Online Markov Decision Processes with Kullback–Leibler Control Cost

Peng Guan, *Student Member, IEEE*, Maxim Raginsky, *Member, IEEE*, and Rebecca M. Willett, *Senior Member, IEEE*

Abstract

This paper considers an online (real-time) control problem that involves an agent performing a discrete-time random walk over a finite state space. The agent's action at each time step is to specify the probability distribution for the next state given the current state. Following the set-up of Todorov, the state-action cost at each time step is a sum of a state cost and a control cost given by the Kullback-Leibler (KL) divergence between the agent's next-state distribution and that determined by some fixed passive dynamics. The online aspect of the problem is due to the fact that the state cost functions are generated by a dynamic environment, and the agent learns the current state cost only after selecting an action. An explicit construction of a computationally efficient strategy with small regret (i.e., expected difference between its actual total cost and the smallest cost attainable using noncausal knowledge of the state costs) under mild regularity conditions is presented, along with a demonstration of the performance of the proposed strategy on a simulated target tracking problem. A number of new results on Markov decision processes with KL control cost are also obtained.

I. Introduction

Markov decision processes (MDPs) [1], [2], [3] comprise a popular framework for sequential decision-making in a random dynamic environment. At each time step, an agent observes the state

This work was supported by NSF grant CCF-1017564 and by AFOSR grant FA9550-10-1-0390. A preliminary version of this work was presented at the American Control Conference, Montreal, Canada, June 2012.

P. Guan is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: peng.guan@duke.edu).

M. Raginsky is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: maxim@illinois.edu).

R. Willett is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: willett@duke.edu).

of the system of interest and chooses an action. The system then transitions to its next state, with the transition probability determined by the current state and the action taken. There is a (possibly time-varying) cost associated with each admissible state-action pair, and a policy (feedback law) for mapping states to actions is selected to minimize average cost. In the basic MDP framework, it is assumed that the cost functions and the transition probabilities are known in advance, the policy is designed "offline" (e.g., using dynamic programming), and the optimality criterion is forward-looking, taking into account the effect of past actions on future costs. In many practical problems, however, this degree of advance knowledge is unavailable. When neither the transition probability nor the cost functions are known in advance, various reinforcement learning (RL) methods, such as the celebrated Q-learning algorithm [4], [5] and its variants, can be used to learn an optimal policy in an online regime. However, the key assumptions underlying RL are that the agent is operating in a stochastically stable environment, and that the state-action costs (or at least their expected values with respect to any environmental randomness) do not vary with time. These assumptions are needed to ensure that the agent is eventually able to learn an optimal stationary control policy.

Another framework for sequential decision-making, dating back to the seminal work of Robbins [6] and Hannan [7] and now widely used in the machine learning community [8], deals with nonstochastic, unpredictable environments. In this *online learning* (or *sequential prediction*) framework, the effects of the environment are modeled by an arbitrarily varying sequence of cost functions, where the cost function at each time step is revealed to the agent only *after* an action has been taken. There is no state, and the goal of the agent is to minimize *regret*, i.e., the difference between the total cost incurred using causally available information and the total cost of the best single action that could have been chosen in hindsight. In contrast with MDPs, the regret-based optimality criterion is necessarily myopic and backward-looking, since the cost incurred at each time step depends only on the action taken at that time step, so past actions have no effect on future costs. There is also a more stringent model of online learning, in which the agent observes not the entire cost function for each time step, but only the value of this cost at the currently taken action [9]. This model is inspired by the celebrated *multiarmed bandit* problem first introduced by Robbins [10], and is referred to as the *nonstochastic bandit problem*. One widely used way of constructing regret-minimizing strategies for such bandit problems is to randomize the agent's actions (*exploration*) so that the random cost value revealed to the agent

can be used to construct an unbiased estimate of the full cost function, which is then fed into a suitable strategy that minimizes regret under the assumption of full information (*exploitation*). We will not consider nonstochastic bandit problems in this paper. Instead, we refer the reader to a recent survey by Bubeck and Cesa-Bianchi [11] that discusses both stochastic and nonstochastic bandit problems.

Recent work by Even-Dar et al. [12] and Yu et al. [13] combines the MDP and the online learning frameworks into what may be described as *online MDPs* with finite state and action spaces. Like in the traditional MDP setting, the agent observes the current state and chooses an action, and the system transitions to the next state according to a fixed and known Markov law. However, like in the online framework, the one-step cost functions form an arbitrarily varying sequence, and the cost function corresponding to each time step is revealed to the agent after the action has been taken. The objective of the agent is to minimize regret relative to the best stationary Markov policy that could have been selected with full knowledge of the cost function sequence over the horizon of interest. The time-varying cost functions may represent unmodeled aspects of the environment or collective (and possibly irrational) behavior of any other agents that may be present; the regret minimization viewpoint then ensures that the agent's *online* policy is robust against these effects.

A. *Brief problem statement and motivating examples*

We give here a brief statement of the problem of interest in order to fix ideas; a more detailed formulation is given later on. The reader may wish to consult Section I-D for notation.

The set-up considered in [12], [13] is motivated by problems in machine learning and artificial intelligence, where the actions are the main object of interest, and the state merely represents memory effects present in the system. In this paper, we take a more control-oriented view: the emphasis is on steering the system along a desirable state trajectory through actions selected according to a state feedback law. Following the formulation proposed recently by Todorov [14], [15], [16], we allow the agent to modulate the state transitions directly, so that actions (resp., state feedback laws) correspond to probability distributions (resp., Markov kernels) on the underlying state space. As in [14], [15], [16], the one-step cost is a sum of two terms: the state cost, which measures how "desirable" each state is, and the control cost, which measures the deviation of the transition probabilities specified by the chosen action from some fixed *default* or *passive*

dynamics. (We also refer the reader to a recent paper by Kappen et al. [17], which interprets Todorov's set-up as an inference problem for probabilistic graphical models.)

More precisely, we consider an MDP with a finite state space X, where the action space U is the simplex $\mathcal{P}(\mathsf{X})$ of probability distributions over X. A fixed Markov matrix (transition kernel) $P^* = [P^*(x,y)]_{x,y\in\mathsf{X}}$ is given. A stationary Markov policy (state feedback law) is a mapping $w : \mathsf{X} \to \mathcal{P}(\mathsf{X})$, so if the system is in state $x \in \mathsf{X}$, then the transition to the next state is stochastic, as determined by the probability distribution $u(\cdot) = w(x) \in \mathcal{P}(\mathsf{X})$. In other words, if we denote the next state by X^+, then the state transitions induced by the action u are governed by the conditional probability law

$$\Pr\{X^+ = x^+ | X = x\} = P(x, x^+) = u(x^+) = [w(x)](x^+).$$

The one-step state-action cost $c(x,u)$ consists of two terms, the *state cost* $f(x)$, where $f : \mathsf{X} \mapsto \mathbb{R}_+$ is a given function, and the *control cost*, which penalizes any deviation of the next-state distribution $u(\cdot) = w(x)$ from the one prescribed by $P^*(x,\cdot)$, the row of P^* corresponding to x. To motivate the introduction of such control costs, we can imagine the situation, in which implementing the state transitions according to P^* can be done "for free." However, it may very well be the case that following P^* will be in conflict with the goal of keeping the state cost low. From this perspective, it may actually be desirable to deviate from P^*. Any such deviation may be viewed as an *active perturbation* of the *passive dynamics* prescribed by P^*, and the agent should attempt to balance the tendency to keep the state costs low against allowing too strong of a perturbation of P^*. Our choice of control cost is inspired by the work of Todorov [14], [16], and is given by the *Kullback–Leibler divergence* (or the *relative entropy*) [18] $D(u\|P^*(x,\cdot))$ between the proposed next-state distribution $u(\cdot)$ and the next-state distribution prescribed by the passive dynamics P^*. One useful property of this control cost is that it automatically forbids all those state transitions that are already forbidden by P^*. Indeed, if for a given $x \in \mathsf{X}$ there exists some $y \in \mathsf{X}$ such that $u(y) = [w(x)](y) > 0$, while $P^*(x,y) = 0$, then $D(u\|P^*(x,\cdot)) = +\infty$. Thus, the overall one-step state-action cost is given by

$$c(x,u) = f(x) + D(u\|P^*(x,\cdot)), \quad \forall x \in \mathsf{X}, u \in \mathcal{P}(\mathsf{X}). \tag{1}$$

In the online version of this problem (detailed in Section II-A), the state costs form an arbitrarily varying sequence $\{f_t\}_{t=1}^{\infty}$, and the agent learns the state cost for each time step only after having

selected the transition law to determine the next state. For any given value of the horizon, the regret is computed with respect to the best stationary Markov policy (state feedback law) that could have been chosen in hindsight. The precise definition of regret is given in Section II-B.

Since this is a nonstandard set-up, we take a moment to situate it in the context of usual models of MDPs. In a standard MDP with finite state and action spaces, we have a finite collection of Markov matrices P_u on X indexed by the actions $u \in \mathsf{U}$. State feedback laws are functions $w : \mathsf{X} \to \mathsf{U}$, and the set of all such functions is finite with cardinality $|\mathsf{U}|^{|\mathsf{X}|}$. Therefore, in each state $x \in \mathsf{X}$ the agent has at most $|\mathsf{U}|^{|\mathsf{X}|}$ choices for the distribution of the next state X^+, and we may equivalently represent each state feedback law w as a mapping from X into $\mathcal{P}(\mathsf{X})$ with $x \mapsto P_{w(x)}(x, \cdot)$. Since the state space U is finite, the range of this mapping is a finite subset of the probability simplex $\mathcal{P}(\mathsf{X})$. The criterion for selecting this next-state distribution pertains to minimization of the expectation of the immediate state-action cost plus a suitable value function that accounts for the effect of the current action on future costs. In many cases, the one-step state-action cost $c(x, u)$ decomposes into a sum of state cost $f(x)$ and control cost $g(x, u)$, where $f(x)$ quantifies the (un)desirability of the state x, while $g(x, u)$ represents the effort required to apply action u in state x.

In the set-up of [14], [15], [16], the collection of all possible next-state distributions is unrestricted. As a consequence, any mapping $w : \mathsf{X} \to \mathcal{P}(\mathsf{X})$ is a feasible stationary Markov policy. Since any Markov matrix P on X can be equivalently represented as a mapping from X into $\mathcal{P}(\mathsf{X})$ with $x \mapsto P(x, \cdot)$, there is thus a one-to-one correspondence between state feedback laws and Markov matrices on X. In contrast to the case when the agent may choose among a finite set of actions, the probability simplex $\mathcal{P}(\mathsf{X})$ is an uncountable set, so the agent has considerably greater freedom to choose the next state distribution. As before, we introduce a state-action cost $c(x, u) = f(x) + g(x, u)$, where $f(x)$ measures the (un)desirability of state x, while $g(x, u)$ quantifies the difficulty of executing action u in state x. Since actions u now correspond to probability distributions, and we choose the Kullback–Leibler control cost $g(x, u) = D(u \| P^*(x, \cdot))$, where P^* is a fixed Markov matrix on the state space X that may represent, e.g., the "free" dynamics of the system in the absence of external controls.

The Kullback–Leibler divergence is widely used in stochastic control and inference. First of all, it has many desirable properties, such as nonnegativity and convexity [18]. Secondly, if we adopt the viewpoint that the purpose of a control policy is to shape the joint distribution of

all relevant variables describing the closed-loop behavior of the system, then using the relative entropy to compare the distribution induced by any control law to some reference model leads to functional equations for the optimal policy that are often easier to solve than the corresponding dynamical programming recursion [19], [20] (e.g., see [19] for an alternative derivation of the optimal controller in an LQG problem using relative entropy instead of dynamic programming); similar ideas are fruitful in the context of robust control, where the relative entropy is used to quantify the radius of uncertainty around some nominal system [21], [22], [23]. Moreover, the relative entropy is a canonical regularization functional for stochastic nonlinear filtering problems [24]: an optimal Bayesian filter is the solution of a variational problem that entails minimization of the sum of expected negative log-likelihood (which can be interpreted as state cost) and a relative entropy with respect to the prior measure on the state space.

To further motivate our interest in problems of this sort, let us consider two examples. One is *target tracking* with an arbitrarily moving target (or multiple targets). In this example, the state space X is the vertex set of an undirected graph, and the passive dynamics P^* specifies some default *random walk* on this graph. The tracker's discrete-time motion is constrained by the topology of the graph, while the targets' motions are not. At each time t, the state cost f_t is the tracking error, which quantifies how far the tracker is from the targets. For instance, it may be given by the graph distance (length of shortest path) between the tracker's current location and the location of the closest target. Other possibilities can also be considered, including some based on noisy information on the location of the targets. The control cost penalizes the tracker's deviation from P^* as it attempts to track the targets. The passive dynamics P^* can be seen as the tracker's prior model for the targets' motion. Moreover, if P^* is sufficiently rapidly mixing, then any tracker that follows P^* will visit every vertex of the graph infinitely often with probability one; however, there is no guarantee that the tracker's prior model is correct (i.e., that the tracker will be anywhere near the targets). Hence, the state-action cost will trade off the tendency of the tracker to "cover" the graph as much as possible (exploration) against the tendency to follow a potentially faulty model of the targets (exploitation).

Another example setting is real-time control of a *brain-machine interface*. There, the state space X may be the set of possible positions or modes of a neural prosthetic device, and the passive dynamics P^* may encode the "natural" (free) dynamics of the device in the absence of user control; we may assume, for instance, that the state transitions prescribed by P^* correspond

to "minimum-energy" operating mode of the device. If the user wishes to make the device execute some trajectory, the state cost f_t at time t may represent the deviation of the current point on the trajectory from the one intended by the user. Since the user is a human operator with conscious intent, we may not want to ascribe an a priori model to her intended trajectory, and instead treat it as an individual sequence modulating the state costs $\{f_t\}_{t=1}^{\infty}$. In this setting, the Kullback–Leibler control cost penalizes significant deviations from the free dynamics P^*, since these will typically be associated with energy expenditures.

The common thread running through these two examples (and it is certainly possible to construct many others) is that they model real-time interaction of a particular system with some well-defined "reference" or "nominal" dynamics P^* with a potentially unpredictable environment (which may include hard-to-model adversaries or rational agents, etc.), and we must balance the tendency to respond to immediate changes in the environment against the need to operate the system near the nominal mode. Since no offline policy design is possible in such circumstances, the regret minimization framework offers a meaningful alternative.

B. Our contributions and comparison with relevant literature

In this paper, we give an explicit construction of a strategy for the agent, such that the regret relative to any uniformly ergodic class of stationary Markov policies grows *sublinearly* as a function of the horizon. The only regularity conditions needed for this result to hold are (a) uniform boundedness of the state costs (the agent need not know the bound, only that it exists); and (b) ergodicity of the passive dynamics. Moreover, our strategy is computationally efficient: the time is divided into phases of increasing length, and during each phase the agent applies a stationary Markov policy optimized for the average of the state cost functions revealed during all of the preceding phases. Thus, our strategy belongs to the class of so-called "lazy" strategies for online decision-making problems [25], [26], [27]; a similar approach was also taken by Yu et al. [13] in their paper on online MDPs with finite state and action spaces. The main advantage of lazy strategies is their computational efficiency, which, however, comes at the price of suboptimal scaling of the regret with the time horizon. We comment on this issue further in the sequel.

Our main contribution is an extension of the theory of online MDPs to a wide class of control problems that lie outside the scope of existing approaches [12], [13]. More specifically:

1) While in [12], [13] both the state and the action spaces are finite, we only assume this for

the state space. Our action space is the simplex of probability distributions on the state space, which is a compact subset of a Euclidean space. Hence, the techniques used in the existing literature are no longer directly applicable. (It is also possible to extend our approach to continuous state spaces, but additional regularity conditions will be needed. This extension will be the focus of our future work.)

2) Yu et al. [13] assume that the underlying MDP is unichain [1, Sec. 8.3] and satisfies a certain uniform ergodicity condition (a similar assumption is also needed by Even-Dar et al. [12]). The unichain assumption is rather strong, since it places significant simultaneous restrictions on an exponentially large family of Markov chains on the state space (each chain corresponds to a particular choice of state feedback law, and there are $|\mathsf{U}|^{|\mathsf{X}|}$ such laws). It is also difficult to verify, since the problem of determining whether an MDP is unichain is NP-hard [28]. By contrast, our ergodicity assumption pertains to only *one* Markov chain (the passive dynamics P^*), it can be efficiently verified in polynomial time, and we prove that it automatically implies uniform ergodicity of all stationary control laws that could possibly be invoked by our strategy.
3) Because these stationary control laws correspond to solutions of certain average-cost optimality equations (ACOEs) in the set-up of Todorov [14], [15], [16], we establish and subsequently exploit several useful and previously unknown results concerning the continuity and uniform ergodicity of optimal policies for Todorov's problem. These results, as well as the techniques used to prove them, play a very important role in our overall contribution. Indeed, in the online setting, the state cost functions are revealed to the agent in real time. Hence, any policy used by the agent must rely on estimates (or forecasts) of future state costs based on currently available information. Our new results on Todorov's optimal control laws provide sharp bounds on the sensitivity of these laws to misspecification of state costs, and may be of independent interest.
4) In [13], the policy computation at the beginning of each phase requires solving a linear program and then adding a carefully tuned random perturbation to the solution. As a result, the performance analysis in [13] is rather lengthy and technical (in particular, it invokes several advanced results from perturbation theory for linear programs). By contrast, even though we are working with a continuous action space, all policy computations in our case reduce to solving finite-dimensional eigenvalue problems, without any need for additional

randomization. Moreover, even though the overall scheme of our analysis is similar to the one in [13] (which, in turn, is inspired by existing work on lazy strategies [25], [27], [26]), the proof is self-contained and much less technical, relying on our new results pertaining to Todorov-type optimal control laws.

A preliminary version of this work has appeared in a conference publication [29], and most of the proofs were omitted due to space limitations. Since a major part of our contribution is a set of probabilistic analysis techniques for MDPs with Kullback–Leibler control cost (in both online and offline settings), the present paper fills in the missing details. In addition, most of our new results on the sensitivity of Todorov-type optimal controllers to perturbations of state costs were omitted from [29]. The present paper not only gives a self-contained treatment of these results, but also demonstrates their crucial role in performance analysis of online strategies for Todorov-type MDPs. Finally, compared to [29], the present paper reports a more thorough and improved empirical evaluation of our proposed strategy in the context of target tracking on a large graph. In particular, we report the results of Monte-Carlo simulation of our strategy (with error bars) and compare it to two baseline strategies: (a) the best stationary policy that could be chosen with full prior knowledge of the state cost sequence and (b) the best stationary policy chosen from a large pool of randomly sampled policies *without* advance knowledge of state costs. In [29], we only compared our strategy to the passive dynamics P^*. The experimental results reported here show that (a) the regret of our strategy w.r.t. the best stationary policy chosen in hindsight is nonnegative and grows sublinearly with time (thus validating our theoretical bound), and (b) in simulations, our strategy performs strictly better than any randomly sampled stationary policy.

C. Organization of the paper

The remainder of the paper is organized as follows. We close this section with a brief summary of frequently used notation. Section II contains precise formulation of the online MDP problem and presents our main result, Theorem 1. In preparation for the proof of the theorem, Section III contains preliminaries on MDPs with KL control cost [14], [15], [16], including a number of new results pertaining to optimal policies. Section IV then describes our proposed strategy, whose performance is then analyzed in Section V in order to prove Theorem 1. Some simulation results are presented in Section VI. We close by summarizing our contributions and outlining some directions for future work. Proofs of all intermediate results are relegated to the Appendix.

D. Notation

We will denote the underlying finite state space by X. A matrix $P = [P(x,y)]_{x,y\in\mathsf{X}}$ with nonnegative entries, and with the rows and the columns indexed by the elements of X, is called *stochastic* (or *Markov*) if its rows sum to one: $\sum_{y\in\mathsf{X}} P(x,y) = 1, \forall x \in \mathsf{X}$.

We will denote the set of all such stochastic matrices by $\mathcal{M}(\mathsf{X})$, the set of all probability distributions over X by $\mathcal{P}(\mathsf{X})$, the set of all functions $f : \mathsf{X} \to \mathbb{R}$ by $\mathcal{C}(\mathsf{X})$, and the cone of all nonnegative functions $f : \mathsf{X} \to \mathbb{R}_+$ by $\mathcal{C}_+(\mathsf{X})$. We will represent the elements of $\mathcal{P}(\mathsf{X})$ by row vectors and denote them by π, μ, ν, etc., and the elements of $\mathcal{C}(\mathsf{X})$ by column vectors and denote them by f, g, h, etc. The total variation (or L_1) distance between $\mu, \nu \in \mathcal{P}(\mathsf{X})$ is

$$\|\mu - \nu\|_1 \triangleq \sum_{x\in\mathsf{X}} |\mu(x) - \nu(x)|.$$

The *Kullback–Leibler divergence* (or *relative entropy*) [18] between μ and ν is

$$D(\mu\|\nu) \triangleq \begin{cases} \sum_{x\in\mathsf{X}} \mu(x) \log \dfrac{\mu(x)}{\nu(x)} & \text{if } \operatorname{supp}(\mu) \subseteq \operatorname{supp}(\nu) \\ +\infty & \text{otherwise} \end{cases}$$

where $\operatorname{supp}(\mu) \triangleq \{x \in \mathsf{X} : \mu(x) > 0\}$ is the *support* of μ. Here and in the sequel, we work with natural logarithms. The *span seminorm* (also called the *oscillation*) of $f \in \mathcal{C}(\mathsf{X})$ is defined as

$$\|f\|_s \triangleq \max_{x\in\mathsf{X}} f(x) - \min_{x\in\mathsf{X}} f(x).$$

Note that $\|f\|_s = 0$ if and only if $f(x) = c$ for some constant $c \in \mathbb{R}$ and all $x \in \mathsf{X}$; $\|f\|_s = \|f + c\|_s$ for any $f \in \mathcal{C}(\mathsf{X})$ and $c \in \mathbb{R}$. We also define the *sup norm* $\|f\|_\infty \triangleq \max_{x\in\mathsf{X}} |f(x)|$ and note that $\|f\|_s \le 2\|f\|_\infty$.

Any Markov matrix $P \in \mathcal{M}(\mathsf{X})$ acts on probability distributions from the right and on functions from the left:

$$\mu P(y) = \sum_{x\in\mathsf{X}} \mu(x) P(x,y), \qquad Pf(x) = \sum_{y\in\mathsf{X}} P(x,y) f(y).$$

We say that P is *unichain* [30] if the corresponding Markov chain has a single recurrent class of states (plus a possibly empty transient class). The is equivalent to P having a unique invariant distribution π_P (i.e. $\pi_P P = \pi_P$) [31]. We will denote the set of all such Markov matrices over X by $\mathcal{M}_1(\mathsf{X})$. Given $\rho \in [0,1]$, we say that P is *ρ-contractive* if

$$\|\mu P - \nu P\|_1 \le \rho \|\mu - \nu\|_1, \qquad \forall \mu, \nu \in \mathcal{P}(\mathsf{X})$$

(in fact, every $P \in \mathcal{M}(\mathsf{X})$ is 1-contractive). We will denote the set of ρ-contractive Markov matrices by $\mathcal{M}_1^\rho(\mathsf{X})$. It is easy to show that, for every $0 \leq \rho < 1$, $\mathcal{M}_1^\rho(\mathsf{X}) \subset \mathcal{M}_1(\mathsf{X})$. The *Dobrushin ergodicity coefficient* [31], [32] of $P \in \mathcal{M}(\mathsf{X})$ is given by

$$\alpha(P) \triangleq \frac{1}{2} \max_{x,x' \in \mathsf{X}} \|P(x,\cdot) - P(x',\cdot)\|_1,$$

and it can be shown that any $P \in \mathcal{M}(\mathsf{X})$ is $\alpha(P)$-contractive [31], [32]. Finally, for any $P, P' \in \mathcal{M}(\mathsf{X})$ we define the supremum distance

$$\|P - P'\|_\infty \triangleq \max_{x \in \mathsf{X}} \|P(x,\cdot) - P'(x,\cdot)\|_1.$$

II. Problem formulation and the main result

A. The model

Given the finite state space X, let $\mathcal{F}$ be a fixed subset of $\mathcal{C}_+(\mathsf{X})$, and let $x_1 \in \mathsf{X}$ be a fixed initial state. Consider an agent (A) performing a controlled random walk on X in response to a dynamic environment (E). The interaction between A and E proceeds as follows:

$X_1 = x_1$
for $t = 1, 2, \ldots$
  A selects $P_t \in \mathcal{M}(\mathsf{X})$ and draws $X_{t+1} \sim P_t(X_t, \cdot)$
  E selects $f_t \in \mathcal{F}$ and announces it to A
end for

At each $t \geq 1$, A selects the transition probabilities $P_t(x,y) = \Pr\{X_{t+1} = y | X_t = x\}$ based on his knowledge $f^{t-1} = (f_1, \ldots, f_{t-1})$, and incurs the *state cost* $f_t(X_t)$ and the *control cost* $D(P_t(X_t,\cdot)\|P^*(X_t,\cdot))$. The total cost incurred by the agent A at time t is given by

$$c_t(X_t, P_t) = f_t(X_t) + D(P_t(X_t,\cdot)\|P^*(X_t,\cdot)).$$

and the objective is to minimize a suitable notion of regret.

B. Strategies and regret

A *strategy* for the agent A is a collection of mappings $\gamma = \{\gamma_t\}_{t=1}^\infty$ where $\gamma_t : \mathcal{F}^{t-1} \to \mathcal{M}(\mathsf{X})$, so that $P_t = \gamma_t(f^{t-1})$. This means our strategy is based on the complete knowledge of all the

past cost functions. The cumulative cost of γ after T steps is

$$C_T = \sum_{t=1}^{T} c_t(X_t, P_t) = \sum_{t=1}^{T} c_t(X_t, \gamma_t(f^{t-1})).$$

To define the regret after T steps, we will consider the gap between C_T and the expected cumulative cost that A could have achieved in hindsight by using a stationary unichain random walk on X (with full knowledge of f^T). This gap arises through the agent's lack of prior knowledge on the sequence of state cost functions. Formally, we define the regret of γ after T steps w.r.t. a particular $P \in \mathcal{M}_1(\mathsf{X})$ by[1]

$$R_T(P) \triangleq C_T - \mathbb{E}_{x_1}^{P}\left[\sum_{t=1}^{T} c_t(X_t, P)\right],$$

where the expectation is taken over the Markov chain induced the by *comparison* transition kernel P with initial state $X_1 = x_1$. In this work, we make the following basic assumption concerning the environment E:

Assumption 0 (Oblivious environment)**.** *The environment* E *is* oblivious *(or* nonadversarial*), i.e., for every t, f_t depends only on f^{t-1}, but not on X^t.*

Assumption 0 is standard in the literature on sequential prediction [8] (in particular, it is also imposed by Yu et al. [13]). In our case, it implies that, for a fixed sequence $f_1, f_2, \ldots$ of state costs chosen by E, the state process $\boldsymbol{X} = \{X_t\}_{t=1}^{\infty}$ induced by A's choices $P_1, P_2, \ldots$ is a (time-inhomogeneous) Markov chain. Now consider some set $\mathcal{N} \subset \mathcal{M}_1(\mathsf{X})$. Adopting standard terminology [8], we will say that γ is *Hannan-consistent* w.r.t. $\mathcal{N}$ if

$$\limsup_{T\to\infty} \sup_{P\in\mathcal{N}} \sup_{f_1,\ldots,f_T\in\mathcal{F}} \frac{\mathbb{E}R_T(P)}{T} \leq 0, \tag{2}$$

where the expectation is w.r.t. the law of the process $\boldsymbol{X}$ starting at $X_1 = x_1$. In other words, a strategy is Hannan-consistent if its worst case (over $\mathcal{F}$) expected per-round regret converges to zero uniformly over $\mathcal{N}$. While it is certainly true that some nonstationary policy with complete prior knowledge of the state cost sequence may (and will) outperform any stationary policy, we limit our consideration to stationary reference policies for two reasons. One is the need to have a fair comparison: indeed, no truly online strategy could compete with the best (i.e., omniscient)

[1]To keep the notation clean, we will suppress the dependence of the cumulative cost C_T and the regret R_T on the strategy γ and on the state costs $f_1, \ldots, f_T$.

nonstationary policy. The other is that we can alternatively interpret the Hannan consistency condition (2) as follows: as the horizon T increases, the smallest average cost achievable by a strategy which is Hannan-consistent w.r.t. $\mathcal{N}$ will converge to the smallest long-term average cost achievable by any stationary Markov strategy in $\mathcal{N}$ on an MDP with the state cost given by the *empirical average* $(1/T)\sum_{t=1}^{T} f_t$ of the state costs revealed up to time T.

C. The main result

Our main result (Theorem 1 below) guarantees the existence of a Hannan-consistent strategy against any uniformly ergodic collection of stationary unichain policies under the following two assumptions on the passive dynamics P^*:

Assumption 1 (Irreducibility and aperiodicity)**.** *The passive dynamics P^* is* irreducible *and* aperiodic*, where the former means that, for every $x, y \in \mathsf{X}$, there exists some $n \in \mathbb{N}$ such that $(P^*)^n(x, y) > 0$, while the latter means that, for every $x \in \mathsf{X}$, the greatest common divisor of the set $\{n \in \mathbb{N} : (P^*)^n(x, x) > 0\}$ is equal to* 1.

Assumption 2 (Ergodicity)**.** *The Dobrushin ergodicity coefficient $\alpha(P^*)$ is strictly less than* 1.

Assumption 1 ensures that P^* has a unique everywhere positive invariant distribution π^* [31] and, for a finite X, it is equivalent to the existence of some $\bar{n} \in \mathbb{N}$, such that

$$\theta \triangleq \min_{x,y \in \mathsf{X}} (P^*)^{\bar{n}}(x, y) > 0$$

(see, e.g., Theorem 1.4 in [31]). Assumption 2, which is frequently used in the study of MDPs with average cost criterion [33], [34], [3], guarantees that the convergence to π^* is exponentially fast (so that P^* is geometrically ergodic), and it also imposes a stronger type of ergodicity, since a Markov matrix $P \in \mathcal{M}(\mathsf{X})$ has $\alpha(P) < 1$ if and only if for any pair $x, x' \in \mathsf{X}$ there exists at least one $y \in \mathsf{X}$, such that y can be reached from both x and x' in one step with strictly positive probability. For example, if P^* satisfies the *Doeblin minorization condition* [32], [35], i.e., if there exist some $\delta \in (0, 1]$ and some $\mu \in \mathcal{P}(\mathsf{X})$, such that $P^*(x, y) \geq \delta\mu(y)$ for all $x, y \in \mathsf{X}$, then we will have $\alpha(P^*) \leq 1 - \delta < 1$ (see, e.g., Lemma 4.3.13 in [32]). Moreover, it is not hard to show that any Markov matrix $P \in \mathcal{M}(\mathsf{X})$ can be approximated arbitrarily well by another Markov matrix P' with $\alpha(P') < 1$.

With these assumptions in place, we are now ready to state our main result:

Theorem 1. *Let $\mathcal{F}$ consist of all $f \in \mathcal{C}_+(\mathsf{X})$ with $\|f\|_\infty \leq 1$. Fix an arbitrary $\epsilon \in (0, 1/3)$. Under Assumptions 0–2, there exists a strategy γ, such that for any $\rho \in [0, 1)$,*

$$\sup_{P \in \mathcal{M}_1^\rho(\mathsf{X})} \sup_{f_1,\ldots,f_T \in \mathcal{F}} \frac{\mathbb{E}R_T(P)}{T} = O(T^{-1/4+\epsilon}). \tag{3}$$

As a consequence, the strategy γ is Hannan-consistent w.r.t. $\mathcal{M}_1^\rho(\mathsf{X})$.

Remark 1. The constant hidden in the $O(\cdot)$ notation depends only on the passive dynamics P^* and on the contraction rate ρ of the baseline policies in $\mathcal{M}_1^\rho(\mathsf{X})$; cf. Eq. (22), and the discussion preceding it, for details.

III. Preliminaries

Our construction of a Hannan-consistent strategy in Theorem 1 relies on Todorov's theory of MDPs with KL control cost [14], [15], [16]. In this section, we give an overview of this theory and present several new results that will be used later on.

First, let us recall the general set-up for MDPs with finite state space X and compact action space U under the average cost criterion (see e.g., [2] or [3]). It involves a family of Markov matrices $P_u \in \mathcal{M}(\mathsf{X})$ indexed by actions $u \in \mathsf{U}$. The (long-term) *average cost* of a stationary Markov policy (state feedback law) $w : \mathsf{X} \to \mathsf{U}$ with initial state $X_1 = x_1$ is given by

$$J(w, x_1) \triangleq \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{x_1}^w \left[\sum_{t=1}^{T} c(X_t, w(X_t)) \right], \tag{4}$$

where the expectation $\mathbb{E}_{x_1}^w[\cdot]$ is w.r.t. the law of the Markov chain $\boldsymbol{X} = \{X_t\}$ with controlled transition probabilities

$$\Pr\{X_{t+1} = y | X_t = x\} = P_{w(x)}(x, y), \qquad X_1 = x_1$$

and $c : \mathsf{X} \times \mathsf{U} \to \mathbb{R}_+$ is the one step state-action cost. The construction of an optimal policy to minimize (4) for every x_1 revolves around the *average-cost optimality equation* (ACOE)

$$h(x) + \lambda = \min_{u \in \mathsf{U}(x)} \{c(x, u) + P_u h(x)\}, \ x \in \mathsf{X} \tag{5}$$

where $\mathsf{U}(x) \subseteq \mathsf{U}$ is the set of allowable actions in state x. If a solution pair $(\lambda, h) \in \mathbb{R}_+ \times \mathcal{C}(\mathsf{X})$ exists with $\|h\|_s < +\infty$, then it can be shown [2], [3] that the stationary policy

$$w_*(x) = \arg\min_{u \in \mathsf{U}(x)} \{c(x, u) + P_u h(x)\}$$

is optimal, and has average cost λ for every x. The function h is called the *relative value function*.

A. *Linearly solvable MDPs*

In a series of papers [14], [15], [16], Todorov has introduced a class of Markov decision processes, for which solving the ACOE reduces to solving an eigenvalue problem. In this set-up, which we have described informally in Section I-A, the action space U is the probability simplex $\mathcal{P}(\mathsf{X})$, which is compact in the Euclidean topology, and for each $u \in \mathcal{P}(\mathsf{X})$ we have $P_u(x,y) \triangleq u(y), \forall (x,y) \in \mathsf{X} \times \mathsf{X}$. Thus, any state feedback law (Markov policy) $w : \mathsf{X} \to \mathcal{P}(\mathsf{X})$ induces the state transitions directly via

$$\Pr\{X_{t+1} = y | X_t = x\} = P_{w(x)}(x,y) = [w(x)](y), \quad t \geq 1.$$

In other words, if $X_t = x$, then $u(\cdot) = w(x)$ is the probability distribution of the next state X_{t+1}. Hence, there is a one-to-one correspondence between Markov policies w and Markov matrices $P \in \mathcal{M}(\mathsf{X})$, given by $w(x) = P(x,\cdot)$.

To specify an MDP, we fix a state cost function $f \in \mathcal{C}_+(\mathsf{X})$ and a Markov matrix P^* as the passive dynamics, which specifies the state transition probabilities in the absence of control. The one-step state-action cost function $c(x,u)$ is given by (1). If we use the shorthand $c(x,P)$ for $c(x,P(x,\cdot))$, then the average cost of a policy $P \in \mathcal{M}(\mathsf{X})$ starting at $X_1 = x_1$ can be written as

$$J(P,x_1) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^P_{x_1} \left[\sum_{t=1}^{T} c(X_t, P) \right].$$

Intuitively, if P has a small average cost, then the induced Markov chain $\boldsymbol{X} = \{X_t\}$ has a small average state cost, and its one-step transitions stay close to those prescribed by P^*.

The ACOE for this problem takes the form

$$h(x) + \lambda = f(x) + \min_{u \in \mathcal{P}(\mathsf{X})} \{ D(u \| P^*(x,\cdot)) + \mathbb{E}_u h \}. \tag{6}$$

For a given $h \in \mathcal{C}(\mathsf{X})$, the minimization of the right-hand side of (6) can be done in closed form. To see this, let us define, for every $\varphi \in \mathcal{C}(\mathsf{X})$, the *twisted kernel* [36]

$$\check{P}_\varphi(x,\cdot) \triangleq \frac{P^*(x,\cdot) e^{-\varphi(\cdot)}}{P^* e^{-\varphi}(x)}, \qquad x \in \mathsf{X} \tag{7}$$

which is obviously an element of $\mathcal{M}(\mathsf{X})$. Then we have

$$\begin{aligned}
&\min_{u\in\mathcal{P}(\mathsf{X})}\{D(u\|P^*(x,\cdot))+\mathbb{E}_u h\}\\
&=\min_{u\in\mathcal{P}(\mathsf{X})}\left\{\mathbb{E}_u\left[\log\frac{u(Y)}{P^*(x,Y)}+h(Y)\right]\right\}\\
&=\min_{u\in\mathcal{P}(\mathsf{X})}\left\{\mathbb{E}_u\left[\log\frac{u(Y)}{\check{P}_h(x,Y)}\right]-\log P^*e^{-h}(x)\right\} \qquad (8)
\end{aligned}$$

If we further define $\Lambda_h(x) \triangleq P^*e^{-h}(x)$, then the quantity in braces in (8) can be written as $D(u\|\check{P}_h(x,\cdot)) - \log\Lambda_h(x)$. Using the fact that the divergence $D(\mu\|\nu)$ between any two $\mu,\nu \in \mathcal{P}(\mathsf{X})$ is nonnegative and equal to zero if and only if $\mu = \nu$ [18], we see that the minimum value in (8) is uniquely achieved by $u_*(x) = \check{P}_h(x,\cdot)$ and is equal to $-\log\Lambda_h(x)$. Thus, we can rewrite the ACOE (6) as

$$h(x)+\lambda = f(x) - \log\Lambda_h(x), \qquad \forall x\in\mathsf{X}. \qquad (9)$$

If we now consider the *exponentiated* relative value function $V \triangleq e^{-h}$, then (9) can be also written as $e^{-f}P^*V(x) = e^{-\lambda}V(x)$. Expressing this in vector form, we obtain the so-called *multiplicative Poisson equation* (MPE) [36]:

$$e^{-f}P^*V = e^{-\lambda}V \qquad (10)$$

To construct the optimal policy for our MDP, we first solve the MPE (10) for λ and V, obtain h, and then compute the twisted kernel $\check{P}_h(x,\cdot)$ for every $x \in \mathsf{X}$. The MPE is an instance of a so-called *Frobenius–Perron eigenvalue* (FPE) problem [31]; there exist efficient methods for solving such problems, e.g., a recent algorithm due to Chanchana [37]. We also should point out that, for each $x \in \mathsf{X}$, the twisted kernel (7) is a *Boltzmann–Gibbs distribution* on the state space X with energy function h and base measure $P^*(x,\cdot)$. Boltzmann–Gibbs distributions arise in various contexts, e.g., in statistical physics and in the theory of large deviations [38], [39], as solutions of variational problems over the space of probability measures that involve minimization of a Gibbs-type free energy functional, consisting of an affine "energy" term and a convex "entropy" term (given by the divergence relative to the base measure). Indeed, the functional being minimized on the right-hand side of (6) is precisely of this form.

In the sequel, we will often need to consider simultaneously several MDPs with different state costs f. Thus, whenever need arises, we will indicate the dependence on f using appropriate

subscripts, as in c_f, λ_f, h_f, V_f, etc. For instance, the MPE (10) for a given state cost f is

$$P_f^* V_f = e^{-\lambda_f} V_f, \tag{11}$$

where $P_f^* \triangleq e^{-f} P^*$, i.e., $P_f^*(x,y) = e^{-f(x)} P^*(x,y)$ for all $x, y \in \mathsf{X}$.

B. Some properties of Todorov's optimal policy

We now investigate the properties of Todorov's optimal policy under the assumptions on the passive dynamics P^* that are listed in Section II-C. Most of the results of this section are new (with some exceptions, which we point out explicitly); the proofs are given in the Appendix.

We start with the following basic existence and uniqueness result, which is implicit in [14]:

Proposition 1. *Under Assumption 1, for any state cost $f \in \mathcal{C}_+(\mathsf{X})$ the MPE (11) has a strictly positive solution $V_f \in \mathcal{C}_+(\mathsf{X})$ with the associated strictly positive eigenvalue $e^{-\lambda_f}$, and the only nonnegative solutions of (11) are positive multiples of V_f. Moreover, the corresponding twisted kernel $\check{P}_{h_f}$ is also irreducible and aperiodic, and has a unique invariant distribution $\check{\pi}_f = \check{\pi}_f \check{P}_f \in \mathcal{P}(\mathsf{X})$.*

Proof: Appendix A. ■

Since $V_f = e^{-h_f}$, the fact that any positive multiple of V_f is a solution of the MPE is equivalent to the well-known fact that the relative value function h_f as a solution of the ACOE (6) is unique up to additive constants. That is, if a particular h_f solves (6), then so does any $h_f + c$ for any additive constant $c \in \mathbb{R}$. For this reason, we can fix an arbitrary $x^\circ \in \mathsf{X}$ and assume that $h_f(x^\circ) = 0$ for any f. This ensures that the mapping

$$f \longmapsto h_f, \qquad h_f(x^\circ) = 0 \tag{12}$$

is well-defined. The following results are new:

Proposition 2. *Under Assumption 1, the mapping (12) is bounded on compact subsets of the cone $\mathcal{C}_+(\mathsf{X})$: for any $f \in \mathcal{C}_+(\mathsf{X})$,*

$$\|h_f\|_s \le \log \theta^{-1} + \bar{n} \|f\|_\infty, \tag{13}$$

where $\bar{n}$ and θ are defined in Section II-C. Hence,

$$\sup_{f \in \mathcal{C}_+(\mathsf{X});\, \|f\|_\infty \le C} \|h_f\|_s \le \log \theta^{-1} + \bar{n} C. \tag{14}$$

Proof: Appendix B. ∎

Moreover, the dependence of the relative value function h_f on the state cost f is *continuous*:

Proposition 3. *Under Assumptions 1 and 2, the mapping (12) is Lipschitz-continuous on compact subsets of $\mathcal{C}_+(\mathsf{X})$: for every $C > 0$ there exists a constant $K = K(C) > 0$, such that for any two $f, g \in \mathcal{C}_+(\mathsf{X})$ with $\|f\|_\infty, \|g\|_\infty \leq C$ we have*

$$\|h_f - h_g\|_s \leq K \|f - g\|_\infty. \tag{15}$$

Proof: Appendix C. ∎

More generally, the twisted kernel $\check{P}_\varphi$ depends smoothly on the "twisting function" φ:

Proposition 4. *Fix any two functions $\varphi, \varphi' \in \mathcal{C}(\mathsf{X})$. Then the twisted kernels (7) have the following properties: for any $x \in \mathsf{X}$,*

$$D(\check{P}_\varphi(x,\cdot) \| \check{P}_{\varphi'}(x,\cdot)) \leq \frac{1}{8} \|\varphi - \varphi'\|_s^2 \tag{16}$$

$$\|\check{P}_\varphi(x,\cdot) - \check{P}_{\varphi'}(x,\cdot)\|_1 \leq \frac{1}{2} \|\varphi - \varphi'\|_s. \tag{17}$$

Moreover, if Assumptions 1 and 2 hold, then there exists a mapping $\kappa : \mathbb{R}_+ \to [0,1)$, such that

$$\|\varphi\|_s \leq C \quad \Longrightarrow \quad \alpha(\check{P}_\varphi) \leq \kappa(C). \tag{18}$$

Proof: Appendix D. ∎

We close with the following basic but important result on steady-state optimality:

Proposition 5. *For any $f \in \mathcal{C}_+(\mathsf{X})$ and any $P \in \mathcal{M}_1(\mathsf{X})$, define*

$$\bar{J}_f(P) \triangleq \mathbb{E}_{\pi_P}[c_f(X,P)] \equiv \mathbb{E}_{\pi_P}[J_f(P,X)].$$

Then

$$\bar{J}_f(\check{P}_{h_f}) = \inf_{P \in \mathcal{M}_1(\mathsf{X})} \bar{J}_f(P).$$

Proof: Appendix E. ∎

IV. The proposed strategy

Our construction of a Hannan-consistent strategy for the problem of Section II is similar to the approach of Yu et al. [13]. The main idea behind it is as follows. We partition the set of

time indices $1, 2, \ldots$ into nonoverlapping contiguous segments (phases) of increasing duration and, during each phase, use Todorov's optimal policy matched to the average of the state cost functions revealed during the preceding phases. As in [13], the phases are sufficiently long to ensure convergence to the steady state within each phase, and yet are sufficiently short, so that the policies used during successive phases are reasonably close to one another.

The phases are indexed by $m \in \mathbb{N}$, where we denote the mth phase by $\mathcal{T}_m$ and its duration by τ_m. Given $\epsilon \in (0, 1/3)$, we let $\tau_m = \lceil m^{1/3-\epsilon} \rceil$. The parameter ϵ is needed to control the growth of the total length of each fixed number of phases relative to the length of the most recent phase (we comment upon this in more detail in the next section). We also define $\mathcal{T}_{1:m} \triangleq \mathcal{T}_1 \cup \ldots \cup \mathcal{T}_m$ (the union of phases 1 through m) and denote its duration by $\tau_{1:m}$. Given a sequence $\{f_t\}$ of state cost functions, we define for each m the average state costs

$$\widehat{f}^{(m)} \triangleq \frac{1}{\tau_m} \sum_{t \in \mathcal{T}_m} f_t, \qquad \widehat{f}^{(1:m)} \triangleq \frac{1}{\tau_{1:m}} \sum_{t \in \mathcal{T}_{1:m}} f_t$$

and let $\widehat{f}^{(0)} = \widehat{f}^{(1:0)} = 0$. Our strategy takes the following form:

for $m = 1, 2, \ldots$
  solve the MPE $e^{-\widehat{f}^{(1:m-1)}} P^* e^{-h^{(m)}} = e^{-\lambda^{(m)}} e^{-h^{(m)}}$
  let $P^{(m)} = \check{P}_{h^{(m)}}$
  for $t \in \mathcal{T}_m$
    draw $X_{t+1} \sim P^{(m)}(X_t, \cdot)$
  end for
end for

Since we use the same policy throughout each phase, the evolution of the state induced by the above algorithm is described by the following inhomogeneous Markov chain:

$$X_1 \xrightarrow{P^{(1)}} X_2 \xrightarrow{P^{(1)}} \ldots \xrightarrow{P^{(1)}} X_{\tau_1} \xrightarrow{P^{(2)}} X_{\tau_1+1} \xrightarrow{P^{(2)}} \ldots$$

The implementation of this strategy reduces to solving a finite-dimensional Frobenius–Perron eigenvalue (FPE) problem [31] at the beginning of each phase to obtain a Todorov-type relative value function. The corresponding twisted kernel then determines the stationary policy to be followed throughout that phase. An efficient method for solving FPE problems was recently developed by Chanchana [37]. This method makes use of the well-known Collatz formula

for the FPE [31] and Elsner's inverse iteration algorithm for computing the spectral radius of a nonnegative irreducible matrix [40]. It is an iterative algorithm, which at each iteration performs an LU factorization of an $|\mathsf{X}| \times |\mathsf{X}|$ matrix. The time complexity of each iteration is $O(|\mathsf{X}|^3)$. Chanchana's algorithm outperforms the three best known algorithms for solving FPE problems, which all rely on Elsner's inverse iteration and have quadratic convergence. Numerical experimental results can be found in [37, Section 3.5].

V. Proof of Theorem 1

A. The main idea

Following the general outline in [13], the proof of Theorem 1 can be divided into four major steps. The first step is to show that there is no loss of generality in considering a different notion of regret, i.e., the *steady-state regret*, which is the difference between the cumulative cost of the proposed strategy and the steady-state cost of a fixed stationary policy. The second step is to bound the difference between the expected total cost of our strategy and the sum of expected steady-state costs within each phase. That is, for each m, the steady-state expectation of the cost incurred in phase m is taken w.r.t. the unique invariant distribution of $P^{(m)}$. After this step, we may only concentrate on expectations over invariant state distributions, which renders the problem much easier. For the third step, we show that the sum of steady-state expected costs is not much worse than what we would get if, at the start of each phase m, we also knew all the state cost functions to be revealed during phase m, i.e., if we used the "clairvoyant" policy $P^{(m+1)}$ in phase m. In the fourth step, we consider the sum of expected costs in each phase that could be attained if we knew all the state cost functions in advance and used the optimal policy w.r.t. the average of all the state cost functions throughout all the phases. We show that this expected cost is actually greater than the sum of expected costs of each phase when we only know the state cost functions one phase ahead. We then assemble the bounds obtained in these four steps to obtain the final bound on the regret of our strategy.

B. Preliminary lemmas

Before proceeding to the proof of Theorem 1, we present two lemmas that will be used throughout. The proofs of the lemmas rely heavily on the results of Section III-B, and are detailed in Appendices F and G.

Lemma 1 (Uniform bounds)**.** *There exists constants $K_0 \geq 0$, $K_1 \geq 0$ and $0 \leq \alpha < 1$, such that, for every $f \in \mathcal{F}$ and every $m \in \mathbb{N}$,*

$$\|c_f(\cdot, P^{(m)})\|_\infty \leq K_0, \quad \|h^{(m)}\|_s \leq K_1, \quad \alpha(P^{(m)}) \leq \alpha$$

Moreover, the bound $\|c_f(\cdot, P)\|_\infty \leq K_0$ holds for all $P \in \mathcal{M}_1(\mathsf{X})$, such that $D(P(x,\cdot)\|P^*(x,\cdot)) < \infty$ for all $x \in \mathsf{X}$.

Lemma 2 (Policy continuity)**.** *There exists a constant $K_2 \geq 0$, such that, for every $m \in \mathbb{N}$,*

$$\|P^{(m+1)}(x,\cdot) - P^{(m)}(x,\cdot)\|_1 \leq \frac{K_2 \tau_m}{\tau_{1:m}}, \tag{19}$$

and

$$\|\pi^{(m+1)} - \pi^{(m)}\|_1 \leq \frac{K_2 \tau_m}{(1-\alpha)\tau_{1:m}} \tag{20}$$

where $\pi^{(m)}$ is the unique invariant distribution of $P^{(m)}$. Moreover, there exists a constant $K_3 \geq 0$, such that for $D^{(m)}(x) \triangleq D(P^{(m)}(x,\cdot)\|P^*(x,\cdot)), \forall x \in \mathsf{X}$, we have

$$\|D^{(m)}(x) - D^{(m+1)}(x)\|_1 \leq \frac{K_3 \tau_m}{\tau_{1:m}}. \tag{21}$$

Remark 2. As will be evident from the proof below, we can specify the precise form of the regret bound in (3) using the constants from the above lemmas:

$$\begin{aligned} &\sup_{P \in \mathcal{M}_1^\rho(\mathsf{X})} \sup_{f_1,\ldots,f_T \in \mathcal{F}} \frac{\mathbb{E}R_T(P)}{T} \\ &\leq \frac{4}{3}\left(\frac{K_0(K_2+2)}{1-\alpha} + K_0 + K_3\right) T^{-1/4+\epsilon} + \frac{2K_0}{(1-\rho)T}. \end{aligned} \tag{22}$$

C. Details

We are now ready to present the detailed proof of Theorem 1.

Step 1: Reduction to the steady-state case. For any $P \in \mathcal{M}_1(\mathsf{X})$, let us define the *steady-state regret* of our strategy γ w.r.t. P by

$$R_T^{\mathrm{ss}}(P) \triangleq C_T - \mathbb{E}_{\pi_P}\left[\sum_{t=1}^T c_t(X,P)\right],$$

which is the difference between the actual cumulative cost of γ and the steady-state cost of the stationary unichain policy P initialized with π_P. Now let us fix some $\rho \in [0,1)$ and consider an arbitrary $P \in \mathcal{M}_1^\rho(\mathsf{X})$, where without loss of generality we can assume $D(P(x,\cdot)\|P^*(x,\cdot)) < \infty$

for all $x \in \mathsf{X}$. For each $t \geq 1$, let $\nu_t = \delta_{x_1} P^{t-1}$ be the distribution of X_t in the Markov chain induced by the transition matrix P and initial state $X_1 = x_1$. For any T, we have

$$\begin{aligned}
&|R_T^{\mathrm{ss}}(P) - R_T(P)| \\
&= \left| \mathbb{E}_{x_1}^P \left[\sum_{t=1}^T c_t(X_t, P) \right] - \mathbb{E}_{\pi_P} \left[\sum_{t=1}^T c_t(X, P) \right] \right| \\
&= \left| \sum_{t=1}^T \{ \mathbb{E}_{\nu_t}[c_t(X_t, P)] - \mathbb{E}_{\pi_P}[c_t(X, P)] \} \right| \\
&\leq \sum_{t=1}^T \|c_t(\cdot, P)\|_\infty \|\nu_t - \pi_P\|_1 \\
&\leq 2K_0 \sum_{t=1}^T \rho^{t-1} \leq \frac{2K_0}{1-\rho}, \qquad (23)
\end{aligned}$$

where the second inequality is by Lemma 1 and the fact that $P \in \mathcal{M}_1^\rho(\mathsf{X})$. Therefore, it suffices to show that the bound in (3) holds with $\mathbb{E}R_T^{\mathrm{ss}}(P)$ in place of $\mathbb{E}R_T(P)$.

Step 2: Steady-state approximation within phases. In this step, we approximate the cumulative cost within each phase by its steady-state value. Let M denote the number of complete phases up to time T, i.e. $\tau_{1:M} \leq T < \tau_{1:M+1}$ (simple algebra gives $M \leq (4/3)T^{3/4+\epsilon}$). Then we can decompose the total cost as

$$\begin{aligned}
C_T &= \sum_{t=1}^{\tau_{1:M}} c_t(X_t, P_t) + \sum_{t=\tau_{1:M}+1}^{T} c_t(X_t, P_t) \\
&\leq \sum_{t=1}^{\tau_{1:M}} c_t(X_t, P_t) + K_0 \tau_{M+1} = C_{\tau_{1:M}} + K_0 \tau_{M+1},
\end{aligned}$$

where the inequality is by Lemma 1. Since all state costs are nonnegative by hypothesis,

$$\mathbb{E}_{\pi_P} \left[\sum_{t=1}^T c_t(X, P) \right] \geq \mathbb{E}_{\pi_P} \left[\sum_{t=1}^{\tau_{1:M}} c_t(X, P) \right],$$

which implies that

$$R_T^{\mathrm{ss}}(P) \leq R_{\tau_{1:M}}^{\mathrm{ss}}(P) + K_0 \tau_{M+1}. \qquad (24)$$

For every time step t, let μ_t be the state distribution induced by our strategy when starting from initial state distribution $\mu_1 = \delta_{x_1}$. Note that the transition matrix at time t is $P_t = P^{(m)}$ if $t \in \mathcal{T}_m$.

We can decompose the expected cost in the first M phases as

$$\mathbb{E}C_{\tau_{1:M}} = \sum_{m=1}^{M} \sum_{t \in \mathcal{T}_m} \mathbb{E}_{\mu_t} \left[c_t(X, P^{(m)}) \right], \tag{25}$$

and for every $t \in \mathcal{T}_m$ we have

$$\begin{aligned}
&\mathbb{E}_{\mu_t} \left[c_t(X, P^{(m)}) \right] \\
&\leq \mathbb{E}_{\pi^{(m)}} \left[c_t(X, P^{(m)}) \right] + \| c_t(\cdot, P^{(m)}) \|_\infty \| \mu_t - \pi^{(m)} \|_1 \\
&\leq \mathbb{E}_{\pi^{(m)}} \left[c_t(X, P^{(m)}) \right] + K_0 \| \mu_t - \pi^{(m)} \|_1,
\end{aligned}$$

where the last step is by Lemma 1. In addition, for every $k \in \{0, 1, \ldots, \tau_m - 1\}$, we have

$$\begin{aligned}
&\| \mu_{\tau_{1:m-1}+k+1} - \pi^{(m)} \|_1 \\
&= \left\| \mu_{\tau_{1:m-1}+1} (P^{(m)})^k - \pi^{(m)} (P^{(m)})^k \right\|_1 \\
&\leq \alpha^k \| \mu_{\tau_{1:m-1}+1} - \pi^{(m)} \|_1 \leq 2\alpha^k,
\end{aligned}$$

where the first inequality is due to Lemma 1. Hence,

$$\begin{aligned}
&\sum_{t \in \mathcal{T}_m} \mathbb{E}_{\mu_t} \left[c_t(X, P^{(m)}) \right] \\
&\leq \sum_{t \in \mathcal{T}_m} \mathbb{E}_{\pi^{(m)}} \left[c_t(X, P^{(m)}) \right] + 2K_0 \sum_{k=0}^{\tau_m - 1} \alpha^k \\
&\leq \sum_{t \in \mathcal{T}_m} \mathbb{E}_{\pi^{(m)}} \left[c_t(X, P^{(m)}) \right] + \frac{2K_0}{1-\alpha}.
\end{aligned}$$

Substituting this into (25), we have

$$\mathbb{E}C_{\tau_{1:M}} \leq \sum_{m=1}^{M} \sum_{t \in \mathcal{T}_m} \mathbb{E}_{\pi^{(m)}} \left[c_t(X, P^{(m)}) \right] + \frac{2K_0 M}{1-\alpha}.$$

Step 3: Looking one phase ahead. In this step, we show that the steady-state cost in each phase is not much worse than what we could get if we knew everything one phase ahead. For

every $m \in \{1, \ldots, M\}$, we have

$$\begin{aligned}
&\sum_{t \in \mathcal{T}_m} \mathbb{E}_{\pi^{(m)}} \left[c_t(X, P^{(m)}) \right] \\
&\leq \sum_{t \in \mathcal{T}_m} \mathbb{E}_{\pi^{(m+1)}} \left[c_t(X, P^{(m)}) \right] + K_0 \tau_m \| \pi^{(m+1)} - \pi^{(m)} \|_1 \\
&\leq \tau_m \mathbb{E}_{\pi^{(m+1)}} \left[\widehat{f}^{(m)} + D^{(m)} \right] + \frac{K_0 K_2 \tau_m^2}{(1-\alpha)\tau_{1:m}} \\
&= \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) + \tau_m \mathbb{E}_{\pi^{(m+1)}} \left[D^{(m)} - D^{(m+1)} \right] \\
&\qquad + \frac{K_0 K_2 \tau_m^2}{(1-\alpha)\tau_{1:m}} \\
&\leq \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) + \left(\frac{K_0 K_2}{1-\alpha} + K_3 \right) \frac{\tau_m^2}{\tau_{1:m}},
\end{aligned}$$

where the first inequality is by Lemma 1, the second inequality is by Lemma 2, and the last inequality is due to (21) in Lemma 2. So we now have

$$\begin{aligned}
\mathbb{E} C_{\tau_{1:M}} \leq &\sum_{m=1}^{M} \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) \\
&+ \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3 \right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2 K_0 M}{1-\alpha}.
\end{aligned} \tag{26}$$

Step 4: Looking M phases ahead. In this step, we consider the fictitious situation where we know everything M phases ahead, and show that the resulting steady-state value is actually greater than what we could get if we knew everything just one phase ahead. In other words, we claim that

$$\sum_{m=1}^{M} \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) \leq \sum_{m=1}^{M} \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(M+1)}). \tag{27}$$

To see that this claim is true, we apply backward induction:

$$\begin{aligned}
&\sum_{m=1}^{M} \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(M+1)}) \\
&= \sum_{m=1}^{M-1} \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(M+1)}) + \tau_M \bar{J}_{\widehat{f}^{(M)}}(P^{(M+1)}) \\
&= \tau_{1:M-1} \bar{J}_{\widehat{f}^{(1:M-1)}}(P^{(M+1)}) + \tau_M \bar{J}_{\widehat{f}^{(M)}}(P^{(M+1)}) \\
&\geq \tau_{1:M-1} \bar{J}_{\widehat{f}^{(1:M-1)}}(P^{(M)}) + \tau_M \bar{J}_{\widehat{f}^{(M)}}(P^{(M+1)}) \\
&= \sum_{m=1}^{M-1} \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(M)}) + \tau_M \bar{J}_{\widehat{f}^{(M)}}(P^{(M+1)}),
\end{aligned}$$

where the second equality is due to the fact that $\tau_{1:M-1}\widehat{f}^{(1:M-1)} = \sum_{t\in\mathcal{T}_{1:M-1}} f_t = \sum_{m=1}^{M} \tau_m \widehat{f}^{(m)}$, while the inequality is by Proposition 5 and the fact that $P^{(M)} = \check{P}_{h^{(M)}}$, where $h^{(M)}$ is the relative value function for state cost $\widehat{f}^{(1:M-1)}$. Repeating this argument, we obtain (27). Moreover,

$$\begin{aligned}
\sum_{m=1}^{M} \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(M+1)}) &= \tau_{1:M} \bar{J}_{\widehat{f}^{(1:M)}}(P^{(M+1)}) \\
&= \tau_{1:M} \inf_{P\in\mathcal{M}_1(\mathsf{X})} \bar{J}_{\widehat{f}^{(1:M)}}(P) \\
&= \inf_{P\in\mathcal{M}_1(\mathsf{X})} \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right]. \qquad (28)
\end{aligned}$$

After these four steps, we are finally in a position to bound the expected steady-state regret. Combining (26)–(28), we can write

$$\begin{aligned}
\mathbb{E}C_{\tau_{1:M}} \leq \inf_{P\in\mathcal{M}_1(\mathsf{X})} \mathbb{E}_{\pi_P}&\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right] \\
&+ \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3\right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}R^{\mathrm{ss}}_{\tau_{1:M}}(P) &= \mathbb{E}C_{\tau_{1:M}} - \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right] \\
&\leq \mathbb{E}C_{\tau_{1:M}} - \inf_{P\in\mathcal{M}_1(\mathsf{X})} \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{\tau_{1:M}} c_t(X,P)\right] \\
&\leq \sum_{m=1}^{M} \left(\frac{K_0 K_2}{1-\alpha} + K_3\right) \frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha}. \qquad (29)
\end{aligned}$$

Next we show that the right-hand side of (29) can be bounded by a quantity that is sublinear in T. From (24), we have

$$\begin{aligned}\mathbb{E}R_T^{\text{ss}}(P) &\leq \mathbb{E}R_{\tau_{1:M}}^{\text{ss}}(P) + K_0\tau_{M+1} \\ &\leq \sum_{m=1}^{M}\left(\frac{K_0K_2}{1-\alpha}+K_3\right)\frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0M}{1-\alpha} + K_0\tau_{M+1}.\end{aligned}$$

Due to our construction of the phases, $M \leq (4/3)T^{3/4+\epsilon}$ and $\tau_{M+1} \leq M$ if $M > 1$. Moreover, it is a matter of routine but tedious algebraic calculations to show that the choice $\tau_m = \lceil m^{\frac{1}{3}-\epsilon}\rceil$ for $m = 1, \ldots, M$ for any $\epsilon \in (0, 1/3)$ is sufficient to guarantee that $\tau_m^2 \leq \sqrt{\tau_{1:m}}$. Thus, we obtain

$$\begin{aligned}\mathbb{E}R_T^{\text{ss}}(P) &\leq \sum_{m=1}^{M}\left(\frac{K_0K_2}{1-\alpha}+K_3\right)\frac{\tau_m^2}{\tau_{1:m}} + \frac{2K_0M}{1-\alpha} + K_0M \\ &\leq M\left(\frac{K_0(K_2+2)}{1-\alpha}+K_0+K_3\right) \\ &\leq \frac{4}{3}\left(\frac{K_0(K_2+2)}{1-\alpha}+K_0+K_3\right)T^{3/4+\epsilon}.\end{aligned}$$

Therefore, recalling (23), we finally obtain

$$\begin{aligned}&\frac{\mathbb{E}R_T(P)}{T} \\ &\leq \frac{\mathbb{E}R_T^{\text{ss}}(P)}{T} + \frac{2K_0}{T(1-\rho)} \\ &\leq \frac{4}{3}\left(\frac{K_0(K_2+2)}{1-\alpha}+K_0+K_3\right)T^{-1/4+\epsilon} + \frac{2K_0}{T(1-\rho)},\end{aligned}$$

which completes the proof of Theorem 1.

VI. Simulations

In this section, we demonstrate the performance of our proposed strategy on a simulated problem involving online (real-time) tracking of a moving target on a large, connected, undirected graph G, which models a terrain with obstacles. The state space is the set of all vertices (nodes) of G. The target is executing a stationary random walk on G with a randomly sampled transition probability matrix, which is different from the one that governs the passive dynamics P^*. The motion of both the tracking agent and the target must conform to the topology of G, in the sense that both can only move between neighboring vertices. The graph used in our simulation has 564 vertices.

To make sure that Assumptions 1 and 2 are satisfied, we construct the passive dynamics in the form $P^* = (1-\delta)P_1 + \delta P_0$ for some $\delta \in (0,1)$. Here, P_1 is a random walk that represents environmental constraints, allowing the agent to go from a given node either to any adjacent node (with equal probability) or to remain at the current location. To ensure that the agent is sufficiently mobile, the probability of not moving is chosen to be relatively small (in our case, 0.01) compared to the probability of transitioning to any of the neighboring nodes. Since the underlying graph is connected, the random walk P_1 is irreducible; it is also aperiodic since $P_1(x,x) > 0$ for all vertices x. We also add a perturbation random walk P_0, which has a fixed column of ones (we can think of the node indexing that column as a "home base" for the agent), and zeros elsewhere. The "size" of the perturbation is controlled by δ, which is set to be small (we have chosen $\delta = 0.01$), so the agent only has a slight chance of returning to "home base" from any given node within one step. This perturbation ensures that no two rows of P^* are orthogonal, and $\alpha(P^*) \leq 1 - \delta = 0.99$.

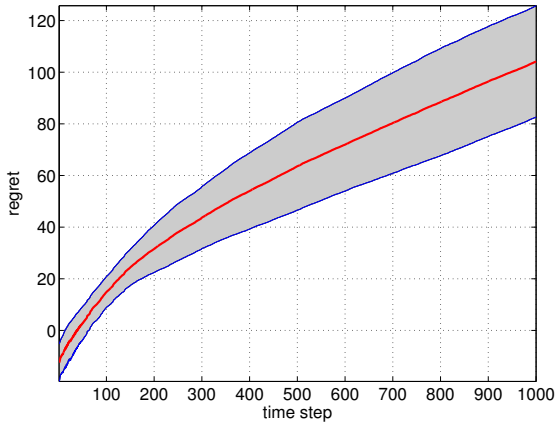

Fig. 1. Regret versus time. The red curve shows the average of the regret (the difference between the total cost of our strategy up to each time t and the total cost of the best stationary policy up to that time) over 100 independent realizations of the simulation. At each time t, the height of the gray region corresponds to one sample standard deviation.

The simulation consists of a number of independent experiments. Each individual experiment runs for $T = 1000$ time steps. We first randomly sample a transition matrix for the target motion. After simulating the target's random walk for T steps, we record the target locations and use them to generate a sequence of state cost functions $\{f_t\}_{t=1}^T$. Then we feed these 1000 state cost functions sequentially to our online algorithm and compute the resulting cumulative cost C_T.

At each time t, the tracking agent is in state (location) x_t, the target is at location s_t, and the agent's action is P_t. The cumulative cost after T time steps is

$$C_T = \sum_{t=1}^{T} \left[f_t(x_t) + D\left(P_t(x_t, \cdot) \| P^*(x_t, \cdot)\right) \right], \tag{30}$$

with state costs $f_t(x_t) = d_G(x_t, s_t)$, where $d_G(\cdot, \cdot)$ is the graph distance (number of edges in the shortest path) between the agent's current location and the location of the target, normalized by the diameter of G. Then we compute the best stationary policy P in hindsight for the average of all the state costs by solving the MPE

$$e^{-\widehat{f}} P^* e^{-h} = e^{-\lambda} e^{-h}$$

for the relative value function h, where $\widehat{f} = \frac{1}{T}\sum_{t=1}^{T} f_t$, and then setting

$$P(x, \cdot) = \frac{P^*(x, \cdot) e^{-h(\cdot)}}{P^* e^{-h}(x)}, \qquad x \in \mathsf{X}$$

The regret is then computed with respect to the steady-state cost of this best stationary policy:

$$R_T(P) = C_T - \mathbb{E}_{\pi_P}\left[\sum_{t=1}^{T} c_t(X, P) \right],$$

where

$$c_t(X, P) = f_t(X) + D\left(P(X, \cdot) \| P^*(X, \cdot)\right),$$

and π_P is the unique invariant distribution of P.

To plot the regret versus time with error bars, we implement the experiment 100 times and compute the empirical average of the regret across experiments. For each realization the agent was initialized with the same starting state. The evolution of the regret versus time is shown in Figure 1, where the regret at time t is defined as the total cost of our strategy up to time t minus the total cost of the best stationary policy up to time t. We can see that the regret is growing sublinearly, as stated in Theorem 1.

We also compare the total cost of our strategy to that of the best stationary baseline policy among a set $\tilde{\mathcal{N}}$ of 10^5 randomly sampled stationary policies. Once again, each experiment runs for $T = 1000$ time steps. The baseline policy P_{baseline} is the one that has the smallest total cost

$$C_T(P_{\text{baseline}}) = \min_{P \in \tilde{\mathcal{N}}} \sum_{t=1}^{T} \left[f_t(x_t) + D\left(P(x_t, \cdot) \| P^*(x_t, \cdot)\right) \right].$$

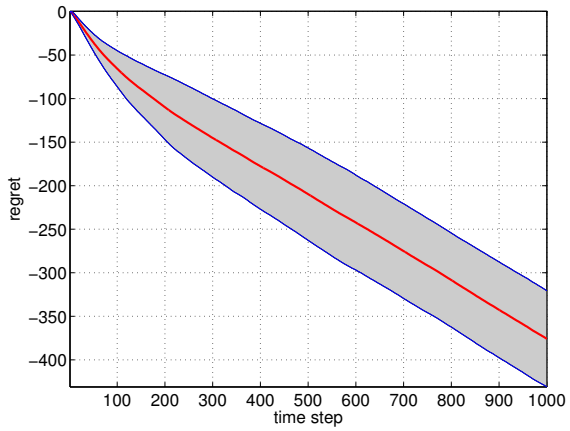

Fig. 2. Comparison of our proposed strategy to the best stationary policy in a set of 10^5 randomly sampled policies. The red curve shows the average of the regret (the difference between the total cost of our strategy up to each time t and the total cost of the best stationary policy up to that time) over 100 independent realizations of the simulation. At each time t, the height of the gray area corresponds to one sample standard deviation.

among the 10^5 randomly sampled policies. The regret of our adaptive strategy is thus given by $C_T - C_T(P_{\text{baseline}})$.

As before, there are 100 independent experiments, where in each experiment the agent using our strategy and the agent using the best sampled stationary policy were initialized with the same starting state. The evolution of the regret versus time is shown in Figure 2, where the regret at time t is defined as the total cost of our strategy up to time t minus the total cost of the best sampled stationary policy up to time t. We can see that the regret is negative, which implies that our strategy outperforms the best sampled stationary policy for each particular realization of the state cost sequence.

VII. Conclusion and Future Work

The problem studied in this paper combines aspects of both stochastic control and online learning. In particular, our construction of a Hannan-consistent strategy (a concept from the theory of online learning [8]) uses several ideas and techniques from the theory of MDPs with average cost criterion, including some new results concerning optimal policies for MDPs with KL control costs [14], [15], [16].

We have proved that, for any horizon T, our strategy achieves sublinear $O(T^{3/4})$ regret relative to any uniformly ergodic class of stationary policies, which is similar to the results of Yu et

al. [13] for online MDPs with finite state and action spaces. However, while our strategy (like that of [13]) is computationally efficient, we believe that the $O(T^{3/4})$ scaling of regret with T is suboptimal. Indeed, in the case when both the state and the action spaces are finite, Even-Dar et al. [12] present a strategy that achieves a much better $O(\sqrt{T})$ regret. Of course, the strategy of [12] involves recomputing the policy at *every* time step (rather than in phases, as is done here and in [13]), which results in a significant loss of efficiency. An interesting open question, which we plan to address in our future work, is whether it is possible to attain $O(\sqrt{T})$ regret for online MDPs with KL control costs. A related challenge is to study these online MDPs in the (nonstochastic) bandit setting, where at each time step the agent only learns the value $f_t(X_t)$ of the state cost at time t at the current state X_t, rather than the full state cost function $f_t \in \mathcal{C}(\mathsf{X})$. While this bandit setting is more realistic, very little is known about it even for online MDPs with finite state and action spaces — Neu et al. [41] constructed a strategy that achieves $O(T^{2/3}(\log T)^{1/3})$ regret, but it is not known whether this is optimal.

Another promising avenue for further research has to do with the apparent duality between our set-up and the theory of *risk-sensitive control* of Markov processes [42], [43]. Indeed, the ACOE (6) can be viewed as a special case of the Isaacs equation for a certain dynamic two-player game with average cost criterion, in which Player 1 generates state cost functions, while Player 2 generates distributions over the state space (cf., e.g., [43, p. 1805]). In the set-up of our Section II, Player 1 would correspond to the environment E, while Player 2 would be the agent A. We plan to explore this connection further.

Finally, as mentioned in the Introduction, we would like to extend our results to more general (e.g., compact) state spaces. This will require more sophisticated machinery, e.g., Foster–Lyapunov criteria and ergodicity w.r.t. weighted norms [30], [35], as well as spectral theory of the MPE for Markov chains with general state spaces [44].

APPENDIX

A. *Proof of Proposition 1*

Consider the matrix $P_f^* \triangleq e^{-f}P^*$ with entries $P_f^*(x,y) = e^{-f(x)}P^*(x,y)$. For any $n \in \mathbb{N}$ and any $x, y \in \mathsf{X}$, $(P_f^*)^n(x,y) \geq e^{-n\|f\|_\infty}(P^*)^n(x,y)$. Since P^* is irreducible (Assumption 1), for any pair $x, y \in \mathsf{X}$ of states there exists some $n \in \mathbb{N}$, such that $(P^*)^n(x,y) > 0$. But then

$(P_f^*)^n(x,y) > 0$ as well, which means that P_f^* is also irreducible. Therefore, by the Frobenius–Perron theorem [31], P_f^* has a strictly positive right eigenvector V_f with a positive eigenvalue r (the Frobenius-Perron eigenvalue): $P_f^* V_f = r V_f$. Thus, $e^{-\lambda_f} = r$. Moreover, the FP eigenvalue is simple, and P_f^* has no nonnegative right eigenvectors other than the positive multiples of V_f [31]. This proves the existence and uniqueness part.

Now, using the fact that $V_f = e^{-h_f}$ solves the MPE (11), we can show that

$$\check{P}_{h_f}(x,y) = e^{\lambda_f} \frac{V_f(y)}{V_f(x)} P_f^*(x,y),$$

whence it follows that

$$(\check{P}_{h_f})^n(x,y) = e^{n\lambda_f^*} \frac{V_f(y)}{V_f(x)} (P_f^*)^n(x,y),$$

As was just proved, P_f^* is irreducible, and V_f is strictly positive. Hence, for any pair $(x,y) \in \mathsf{X}\times\mathsf{X}$ there exists some $n \in \mathbb{N}$, such that $(\check{P}_{h_f})^n(x,y) > 0$ as well. This proves the irreducibility of $\check{P}_{h_f}$. Now, since P_f^* is irreducible, the Frobenius–Perron theorem says that there exists a unique strictly positive $\mu \in \mathcal{P}(\mathsf{X})$, such that $\mu P_f^* = e^{-\lambda_f}\mu$ [31]. Now define $\check{\pi}_f \in \mathcal{P}(\mathsf{X})$ through

$$\check{\pi}_f(x) \triangleq \frac{\mu(x)V_f(x)}{\sum_{y\in\mathsf{X}} \mu(y)V_f(y)} \equiv \frac{\mu(x)V_f(x)}{\mathbb{E}_\mu V_f}, \qquad x \in \mathsf{X}.$$

A straightforward calculation shows that $\check{\pi}_f$ is an invariant distribution of $\check{P}_{h_f}$. The uniqueness of $\check{\pi}_f$ follows from the irreducibility of $\check{P}_{h_f}$.

B. Proof of Proposition 2

We essentially follow the proof of Theorem 3.2 in [43], with some simplifications. For each $T \in \mathbb{N}$, define the function $W_T : \mathsf{X} \to \mathbb{R}$ via

$$e^{-W_T(x)} \triangleq \mathbb{E}_x \left[\exp\left(-\sum_{t=1}^{T} f(X_t) - h_f(X_{T+1}) \right) \right],$$

where $\mathbb{E}_x[\cdot]$ denotes the expectation w.r.t. the Markov chain $\boldsymbol{X} = (X_1, X_2, \ldots)$ with initial state $X_1 = x$ and transition matrix P^*. Then a simple inductive argument shows that

$$e^{-W_T(x)} = e^{-T\lambda_f - h_f(x)}. \tag{31}$$

Indeed, for each t let $\Psi_t \triangleq \prod_{s=1}^{t} \frac{V_f(X_s)}{P^* V_f(X_s)}$. Then, since $e^{-f(x)} = \frac{e^{-\lambda_f} V_f(x)}{P^* V_f(x)}$ by (11), we can write

$$e^{-W_T(x)} = e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_T V_f(X_{T+1}) \right] \tag{32}$$

$$= e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_T \mathbb{E}[V_f(X_{T+1}) | X_T] \right] \tag{33}$$

$$= e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_T P^* V_f(X_T) \right] \tag{34}$$

$$= e^{-T\lambda_f} \mathbb{E}_x \left[\Psi_{T-1} V_f(X_T) \right], \tag{35}$$

where (32) follows from definitions, (33) and (34) use the Markov property, and (35) again follows from definitions. Proceeding backwards, we get

$$\mathbb{E}_x[\Psi_T V_f(X_{T+1})] = \mathbb{E}_x[\Psi_1 V_f(X_2)] = \mathbb{E}_x[\Psi_1 P^* V_f(X_1)]$$
$$= V_f(x) = e^{-h_f(x)}.$$

Substituting this into (32), we get (31), which in turn implies that $h_f(x) = W_T(x) - T\lambda_f$ for all $x \in \mathsf{X}, T \in \mathbb{N}$. Since $h_f(x^\circ) = 0$, we can write $h_f(x) = W_T(x) - W_T(x^\circ), \forall x \in \mathsf{X}, T \in \mathbb{N}$. Let ν (respectively, ν°) be the distribution of $X_{\bar{n}+1}$ in the Markov chain with transition matrix P^* and initial state $X_1 = x$ (respectively, $X_1 = x^\circ$). Then

$$\frac{\nu(y)}{\nu^\circ(y)} = \frac{(P^*)^{\bar{n}}(x, y)}{\nu^\circ(y)} \geq \theta > 0 \tag{36}$$

for every $y \in \mathsf{X}$. Consequently, for any $T > \bar{n}$ we have

$$e^{-W_T(x)}$$

$$= \mathbb{E}_x \left[e^{-\sum_{t=1}^{\bar{n}} f(X_t)} e^{-\sum_{t=\bar{n}+1}^{T} f(X_t) - h_f(X_{T+1})} \right] \tag{37}$$

$$\geq e^{-\bar{n}\|f\|_\infty} \mathbb{E}_x \left[e^{-\sum_{t=\bar{n}+1}^{T} f(X_t) - h_f(X_{T+1})} \right] \tag{38}$$

$$= e^{-\bar{n}\|f\|_\infty} \mathbb{E}_{x^\circ} \left[e^{-\sum_{t=\bar{n}+1}^{T} f(X_t) - h_f(X_{T+1})} \frac{\nu(X_{\bar{n}+1})}{\nu^\circ(X_{\bar{n}+1})} \right] \tag{39}$$

$$\geq \theta e^{-\bar{n}\|f\|_\infty} \mathbb{E}_{x^\circ} \left[e^{-\sum_{t=1}^{T} f(X_t) - h_f(X_{T+1})} \right] \tag{40}$$

$$= \theta e^{-\bar{n}\|f\|_\infty} e^{-W_T(x^\circ)}, \tag{41}$$

where (37) is by definition, (39) follows from the Markov property and a change of measure, (40) follows from (36) and from the fact that $f \geq 0$, and (41) is again by definition. Taking logarithms, we get $W_T(x) - W_T(x^\circ) \leq \log \theta^{-1} + \bar{n}\|f\|_\infty, \forall T > \bar{n}$. Interchanging the roles of x and x°, we get $|h_f(x)| \leq \log \theta^{-1} + \bar{n}\|f\|_\infty$. This proves (13); (14) follows immediately.

C. Proof of Proposition 3

The basic idea is as follows. For a given $f \in \mathcal{C}_+(\mathsf{X})$, let us introduce the dynamic programming operator $\mathbb{T}_f$ that maps any $\varphi \in \mathcal{C}(\mathsf{X})$ to $\mathbb{T}_f\varphi \in \mathcal{C}(\mathsf{X})$, where $\forall \varphi \in \mathcal{C}(\mathsf{X}), x \in \mathsf{X}$,

$$\mathbb{T}_f\varphi(x) \triangleq f(x) + \inf_{\mu \in \mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu \varphi + D(\mu \| P^*(x, \cdot))\}.$$

Then we can express the ACOE (6) as $h_f + \lambda_f = \mathbb{T}_f h_f$. Hence, for any $f, g \in \mathcal{C}_+(\mathsf{X})$,

$$\begin{aligned} \|h_f - h_g\|_s &= \|(\mathbb{T}_f h_f - \lambda_f) - (\mathbb{T}_g h_g - \lambda_g)\|_s \\ &= \|\mathbb{T}_f h_f - \mathbb{T}_g h_g\|_s \qquad (42) \\ &\leq \|\mathbb{T}_f h_f - \mathbb{T}_g h_f\|_s + \|\mathbb{T}_g h_f - \mathbb{T}_g h_g\|_s, \qquad (43) \end{aligned}$$

where (42) uses the fact that the span seminorm is unchanged after adding a constant, and (43) is by the triangle inequality. We will then show the following:

1) For any $\varphi \in \mathcal{C}(\mathsf{X})$ and any $f, g \in \mathcal{C}_+(\mathsf{X})$,

$$\|\mathbb{T}_f\varphi - \mathbb{T}_g\varphi\|_s \leq 2\|f - g\|_\infty. \qquad (44)$$

2) For a fixed $f \in \mathcal{C}_+(\mathsf{X})$, the dynamic programming operator $\mathbb{T}_f : \mathcal{C}(\mathsf{X}) \to \mathcal{C}(\mathsf{X})$ is a contraction in the span seminorm: for every $M > 0$, there exists a constant $K' = K'(M) \in (0, 1)$, such that for any two $\varphi, \varphi' \in \mathcal{C}(\mathsf{X})$ with $\|\varphi\|_s, \|\varphi'\|_s \leq M$ we have

$$\|\mathbb{T}_f\varphi - \mathbb{T}_f\varphi'\|_s \leq K'\|\varphi - \varphi'\|_s. \qquad (45)$$

Assuming that items 1) and 2) above are proved, we proceed as follows. First of all, the first term in (43) is bounded by $2\|f - g\|_\infty$ by (44). Next, since $\|f\|_\infty, \|g\|_\infty \leq C$, Proposition 2 guarantees that there exists some $M = M(C) < \infty$, such that $\|h_f\|_s, \|h_g\|_s \leq M$. Therefore, there exists a constant $K' = K'(M) < 1$, such that the second term in (43) is bounded by $K'\|h_f - h_g\|_s$. Therefore, $\|h_f - h_g\|_s \leq \frac{2}{1-K'}\|f - g\|_\infty$, which gives (15) with $K = 2/(1 - K')$.

We now prove 1) and 2). For any function $\varphi \in \mathcal{C}(\mathsf{X})$ and any two $f, g \in \mathcal{C}_+(\mathsf{X})$, we have

$$\begin{aligned}
&\max_{x\in\mathsf{X}} \{\mathbb{T}_f\varphi(x) - \mathbb{T}_g\varphi(x)\} \\
&= \max_{x\in\mathsf{X}} \left\{ \left[f(x) + \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x,\cdot))\} \right] \right. \\
&\qquad\qquad \left. - \left[g(x) + \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x,\cdot))\} \right] \right\} \\
&= \max_{x\in\mathsf{X}} [f(x) - g(x)] .
\end{aligned}$$

Similarly, we get $\min_{x\in\mathsf{X}} \{\mathbb{T}_f\varphi(x) - \mathbb{T}_g\varphi(x)\} = \min_{x\in\mathsf{X}} [f(x) - g(x)]$. Thus, $\|\mathbb{T}_f\varphi - \mathbb{T}_g\varphi\|_s = \|f - g\|_s \le 2\|f - g\|_\infty$, so we have proved (44).

To establish (45), we follow the proof of Proposition 2.2 in [34] with some simplifications. Pick any $x, x' \in \mathsf{X}$ and let

$$\begin{aligned}
\nu &= \arg\min_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi' + D(\mu\|P^*(x,\cdot))\} , \\
\nu' &= \arg\min_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x',\cdot))\} ,
\end{aligned}$$

where explicitly $\nu(\cdot) = \check{P}_{\varphi'}(x,\cdot)$ and $\nu'(\cdot) = \check{P}_{\varphi}(x',\cdot)$. Then

$$\begin{aligned}
&[\mathbb{T}_f\varphi(x) - \mathbb{T}_f\varphi'(x)] - [\mathbb{T}_f\varphi(x') - \mathbb{T}_f\varphi'(x')] \\
&= \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x,\cdot))\} \\
&\qquad - \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi' + D(\mu\|P^*(x,\cdot))\} \\
&\qquad - \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi + D(\mu\|P^*(x',\cdot))\} \\
&\qquad + \inf_{\mu\in\mathcal{P}(\mathsf{X})} \{\mathbb{E}_\mu\varphi' + D(\mu\|P^*(x',\cdot))\} \\
&\le \mathbb{E}_\nu\varphi + D(\nu\|P^*(x,\cdot)) - \mathbb{E}_\nu\varphi' - D(\nu\|P^*(x,\cdot)) \\
&\qquad - \mathbb{E}_{\nu'}\varphi - D(\nu'\|P^*(x',\cdot)) + \mathbb{E}_{\nu'}\varphi' + D(\nu'\|P^*(x',\cdot)) \\
&= \int (\varphi - \varphi')\mathrm{d}(\nu - \nu').
\end{aligned}$$

A standard argument shows that that $\int(\varphi-\varphi')\mathrm{d}(\nu-\nu') \leq \frac{1}{2}\|\varphi-\varphi'\|_s\|\nu-\nu'\|_1$. Consequently,

$$\begin{aligned}&\|\mathbb{T}_f\varphi - \mathbb{T}_f\varphi'\|_s \\ &\leq \frac{1}{2}\|\varphi-\varphi'\|_s \cdot \max_{x,x'\in\mathsf{X}} \left\|\check{P}_\varphi(x,\cdot) - \check{P}_{\varphi'}(x',\cdot)\right\|_1.\end{aligned}$$

Then the proof of (45) will be complete if we can show that

$$\begin{aligned}K'(M) &\triangleq \frac{1}{2}\sup_{\varphi,\varphi';\|\varphi\|_s,\|\varphi'\|_s\leq M}\max_{x,x'\in\mathsf{X}}\left\|\check{P}_\varphi(x,\cdot)-\check{P}_{\varphi'}(x',\cdot)\right\|_1 \\ &< 1. \end{aligned} \tag{46}$$

Suppose that (46) does not hold. Then there exist sequences $\{\varphi_n\}$, $\{\varphi'_n\}$ of functions with $\|\varphi_n\|_s, \|\varphi'_n\|_s \leq M, \forall n$, a set $B\subset\mathsf{X}$, and a pair of points $x,x'\in\mathsf{X}$, such that

$$\lim_{n\to\infty}\left[\check{P}_{\varphi_n}(x,B) - \check{P}_{\varphi'_n}(x',B)\right] = 1,$$

where for any $P\in\mathcal{M}(\mathsf{X})$ we denote $P(x,B)\triangleq\sum_{y\in B}P(x,y)$. This implies in turn that

$$\lim_{n\to\infty}\check{P}_{\varphi_n}(x,\mathsf{X}\backslash B) = \lim_{n\to\infty}\check{P}_{\varphi'_n}(x',B) = 0. \tag{47}$$

Since $\check{P}_\varphi(x,B)\geq e^{-\|\varphi\|_s}P^*(x,B)$, (47) implies that $P^*(x,\mathsf{X}\backslash B) = P^*(x',B) = 0$. But this means that $P^*(x,B)-P^*(x',B)=1$, which contradicts Assumption 2. Hence, (46) holds.

D. Proof of Proposition 4

We begin with (16). From definitions, we have

$$\begin{aligned}&D(\check{P}_\varphi(x,\cdot)\|\check{P}_{\varphi'}(x,\cdot)) \\ &\qquad = \mathbb{E}_{\check{P}_\varphi(x,\cdot)}[\varphi'(Y)-\varphi(Y)] + \log\frac{\Lambda_{\varphi'}(x)}{\Lambda_\varphi(x)}.\end{aligned} \tag{48}$$

A simple change-of-measure calculation shows that

$$\begin{aligned}\frac{\Lambda_{\varphi'}(x)}{\Lambda_\varphi(x)} &= \frac{\sum_y P^*(x,y)e^{-\varphi'(y)}}{\Lambda_\varphi(x)} \\ &= \frac{\sum_y e^{\varphi(y)-\varphi'(y)}P^*(x,y)e^{-\varphi(y)}}{\Lambda_\varphi(x)} \\ &= \mathbb{E}_{\check{P}_\varphi(x,\cdot)}[e^{\varphi(Y)-\varphi'(Y)}].\end{aligned} \tag{49}$$

To bound the right-hand side of (49), we recall the well-known *Hoeffding bound* [45], which for our purposes can be stated as follows: For any $\mu \in \mathcal{P}(\mathsf{X})$ and any $\psi \in \mathcal{C}(\mathsf{X})$,

$$\log \mathbb{E}_\mu e^\psi \leq \mathbb{E}_\mu \psi + \frac{\|\psi\|_s^2}{8}.$$

Applying this bound gives

$$\log \frac{\Lambda_{\varphi'}(x)}{\Lambda_\varphi(x)} \leq \mathbb{E}_{\check{P}_\varphi(x,\cdot)}[\varphi(Y) - \varphi'(Y)] + \frac{\|\varphi - \varphi'\|_s^2}{8}.$$

Substituting this bound into (48), we see that the terms involving the expectation of the difference $\varphi - \varphi'$ cancel, and we are left with (16). To prove (17), we use Pinsker's inequality, $\|P_1 - P_2\|_1 \leq \sqrt{2D(P_1\|P_2)}$ [18]. To prove (18), we follow essentially the same strategy as in the proof of Proposition 3 to show that $\kappa(C) \triangleq \sup_{\varphi;\|\varphi\|_s \leq C} \alpha(\check{P}_\varphi) < 1$ for every $C > 0$.

E. Proof of Proposition 5

Fix some $P \in \mathcal{M}_1(\mathsf{X})$. If there exists some $x \in \mathsf{X}$ such that $\pi_P(x) > 0$ and $D(P(x,\cdot)\|P^*(x,\cdot)) = +\infty$, then Proposition 5 holds trivially. Thus, there is no loss of generality if we assume that $D(P(x,\cdot)\|P^*(x,\cdot)) < +\infty, \forall x \in \mathsf{X}$. Then

$$\begin{aligned}
\bar{J}_f(P) &= \mathbb{E}_{\pi_P}[f(X) + D(P(X,\cdot)\|P^*(X,\cdot))] \\
&= \sum_{x\in\mathsf{X}} \pi_P(x) \Bigg[f(x) + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{P(x,y)}{\check{P}_{h_f}(x,y)} \\
&\qquad + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{\check{P}_{h_f}(x,y)}{P^*(x,y)} \Bigg] \\
&= \sum_{x\in\mathsf{X}} \pi_P(x) \Bigg[f(x) + \mathbb{E}_{\pi_P} D(P(X,\cdot)\|\check{P}_{h_f}(X,\cdot)) \\
&\qquad + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{e^{-h_f(y)}}{\Lambda_{h_f}(x)} \Bigg] \\
&\geq \sum_{x\in\mathsf{X}} \pi_P(x) \Bigg[f(x) + \sum_{y\in\mathsf{X}} P(x,y) \log \frac{e^{-h_f(y)}}{\Lambda_{h_f}(x)} \Bigg] \\
&= \mathbb{E}_{\pi_P}[f(X) - Ph_f(X) - \log \Lambda_{h_f}(X)] \\
&= \mathbb{E}_{\pi_P}[f(X) - h_f(X) - \log \Lambda_{h_f}(X)],
\end{aligned}$$

where the inequality is due to the fact that the KL divergence is always nonnegative, and the last step is due to the fact that π_P is the invariant distribution of P. By the ACOE (9), we know that $f(x) - h_f(x) - \log \Lambda_{h_f}(x) = \lambda_f$ for every $x \in \mathsf{X}$. So we have $\bar{J}_f(P) \geq \lambda_f, \forall P \in \mathcal{M}_1(\mathsf{X})$. Note that if we take the expectation $\mathbb{E}_{\check{\pi}_f}[\cdot]$ of both sides of the ACOE (6), we get

$$\begin{aligned}
&\mathbb{E}_{\check{\pi}_f}[h_f(X) + \lambda_f] \\
&= \mathbb{E}_{\check{\pi}_f}[f(X) + D(\check{P}_{h_f}(X,\cdot)\|P^*(X,\cdot)) + \check{P}_{h_f} h_f(X)] \\
&= \mathbb{E}_{\check{\pi}_f}[f(X) + D(\check{P}_{h_f}(X,\cdot)\|P^*(X,\cdot))] + \mathbb{E}_{\check{\pi}_f}[h_f(X)],
\end{aligned}$$

where the last equality is due to the fact that $\check{\pi}_f$ is the invariant distribution of $\check{P}_{h_f}$. Therefore, $\mathbb{E}_{\check{\pi}_f}[f(X) + D(\check{P}_{h_f}(X,\cdot)\|P^*(X,\cdot))] = \bar{J}_f(\check{P}_{h_f}) = \lambda_f$. So we now have $\bar{J}_f(P) \geq \bar{J}_f(\check{P}_{h_f})$ for any $P \in \mathcal{M}_1(\mathsf{X})$, which completes the proof of Proposition 5.

F. Proof of Lemma 1

For every state $x \in \mathsf{X}$, let G_x denote the set of states that can be reached from x in one step by the passive dynamics P^*, i.e., $G_x \triangleq \{y : P^*(x,y) > 0\}$. Let us also define $p^*_x = \min_{y \in G_x} P^*(x,y)$ and $p^* = \min_{x \in \mathsf{X}} p^*_x$. Since $P^{(m)} = \check{P}_{h^{(m)}}$, and $h^{(m)}$ is bounded by Proposition 2, we have $\mathrm{supp}(P^{(m)}(x,\cdot)) \subseteq \mathrm{supp}(P^*(x,\cdot)) \equiv G_x$. Therefore,

$$\begin{aligned}
D(P^{(m)}(x,\cdot)\|P^*(x,\cdot)) &= \sum_{y \in G_x} P^{(m)}(x,y) \log \frac{P^{(m)}(x,y)}{P^*(x,y)} \\
&\leq \log \frac{1}{p^*_x}, \qquad \forall x \in \mathsf{X}, m \in \mathbb{N}
\end{aligned}$$

and for any $f \in \mathcal{F}$

$$\begin{aligned}
&\|c_f(\cdot, P^{(m)})\|_\infty \\
&\leq \|f\|_\infty + \max_{x \in \mathsf{X}} D(P^{(m)}(x,\cdot)\|P^*(x,\cdot)) \leq 1 + \log \frac{1}{p^*}.
\end{aligned}$$

Thus, the first bound of Lemma 1 holds with $K_0 = 1 + \log \frac{1}{p^*}$. The same argument works for any $P \in \mathcal{M}_1(\mathsf{X})$ that satisfies $D(P(x,\cdot)\|P^*(x,\cdot)) < \infty, \forall x \in \mathsf{X}$. The second bound holds by Proposition 2, where $K_1 = \log \theta^{-1} + \bar{n}$. The third bound follows from the second bound, $\|h^{(m)}\|_s \leq K_1$, and by Proposition 4 with $\alpha = \kappa(K_1)$.

G. Proof of Lemma 2

Let us recall that each $P^{(m)}$ is given by the twisted kernel $\check{P}_{h^{(m)}}$, where the relative value function $h^{(m)}$ arises from the solution of the MPE (11) with state cost $\widehat{f}^{(1:m-1)}$. Then

$$\begin{aligned}\|P^{(m+1)}(x,\cdot) - P^{(m)}(x,\cdot)\|_1 &\leq \frac{1}{2}\|h^{(m+1)} - h^{(m)}\|_s \\ &\leq \frac{K_2}{2}\|\hat{f}^{(1:m)} - \hat{f}^{(1:m-1)}\|_\infty \leq \frac{K_2\tau_m}{\tau_{1:m}}\end{aligned} \tag{50}$$

where the first step is by Proposition 4, the second by Proposition 3 with $K_2 = K(1)$, and the third by Lemma 4.3 in [13]. This proves (19). Moreover, Proposition 1 guarantees that $P^{(m)} = \check{P}_{h^{(m)}}$ has a unique invariant distribution $\pi^{(m)}$. Therefore,

$$\begin{aligned}&\|\pi^{(m)} - \pi^{(m+1)}\|_1 \\ &= \|\pi^{(m)}P^{(m)} - \pi^{(m+1)}P^{(m+1)}\|_1 \\ &\leq \|\pi^{(m)}P^{(m)} - \pi^{(m)}P^{(m+1)}\|_1 \\ &\qquad + \|\pi^{(m)}P^{(m+1)} - \pi^{(m+1)}P^{(m+1)}\|_1 \\ &\leq \|P^{(m)} - P^{(m+1)}\|_\infty + \alpha\|\pi^{(m)} - \pi^{(m+1)}\|_1 \\ &\leq \frac{K_2\tau_m}{\tau_{1:m}} + \alpha\|\pi^{(m)} - \pi^{(m+1)}\|_1,\end{aligned}$$

where the third inequality follows from (50). Rearranging, we get (20).

Next, from the form of $P^{(m)}$ and $P^{(m+1)}$ we have

$$\begin{aligned}&D^{(m)}(x) - D^{(m+1)}(x) \\ &= \mathbb{E}_x^{(m+1)}[h^{(m+1)}] - \mathbb{E}_x^{(m)}[h^{(m)}] + \log\frac{\Lambda_{h^{(m+1)}}(x)}{\Lambda_{h^{(m)}}(x)},\end{aligned} \tag{51}$$

where $\mathbb{E}_x^{(m)}[\cdot]$ denotes expectation w.r.t. $P^{(m)}(x,\cdot)$, and we can follow the same steps we have used in (49) to show that

$$\frac{\Lambda_{h^{(m+1)}}(x)}{\Lambda_{h^{(m)}}(x)} = \mathbb{E}_x^{(m)}\left[e^{h^{(m)} - h^{(m+1)}}\right].$$

Using the Hoeffding bound [45], we can write

$$\log\frac{\Lambda_{h^{(m+1)}}(x)}{\Lambda_{h^{(m)}}(x)} \leq \mathbb{E}_x^{(m)}[h^{(m)} - h^{(m+1)}] + \frac{\|h^{(m)} - h^{(m+1)}\|_s^2}{8} \tag{52}$$

Substituting (52) into (51) and simplifying, we get

$$\begin{aligned}
&D^{(m)}(x) - D^{(m+1)}(x) \\
&\leq \mathbb{E}_x^{(m+1)}[h^{(m+1)}] - \mathbb{E}_x^{(m)}[h^{(m+1)}] + \frac{1}{8}\|h^{(m)} - h^{(m+1)}\|_s^2 \\
&\leq \|h^{(m+1)}\|_s \cdot \|P^{(m)}(x,\cdot) - P^{(m+1)}(x,\cdot)\|_1 \\
&\qquad + \frac{1}{8}\|h^{(m)} - h^{(m+1)}\|_s^2 \\
&\leq \frac{K_1 K_2 \tau_m}{\tau_{1:m}} + \frac{1}{8}\|h^{(m)} - h^{(m+1)}\|_s^2 \\
&\leq \frac{K_1 K_2 \tau_m}{\tau_{1:m}} + \frac{K_2^2 \tau_m^2}{2\tau_{1:m}^2} \\
&\leq \left(K_1 K_2 + \frac{K_2^2}{2}\right)\frac{\tau_m}{\tau_{1:m}}.
\end{aligned}$$

Here, the third step uses the fact that $\|h^{(m)}\|_s \leq K_1$ (Lemma 2) and (50), the fourth also uses (50), and the last is due to the fact that $\frac{\tau_m}{\tau_{1:m}} < 1$. Letting $K_3 = K_1 K_2 + \frac{K_2^2}{2}$, we get (21).

REFERENCES

[1] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

[2] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

[3] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: a survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, 1993.

[4] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[5] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, pp. 185–202, 1994.

[6] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," in *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability 1950*. Berkeley, CA: University of California Press, 1951, pp. 131–148.

[7] J. Hannan, "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games*. Princeton Univ. Press, 1957, vol. 3, pp. 97–139.

[8] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games*. Cambridge Univ. Press, 2006.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[10] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, no. 55, pp. 527–535, 1952.

[11] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multiarmed bandit problems," *Foundations and Trends in Machine Learning*, 2012, to appear.

[12] E. Even-Dar, S. M. Kakade, and Y. Mansour, "Online Markov decision processes," *Math. Oper. Res.*, vol. 34, no. 3, pp. 726–736, 2009.

[13] J. Y. Yu, S. Mannor, and N. Shimkin, "Markov decision processes with arbitrary reward processes," *Math. Oper. Res.*, vol. 34, no. 3, pp. 737–757, 2009.

[14] E. Todorov, "Linearly-solvable Markov decision problems," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 1369–1376.

[15] ——, "General duality between optimal control and estimation," in *Proc. 47th IEEE Conf. on Decision and Control*, 2008, pp. 4286–4292.

[16] ——, "Efficient computation of optimal actions," *Proc. Nat. Acad. Sci.*, vol. 106, no. 28, pp. 11 478–11 483, 2009.

[17] B. Kappen, V. Gomez, and M. Opper, "Optimal control as a graphical model inference problem," *Machine Learning*, vol. 87, no. 2, pp. 159–182, 2012.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[19] M. Kárný, "Towards fully probabilistic control design," *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.

[20] J. Šindelář, I. Vajda, and M. Kárný, "Stochastic control optimal in the Kullback sense," *Kybernetika*, vol. 44, no. 1, pp. 53–60, 2008.

[21] I. R. Petersen, M. R. James, and P. Dupuis, "Minimax optimal control of stochastic systems with relative entropy constraints," *IEEE Trans. Automat. Control*, vol. 45, no. 3, pp. 398–412, March 2000.

[22] C. D. Charalambous and F. Rezaei, "Stochastic uncertain systems subject to relative entropy constraints: induced norms and monotonicity properties of minimax games," *IEEE Trans. Automat. Control*, vol. 52, no. 4, pp. 647–660, April 2007.

[23] L. P. Hansen and T. J. Sargent, *Robustness*. Princeton University Press, 2008.

[24] S. K. Mitter and N. J. Newton, "A variational approach to nonlinear estimation," *SIAM J. Control Optim.*, vol. 42, no. 5, pp. 1813–1833, 2003.

[25] V. G. Vovk, "Aggregating strategies," in *Proc. 3rd Annual Workshop on Computational Learning Theory*, San Mateo, CA, 1990, pp. 372–383.

[26] N. Merhav, E. Ordentlich, G. Seroussi, and M. J. Weinberger, "On sequential strategies for loss functions with memory," *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1947–1958, July 2002.

[27] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems," *J. Comput. Sys. Sci.*, vol. 71, pp. 291–307, 2005.

[28] J. N. Tsitisklis, "NP-hardness of checking the unichain condition in average cost MDPs," *Oper. Res. Lett.*, vol. 35, no. 3, pp. 319–323, 2007.

[29] P. Guan, M. Raginsky, and R. Willett, "Online Markov decision processes with Kullback-Leibler control cost," *Proceedings of the American Control Conference*, 2012.

[30] O. Hernández-Lerma and J. B. Lasserre, *Markov Chains and Invariant Probabilities*. Birkhäuser, 2003.

[31] E. Seneta, *Nonnegative Matrices and Markov Chains*. Springer, 2006.

[32] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. Springer, New York, 2005.

[33] O. Hernández-Lerma, *Adaptive Markov Control Processes*. Springer, 1989.

[34] G. B. Di Masi and L. Stettner, "Risk-sensitive control of discrete-time Markov processes with infinite horizon," *SIAM J. Control Optim.*, vol. 38, no. 1, pp. 61–78, 1999.

[35] S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. Cambrdige Univ. Press, 2009.

[36] S. Balaji and S. P. Meyn, "Multiplicative ergodicity and large deviations for an irreducible Markov chain," *Stochastic Process. Appl.*, vol. 90, pp. 123–144, 2000.

[37] P. Chanchana, "An algorithm for computing the Perron root of a non-negative irreducible matrix," Ph.D. dissertation, North Carolina State University, Raleigh, NC, 2007.

[38] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. Springer, 1985.

[39] R. F. Streater, *Statistical Dynamics: A Stochastic Approach to Nonequilibrum Thermodynamics*, 2nd ed. London: Imperial College Press, 2009.

[40] L. Elsner, "Inverse iteration for calculating the spectral radius of a non-negative irreducible matrix," *Lin. Algebra Appl.*, no. 15, pp. 235–242, 1976.

[41] G. Neu, A. György, C. Szepesvári, and A. Antos, "Online Markov decision processes under bandit feedback," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 1804–1812.

[42] D. Hernández-Hernández and S. I. Marcus, "Risk sensitive control of Markov processes in countable state space," *Systems Control Lett.*, vol. 29, pp. 147–155, 1996.

[43] W. H. Fleming and D. Hernández-Hernández, "Risk-sensitive control of finite state machines on an infinite horizon I," *SIAM J. Control Optim.*, vol. 35, no. 5, pp. 1790–1810, September 1997.

[44] I. Kontoyiannis and S. P. Meyn, "Spectral theory and limit theorems for geometrically ergodic Markov processes," *Ann. Appl. Probab.*, vol. 13, no. 1, pp. 304–362, 2003.

[45] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.