Learning and Planning for Time-Varying MDPs Using Maximum Likelihood Estimation

Melkior Ornik and Ufuk Topcu

Abstract

This paper proposes a formal approach to learning and planning for agents operating in a priori unknown, time-varying environments. The proposed method computes the maximally likely model of the environment, given the observations about the environment made by an agent earlier in the system run and assuming knowledge of a bound on the maximal rate of change of system dynamics. Such an approach generalizes the estimation method commonly used in learning algorithms for unknown Markov decision processes with time-invariant transition probabilities, but is also able to quickly and correctly identify the system dynamics following a change. Based on the proposed method, we generalize the exploration bonuses used in learning for time-invariant Markov decision processes by introducing a notion of uncertainty in a learned time-varying model, and develop a control policy for time-varying Markov decision processes based on the exploitation and exploration trade-off. We demonstrate the proposed methods on four numerical examples: a patrolling task with a change in system dynamics, a two-state MDP with periodically changing outcomes of actions, a wind flow estimation task, and a multi-arm bandit problem with periodically changing probabilities of different rewards.

I. Introduction

A variety of intelligent agents — notably, autonomous systems — are commonly required to operate in unknown environments [1]–[3], necessitating the use of learning in order to complete their tasks. While methods for learning and planning for agents in unknown environments exist in a variety of frameworks [4]–[8], most of them assume that the environment in which the agent operates is unchanged over the course of the agent's operation. Such an assumption is useful for two reasons. Firstly, it ensures predictability of the outcomes of the actions performed by the agent after it ceases to learn. Secondly, in environments with stochastic dynamics, it allows construction of an estimate of the dynamics by performing repeated experiments and observing the outcomes [6], [8]–[14]. However, this assumption is often not realistic for systems operating on long-term missions outside a strictly controlled environment, e.g., a laboratory testbed. Taking an example of an extraterrestrial rover mission, changes in the environment may be a consequence of regular, predictable events such as intra-day or seasonal temperature variations [15] or may result from more complex phenomena that are difficult to predict: e.g., terrain changes due to wind (see references in [16] for a detailed study).

In contrast to assuming time-invariance, accounting for time-varying changes in the environment presents a major challenge to learning and planning. A naive approach — restarting the learning process whenever the environment changes — does not make sense: the environment is possibly continually changing. Restarting the learning process whenever the environment *sufficiently* changes, or sufficient length of time passes, would both neglect the environmental changes between the process restarts and rely on heuristics in deciding when to restart learning. Restarting too often will lead to the agent spending too much time on learning, and lacking time to perform its task. On the other hand, restarting the learning process too rarely can lead to unreliable learning outcomes.

This paper develops a method that neither assumes that the environment is time-invariant, nor uses discrete learning episodes to artificially adapt the agent to a changing environment while discarding all old learned information. The framework of this paper is one of time-varying Markov decision processes (TVMDPs) [17], i.e., discrete-time, finite-state stochastic control processes where transitions from one state to another are governed by a time-varying transition probability function. Building on the maximum-likelihood approach to learning and planning in unknown time-invariant environments [18]–[20], we propose a *change-conscious maximum likelihood estimate* (CCMLE) that computes a time-varying transition probability function that is *maximally likely*, given (i) the previously observed outcomes of the agent's actions and (ii) a priori known bounds on the rate of change of the transition probabilities. In our motivational narrative, such bounds may come from prior study of the causes behind the changes in transition probabilities — for example, wind or temperature change [21].

Using the proposed estimation method, we additionally propose an *active learning* policy, seeking to ensure that the agent estimates the system dynamics as quickly as possible during a single system run, and incorporate such a policy in a joint learning and planning mechanism, enabling the agent to use active learning in order to accomplish its control objective.

An attractive feature of the proposed estimation method for TVMDPs is its interpretation as a generalization of a standard estimation method on time-invariant MDPs, described in [11]. Namely, if the environment is time-invariant, i.e., the rate of

M. Ornik is with the Department of Aerospace Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. e-mail: mornik@illinois.edu

U. Topcu is with the Department of Aerospace Engineering and Engineering Mechanics and the Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX 78712, USA. e-mail: utopcu@utexas.edu

change of the environment is known to be 0, the proposed method provides estimates that match the estimates provided by the method used for time-invariant MDPs. Such a property ensures that the proposed method retains all the appealing theoretical results afforded to the previously used estimation method when the environment is time-invariant, while allowing for successful estimation when the environment is known to change over time.

While the framework of TVMDPs has been described in [17], the work contained therein solely discusses optimal control policies for a priori known TVMDPs. Our interest is in learning and planning for unknown TVMDPs; to the best of our knowledge, such a problem has not been discussed previously.

Frameworks similar to TVMDPs include the following:

- Time-dependent MDPs [22], where the dependence of transition probabilities on time is encoded by appending a continuous time stamp as a coordinate in the state space, thus yielding a continuous-state MDP.
- Time-varying Markov-switching models [23], which do not include a notion of a control action.
- Semi-Markov decision processes and related frameworks [9], [24], [25], where the transition probabilities themselves are time-invariant, but the time needed to perform a transition may vary.
- ε-stationary MDPs [26]–[28], which allow for time-varying transition probabilities only inasmuch as they remain close to constant over time.

Learning and planning for agents operating in the last two frameworks have been discussed at length; see, e.g., [9], [27]. However, the nature of these frameworks — with transition probabilities that, potentially disregarding bounded disturbances, do not change over time — yields reinforcement learning methods that are not useful in the setting where transition probabilities may significantly vary over long periods of time. Learning of time-varying Markov-switching models (see, e.g., [29]) is more similar to the problem of learning for TVMDPs. However, as that framework does not include explicit decision-making, learning simply relies on passively collecting data from multiple system runs. While similar in the estimation part — although with technical differences due to different assumptions on previous knowledge — our proposed method seeks to make the agent actively learn by performing those actions that are expected to reduce the uncertainty in the learned model. Finally, time-dependent MDPs naturally fall into the category of continuous-state MDPs. However, learning methods for continuous-state MDPs are generally computationally intractable [30].

The organization of this paper is as follows. Section II recalls the definition of a TVMDP and poses problems of optimal learning — during a single system run — and optimal control for an agent operating in a TVMDP with a priori unknown transition probabilities. Section III introduces the key element of the proposed learning method: a change-conscious maximum likelihood estimate (CCMLE) of the TVMDP's transition probabilities, given prior observations and knowledge about the rate of change of probabilities. It additionally relates the CCMLE to estimates produced by standard estimation in the case of time-invariant MDPs, and provides theoretical results describing the CCMLE for the case when the transition probabilities change over time. Section IV develops a measure of uncertainty of a CCMLE and relates it to measures of uncertainty used in learning and planning techniques for time-invariant MDPs. Section V uses this notion of uncertainty to propose optimal learning and control policies for an agent operating in an unknown, time-varying environment. In particular, Section V-A considers a policy that minimizes the agent's uncertainty, while Section V-B uses the measure of uncertainty as an "exploration bonus" in proposing a control policy based on the exploration-exploitation framework, seeking to minimize the uncertainty while directing the agent to progress towards its objective. Building on the theoretical results of the previous section, Section VI illustrates the developed theory by considering learning and planning for an agent in two simple scenarios: Section VI-A discusses a scenario of a one-off change in transition probabilities during a patrolling mission, while Section VI-B considers a setting of regular, periodic changes in transition probabilities on a two-state MDP that models a two-arm bandit. In these sections, we compare results attained by the proposed method to the methods introduced in previous work, and show that the proposed method indeed leads to smaller estimation errors and expedited completion of control objectives. Proofs of theoretical results are provided in Appendix A. Finally, Appendix B builds on Section VI to describe numerical results obtained in two more complicated settings: Appendix B-A describes estimation of wind flow using a pilot balloon with CCMLE, and Appendix B-B modifies the two-arm bandit setting to a multi-arm bandit.

Notation. Symbol $\mathbb{N}_0$ denotes all nonnegative integers. Function $d : \mathbb{R}^n \times \mathbb{R}^n \to [0, +\infty)$ denotes the Euclidean distance on $\mathbb{R}^n$. For a set $\mathcal{P} \subseteq \mathbb{R}^n$, $\mathrm{diam}(\mathcal{P})$ denotes the diameter of the set: $\mathrm{diam}(\mathcal{P}) = \sup\{d(x, y) \mid x, y \in \mathcal{P}\}$. For a set X, $|X|$ denotes its cardinality.

II. Problem Statement

Consider a time-varying Markov decision process (TVMDP) [17] $\mathcal{M} = (S, A, P)$, where $S = \{s^1, \ldots, s^n\}$ is the state space, A is the set of actions, and

$$P : S \times A \times S \times \mathbb{N}_0 \to [0, 1]$$

is a transition probability function. Namely, $P(s, a, s', t)$ denotes the probability that an agent positioned at state $s \in S$ at time $t \in \mathbb{N}_0$ will, after performing action $a \in A$, transition to a state $s' \in S$ at time $t + 1$. Naturally, $\sum_{s' \in S} P(s, a, s', t) = 1$ for all $s \in S$, $a \in A$, $t \in \mathbb{N}_0$. If s_0 is the agent's initial state, the agent's *path* until time $T \in \mathbb{N}_0$ is denoted by $\sigma = (s_0, \ldots, s_T)$, while the agent's corresponding actions are given by $\alpha = (a_0, \ldots, a_{T-1})$. A *time-varying policy* on a TVMDP $\mathcal{M}$ is a sequence $\pi = (\pi_1, \pi_2, \ldots)$, where $\pi_t \in A$ depends on time t, the agent's position s_t, as well as the agent's previous positions $s_0, \ldots, s_{t-1}$. If the transition probability function P is a priori unknown, π_t may also depend on the agent's estimate of P at time t. A *time-invariant policy* on a TVMDP $\mathcal{M}$ is a policy that depends solely on the agent's position s_t, i.e., with a slight abuse of notation, $\pi : S \to A$.

TVMDPs seek to model an environment in which the agent dynamics may change over time. This framework is a generalization of classical time-invariant Markov decision processes (MDPs); in standard MDPs, transition probability function P is not dependent on time. In the remainder of the paper, if a transition probability is time-invariant, we will denote it by $P(s, a, s', *)$.

We assume that the transition probability function P is unknown to an agent at the beginning of the system run, i.e., prior to $t = 0$. As in the previous work [6], [10], [12], [14], [31] on time-invariant MDPs, we study two objectives:

(i) learning the transition probabilities as efficiently and correctly as possible during a single system run, and
(ii) for a reward function $R : S \times A \to \mathbb{R}$, maximizing the agent's expected collected reward over a period of time.

In time-invariant MDPs, because the transition probabilities do not change over time, it is possible to learn the transition probabilities at every state-action pair (s, a) with an arbitrarily small error, by repeatedly visiting the state s, performing the action a, and observing the outcome. By the law of large numbers,

$$\lim_{\#(s,a)\to\infty} \frac{\#(s,a,s')}{\#(s,a)} = P(s,a,s',*) \tag{1}$$

with probability 1, where $\#(s,a)$ denotes the number of times that action a was performed at state s, and $\#(s,a,s')$ denotes the number of times that performing action a at state s led to the agent immediately moving to state s'. Hence, while the meaning of learning "as efficiently and correctly as possible" in (i) depends on the particular formal definition, it is — under some ergodicity assumptions — possible to learn the transition probabilities $P(\cdot,\cdot,\cdot,*)$ within one system run and with an arbitrarily small error. After learning these probabilities, it is then straightforward (see, e.g., [32]) to determine a policy that comes arbitrarily close to maximizing the agent's expected collected reward, thus solving objective (ii).

In the case of TVMDPs, it is impossible to learn the transition probabilities during a single system run with an arbitrarily small error, as these probabilities may continually change. In fact, if the transition probabilities at different times are entirely independent, the agent would only have one time step (i.e., one action) to learn the transition probabilities at $|S||A|$ state-action pairs. In such a case, any attempt at learning is meaningless. Even if the transition probabilities are not independent, i.e., it is known that there exists $\varepsilon_t \in [0, 1)$ such that

$$|P(s,a,s',t+1) - P(s,a,s',t)| \leq \varepsilon_t \tag{2}$$

for all $s, s' \in S$, $a \in A$ and $t \in T$, perfect knowledge of *all* transition probabilities $P(s,a,s',\tau)$ for $\tau \neq T$ only implies that

$$\begin{aligned} &P(s,a,s',T) \in [P(s,a,s',T-1) - \varepsilon_{T-1}, P(s,a,s',T-1) + \varepsilon_{T-1}] \text{ and} \\ &P(s,a,s',T) \in [P(s,a,s',T+1) - \varepsilon_T, P(s,a,s',T+1) + \varepsilon_T], \end{aligned}$$

and just one sample collected from (s,a) at time T is not sufficient to determine the value of $P(s,a,s',T)$.

The above discussion behooves us to interpret objective (i) in the following way.

Problem 1 (Optimal learning in TVMDPs)**.** *Determine a policy π^* such that, at every time $t \geq 0$, after taking action $\pi^*_t \in A$, the* uncertainty *in the estimated transition probabilities $P(\cdot,\cdot,\cdot,t)$ is minimized.*

We purposefully leave the notion of uncertainty vague at this point. Sections III, IV, and V-A of this paper will be dedicated to designing a meaningful estimate of transition probabilities, defining the notion of uncertainty of such an estimate, and determining a policy π that solves the optimal learning problem.

Our interpretation of objective (ii) also follows from the time-varying nature of the environment. Since the transition probabilities change, the optimal policy should also be time-varying. Hence, we are interested in maximizing the expected total reward collected during a single system run. Largely for notational purposes, we express the problem in terms of expected average rewards on an infinite system run.

Problem 2 (Optimal control in TVMDPs)**.** *Determine a policy π^* that maximizes*

$$\mathbb{E}\left[\liminf_{T\to\infty} \frac{\sum_{t=0}^{T} R(s_t, \pi^*_t, s_{t+1})}{T}\right],$$

where s_t is the agent's state at time t.

In both the optimal learning and optimal control problems, we allow π^*_t to depend on the agent's path until time t and its estimates of transition probabilities.

By appending the state space S by the time coordinate and interpreting the transition probabilities and rewards as being defined on the state space $S \times \mathbb{N}_0$, the optimal control problem on TVMDP $\mathcal{M}$ is equivalent to a standard optimal control problem on a countably infinite MDP $\hat{\mathcal{M}}$, with a finite set of actions A and an averaged reward objective. A detailed discussion of such a problem in the context of infinite MDPs is given in [32]. With a slight change to a discounted reward objective instead of an averaged reward, the work in [32] shows that such a problem admits a stationary optimal policy, which naturally translates to a time-varying optimal policy on $\mathcal{M}$. However, such a policy can only be found if the transition probabilities are a priori *known*. Finding an optimal control policy under the stipulation that P is unknown at the beginning of the system run is clearly impossible. In Section V-B we will propose a method motivated by the exploration-exploitation framework of previous work [6], [10], [12], [14], [31], seeking to actively learn about the transition probabilities in order to be able to collect higher rewards.

We now proceed to discuss the initial building block of our method for learning and planning in TVMDPs: estimating the transition probabilities.

III. Change-Conscious Maximum Likelihood Estimate

The objective of this section is to develop a method for estimating the transition probabilities $P(s,a,s',t)$, $t \leq T$, given the observations of the agent's motion until time T. To this end, we develop a *change-conscious maximum likelihood estimate* (CCMLE) which produces a set of probability distributions $P(s,a,\cdot,t)$ for all $s \in S$, $a \in A$, and $t < T$, for which the probability of the agent's observed path $\sigma = (s_0, s_1, \ldots, s_T)$ until time T is maximal, given the agent's actions $\alpha = (a_0, \ldots, a_{T-1})$ and an known a priori on the rate of change of transition probabilities over time.

Let us now consider the path $\sigma = (s_0, s_1, \ldots, s_T)$. For ease of notation, we assume that $A = \{a\}$, i.e., that the TVMDP $\mathcal{M}$ is a time-varying Markov chain; if $|A| > 1$, we can separate the agent's paths into $|A|$ (possibly disconnected) paths, one for each action. The probability of the agent following the path σ is

$$\mathbb{P}(\sigma) = \prod_{t=0}^{T-1} P(s_t, a, s_{t+1}, t). \tag{3}$$

We note that $\mathbb{P}(\sigma)$ as defined in (3) depends on the values $P(s,a,s',t)$, with $t < T$. Given the agent's path σ, we want to determine the parameter set $\{\tilde{P}(s,a,s',t) \mid s,s' \in S, t < T\}$ (in future to be denoted by $\tilde{P}_{t<T}$) which is *most likely* to have produced such a path [33]. In other words, we want to find a transition probability function $\tilde{P}_{t<T}$ that maximizes $\mathbb{P}(\sigma)$, i.e., a solution to the problem

$$\begin{array}{lll} \max\limits_{\tilde{P}_{t<T}} & \prod\limits_{t=0}^{T-1} \tilde{P}(s_t, a, s_{t+1}, t) & \\ \text{s.t.} & \tilde{P}(s,a,s',t) \geq 0 & \text{for all } s,s' \in S,\ t < T, \\ & \sum\limits_{s' \in S} \tilde{P}(s,a,s',t) = 1 & \text{for all } s \in S,\ t < T. \end{array} \tag{4}$$

The set of solutions to problem (4) is trivially given by

$$\left\{ \tilde{P}_{t<T} \mid \tilde{P}(s_t, a, s_{t+1}, t) = 1 \text{ for all } t < T \right\}.$$

Such a result is intuitive: without any additional constraints, transition probabilities at different times are possibly independent. Thus, the transition probability function that will generate the observed outcomes with the highest probability is the one that ensures that all the observed outcomes happen with probability 1. However, learning transition probabilities is impossible in this framework, as all observations at times $t < T$ make no impact on the estimate of transition probabilities for time $t = T$. Thus, learning needs to be based on additional knowledge of the relationship between transition probabilities at different times. We do not assume any specific knowledge about the changes other than the maximal rate of change of transition probabilities, i.e., $\varepsilon_t \in [0,1]$, $t \in \mathbb{N}_0$, which satisfy (2). Thus, the CCMLE problem is given by

$$\begin{array}{lll} \max\limits_{\tilde{P}_{t<T}} & \prod\limits_{t=0}^{T-1} \tilde{P}(s_t, a, s_{t+1}, t) & \\ \text{s.t.} & \tilde{P}(s,a,s',t) \geq 0 & \text{for all } s,s' \in S,\ t < T, \\ & \sum\limits_{s' \in S} \tilde{P}(s,a,s',t) = 1 & \text{for all } s \in S,\ t < T. \\ & \tilde{P}(s,a,s',t+1) - \tilde{P}(s,a,s',t) \leq \varepsilon_t & \text{for all } s,s' \in S,\ t < T, \\ & \tilde{P}(s,a,s',t) - \tilde{P}(s,a,s',t+1) \leq \varepsilon_t & \text{for all } s,s' \in S,\ t < T, \end{array} \tag{5}$$

where the decision variables are $\tilde{P}(s,a,s',t)$ for all $s,s' \in S$, $t < T$.

Noting that the product in the objective function of (5) is nonnegative, and the logarithm function is monotonic, (5) can be replaced by the constrained log-likelihood problem

$$\begin{array}{lll} \min\limits_{\tilde{P}_{t<T}} & -\sum\limits_{t=0}^{T-1} \log \tilde{P}(s_t,a,s_{t+1},t) & \\ \text{s.t.} & \tilde{P}(s,a,s',t) \geq 0 & \text{for all } s,s' \in S,\ t<T, \\ & \sum\limits_{s' \in S} \tilde{P}(s,a,s',t) = 1 & \text{for all } s \in S,\ t<T, \\ & \tilde{P}(s,a,s',t+1) - \tilde{P}(s,a,s',t) \leq \varepsilon_t & \text{for all } s,s' \in S,\ t<T, \\ & \tilde{P}(s,a,s',t) - \tilde{P}(s,a,s',t+1) \leq \varepsilon_t & \text{for all } s,s' \in S,\ t<T, \end{array} \tag{6}$$

with the understanding that $\log 0 = -\infty$.

The optimization problem in (6) is a convex optimization problem with a linear set of constraints and $T|S|^2$ decision variables $\tilde{P}(s,a,s',t)$; for $|A|>1$, there would be $T|S|^2|A|$ decision veriables. Alternatively, as discrete distributions $\tilde{P}(s,a,\cdot,t)$ for different s are not coupled by any of the constraints, we can instead treat (6) as $|S|$ problems with $T|S|$ decision variables, which we will do in the remainder of the paper.

The maximal value of the objective function in (6) is not $+\infty$ because $\tilde{P}_{t<T}$ defined by $\tilde{P}(s,a,s',t) = 1/|S|$ for all $s,s' \in S$ and $t<T$ is in the feasible set, and produces a real value for the objective function. Thus, by continuity, the objective function attains a minimum in the feasible set. Such a minimum may not be unique. In the remainder of the paper, we use $\tilde{P}^T : S \times A \times S \times \{0,1,\ldots,T-1\} \to [0,1]$ or $\tilde{P}^T_{t<T}$ to denote any CCMLE obtained from the observations until time T, i.e., immediately before taking action a_T.

The following result, with the proof in Appendix A, shows that the CCMLE directly generalizes the estimate from (1) for the case of time-invariant transition probabilities.

Proposition 3. *Assume that $\varepsilon_t = 0$ for all $t \in \mathbb{N}_0$. Then, $\tilde{P}^T(s,a,s',*) = \#(s,a,s')/\#(s,a)$ for all $s,s' \in S$, $a \in A$, $T \in \mathbb{N}_0$, where $\#(s,a,s') = |\{t \in \{0,\ldots,T-1\} \mid s_t = s, a_t = a, s_{t+1} = s'\}|$ and $\#(s,a) = |\{t \in \{0,\ldots,T-1\} \mid s_t = s, a_t = a\}|$.*

As discussed in Section II, due to the time-varying nature of transition probabilities, it is generally not possible to ensure that the solution to (6), or any other estimation method, indeed correctly estimates the transition probabilities of the TVMDP. Nonetheless, Proposition 3 shows that, if the transition probabilities are known to be time-invariant, the produced estimates will be asymptotically correct with probability 1. We now generalize this claim to the case in which the transition probabilities are known to become time-invariant after finitely many time steps.

Theorem 4. *Let $N \in \mathbb{N}_0$. Assume that $\varepsilon_t = 0$ for all $t \geq N$. Then,*

$$\lim_{\#(s,a)\to\infty} \tilde{P}^T(s,a,s',T-1) = P(s,a,s',T-1)$$

for all $s,s' \in S$, $a \in A$ with probability 1.

The proof of Theorem 4 is in Appendix A. Theorem 4 states that, asymptotically, the CCMLE will disregard the possible changes in the transition probabilities that occur at the beginning of the system run, *as long as the transition probabilities are known to be time-invariant after a finite time*. Such a property is shared with the estimate $\#(s,a,s')/\#(s,a) \approx P(s,a,s',t)$ which implicitly assumes that the transition probabilities are time-invariant from the start of the system run. The following theorem shows that, under the condition that $P(s,a,s',t) = 1$ after some $t = N$, the CCMLE actually learns $P(s,a,s',t)$ correctly *in finite time*, as opposed to asymptotically, and without requiring knowledge that $P(s,a,s',t)$ is constant in t.

Theorem 5. *Let $N \in \mathbb{N}_0$, $s,s' \in S$, and $a \in A$. Assume that $\varepsilon_t = \varepsilon \in (0,1]$ for all $t \in \mathbb{N}_0$. Let $P(s,a,s',t) = 1$ for all $t \geq N$. Then, $\tilde{P}^{T+1}(s,a,s',T) = 1$ for all $T \geq N + 1/\varepsilon$ such that $(s_T,a_T) = (s,a)$.*

We again invite the reader to see Appendix A for the proof of Theorem 5.

Remark 6. *If (s,a) is not visited at time T, any probability distribution $\tilde{P}^{T+1}(s,a,\cdot,T)$ which satisfies the constraints in* (6) *can be chosen without any impact on the objective function. Such a possibility reflects the agent's lack of knowledge about the drift of $P(s,a,s',\cdot)$ since the last time that the agent obtained any information about it.*

The conditions of Theorem 4 and Theorem 5 require the system to be *eventually time-invariant* (ETI). ETI systems appear naturally in settings where changes occur on short time scales between long periods of unchanged behavior, e.g., weather fronts [34]. ETI TVMDPs are also a stochastic discretization of classical ETI control systems [35].

Theorem 5 proves that the CCMLE holds a significant advantage over the estimate given by $\#(s,a,s')/\#(s,a) \approx P(s,a,s',t)$ in the case where a transition probability changes over time and ultimately becomes 1. Although $\#(s,a,s')/\#(s,a)$ will converge to 1 as $\#(s,a) \to \infty$, such convergence will be slow: if (s,a) has been visited v times prior to $P(s,a,s',t)$ becoming constantly 1, it is simple to see that it can take up to $v(1-\eta)/\eta$ additional visits to (s,a) for the estimate of $P(s,a,s',T)$ to have an error no larger than η, and the error may never become 0. On the other hand, after a *single visit* to (s,a) at time no earlier than $1/\varepsilon$ after $P(s,a,s',t)$ becomes constantly 1, the estimating procedure (6) is guaranteed to produce the correct transition probability.

It is possible to modify the classical estimate $\#(s,a,s')/\#(s,a) \approx P(s,a,s',t)$ in order to satisfy Theorem 5: we can simply make the agent "forgetful" (as discussed in related frameworks in, e.g., [36], [37]) and only calculate the estimate based on the outcomes of actions performed in the last $1/\varepsilon$ time steps. In that case, assuming that the transition probabilities satisfy the very specific condition of Theorem 5, the estimate produced in such a way would satisfy the claim of Theorem 5. However, choice of $1/\varepsilon$ would be arbitrary; an analogous claim to that of Theorem 5 would hold for an estimate with any finite memory length, and it is possible that an estimate with a longer or shorter memory would provide better results in general.

The same notion of forgetfulness gives rise to an attractive heuristic to reduce the complexity of computing the CCMLE. Instead of solving an optimization problem with up to $T|S|$ variables at every time step — thus, a problem that grows without bound in size as the system run progresses — we can choose to "forget" the variables, i.e., transition probabilities, that are far enough before the current time. Theorem 5 guarantees that, if we choose to exclude all variables that correspond to time steps that occurred more than $1/\varepsilon$ steps ago, such a *forgetful CCMLE* of $\tilde{P}^T(s,a,s',T-1)$ will not be affected under the conditions of Theorem 6. In general, forgetfulness is not without effect — and, unlike for the classical estimate — we have no reason to believe that the error of the estimates produced by a forgetful CCMLE will be smaller than the one produced by a CCMLE without forgetting. However, the simulations in Section VI will show that the difference between a CCMLE and a forgetful CCMLE may be small, while the CCMLE requires significantly more computational power. In particular, for each action the number of decision variables in the forgetful CCMLE is $|S|/\varepsilon$ — a value independent of T — while in the CCMLE it is $T|S|$.

Having described the CCMLE method for estimating the time-varying transition probabilities, we now continue to the second step in our solution of the optimal learning problem: quantifying how unsure we are about transition probabilities given the agent path.

IV. Uncertainty in Estimation

The definition of uncertainty proposed in this paper arises out of two previously identified intuitive causes for uncertainty of an estimate [38]. Namely, (i) if a single additional observation makes a large difference in an estimate, such an estimate is highly unstable [38], and (ii) if there multiple parameter sets that produce the observed data with the same likelihood — as mentioned, a CCMLE, produced by solving (6), is not necessarily unique. In case (i), while an estimate might be unique, its instability indicates that it is not credible [38], while in case (ii) it is uncertain which of the produced estimates, if any, is the correct one. In order to define uncertainty, we first provide a simple description of the set of all solutions to (6).

Lemma 7. *Let the state-action pair (s,a) be visited at times $0 \leq T_0 < T_1 < \ldots < T_k < T$. Then, $\tilde{P}^T(s,a,s_{T_i+1},T_i)$ are uniquely defined.*

As with nearly all theoretical proofs, the proof of Lemma 7 is in Appendix A. We do provide the proof of the following result within this section, as it provides a method for computation of all CCMLEs.

Theorem 8. *The set of solutions to* (6) *is a polytope.*

Proof. By Lemma 7, the estimates of the transition probabilities $\tilde{P}^T(s_t,a_t,s_{t+1},t)$ obtained from (6) are uniquely defined for all $t \in \{0,\ldots,T-1\}$. Since the other values in $\tilde{P}^T_{t<T}$ do not feature at all in the objective function, the set of solutions to (6) is given by any $\tilde{P}^T_{t<T}$ where $\tilde{P}^T(s_t,a_t,s_{t+1},t)$ are the uniquely defined optimal solutions, and all other $\tilde{P}^T(s,a,s',t)$ satisfy the constraints in (6). All of those constraints are affine, i.e., the set of solutions to (6) is a bounded intersection of finitely many halfspaces. Hence, it is a polytope. □

By Lemma 7 and the proof of Theorem 8, in order to compute the set of all CCMLEs $\mathcal{P}^T$, we first determine a single solution $\tilde{P}^{T*}_{t<T}$ by solving a convex optimization problem. Then, the set $\mathcal{P}^T$ of all other solutions $\tilde{P}^T_{t<T}$ is a polytope in $\mathbb{R}^{|S|^2|A|T}$ given by constraints

$$
\begin{aligned}
\tilde{P}^T(s_t,a_t,s_{t+1},t) = \tilde{P}^{T*}(s_t,a_t,s_{t+1},t) \quad & \text{for all } t < T, \\
\tilde{P}^T(s,a,s',t) \geq 0 \quad & \text{for all } s,s' \in S,\ a \in A,\ t < T, \\
\sum_{s' \in S} \tilde{P}^T(s,a,s',t) = 1 \quad & \text{for all } s \in S,\ a \in A,\ t < T, \\
\tilde{P}^T(s,a,s',t+1) - \tilde{P}(s,a,s',t) \leq \varepsilon_t \quad & \text{for all } s,s' \in S,\ a \in A,\ t < T, \\
\tilde{P}^T(s,a,s',t) - \tilde{P}(s,a,s',t+1) \leq \varepsilon_t \quad & \text{for all } s,s' \in S,\ a \in A,\ t < T.
\end{aligned}
\tag{7}
$$

We are now ready to define the uncertainty in the estimate $\tilde{P}^T(\cdot,\cdot,\cdot,T-1)$.

Definition 9. *For all $s \in S$, $a \in A$, $t \leq T$, let $\mathcal{P}^T_{\sigma,\alpha}(s,a,t) \subseteq \mathbb{R}^{|S|}$ be the polytope where each point is a probability distribution $\tilde{P}^T(s,a,\cdot,t)$ obtained from the set of all CCMLEs based on the agent's previous trajectory $\sigma = (s_0,\ldots,s_T)$ and actions $\alpha = (a_0,\ldots,a_{T-1})$. The* uncertainty of estimates *$\tilde{P}^T(s,a,\cdot,T)$ under (σ,α), denoted by $U_{\sigma,\alpha}(s,a)$, is defined by*

$$U_{\sigma,\alpha}(s,a) = \max\left(\operatorname{diam}\left(\mathcal{P}^T_{\sigma,\alpha}(s,a,T)\right), \max_{s'\in S} \max_{\substack{x\in\mathcal{P}^T_{\sigma,\alpha}(s,a,T)\\ y\in\mathcal{P}^{T+1}_{\overline{\sigma}_{s'},\overline{\alpha}}(s,a,T)}} d(x,y)\right), \tag{8}$$

where $\overline{\sigma}$ and $\overline{\alpha}$ denote a trajectory and set of actions equal to σ and α, with an additional transition (s,a,s') observed at time T and an action a performed at the same time.

In Definition 9 we make use of estimates $\tilde{P}^T(\cdot,\cdot,\cdot,T)$. While (6) only considered $\tilde{P}^T_{t<T}$, we can introduce additional variables $\tilde{P}^T(\cdot,\cdot,\cdot,T)$ which still need to satisfy the constraints of (6). A CCMLE produced by (6) is then certainly not unique, as $\tilde{P}^T(\cdot,\cdot,\cdot,T)$ can be freely chosen, as long as they respect the constraints. This lack of uniqueness is intuitive: it represents the agent's lack of certainty about the current transition probabilities, even if it has all the possible knowledge about transition probabilities at the previous times. We also note that $\overline{\sigma}$ is not necessarily a legitimate path for an agent, as the agent's position s_T at time T in σ does not necessarily equal the starting position for the transition (s,a,s') observed at time T. Nonetheless, a CCMLE can be equally produced using (6), with the objective function

$$-\sum_{t=0}^{T-1} \log \tilde{P}^{T+1}(s_t,a_t,s_{t+1},t) - \log \tilde{P}^{T+1}(s,a,s',T).$$

An intuitive explanation of formula (8) is as follows: the first term in the $\max$ represents the size of the current set of CCMLEs. If two very distant probability transition functions produce the observed path with the same probability, this term will be large. However, even if the first term is small, it is possible that an additional observation will significantly change the estimate. We are more certain in our knowledge of the transition probabilities if a single "outlier" observation cannot significantly change the estimate. We note that $U_{\sigma,\alpha}(s,a) \leq \sqrt{2}$, as all probability distributions $\tilde{P}^T(s,a,\cdot,t)$ necessarily belong to the probability simplex, which has the diameter $\sqrt{2}$.

Remark 10. *As all sets $\mathcal{P}^T_{\sigma,\alpha}(s,a,T)$ are polytopes, u_1 is the maximal distance between the vertices of $\mathcal{P}^T_{\sigma,\alpha}$. On the other hand, u_2 is the maximum of distances between any point of $\mathcal{P}^T_{\sigma,\alpha}(s,a,T)$ and any point in $\cup_{s'\in S}\mathcal{P}^{T+1}_{\overline{\sigma}_{s'},\overline{\alpha}}(s,a,T)$. The maximum distance between points in any two polytopes is, by convexity, achieved when both of those points are vertices of corresponding polytopes. Thus, computing u_2 also reduces to checking all distances between vertices of $\mathcal{P}^T_{\sigma,\alpha}(s,a,T)$ and vertices of $|S|$ polytopes $\mathcal{P}^{T+1}_{\overline{\sigma}_{s'},\overline{\alpha}}(s,a,T)$. Hence, $U_{\sigma,\alpha}(s,a)$ can be computed by determining vertices of $1+|S|$ polytopes and their pairwise distances.*

The notion of uncertainty emulates the role of functions $1/(1+\#(s,a))$ and $1/\sqrt{\#(s,a)}$ in the setting of time-invariant MDPs, where $\#(s,a)$ denotes the number of times a pair (s,a) has been visited until the current time. In [10] and [11], respectively, those functions — multiplied by a tuning parameter β — are used to determine which transition probabilities $P(s,a,\cdot,*)$ are not yet known and should be visited. The intuition behind these functions relies on the law of large numbers: as previously discussed, as $\#(s,a)\to\infty$, the estimate $\#(s,a,s')/\#(s,a)$ converges to $P(s,a,s',*)$ with probability 1, while both of the above functions converge to 0. Theorem 11, proved in Appendix A, shows that U defined in (8) satisfies the same property, and in particular relates U to the function $\beta/(1+\#(s,a))$ used by [10].

Theorem 11. *Assume that $\varepsilon_t = 0$ for all $t \in \mathbb{N}_0$. Let σ be an agent's path until time T, and let α be the actions that the agent takes until time $T-1$. Then, for all $T \geq 0$, $s \in S$, and $a \in A$,*

$$\frac{\sqrt{1-1/|S|}}{1+\#(s,a)} \leq U_{\sigma,\alpha}(s,a) \leq \frac{\sqrt{2}}{1+\#(s,a)},$$

where $\#(s,a)$ denotes the number of times that (s,a) has been visited until time $T-1$.

We note that, from the proof of Theorem 4, it also directly follows that $U_{\sigma,\alpha}(s,a)\to 0$ as $\#(s,a)\to\infty$ for the case of transition probabilities which eventually become time-invariant.

We now proceed to the final step of developing an optimal learning and control policy for an agent in a TVMDP: using the uncertainties of state-action pairs to determine which action to perform.

V. Control Policy Design

In designing the optimal policy, we follow the "exploration bonus" framework used by previous work [10], [11], [14], [31]. In other words, instead of always pursuing actions estimated to be the most useful to satisfying its control objectives, the agent may take into account the value of a particular action to the learning process, with the goal of minimizing the error in the estimated transition probabilities, hence again helping to achieve the control objectives.

A. Optimal Learning Policy

As discussed in Section II, we interpret the agent's objective as minimization of the total uncertainty $\mathbb{U}(t) = \sum_{(s,a)} U_{\sigma,\alpha}(s,a)$ about the environment, where σ and α are the sequences of agent's states and actions, respectively, until the beginning of time step t. That is, we wish to determine a policy π to minimize $\sum_t \mathbb{U}(t)$. Such a sum may be finite, discounted-infinite, or, alternatively, we may only be interested in retaining $\mathbb{U}(t)$ to be as low as possible as $t \to \infty$.

Analogously to the requirement that arises when attempting to construct an optimal policy in an unknown time-invariant MDP, determining an optimal policy in a TVMDP requires knowledge of the time-varying evolution of $\mathbb{U}$. This evolution depends on the transition probabilities in the underlying TVMDP, which we do not know a priori and are changing over time.

Faced with the lack of knowledge about the future uncertainties, we propose the following time-varying policy. At time T, given a CCMLE of current and future transition probabilities, construct a time-invariant policy

$$\psi_T^*(s) = \operatorname*{argmin}_{\psi} \mathbb{E}\left[\sum_{t\geq 0} \mathbb{U}_\psi(T+t)\right], \tag{9}$$

where $\mathbb{U}_\psi$ denotes the uncertainty if the agent follows a time-invariant policy ψ starting at s. Then, $\pi_T = \psi_T^*(s_T)$.

Naturally, there are no guarantees that the proposed policy π will indeed minimize the total uncertainty about the system. Policy π represents a heuristic attempt to always follow whichever action seems to be optimal at reducing the future uncertainty. We note that at every time T it depends on the current CCMLE of transition probabilities, including probabilities $P(\cdot,\cdot,\cdot,T+t)$ for $t \geq 0$. As discussed before, such a CCMLE may not be unique. We also note that we are being non-committal about the horizon length (or possible discounts) for the sum in (9). Its choice depends on the horizon of our learning objective; we reserve further discussion for the subsequent section.

B. Optimal Control Policy

We now amend the above discussion about optimal active learning by considering an agent that desires to maximize the collected state-action-based rewards. This framework is the setting of [6], [10], [31], which deal with the possibly conflicting exploration and exploitation objectives — learning to improve the accuracy of the estimated transition probabilities (and thus lead to better planning in the future) and attempting to collect rewards using the current estimates — by adding a *learning bonus* to the agent's collected reward. In other words, instead of using the policy

$$\pi_T^* = \operatorname*{argmax}_{\pi} \mathbb{E}\left[\sum_{t\geq 0} R(s_{T+t}, \pi_{T+t})\right],$$

the agent determines a policy

$$\pi_T^* = \operatorname*{argmax}_{\pi} \mathbb{E}\left[\sum_{t\geq 0} R(s_{T+t}, \pi_{T+t}) + f(s_{T+t}, \pi_{T+t}, T+t)\right],$$

where $f(t)$ relates to the "amount of information" that the agent will collect by visiting (s_t, π_t) at time t. As mentioned, [10] defines $f(s,a,t) = \beta/(\#(s,a)+1)$, where $\#(s,a)$ denotes the number of times $\#(s,a)$ has been visited before time t. The results in [10] show that, for time-invariant MDPs, such a bonus will, with high probability, lead to eventual learning of an almost-optimal policy (in the Bayesian sense, as defined in [10]). Our framework does not allow for such a result, as there is a constant need for learning due to the change in transition probabilities. Nonetheless, we adapt the approach of [6], [10], [31] and define an optimal control policy to be a policy

$$\pi_T^* = \operatorname*{argmax}_{\pi} \mathbb{E}\left[\sum_{t\geq 0} R(s_{T+t}, \pi_{T+t}) + \beta U_{T+t}(s_{T+t}, \pi_{T+t})\right], \tag{10}$$

where $\beta \geq 0$ and U_t denotes the uncertainty of the agent about the estimates of transition probabilities associated with the particular state-action pair, given the agent's motion and actions until time t. As shown in Theorem 11, in the case of time-invariant MDPs policy, the bonus defined in (10) is indeed similar to the form of the bonus in [10]. We again note that the expectation in (10) is computed using the CCMLE computed at time t. Thus, as in the previous section, the agent needs to recompute and reapply the proposed policy during the system run in order to make use of its learning.

Naturally, policy proposed in (10) depends on the length of the horizon that is considered; [10] provides a theoretical discussion of an optimal horizon length for the policy used in that work. As computing predictions of the future uncertainties is computationally difficult, solving (10) comes with a heavy computational burden for long horizons. Predictions of future uncertainties also become increasingly unreliable with the horizon length, so $\mathbb{E}[U_t(s_t, \pi_t)]$ may not be meaningful for large t. For this reason, in the work of Section VI and Appendix A we concentrate on (i) computing policy (10) with horizon 1 and (ii) computing policy (10) with $\beta = 0$, the latter of which does not promote active learning, but nonetheless enables the agent to improve its estimates of transition probabilities by observing new transitions.

VI. Simulations

In this section, we illustrate the proposed CCMLE method on two numerical examples. The first example is within the classical gridworld-based patrolling (or pickup-delivery) domain [39]–[41], where probabilities of the agent moving in a particular direction when using a given action does not depend on the system state. For the purposes of our investigation, we consider the case where those probabilities change over time, and investigate the agent's success at learning the transition probabilities and performing its task. The second example is that of a two-state MDP with periodically changing transition probabilities. In addition to learning transition probabilities, we consider a reward maximization objective. Such a problem can be interpreted as a semi-Markov two-armed bandit problem [42], [43]: a two-armed bandit where the arms take different lengths of time to pull. In our setting, the reward on one of the arms periodically varies over time. For more involved examples, we avert the reader's attention to Appendix B, where we discuss more realistic scenarios, with complicated rules governing environmental changes, and poor choices of bound ε.

A. Patrol with One-Time Change of Dynamics

The scenario we are simulating is as follows: an agent is moving on an $n \times n$ grid, starting in one of the grid corners. A 5×5 illustration is shown in Figure 1. At every time step, if the agent is at a non-edge tile in the grid, it will move north, east, south, west, or stay in place. At the grid edges ("walls"), the agent "bounces back", i.e., moves to the nearest non-edge tile in the next step.

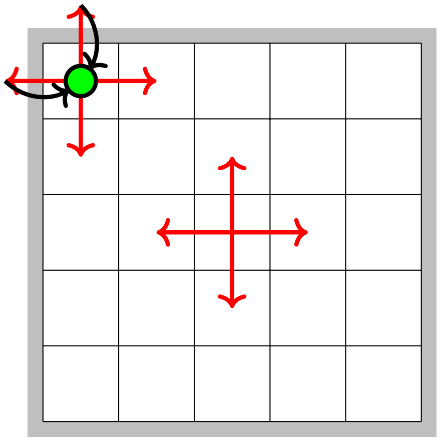

Fig. 1. An illustration of the grid world in the simulated scenario. The walls are denoted in gray. The possible motions of an agent during one time step are denoted in red; an agent can also remain in place during a time step. If the agent's action results in the agent moving into a wall, the agent automatically returns to its last non-wall position in the subsequent time step (denoted in black).

The transition probabilities at every non-edge tile in the grid are known to the agent to be the same, and the agent can use one of 5 actions at every point in time: $A = \{1, 2, 3, 4, 5\}$. At the beginning of the system run, action 1 results in the agent moving north (if it is at a non-edge tile), 2 in moving east, 3 in moving south, 4 in moving west, and 5 in remaining in place. However, during the system run, actions 1 and 3 are slowly switching their outcomes, and so are actions 2 and 4. More precisely, at time t, action 1 (2, 3, 4, respectively) will result in the agent moving north (east, south, west, respectively) with probability $1 - t/100$ for $t \leq 100$ and 0 for $t > 100$ and in the agent moving south (west, north, east, respectively) with probability $t/100$ for $t \leq 100$ and 1 for $t > 100$.

We consider two settings: in the first one, the agent solely seeks to learn the transition probabilities, while in the second one, it seeks to satisfy a patrolling objective.

1) Learning: In this setting, the agent's sole goal is to learn the transition probabilities: its control action is always the action that has been least used so far in the system run. We compare two ways of agent's learning: (i) by performing classical estimation — assuming that the transition probabilities are time-invariant, counting outcomes, and dividing by the number of times that an action was taken — and (ii) by using the knowledge that the change in transition probabilities between consecutive time steps is no larger than 0.01 and obtaining the CCMLE. While method (i) will lead the agent to converge to the correct transition probabilities (as they are time-invariant after $t = 100$), such convergence is only asymptotic. On the other hand, the CCMLE method in (ii) takes into account the possible change in transition probabilities, and implicitly quickly rejects those samples that were collected much earlier during the system run. The average error $\sum_{s,a,s'} |\tilde{P}(s,a,s',t-1) - P(s,a,s',t-1)|/(|S|^2|A|)$ in estimated transition probabilities at time t is given in Figure 2. We note that all the results in this section correspond to the 5×5 grid. However, since the probabilities of moving in a particular direction in the grid do not depend on the state, the results for grids of all sizes are largely the same.

The CCMLE method produces significantly better results than classical estimation. Near the start of the system run, the importance of changes over time is small, and the two methods perform similarly. As the transition probabilities continue

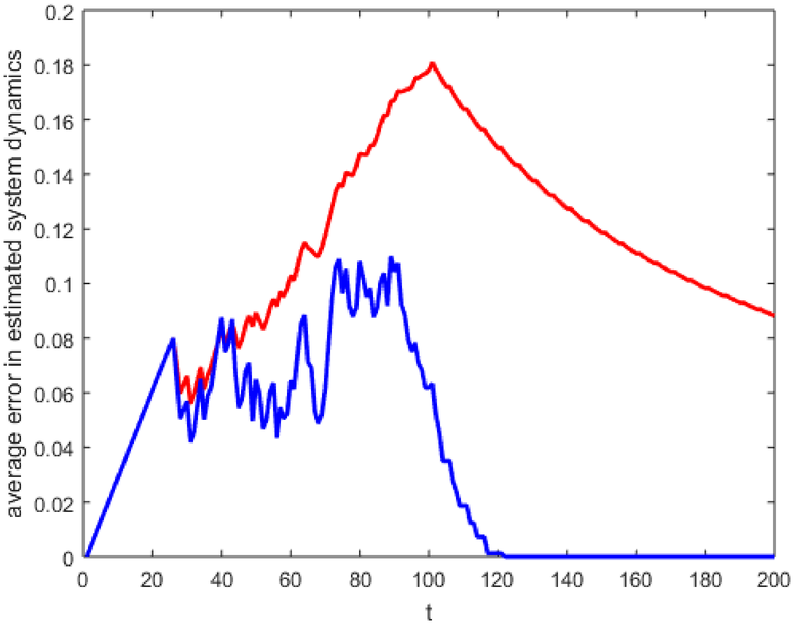

Fig. 2. Average error in the transition probability estimates. The red curve indicates the error with classical estimation that assumes time-invariant transition probabilities. The blue curve indicates the error with the CCMLE method.

changing, classical estimation is unable to adapt, and the error in the estimated transition probabilities continues growing. While the CCMLE method is also unable to provide entirely correct estimates, its estimates are generally better than the classical one, and, at around $t = 90$ — even before the transition probabilities stop changing — the error begins to quickly decrease. While Theorem 5 only guarantees that the CCMLE method will produce correct estimates of transition probabilities at time $t = 200$, this result already occurs $t = 120$, i.e., only 20 time steps after the transition probabilities cease changing. On the other hand, classical estimation continues having a comparatively large (albeit diminishing) average error.

We note that the *average* error for both methods is no larger than 0.2 at any point in time, which may not seem excessive. However, such a measure is somewhat deceptive, as there may be one or two transition probabilities with a significantly larger error. Fig. 3 illustrates the maximal error $\max_{s,a,s'} |\tilde{P}(s,a,s',t-1) - P(s,a,s',t-1)|$ in the transition probabilities at time t. The maximal error for both methods goes up to 0.7. However, while the two methods are roughly similar in terms of maximal error until $t = 90$, once the transition probabilities become time-invariant, the error using the CCMLE quickly decreases to 0, while the maximal error with classical estimation remains large. The effect of such an error will be explored in the setting with a control objective.

Finally, we consider a situation that is not covered by the theoretical work of previous sections: the transition probabilities changing more rapidly than the bound ε_t. In particular, we retain the above scenario, and the agent still counts on the transition probabilities changing by no more than 0.01 between two consecutive steps. However, the transition probabilities proceed to change by 1 in a single step: instead of shifting in slow increments as before, the switch in outcomes between actions 1 and 3, as well as actions 2 and 4, happens instantaneously at time $t = 20$. Fig. 4 illustrates the average error $\sum_{s,a,s'} |\tilde{P}(s,a,s',t-1) - P(s,a,s',t-1)|/(|S|^2|A|)$ in the estimated transition probabilities at time t using classical estimation, unconscious of possible changes in transition probabilities, and the CCMLE.

While the error in the estimates was naturally high immediately after the switch for both methods, the CCMLE method identifies the correct transition probabilities significantly more quickly, and in fact obtains an entirely correct model around 40 time steps after the sudden, extreme change. The estimated transition probabilities remain correct at all times afterwards. On the other hand, while classical estimation converges towards the correct transition probabilities, it does so slowly and asymptotically, with the rate of convergence as discussed under Remark 6. For instance, 50 time steps after the switch, the error is still greater than the error of the CCMLE at 25 time steps after the switch, and the error of the classical estimation will never reach 0.

2) Planning: In this setting, the agent seeks to satisfy the following control objective: reach the eastern wall of the grid, then reach the southern wall, then the western wall, then the northern wall (in this order), and repeat the process indefinitely. Such an objective is a patrolling task in the sense of [41], and is a version of the pickup-delivery objective from [40], where there are multiple pickup and delivery points.

If we encode the described control objective into a reward function, the rewards that the agent obtains are time-varying, i.e., depend on the agent's previous path. While such a framework technically differs from the setting of the previous sections, the

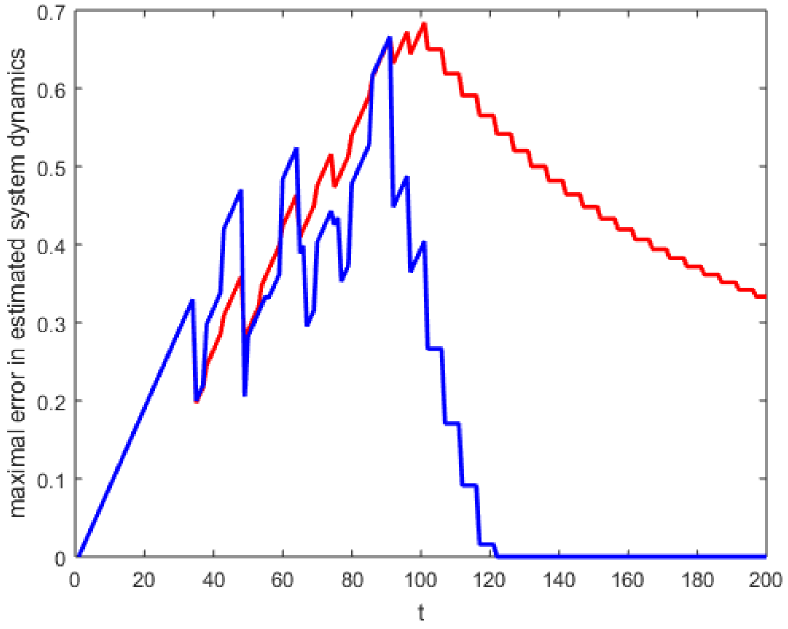

Fig. 3. Maximal error in the transition probabilities estimates. The red curve indicates the error with classical estimation that assumes time-invariant transition probabilities. The blue curve indicates the error with the CCMLE method.

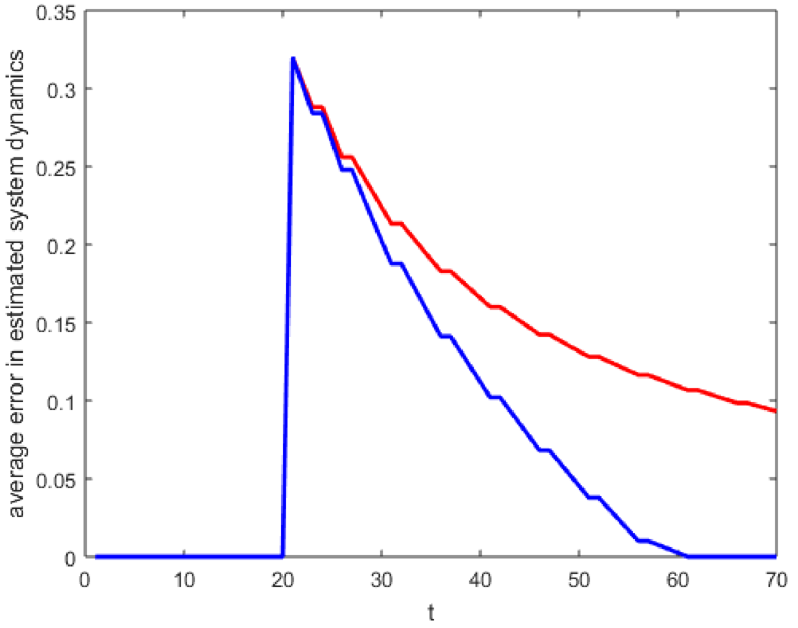

Fig. 4. Average error in the transition probability estimates, in the scenario where the change occurs more rapidly than expected. The red curve indicates the error with classical estimation that assumes time-invariant transition probabilities. The blue curve indicates the error with the CCMLE method.

agent is still able to compute the optimal policy (10) at every time, given the rewards at that time, and then recompute it once the rewards change. In our framework, the agent uses the policy (10) with $\beta = 0$ and with a horizon long enough to ensure that it has the incentive to visit the desired wall as soon as possible.

Given that we set $\beta = 0$, the agent seeks to achieve its objective without any regard to conscious exploration: it always takes an action that, by its current estimate, should take it in the desired direction with the highest probability. Fig. 5 illustrates how often the agent is able to reach its objective wall. The two methods enjoy a comparable success rate at the beginning. However, the agent that learned using the CCMLE is able to adapt to the changes in transition probabilities much more quickly than the agent using a classical estimation method.

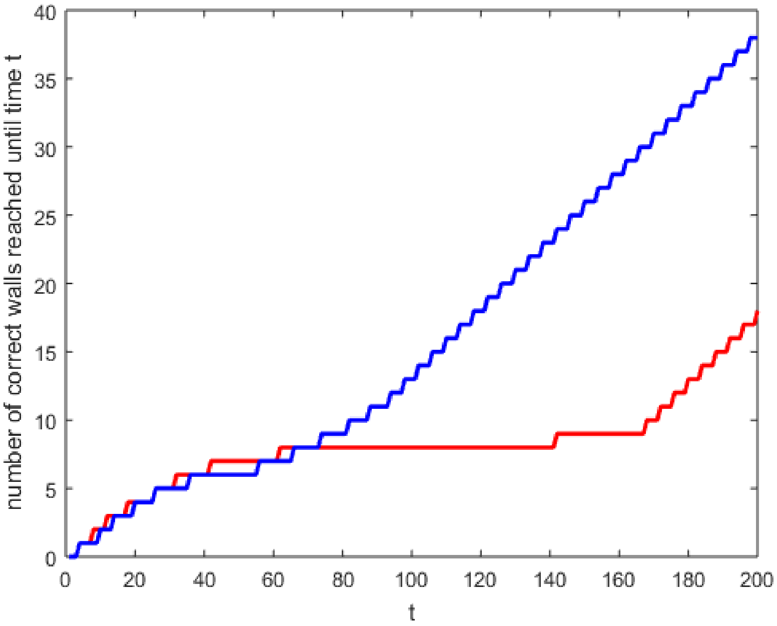

Fig. 5. Satisfaction of control objectives. The red curve indicates the number of times the agent reached a correct wall until time t when using the plan that assumes time-invariant transition probabilities, and the blue curve indicates the number of times the agent reached a correct wall until time t when using the CCMLE-based plan conscious of changes in dynamics.

While the above simulation, along with other simulations in this section and Appendix B, does not present the effect of different initial states on the CCMLE and subsequent learning and control policies, we remark that such an effect is not difficult to deduce — at least on an informal level — given the time-varying nature of dynamics. Namely, for a state space where an agent may not visit a particular state-action pair for a substantial number of steps, its estimate of states that have not been recently visited will necessarily be incorrect using *any* estimation method, as the dynamics will have changed since the last visit. Thus, two agents with different initial states and different trajectories will necessarily have different estimates for those parts of the state space that they last visited at substantially different times. On the other hand, for a "small" state space where an agent frequently visits each state-action pair regardless of its starting state, the estimates will be similar for all initial states. We thus omit in-depth discussions of different initial states from the remainder of the section.

B. Periodically Changing Transition Probabilities

The scenario we simulate in this section is that of a 2-state TVMDP illustrated in Fig. 6. As in Section VI-A, we first consider solely estimation, i.e., learning, and then joint learning and planning. Since the TVMDP given in Fig. 6 only contains a single action, for the discussion of planning we will append an additional action.

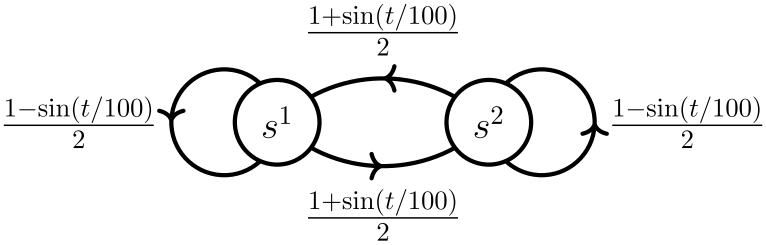
$\frac{1+\sin(t/100)}{2}$
$\frac{1-\sin(t/100)}{2}$
s^1
s^2
$\frac{1-\sin(t/100)}{2}$
$\frac{1+\sin(t/100)}{2}$

Fig. 6. The scenario of periodically changing transition probabilities. The time-varying transition probabilities, a priori unknown to the agent, are indicated next to the arrows indicating transitions.

1) Learning: As in the previous section, we compare classical estimates, produced by assuming that the transition probabilities are time-invariant, to the CCMLE. In particular, we assume that it is a priori known that the transition probabilities change by no more than

$$\varepsilon = \frac{1}{2}\left(1 - \cos\frac{1}{100} + \sin\frac{1}{100}\right). \tag{11}$$

However, the agent does not know the exact change in the transition probabilities, nor is it aware that the transition probabilities are periodic.

Fig. 7 shows the estimate of the probability of a *switch*: a transition that moves the agent from s^1 to s^2 or vice versa. Classical estimation that assumes time-invariant transition probabilities obviously produces estimates that converge towards the mean of the time-varying transition probability. Thus, the obtained estimates are increasingly less accurate as time progresses. On the other hand, the CCMLE tracks the true transition probability with remarkable accuracy.

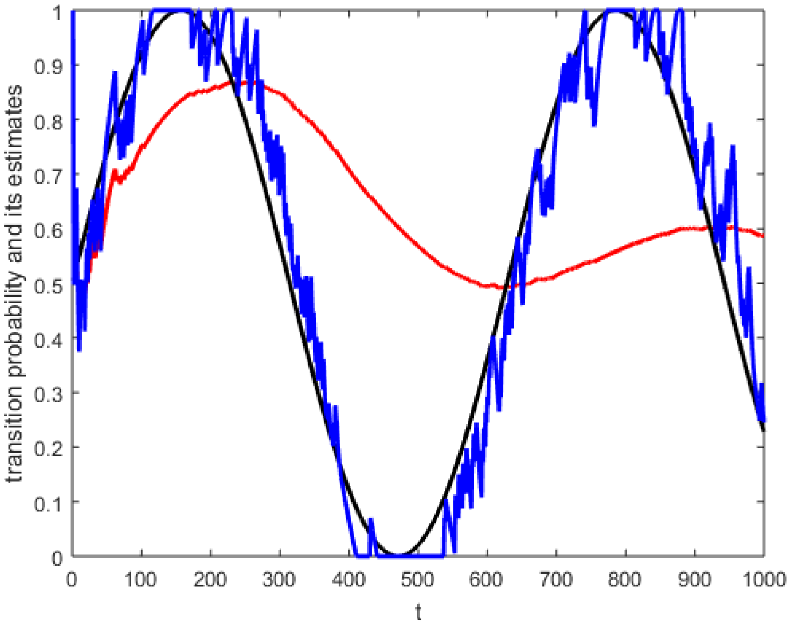

Fig. 7. Estimates of the transition probability which results in a switch. The black curve indicates the true probability. The red curve indicates the estimates with classical estimation that assumes time-invariant transition probabilities. The blue curve indicates the estimates with the CCMLE method.

As mentioned in Section III, classical learning can be heuristically made aware of the changes in the transition probabilities by making the estimation forgetful, i.e., counting only the samples obtained within finitely many previous steps (“sliding window”). On the other hand, making the CCMLE method — which is already explicitly aware of the bound on the changes of transition probabilities — forgetful can be useful to reduce the complexity of the relevant optimization problems. The results of Section III indicate that $1/\varepsilon$ may be an appropriate amount of memory. Fig. 8 gives the comparison between the two forgetful learning methods. The classical estimation method, now made forgetful, identifies the transition probabilities with a delay — by the time it collects enough samples about a particular transition probability, the probability has changed. On the other hand, introduction of forgetfulness into the CCMLE does not result in a significant impact on its quality. The CCMLE, with or without forgetfulness, still largely outperforms the classical estimation, even when the classical method is improved by introducing forgetfulness.

2) Planning: In order to discuss the optimal control policy, we slightly modify the TVMDP introduced in Fig. 6, as shown in Fig. 9. The setting in Fig. 9 introduces a single deterministic action ($black$) available at state s^2. We define the reward for such an action by $R(s^2, black) = 3$. State s^1 admits two available actions: action $blue$ is deterministic and we define $R(s^1, blue) = 1$, while action red may end up in two outcomes, as shown in Fig. 9, and its reward is given by $R(s^1, red) = 0$. The agent's starting position is s^1.

We note that the described setting corresponds to the two-armed bandit problem in the sense of [42], [43]: action $blue$ represents one of the bandit arms, which produces a reward of 1 and remains active for 1 time step, while action red (and potential subsequent action $black$) represents the other bandit arm, which produces either a reward of 0 or 3, and remains active for 1 or 2 time steps, respectively. However, in keeping with the narrative of the remainder of the paper, we continue using the notation from the MDP setting throughout this section. We consider a multi-armed bandit in Appendix B-B, using the standard notions of bandit problems.

Throughout this example, we assume that the agent is a priori aware that the actions $blue$ and $black$ are deterministic. Hence, its uncertainty is only about action red; again, the agent knows that the rate of change does not exceed ε given in (11). The problem of computing the agent's uncertainty, as given in Definition 9, can be made computationally more simple owing to the fact that action red only has two possible outcomes: agent's total uncertainty and the uncertainty in the estimated probability of a switch are scalar multiples, the former being larger by a factor of $\sqrt{2}$.

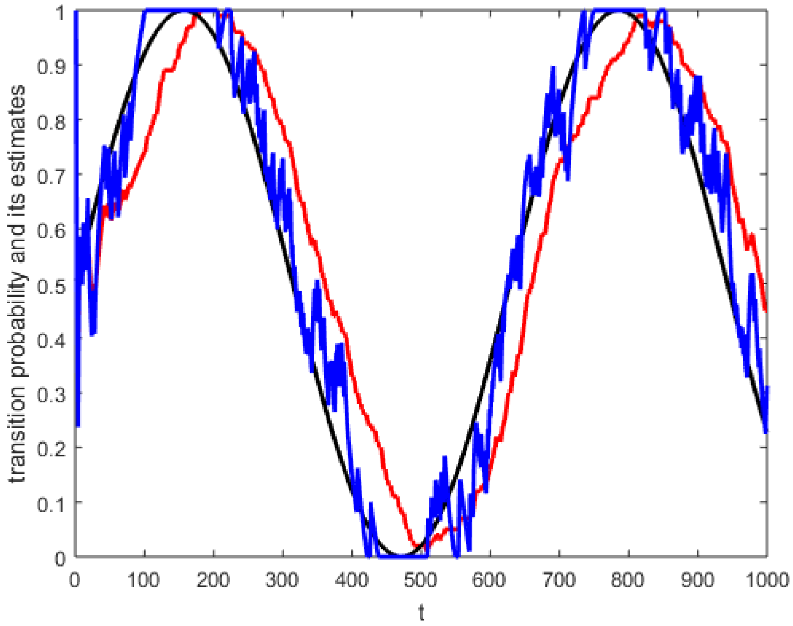

Fig. 8. Estimates of the transition probability which results in a switch, using forgetful learning methods. The black curve indicates the true probability. The red curve indicates the estimates with forgetful classical estimation that assumes time-invariant transition probabilities within the time period in its memory. The blue curve indicates the estimates with the forgetful CCMLE.

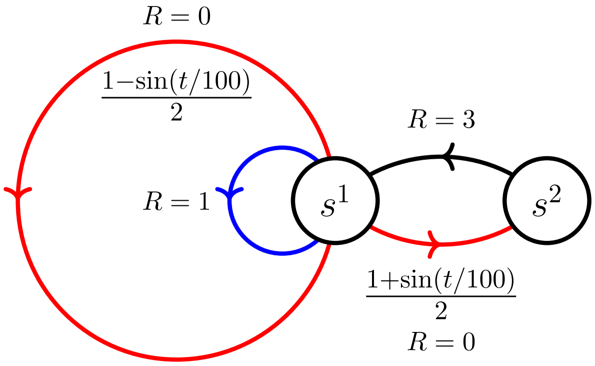

Fig. 9. The scenario of periodically changing transition probabilities, with control actions. The transition probabilities when using the black and blue actions are 1; those actions are deterministic. The transition probabilities when using the red action are as indicated.

In order to maximize its average collected reward, the agent applies the optimal policy given in (10), with the horizon length equal to 1, and recomputes and reapplies it at every time step. In other words, at every time the agent cares only about the results of its current and next step. We compare the agent's average reward between three cases: (a) when the agent bases its decision on the estimates obtained by classical estimation, (b) when the agent uses CCMLE, but does not use active learning, and (c) when the agent uses CCMLE with active learning, i.e., $\beta > 0$.

An agent that uses learning based on classical estimation, with or without a learning bonus as used in [10], [11], would choose action red when its estimate of transitioning to state s^2 with the red action is greater than $2/3$, or it receives a sufficient bonus, and choose action $blue$ otherwise. Note that such an agent will eventually always choose action $blue$: learning bonuses will eventually converge to 0, and the estimate of the probability of a switch for an agent that uses classical learning eventually converges to $1/2 < 2/3$. Thus, the average reward of an agent that assumes that probabilities are not changing will converge to $1 = R(s^1, blue)$.

An agent that uses the CCMLE without active learning, i.e., applies (10) with $\beta = 0$, would essentially fall into the same trap. Once its estimate of the probability of a switch falls under $2/3$ once — which is likely to happen, due to the oscillating nature of the transition probabilities and the good quality of estimation exhibited by an MLE-based learner in previous examples — the agent will again cease to learn, and will use solely action $blue$. Thus, its average reward will converge to 1.

However, if $\beta > 0$, the agent may choose action red in order to reduce its uncertainty. Thus, even after its estimate of the probability of a switch falls under $2/3$, it may still occasionally perform action red, thus allowing itself to observe the changed transition probabilities and restart collecting higher rewards in the periods when the transition probability is higher than $2/3$. An example of such behavior is exhibited in Fig. 10, for $\beta = 3/\sqrt{2}$.

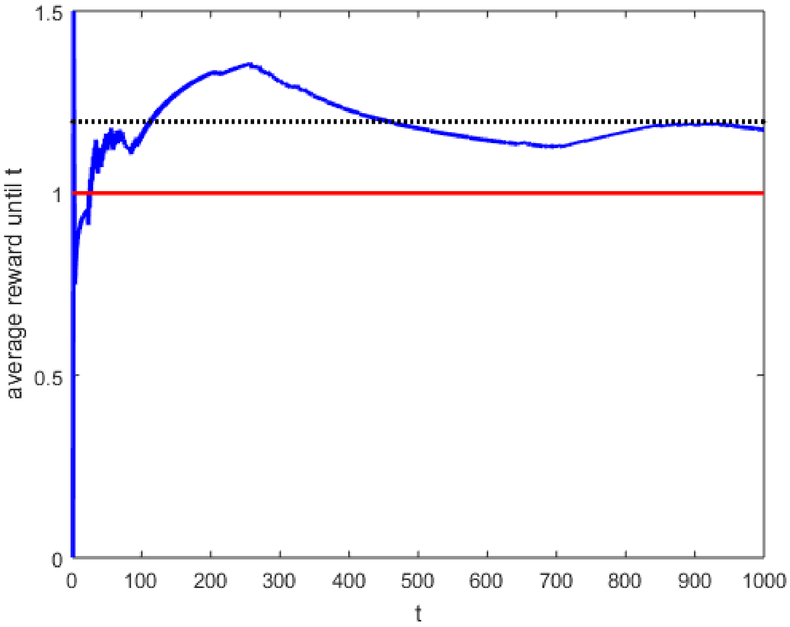

Fig. 10. The blue line indicates the average reward obtained by an agent who uses the CCMLE with an uncertainty-based learning bonus. The dotted line denotes indicates the asymptotic average reward obtained by an agent who has perfect a priori knowledge of the transition probabilities. The red line indicates the asymptotic average reward obtained by an agent that uses classic estimation unconscious of the change in the transition probabilities and the CCMLE without active learning.

As Fig. 10 shows, the average reward obtained by an agent who uses policy 10 with $\beta = 3$ converges to around 1.15. This is a significant improvement over the average reward of 1: an agent with perfect knowledge who chooses to use action red whenever the transition probability of a switch is greater than $2/3$ and action $blue$ otherwise will obtain an asymptotic average reward of around 1.2. Even so, slightly higher gains may be obtained by a different choice of β.

VII. Conclusions

The work in this paper concentrated on presenting an integrative method for estimation, learning, and planning of an agent operating in an unknown TVMDP. The proposed method is founded on introducing three notions:

- *change-conscious maximal likelihood estimation (CCMLE)*, which exploits the knowledge on the maximal possible rate of change of transition probabilities to produce time-varying estimates based on the observed outcomes of agent's actions;
- *uncertainty of an estimate*, which quantifies the lack of knowledge about transition probabilities at a particular time during the system run; and
- *optimal control policy with an uncertainty-based learning bonus*, which aims to enable the agent to actively learn about transition probabilities in order to increase its long-term attained reward.

As shown in Proposition 3, when used in a time-invariant MDP, the CCMLE produces the same estimates as the classical method based on the frequency of all previously observed outcomes. On the other hand, as indicated by the theoretical results of Section III and validated on the numerical examples in Section VI and Appendix B, in a time-varying setting the CCMLE produces significantly better estimates of transition probabilities than the method based on the frequency of all previously observed outcomes, which implicitly assumes that the transition probabilities are time-invariant. Similarly, the notion of uncertainty introduced in Section IV reduces to previous methods for describing the lack of knowledge about transition probabilities in time-invariant MDPs, while providing a novel measure of the lack of knowledge about transition probabilities in the time-varying setting. As proposed in Section V, such a measure can then be used to design a learning bonus for a learning and control policy of an agent operating in a TVMDP, once again generalizing the active learning and control policies previously introduced for time-invariant MDPs. Numerical examples of Section VI and Appendix B show that the proposed policies enable an agent to successfully learn the transition probabilities and achieve its control objective, both in comparison with classical method lead to successful learning, and to theoretical maxima achievable by an agent with a priori knowledge of transition probabilities.

The work presented in this paper presents an initial discussion of optimal estimation and learning methods for time-varying stochastic control processes. While theoretical results and numerical examples encourage future exploration of the CCMLE and

CCMLE-based learning and planning, these results are by no means exhaustive. Namely, the behavior of the CCMLE in the case of changing transition probabilities has been theoretically explored only in the case when one of the transition probabilities eventually equals 1. A natural next step would be to obtain theoretical bounds for estimation error and convergence to correct transition probabilities in the case when all transition probabilities are in $[0, 1)$. On the side of active learning, Theorem 11 relates the measure of uncertainty introduced in this paper to exploration bonuses used in time-invariant MDPs. However, there are currently no formal guarantees — parallel to the work that uses classical exploration bonuses — on the convergence to optimal control policy when using an uncertainty-based learning bonus, either in a time-varying or time-invariant setting. The role of bonus multiplier β remains to be discussed — while for $\beta = 0$ the agent does not actively learn, and for $\beta \to \infty$ the agent solely learns and does not have any incentive to maximize its collective reward, it is currently unknown how β should be chosen to obtain an optimal policy.

Finally, we note that while the current paper presents the CCMLE solely in the framework of TVMDPs. The same method, however, may be easily adapted to the framework of general online learning, where the problem is to estimate a time-varying parameter from a time series of observations, where the parameter is known to change with a rate no greater than some a priori known bound. A similar problem has been considered in [44]; however, instead of the CCMLE approach of attempting to find the most likely time-varying parameter satisfying the constraint on the maximal rate of change, [44] attempts to find the time-varying parameter that offers the least regret against a comparison sequence that satisfies a similar constraint. The CCMLE problem for general online learning, with all the questions resolved and opened in this paper, thus remains wide open.

ACKNOWLEDGMENTS

This research was partly performed while the first author was a research associate at the Institute for Computational Engineering and Sciences, University of Texas at Austin. This work was partly supported by an Early Stage Innovations grant from NASA's Space Technology Research Grants Program (grant no. 80NSSC19K0209), Sandia National Laboratories grant no. 2100656, National Science Foundation grants no. 1646522 and no. 1652113, and Defense Advanced Research Projects Agency grant no. W911NF-16-1-0001.

REFERENCES

[1] A. Elfes, "Sonar-based real-world mapping and navigation," in *Autonomous Robot Vehicles*, I. J. Cox and G. T. Wilfong, Eds., 1990, pp. 233–249.

[2] H. Gao, X. Song, L. Ding, K. Xia, N. Li, and Z. Deng, "Adaptive motion control of wheeled mobile robot with unknown slippage," *International Journal of Control*, vol. 87, no. 8, pp. 1513–1522, 2014.

[3] J. D. Hernández, E. Vidal, G. Vallicrosa, E. Galceran, and M. Carreras, "Online path planning for autonomous underwater vehicles in unknown environments," in *IEEE International Conference on Robotics and Automation*, 2015, pp. 1152–1157.

[4] J. Forbes, T. Huang, K. Kanazawa, and S. Russell, "The BATmobile: Towards a Bayesian automated taxi," in *14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1878–1885.

[5] J. C. Santamaría, R. S. Sutton, and A. Ram, "Experiments with reinforcement learning in problems with continuous state and action spaces," *Adaptive behavior*, vol. 6, no. 2, pp. 163–217, 1997.

[6] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Machine Learning*, vol. 49, pp. 209–232, 2002.

[7] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[9] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial Intelligence*, vol. 112, pp. 181–211, 1999.

[10] J. Z. Kolter and A. Y. Ng, "Near-Bayesian exploration in polynomial time," in *26th International Conference on Machine Learning*, 2009, pp. 513–520.

[11] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for Markov decision processes," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008.

[12] J. Fu and U. Topcu, "Probably approximately correct MDP learning and control with temporal logic constraints," in *Robotics: Science and Systems*, 2014.

[13] M. J. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *AAAI Fall Symposia*, 2015.

[14] M. Ornik, J. Fu, N. T. Lauffer, W. K. Perera, M. Alshiekh, M. Ono, and U. Topcu, "Expedited learning in MDPs with side information," in *57th IEEE Conference on Decision and Control*, 2018, pp. 1941–1948.

[15] A. R. Vasavada, S. Piqueux, K. W. Lewis, M. T. Lemmon, and M. D. Smith, "Thermophysical properties along Curiosity's traverse in Gale crater, Mars, derived from the REMS ground temperature sensor," *Icarus*, vol. 284, pp. 372–386, 2017.

[16] J. R. Zimbelman, "Introduction to the Planetary Dunes special issue, and the aeolian career of Ronald Greeley," *Icarus*, vol. 230, pp. 1–4, 2014.

[17] L. Liu and G. S. Sukhatme, "A solution to time-varying Markov decision processes," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1631–1638, 2018.

[18] M. Vergassola, E. Villermaux, and B. I. Shraiman, "'Infotaxis' as a strategy for searching without gradients," *Nature*, vol. 445, pp. 406–409, 2007.

[19] A. L. Strehl, L. Li, and M. L. Littman, "Reinforcement learning in finite MDPs: PAC analysis," *Journal of Machine Learning Research*, vol. 10, pp. 2413–2444, 2009.

[20] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," in *Neural Information Processing Systems*, 2010, pp. 586–594.

[21] L. K. Fenton, A. D. Toigo, and M. I. Richardson, "Aeolian processes in Proctor Crater on Mars: Mesoscale modeling of dune-forming winds," *Journal of Geophysical Research: Planets*, vol. 110, no. E6, 2005.

[22] J. A. Boyan and M. L. Littman, "Exact solutions to time-dependent MDPs," in *Neural Information Processing Systems*, 2001, pp. 1026–1032.

[23] A. J. Filardo, "Business-cycle phases and their transitional dynamics," *Journal of Business & Economic Statistics*, vol. 12, no. 3, pp. 299–308, 1994.

[24] S. M. Ross, *Applied Probability Models with Optimization Applications*. Holden-Day, 1970.

[25] H. L. S. Younes and R. G. Simmons, "Solving generalized semi-Markov decision processes using continuous phase-type distributions," in *AAAI-04: 19th National Conference on Artificial Intelligence*, 2004, pp. 742–747.

[26] Z. Kalmár, C. Szepesvári, and A. Lőrincz, "Module-based reinforcement learning: Experiments with a real robot," *Machine Learning*, vol. 31, no. 1–3, pp. 55–85, 1998.

[27] I. Szita, B. Takács, and A. Lörincz, "ε-MDPs: Learning in varying environments," *Journal of Machine Learning Research*, vol. 3, pp. 145–174, 2002.

[28] B. C. Csáji and L. Monostori, "Value function based reinforcement learning in changing Markovian environments," *Journal of Machine Learning Research*, vol. 9, pp. 1679–1709, 2008.

[29] F. X. Diebold, J.-H. Lee, and G. C. Weinbach, "Regime switching with time-varying transition probabilities," in *Non-Stationary Time Series Analysis and Cointegration*, C. P. Hargreaves, Ed., 1994, pp. 283–302.

[30] H. van Hasselt, "Reinforcement learning in continuous state and action spaces," in *Reinforcement Learning: State-of-the-Art*, M. Wiering and M. van Otterlo, Eds., 2012, pp. 207–251.

[31] R. I. Brafman and M. Tennenholtz, "R-MAX — a general polynomial time algorithm for near-optimal reinforcement learning," *Journal of Machine Learning Research*, vol. 3, pp. 213–231, 2002.

[32] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2005.

[33] N. Balakrishnan and V. B. Nevzorov, *A Primer on Statistical Distributions*. Wiley, 2004.

[34] T. Gregorius and G. Blewitt, "The effects of weather fronts on GPS measurements," *GPS World*, vol. 9, pp. 52–60, 1998.

[35] A. Feintuch, "Stabilization and sensitivity for eventually time-invariant systems," *Linear Algebra and its Applications*, vol. 122–124, pp. 105–114, 1989.

[36] D. Ryabko and M. Hutter, "On the possibility of learning in reactive environments with arbitrary dependence," *Theoretical Computer Science*, vol. 405, no. 3, pp. 274–284, 2008.

[37] P. Wagner, "Optimistic policy iteration and natural actor-critic: A unifying view and a non-optimality result," in *Neural Information Processing Systems*, 2013, pp. 1592–1600.

[38] E. Andries, "Statistical modelling strategies for reliability data on physical components with possibly multiple causes of failure," Ph.D. dissertation, Limburgs Universitair Centrum, 2004.

[39] H. Santana, G. Ramalho, V. Corruble, and B. Ratitch, "Multi-agent patrolling with reinforcement learning," in *3rd International Joint Conference on Autonomous Agents and Multiagent Systems*, 2004, pp. 1122–1129.

[40] Y. Chen, J. Tůmová, and C. Belta, "LTL robot motion control based on automata learning of environmental dynamics," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 5177–5182.

[41] R. Toro Icarte, T. Q. Klassen, R. Valenzano, and S. A. McIlraith, "Using reward machines for high-level task specification and decomposition in reinforcement learning," in *35th International Conference on Machine Learning*, 2018, pp. 2107–2116.

[42] J. N. Tsitsiklis, "A short proof of the Gittins index theorem," *Annals of Applied Probability*, vol. 4, no. 1, pp. 194–199, 1994.

[43] M. O. Duff and A. G. Barto, "Local bandit approximation for optimal learning problems," in *Neural Information Processing Systems*, 1997, pp. 1019–1025.

[44] J. Yuan and A. Lamperski, "Trading-off static and dynamic regret in online least-squares and beyond," *arXiv:1909.03118 [cs.LG]*, 2019, preprint.

[45] D. Zelterman, *Models for Discrete Data*. Oxford University Press, 2006.

[46] R. T. Rockafellar and R. J. Wets, *Variational Analysis*. Springer, 1998.

[47] A. Beck, *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. Society for Industrial and Applied Mathematics, 2004.

[48] P. S. Bullen, *Handbook of Means and Their Inequalities*. Springer, 2003.

[49] A. Hickman, "History of pilot ballooning," *Weather*, vol. 70, pp. 521–523, 2015.

[50] W. H. Al-Sabban, L. F. Gonzalez, and R. N. Smith, "Wind-energy based path planning for unmanned aerial vehicles using Markov decision processes," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 784–789.

Appendix A
Proofs of Theoretical Results

Proof of Proposition 3. To simplify notation, assume that $|A| = 1$, i.e., $A = \{a\}$. When $\varepsilon_t = 0$, (6) devolves into

$$\begin{array}{lll} \min\limits_{\tilde{P}^T(\cdot,a,\cdot,*)} & -\sum\limits_{t=0}^{T-1} \log \tilde{P}^T(s_t, a, s_{t+1}, *) & \\ \text{s.t.} & \tilde{P}^T(s, a, s', *) \geq 0 & \text{for all } s, s' \in S, \\ & \sum\limits_{s' \in S} \tilde{P}^T(s, a, s', *) = 1 & \text{for all } s \in S. \end{array} \tag{12}$$

As discrete distributions $\tilde{P}(s, a, \cdot, *)$ for different $s \in S$ never appear together in the constraints of (12), this problem can be separated into $|S|$ problems

$$\begin{array}{lll} \min\limits_{\tilde{P}^T(s,a,\cdot,*))} & -\sum\limits_{i=0}^{\tau_s} \log \tilde{P}^T(s, a, s_{\nu_{i;s}}, *) & \\ \text{s.t.} & \tilde{P}^T(s, a, s', *) \geq 0 & \text{for all } s' \in S, \\ & \sum\limits_{s' \in S} \tilde{P}^T(s, a, s', *) = 1, & \end{array} \tag{13}$$

where $\nu_{i;s}$ designates the time at which state s has been visited i-th time (out of τ_s). Problem (13) is a standard maximum likelihood estimation problem for a multinomial distribution, and it can be easily shown (see, e.g., [45]) that the minimum of the objective function is achieved if and only if $\tilde{P}^T(s, a, s', *) = |\{i \in \{0, \ldots, T_i\} \mid s_{\nu_{i;s}} = s'\}|/\tau_s = \#(s, a, s')/\#(s, a)$. $\square$

Proof of Theorem 4. Let $T > N$. As in the proof of Proposition 3, we decouple the transition probabilities at different state-action pairs (s, a). To emphasize that the transition probabilities $P^T(s, a, \cdot, t)$ for $t \geq n$ are time-invariant for $t \geq N$, we denote them by $P(s,a,\cdot,*_{t\geq N})$, and, analogously, their estimates by $\tilde{P}(s,a,\cdot,*_{t\geq N})$. By [45], or easily by direct computations, the probabilities $P(s,a,\cdot,*_{t\geq N})$ are a unique solution to

$$\begin{aligned} \min_{\tilde{P}^T(s,a,\cdot,*_{t\geq N})} \quad & -\sum_{s'\in S} P(s,a,s',*_{t\geq N}) \log \tilde{P}^T(s,a,s',*_{t\geq N}) && (14)\\ \text{s.t.} \quad & \tilde{P}^T(s,a,s',*_{t\geq N}) \geq 0 \qquad \text{for all } s' \in S, && (15)\\ & \sum_{s'\in S} \tilde{P}^T(s,a,s',*_{t\geq N}) = 1. && (16) \end{aligned}$$

Hence, they also make up a solution to the optimization problem with the same objective function, but with additional decision variables $\tilde{P}^T(s,a,s',t)$, $t<N$, that satisfy

$$\begin{aligned} \tilde{P}^T(s,a,s',t) \geq 0 \quad & \text{for all } s' \in S,\ t<n,\\ \sum_{s'\in S} \tilde{P}^T(s,a,s',t) = 1 \quad & \text{for all } t<n,\\ \left|\tilde{P}^T(s,a,s',t+1) - \tilde{P}^T(s,a,s',t)\right| \leq \varepsilon_t \quad & \text{for all } s' \in S,\ t<n-1, \text{ and}\\ \left|\tilde{P}^T(s,a,s',*_{t\geq N}) - \tilde{P}^T(s,a,s',N-1)\right| \leq \varepsilon_{N-1} \quad & \text{for all } s' \in S. \end{aligned} \tag{17}$$

Additionally, for all solutions to this modified problem, $P(s,a,\cdot,*_{t\geq N})$ are always the same by above.

Now, at time T, the objective function in (6) will equal

$$-\sum_{i=0}^{\tau'_s} \log \tilde{P}^T(s,a,s_{\nu_{i;s}},\nu_{i;s}) - \sum_{s'\in S} (\#(s,a,s') - |\{t \leq N \mid s_t = s'\}|) \log \tilde{P}^T(s,a,s',*_{t\geq N}), \tag{18}$$

where τ'_s indicates the number of times that s is visited until $t=N$. We note that τ'_s and $|\{t \leq N \mid s_t = s'\}|$ are independent of T. Hence, by the law of large numbers, the objective function (18), when divided by $\#(s,a)$, converges with probability 1 on the compact set given by (15)–(17) pointwise to the objective function in (14) as $\#(s,a) \to \infty$. Since there is a unique solution $\tilde{P}^T(s,a,s',*_{t\geq N}) = P(s,a,s',*_{t\geq N})$ of (14)–(17), it can be shown (e.g., by Theorem 7.33 in [46]) that any solution $\tilde{P}^T(s,a,s',*_{t\geq N}) = \tilde{P}^T(s,a,s',T-1)$ to the problem of minimizing (18) under conditions (15)–(17) converges to $P(s,a,s',*_{t\geq N}) = P(s,a,s',T-1)$ as $\#(s,a) \to \infty$. □

Proof of Theorem 5. Suppose first that (s,a) has not been visited before $t=N$. Let $N \leq T_0 < T_1 < \ldots$ denote the times at which (s,a) is been visited. After decoupling (6) into $|A||S|$ optimization problems by fixing $s \in S$ and $a \in A$, the choice $\tilde{P}^{T_i+1}(s,a,s',\cdot) = 1$ and $\tilde{P}^{T_i+1}(s,a,s'',\cdot) = 0$ for all $s'' \neq s'$, for all $i \geq 0$, clearly minimizes the objective function

$$-\sum_{r=0}^{i} \log \tilde{P}^{T_i+1}(s,a,s',T_r)$$

for all $i \geq 0$, while satisfying the constraints of (6).

We now consider the case when (s,a) has been visited before $t=N$. Let T_0 be the last time at which (s,a) is visited before $t=N$, and let $T_0 < T_1 < \ldots < T_{k-1} < N + 1/\varepsilon \leq T_k < T_{k+1} < \ldots$, $k \geq 1$, denote all times at which (s,a) is visited starting at T_0. We claim that

$$\tilde{P}^{T_i+1}(s,a,s',T_i) \in \left[\min\left(1, \tilde{P}^{T_{i-1}+1}(s,a,s',T_{i-1}) + (T_i - T_{i-1})\varepsilon\right), 1\right] \tag{19}$$

for all $i \geq 1$.

Assume that (19) does not hold. Since $\tilde{P}^{T_i+1}(s,a,s',T_i) \leq 1$, we have

$$\tilde{P}^{T_i+1}(s,a,s',T_i) < \min\left(1, \tilde{P}^{T_{i-1}+1}(s,a,s',T_{i-1}) + (T_i - T_{i-1})\varepsilon\right). \tag{20}$$

Now, define an alternative choice of transition probabilities as follows:

$$\overline{P}^{T_i+1}(s,a,s^*,T) = \begin{cases} \tilde{P}^{T_{i-1}+1}(s,a,s^*,T) & \text{for all } s^* \in S,\ T \leq T_{i-1},\\ \min\left(1, \tilde{P}^{T_{i-1}+1}(s,a,s',T_{i-1}) + (T_i - T_{i-1})\varepsilon\right) & \text{if } s^* = s' \text{ and } T = T_i. \end{cases} \tag{21}$$

Let $d_P^1 = \overline{P}^{T_i+1}(s,a,s',T_i) - \tilde{P}^{T_{i-1}+1}(s,a,s',T_{i-1}) > 0$. We define $\overline{P}^{T_i+1}(s,a,s^*,T_i)$ for $s^* \neq s'$ in the following way: if $S = \{s^1, \ldots, s^n\}$, where $s' = s^1$, then recursively define

$$\begin{aligned} &\overline{P}^{T_i+1}(s,a,s^r,T_i) = \max\left(0, \tilde{P}^{T_{i-1}+1}(s,a,s^r,T_{i-1}) - d_P^{r-1}\right),\\ &d^r = \tilde{P}^{T_{i-1}+1}(s,a,s^r,T_{i-1}) - \overline{P}^{T_i+1}(s,a,s^r,T_i),\\ &d_P^r = d_P^{r-1} - d^r \end{aligned} \tag{22}$$

for $r \geq 2$. We also define $d^1 = d^1_P$.

We will show that $\overline{P}^{T_i+1}(s, a, \cdot, T_i) \geq 0$ as defined in (21)–(22) is a legitimate discrete probability distribution:

1) By definition, $\overline{P}^{T_i+1}(s, a, s^r, T_i) \geq 0$ for all r.
2) By (21) and (22),

$$\sum_{r=1}^{n} \overline{P}^{T_i+1}(s, a, s^r, T_i) = \sum_{r=1}^{n} \tilde{P}^{T_{i-1}+1}(s, a, s^r, T_{i-1}) + d^1 - \sum_{r=2}^{n} d^r = 1 + d^1 - \sum_{r=2}^{n} d^r = 1 + d^n_P \geq 1.$$

We claim that $d^n_P = 0$. If $\tilde{P}^{T_{i-1}+1}(s, a, s^r, T_{i-1}) \geq d^{r-1}_P$ for any $r \geq 2$, then $d^r_P = 0$, and $d^n_P = d^{n-1}_P = \cdots = d^r_P = 0$. Thus, if $d^n_P > 0$, then $\tilde{P}^{T_{i-1}+1}(s, a, s^r, T_{i-1}) < d^{r-1}_P$ for all $r \geq 2$. Hence, by (22), $\overline{P}^{T_i+1}(s, a, s^r, T_i) = 0$ for all $r \geq 2$. Then,

$$1 + d^n_P = \sum_{r=1} \overline{P}^{T_i+1}(s, a, s^r, T_i) = \overline{P}^{T_i+1}(s, a, s^1, T_i) \leq 1,$$

where the inequality holds by (21). Thus, $d^n_P \leq 0$, contradicting the assumption $d^n_P > 0$.

Additionally, by (22), $d^r \geq 0$ and $d^r \leq d^{r-1}_P \leq d^{r-2}_P \leq \cdots \leq d^1_P \leq (T_i - T_{i-1})\varepsilon$ for all r. Thus,

$$\left| \overline{P}^{T_i+1}(s, a, s^r, T_i) - \tilde{P}^{T_{i-1}+1}(s, a, s^r, T_{i-1}) \right| \leq (T_i - T_{i-1})\varepsilon$$

for all r. Hence, for all $s^* \in S$, we define $\overline{P}^{T_i+1}(s, a, s^*, T)$ for $T \in \{T_{i-1}+1, \ldots, T_i - 1\}$ by

$$\overline{P}^{T_i+1}(s, a, s^*, T) = \frac{T_i - T}{T_i - T_{i-1}} \overline{P}^{T_i+1}(s, a, s^*, T_{i-1}) + \frac{T - T_{i-1}}{T_i - T_{i-1}} \overline{P}^{T_i+1}(s, a, s^*, T_i), \tag{23}$$

thus ensuring that $|\overline{P}^{T_i+1}(s, a, s^*, T+1) - \overline{P}^{T_i+1}(s, a, s^*, T)| \leq \varepsilon$ remains satisfied for all T. It can also be easily verified that $\overline{P}^{T_i+1}(s, a, s^*, T) \geq 0$ for all s^*, and that these values sum up to 1.

We verified that $\overline{P}^{T_i+1}_{t<T_i+1}$, as defined in (21)–(23), satisfies all the constraints in (6). The value of the objective function for $\overline{P}^{T_i+1}_{t<T_i+1}$ is strictly lower than for $\tilde{P}^{T_i+1}_{t<T_i+1}$: the values for $\overline{P}^{T_i+1}_{t<T_{i-1}+1}$ have been chosen to be optimal, and the only other element present in the objective function, $\overline{P}^{T_i+1}(s, a, s', T_i)$, satisfies $\overline{P}^{T_i+1}(s, a, s', T_i) > \tilde{P}^{T_i+1}(s, a, s', T_i)$ by (20) and (21). Thus, we reached a contradiction with $\tilde{P}^{T_i+1}_{t<T_i+1}$ being a CCMLE.

Claim (19) is thus proved. Now, for each $i \geq 1$ we either have $\tilde{P}^{T_i+1}(s, a, s', T_i) \geq \tilde{P}^{T_{i-1}+1}(s, a, s', T_{i-1}) + (T_i - T_{i-1})\varepsilon$ or $\tilde{P}^{T_i+1}(s, a, s', T_i) = 1$. Assume that $\tilde{P}^{T_i+1}(s, a, s', T_i) < 1$ for some $i \geq k$, i.e., for $T_i \geq N + 1/\varepsilon$. Then, $\tilde{P}^{T_{i-1}+1}(s, a, s', T_{i-1}) < 1 - (T_i - T_{i-1})\varepsilon$. Continuing onwards, we obtain that $\tilde{P}^{T_0+1}(s, a, s', T_0) < 1 - (T_i - T_0)\varepsilon$. Since $T_i - T_0 \geq 1/\varepsilon$, we obtain $\tilde{P}^{T_0+1}(s, a, s', T_0) < 1$, i.e., a contradiction. □

Proof of Lemma 7. After decoupling (6), the objective function for (s, a) equals

$$-\sum_{i=0}^{k} \log \tilde{P}^T(s, a, s_{T_i+1}, T_i).$$

Assume that there exist two solutions $\tilde{P}^T_1(s, a, s_{T_i+1}, T_i)$ and $\tilde{P}^T_2(s, a, s_{T_i+1}, T_i)$ yielding the same minimal value of the objective function. Then, by a simple convexity argument (see, e.g., [47]), $\lambda\tilde{P}^T_1(s, a, s_{T_i+1}, T_i) + (1-\lambda)\tilde{P}^T_2(s, a, s_{T_i+1}, T_i)$ all need to yield the same value, for all $\lambda \in [0, 1]$. Thus,

$$-\sum_{i=0}^{k} \log\left(\lambda\tilde{P}^T_1(s, a, s_{T_i+1}, T_i) + (1-\lambda)\tilde{P}^T_2(s, a, s_{T_i+1}, T_i)\right) \tag{24}$$

is a constant function of $\lambda \in [0, 1]$. Taking the derivative of (24) with respect to λ, we obtain

$$\sum_{i=0}^{k} \frac{\tilde{P}^T_1(s, a, s_{T_i+1}, T_i) - \tilde{P}^T_2(s, a, s_{T_i+1}, T_i)}{\lambda\tilde{P}^T_1(s, a, s_{T_i+1}, T_i) + (1-\lambda)\tilde{P}^T_2(s, a, s_{T_i+1}, T_i)} = 0$$

for all $\lambda \in (0, 1)$. Taking the second derivative, we obtain

$$\sum_{i=0}^{k} \frac{(\tilde{P}^T_1(s, a, s_{T_i+1}, T_i) - \tilde{P}^T_2(s, a, s_{T_i+1}, T_i))^2}{(\lambda\tilde{P}^T_1(s, a, s_{T_i+1}, T_i) + (1-\lambda)\tilde{P}^T_2(s, a, s_{T_i+1}, T_i))^2} = 0.$$

In other words, $\tilde{P}^T_1(s, a, s_{T_i+1}, T_i) = \tilde{P}^T_2(s, a, s_{T_i+1}, T_i)$ for all i. □

Proof of Theorem 11. If $\#(s, a) = 0$, the claim is obvious, as $\mathcal{P}^T_{\sigma,\alpha}(s, a, *)$ is the entire probability simplex, so $U_{\sigma,\alpha} = \sqrt{2}$.

Assume now that $\#(s,a) \geq 1$. By the proof of Proposition 3, $\mathcal{P}^T_{\sigma,\alpha}(s,a,*) \subseteq \mathbb{R}^{|S|}$ contains a single element given by components $\tilde{P}^T(s,a,s',*) = \#(s,a,s')/\#(s,a)$, where $\#(s,a,s')$ denotes the number of times that transition (s,a,s') has occurred among the first $T-1$ transitions. Thus, the diameter of $\mathcal{P}^T_{\sigma,\alpha}(s,a,*)$ is 0.

For each $s'' \in S$, again by the proof of Proposition 3, set $\mathcal{P}^{T+1}_{\overline{\sigma}_{s''},\overline{\alpha}}(s,a,*)$ also contains a single element given by $\tilde{P}^{T+1}(s,a,s'',*) = (\#(s,a,s'')+1)/(\#(s,a)+1)$ and $\tilde{P}^{T+1}(s,a,s',*) = \#(s,a,s')/(\#(s,a)+1)$ for all $s' \neq s''$.

Thus, the maximum distance between elements of polytopes $\mathcal{P}^T_{\sigma,\alpha}(s,a,*)$ and $\mathcal{P}^{T+1}_{\overline{\sigma}_{s''},\overline{\alpha}}(s,a,*)$ equals

$$\begin{aligned}&\sqrt{\left(\frac{\#(s,a,s'')+1}{\#(s,a)+1} - \frac{\#(s,a,s'')}{\#(s,a)}\right)^2 + \sum_{s' \neq s''}\left(\frac{\#(s,a,s')}{\#(s,a)+1} - \frac{\#(s,a,s')}{\#(s,a)}\right)^2} = \\ &\frac{1}{\#(s,a)(\#(s,a)+1)}\sqrt{\sum_{s'\neq s''} \#(s,a,s')^2 + (\#(s,a) - \#(s,a,s''))^2}.\end{aligned} \tag{25}$$

On one hand, by the power mean inequality [48], the value of (25) is greater than or equal to

$$\frac{1}{\#(s,a)(\#(s,a)+1)}\sqrt{\frac{\left(\sum_{s'\neq s''} \#(s,a,s')\right)^2}{|S|-1} + (\#(s,a) - \#(s,a,s''))^2} = \frac{\sqrt{|S|}\,(\#(s,a) - \#(s,a,s''))}{\#(s,a)(\#(s,a)+1)\sqrt{|S|-1}}. \tag{26}$$

We are interested in determining the maximal value of (25) over different $s'' \in S$. There exists s'' such that $\#(s,a,s'') \leq \#(s,a)/|S|$. By plugging in this s'' into (26), we obtain that the value of (25) is greater than or equal to $\sqrt{1-1/|S|}/(1+\#(s,a))$.

On the other hand, the value of (25) is trivially less than or equal to

$$\frac{1}{\#(s,a)(\#(s,a)+1)}\sqrt{\left(\sum_{s'\in S} \#(s,a,s')\right)^2 + \#(s,a)^2} = \frac{\sqrt{2}}{\#(s,a)+1}$$

for any s''. By (8), we obtain the desired claim. □

Appendix B
Additional Simulations

A. Wind Flow Estimation

While the examples of Section VI were largely constructed to vividly describe and support the obtained theoretical results, we now provide a more realistic estimation example, simultaneously featuring (i) changes of multiple transition probabilities by different amounts, (ii) complex, time-varying, and partly random rates of change, and (iii) a poorly chosen a priori bound ε on the rate of change of transition probabilities, i.e., a bound that is sometimes overly conservative — allowing for a larger change than it actually is — and sometimes incorrect, expecting only changes smaller than the actual ones.

The setting of this simulation is that of an unpowered aerial vehicle without extensive instrumentation, i.e., a pilot balloon. The movement of such a vehicle — tracked from the ground or using a GPS receiver as its sole instrument — is commonly used to estimate the dynamics in the balloon's area of operation [49]. As the balloon is not equipped with any instruments apart from possibly a GPS receiver, the estimation is based solely on the observed trajectory. In this section, we will develop a CCMLE-based method for estimating the balloon's dynamics in a changing wind flow.

We develop our model of the system dynamics based on the work of [50], which devises and discusses a scheme to convert wind direction into transition probabilities on an MDP. Given that the balloon is unpowered, an MDP reduces to a Markov chain, i.e., to an MDP with a single action. Additionally, as the balloon cannot intentionally keep revisiting the same state, the natural underlying assumption is that the dynamics in the area of operation are *uniform*, i.e., do not vary across the state space. However, the wind flow and the subsequent dynamics are time-varying.

We discretize the state space as a grid similar to the one used in Section VI-A, although we allow such a grid to be infinite or arbitrarily large. At every time step, the balloon moves by one tile in one of four possible directions, denoted by their angle with the positive ray of the x-axis: north ($\pi/2$), east (0), south ($3\pi/2$), or west (π). Combined with the assumptions of uniformity and time-varying nature of the wind flow, the presented setting yields a probability transition function $P : \{0, \pi/2, \pi, 3\pi/2\} \times \mathbb{N}_0 \to [0,1]$, where $P(o,t)$ signifies the probability at time t that the balloon will move in the direction denoted by o.

For simplicity, we consider wind speed to be constant, and only consider changes in its direction $d(t) \in [0, 2\pi)$. Adapting the work in [50], we develop the following model for transition probabilities. Given the wind direction d, the balloon movement

$o \in \{0, \pi/2, \pi, 3\pi/2\}$ is a discrete random variable obtained by rounding O to the nearest $\pi/2$ (modulo 2π), where O is a normal random variable with mean $\mathbb{E}[O] = d$ and variance $\sigma^2 = 1/2$. In other words,

$$\mathbb{P}(o|d) = \sum_{n=-\infty}^{\infty} \int_{2n\pi+o-\pi/4-d}^{2n\pi+o+\pi/4-d} \frac{2}{\sqrt{\pi}} e^{-t^2} dt.$$

The transition probability $P(o,t)$ is then naturally given by $P(o,t) = \mathbb{P}(o|d(t))$. We emphasize that the estimator is neither aware of the wind direction at any time, not aware of the relationship between d and o. The estimator is not attempting to establish the wind direction $d(t)$, but solely estimate the transition probabilities $P(\cdot, t)$.

We simulate the wind flow given by

$$\begin{aligned} d(t+1) &= d(t) - 3\pi/180 + X(t), \\ d(0) &= D_0, \end{aligned}$$

where $X(t)$ is a random variable whose value is drawn from a uniform distribution on $[-\pi/180, \pi/180]$, and $Y(t)$ is a random variable whose value is drawn from a uniform distribution on $[0, 2\pi]$. In other words, the wind changes direction by 2 to 4 degrees at every time step, resulting in continually changing transition probabilities.

To demonstrate the strength of the CCMLE approach, we chose $\varepsilon = 0.03$ as the bound on rate of change in P. For greater realism of the example, such a bound is — as in the setting illustrated in Fig. 4 — intentionally incorrect. the actual change in P can be computed from the equations above to be between 0.01 and 0.04. Fig. 11 compares the maximal error in the transition probabilities of a CCMLE approach and a classical approach. It illustrates the estimation errors for 240 time steps, allowing the wind direction vectors to make around two full circles.

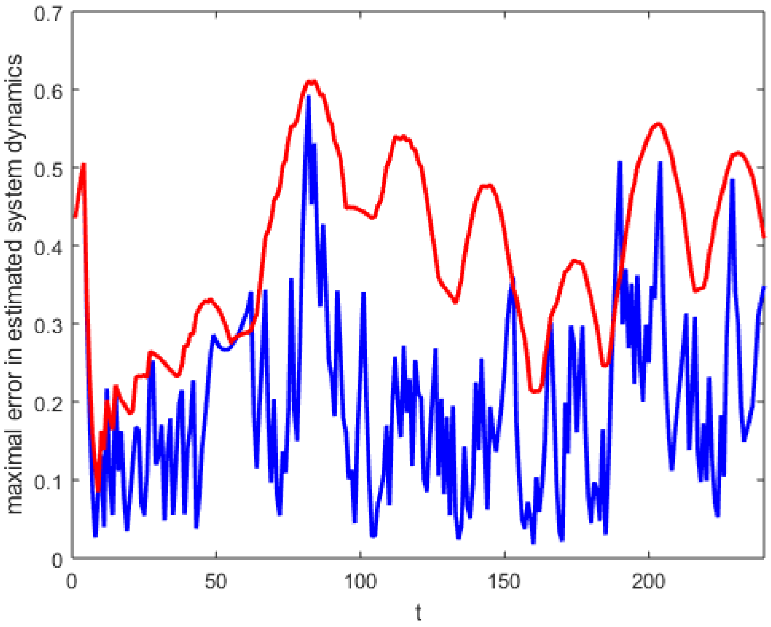

Fig. 11. Maximal error in the transition probability estimates for the wind flow estimation problem. The red curve indicates the error with classical estimation that assumes time-invariant wind flow. The blue curve indicates the error with the CCMLE method.

Even though the simulated setting is not covered by the developed theory of CCMLE, the CCMLE approach results in significantly better estimation than the classical estimation method. While "spikes" in errors are provably unavoidable — when the transitions are non-deterministic, any arbitrarily large sequence of low-probability outcomes will eventually occur — the error from the CCMLE exceeds the error of classical estimation only on a handful of occasions. The average of the maximal CCMLE error over time equals 0.2, while the average of the classical estimation method is 0.39. We emphasize that these results were obtained in the case where ε was not given correctly, and none of the developed theoretical results were thus guaranteed to hold.

B. Multi-Armed Bandit

In this example we modify the two-armed bandit setup used in Section VI-B. Unlike Section VI-B, we will use the standard terminology of theory of multi-armed bandits [42]. We consider n arms. Each arm pull lasts a single time step, and at each time step, exactly one bandit arm must be pulled. Arm 1 always produces the reward of 1. For each i, $i \in \{2, \ldots, n\}$, arm i

produces one of the two rewards: a reward of i with probability $0.95(\sin(\alpha_i t+\beta_i)+1)/i$, and a reward of 0 with probability $1-0.95(\sin(\alpha_i t+\beta_i)+1)/i$. While the possible rewards are known to the planner, their probabilities are not.

It is clear that, as $t\to\infty$, the average reward produced by each arm i, $i\in\{2,\ldots,n\}$, converges to 0.95. Thus, an agent that uses classical estimation for learning and planning and pulls the arm that is estimated to bring the highest reward will always eventually begin solely using arm 1. The average collected reward will thus converge to 1.

On the other hand, if sets $\{\alpha_i \mid i\geq 2\}$ or $\{\beta_i \mid i\geq 2\}$ are "sufficiently different", at every time there will exist an arm producing a reward greater than 1. Thus, by choosing the bandit arms wisely, an agent may collect an average reward greater than 1: the primary challenge is in deciding when to choose which arm to pull. In combination with the CCMLE, the notion of uncertainty introduced in Section IV and the subsequent control policy introduced in Section V-B provide a possible solution. By performing exploratory pulls when the uncertainty of probabilities in a particular arm becomes high, the agent is able to detect when an arm yields a high probability of rewards.

We simulated a 5-arm bandit during 10000 time steps, with all α_i chosen uniformly at random in $[0,1/5]$ and β_i chosen uniformly at random in $[0,2\pi)$. We use the same horizon length as in Section VI-B. Bound ε is 0.25. Such a bound is extremely loose — namely, more than three times larger than the amount of maximal change on any arm, and more than 40 times larger than the amount of the maximal change on the third arm. Uncertainty weight β is chosen to be $3/(2\sqrt{2})$. We remark that the computational complexity of the CCMLE method and subsequent planning depends only linearly on the number of bandits, as the optimization problems for computation of a CCMLE and uncertainties are performed separately on each bandit. While the number of variables in the optimization problems grows linearly with elapsed time, this dependence can be removed by adopting a notion of forgetfulness explored in Section VI-B. We adopt such a notion in the simulation, and limit the memorized history of outcomes for each bandit to $1+1/\varepsilon$, in line with the discussion at the end of Section 5.

Fig. 12 shows the collected average reward collected by an agent that follows the CCMLE method of estimation and decides which arms to pull based on the sum of expected reward and uncertainty bonus. The planner achieves a reward that is 25% higher than the average reward of an agent that does not use CCMLE or an uncertainty bonus. We emphasize that these results were obtained without any tuning of the weight β or bound ε, the latter of which was intentionally poorly chosen.

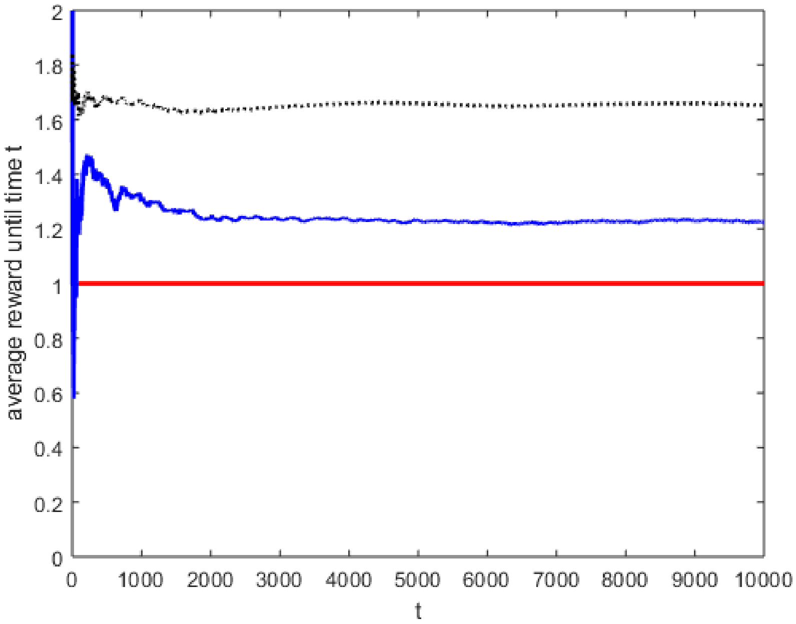

Fig. 12. The blue line indicates the average reward obtained by an agent who uses the forgetful CCMLE with an uncertainty-based learning bonus to choose the best among the 5 bandit arms. The dotted line denotes indicates the expected average reward obtained by an agent who has perfect a priori knowledge of the transition probabilities. The red line indicates the asymptotic average reward obtained by an agent that uses classic estimation unconscious of the change in the transition probabilities.