}$ and let $\tilde{\rho}(b)$ denote the PWLC approximated belief-based reward and $\tilde{Q}_t^*(b, \mathbf{a}) = \tilde{\rho}(b) + \sum_{\mathbf{z}} \Pr(\mathbf{z}|b, \mathbf{a}) \tilde{V}_{t-1}^*(b^{\mathbf{z}, \mathbf{a}})$ denote the value of taking action \mathbf{a} in belief b under an optimal policy. Let $\tilde{Q}_t^G(b, \mathbf{a})$ be the action-value function computed by greedy PBVI for the dynamic sensor selection POMDP. As mentioned before, it is not guaranteed that $\tilde{Q}_t^G(b, \mathbf{a})$ is submodular. Instead, we show that it is ϵ -submodular:

Definition 2. The set function $f(\mathbf{a})$ is ϵ -submodular in \mathbf{a} , if for every $\mathbf{a}_M \subseteq \mathbf{a}_N \subseteq A^+$, $a_e \in A^+ \setminus \mathbf{a}_N$ and $\epsilon \geq 0$,

$$f(a_e \cup \mathbf{a}_M) - f(\mathbf{a}_M) \geq f(a_e \cup \mathbf{a}_N) - f(\mathbf{a}_N) - \epsilon.$$

Lemma 7. *If $\|V_{t-1}^\pi - \tilde{V}_{t-1}^\pi\|_\infty \leq \eta$, and $Q_t^\pi(b, \mathbf{a})$ is submodular in \mathbf{a} , then $\tilde{Q}_t^\pi(b, \mathbf{a})$ is ϵ' -submodular in \mathbf{a} for all b , where $\epsilon' = 4(\gamma + 1)\eta$.*

Proof. Since, $\|V_{t-1}^\pi - \tilde{V}_{t-1}^\pi\|_\infty \leq \eta$, then for all beliefs b ,

$$V_{t-1}^\pi(b) - \tilde{V}_{t-1}^\pi(b) \leq \eta, \quad (42)$$

For a given \mathbf{a} , on both sides multiply $\gamma \geq 0$, take the expectation over \mathbf{z} ,

$$\gamma \mathbb{E}_{\mathbf{z}|b, \mathbf{a}} V_{t-1}^\pi(b) - \gamma \mathbb{E}_{\mathbf{z}|b, \mathbf{a}} \tilde{V}_{t-1}^\pi(b) \leq \gamma \eta \quad (43)$$

Since $\rho(b) - \tilde{\rho}(b) \leq \eta$,

$$\rho(b) - \tilde{\rho}(b) + \gamma \mathbb{E}_{\mathbf{z}|b, \mathbf{a}} V_{t-1}^\pi(b) - \gamma \mathbb{E}_{\mathbf{z}|b, \mathbf{a}} \tilde{V}_{t-1}^\pi(b) \leq \gamma \eta + \eta \quad (44)$$

Therefore for all b, \mathbf{a} ,

$$Q_t^\pi(b, \mathbf{a}) - \tilde{Q}_t^\pi(b, \mathbf{a}) \leq (\gamma + 1)\eta \quad (45)$$

Now since $Q_t^\pi(b, \mathbf{a})$ is submodular, it satisfies the following equation,

$$Q_t^\pi(b, a_e \cup \mathbf{a}_M) - Q_t^\pi(b, \mathbf{a}_M) \geq Q_t^\pi(b, a_e \cup \mathbf{a}_N) - Q_t^\pi(b, \mathbf{a}_N), \quad (46)$$

for every $\mathbf{a}_M \subseteq \mathbf{a}_N \subseteq A^+$, $a_e \in A^+ \setminus \mathbf{a}_N$. For each action that appear in (46), that is, $\{a_e \cup \mathbf{a}_M\}, \mathbf{a}_M, \{a_e \cup \mathbf{a}_N\}$ and \mathbf{a}_N , the value computed by \tilde{Q}_t^π for belief b will be an approximation to Q_t^π . Thus the inequality in (46) that holds for Q_t^π , may not hold for \tilde{Q}_t^π . The worst case possible is, for some combination of $b, \{a_e \cup \mathbf{a}_M\}, \mathbf{a}_M, \{a_e \cup \mathbf{a}_N\}, \tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_M)$ and $Q_t^\pi(b, \mathbf{a}_N)$ underestimates the true value of $Q_t^\pi(b, a_e \cup \mathbf{a}_M)$ and $\tilde{Q}_t^\pi(b, \mathbf{a}_N)$ by $(\gamma + 1)\eta$ each and $\tilde{Q}_t^\pi(b, \mathbf{a}_M)$ and $\tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_N)$ overestimates the value of $Q_t^\pi(b, \mathbf{a}_M)$ and $Q_t^\pi(b, a_e \cup \mathbf{a}_N)$ by $(\gamma + 1)\eta$ each.

The worse case described above can be written formally as:

Since $\tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_M)$ and $\tilde{Q}_t^\pi(b, \mathbf{a}_N)$ underestimates the value by $(\gamma + 1)\eta$,

$$\begin{aligned} \tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_M) + (\gamma + 1)\eta &= Q_t^\pi(b, a_e \cup \mathbf{a}_M) \\ \tilde{Q}_t^\pi(b, \mathbf{a}_N) + (\gamma + 1)\eta &= Q_t^\pi(b, \mathbf{a}_N) \end{aligned} \quad (47)$$

Since $\tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_N)$ and $\tilde{Q}_t^\pi(b, \mathbf{a}_M)$ overestimates the value by $(\gamma + 1)\eta$,

$$\begin{aligned} \tilde{Q}_t^\pi(b, \mathbf{a}_M) - (\gamma + 1)\eta &= Q_t^\pi(b, \mathbf{a}_M) \\ \tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_N) - (\gamma + 1)\eta &= Q_t^\pi(b, a_e \cup \mathbf{a}_N) \end{aligned} \quad (48)$$

Substituting these equations in (46),

$$\tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_M) - \tilde{Q}_t^\pi(b, \mathbf{a}_M) \geq \tilde{Q}_t^\pi(b, a_e \cup \mathbf{a}_N) - \tilde{Q}_t^\pi(b, \mathbf{a}_N) - 4(\gamma + 1)\eta. \quad (49)$$

□

Lemma 8. *If $\tilde{Q}_t^\pi(b, \mathbf{a})$ is non-negative, monotone and ϵ -submodular in \mathbf{a} , then*

$$\tilde{Q}_t^\pi(b, \mathbf{a}^G) \geq (1 - e^{-1})\tilde{Q}_t^\pi(b, \mathbf{a}^*) - 4\chi_K\epsilon, \quad (50)$$

where $\chi_K = \sum_{p=0}^{K-1} (1 - K^{-1})^p$.

The proof for above theorem follows the same strategy as shown by [Krause and Guestrin, 2007]

Proof. Let \mathbf{a}^* be the optimal set of action features of size K , $\mathbf{a}^* = \arg \max_{\mathbf{a}} \tilde{Q}_t^\pi(b, \mathbf{a})$ and let \mathbf{a}^l be the greedily selected set of size l , that is, $\mathbf{a}^l = \text{greedy-argmax}(Q_t^\pi(b, \cdot), A^+, l)$. Also, let $\mathbf{a}^* = \{a_1^* \dots a_K^*\}$ be the elements of set \mathbf{a}^* .

Then,

$$\begin{aligned}
& \text{By monotonicity of } \tilde{Q}_t^\pi(b, \mathbf{a}) \\
& \tilde{Q}_t^\pi(b, \mathbf{a}^*) \leq \tilde{Q}_t^\pi(b, \mathbf{a}^* \cup \mathbf{a}^l) \\
& \text{Re-writing as a telescoping sum} \\
& = \tilde{Q}_t^\pi(b, \mathbf{a}^l) + \sum_{j=1}^K \Delta_{\tilde{Q}_b}(a_j^* | \mathbf{a}^l \cup \{a_1^* \dots a_{j-1}^*\}) \\
& \text{Using (49)} \\
& \leq \tilde{Q}_t^\pi(b, \mathbf{a}^l) + \sum_{j=1}^K \Delta_{\tilde{Q}_b}(a_j^* | \mathbf{a}^l) + 4K\epsilon \tag{51} \\
& \text{As } \mathbf{a}^{l+1} \text{ is built greedily from } \mathbf{a}^l \text{ in order to maximize } \Delta_{\tilde{Q}_b} \\
& \leq \tilde{Q}_t^\pi(b, \mathbf{a}^l) + \sum_{j=1}^K (\tilde{Q}_t^\pi(b, \mathbf{a}^{l+1}) - \tilde{Q}_t^\pi(b, \mathbf{a}^l)) + 4K\epsilon \\
& \text{As } |\mathbf{a}^*| = K \\
& = \tilde{Q}_t^\pi(b, \mathbf{a}^l) + K(\tilde{Q}_t^\pi(b, \mathbf{a}^{l+1}) - \tilde{Q}_t^\pi(b, \mathbf{a}^l)) + 4K\epsilon
\end{aligned}$$

Let $\delta_l := \tilde{Q}_t^\pi(b, \mathbf{a}^*) - \tilde{Q}_t^\pi(b, \mathbf{a}^l)$, which allows us to rewrite above equation as: $\delta_l \leq K(\delta_l - \delta_{l+1}) + 4K\epsilon$.

Hence, $\delta_{l+1} \leq (1 - \frac{1}{K})\delta_l + 4\epsilon$.

Using this relation recursively, we can write, $\delta_K \leq (1 - \frac{1}{K})^K \delta_0 + 4 \sum_{p=0}^{K-1} (1 - \frac{1}{K})^p \epsilon$.

Also, $\delta_0 = \tilde{Q}_t^\pi(b, \mathbf{a}^*) - \tilde{Q}_t^\pi(b, \mathbf{a}^0)$ and using the inequality $1 - x \leq e^{-x}$, we can write $\delta_K \leq e^{-\frac{K}{K}} \tilde{Q}_t^\pi(b, \mathbf{a}^*) + 4 \sum_{p=0}^{K-1} (1 - K^{-1})^p \epsilon$. Substituting δ_K and rearranging terms (Also $\chi_K = \sum_{p=0}^{K-1} (1 - \frac{1}{K})^p$).

$$\begin{aligned}
\tilde{Q}_t^\pi(b, \mathbf{a}^K) & \geq (1 - e^{-1}) \tilde{Q}_t^\pi(b, \mathbf{a}^*) - 4\chi_K \epsilon \\
\tilde{Q}_t^\pi(b, \mathbf{a}^G) & \geq (1 - e^{-1}) \tilde{Q}_t^\pi(b, \mathbf{a}^*) - 4\chi_K \epsilon
\end{aligned} \tag{52}$$

□

Theorem 5. For all beliefs, the error between $\tilde{V}_t^G(b)$ and $\tilde{V}_t^*(b)$ is bounded, if $\rho(b) = -H_b(s)$, V_t^π is convex over the belief space for all t, π , and $\mathbf{z}^{t:k}$ is conditionally independent given s^k .

Proof. Theorem 4 shows that, if $\rho(b) = -H_b(s)$, and $\mathbf{z}^{t:k}$ is conditionally independent given s^k , then $Q_t^G(b, \mathbf{a})$ is submodular. Using Lemma 7, for $V_t^\pi = V_t^G$, $\tilde{V}_t^\pi = \tilde{V}_t^G$, $Q_t^\pi(b, \mathbf{a}) = Q_t^G(b, \mathbf{a})$ and $\tilde{Q}_t^\pi(b, \mathbf{a}) = \tilde{Q}_t^G(b, \mathbf{a})$, it is easy to see that $\tilde{Q}_t^G(b, \mathbf{a})$ is ϵ -submodular. This satisfies one condition of Lemma 8. The convexity of $\tilde{V}_t^G(b)$ follows from Lemma ?? and that $\tilde{\rho}(b)$ is convex. Given that $\tilde{V}_t^G(b)$ is convex, the monotonicity of $\tilde{Q}_t^G(b, \mathbf{a})$ follows from Lemma 5. Since $\tilde{\rho}(b)$ is non-negative, $\tilde{Q}_t^G(b, \mathbf{a})$ is non-negative too. Now we can apply Lemma 9 to prove that the error generated by a one-time application of the greedy Bellman operator to $\tilde{V}_t^G(b)$, instead of the Bellman optimality operator, is bounded. It is thus easy to see that the error between $\tilde{V}_t^G(b)$, produced by multiple applications of the greedy Bellman operator, and $\tilde{V}_t^*(b)$ is bounded for all beliefs. □

7 Experiments

To empirically evaluate greedy PBVI, we tested it on the problem of tracking either one or multiple people using a multi-camera system. The problem was extracted from a real-world

dataset collected in a shopping mall [Bouma et al., 2013]. The dataset was gathered over 4 hours using 13 CCTV cameras. Each camera uses a *FPDW* pedestrian detector [Dollár et al., 2010] to detect people in each camera image and *in-camera tracking* [Bouma et al., 2013] to generate tracks of the detected people’s movement over time. The dataset thus consists of 9915 tracks, each specifying one person’s x - y position throughout time. Figure 2 shows sample tracks from all of the cameras.

To model this setting as a dynamic sensor selection POMDP, we discretize the continuous x - y plane into $C = 20$ cells. The state s thus describes the person’s location and $|S| = 21$ (the 20 cells plus an ‘external’ state indicating the person has left the shopping mall). Using the track data, we learned a maximum-likelihood tabular transition function. Since we do not have ground truth data about people’s location, it is not possible to learn an observation function from the tracks. Instead, we constructed an observation function by generating the false positive and false negative probabilities randomly. For each camera and each cell in that camera’s region, the probability of a false positive and false negative were set by sampling a number uniformly randomly from the interval $[0.15, 0.25]$. The reward function is described as a set of $|S|$ vectors, $\Gamma^p = \{\alpha_1 \dots \alpha_{|S|}\}$, with $\alpha_i(s) = 1$ if $s = i$ and $\alpha_i(s) = 0$ otherwise. The initial belief is uniform across all states. We planned for horizon $h = 10$ with $\gamma = 0.99$.

To address the blowup in the size of the state space for multi-person tracking, we use a variant of *transfer planning* [Oliehoek et al., 2013]. We divide the multi-person problem into several *source* problems, one for each person, and solve them independently to compute $V_t(b) = \sum V^i(b_i)$, where $V^i(b_i)$ is the value of the current belief b_i about the i -th person’s location. Thus $V^i(b_i)$ only needs to be computed once, by solving POMDP of the same size as that in the single-person setting. During action selection, $V_t(b)$ is computed using the current b_i for each person. Computing $V_t(b)$ as the linear combination of $V^i(b_i)$ corresponds to the assumption that each person’s movement is independent of that of other people. Although violated in practice, such assumptions nonetheless often yield good approximations. In our setting, since this assumption is applied to both regular and greedy PBVI, it allows a fair comparison of the two methods in POMDPs with larger state spaces than would otherwise be not feasible.

As baselines, we tested against regular PBVI and *myopic* versions of both greedy and regular PBVI that compute a policy assuming $h = 1$ and use it at each timestep.

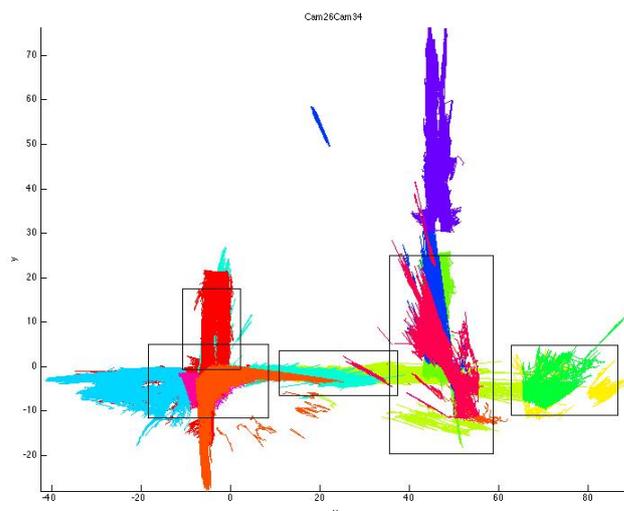


Figure 2: Sample tracks for all the cameras. Each color represents all the tracks observed by a given camera. The boxes denote regions of high overlap between cameras.

Figure 3 shows runtimes under different values of N and K . Since multi-person tracking uses the value function obtained by solving a single-person POMDP, single and multi-person tracking have the same runtimes. These results demonstrate that greedy PBVI requires only a fraction of the computational cost of regular PBVI. In addition, the difference in runtime grows quickly as the action space gets larger: for $N = 5$ and $K = 2$ greedy PBVI is twice as fast, while for $N = 11, K = 3$ it is approximately nine times as fast. Thus, greedy PBVI enables much better scalability in the action space.

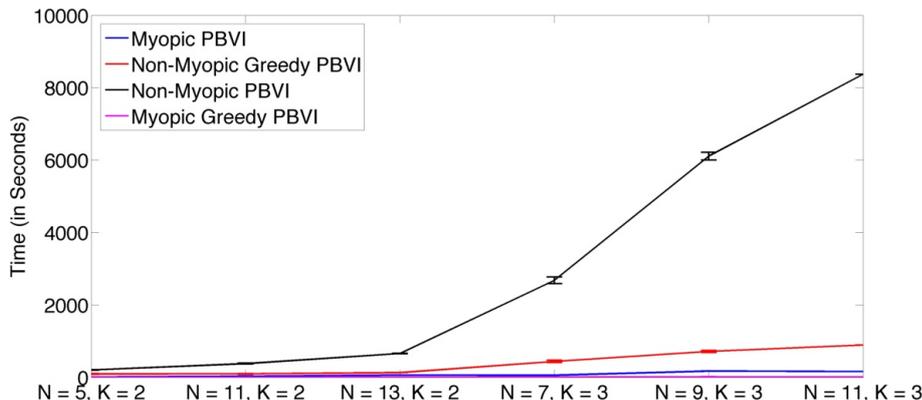


Figure 3: Runtimes for the different methods.

Figure 4, which shows the cumulative reward under different values of N and K for single-person (left) and multi-person (right) tracking, verifies that greedy PBVI’s speedup does not come at the expense of performance, as greedy PBVI accumulates nearly as much reward as regular PBVI. They also show that both PBVI and greedy PBVI benefit from non-myopic planning. While the performance advantage of non-myopic planning is relatively modest, it increases with the number of cameras and people, which suggests that non-myopic planning is important to making active perception scalable.

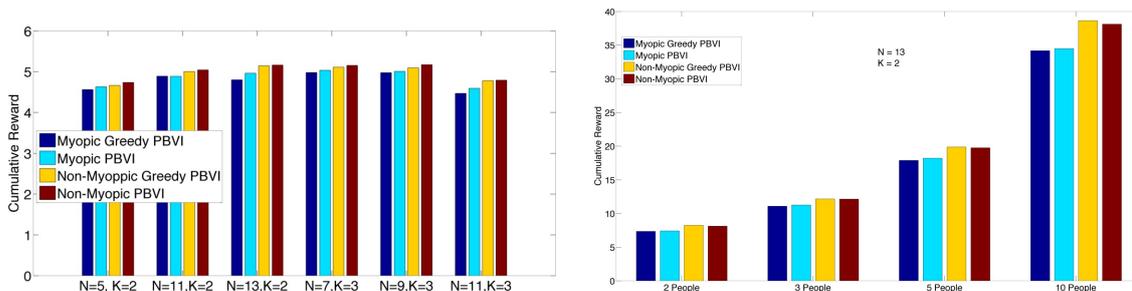


Figure 4: Cumulative reward for single-person (left) and multi-person (right) tracking.

Furthermore, an analysis of the resulting policies showed that myopic and non-myopic policies differ qualitatively. A myopic policy, in order to minimise uncertainty in the next step, tends to look where it believes the person to be. By contrast, a non-myopic policy tends to proactively look where the person might go next, so as to more quickly detect her new location when she moves. Consequently, non-myopic policies exhibit less fluctuation in belief and accumulate more reward, as illustrated in Figure 5. The blue lines marks when the agent happens to choose the camera which can observe the cell occupied by the person. The red line plots the max of the agent’s belief. The difference in fluctuation in belief is evident from the figure as the max of the belief often drops below 0.5 for the myopic policy but rarely does so for the non-myopic policy.

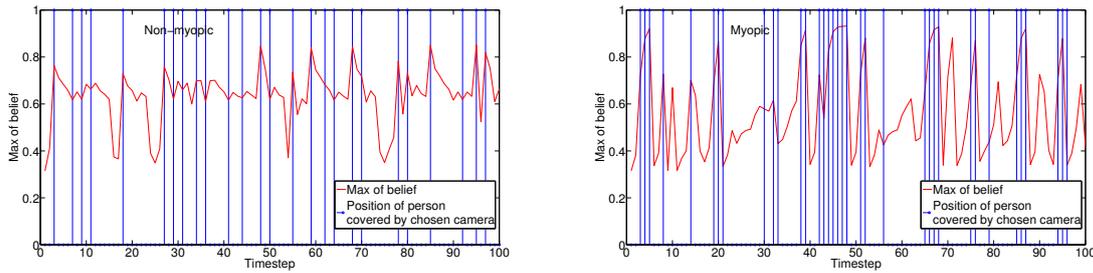


Figure 5: Behaviour of myopic vs. non-myopic policy.

8 Related Work

Dynamic sensor selection has been studied in many contexts. Most work focuses on either open-loop or myopic solutions, e.g., [Kreucher et al., 2005, Williams et al., 2007, Joshi and Boyd, 2009]. By contrast, our POMDP-approach enables a closed-loop, non-myopic approach that can lead to better performance when the underlying state of the world changes over time.

Spaan (2008) and Spaan and Lima (2009) also consider a POMDP approach to dynamic sensor selection. However, they apply their method only to small POMDPs without addressing scalability with respect to the action space. Such scalability, which greedy PBVI makes possible, is central to the practical utility of POMDPs for sensor selection. Other work using POMDPs for sensor selection [Krishnamurthy and Djonin, 2007, Ji et al., 2007] also does not consider scalability in the action space. Krishnamurthy and Djonin (2007) consider a non-standard POMDP in which, unlike in our setting, the reward is not linear in the belief.

In recent years, applying greedy maximization to submodular functions has become a popular and effective approach to sensor selection [Krause and Guestrin, 2005, 2007]. However, such work focuses on myopic or fully observable settings [Kumar and Zilberstein, 2009] and thus does not enable the long-term planning required to cope with dynamic state in a POMDP.

Adaptive submodularity [Golovin and Krause, 2011] is a recently developed extension that addresses these limitations by allowing action selection to condition on previous observations. However, it assumes a static state and thus cannot model the dynamics of a POMDP across timesteps. Therefore, in a POMDP, adaptive submodularity is only applicable *within* a timestep, during which state does not change but the agent can sequentially add sensors to a set. In principle, adaptive submodularity could enable this intra-timestep sequential process to be adaptive, i.e., the choice of later sensors could condition on the observations generated by earlier sensors. However, this is not possible in our setting because we assume that, due to computational costs, all sensors must be selected simultaneously. Consequently, our analysis considers only classic, non-adaptive submodularity.

To our knowledge, our work is the first to establish the submodularity of POMDP value functions for dynamic sensor selection POMDPs and thus leverage greedy maximization to scalably compute bounded approximate policies for dynamic sensor selection modeled as a full POMDP.

9 Conclusions & Future Work

This paper proposed greedy PBVI, a new POMDP planning method for dynamic sensor selection that exploits greedy maximization to improve scalability in the action space. We showed that the value function computed in this way has bounded error if certain conditions including submodularity are met. We also showed that such conditions are met, or approximately

met, if reward is defined using negative belief entropy or an approximation thereof. Experiments on a real-world dataset from a multi-camera tracking system show that it achieves similar performance to existing methods but incurs only a fraction of the computational cost.

One avenue for future work includes quantifying the error bound between $\tilde{V}_t^G(b)$ and $\tilde{V}_t^*(b)$, as our current results (Theorem 5) show only that it is bounded. We also intend to consider cases where its possible to sequentially process information from sensors and thus integrate our approach with adaptive submodularity.

References

- M. Araya, O. Buffet, V. Thomas, and F. Charpillet. A POMDP extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems*, pages 64–72, 2010.
- K. J. Aström. Optimal control of Markov decision processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.
- D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 3rd edition, 2007.
- H. Bouma, J. Baan, S. Landsmeer, C. Kruszynski, G. van Antwerpen, and J. Dijk. Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall. In *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2013.
- A. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proceedings of the Thirteenth conference on Uncertainty in Artificial Intelligence*, pages 54–61, 1997.
- T. M. Cover and J. A. Thomas. Entropy, relative entropy and mutual information. *Elements of Information Theory*, pages 12–49, 1991.
- P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.
- D. Gilbarg and N. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. U.S. Government Printing Office, 2001.
- D. Golovin and A. Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.
- S. Ji, R. Parr, and L. Carin. Nonmyopic multiaspect sensing with partially observable Markov decision processes. *IEEE Transactions on Signal Processing*, 55(6):2720–2730, 2007.
- S. Joshi and S. Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.
- L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.
- A. Krause and D. Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2014.
- A. Krause and C. Guestrin. Optimal nonmyopic value of information in graphical models - efficient algorithms and theoretical limits. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, 2005.

- A. Krause and C. Guestrin. Near-optimal observation selection using submodular functions. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, 2007.
- C. Kreucher, K. Kastella, and A. O. Hero, III. Sensor management using an active sensing approach. *Signal Processing*, 85(3):607–624, 2005.
- V. Krishnamurthy and D. V. Djonin. Structured threshold policies for dynamic sensor scheduling—a partially observed Markov decision process approach. *IEEE Transactions on Signal Processing*, 55(10):4938–4957, 2007.
- A. Kumar and S. Zilberstein. Event-detecting multi-agent MDPs: Complexity and constant-factor approximation. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, pages 201–207, 2009.
- W. S. Lovejoy. Computationally feasible bounds for partially observed Markov decision processes. *Operations Research*, 39(1):162–175, 1991.
- G. E. Monahan. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.
- G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.
- F. A. Oliehoek, S. Whiteson, and M. T. J. Spaan. Approximate solutions for factored Dec-POMDPs with many agents. In *Proceedings of the Twelfth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 563–570, 2013.
- J. Pineau, G. J. Gordon, and S. Thrun. Anytime point-based approximations for large POMDPs. *Journal of Artificial Intelligence Research*, 27:335–380, 2006.
- Y. Satsangi, S. Whiteson, and F. Oliehoek. Exploiting submodular value functions for faster dynamic sensor selection. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. To appear.
- R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.
- M. T. J. Spaan. Cooperative active perception using POMDPs. In *AAAI 2008 Workshop on Advancements in POMDP Solvers*, 2008.
- M. T. J. Spaan and P. U. Lima. A decision-theoretic approach to dynamic sensor selection in camera networks. In *Proceedings of the Nineteenth International Conference on Automated Planning and Scheduling*, pages 279–304, 2009.
- J. Williams, J. Fisher, and A. Willsky. Approximate dynamic programming for communication-constrained sensor network management. *IEEE Transactions on Signal Processing*, 55(8):4300–4311, 2007.

Acknowledgements

We thank Henri Bouma and TNO for providing us with the dataset used in our experiments. We also thank the STW User Committee for its advice regarding active perception for multi-camera tracking systems. This research is supported by the Dutch Technology Foundation STW (project #12622), which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs. Frans Oliehoek is funded by NWO Innovational Research Incentives Scheme Veni #639.021.336.

IAS reports

This report is in the series of IAS technical reports. The series editor is Bas Terwijn (B.Terwijn@uva.nl). Within this series the following titles appeared:

F.A. Oliehoek and C. Amato *Dec-POMDPs as Non-Observable MDPs* Technical Report IAS-UVA-14-01, Informatics Institute, University of Amsterdam, The Netherlands, November 2014.

A. Visser *UvA Rescue Technical Report: a description of the methods and algorithms implemented in the UvA Rescue code release* Technical Report IAS-UVA-12-02, Informatics Institute, University of Amsterdam, The Netherlands, September 2012.

A. Visser *A survey of the architecture of the communication library LCM for the monitoring and control of autonomous mobile robots* Technical Report IAS-UVA-12-01, Informatics Institute, University of Amsterdam, The Netherlands, September 2012.

All IAS technical reports are available for download at the ISLA website:
<http://isla.science.uva.nl/node/85>