

Active Perception With Initial-State Uncertainty: A Policy Gradient Method

Chongyang Shi¹, Graduate Student Member, IEEE, Shuo Han², Member, IEEE,
Michael Dorothy³, and Jie Fu⁴, Member, IEEE

Abstract—This letter studies the synthesis of an active perception policy that maximizes the information leakage of the initial state in a stochastic system modeled as a hidden Markov model (HMM). Specifically, the emission function of the HMM is controllable with a set of perception or sensor query actions. Given the goal is to infer the initial state from partial observations in the HMM, we use Shannon conditional entropy as the planning objective and develop a novel policy gradient method with convergence guarantees. By leveraging a variant of observable operators in HMMs, we prove several important properties of the gradient of the conditional entropy with respect to the policy parameters, which allow efficient computation of the policy gradient and stable and fast convergence. We demonstrate the effectiveness of our solution by applying it to an inference problem in a stochastic grid world environment.

Index Terms—Information theory and control, Markov processes, optimization.

I. INTRODUCTION

THIS letter studies the synthesis of an active perception strategy that maximizes the transparency of an initial state in a stochastic system given partial observations. We introduce a set of active perception actions into the system modeled as a hidden Markov model (HMM) such that the emission at a given state is jointly determined by the state and a perception action. The goal is to compute an active perception strategy that maximizes the information leakage about the initial state S_0 given observations Y , which is measured by the negative conditional entropy, $-H(S_0|Y)$.

The contributions of this letter are summarized as follows: First, we prove that active perception planning to minimize the

conditional entropy $H(S_0|Y)$ cannot be reduced to a partially observable Markov decision process (POMDP) with a belief-based reward function. Leveraging a variant of observable operators [10], we develop an efficient algorithm to compute the posterior distribution $\mathbb{P}(S_0|Y)$ given a perception policy. Additionally, it is shown that the policy gradient of conditional entropy depends only on the posterior distribution $\mathbb{P}(S_0|Y)$ and the gradient of the policy with respect to its parameters. We prove that the entropy is Lipschitz continuous and Lipschitz smooth in the policy parameters under some assumptions for policy search space and thus ensure the convergence of the gradient-based planning. Finally, we evaluate the performance in a stochastic grid world environment.

Related Work: Active perception [4] involves selectively gathering information to enhance the task performance of an autonomous system. The applications of active perception range from object localization [2], target tracking [6], [19] to mission planning for surveillance and monitoring [7], [11]. Information-theoretic metrics are introduced as planning objectives in various active perception problems. One approach [3], [8] to active perception is to formulate a partially observable Markov decision process (POMDP) with a reward function that depends on the information state or the belief of an agent. For example, in target surveillance, a patrolling team is rewarded by uncertainty reduction of the belief about the intruder's position/state. In intent inference, the authors [15] used the negative entropy of the belief over the opponent's intent as a reward and maximized the total reward for active perception. Another approach is to apply the policy gradient method in active estimation [17], [18].

Besides active perception, several studies have explored decision-making with information-theoretic objectives. In [12], the authors introduce a method for obfuscating/estimating state trajectories in POMDPs and show that the causal conditional entropy of the state trajectory given observations and controls can be reformulated as a cumulative sum, allowing the use of standard POMDP solvers. The work [16] develops a method to maximize the conditional entropy of a secret variable in an MDP to a passive observer. Another work [14] proposes entropy maximization in POMDPs to minimize the predictability of an agent's trajectories to an outside observer. In both cases, the planning agent controls the stochastic dynamics and the observer is passive. In comparison, this letter studies the

Received 16 September 2024; revised 20 November 2024; accepted 25 November 2024. Date of publication 9 December 2024; date of current version 6 January 2025. This work was supported in part by NSF under Award 2024802; in part by the Army Research Laboratory under Grant W911NF-22-2-0233; and in part by the Army Research Office under Grant W911NF-22-1-0166. Recommended by Senior Editor C. Briat. (Corresponding author: Jie Fu.)

Chongyang Shi and Jie Fu are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32608 USA (e-mail: c.shi@ufl.edu; fujie@ufl.edu).

Shuo Han is with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: hanshuo@uic.edu).

Michael Dorothy is with the Aerospace Engineering, DEVCOM Army Research Laboratory, Adelphi, MD 21001 USA (e-mail: michael.r.dorothy.civ@army.mil).

Digital Object Identifier 10.1109/LCSYS.2024.3513896

dual problem when the control system is autonomous but the observer is active. Thus, the observer's policy is restricted to be observation-based. Unlike [14] where the goal is to maximize the sum of the conditional entropy of current states given the historical states, our goal is to maximize the information about some past state (initial state) given the observations received by the observer. See Remark 1 for more discussion.

II. PRELIMINARIES AND PROBLEM FORMULATION

Notation: The set of real numbers is denoted by \mathbb{R} . Random variables will be denoted by capital letters, and their realizations by lowercase letters (e.g., X and x). A sequence of random variables and their realizations with length T are denoted as $X_{0:T}$ and $x_{0:T}$. The notation x_i refers to the i -th component of a vector $x \in \mathbb{R}^n$ or to the i -th element of a sequence x_0, x_1, \dots , which will be clarified by the context. Given a finite set \mathcal{S} , let $\mathcal{D}(\mathcal{S})$ be the set of all probability distributions over \mathcal{S} . The set \mathcal{S}^T denotes the set of sequences with length T composed of elements from \mathcal{S} , and \mathcal{S}^* denotes the set of all finite sequences generated from \mathcal{S} . The empty string in \mathcal{S}^* is denoted by λ .

We introduce a class of active perception problems where the agent cannot control the dynamical system but can select perception actions to monitor it in order to infer some unknown state variables. The class of perception actions can be the choices of sensors to query in a distributed sensor network or the choices of poses for cameras with a limited field of view (FoV).

Definition 1: An HMM with a controllable emission function is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{O}, \mathcal{A}, P, E, S_0 \rangle$ where 1) $\mathcal{S} = \{1, \dots, N\}$ is a finite state space. 2) \mathcal{O} is a finite set of observations. 3) \mathcal{A} is a finite set of *perception actions*. 4) $P : \mathcal{S} \rightarrow \mathcal{D}(\mathcal{S})$ is the probabilistic transition function. 5) $E : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{D}(\mathcal{O})$ is the emission function (observation function) that takes a state s and a *perception action* a , outputs a distribution over observations. 6) S_0 is a random variable representing the initial state. The distribution of S_0 is denoted by μ_0 . And \mathcal{S}_0 denotes the set of initial states.

A *non-stationary, observation-based* perception policy is a function $\pi : \mathcal{O}^* \rightarrow \mathcal{D}(\mathcal{A})$ that maps a history $o_{0:t}$ of observations to a distribution $\pi(\cdot | o_{0:t})$ over perception actions.

For a given HMM \mathcal{M} , a perception policy π induces a discrete stochastic process $M_\pi := \{S_t, A_t, O_t, t \in \mathbb{N}\}$, where $S_t \in \mathcal{S}$ and $O_t \in \mathcal{O}$ are the underlying hidden state and observation at the t -th time step, and

$$\begin{aligned} \mathbb{P}^\pi(O_t = o | O_{0:t-1} = o_{0:t-1}, S_t = s_t) \\ = \sum_{a \in \mathcal{A}} E(o | s_t, a) \pi(a | o_{0:t-1}), \forall t > 0. \end{aligned}$$

when the perception policy π is understood from the context, we write \mathbb{P} instead of \mathbb{P}^π for clarity.

The conditional entropy of X_2 given X_1 is defined by

$$H(X_2 | X_1) = - \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} p(x_1, x_2) \log p(x_2 | x_1).$$

The conditional entropy measures the uncertainty about X_2 given knowledge of X_1 . A lower conditional entropy makes it easier to learn X_2 from observing a sample of X_1 .

For any finite horizon T , the agent's partial observation about a path in the HMM includes a sequence $o_{0:T}$ of observations and a sequence $a_{0:T}$ of perception actions. We denote the agent's information by $y_{0:T} = (o_{0:T}, a_{0:T})$. When the length of $y_{0:T}$ is clear from the context, we omit the subscript and use y to denote the sequence. In the following, we refer to y as an observation sequence with the understanding that it is the joint observation and perception action sequence.

Problem 1: Let an HMM \mathcal{M} and a finite horizon T be given. Let Π be a policy space. Compute an active perception policy $\pi \in \Pi$ that minimizes the conditional entropy of the initial state S_0 given the partial information $Y_{0:T} = (O_{0:T}, A_{0:T})$ induced by π . That is,

$$\underset{\pi \in \Pi}{\text{minimize}} \quad H(S_0 | Y_{0:T}; M_\pi),$$

where the conditional entropy term $H(S_0 | Y_{0:T}; M_\pi) := - \sum_{s_0 \in \mathcal{S}_0, y_{0:T} \in \mathcal{O}^T \times \mathcal{A}^T} \mathbb{P}^\pi(s_0, y_{0:T}) \log \mathbb{P}^\pi(s_0 | y_{0:T})$ and $\mathbb{P}^\pi(s_0, y_{0:T})$ is the joint probability of starting from s_0 and observing $y_{0:T}$ under the policy π and $\mathbb{P}^\pi(s_0 | y_{0:T})$ is the conditional probability of starting from s_0 given observation $y_{0:T}$.

Remark 1: This letter considers an information-theoretic, quantitative measure of initial-state opacity/transparency due to uncertainty in transition dynamics and emission function. Qualitative initial-state opacity has been studied in supervisory control [5], [13], with applications to intent inference, tracking, and system diagnosis. A system is initial-state opaque if the observer cannot infer if a path started from a set of secret initial states. Take the intent recognition for example, consider a surveillance sensor network that can be actively queried to identify if an agent is benign or adversarial. This problem is reduced to an initial-state inference problem by constructing an HMM composed of two HMMs, one for the benign agent and another for the adversary. The initial state of the composite HMM determines which HMM (benign or adversarial) is selected to generate the observation. Thus, intent inference is equivalent to inferring the initial state of the composite HMM.

Remark 2: A constrained formulation to minimize entropy given a bounded perception cost can also be formulated. Because policy gradient methods [1] for optimizing a cumulative reward/cost are well-understood, we focus on solving this entropy minimization problem and expect that only a small modification to the gradient computation is needed to minimize a weighted sum of the conditional entropy and the expected total cost of perception actions.

III. MAIN RESULT

Given observations $y_{0:t} = (o_{0:t}, a_{0:t})$ and a perception policy π , the agent can compute a posterior distribution about the initial state S_0 , referred to as its *belief* of the initial state, such that $b_t(s) = \mathbb{P}^\pi(S_0 = s | o_{0:t}, a_{0:t})$.

First, we show that the problem cannot be reduced to a ρ -POMDP [3] which is a POMDP with a belief-based reward.

Proposition 1: There is no belief-based reward $R : \mathcal{D}(\mathcal{S}) \rightarrow \mathbb{R}$ such that $\sum_{i=0}^t R(b_i) = -H(S_0 | y_{0:t})$.

Proof: Suppose, by way of contradiction, such a belief-based reward function exists. Then, $R(b_t) = \sum_{i=0}^t R(b_i) - \sum_{i=0}^{t-1} R(b_i) = -H(S_0|y_{0:t}) + H(S_0|y_{0:t-1})$. We show that it is possible to reach the same belief with two different observations $y_{0:t}$ and $y'_{0:t}$, but $-H(S_0|y_{0:t}) + H(S_0|y_{0:t-1}) \neq -H(S_0|y'_{0:t}) + H(S_0|y'_{0:t-1})$. Consider an example of HMM where the initial state can be either 0 or 1 with a prior distribution $[0.5, 0.5]$; namely, the initial belief is given by $b_0 = [0.5, 0.5]$. If $S_0 = 0$, the next state S_1 will always emit an observation 0. If $S_0 = 1$, the next state S_1 will always emit an observation 1. Thus, if 0 is observed, then $b_1 = [1, 0]$; if 1 is observed, then $b_1 = [0, 1]$. The reward $R(b_1) = -H(S_0|o) + H(S_0|\lambda) = 1$ regardless whether $o = 0$ or $o = 1$ is observed. After reaching state S_1 , the system reaches S_2 next and yields some observation o' . The belief does not change, i.e., $b_2 = b_1$ because the initial state is known with certainty. However, based on the formula $R(b_2) = -H(S_0|o'o) + H(S_0|o) = -0 + 0 = 0$, which is different from $R(b_1)$. A contradiction is established. ■

In the next, we develop a policy gradient method to solve Problem 1. Consider a class of parametrized (stochastic) policies $\{\pi_\theta \mid \theta \in \Theta\}$. We denote by M_θ the stochastic process $\{S_t, A_t, O_t, t \in \mathbb{N}\}$ induced by a policy π_θ , and $\mathbb{P}_\theta(\cdot)$ the corresponding probability measure.

For a given parameterized policy π_θ , the gradient of $H(S_0|Y; \theta) := H(S_0|Y; \pi_\theta)$ is given by

$$\begin{aligned} \nabla_\theta H(S_0|Y; \theta) &= - \sum_{y \in \mathcal{O}^T \times \mathcal{A}^T} \sum_{s_0 \in \mathcal{S}_0} \left[\nabla_\theta \mathbb{P}_\theta(s_0, y) \log \mathbb{P}_\theta(s_0|y) \right. \\ &\quad \left. + \mathbb{P}_\theta(s_0, y) \nabla_\theta \log \mathbb{P}_\theta(s_0|y) \right] \\ &= - \sum_{y \in \mathcal{O}^T \times \mathcal{A}^T} \sum_{s_0 \in \mathcal{S}_0} \left[\mathbb{P}_\theta(y) \nabla_\theta \mathbb{P}_\theta(s_0|y) \log \mathbb{P}_\theta(s_0|y) \right. \\ &\quad \left. + \mathbb{P}_\theta(s_0|y) \nabla_\theta \mathbb{P}_\theta(y) \log \mathbb{P}_\theta(s_0|y) + \mathbb{P}_\theta(y) \nabla_\theta \mathbb{P}_\theta(s_0|y) \right] \\ &= - \sum_{y \in \mathcal{Y}_\theta} \mathbb{P}_\theta(y) \sum_{s_0 \in \mathcal{S}_0} \left[\log \mathbb{P}_\theta(s_0|y) \nabla_\theta \mathbb{P}_\theta(s_0|y) \right. \\ &\quad \left. + \mathbb{P}_\theta(s_0|y) \log \mathbb{P}_\theta(s_0|y) \nabla_\theta \log \mathbb{P}_\theta(y) + \nabla_\theta \mathbb{P}_\theta(s_0|y) \right]. \end{aligned} \quad (1)$$

where $\mathcal{Y}_\theta = \{y \in \mathcal{O}^T \times \mathcal{A}^T \mid P_\theta(y) > 0\}$ the set of possible observations under the perception policy π_θ . In the following, Propositions 5 and 4 will allow us to further simplify the computation of gradient.

To derive the results, we introduce the *observable operator* augmented with perception actions. The observable operator [9] has been proposed to represent a discrete HMM and used to calculate the probability of an observation sequence in an HMM using matrix multiplications.

Let the random variable of state, observation, and action, at time point t be denoted as X_t, O_t, A_t , respectively. Let $\mathbf{T} \in \mathbb{R}^{N \times N}$ be the transposed state transition matrix with

$$\mathbf{T}_{i,j} = \mathbb{P}(X_{t+1} = i \mid X_t = j).$$

For each $a \in \mathcal{A}$, Let $\mathbf{O}^a \in \mathbb{R}^{M \times N}$ be the observation probability matrix with $\mathbf{O}_{o,j}^a = E(o|j, a)$.

Definition 2: Given the HMM with controllable emissions \mathcal{M} , for any pair of observation and perception action (o, a) , the observable operator given perception actions $\mathbf{A}_{o|a}$ is a matrix of size $N \times N$ with its ij -th entry defined as

$$\mathbf{A}_{o|a}[i, j] = \mathbf{T}_{i,j} \mathbf{O}_{o,j}^a$$

which is the probability of transitioning from state j to state i and at the state j , an observation o is emitted given perception action a . In matrix form,

$$\mathbf{A}_{o|a} = \mathbf{T} \text{diag}(\mathbf{O}_{o,1}^a, \dots, \mathbf{O}_{o,N}^a).$$

Proposition 2: The probability of an observation sequence $o_{0:t}$ given a sequence of perception actions $a_{0:t}$, can be written as matrix operations,

$$\mathbb{P}(o_{0:t}|a_{0:t}) = \mathbf{1}_N^\top \mathbf{A}_{o_t|a_t} \dots \mathbf{A}_{o_0|a_0} \mu_0. \quad (2)$$

In addition, for a fixed initial state $s_0 \in \mathcal{S}_0$,

$$\mathbb{P}(o_{0:t}|a_{0:t}, s_0) = \mathbf{1}_N^\top \mathbf{A}_{o_t|a_t} \dots \mathbf{A}_{o_0|a_0} \mathbf{1}_{s_0}. \quad (3)$$

where $\mathbf{1}_N$ is a vector of size N with all entries equal to one and $\mathbf{1}_{s_0}$ is a one-hot vector which assigns 1 to the s_0 -th entry.

Proof: Since the transition does not depend on the perception actions, the following equation holds according to the matrix notation of the well-known “forward algorithm” in [9],

$$\begin{aligned} \mathbb{P}(o_{0:t}|a_{0:t}) &= \mathbf{1}_N^\top \mathbf{T} \text{diag}(\mathbf{O}_{o_t,1}^{a_t}, \dots, \mathbf{O}_{o_t,N}^{a_t}) \dots \\ &\quad \cdot \text{diag}(\mathbf{O}_{o_0,1}^{a_0}, \dots, \mathbf{O}_{o_0,N}^{a_0}) \mu_0. \end{aligned}$$

Eq. (3) is derived by replacing the initial distribution μ_0 with the one-hot distribution. ■

To compute the gradient $\nabla_\theta H(S_0|Y; \theta)$, we will need the value of $\nabla_\theta \log \mathbb{P}_\theta(y)$. We start by calculating the probability of an observation sequence $y = (o_{0:t}, a_{0:t})$ in M_θ .

Proposition 3: The probability of a sequence $y = (o_{0:t}, a_{0:t})$ of observations and perception actions in M_θ can be computed as follows:

$$\mathbb{P}_\theta(o_{0:t}, a_{0:t}) = \frac{\mathbb{P}(o_{0:t}|a_{0:t})}{\mathbb{P}(o_0|a_0)} \prod_{i=0}^t \pi_\theta(a_i|o_{0:i-1}). \quad (4)$$

where $o_{0:-1} := \lambda$ is the initial empty observation.

Proof: By the product rule of probability and causality, we can write the probability in the form,

$$\begin{aligned} \mathbb{P}_\theta(o_{0:t}, a_{0:t}) &= \mathbb{P}_\theta(o_t, a_t | o_{0:t-1}, a_{0:t-1}) \\ &\quad \cdot \mathbb{P}_\theta(o_{t-1}, a_{t-1} | o_{0:t-2}, a_{0:t-2}) \dots \mathbb{P}_\theta(o_1, a_1 | o_0, a_0) \\ &= \prod_{i=1}^t \mathbb{P}_\theta(o_i, a_i | o_{0:i-1}, a_{0:i-1}). \end{aligned} \quad (5)$$

For any $1 \leq i \leq t$, $i \in \mathbb{N}$, based on the multiplication rule of probability, we can decompose the conditional probability $\mathbb{P}(o_i, a_i | o_{0:i-1}, a_{0:i-1})$ as

$$\mathbb{P}_\theta(o_i, a_i | o_{0:i-1}, a_{0:i-1}) = \mathbb{P}(o_i | a_{0:i}, o_{0:i-1}) \pi_\theta(a_i | o_{0:i-1}). \quad (6)$$

Note that the probability of observing o_i given the sequence of actions $a_{0:i}$ and past observations $o_{0:i-1}$ is independent from

the policy π_θ . And the conditional probability $\mathbb{P}(o_i|a_{0:i}, o_{0:i-1})$ is derived as

$$\begin{aligned}
 & \mathbb{P}(o_i | a_{0:i}, o_{0:i-1}) \\
 &= \sum_{s_i \in \mathcal{S}} \mathbb{P}(o_i|s_i, a_{0:i}, o_{0:i-1}) \mathbb{P}(s_i|a_{0:i}, o_{0:i-1}) \\
 &\stackrel{(i)}{=} \sum_{s_i \in \mathcal{S}} E(o_i|s_i, a_i) \mathbb{P}(s_i|a_{0:i-1}, o_{0:i-1}) \\
 &= \sum_{s_i \in \mathcal{S}} E(o_i|s_i, a_i) \frac{\mathbb{P}(s_i, o_{0:i-1}|a_{0:i-1})}{\mathbb{P}(o_{0:i-1}|a_{0:i-1})} \\
 &= \frac{1}{\mathbb{P}(o_{0:i-1}|a_{0:i-1})} \sum_{s_i \in \mathcal{S}} E(o_i|s_i, a_i) \mathbb{P}(s_i, o_{0:i-1}|a_{0:i-1}) \\
 &\stackrel{(ii)}{=} \frac{\mathbb{P}(o_{0:i}|a_{0:i})}{\mathbb{P}(o_{0:i-1}|a_{0:i-1})}. \tag{7}
 \end{aligned}$$

where (i) is because 1) the probability of observing o_i is determined by state s_i and perception action a_i given the emission function $E(\cdot)$; and 2) the probability of reaching state s_i at the i -th time step does not depend on the perception action a_i when the action sequence $a_{0:i}$ is fixed, which can be calculated by equation (2). The equality (ii) is established by the definition of observable operators and Proposition 2. Substituting (7) into (6) and rewrite (5),

$$\begin{aligned}
 \mathbb{P}_\theta(o_{0:t}, a_{0:t}) &= \prod_{i=1}^t \frac{\mathbb{P}(o_{0:i}|a_{0:i})}{\mathbb{P}(o_{0:i-1}|a_{0:i-1})} \pi_\theta(a_i|o_{0:i-1}) \\
 &= \frac{\mathbb{P}(o_{0:t}|a_{0:t})}{\mathbb{P}(o_0|a_0)} \prod_{i=0}^t \pi_\theta(a_i|o_{0:i-1}). \tag{8}
 \end{aligned}$$

where $\mathbb{P}(o_{0:t}|a_{0:t})$ can be computed (6). ■

For a fixed initial state $s_0 \in \mathcal{S}_0$, the result of (4) becomes

$$\mathbb{P}_\theta(y|s_0) = \frac{\mathbb{P}(o_{0:t}|a_{0:t}, s_0)}{\mathbb{P}(o_0|a_0, s_0)} \prod_{i=0}^t \pi_\theta(a_i|o_{0:i-1}), \tag{9}$$

where the calculation of term $\mathbb{P}(o_{0:t}|a_{0:t}, s_0)$ is given in (3).

Numerical issues may arise in computing $\mathbb{P}_\theta(y|s_0)$ and $\mathbb{P}_\theta(y)$ because the probabilities can be close to 0 for a long horizon t . We can avoid these issues by taking the logarithm of both sides of equation (9). The following properties further show the gradient calculation can be simplified.

Proposition 4: Given $y = (o_{0:T}, a_{0:T})$, the gradient of $\log \mathbb{P}_\theta(y|s_0)$ can be computed as

$$\nabla_\theta \log \mathbb{P}_\theta(y|s_0) = \nabla_\theta \log \mathbb{P}_\theta(y) = \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t|o_{0:t-1}). \tag{10}$$

which is invariant with respect to the initial state s_0 .

Proof: First, let us take the logarithm on both sides in (9),

$$\begin{aligned}
 \log \mathbb{P}_\theta(y|s_0) &= \log \mathbb{P}(o_{0:t}|a_{0:t}, s_0) - \log \mathbb{P}(o_0|a_0, s_0) \\
 &\quad + \sum_{t=0}^T \log \pi_\theta(a_t|o_{0:t-1}).
 \end{aligned}$$

Then, we calculate the gradient on both sides,

$$\nabla_\theta \log \mathbb{P}_\theta(y|s_0) = \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t|o_{0:t-1}). \tag{11}$$

which can be shown to equal $\nabla_\theta \log \mathbb{P}_\theta(y)$ using (4). ■

The following proposition states that changing the policy does not affect the belief given a fixed observation.

Proposition 5: For any $y \in \mathcal{Y}_\theta$, the gradient of the logarithm of the posterior probability $\mathbb{P}_\theta(s_0|y)$ with respect to θ is 0, i.e., $\nabla_\theta \log \mathbb{P}_\theta(s_0|y) = 0$. Further, when $\mathbb{P}_\theta(s_0|y) \neq 0$, $\nabla_\theta \mathbb{P}_\theta(s_0|y) = 0$.

Proof: First, using the Bayes' rule, $\mathbb{P}_\theta(s_0|y) = \frac{\mathbb{P}_\theta(y|s_0)\mu_0(s_0)}{\mathbb{P}_\theta(y)}$. Taking the logarithm on both sides:

$$\log \mathbb{P}_\theta(s_0|y) = \log \mathbb{P}_\theta(y|s_0) + \log \mu_0(s_0) - \log \mathbb{P}_\theta(y),$$

and then taking the gradient on both sides with respect to θ ,

$$\begin{aligned}
 \nabla_\theta \log \mathbb{P}_\theta(s_0|y) &= \nabla_\theta \log \mathbb{P}_\theta(y|s_0) \\
 &\quad + \nabla_\theta \log \mu_0(s_0) - \nabla_\theta \log \mathbb{P}_\theta(y),
 \end{aligned}$$

From Proposition 4, we derive $\nabla_\theta \log \mathbb{P}_\theta(s_0|y) = 0$ because $\nabla_\theta \log \mathbb{P}_\theta(y|s_0) - \nabla_\theta \log \mathbb{P}_\theta(y) = 0$ and $\log \mu_0(s_0)$ is a constant. Furthermore, because $\nabla_\theta \log \mathbb{P}_\theta(s_0|y) = \frac{\nabla_\theta \mathbb{P}_\theta(s_0|y)}{\mathbb{P}_\theta(s_0|y)}$, when $\mathbb{P}_\theta(s_0|y) \neq 0$ but $\nabla_\theta \log \mathbb{P}_\theta(s_0|y) = 0$, we can derive that $\nabla_\theta \mathbb{P}_\theta(s_0|y) = 0$. ■

Theorem 1: The gradient of the conditional entropy w.r.t. the policy parameter θ is

$$\nabla_\theta H(S_0|Y; \theta) = \mathbb{E}_{y \sim \mathcal{M}_\theta} [H(S_0|Y = y; \theta) \nabla_\theta \log \mathbb{P}_\theta(y)]. \tag{12}$$

Proof: The conditional entropy is given by:

$$H(S_0|Y; \theta) = - \sum_{y \in \mathcal{O}^T} \mathbb{P}_\theta(y) \sum_{s_0 \in \mathcal{S}_0} \mathbb{P}_\theta(s_0|y) \log \mathbb{P}_\theta(s_0|y).$$

When computing the gradient using (1), the terms in the summation corresponding to $\mathbb{P}_\theta(s_0|y) = 0$ vanish. Let $\mathcal{S}_{0,y} = \{s_0 \in \mathcal{S} \mid \mathbb{P}_\theta(s_0|y) > 0\}$ be the set of initial states from which an observation y is possible. Then,

$$\begin{aligned}
 & \nabla_\theta H(S_0|Y; \theta) \\
 &= - \sum_{y \in \mathcal{Y}_\theta} \mathbb{P}_\theta(y) \sum_{s_0 \in \mathcal{S}_{0,y}} \left[\log \mathbb{P}_\theta(s_0|y) \nabla_\theta \mathbb{P}_\theta(s_0|y) \right. \\
 &\quad \left. + \mathbb{P}_\theta(s_0|y) \log \mathbb{P}_\theta(s_0|y) \nabla_\theta \log \mathbb{P}_\theta(y) + \nabla_\theta \mathbb{P}_\theta(s_0|y) \right]
 \end{aligned}$$

From Proposition 5, the gradient $\nabla_\theta \mathbb{P}_\theta(s_0|y) = 0$ when $\mathbb{P}_\theta(s_0|y) \neq 0$. Thus, we have

$$\begin{aligned}
 \nabla_\theta H(S_0|Y; \theta) &= - \sum_{y \in \mathcal{Y}_\theta} \mathbb{P}_\theta(y) \\
 &\quad \sum_{s_0 \in \mathcal{S}_{0,y}} \left[\mathbb{P}_\theta(s_0|y) \log \mathbb{P}_\theta(s_0|y) \nabla_\theta \log \mathbb{P}_\theta(y) \right], \\
 &= \mathbb{E}_{y \sim \mathcal{M}_\theta} [H(S_0|Y = y; \theta) \nabla_\theta \log \mathbb{P}_\theta(y)] \tag{13}
 \end{aligned}$$

where the expectation is taken with respect to the stochastic process of observations induced by the perception policy π_θ . Note that $\nabla_\theta \log \mathbb{P}_\theta(y)$ can be computed using (10). ■

Next, we show the convergence of a gradient-descent method under a common assumption of the policy space.

Assumption 1: For any time step $t \in [0, T]$, for any $(o_{0:t-1}, a) \in \mathcal{O}^t \times A$, both $\nabla_{\theta} \log \pi_{\theta}(a|o_{0:t-1})$ and $\nabla_{\theta}^2 \log \pi_{\theta}(a|o_{0:t-1})$ are bounded.

Theorem 2: Under Assumption 1, the entropy $H(S_0|Y; \theta)$ is Lipschitz-continuous and Lipschitz-smooth in θ .

Proof: We prove the theorem by showing that the gradient $\nabla_{\theta} H(S_0|Y; \theta)$ and Hessian $\nabla_{\theta}^2 H(S_0|Y; \theta)$ are both bounded. Referring to (13), by Jensen's inequality, we obtain

$$H(S_0|Y = y; \theta) \leq \log |S_0|. \quad (14)$$

The results of Proposition 4 show that

$$\nabla_{\theta} \log \mathbb{P}_{\theta}(y) = \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t|o_{0:t-1}). \quad (15)$$

which is bounded given Assumption 1 and the triangle inequality. Due to the boundedness of $H(S_0|Y = y; \theta)$ and $\nabla_{\theta} \log \mathbb{P}_{\theta}(y)$, the gradient $\nabla_{\theta} H(S_0|Y = y; \theta)$ is also bounded because $\mathbb{E}_{y \sim M_{\theta}}[H(S_0|Y = y; \theta) \nabla_{\theta} \log \mathbb{P}_{\theta}(y)] \leq \mathbb{E}_{y \sim M_{\theta}}(\beta) = \beta$ where $\beta = \max_{y \in \mathcal{Y}} \|H(S_0|Y = y; \theta) \nabla_{\theta} \log \mathbb{P}_{\theta}(y)\|_{\infty}$. Next, consider the Hessian

$$\begin{aligned} \nabla_{\theta}^2 H(S_0|Y; \theta) &= \mathbb{E}_{y \sim M_{\theta}} \left[\nabla_{\theta} H(S_0|Y = y; \theta) \nabla_{\theta} \log \mathbb{P}_{\theta}(y) \right. \\ &\quad \left. + H(S_0|Y = y; \theta) \nabla_{\theta}^2 \log \mathbb{P}_{\theta}(y) \right] \\ &= \mathbb{E}_{y \sim M_{\theta}} \left[H(S_0|Y = y; \theta) \nabla_{\theta}^2 \log \mathbb{P}_{\theta}(y) \right] \end{aligned} \quad (16)$$

where the last equality is because $\nabla_{\theta} \mathbb{P}_{\theta}(s_0|y) = \nabla_{\theta} \log \mathbb{P}_{\theta}(s_0|y) = 0$ and thus $\nabla_{\theta} H(S_0|Y = y; \theta) = 0$ by Proposition 5. Further, when $\nabla_{\theta}^2 \log \pi_{\theta}(a_t|o_{0:t-1})$ is bounded for all $t \in [0, T]$, $o_{0:t-1} \in \mathcal{O}^t$, and $a \in A$,

$$\nabla_{\theta}^2 \log \mathbb{P}_{\theta}(y) = \sum_{t=0}^T \nabla_{\theta}^2 \log \pi_{\theta}(a_t|o_{0:t-1}) \quad (17)$$

is bounded. Thus, $\nabla_{\theta}^2 \log \mathbb{P}_{\theta}(y)$ is bounded by the triangle inequality. Combining (14) and (17), with similar reasoning for the boundedness of the gradient, it holds that $\nabla_{\theta}^2 H(S_0|Y; \theta)$ is bounded. ■

To obtain the locally optimal policy parameter θ , we initialize a policy parameter θ_0 and carry out the gradient descent. At each iteration $\tau \geq 1$, $\theta_{\tau+1} = \theta_{\tau} - \eta \nabla_{\theta} H(S_0|Y; \theta_{\tau})$, where η is the step size. When using the *gradient descent* algorithm to compute the optimal θ , it may be computationally expensive to compute $\nabla_{\theta} H(S_0|Y; \theta)$ by enumerating all possible observations y . A sample approximation is employed such that at each iteration, we collect M sequences of observations $\{y_1, \dots, y_M\}$, and compute

$$\begin{aligned} \nabla_{\theta} H(S_0|Y; \theta) &\approx \hat{\nabla}_{\theta} H(S_0|Y; \theta) \\ &= -\frac{1}{M} \sum_{k=1}^M \sum_{s_0 \in \mathcal{S}_0} \left[\mathbb{P}_{\theta}(s_0|y_k) \log \mathbb{P}_{\theta}(s_0|y_k) \nabla_{\theta} \log \mathbb{P}_{\theta}(y_k) \right]. \end{aligned}$$

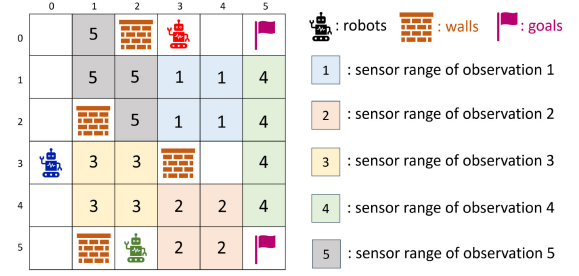


Fig. 1. A stochastic grid world monitored by a set of sensors.

IV. EXPERIMENTS

Consider a stochastic grid world environment (Fig. 1) with three types of robots, each starting from different positions to reach specific goals (flags). The blue robot (type 1) starts at (0, 3), the red (type 2) at (3, 0), and the green (type 3) at (5, 2). In line with standard grid world dynamics, each robot has a 20% chance of moving to one of the two nearest cells instead of the intended direction. Robots hitting walls or boundaries remain in place. Their policies are computed to maximize the probability of reaching their goals from their starting positions.

The environment has five sensors, each with a distinct range. The observer can query only one sensor at a time. For sensor $i \in \{1, \dots, 5\}$, if a robot is within its range and the observer queries it, the observer receives observation i with 90% probability and a null observation (“n”) with 10% probability (false negative). Otherwise, the observer also receives a null observation. We employ a *finite-state, observation-based* perception policy π with deterministic transitions, defined as a tuple $\pi := \langle Q, \mathcal{O}, \mathcal{A}, \delta, \psi, q_0 \rangle$ where 1) Q is a set of memory states. 2) \mathcal{O}, \mathcal{A} are a set of inputs and a set of outputs, respectively. 3) $\delta : Q \times \mathcal{O} \rightarrow Q$ is a deterministic transition function that maps a state and an input (q, o) to a next state $\delta(q, o)$. 4) $\psi : Q \rightarrow \mathcal{D}(\mathcal{A})$ is a probabilistic output function. 5) q_0 is the initial state. Specifically, given an integer $K \geq 0$, the state set $Q = \{O^{\leq K}\}$ are the set of observations with length $\leq K$. For each $q \in Q$, $\delta(q, o) = q'$ is defined such that $q' = \text{suffix}^{=K}(q \cdot o)$ ¹ is the last (up to) K observations after appending the observation o to q . The probabilistic output function is parameterized as, $\psi_{\theta}(a|q) = \frac{\exp(\theta_{q,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{q,a'})}$, where $\theta \in \mathbb{R}^{|Q \times \mathcal{A}|}$ is the policy parameter vector. Given observations $o_{0:t}$, the policy $\pi_{\theta}(a|o_{0:t}) = \psi_{\theta}(a|\delta(q_0, o_{0:t}))$. The softmax policy satisfies Assumption 1 and is differentiable. In the experiments, we set the length of memory $K = 2$.

The initial distribution μ_0 is 0.1, 0.4, 0.5 for type 1, type 2, and type 3, respectively. Figure 2a illustrates the convergence of the policy gradient.² When the algorithm converges, the conditional entropy $H(S_0|Y; \theta)$ approaches approximately

¹ $\text{suffix}^{=K}(w)$ is the last K symbols of string w if $|w| \geq K$ or w itself otherwise.

²We sample $M = 2000$ trajectories and set the horizon $T = 10$ for each iteration. The fixed step size of the gradient descent algorithm is set to be 0.5. We run $N = 2000$ iterations on the 12th Gen Intel Core i7-12700, the average time consumed for one iteration is 6.7 seconds.

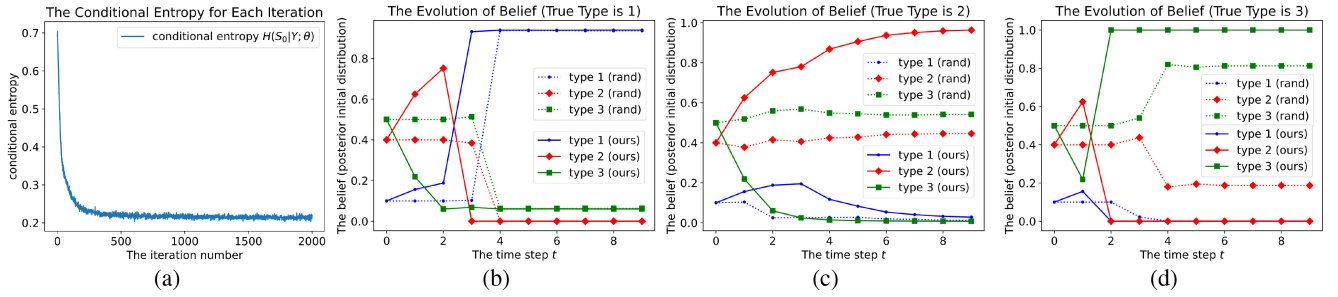


Fig. 2. The convergence results of policy gradients algorithm and beliefs evolution for different true types of robots.

0.22. This indicates that the observations provide substantial information about the robot's type on average.

To provide an intuitive understanding, we use the computed perception policy θ^* , called the “min-entropy” policy, to evaluate the posterior belief $P_{\theta^*}(S_0|o_{0:i})$ of the robot's type given observations $o_{0:i}$ for $0 \leq i \leq T$. For comparison, we also generated random policies by randomly selecting policy parameter θ and chose the one with the lowest conditional entropy, referred to as the “random policy.” We then sampled an initial state/type and used both policies to collect observations. Figures 2b, 2c, and 2d show how the agent's belief evolves for different types under the random and min-entropy policies. For each robot type, the min-entropy policy allows for accurate type identification by time $T = 9$, with probabilities of 0.94 for type 1, 0.96 for type 2, and 1.00 for type 3. In contrast, the random policy results in lower probabilities: 0.93 for type 1, 0.45 for type 2, and 0.81 for type 3. Both policies perform well for type 1, but the min-entropy policy significantly outperforms the random policy for identifying type 2 and type 3 robots.

V. CONCLUSION AND FUTURE WORK

In this letter, we introduce conditional entropy to quantify uncertainty and formulate a problem of minimizing the uncertainty of the initial state in an HMM. To solve this optimization problem, we develop a gradient descent algorithm and derive the gradient of conditional entropy using observable operators. Further, we prove that the conditional entropy is Lipschitz-continuous and L -smooth with respect to the policy parameters under certain assumptions for the policy search space. An interesting direction for future research would be to explore active perception under varying assumptions for the perception agent. For example, the agent with imprecise knowledge about the model dynamics. It is also interesting to consider a general formulation where the active perception agent can change both the transition dynamics and emission function of the partially observable systems.

REFERENCES

- [1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, “On the theory of policy gradient methods: Optimality, approximation, and distribution shift,” *J. Mach. Learn. Res.*, vol. 22, no. 98, pp. 1–76, 2021.
- [2] A. Andreopoulos and J. K. Tsotsos, “A theory of active object localization,” in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 903–910.
- [3] M. Araya, O. Buffet, V. Thomas, and F. Charpillet, “A POMDP extension with belief-dependent rewards,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1–9.
- [4] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, “Revisiting active perception,” *Auton. Robots*, vol. 42, no. 2, pp. 177–196, 2018.
- [5] F. Basile, G. De Tommasi, and C. Motta, “Assessment of initial-state-opacity in live and bounded labeled petri net systems via optimization techniques,” *Automatica*, vol. 152, Jun. 2023, Art. no. 110911.
- [6] S. Casao, Á. Serra-Gómez, A. C. Murillo, W. Böhrmer, J. Alonso-Mora, and E. Montijano, “Distributed multi-target tracking and active perception with mobile camera networks,” *Comput. Vis. Image Underst.*, vol. 238, Jan. 2024, Art. no. 103876.
- [7] M. Dunbabin and L. Marques, “Robots for environmental monitoring: Significant advancements and applications,” *IEEE Robot. Autom. Mag.*, vol. 19, no. 1, pp. 24–39, Mar. 2012.
- [8] M. Egorov, M. Kochenderfer, and J. Uudmae, “Target surveillance in adversarial environments using POMDPs,” in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2473–2479.
- [9] H. Jaeger, “Observable operator models for discrete stochastic time series,” *Neural Comput.*, vol. 12, no. 6, pp. 1371–1398, 2000.
- [10] H. Jaeger and S. Augustin, “Observable operator processes and conditioned continuation representations,” GMD, São Paulo, Brazil, Working Paper, Jan. 1997.
- [11] M. Lauri and F. Oliehoek, “Multi-agent active perception with prediction rewards,” in *Proc. 34th Conf. Neural Inf. Process. Syst.*, 2020, pp. 13651–13661.
- [12] T. L. Molloy and G. N. Nair, “Smoother entropy for active state trajectory estimation and obfuscation in POMDPs,” *IEEE Trans. Autom. Control*, vol. 68, no. 6, pp. 3557–3572, Jun. 2023.
- [13] A. Saboori and C. N. Hadjicostis, “Verification of initial-state opacity in security applications of discrete event systems,” *Inf. Sci.*, vol. 246, pp. 115–132, Oct. 2013.
- [14] Y. Savas, M. Hibbard, B. Wu, T. Tanaka, and U. Topcu, “Entropy maximization for partially observable Markov decision processes,” *IEEE Trans. Autom. Control*, vol. 67, no. 12, pp. 6948–6955, Dec. 2022.
- [15] M. Shen and J. P. How, “Active perception in adversarial scenarios using maximum entropy deep reinforcement learning,” in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 3384–3390.
- [16] C. Shi, Y. Bu, and J. Fu, “Information-theoretic opacity-enforcement in Markov decision processes,” in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 6779–6787.
- [17] P. Yang, S. Koga, A. Asgharivaskasi, and N. Atanasov, “Policy learning for active target tracking over continuous $SE(3)$ trajectories,” in *Proc. 5th Annu. Learn. Dyn. Control Conf.*, Jun. 2023, pp. 64–75.
- [18] P. Yang, Y. Liu, S. Koga, A. Asgharivaskasi, and N. Atanasov, “Learning continuous control policies for information-theoretic active perception,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 2098–2104.
- [19] K. Zhou and S. I. Roumeliotis, “Multirobot active target tracking with combinations of relative observations,” *IEEE Trans. Robot.*, vol. 27, no. 4, pp. 678–695, Aug. 2011.