2012 American Control Conference
Fairmont Queen Elizabeth, Montréal, Canada
June 27-June 29, 2012

Online Markov Decision Processes with Kullback–Leibler Control Cost

Peng Guan, Maxim Raginsky, and Rebecca Willett

Abstract— **We consider 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 (2007, 2009), the state-action cost at each time step is a sum of a nonnegative state cost and a control cost given by the Kullback–Leibler 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 having selected the corresponding action. We give an explicit construction of an efficient strategy that has small regret (i.e., the difference between the total state-action cost incurred causally and the smallest cost attainable using noncausal knowledge of the state costs) under mild regularity conditions on the passive dynamics. We demonstrate the performance of our proposed strategy on a simulated target tracking problem.**

I. Introduction

Markov decision processes (MDPs) [1], [2] are a popular framework for sequential decision-making in a random dynamic environment. At each time step, an agent observes the state of the system of interest and chooses an action. The system then transitions stochastically 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 relevant optimality criterion is forward-looking, taking into account the effect of past actions on future costs.

Another framework for sequential decision-making, dating back to the seminal work of Hannan [3] and now widely used in the machine learning community [4], models the effects of the environment by an arbitrarily varying sequence of cost functions, where the cost function pertaining to each time step is revealed to the agent only *after* the corresponding 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.

Recent work by Even-Dar et al. [5] and Yu et al. [6] combines these two 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.

Our contributions. The set-up considered in [5], [6] is motivated by problems in machine learning, 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 [7], [8], 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 [7], [8], 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*.

In the online version of this problem, the state costs form an arbitrarily varying sequence, 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. Our main contribution is an explicit construction of a strategy for the agent, such

This work was supported by NSF grant CCF-1017564 and by AFOSR grant FA9550-10-1-0390.

P. Guan and R. Willett are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA; e-mail: {peng.guan,willett}@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.

978-1-4577-1096-4/12/$26.00 ©2012 AACC

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 costs revealed during all of the preceding phases.

Our construction is inspired by the approach of Yu et al. [6], but there are significant differences:

1) While in [6] 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. (It is possible to extend our approach to more general, e.g., compact, state spaces, but additional regularity conditions will be needed. This extension will be the focus of our future work.)
2) Yu et al. [6] 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. [5]). Since we are working with a continuous action space, we need to explicitly prove uniform ergodicity of any stationary policy that could conceivably be used by our strategy.
3) In [6], 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. By contrast, because of the specific structure of the state-action costs we use, all policy computations reduce to solving finite-dimensional eigenvalue problems, without any need for additional randomization.

Notation. The underlying finite state space is denoted by X. The set of all *Markov* (or *stochastic*) matrices over X is denoted 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 represent the elements of $\mathcal{P}(\mathsf{X})$ by row vectors, and the elements of $\mathcal{C}(\mathsf{X})$ by column vectors. 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*) between μ and ν is

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

(here and elsewhere, we work with natural logarithms). The *span seminorm* of $f \in \mathcal{C}(\mathsf{X})$ is $\|f\|_s \triangleq \max_{x\in\mathsf{X}} f(x) - \min_{x\in\mathsf{X}} f(x)$, where $\|f\|_s = 0$ iff $f \equiv c$ for some constant $c \in \mathbb{R}$. Also, $\|f\|_\infty \triangleq \max_{x\in\mathsf{X}} |f(x)|$ is the sup norm.

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

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

We say that P is *unichain* [9, Ch. 3] if the corresponding Markov chain has a single recurrent class of states (plus a possibly empty transient class). This is equivalent to P having a *unique* invariant distribution π_P (i.e., $\pi_P P = \pi_P$) [10, Sec. 4.2]. 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 $\|\mu P - \nu P\|_1 \le \rho\|\mu - \nu\|_1, \forall \mu, \nu \in \mathcal{P}(\mathsf{X})$ (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 $0 \le \rho < 1$, $\mathcal{M}_1^\rho(\mathsf{X}) \subset \mathcal{M}_1(\mathsf{X})$. The *Dobrushin ergodicity coefficient* of $P \in \mathcal{M}(\mathsf{X})$ is

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

and P is $\alpha(P)$-contractive [10]. Finally, for any $P, P' \in \mathcal{M}(\mathsf{X})$ we let $\|P - P'\|_\infty \triangleq \max_{x\in\mathsf{X}} \|P(x,\cdot) - P'(x,\cdot)\|_1$.

II. Problem statement and main result

The model. We start by specifying our online MDP 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 $t = 1, 2, \dots$
  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 $t \ge 1$, A selects the transition probabilities $P_t(x,y) = \Pr\{X_{t+1} = y | X_t = x\}$ based on $f^{t-1} = (f_1, \dots, f_{t-1})$, and incurs the state cost $f_t(X_t)$. Following Todorov [7], [8], we use the Kullback–Leibler control cost, $D(P_t(X_t,\cdot)\|P^*(X_t,\cdot))$, where $P^* \in \mathcal{M}(\mathsf{X})$ is a fixed Markov matrix that models the passive (or reference) dynamics. Thus, the total cost at time t is

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

and the goal is to minimize a suitable notion of regret. We assume that the environment E is *oblivious* (or nonadaptive), i.e., each f_t may depend on f^{t-1}, but not on X^t.

Strategies and regret. A *strategy* for the agent is a sequence $\gamma = \{\gamma_t\}_{t=1}^\infty$ of mappings $\gamma_t : \mathcal{F}^{t-1} \to \mathcal{M}(\mathsf{X})$, so that $P_t = \gamma_t(f^{t-1})$. The cumulative cost of γ after T steps is

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

The regret at time T is the difference between C_T and the expected cumulative cost that A could achieve *in hindsight* (with full knowledge of f^T) using a stationary unichain random walk on X. Formally, we define the regret of γ at time T w.r.t. $P \in \mathcal{M}_1(\mathsf{X})$ by[1]

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

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

where $\mathbb{E}^P_{x_1}[\cdot]$ denotes expectation w.r.t. the Markov chain with initial state $X_1 = x_1$ and transition matrix P. Adopting standard terminology [4], we say that γ is *Hannan-consistent* w.r.t. $\mathcal{N} \subseteq \mathcal{M}_1(\mathsf{X})$ if

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

i.e., if the worst-case (over $\mathcal{F}$) expected average regret converges to zero uniformly over $\mathcal{N}$.

The main result. We make the following two assumptions:

Assumption 1. *The passive dynamics P^* is ergodic (i.e., irreducible and aperiodic).*

Assumption 2. $\alpha(P^*) < 1$.

Assumption 1 ensures that P^* has a unique everywhere positive invariant distribution π^*. Assumption 2 guarantees that the convergence to π^* is exponentially fast (so that P^* is geometrically ergodic), but 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 both $P(x, y)$ and $P(x', y)$ are strictly positive. We are now in a position to state our main result:

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

$$\sup_{P\in\mathcal{M}^\rho_1(\mathsf{X})} \sup_{f_1,\ldots,f_T\in\mathcal{F}} \frac{\mathbb{E}R_T(P)}{T} = O(T^{-1/4+\epsilon}) \qquad (1)$$

i.e., γ is Hannan-consistent w.r.t. $\mathcal{M}^\rho_1(\mathsf{X})$.

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 comparison policies in $\mathcal{M}^\rho_1(\mathsf{X})$.

III. Preliminaries: MDPs with KL control cost

Our construction of a Hannan-consistent strategy uses Todorov's theory of MDPs with KL control cost [7], [8]. In this section, we give a brief overview of this theory and present several new results that will be used in the sequel.

The standard set-up for an MDP with a finite state space X and a compact action space U (see, e.g., [2] or [11]) involves a family of Markov matrices $P_u \in \mathcal{M}(\mathsf{X})$ indexed by $u \in \mathsf{U}$. The *average cost* of a stationary Markov policy $w : \mathsf{X} \to \mathsf{U}$ with initial condition $X_1 = x_1$ is given by

$$J(w, x_1) \triangleq \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^w_{x_1}\left[\sum_{t=1}^T c(X_t, w(X_t))\right], \qquad (2)$$

where $\mathbb{E}^w_{x_1}[\cdot]$ is the expectation w.r.t. 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 that minimizes (2) for every initial condition x_1 revolves around the *average-cost optimality equation* (ACOE)

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

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, then it can be shown [2], [11] that the stationary policy

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

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

MDPs with KL control cost. In the set-up of [7], [8], 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 $P_u(x, y) \triangleq u(y)$, $(x, y) \in \mathsf{X} \times \mathsf{X}$. Thus, any state feedback law $w : \mathsf{X} \to \mathcal{P}(\mathsf{X})$ induces the state transitions

$$\mathbb{P}\{X_{t+1} = y | X_t = x\} = P_{w(x)}(x, y) \equiv u(y), \; t \ge 1$$

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

To specify an MDP, we fix a function $f \in \mathcal{C}_+(\mathsf{X})$ and a Markov matrix $P^* \in \mathcal{M}(\mathsf{X})$, and define the one-step state-action cost $c : \mathsf{X} \times \mathcal{P}(\mathsf{X}) \to \mathbb{R}_+ \cup \{+\infty\}$ by

$$c(x, u) \triangleq f(x) + D(u\|P^*(x, \cdot)), \quad x \in \mathsf{X}, u \in \mathcal{P}(\mathsf{X})$$

where $f(x)$ is the *state cost* that penalizes the "undesirability" of x, while the KL divergence $D(u\|P^*(x, \cdot))$ is the *control cost* that penalizes deviations of $u \in \mathcal{P}(\mathsf{X})$ from $P^*(x, \cdot)$. Here, P^* is the *passive dynamics*, which may be thought of as specifying the state transition probabilities in the absence of control. Any state feedback law *perturbs* these transition probabilities, and the KL control cost ensures that these perturbations are as small as possible.

Following common practice, we will 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 $X_1 = x_1$ is given by

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

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

The ACOE for this problem takes the form

$$h(x) + \lambda = f(x) + \min_{u\in\mathcal{P}(\mathsf{X})} \{D(u\|P^*(x, \cdot)) + \mathbb{E}_u h\}, \quad (3)$$

and it is easy to show, using the fact that the KL divergence is nonnegative, that the optimal policy is given by

$$P_*(x, y) = \frac{P^*(x, y)e^{-h(y)}}{\Lambda(x)}, \qquad (4)$$

where $\Lambda(x) = P^* e^{-h}(x)$ is a normalization factor. If we define for every $\varphi \in \mathcal{C}(\mathsf{X})$ the *twisted kernel* [12]

$$\check{P}_\varphi(x,y) \triangleq \frac{P^*(x,y)e^{-\varphi(y)}}{P^*e^{-\varphi}(x)},$$

then we can express the optimal policy (4) more succinctly as $P_* = \check{P}_h$. Moreover, substituting (4) into (3), we get $h(x) + \lambda = f(x) - \log \Lambda(x), \forall x \in \mathsf{X}$, which implies in turn that the exponentiated relative value function $V \triangleq e^{-h}$ solves the so-called *multiplicative Poisson equation* (MPE) $e^{-f}P^*V = e^{-\lambda}V$ [12].

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.

Some properties of Todorov's optimal policy. Our proof of Theorem 1 exploits several key properties enjoyed by the policy (4) under the assumptions of Section II. The relevant results are stated and proved in detail in the full version of this paper [13]; here we only state two representative ones:

Proposition 1 (Existence, uniqueness, ergodicity)**.** *Under Assumption 1, for any state cost $f \in \mathcal{C}_+(\mathsf{X})$ the MPE 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 the MPE 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 $\check{\pi}_f = \check{\pi}_f \check{P}_f \in \mathcal{P}(\mathsf{X})$.*

Remark 2. This result is implicit in [7].

Proposition 2 (Steady-state optimality)**.** *For any $f \in \mathcal{C}_+(\mathsf{X})$ and any $P \in \mathcal{M}_1(\mathsf{X})$ define*

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

Then

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

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. [6]. 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 [6], 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 $\tau_m = \lceil m^{1/3-\epsilon} \rceil$. 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 is as follows:

```
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
```

The implementation of this strategy reduces to solving a finite-dimensional Frobenius–Perron eigenvalue (FPE) problem [10] at the beginning of each phase to obtain a Todorov-type relative value function. The corresponding twisted kernel (which is ergodic by virtue of Proposition 1) then determines the stationary policy to be followed throughout that phase. If it is infeasible to find the exact solution of the FPE problem, one can use an iterative procedure described in Section 2.1 of [7] which has an exponential rate of convergence.

V. Proof of Theorem 1

A. The main idea

The proof follows the same general outline as in [6] and consists of a series of steps. First, we demonstrate that there is no loss of generality in considering a different notion of regret, where the cumulative cost of the strategy of interest is compared to the *steady-state cost* of a fixed stationary policy. Next, we decompose the total cost up to time T into the contributions of the individual phases and use uniform ergodicity of the Markov matrices $P^{(1)}, P^{(2)}, \ldots$ (which we first prove) to approximate the total cost within each phase by its steady-state value (i.e., when the state within the mth phase is sampled from the unique invariant distribution of $P^{(m)}$). Once this is done, we show that this steady-state value is not too different from what one would get with full knowledge of all state cost functions within the corresponding phase, i.e., if in phase m one used the next-phase policy $P^{(m+1)}$ instead of $P^{(m)}$. We conclude the proof by showing that the cumulative "one-phase-ahead" cost is bounded above by the total steady-state cost of the stationary policy optimized for the average of all state cost functions revealed within the horizon of interest.

B. Two preliminary lemmas

The proof makes extensive use of the following two lemmas (see the full version of this paper [13] for details):

Lemma 1 (Uniform bounds)**.** *There exist 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 first bound holds for any $P \in \mathcal{M}_1(\mathsf{X})$, such that $D(P(x,\cdot)\|P^*(x,\cdot)) < +\infty$ for all $x \in \mathsf{X}$.

Lemma 2 (Policy continuity)**.** *There is a constant $K_2 \geq 0$, such that*

$$\|P^{(m+1)} - P^{(m)}\|_\infty \leq \frac{K_2 \tau_m}{\tau_{1:m}}$$

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

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

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

C. Details

Due to space limitations, we only present a brief sketch of the main steps; the full details can be found in [13].

Step 1: Reduction to the steady-state case. For any $P \in \mathcal{M}_1(\mathsf{X})$, we define the *steady-state regret*

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

Now let us fix some $\rho \in [0,1)$ and consider an arbitrary $P \in \mathcal{M}^\rho_1(\mathsf{X})$. Using Lemma 1, we can show that

$$|R^{\text{ss}}_T(P) - R_T(P)| \leq \frac{2K_0}{1-\rho}. \qquad (5)$$

Therefore, it suffices to show that the bound (1) holds with $\mathbb{E}R^{\text{ss}}_T(P)$ in place of $\mathbb{E}R_T(P)$.

Step 2: Steady-state approximation within phases. Let M denote the number of complete phases up to time T (so that $\tau_{1:M} \leq T < \tau_{1:M+1}$); simple algebra shows that $M = O(T^{3/4+\epsilon})$. Then using Lemma 1 and the facts that $\tau_{M+1} = O(M^{1/3-\epsilon})$ and all f_t's are nonnegative, we have

$$R^{\text{ss}}_T(P) \leq R^{\text{ss}}_{\tau_{1:M}}(P) + O(T^{1/4-\epsilon}). \qquad (6)$$

Let $C_{\tau_{1:M}} = \sum_{t=1}^{\tau_{1:M}} c_t(X_t,P_t)$. Applying Lemma 1, we get

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

Step 3: Looking one phase ahead. Next, for every $1 \leq m \leq M$, using Lemma 1 and Lemma 2, we have,

$$\sum_{t\in\mathcal{T}_m} \mathbb{E}_{\pi^{(m)}} c_t(X,P^{(m)}) \leq \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) + \frac{K_4 \tau_m^2}{\tau_{1:m}},$$

where $K_4 = K_0 K_2/(1-\alpha) + K_3$. Therefore,

$$\mathbb{E}C_{\tau_{1:M}} \leq \sum_{m=1}^M \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) + \sum_{m=1}^M \frac{K_4 \tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha}.$$

Step 4: Looking M phases ahead. Using backward induction and Proposition 2, we can show

$$\sum_{m=1}^M \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(m+1)}) \leq \sum_{m=1}^M \tau_m \bar{J}_{\widehat{f}^{(m)}}(P^{(M+1)}). \qquad (7)$$

Moreover, since

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

we can bound the expected steady-state regret as

$$\mathbb{E}R^{\text{ss}}_{\tau_{1:M}}(P) \leq \sum_{m=1}^M \frac{K_4 \tau_m^2}{\tau_{1:m}} + \frac{2K_0 M}{1-\alpha}. \qquad (8)$$

We can show that the sum on the r.h.s. of (8) is $O(M) = O(T^{3/4+\epsilon})$. Therefore, $\mathbb{E}R^{\text{ss}}_{\tau_{1:M}}(P) = O(T^{3/4+\epsilon})$. Combining this with (6), we see that $\mathbb{E}R^{\text{ss}}_T(P) = O(T^{3/4+\epsilon})$, which, together with (5), gives (1). This completes the proof of Theorem 1.

VI. Simulation results

In this section, we demonstrate the performance of our proposed strategy on a simulated problem involving online (real-time) tracking of multiple targets on a large connected graph, which models a terrain with obstacles. The state space is the set of all nodes of the graph. The tracking agent's motion is constrained by the topology of the graph, while the targets' motion is unconstrained and arbitrarily time-varying. Thus, the agent has to bypass the obstacles to reach desired locations, while the targets can go to any place in one step.

To make sure that our assumptions are satisfied, we construct the passive dynamics in the form $P^* = (1-\delta)P_1 + \delta P_0$. Here, P_1 is a random walk that represents environmental constraints, allowing the agent to go from a given node either to the node's nearest neighbors (with equal probability) or to stay put. To ensure that the agent is sufficiently mobile, the probability of staying put is chosen to be relatively small compared to the probability of transitioning to any of the neighboring nodes. Since the underlying graph is connected, the random walk P_1 is ergodic. We also add a perturbation P_0, which has a fixed column of ones (we can think of the node indexing that column as a "home base" for the tracker), and zeros elsewhere. This perturbation ensures that no two rows of P^* are orthogonal, so $\alpha(P^*) < 1$. The "size" of the perturbation is controlled by $\delta \in (0,1)$, which is set to be small, so the agent only has a slight chance to go back to "home base" from any given node. Note that $\alpha(P^*) \leq 1-\delta$.

We implemented our simulation with two targets. We ran $K = 1000$ independent experiments, each consisting of $T = 1000$ time steps. At time t in the kth experiment, the tracking agent is in state (location) $x^{(k)}_t$, the two targets are at locations $s^{(k)}_{i,t}, i = 1,2$, and the tracker's action is $P^{(k)}_t$. The cumulative cost after T time steps is

$$C^{(k)}_T = \sum_{t=1}^T \left[f^{(k)}_t\left(x^{(k)}_t\right) + D\left(P^{(k)}_t(x^{(k)}_t,\cdot)\middle\| P^*(x^{(k)}_t,\cdot)\right)\right],$$

where the state cost

$$f^{(k)}_t\left(x^{(k)}_t\right) = \min_{i=1,2} \left\| x^{(k)}_t - s^{(k)}_{i,t} \right\| \qquad (9)$$

is the Euclidean distance between the agent and the closer of the two targets. Each experiment was initialized with a different starting state of the tracker. We compared the performance of our adaptive strategy to a nonadaptive one, in which the agent simply performs a random walk on the graph according to the passive dynamics P^*. The results are reported in the two figures below.

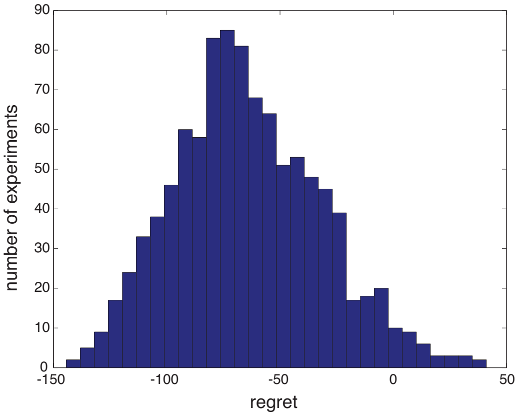

Fig. 1. Histogram of regret in 1000 experiments, evaluated w.r.t. the passive dynamics at $T = 1000$ in each experiment. The regret of our strategy is negative 98 percent of the time, which implies that our adaptive strategy outperforms the passive dynamics 98 percent of the time.

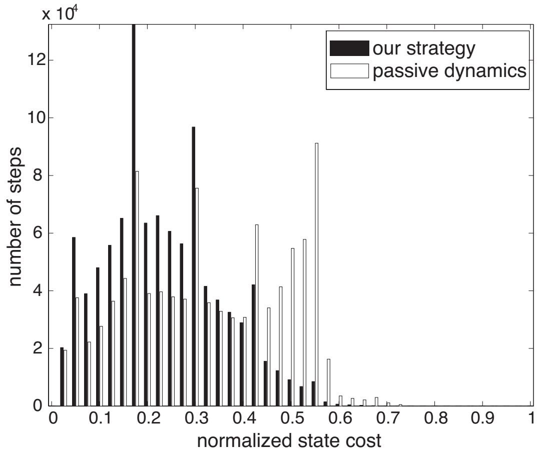

Fig. 2. Histogram of normalized state cost (9) over all experiments (total of $KT = 10^6$ time steps). The horizontal axis corresponds to the distance between the agent and closer of the two targets; the vertical axis corresponds to the number of steps in which the agent is within the corresponding distance range. We can see that, compared to the passive dynamics, the agent implementing our adaptive strategy spends more time in close proximity to (one of the) targets.

VII. Conclusion

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 [4]) 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 [7], [8].

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. [6] for online MDPs with finite state and action spaces. However, while our strategy (like that of [6]) 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. [5] present a strategy that achieves a much better $O(\sqrt{T})$ regret. Of course, the strategy of [5] involves recomputing the policy at *every* time step (rather than in phases, as is done here and in [6]), which results in a significant loss of efficiency. An interesting open question, which we plan to address in future work, is whether it is possible to attain $O(\sqrt{T})$ regret for online MDPs with KL control costs.

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 [14], [15]. Indeed, the ACOE (3) 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., [15, 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.

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] 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.

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

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

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] P. Guan, M. Raginsky, and R. Willett, "Online Markov decision processes with Kullback–Leibler control cost," 2012, preprint.

[14] 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.

[15] 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.