Universal Reinforcement Learning

Vivek F. Farias, Ciamac C. Moallemi, *Member, IEEE*, Benjamin Van Roy, *Senior Member, IEEE*, and Tsachy Weissman, *Senior Member, IEEE*

***Abstract*—We consider an agent interacting with an unmodeled environment. At each time, the agent makes an observation, takes an action, and incurs a cost. Its actions can influence future observations and costs. The goal is to minimize the long-term average cost. We propose a novel algorithm, known as the active LZ algorithm, for optimal control based on ideas from the Lempel-Ziv scheme for universal data compression and prediction. We establish that, under the active LZ algorithm, if there exists an integer K such that the future is conditionally independent of the past given a window of K consecutive actions and observations, then the average cost converges to the optimum. Experimental results involving the game of Rock-Paper-Scissors illustrate merits of the algorithm.**

***Index Terms*—Context tree, dynamic programming, Lempel-Ziv, optimal control, reinforcement learning, value iteration.**

I. Introduction

CONSIDER an agent that, at each integer time t, makes an observation X_t from a finite observation space $\mathbb{X}$, and takes an action A_t selected from a finite action space $\mathbb{A}$. The agent incurs a bounded cost $g(X_t, A_t, X_{t+1}) \in [-g_{\max}, g_{\max}]$. The goal is to minimize the long-term average cost

$$\limsup_{T\to\infty} \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T} g(X_t, A_t, X_{t+1})\right].$$

Here, the expectation is over the randomness in the X^t process,[1] and, at each time t, the action A_t is selected as a function of the prior observations X^t and the prior actions A^{t-1}.

We will propose a general action-selection strategy called the heightdepth active LZ algorithm. In addition to the new strategy, a primary contribution of this paper is a theoretical guarantee that this strategy attains optimal average cost under weak assumptions about the environment. The main assumption is that there exists an integer K such that the future is conditionally independent of the past given a window of K consecutive actions and observations. In other words

$$\Pr(X_t = x_t | \mathcal{F}_{t-1}) = P\left(x_t | X_{t-K}^{t-1}, A_{t-K}^{t-1}\right) \quad (1)$$

where P is a transition kernel and $\mathcal{F}_t$ is the σ-algebra generated by (X^t, A^t). We are particularly interested in situations where neither P nor even K are known to the agent. That is, where there is a finite but unknown dependence on history.

Consider the following examples, which fall into the above formalism.

Example 1 (Rock-Paper-Scissors): Rock-Paper-Scissors is a two-person, zero-sum matrix game that has a rich history as a reinforcement learning problem. The two players play a series of games indexed by the integer t. Each player must generate an action—rock, paper, or scissors—for each game. He then observes his opponent's hand and incurs a cost of -1, 1, or 0, depending on whether the pair of hands results in a win, loss, or draw. The game is played repeatedly and the player's objective is to minimize the average cost.

Define X_t to be the opponent's choice of action in game t, and A_{t-1} to be the player's choice of action in game t. The action and observation spaces for this game are

$$\mathbb{A} \triangleq \mathbb{X} \triangleq \{\text{ rock, paper, scissors}\}.$$

Identifying these with the integers $\{1, 2, 3\}$, the cost function is

$$g(x_t, a_t, x_{t+1}) \triangleq \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}_{x_{t+1}, a_t}$$

Assuming that the opponent uses a mixed strategy that depends only on information from the last $K-1$ games, such a strategy defines a transition kernel P over the opponent's play X_t in game t of (1). (Note that such a P has special structure in that, for example, it has no dependence on the player's action A_{t-1} in game t, since this is unknown to the opponent until after game t is played.) Then, the problem of finding the optimal strategy against an unknown, finite-memory opponent falls within our framework.

Example 2 (Joint Source-Channel Coding With a Fixed Decoder): Let $\mathbb{S}$ and $\mathbb{Y}$ be finite source and channel alphabets, respectively. Consider a sequence of symbols $\{S_t\}$ from the source alphabet $\mathbb{S}$ which are to be encoded for transmission across a channel. Let $Y_t \in \mathbb{Y}$ represent the choice of encoding at time t, and let $\hat{Y}_t \in \mathbb{Y}$ be the symbol received after corruption by the channel. We will assume that this channel has a finite memory of order M. In other words, the distribution of $\hat{Y}_t$ is a function of Y_{t-M+1}^t. For all times t, let $d : \mathbb{Y}^L \to \mathbb{S}$ be some fixed decoder that decodes the symbol at time t based on the most recent L symbols received $\hat{Y}_{t-L+1}^t$. Given a single letter

Manuscript received July 20, 2007; revised June 09, 2009. Current version published April 21, 2010. The work of V. F. Farias was supported by a supplement to the NSF by Grant ECS-9985229 provided by the MKIDS Program. The work of was C. C. Moallemi was supported by a Benchmark Stanford Graduate Fellowship. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Adelaide, Australia, September 2005.

V. F. Farias is with the Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: vivekf@mit.edu).

C. C. Moallemi is with the Graduate School of Business, Columbia University, New York, NY 10027 USA (e-mail: ciamac@gsb.columbia.edu).

B. Van Roy is with the Department of Management Science and Engineering and the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: bvr@stanford.edu).

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

Communicated by W. Szpankowski, Associate Editor for Source Coding.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2010.2043762

[1]For a sequence such as $\{X_t\}$, X_s^t denotes the vector $(X_s, \ldots, X_t)$. We also use the notation $X^t = X_1^t$.

0018-9448/$26.00 © 2010 IEEE

distortion measure $\rho : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$, define the expected distortion at time t by

$$g\left(s_t, y_{t-L-M+2}^{t}\right) \triangleq \mathsf{E}\left[\rho\left(d\left(\hat{Y}_{t-L+1}^{t}\right), s_t\right) \middle| Y_{t-L-M+2}^{t} = y_{t-L+M+2}^{t}\right].$$

The optimization problem is to find a sequence of functions $\{\mu_t\}$, where each function $\mu_t : \mathbb{X}^t \to \mathbb{A}$ specifies an encoder at time t, so as to minimize the long-term average distortion

$$\limsup_{T\to\infty} \mathsf{E}_{\mu}\left[\frac{1}{T}\sum_{t=1}^{T} g(S_t, Y_{t-L-M+2}^{t})\right].$$

Assume that the source is Markov of order N, but that both the transition probabilities for the source and the order N are unknown. Setting $K = \max(L + M - 1, N)$, define the observation at time t to be the vector $X_t = \left(S_t, Y_{t-L-M+2}^{t-1}\right)$ and the action at time t to be $A_t = Y_t$. Then, optimal coding problem at hand falls within our framework (cf. [1] and references therein).

With knowledge of the kernel P (or even just the order of the kernel, K), solving for the average cost optimal policy in either of the examples above via dynamic programming methods is relatively straightforward. This paper develops an algorithm that, without knowledge of the kernel or its order, achieves average cost optimality. The active LZ algorithm we develop consists of two broad components. The first is an efficient data structure, a context tree on the joint process $\left(X^t, A^{t-1}\right)$, to store information relevant to predicting the observation at time $t+1$, X_{t+1}, given the history available up to time t and the action selected at time t, A_t. Our prediction methodology borrows heavily from the Lempel-Ziv algorithm for data compression [2]. The second component of our algorithm is a dynamic programming scheme that, given the probabilistic model determined by the context tree, selects actions so as to minimize costs over a suitably long horizon. Absent knowledge of the order of the kernel, K, the two tasks above—building a context tree in order to estimate the kernel, and selecting actions that minimize long-term costs—must be done continually in tandem which creates an important tension between "exploration" and "exploitation." In particular, on the one hand, the algorithm must select actions in a manner that builds an accurate context tree. On the other hand, the desire to minimize costs naturally restricts this selection. By carefully balancing these two tensions our algorithm achieves an average cost equal to that of an optimal policy with full knowledge of the kernel P.

Related problems have been considered in the literature. Kearns and Singh [3] present an algorithm for reinforcement learning in a Markov decision process. This algorithm can be applied in our context when K is known, and asymptotic optimality is guaranteed. More recently, Even-Dar *et al.* [4] present an algorithm for optimal control of partially observable Markov decision processes, a more general setting than what we consider here, and are able to establish theoretical bounds on convergence time. The algorithm there, however, seems difficult and unrealistic to implement in contrast with what we present here. Further, it relies on knowledge of a constant related to the amount of time a "homing" policy requires to achieve equilibrium. This constant may be challenging to estimate.

Work by de Farias and Megiddo [5] considers an optimal control framework where the dynamics of the environment are not known and one wishes to select the best of a finite set of "experts." In contrast, our problem can be thought of as competing with the set of all possible strategies. The prediction problem for loss functions with memory and a Markov-modulated source considered by Merhav *et al.* [6] is essentially a Markov decision problem as the authors point out; again, in this case, knowing the structure of the loss function implicitly gives the order of the underlying Markov process.

The active LZ algorithm is inspired by the Lempel-Ziv algorithm. This algorithm has been extended to address many problems, such as prediction [7], [8], and filtering [6]. In almost all cases, however, future observations are not influenced by actions taken by the algorithm. This is in contrast to the active LZ algorithm, which proactively anticipates the effect of actions on future observations. An exception is the work of Vitter and Krishnan [9], which considers cache prefetching and can be viewed as a special case of our formulation.

The Lempel-Ziv algorithm and its extensions revolve around a context tree data structure that is constructed as observations are made. This data structure is simple and elegant from an implementational point of view. The use of this data structure in reinforcement learning represents a departure from representations of state and belief state commonly used in the reinforcement learning literature. Such data structures have proved useful in representing experience in algorithms for engineering applications ranging from compression to prediction to denoising. Understanding whether and how some of this value can be extended to reinforcement learning is the motivation for this paper.

The remainder of this paper is organized as follows. In Section II, we formulate our problem precisely. In Section III, we present our algorithm and provide computational results in the context of the rock-paper-scissors example. Our main result, as stated in Theorem 2 in Section IV, is that the algorithm is asymptotically optimal. Section V concludes.

II. Problem Formulation

Recall that we are endowed with finite action and observation spaces $\mathbb{A}$ and $\mathbb{X}$, respectively, and we have

$$\Pr\left(X_t = x_t \middle| \mathcal{F}_{t-1}\right) = P\left(x_t \middle| X_{t-K}^{t-1}, A_{t-K}^{t-1}\right)$$

where P is a stochastic transition kernel. A *policy* μ is a sequence of mappings $\{\mu_t\}$, where for each time t the map $\mu_t : \mathbb{X}^t \times \mathbb{A}^{t-1} \to \mathbb{A}$ determines which action shall be chosen at time t given the history of observations and actions observed up to time t. In other words, under policy μ, actions will evolve according to the rule

$$A_t = \mu_t\left(X^t, A^{t-1}\right).$$

We will call a policy μ stationary if

$$\mu_t\left(X^t, A^{t-1}\right) = \mu\left(X_{t-K+1}^{t}, A_{t-K+1}^{t-1}\right), \quad \text{for all } t \geq K$$

for some function $\mu : \mathbb{X}^K \times \mathbb{A}^{K-1} \to \mathbb{A}$. Such a policy selects actions in a manner that depends only one the current observation X_t and the observations and actions over the most recent

K time steps. It is clear that for a fixed stationary policy μ, the observations and actions for time $t \geq K$ evolve according to a Markov chain on the finite state space $\mathbb{X}^K \times \mathbb{A}^{K-1}$. Given an initial state (x^K, a^{K-1}), we can define the average cost associated with the stationary policy μ by

$$\lambda_\mu(x^K, a^{K-1}) \triangleq \lim_{T\to\infty} \mathsf{E}_\mu \left[\frac{1}{T} \sum_{t=1}^{T} g(X_t, A_t, X_{t+1}) \,\middle|\, x^K, a^{K-1} \right].$$

Here, the expectation is conditioned on the initial state $\left(X^K, A^{K-1}\right) = \left(x^K, a^{K-1}\right)$. Since the underlying state-space, $\mathbb{X}^K \times \mathbb{A}^{K-1}$, is finite, the above limit always exists [10, Proposition 4.1.2]. Since there are finitely many stationary policies, we can define the optimal average cost over stationary policies by

$$\lambda^*\left(x^K, a^{K-1}\right) \triangleq \min_{\mu} \lambda_\mu\left(x^K, a^{K-1}\right)$$

where the minimum is taken over the set of all stationary policies. Again, because of the finiteness of the underlying state space, λ^* is also the optimal average cost that can be achieved using any policy, stationary or not. In other words

$$\lambda^*\left(x^K, a^{K-1}\right) = \inf_{\nu} \limsup_{T\to\infty} \mathsf{E}_\nu \left[\frac{1}{T} \sum_{t=1}^{T} g(X_t, A_t, X_{t+1}) \middle| x^K, a^{K-1} \right] \quad (2)$$

where the infimum is taken over the set of all policies [10, Proposition 4.1.7].

We next make an assumption that will enable us to streamline our analysis in subsequent sections.

Assumption 1: The optimal average cost is independent of the initial state. That is, there exists a constant λ^* so that

$$\lambda^*\left(x^K, a^{K-1}\right) = \lambda^*, \quad \forall \left(x^K, a^{K-1}\right) \in \mathbb{X}^K \times \mathbb{A}^{K-1}.$$

The above assumption is benign and is satisfied for any strictly positive kernel P, for example. More generally, such an assumption holds for the class of problems satisfying a "weak accessibility" condition (see Bertsekas [10] for a discussion of the structural properties of average cost Markov decision problems). In the context of our problem, it is difficult to design controllers that achieve optimal average cost in the absence of such an assumption. In particular, if there exist policies under which the chain has multiple recurrent classes, then the optimal average cost may well depend on the initial state and actions taken very early on might play a critical role in achieving this performance. We note that in such cases the assumption above (and our subsequent analysis) is valid for the recurrent class our controller eventually enters.

If the transition kernel P (and, thereby, K) were known, dynamic programming is a means to finding a stationary policy that achieves average cost λ^*. One approach would be to find a solution $J : \mathbb{X}^K \times \mathbb{A}^{K-1} \to \mathbb{R}$ to the discounted Bellman equation

$$J\left(x^K, a^{K-1}\right) = \min_{a_K} \sum_{x_{K+1}} P\left(x_{K+1} | x^K, a^K\right) \times \left[g(x_k, a_k, x_{K+1}) + \alpha J\left(x_2^{K+1}, a_2^K\right) \right] \quad (3)$$

for all $\left(x^K, a^{K-1}\right) \in \mathbb{X}^K \times \mathbb{A}^{K-1}$. Here, $\alpha \in (0,1)$ is a discount factor. If the discount factor alpha is chosen to be sufficiently close to 1, a solution J_α^* (known as the cost-to-go function) to the Bellman equation can be used to define an optimal stationary policy for the original, average-cost problem (2). In particular, for all $\left(x^K, a^{K-1}\right) \in \mathbb{X}^K \times \mathbb{A}^{K-1}$, define the set $\mathcal{A}_\alpha^*\left(x^K, a^{K-1}\right)$ of α-discounted optimal actions to be the set of minimizers to the optimization program

$$\min_{a_K} \sum_{x_{K+1}} P\left(x_{K+1} | x^K, a^K\right) \times \left[g(x_K, a_K, x_{K+1}) + \alpha J_\alpha^*\left(x_2^{K+1}, a_2^K\right) \right]. \quad (4)$$

At a give time t, $\mathcal{A}_\alpha^*\left(X_{t-K+1}^t, A_{t-K+1}^{t-1}\right)$ is the set of actions obtained acting greedily with respect to J_α^*. These actions seek to minimize the expected value of the immediate cost $g(X_t, A_t, X_{t+1})$ at the current time, plus a continuation cost, which quantifies the impact of the current decision on all future costs, and is captured by J_α^*.

If α is sufficiently close to 1, and μ^* is a policy such that for $t \geq K$

$$\mu_t^*\left(X^t, A^{t-1}\right) \in \mathcal{A}_\alpha^*\left(X_{t-K+1}^t, A_{t-K+1}^{t-1}\right) \quad (5)$$

then, μ^* will achieve the optimal average cost λ^*. Such a policy μ^* is sometimes called a Blackwell optimal policy [10].

We return to our example of the game of Rock-Paper-Scissors, to illustrate the above approach.

Example 1 (Rock-Paper-Scissors): Given knowledge of the opponent's (finite-memory) strategy and, thus the transition kernel P, the Bellman equation (3) can be solved for the optimal cost-to-go function J_α^*. Then, an optimal policy for the player would be, for each game $t+1$, to select an action A_t according to (4)–(5). This action is a function of the entire history of game play only through the sequence $\left(X_{t-K+1}^t, A_{t-K+1}^{t-1}\right)$ of recent game play. The action is selected by optimally accounting for both the expected immediate cost $g(X_t, A_t, X_{t+1})$ of the game at hand, and the impact of the choice of action towards all future games (through the cost-to-go function J_α^*).

III. Universal Scheme

Direct solution of the Bellman (3) requires knowledge of the transition kernel P. Algorithm 1, the active LZ algorithm, is a method that requires no knowledge of P, or even of K. Instead, it simultaneously estimates a probabilistic model for the evolution of the system and develops an optimal control for that model, along the course of a single system trajectory. At a high-level, the two critical components of the active LZ algorithm are the estimates $\hat{P}$ and $\hat{J}$. $\hat{P}$ is our estimate of the true kernel P. This estimate is computed using variable length contexts to dynamically build higher order models of the underlying process, in a manner reminiscent of the Lempel-Ziv scheme used for universal prediction. $\hat{J}$ is the estimate to the optimal cost-to-go function J_α^* that is the solution to the Bellman (3). It is computed in a fashion similar to the value iteration approach to solving the Bellman equation (see [10]). Given the estimates $\hat{P}$ and $\hat{J}$, the algorithm randomizes to strike a balance selecting actions so as to improve the quality of the estimates

(exploration) and acting greedily with respect to the estimates so as to minimize the costs incurred (exploitation).

Algorithm 1 The active LZ algorithm, a Lempel-Ziv inspired algorithm for learning.

Input: a discount factor $\alpha \in (0,1)$ and a sequence of exploration probabilities $\{\gamma_t\}$

1: $c \leftarrow 1$\{ the index of the current phrase \}

2: $\tau_c \leftarrow 1$\{ start time of the cth phrase \}

3: $N(\cdot) \leftarrow 0$\{ initialize context counts \}

4: $\hat{P}(\cdot) \leftarrow 1/|\mathbb{X}|$\{ initialize estimated transition probabilities \}

5: $\hat{J}(\cdot) \leftarrow 0$\{ initialize estimated cost-to-go values \}

6: for each time t do

7: observe X_t

8: **if** $N\left(X_{\tau_c}^{t}, A_{\tau_c}^{t-1}\right) > 0$ then \{ are in a context that we have seen before?\}

9: with probability γ_t, pick A_t uniformly over $\mathbb{A}$ \{ explore independent of history \}

10: with remaining probability, $1-\gamma_t$, pick A_t greedily according to $\hat{P}$, $\hat{J}$

$$A_t \in \arg\min_{a_t} \sum_{x_{t+1}} \hat{P}\left(x_{t+1} \middle| X_{\tau_c}^{t}, (A_{\tau_c}^{t-1}, a_t)\right) \times \left[g(X_t, a_t, x_{t+1}) + \alpha \hat{J}\left(\left(X_{\tau_c}^{t}, x_{t+1}\right), \left(A_{\tau_c}^{t-1}, a_t\right)\right)\right]$$

\{ exploit by picking an action greedily \}

11: else \{ we are in a context not seen before \}

12: pick A_t uniformly over $\mathbb{A}$

13: for s with $\tau_c \le s \le t$, in decreasing order do \{ traverse backward through the current context \}

14: update context count: $N(X_{\tau_c}^{s}, A_{\tau_c}^{s-1}) \leftarrow N(X_{\tau_c}^{s}, A_{\tau_c}^{s-1}) + 1$

15: update probability estimates: for all $x_s \in \mathbb{X}$

$$\hat{P}(x_s | X_{\tau_c}^{s-1}, A_{\tau_c}^{s-1}) \leftarrow \frac{N\left((X_{\tau_c}^{s-1}, x_s), A_{\tau_c}^{s-1}\right) + 1/2}{\sum_{x'} N\left((X_{\tau_c}^{s-1}, x'), A_{\tau_c}^{s-1}\right) + |\mathbb{X}|/2}$$

16: update cost-to-go estimate

$$\hat{J}\left(X_{\tau_c}^{s}, A_{\tau_c}^{s-1}\right) \leftarrow \min_{a_s} \sum_{x_{s+1}} \hat{P}\left(x_{s+1} | X_{\tau_c}^{s}, (A_{\tau_c}^{s-1}, a_s)\right) \times \left[g(X_s, a_s, x_{s+1}) + \alpha \hat{J}\left(\left(X_{\tau_c}^{s}, x_{s+1}\right), \left(A_{\tau_c}^{s-1}, a_s\right)\right)\right]$$

17: **end for**

18: $c \leftarrow c+1$, $\tau_c \leftarrow t+1$ \{ start the next phrase \}

19: **end if**

20: **end for**

The active LZ algorithm takes as inputs a discount factor $\alpha \in (0,1)$, sufficiently close to 1, and a sequence of exploration probabilities $\{\gamma_t\}$. The algorithm proceeds as follows: time is parsed into intervals, or "phrases," with the property that if the cth phrase covers the time intervals $\tau_c \le t \le \tau_{c+1} - 1$, then the observation/action sequence $\left(X_{\tau_c}^{\tau_{c+1}-1}, A_{\tau_c}^{\tau_{c+1}-2}\right)$ will not have occurred as the prefix of any other phrase before time τ_c.

At any point in time t, if the current phrase started at time τ_c, the sequence $\left(X_{\tau_c}^{t}, A_{\tau_c}^{t-1}\right)$ defines a context which is used to estimate transition probabilities and cost-to-go function values. To be precise, given a sequence of observations and actions $\left(x^\ell, a^{\ell-1}\right)$, we say the context at time t is $\left(x^\ell, a^{\ell-1}\right)$ if $\left(X_{\tau_c}^{t}, A_{\tau_c}^{t-1}\right) = \left(x^\ell, a^{\ell-1}\right)$. For each $x_{\ell+1} \in \mathbb{X}$ and $a_\ell \in \mathbb{A}$, the algorithm maintains an estimate $\hat{P}\left(x_{\ell+1} | x^\ell, a^\ell\right)$ of the probability of observing $X_{t+1} = x_{\ell+1}$ at the next time step, given the choice of action $A_t = a_\ell$ and the current context $\left(X_{\tau_c}^{t}, A_{\tau_c}^{t-1}\right) = \left(x^\ell, a^{\ell-1}\right)$. This transition probability is initialized to be uniform, and subsequently updated using an empirical estimator based on counts for various realizations of X_{t+1} at prior visits to the context in question. If $N\left(x^{\ell+1}, a^\ell\right)$ is the number of times the context $\left(x^{\ell+1}, a^\ell\right)$ has been visited prior to time t, then the estimate

$$\hat{P}(x_{\ell+1} | x^\ell, a^\ell) = \frac{N(x^{\ell+1}, a^\ell) + 1/2}{\sum_{x'} N\left((x^\ell, x'), a^\ell\right) + |\mathbb{X}|/2} \tag{6}$$

is used. This empirical estimator is akin to the update of a Dirichlet-$1/2$ prior with a multinomial likelihood and is similar to that considered by Krichevsky and Trofimov [11].

Similarly, at each point in time t, given the context $\left(X_{\tau_c}^{t}, A_{\tau_c}^{t-1}\right) = \left(x^\ell, a^{\ell-1}\right) \in \mathbb{X}^\ell \times \mathbb{A}^{\ell-1}$, for each $x_{\ell+1} \in \mathbb{X}$ and $a_\ell \in \mathbb{A}$, the quantity $\hat{J}\left(x^{\ell+1}, a^\ell\right)$ is an estimate of the cost-to-go if the action $A_t = a_\ell$ is selected and then observation $X_{t+1} = x_{\ell+1}$ is subsequently realized. This estimate is initialized to be zero, and subsequently refined by iterating the dynamic programming operator from (3) backwards over outcomes that have been previously realized in the system trajectory, using $\hat{P}$ to estimate the probability of each possible outcome (line 16).

At each time t, an action A_t is selected either with the intent to explore or to exploit. In the former case, the action is selected uniformly at random from among all the possibilities [line (9)]. This allows the action space to be fully explored and will prove critical in ensuring the quality of the estimates $\hat{P}$ and $\hat{J}$. In the latter case, the impact of each possible action on all future costs is estimated using the transition probability estimates $\hat{P}$ and the cost-to-go estimates $\hat{J}$, and the minimizing action is taken acting greedily with respect to $\hat{P}$ and $\hat{J}$ (line 10). A sequence $\{\gamma_t\}$ controls the relative frequency of actions taken to explore versus exploit; over time, as the system becomes well understood, actions are increasingly chosen to exploit rather than explore.

Note that the active LZ algorithm can be implemented easily using a tree-like data structure. Nodes at depth ℓ correspond to contexts of the form $\left(x^\ell, a^{\ell-1}\right)$ that have already been visited. Each such node can link to at most $|\mathbb{X}||\mathbb{A}|$ child nodes of the form $\left(x^{\ell+1}, a^\ell\right)$ at depth $\ell+1$. Each node $\left(x^{\ell+1}, a^\ell\right)$ maintains a count $N\left(x^{\ell+1}, a^\ell\right)$ of how many times it has been seen as a context and maintains a cost-to-go estimate $\hat{J}\left(x^{\ell+1}, a^\ell\right)$. The

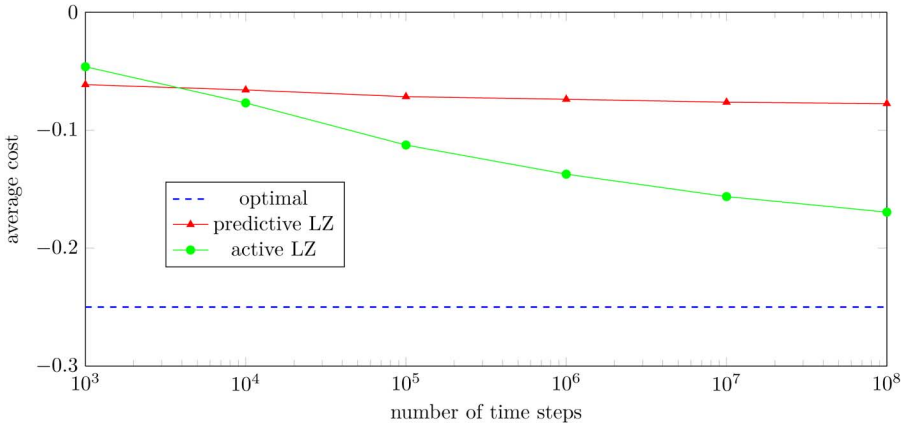

Fig. 1. Performance of the active LZ algorithm on Rock-Paper-Scissors relative to the predictive LZ algorithm and the optimal policy.

probability estimates $\hat{P}$ need not be separately stored, since they are readily constructed from the context counts N according to (6). Each phrase interval amounts to traversing a path from the root to a leaf, and adding an additional leaf. After each such path is traversed, the algorithm moves backwards along the path [lines (11)–(19)] and updates only the counts and cost-to-go estimates along that path. Note that such an implementation has linear complexity, and requires a bounded amount of computation and storage per unit time (or symbol).

We will shortly establish that the active LZ algorithm achieves the optimal long-term average cost. Before launching into our analysis, however, we next consider employing the active LZ algorithm in the context of our running example of the game of Rock-Paper-Scissors. We have already seen how a player in this game can minimize his long-term average cost if he knows the opponent's finite-memory strategy. Armed with the active LZ algorithm, we can now accomplish the same task without knowledge of the opponent's strategy. In particular, as long as the opponent plays using some finite-memory strategy, the active LZ algorithm will achieve the same long-term average cost as an optimal response to this strategy.

Example (1) (Rock-Paper-Scissors): The active LZ algorithm begins with a simple model of the opponent—it assumes that the opponent selects actions uniformly at random in every time step, as per line (4). The algorithm thus does not factor in play in future time steps in making decisions initially, as per line (5). As the algorithm proceeds, it refines its estimates of the opponent's behavior. For game $t+1$, the current context $\left(X_{\tau_c}^t, A_{\tau_c}^{t-1}\right)$ specifies a recent history of the game. Given this recent history, algorithm can make a prediction of the opponent's next play according to $\hat{P}$, and an estimate of the cost-to-go according to $\hat{J}$. These estimates are refined as play proceeds and more opponent behavior is observed. If these estimates converge to their corresponding true values, the algorithm makes decisions [line (10)] that correspond to the optimal decisions that would be made if the true transition kernel and cost-to-go function were known, as in (4)–(5).

A. Numerical Experiments With Rock-Paper-Scissors

Before proceeding with our analysis that establishes the average cost optimality of the active LZ algorithm, we demonstrate its performance on a simple numerical example of the Rock-Paper-Scissors game. The example will highlight the importance of making decisions that optimize long-term costs.

Consider a simple opponent that plays as follows. If, in the previous game, the opponent played rock against scissors, the opponent will play rock again deterministically. Otherwise, the opponent will pick a play uniformly at random. It is easy to see that an optimal strategy against such an opponent is to consistently play scissors until (rock, scissors) occurs, play paper for one game, and then repeat. Such a strategy incurs an optimal average cost of -0.25.

We will compare the performance of the active LZ algorithm against this opponent versus the performance of an algorithm (which we call "predictive LZ") based on the Lempel-Ziv predictor of Martinian [12]. Here, we use the Lempel-Ziv algorithm to predict the opponent's most likely next play based on his history, and play the best response. Since Lempel-Ziv offers both strong theoretical guarantees and impressive practical performance for the closely related problems of compression and prediction, we would expect this algorithm would be effective at detecting and exploiting nonrandom behavior of the opponent. Note, however, such an algorithm is myopic in that it is always optimizing one-step costs and does not factor in the effect of its actions on the opponent's future play.

In Fig. 1, we can see the relative performance of the two algorithms. The predictive LZ algorithm is able to make some modest improvements but gets stuck at a fixed level of performance that is well below optimum. The active LZ algorithm, on the other hand is able to make consistent improvements. The time required for convergence to the optimal cost does, however, appear to be substantial.

IV. Analysis

We now proceed to analyze the active LZ algorithm. In particular, our main theorem, Theorem 2, will show that the average cost incurred upon employing the active LZ algorithm will equal the optimal average cost, starting at any state.

A. Preliminaries

We begin with some notation. Recall that, for each $c \geq 1$, τ_c is the starting time of the cth phrase, with $\tau_1 = 1$. Define $c(t)$ to be index of the current phrase at time t, so that

$$c(t) \triangleq \sup\{c \geq 1 : \tau_c \leq t\}.$$

At time t, the current context will be $\left(X_{\tau_{c(t)}}^{t}, A_{\tau_{c(t)}}^{t-1}\right)$. We define the length of the context at time t to be $d(t) \triangleq t - \tau_{c(t)} + 1$.

The active LZ algorithm maintains context counts N, probability estimates $\hat{P}$, and cost-to-go estimates $\hat{J}$. All of these evolve over time. In order to highlight this dependence, we denote by N_t, $\hat{P}_t$, and $\hat{J}_t$, respectively, the context counts, probability estimates, and cost-to-go function estimates at time t.

Given two probability distributions p and q over $\mathbb{X}$, define $\mathrm{TV}(p,q)$ to be the total variation distance

$$\mathrm{TV}(p,q) \triangleq \tfrac{1}{2}\sum_{x} |p(x) - q(x)|.$$

B. Dynamic Programming Lemma

Our analysis rests on a dynamic programming lemma. This lemma provides conditions on the accuracy of the probability estimates $\hat{P}_t$ at time t that, if satisfied, guarantee that actions generated by acting greedily with respect to $\hat{P}_t$ and $\hat{J}_t$ are optimal. It relies heavily on the fact that the optimal cost-to-go function can be computed by a value iteration procedure that is very similar to the update for $\hat{J}_t$ employed in the active LZ algorithm.

Lemma 1: Under the active LZ algorithm, there exist constants $\bar{K} \geq 1$ and $\bar{\epsilon} \in (0,1)$ so that the following holds: Suppose that, at any time $t \geq K$, when the current context is $\left(X_{\tau_{c(t)}}^{t}, A_{\tau_{c(t)}}^{t-1}\right) = (x^s, a^{s-1})$, we have

(i) The length $s = d(t)$ of the current context is at least K.

(ii) For all ℓ with $s \leq \ell \leq s + \bar{K}$ and all $\left(x_{s+1}^{\ell}, a_s^{\ell-1}\right)$, the context $\left(x^{\ell}, a^{\ell-1}\right)$ has been visited at least once prior to time t.

(iii) For all ℓ with $s \leq \ell \leq s + \bar{K}$ and all $\left(x_{s+1}^{\ell}, a_s^{\ell}\right)$, the distribution $\hat{P}_t\left(\cdot|x^{\ell}, a^{\ell}\right)$ satisfies

$$\mathrm{TV}\left(\hat{P}_t(\cdot|x^{\ell}, a^{\ell}), P(\cdot|x_{\ell-K+1}^{\ell}, a_{\ell-K+1}^{\ell})\right) \leq \bar{\epsilon}.$$

Then, the action selected by acting greedily with respect to $\hat{P}_t$ and $\hat{J}_t$ at time t [as in line (10) of the active LZ algorithm] is α-discounted optimal. That is, such an action is contained in the set of actions $\mathcal{A}_{\alpha}^{*}\left(X_{t-K+1}^{t}, A_{t-K+1}^{t-1}\right)$.

Proof: First, note that there exists a constant $\epsilon > 0$ so that if $\tilde{P} : \mathbb{X}^K \times \mathbb{A}^K \to [0,1]$ and $\tilde{J} : \mathbb{X}^K \times \mathbb{A}^{K-1} \to \mathbb{R}$ are two arbitrary functions with

$$\|\tilde{P}(\cdot|x^K, a^K) - P(\cdot|x^K, a^K)\|_1 < \epsilon, \quad \forall\, x^K,\ a^K \tag{7}$$

$$|\tilde{J}(x^K, a^{K-1}) - J_{\alpha}^{*}(x^K, a^{K-1})| < \epsilon, \quad \forall\, x^K,\ a^{K-1} \tag{8}$$

then acting greedily with respect to $(\tilde{P}, \tilde{J})$ results in actions that are also optimal with respect to (P, J_{α}^{*})—that is, an optimal policy. The existence of such an ϵ follows from the finiteness of the observation and action spaces. Now, suppose that, at time t, the hypotheses of the lemma hold for some $(\bar{\epsilon}, \bar{K})$, and that the current context is $\left(x^s, a^{s-1}\right)$, with $s = d(t)$. If we can demonstrate that, for every $a_s \in \mathcal{A}$

$$\sum_{x_{s+1}} \left|\hat{P}_t(x_{s+1}|x^s, a^s) - P(x_{s+1}|x_{s-K+1}^{s}, a_{s-K+1}^{s})\right| < \epsilon \tag{9}$$

and

$$\max_{x_{s+1}, a_s} \left|\hat{J}_t\left(x^{s+1}, a^s\right) - J_{\alpha}^{*}(x_{s-K+2}^{s+1}, a_{s-K+2}^{s})\right| < \epsilon \tag{10}$$

then, by the discussion above, the conclusion of the lemma holds. Equation (9) is immediate from our hypotheses if $\bar{\epsilon} < \epsilon/2$.

It remains to establish (10). In order to do so, fix a choice of x_{s+1} and a_s. To simplify notation in what follows, we will suppress the dependence of certain probabilities, costs, and value functions on $\left(x^{s+1}, a^s\right)$. In particular, for all x_{s+2} and a_{s+1}, define

$$\hat{P}_t(x_{s+2}|a_{s+1}) \triangleq \hat{P}_t\left(x_{s+2}|x^{s+1}, a^{s+1}\right)$$
$$P(x_{s+2}|a_{s+1}) \triangleq P\left(x_{s+2}|x_{s-K+2}^{s+1}, a_{s-K+2}^{s+1}\right).$$

These are, respectively, estimated and true transition probabilities. Define

$$g_t(a_{s+1}, x_{s+2}) \triangleq g(x_s, a_{s+1}, x_{s+2})$$

to be the current cost, and define the value functions

$$\hat{J}_t(x_{s+2}, a_{s+1}) \triangleq \hat{J}_t\left(x^{s+2}, a^{s+1}\right),$$
$$J_{\alpha}^{*}(x_{s+2}, a_{s+1}) \triangleq J_{\alpha}^{*}\left(x_{s-K+3}^{s+2}, a_{s-K+3}^{s+1}\right).$$

Then, using the fact that J_{α}^{*} solves the Bellman (3) and the recursive definition of $\hat{J}_t$ [line (16) in the active LZ algorithm], we have

$$\begin{aligned}
&|\hat{J}_t\left(x^{s+1}, a^s\right) - J_{\alpha}^{*}\left(x_{s-K+2}^{s+1}, a_{s-K+2}^{s}\right)| \\
&\quad = \Bigg|\min_{a_{s+1}} \sum_{x_{s+2}} \hat{P}_t(x_{s+2}|a_{s+1}) \\
&\qquad\qquad \times \left[g_t(a_{s+1}, x_{s+2}) + \alpha \hat{J}_t(x_{s+2}, a_{s+1})\right] \\
&\qquad\qquad - \min_{a_{s+1}} \sum_{x_{s+2}} P(x_{s+2}|a_{s+1}) \\
&\qquad\qquad \times \left[g_t(a_{s+1}, x_{s+2}) + \alpha J_{\alpha}^{*}(x_{s+2}, a_{s+1})\right]\Bigg|.
\end{aligned}$$

Observe that, for any $v, w : \mathcal{A} \to \mathbb{R}$

$$\left|\min_{a} v(a) - \min_{a} w(a)\right| \leq \max_{a} |v(a) - w(a)|.$$

Then

$$\begin{aligned}
&|\hat{J}_t\left(x^{s+1}, a^s\right) - J_{\alpha}^{*}\left(x_{s-K+2}^{s+1}, a_{s-K+2}^{s}\right)| \\
&\quad \leq \max_{a_{s+1}} \Bigg|\sum_{x_{s+2}} \hat{P}_t(x_{s+2}|a_{s+1}) \\
&\qquad\qquad \times \left[g_t(a_{s+1}, x_{s+2}) + \alpha \hat{J}_t(x_{s+2}, a_{s+1})\right] \\
&\qquad\qquad - \sum_{x_{s+2}} P(x_{s+2}|a_{s+1}) \\
&\qquad\qquad \times \left[g_t(a_{s+1}, x_{s+2}) + \alpha J_{\alpha}^{*}(x_{s+2}, a_{s+1})\right]\Bigg|.
\end{aligned}$$

It follows that

$$
\begin{aligned}
&|\hat{J}_t\left(x^{s+1},a^s\right)-J^*_\alpha\left(x^{s+1}_{s-K+2},a^s_{s-K+2}\right)|\\
&\quad\le 2g_{\max}\bar{\epsilon}\\
&\qquad+\alpha\max_{a_{s+1}}\left|\sum_{x_{s+2}}\left[\hat{P}_t(x_{s+2}|a_{s+1})\hat{J}_t(x_{s+2},a_{s+1})\right.\right.\\
&\qquad\qquad\left.\left.-P(x_{s+2}|a_{s+1})J^*_\alpha(x_{s+2},a_{s+1})\right]\right|\\
&\quad\le 2g_{\max}\bar{\epsilon}\\
&\qquad+\alpha\max_{a_{s+1}}\left|\sum_{x_{s+2}}\hat{J}_t(x_{s+2},a_{s+1})\right.\\
&\qquad\qquad\left.\times\left[\hat{P}_t(x_{s+2}|a_{s+1})-P(x_{s+2}|a_{s+1})\right]\right|\\
&\qquad+\left|\sum_{x_{s+2}}P(x_{s+2}|a_{s+1})\right.\\
&\qquad\qquad\left.\times\left[J^*_\alpha(x_{s+2},a_{s+1})-\hat{J}_t(x_{s+2},a_{s+1})\right]\right|.
\end{aligned}
$$

Using the fact that $|\hat{J}_t|<g_{\max}/(1-\alpha)$, since it represents a discounted sum

$$
\begin{aligned}
&|\hat{J}_t\left(x^{s+1},a^s\right)-J^*_\alpha\left(x^{s+1}_{s-K+2},a^s_{s-K+2}\right)|\\
&\quad\le 2g_{\max}\bar{\epsilon}\left(1+\frac{\alpha}{1-\alpha}\right)\\
&\qquad+\alpha\max_{a_{s+1},x_{s+2}}\left|J^*_\alpha(x_{s+2},a_{s+1})-\hat{J}_t(x_{s+2},a_{s+1})\right|.
\end{aligned}
$$

We can repeat this same analysis on the $|J^*_\alpha(x_{s+2},a_{s+1})-\hat{J}_t(x_{s+2},a_{s+1})|$ term. Continuing this $\bar{K}$ times, we reach the expression

$$
\begin{aligned}
&|\hat{J}_t\left(x^{s+1},a^s\right)-J^*_\alpha\left(x^{s+1}_{s-K+2},a^s_{s-K+2}\right)|\\
&\quad\le 2g_{\max}\bar{\epsilon}\left(1+\frac{\alpha}{1-\alpha}\right)\sum_{\ell=0}^{\bar{K}-1}\alpha^\ell+\frac{\alpha^{\bar{K}}g_{\max}}{1-\alpha}\\
&\quad\le\frac{2g_{\max}\bar{\epsilon}}{1-\alpha}\left(1+\frac{\alpha}{1-\alpha}\right)+\frac{\alpha^{\bar{K}}g_{\max}}{1-\alpha}. \qquad (11)
\end{aligned}
$$

It is clear that we can pick $\bar{\epsilon}$ sufficiently small and $\bar{K}$ sufficiently large so that $\bar{\epsilon}<\epsilon/2$ and the right-hand side of (11) is less than ϵ. Such a choice will ensure that (9)–(10) hold, and hence the requirements of the lemma. ■

Lemma 1 provides sufficient conditions to guarantee when the active LZ algorithm can be expected to select the correct action given a current context of $\left(x^s,a^{s-1}\right)$. The sufficient conditions are a requirement the length of the current context, and on the context counts and probability estimates over all contexts (up to a certain length) that have $\left(x^s,a^{s-1}\right)$ as a prefix.

We would like to characterize when these conditions hold. Motivated by Lemma 1, we define the following events for ease of exposition.

Definition 1 ($\bar{\epsilon}$ -One-Step Inaccuracy): Define $\mathcal{I}^{\bar{\epsilon}}_t$ to be the event that, at time t, at least one of the following holds:

(i) $\mathrm{TV}\left(\hat{P}_t\left(\cdot|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}}\right),P\left(\cdot|X^t_{t-K+1},A^t_{t-K+1}\right)\right)>\bar{\epsilon}$.

(ii) The current context $\left(X^t_{\tau_{c(t)}},A^{t-1}_{\tau_{c(t)}}\right)$ has never been visited prior to time t.

If the event $\mathcal{I}^{\bar{\epsilon}}_t$ holds, then at time t the algorithm either possesses an estimate of the next-step transition probability $\hat{P}_t\left(\cdot|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}}\right)$ that is more than $\bar{\epsilon}$ inaccurate relative to the true transition probabilities, under the total variation metric, or else these probabilities have never been updated from their initial values.

Definition 2 ($\bar{\epsilon}$, $\bar{k}$ -Inaccuracy): Define $\mathcal{B}^{\bar{\epsilon},\bar{K}}_t$ to be the event that, at time $t\ge K$, either

(i) The length $d(t)$ of the current context is less than K.

(ii) There exist ℓ and $\left(x^\ell,a^\ell\right)$ such that

(a) $d(t)\le\ell\le d(t)+\bar{K}$

(b) $\left(x^\ell,a^\ell\right)$ contains the current context $\left(X^t_{\tau_{c(t)}},A^{t-1}_{\tau_{c(t)}}\right)$ as a prefix, that is

$$x^{d(t)}=X^t_{\tau_{c(t)}},\quad a^{d(t)-1}=A^{t-1}_{\tau_{c(t)}}.$$

(c) The estimated transition probabilities $\hat{P}_t\left(\cdot|x^\ell,a^\ell\right)$ are more than $\bar{\epsilon}$ inaccurate, under the total variation metric, and/or the context $\left(x^\ell,a^{\ell-1}\right)$ has never been visited prior to time t.

From Lemma 1, it follows that if the event $\mathcal{B}^{\bar{\epsilon},\bar{K}}_t$ *does not* hold, then the algorithm has sufficiently accurate probability estimates in order to make an optimal decision at time t.

Our analysis of the active LZ algorithm proceeds in two broad steps:

1) In Section IV-C, we establish that $\bar{\epsilon}$-one-step inaccuracy occurs a vanishing fraction of the time. Next, we show that this, in fact, suffices to establish that $\bar{\epsilon},\bar{K}$-inaccuracy also occurs a vanishing fraction of the time. By Lemma 1, this implies that, when the algorithm chooses to exploit, the selected action is suboptimal only a vanishing fraction of the time.
2) In Section IV-D, by further controlling the exploration rate appropriately, we can use these results to conclude that the algorithm attains the optimal average cost.

C. Approximating Transition Probabilities

We digress briefly, to discuss a result from the theory of universal prediction: given an arbitrary sequence $\{y_t\}$, with $y_t\in\mathbb{Y}$ for some finite alphabet $\mathbb{Y}$, consider the problem of making sequential probability assignments $Q_{t-1}(\cdot)$ over $\mathbb{Y}$, given the entire sequence observed up to and including time $t-1$, y^{t-1}, so as to minimize the cost function $\sum_{t=1}^T-\log Q_{t-1}(y_t)$, for some horizon T. It has been shown by Krichevsky and Trofimov [11] that the assignment

$$Q_t(y)\triangleq\frac{N_t(y)+1/2}{t+|\mathbb{Y}|/2} \qquad (12)$$

where $N_t(y)$ is the number of occurrences of the symbol y up to time t, achieves the following.

Lemma 2:

$$-\sum_{t=1}^{t}\log Q_{t-1}(y_t)-\min_{q\in\mathcal{M}(\mathbb{Y})}\left[-\sum_{t=1}^{t}\log Q(y_t)\right]\leq\frac{|\mathbb{Y}|}{2}\log T+O(1)$$

where the minimization in taken over the set $\mathcal{M}(\mathbb{Y})$ of all probability distributions on $\mathbb{Y}$.

Lemma 2 provides a bound on the performance of the sequential probability assignment (12) versus the performance of the best constant probability assignment, made with knowledge of the full sequence y^T. Notice that (12) is precisely the one-step transition probability estimate employed at each context by the active LZ algorithm (line 15).

Returning to our original setting, define $p_{\min}$ to be the smallest element of the set of non-zero transition probabilities

$$\left\{P\left(x_{K+1}|x^K,a^K\right):P\left(x_{K+1}|x^K,a^K\right)>0\right\}.$$

The proof of the following lemma essentially involves invoking Lemma 2 at each context encountered by the algorithm, the use of a combinatorial lemma (Ziv's inequality), and the use of the Azuma-Hoeffding inequality (see, for example, [13]). Part of the proof is motivated by results on Lempel-Ziv based prediction obtained by Feder *et al.* [14].

Lemma 3: For arbitrary $\epsilon'>0$

$$\Pr\left(\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\mathcal{I}_t^{\bar\epsilon}}\geq\frac{K_1}{2\bar\epsilon^2}\frac{\log\log T}{\log T}+\frac{\epsilon'}{2\bar\epsilon^2}\right)\leq\exp\left(-\frac{T\epsilon'^2}{8\log^2}\left((2T+|\mathbb{X}|)/p_{\min}\right)\right)$$

where K_1 is a constant that depends only on $|\mathbb{X}|$ and $|A|$

Proof: See Appendix A. ∎

Lemma 3 controls the fraction of the time that the active LZ algorithm is $\bar\epsilon$-one-step inaccurate. In particular, Lemma 3 is sufficient to establish that this fraction of time goes to 0 (via a use of the first Borel-Cantelli lemma) and also gives us a rate of convergence.

It turns out that if the exploration rate γ_t decays sufficiently slowly, this suffices to ensure that the fraction of time the algorithm is $\bar\epsilon,\bar K$-inaccurate goes to 0 as well. To see this, suppose that the current context at time t is $\left(X_{\tau_{c(t)}}^{t},A_{\tau_{c(t)}}^{t-1}\right)=\left(x^s,a^{s-1}\right)$, and that the algorithm is $\bar\epsilon,\bar K$-inaccurate (i.e., the event $\mathcal{B}_t^{\bar\epsilon,\bar K}$ holds). Then, one of two things must be the case:

- The current context length s is less than $\bar K$. We will demonstrate that this happens only a vanishing fraction of the time.
- There exists $\left(x^\ell,a^\ell\right)$, with $s\leq\ell\leq s+\bar K$, so that either the estimated transition probability distribution $\hat P_t\left(\cdot|x^\ell,a^\ell\right)$ is $\bar\epsilon$ inaccurate under the total variation metric, or the context $\left(x^\ell,a^{\ell-1}\right)$ has never been visited in the past. The probability that the realized sequence of future observations and actions $\left(X_{t+1}^{t+\ell-s},A_t^{t+\ell-s}\right)$ will indeed correspond to $\left(x_{s+1}^{\ell},a_s^{\ell}\right)$ is at least

$$p_{\min}^{\ell-s}\prod_{m=t}^{t+\ell-s}\gamma_m$$

where $p_{\min}$ is the smallest nonzero transition probability. Thus, with this minimum probability, a $\bar\epsilon$-one-step inaccurate time will occur before the time $t+\bar K$. Then, if the exploration probabilities $\{\gamma_m\}$ decays sufficiently slowly, would be impossible for the fraction of $\bar\epsilon$-one-step inaccurate times to go to 0 without the fraction of $\bar\epsilon,\bar K$-inaccurate times also going to 0.

By making these arguments precise we can prove the following lemma. The lemma states that the fraction of time we are at a context wherein the assumptions of Lemma 1 are not satisfied goes to 0 almost surely.

Lemma 4: Assume that

$$\gamma_t\geq(a_1/\log t)^{1/(a_2\bar K)}$$

for arbitrary constants $a_1>0$ and $a_2>1$. Further assume that $\{\gamma_t\}$ is nonincreasing. Then

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\mathcal{B}_t^{\bar\epsilon,\bar K}}=0,\quad a.s.$$

Proof: First, we consider the instances of time where the current context length is less than K. Note that

$$\sum_{t=K}^{T}\mathbb{I}_{\{d(t)<K\}}\leq\sum_{c=1}^{c(T)}\sum_{t=\tau_c}^{\tau_{c+1}-1}\mathbb{I}_{\{t-\tau_c+1<K\}}\leq\sum_{c=1}^{c(T)}K=Kc(T).$$

Applying Ziv's inequality (Lemma 5)

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\{d(t)<K\}}\leq\lim_{T\to\infty}\frac{KC_2}{\log T}=0.\qquad(13)$$

Next, define B_t to be the event that an $\bar\epsilon$-one-step inaccurate time occurs between t and $t+\bar K$ inclusive, that is

$$B_t\triangleq\bigcup_{s=t}^{t+\bar K}\mathcal{I}_s^{\bar\epsilon}.$$

It is easy to see that

$$\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{B_t}\leq\frac{\bar K+1}{T}\sum_{t=K}^{T+\bar K}\mathbb{I}_{\mathcal{I}_t^{\bar\epsilon}}\leq\frac{\bar K+1}{T}\sum_{t=K}^{T}\mathbb{I}_{\mathcal{I}_t^{\bar\epsilon}}+\frac{(\bar K+1)^2}{T}.$$

From Lemma 3, we immediately have, for arbitrary $\epsilon' > 0$

$$\Pr\left(\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{B_t} \geq \frac{(\bar{K}+1)K_1}{2\bar{\epsilon}^2}\frac{\log\log T}{\log T} + \frac{(\bar{K}+1)\epsilon'}{2\bar{\epsilon}^2} + \frac{(\bar{K}+1)^2}{T}\right) \leq \exp\left(-\frac{T\epsilon'^2}{8\log^2\left((2T+|\mathbb{X}|)/p_{\min}\right)}\right). \quad (14)$$

Define H_t to be the event that $\mathcal{B}_t^{\bar{\epsilon},\bar{K}}$ holds, but $d(t) \geq K$. The event H_t holds when, at time t, there exists some context, up to $\bar{K}$ levels below the current context, which is $\bar{\epsilon}$-one-step inaccurate. Such a context will be visited with probability at least

$$p_{\min}^{\bar{K}}\prod_{m=t}^{t+\bar{K}}\gamma_m \geq (p_{\min}\gamma_{t+\bar{K}})^{\bar{K}+1}$$

in which case B_t holds. Consequently

$$\mathsf{E}[\mathbb{I}_{B_t}|\mathcal{F}_t] \geq (p_{\min}\gamma_{t+\bar{K}})^{\bar{K}+1}\mathbb{I}_{H_t}.$$

Since γ_t is nonincreasing

$$\frac{1}{T}\sum_{t=K}^{T}\mathsf{E}[\mathbb{I}_{B_t}|\mathcal{F}_t] \geq \frac{(p_{\min}\gamma_{T+\bar{K}-1})^{\bar{K}+1}}{T}\sum_{t=K}^{T}\mathbb{I}_{H_t}. \quad (15)$$

Now define, for $i = 0, 1, \ldots, \bar{K}-1$ and $n \geq 0$, martingales $M_n^{(i)}$ adapted to $\mathcal{G}_n^{(i)} = \mathcal{F}_{K+n\bar{K}+i}$, according to $M_0^{(i)} = 0$, and, for $n > 0$

$$M_n^{(i)} \triangleq \sum_{j=0}^{n-1}\mathbb{I}_{B_{K+j\bar{K}+i}} - \mathsf{E}[\mathbb{I}_{B_{K+j\bar{K}+i}}|\mathcal{G}_j^{(i)}].$$

Since $\left|M_n^{(i)} - M_{n-1}^{(i)}\right| \leq 2$, we have via the Azuma-Hoeffding inequality, for arbitrary $\epsilon'' > 0$

$$\Pr\left(M_n^{(i)} \geq n\epsilon''\right) \leq \exp\left(-n\epsilon''^2/8\right). \quad (16)$$

For each i, let $n_i(T)$ be the largest integer such that $K + n_i(T)\bar{K} + i \leq T$, so that

$$\sum_{t=K}^{T}\mathbb{I}_{B_t} - \mathsf{E}[\mathbb{I}_{B_t}|\mathcal{F}_t] = \sum_{i=0}^{\bar{K}-1}M_{n_i(T)}^{(i)}.$$

Since $n_i(T) \leq \frac{T}{\bar{K}}$, the union bound along with (16) then implies that

$$\begin{aligned}\Pr&\left(\sum_{t=K}^{T}\mathbb{I}_{B_t} - \mathsf{E}[\mathbb{I}_{B_t}|\mathcal{F}_t] \geq T\epsilon''\right)\\ &\leq \sum_{i=0}^{\bar{K}-1}\Pr\left(M_{n_i(T)}^{(i)} \geq T\epsilon''/\bar{K}\right)\\ &\leq \sum_{i=0}^{\bar{K}-1}\exp\left(-T^2\epsilon''^2/8\bar{K}^2 n_i(T)\right)\\ &\leq \bar{K}\exp\left(-T\epsilon''^2/8\bar{K}\right). \end{aligned} \quad (17)$$

Now, define

$$\kappa(T) \triangleq \frac{1}{(p_{\min}\gamma_{T+\bar{K}-1})^{\bar{K}+1}}\left[\frac{(\bar{K}+1)K_1}{2\bar{\epsilon}^2}\frac{\log\log T}{\log T} + \frac{(\bar{K}+1)\epsilon'(T)}{2\bar{\epsilon}^2} + \frac{(\bar{K}+1)^2}{T} + \epsilon''(T)\right]$$

with

$$\epsilon'(T) \triangleq \frac{1}{\log T}, \quad \epsilon''(T) \triangleq \frac{1}{\log T}.$$

It follows from (14), (15), and (17) that

$$\begin{aligned}\Pr&\left(\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{H_t} \geq \kappa(T)\right)\\ &\leq \exp\left(-\frac{T}{8\log^4\left((2T+|\mathbb{X}|)/p_{\min}\right)}\right)\\ &\quad + \bar{K}\exp\left(-\frac{T}{8\bar{K}\log^2 T}\right).\end{aligned}$$

By the first Borel-Cantelli lemma

$$\Pr\left(\frac{1}{T}\sum_{t=K}^{T-1}\mathbb{I}_{H_t} \geq \kappa(T), \text{ i.o.}\right) = 0.$$

Note that the hypothesis on γ_t implies that $\kappa(T) \to 0$ as $T \to \infty$. Then

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=K}^{T-1}\mathbb{I}_{H_t} = 1, \quad \text{a.s.} \quad (18)$$

Finally, note that

$$\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\mathcal{B}_t^{\bar{\epsilon},\bar{K}}} \leq \frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\{d(t)<K\}} + \frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{H_t}.$$

The result then follows from (13) and (18). ■

D. Average Cost Optimality

Observe that if the active LZ algorithm chooses an action that is nonoptimal at time t, that is

$$A_t \notin \mathcal{A}_\alpha^*\left(X_{t-K+1}^t, A_{t-K+1}^{t-1}\right)$$

then, either the event $\mathcal{B}_t^{\bar{\epsilon},\bar{K}}$ holds or the algorithm chose to explore. Lemma 4 guarantees that the first possibility happens a vanishing fraction of time. Further, if $\gamma_t \downarrow 0$, then the algorithm will explore a vanishing fraction of time. Combining these observations give us the following theorem.

Theorem 1: Assume that

$$\gamma_t \geq (a_1/\log t)^{1/(a_2\bar{K})}$$

for arbitrary constants $a_1 > 0$ and $a_2 > 1$. Further, assume that $\gamma_t \downarrow 0$. Then

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\{A_t \notin \mathcal{A}_\alpha^*(X_{t-K+1}^t, A_{t-K+1}^{t-1})\}} = 0, \quad \text{a.s.}$$

Proof: Given a sequence of independent bounded random variables $\{Z_n\}$, with $\mathsf{E}[Z_n] \to 0$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} Z_n = 0, \quad \text{a.s.}$$

This follows, for example, from the Azuma-Hoeffding inequality followed by the first Borel-Cantelli lemma. This immediately yields

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=k}^{T-1} \mathbb{I}_{\{\text{exploration at time } t\}} \to 0, \quad \text{a.s.} \tag{19}$$

provided $\gamma_t \to 0$ (note that the choice of exploration at each time t is independent of all other events). Now observe that

$$\left\{A_t \notin \mathcal{A}^*_\alpha(X^t_{t-K+1}, A^{t-1}_{t-K+1})\right\} \subset \mathcal{B}^{\bar{\epsilon},\bar{K}}_{t-1} \cup \{\text{ exploration at time } t\}.$$

Combining (19) with Lemma 4, the result follows. ∎

Assumption 1 guarantees the optimal average cost is λ^*, independent of the initial state of the Markov chain, and that there exists a stationary policy that achieves the optimal average cost λ^*. By the ergodicity theorem, under such a optimal policy

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g(X_t, A_t, X_{t+1}) = \lambda^*, \quad \text{a.s.} \tag{20}$$

On the other hand, Theorem 1 suggests that, under the active LZ algorithm, the fraction of time at which nonoptimal decisions are made vanishes asymptotically. Combining these facts yields our main result.

Theorem 2: Assume that

$$\gamma_t \geq (a_1/\log t)^{1/(a_2 \bar{K})}$$

for arbitrary constants $a_1 > 0$ and $a_2 > 1$, and that $\gamma_t \downarrow 0$. Then, for $\alpha \in (0,1)$ sufficiently close to 1

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g(X_t, A_t, X_{t+1}) = \lambda^*, \quad \text{a.s.}$$

under the active LZ algorithm. Hence, the active LZ algorithm achieves an asymptotically optimal average cost regardless of the underlying transition kernel.

Proof: Without loss of generality, assume that the cost $g(X_t, A_t, X_{t+1})$ does not depend on X_{t+1}.

Fix $\epsilon > 0$, and consider an interval of time $T_\epsilon > K$. For each $\left(x^K, a^K\right) \in \mathbb{X}^K \times \mathbb{A}^K$, define a coupled process $\left(\tilde{X}_t\left(x^k, a^k\right), \tilde{A}_t\left(x^K, a^K\right)\right)$ as follows. For every integer n, set

$$\tilde{X}^{(n-1)T_\epsilon+K}_{(n-1)T_\epsilon+1}(x^K, a^K) = x^K_1$$
and
$$\tilde{A}^{(n-1)T_\epsilon+K}_{(n-1)T_\epsilon+1}(x^K, a^K) = a^K_1.$$

For all other times t, the coupled processes will choose actions according to an optimal stationary policy, that is

$$\tilde{A}_t\left(x^K, a^K\right) \in \mathcal{A}^*_\alpha\left(\tilde{X}^t_{t-K+1}\left(x^K, a^K\right), \tilde{A}^{t-1}_{t-K+1}\left(x^K, a^K\right)\right).$$

Without loss of generality, we will assume that the choice of action is unique.

Now, for each n there will be exactly one $\left(x^K, a^K\right)$ that matches the original process (X_t, A_t) over times $(n-1)T_\epsilon + 1 \leq t \leq (n-1)T_\epsilon + K$, that is

$$\left(x^K, a^K\right) = \left(X^{(n-1)T_\epsilon+K}_{(n-1)T_\epsilon+1}, A^{(n-1)T_\epsilon+K}_{(n-1)T_\epsilon+1}\right).$$

For the process indexed by $\left(x^K, a^K\right)$, for $(n-1)T_\epsilon + K < t \leq nT_\epsilon$, if

$$\left(\tilde{X}^{t-1}_{t-K}(x^K, a^K), \tilde{A}^{t-1}_{t-K}(x^K, a^K)\right) = (X^{t-1}_{t-K}, A^{t-1}_{t-K})$$

then set $\tilde{X}_t\left(x^K, a^K\right) = X_t$. Otherwise, allow $\tilde{X}_t\left(x^K, a^K\right)$ to evolve independently according to the process transition probabilities. Similarly, allow all other the processes to evolve independently according to the proper transition probabilities. Define

$$G_n\left(x^k, a^k\right) \triangleq \frac{1}{T_\epsilon} \sum_{t=(n-1)T_\epsilon+1}^{nT_\epsilon} g\left(\tilde{X}_t\left(x^K, a^K\right), \tilde{A}_t\left(x^K, a^K\right)\right).$$

Note that each $G_n\left(x^K, a^K\right)$ is the average cost under an *optimal* policy. Therefore, because of (20), we can pick T_ϵ large enough so that for any n

$$\mathsf{E}\left[\max_{x^K, a^K} \left|G_n(x^K, a^K) - \lambda^*\right|\right] < \epsilon. \tag{21}$$

Define $\mathcal{Z}_n$ to be the event that, within the nth interval, the algorithm chooses a nonoptimal action. That is

$$\mathcal{Z}_n \triangleq \left\{\exists\, t,\ (n-1)T_\epsilon < t \leq nT_\epsilon,\ A_t \notin \mathcal{A}^*_\alpha(X^t, A^{t-1})\right\}.$$

Set

$$E_N = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}_{\mathcal{Z}_n}.$$

Then,

$$\left|\frac{1}{NT_\epsilon} \sum_{t=1}^{NT_\epsilon} (g(X_t, A_t) - \lambda^*)\right| \leq \frac{\max(|g_{\max} - \lambda^*|, \lambda^*) E_N}{N} + \left|\frac{1}{NT_\epsilon} \sum_{n=1}^{N} (1 - \mathbb{I}_{\mathcal{Z}_n}) \sum_{t=(n-1)T_\epsilon+1}^{nT_\epsilon} (g(X_t, A_t) - \lambda^*)\right|.$$

Note that, from Theorem 1, $E_N/N \to 0$ almost surely as $N \to \infty$. Thus

$$\limsup_{N\to\infty}\left|\frac{1}{NT_\epsilon}\sum_{t=1}^{NT_\epsilon}(g(X_t,A_t)-\lambda^*)\right| \le \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}(1-\mathbb{I}_{\mathcal{Z}_n}) \times\left|\frac{1}{T_\epsilon}\sum_{t=(n-1)T_\epsilon+1}^{nT_\epsilon}(g(X_t,A_t)-\lambda^*)\right|.$$

Notice that when $\mathbb{I}_{\mathcal{Z}_n}=0$, we have for some $\left(x^K,a^K\right)$ that $\hat{X}_t\left(x^K,a^K\right)=X_t$ for all $(n-1)T_\epsilon < t \le nT_\epsilon$. Thus

$$\limsup_{N\to\infty}\left|\frac{1}{NT_\epsilon}\sum_{t=1}^{NT_\epsilon}(g(X_t,A_t)-\lambda^*)\right| \le \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}(1-\mathbb{I}_{\mathcal{Z}_n})\max_{x^K,a^K}\left|G_n\left(x^K,a^K\right)-\lambda^*\right| \le \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\max_{x^K,a^K}\left|G_n\left(x^K,a^K\right)-\lambda^*\right|.$$

However, the variables

$$\max_{x^K,a^K}\left|G_n\left(x^K,a^K\right)-\lambda^*\right|$$

are independent and identically distributed as n varies. Thus, by the Strong Law of Large Numbers and (21)

$$\limsup_{T\to\infty}\left|\frac{1}{T}\sum_{t=1}^{T}(g(X_t,A_t)-\lambda^*)\right| \le \epsilon$$

with probability 1. Since ϵ was arbitrary, the result follows. ■

E. Choice of Discount Factor

Given a choice of α sufficiently close to 1, the optimal α-discounted cost policy coincides with the average cost optimal policy. Our presentation thus far has assumed knowledge of such an α. For a given α, under the assumptions of Theorem 1, The active LZ algorithm is guaranteed to take α-discounted optimal actions a fraction 1 of the time which for an *ad hoc* choice of α sufficiently close to 1 is likely to yield good performance. Nonetheless, one may use a "doubling-trick" in conjunction with the active LZ algorithm to attain average cost optimality without knowledge of α. In particular, consider the following algorithm that uses the active LZ algorithm, with the choice of $\{\gamma_t\}$ as stipulated by Theorem 1, as a subroutine.

Here β_k is a sequence that approaches 0 sufficiently slowly. One can show that if $\beta_k = \Omega(1/\log\log k)$, then the above scheme achieves average cost optimality. A rigorous proof of this fact would require repetition of arguments we have used to prove earlier results. As such, we only provide a sketch that outlines the steps required to establish average cost optimality.

We begin by noting that in the kth epoch of Algorithm 2, one choice (so that Lemma 1 remains true) is to let $\bar{\epsilon}_k, \bar{K}_k$ grow as α approaches 1 according to $\bar{\epsilon}_k = \Omega(1)$ and $\bar{K}_k = \Omega(1/\beta_k)$, respectively. If $\beta_k = \Omega(1/\log\log k)$, then for the kth epoch of Algorithm 2, Lemma 4 is easily modified to show that with high probability the greedy action is suboptimal over less than $2^k\kappa\left(2^k\right)$ time steps where $\kappa\left(2^k\right) = O\left(\left(\log\log 2^k\right)^3/\log 2^k\right)$. The Borel-Cantelli Lemma may then be used to establish that beyond some finite epoch, over all subsequent epochs k, the greedy action is suboptimal over at most $2^k\kappa\left(2^k\right)$ time steps. Provided $\beta_k \to 0$, this suffices to show that the greedy action is optimal a fraction 1 of the time. Provided one decreases exploration probabilities sufficiently quickly, this in turn suffices to establish average cost optimality.

Algorithm 2 The active LZ with a doubling scheme.

1: **for** nonnegative integers k do

2: **for** each time $2^k \le t' < 2^{k+1}$ do

3: Apply the active LZ algorithm (Algorithm 1) with $\alpha = 1-\beta_k$, and time index $t = t' - 2^k$.

4: **end for**

5: **end for**

F. On the Rate of Convergence

We limit our discussion to the rate at which the fraction of time the active LZ algorithm takes sub-optimal actions goes to zero; even assuming one selects optimal actions at every point in time, the rate at which average costs incurred converge to λ^* are intimately related to the structure of P which is a somewhat separate issue. Now the proofs of Lemma 4 and Theorem 1 tell us that the fraction of time the active LZ algorithm selects suboptimal actions goes to zero at a rate that is $O((1/\log T)^c)$ where c is some constant less than 1. The proofs of Lemmas 3 and 4 reveal that the determining factor of this rate is effectively the rate at which the transition probability estimates provided by $\hat{P}$ converge to their true values. Thus while the rate at which the fraction of sub-optimal action selections goes to zero is slow, this rate isn't surprising and is shared with many Lempel-Ziv schemes used in prediction and compression.

A natural direction for further research is to explore the effect of replacing the LZ-based context tree data structure by the context-tree weighting method of Willems *et al.* [15]. It seems plausible to expect that such an approach will yield algorithms with significantly improved convergence rates, as is the case in data compression and prediction.

V. Conclusion

We have presented and established the asymptotic optimality of a Lempel-Ziv inspired algorithm for learning. The algorithm is a natural combination of ideas from information theory and dynamic programming. We hope that these ideas, in particular

the use of a Lempel-Ziv tree to model an unknown probability distribution, can find other uses in reinforcement learning.

One interesting special case to consider is when the next observation is Markovian given the past K observations and only the latest action. In this case, a variation of the active LZ algorithm that uses contexts of the form (x^s, a) could be used. Here, the resulting tree would have exponentially fewer nodes and would be much quicker to converge to the optimal policy.

A number of further issues are under consideration. It would be of great interest to develop theoretical bounds for the rate of convergence. Also, it would be natural to extend the analysis of our algorithms to systems with possibly infinite dependence on history. One such extension would be to mixing models, such as those considered by Jacquet, *et al.* [8]. Another would be to consider the optimal control of a partially observable Markov decision process.

APPENDIX
PROOF OF LEMMA 3

An important device in the proof of Lemma 3 the following combinatorial lemma. A proof can be found in Cover and Thomas [16].

Lemma 5 (Ziv's Inequality): The number of contexts seen by time T, $c(T)$, satisfies

$$c(T) \leq \frac{C_2 T}{\log T}$$

where C_2 is a constant that depends only on $|\mathbb{X}|$ and $|\mathbb{A}|$.

Without loss of generality, assume that X_t and A_t take some fixed but arbitrary values of $-K+2 \leq t \leq 0$, so that the expression $P\left(X_{t+1}|X_{t-K+1}^t, A_{t-K+1}^t\right)$ is well-defined for all $t \geq 1$. We will use Lemma 2 to show the following.

Lemma 6:

$$\begin{aligned} &-\sum_{t=1}^{t} \log \hat{p}_t\left(x_{t+1}|x_{\tau_{c(t)}}^t, a_{\tau_{c(t)}}^t\right) \\ &\quad \leq -\sum_{t=1}^{t} \log p\left(x_{t+1}|X_{t-k+1}^t, a_{t-k+1}^t\right) \\ &\qquad + \bar{k}_1 T \frac{\log\log T}{\log T} \end{aligned}$$

where $\bar{K}_1$ is a positive constant that depends only on $|\mathbb{X}|$ and $|\mathbb{A}|$.

Proof: Observe that the probability assignment made by our algorithm is equivalent to using (12) at every context. In particular, at every time t

$$\hat{P}_t(X_{t+1}|X_{\tau_{c(t)}}^t, A_{\tau_{c(t)}}^t) = \frac{N_t(X_{\tau_{c(t)}}^{t+1}, A_{\tau_{c(t)}}^t) + 1/2}{\sum_x N_t((X_{\tau_{c(t)}}^t, x), A_{\tau_{c(t)}}^t) + |\mathbb{X}|/2}.$$

For each $\left(x^j, a^j\right)$, define $\mathcal{T}_T\left(x^j, a^j\right)$ to be the set of times

$$\mathcal{T}_T\left(x^j, a^j\right) \triangleq \left\{t : 1 \leq t \leq T,\ \left(X_{\tau_{c(t)}}^t, A_{\tau_{c(t)}}^t\right) = \left(x^j, a^j\right)\right\}.$$

It follows from Lemma 2 that

$$\begin{aligned} &-\sum_{t \in \mathcal{T}_T(x^j, a^j)} \log \hat{P}_t\left(X_{t+1}|X_{\tau_{c(t)}}^t, A_{\tau_{c(t)}}^t\right) \\ &\quad \leq \min_{p \in \mathcal{M}(\mathbb{X})} -\sum_{t \in \mathcal{T}_T(x^j, a^j)} \log p(X_{t+1}) \\ &\qquad + \frac{|\mathbb{X}|}{2} \log |\mathcal{T}_T\left(x^j, a^j\right)| + C_1. \end{aligned}$$

Summing this expression over all distinct $\left(x^j, a^j\right)$ that have occurred up to time T

$$\begin{aligned} &-\sum_{t=1}^{T} \log \hat{P}_t\left(X_{t+1}|X_{\tau_{c(t)}}^t, A_{\tau_{c(t)}}^t\right) \\ &\quad \leq \sum_{(x^j, a^j)} \left[\min_{p \in \mathcal{M}(\mathbb{X})} -\sum_{t \in \mathcal{T}(x^j, a^j)} \log p(X_{t+1})\right] \\ &\qquad + \sum_{(x^j, a^j)} \left[\frac{|\mathbb{X}|}{2} \log |\mathcal{T}_T\left(x^j, a^j\right)| + C_1\right] \\ &\quad \leq -\sum_{t=1}^{T} \log P\left(X_{t+1}|X_{t-K+1}^t, A_{t-K+1}^t\right) \\ &\qquad + \sum_{(x^j, a^j)} \left[\frac{|\mathbb{X}|}{2} \log |\mathcal{T}_T\left(x^j, a^j\right)| + C_1\right]. \end{aligned} \tag{22}$$

Now, $c(T)$ is the total number of distinct contexts that have occurred up to time T. Note that this is also precisely the number of distinct $\left(x^j, a^j\right)$ with $|\mathcal{T}_T\left(x^j, a^j\right)| > 0$. Then, by the concavity of $\log(\cdot)$

$$\begin{aligned} &\sum_{(x^j, a^j)} \left[\frac{|\mathbb{X}|}{2} \log |\mathcal{T}_T\left(x^j, a^j\right)| + C_1\right] \\ &\qquad \leq \frac{|\mathbb{X}| c(T)}{2} \log \frac{T}{c(T)} + C_1 c(T). \end{aligned}$$

Applying Lemma 5

$$\begin{aligned} &\sum_{(x^j, a^j)} \left[\frac{|\mathbb{X}|}{2} \log |\mathcal{T}_T\left(x^j, a^j\right)| + C_1\right] \\ &\quad \leq \frac{C_2 |\mathbb{X}|}{2} \frac{T}{\log T} \left[\log\log T - \log C_2\right] \\ &\qquad + C_1 C_2 \frac{T}{\log T}. \end{aligned} \tag{23}$$

The lemma follows by combining (22) and (23). ■

For the remainder of this section, define Δ_t to be the Kullback-Leibler distance between the estimated and true transition probabilities at time t, that is

$$\Delta_t \triangleq D\left(P\left(\cdot|X_{t-K+1}^t, A_{t-K+1}^t\right) \| \hat{P}_t\left(\cdot|X_{\tau_{c(t)}}^t, A_{\tau_{c(t)}}^t\right)\right).$$

Lemma 7: For arbitrary $\epsilon' > 0$

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left[\log\frac{\hat{P}_t(X_{t+1}|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}})}{P\left(X_{t+1}|X^t_{t-K+1},A^t_{t-K+1}\right)}+\Delta_t\right]\geq\epsilon'\right)\leq\exp\left(-\frac{T\epsilon'^2}{8\log^2\left((2T+|\mathbb{X}|)/p_{\min}\right)}\right).$$

Proof: Define, for $T \geq 0$, a process $\{M_T\}$ adapted to $\mathcal{F}_{T+1}$ as follows: set with $M_0 = 0$, and, for $T > 1$

$$\begin{aligned}M_T &\triangleq \sum_{t=1}^{T}\log\frac{\hat{P}_t\left(X_{t+1}|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}}\right)}{P\left(X_{t+1}|X^t_{t-K+1},A^t_{t-K+1}\right)}\\&\quad-\sum_{t=1}^{T}\mathsf{E}\left[\log\frac{\hat{P}_t\left(X_{t+1}|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}}\right)}{P(X_{t+1}|X^t_{t-K+1},A^t_{t-K+1})}\middle|\mathcal{F}_t\right]\\&=\sum_{t=1}^{T}\left(\log\frac{\hat{P}_t\left(X_{t+1}|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}}\right)}{P\left(X_{t+1}|X^t_{t-K+1},A^t_{t-K+1}\right)}+\Delta_t\right).\end{aligned}$$

It is clear that M_T is a martingale with $\mathsf{E}[M_T] = 0$. Further

$$0\geq\log\hat{P}_t\left(X_{t+1}|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}}\right)\geq\log(1/(2t+|\mathbb{X}|)).$$

and

$$0\geq\log P\left(X_{t+1}|X^t_{t-K+1},A^t_{t-K+1}\right)\geq\log p_{\min}$$

so that

$$|M_T - M_{T-1}|\leq 2\log\left(\frac{2T+|\mathbb{X}|}{p_{\min}}\right).$$

An application of the Azuma-Hoeffding inequality then yields, for arbitrary $\epsilon' > 0$

$$\Pr\left(\frac{M_T}{T}\geq\epsilon'\right)\leq\exp\left(-\frac{T^2\epsilon'^2}{8\sum_{t=1}^{T}\log^2\left((2T+|\mathbb{X}|)/p_{\min}\right)}\right)\leq\exp\left(-\frac{T\epsilon'^2}{8\log^2\left((2T+|\mathbb{X}|)/p_{\min}\right)}\right).$$

■

We are now ready to prove Lemma 3.

Lemma 3: For arbitrary $\epsilon' > 0$

$$\Pr\left(\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\mathcal{I}_t^{\bar{\epsilon}}}\geq\frac{K_1}{2\bar{\epsilon}^2}\frac{\log\log T}{\log T}+\frac{\epsilon'}{2\bar{\epsilon}^2}\right)\leq\exp\left(-\frac{T\epsilon'^2}{8\log^2\left((2T+|\mathbb{X}|)/p_{\min}\right)}\right)$$

where K_1 is a constant that depends only on $|\mathbb{X}|$ and $|A|$.

Proof: Define

$$\Xi_t\triangleq\mathrm{TV}\left(P(\cdot|X^t_{t-K+1},A^t_{t-K+1}),\hat{P}_t\left(\cdot|X^t_{\tau_{c(t)}},A^t_{\tau_{c(t)}}\right)\right).$$

We have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\Delta_t&\geq\frac{2\bar{\epsilon}^2}{T}\sum_{t=1}^{T}\mathbb{I}_{\{\Delta_t\geq 2\bar{\epsilon}^2\}}\\&\geq\frac{2\bar{\epsilon}^2}{T}\sum_{t=1}^{T}\mathbb{I}_{\{\Xi_t>\bar{\epsilon}\}}.\end{aligned}\tag{24}$$

Here, the first inequality follows by the nonnegativity of Kullback-Leibler distance. The second inequality follows from Pinsker's inequality, which states that $\mathrm{TV}(\cdot,\cdot)\leq\sqrt{D(\cdot\|\cdot)/2}$.

Now, let F_t be the event that the current context at time t, $\left(X^t_{\tau_{c(t)}},A^{t-1}_{\tau_{c(t)}}\right)$ has never been visited in the past. Observe that, by Lemma 5

$$\sum_{t=1}^{T}\mathbb{I}_{F_t}=c(T)\leq\frac{C_2T}{\log T}.\tag{25}$$

Putting together (24) and (25) with the definition of the event $\mathcal{I}_t^{\bar{\epsilon}}$

$$\begin{aligned}\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\mathcal{I}_t^{\bar{\epsilon}}}&\leq\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{I}_{\{\Xi_t>\bar{\epsilon}\}}+\mathbb{I}_{F_t}\right)\\&\leq\frac{1}{2\bar{\epsilon}^2T}\sum_{t=1}^{T}\Delta_t+\frac{C_2}{\log T}.\end{aligned}$$

Then

$$\Pr\left(\frac{1}{T}\sum_{t=K}^{T}\mathbb{I}_{\mathcal{I}_t^{\bar{\epsilon}}}\geq\frac{\bar{K}_1}{2\bar{\epsilon}^2}\frac{\log\log T}{\log T}+\frac{\epsilon'}{2\bar{\epsilon}^2}+\frac{C_2}{\log T}\right)\leq\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\Delta_t\geq\bar{K}_1\frac{\log\log T}{\log T}+\epsilon'\right).$$

By Lemma 6 and Lemma 7, we have

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\Delta_t\geq\bar{K}_1\frac{\log\log T}{\log T}+\epsilon'\right)\leq\exp\left(-\frac{T\epsilon'^2}{8\log^2\left((2T+|\mathbb{X}|)/p_{\min}\right)}\right).$$

This yields the desired result by defining the constant $K_1 \triangleq \bar{K}_1 + C_2/\log\log K$. ■

REFERENCES

[1] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4017–4035, Sep. 2006.

[2] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. 24, pp. 530–536, 1978.

[3] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," in *Proc. 15th Int. Conf. Mach. Learn.*, San Francisco, CA, 1998, pp. 260–268.

[4] E. Even-Dar, S. M. Kakade, and Y. Mansour, "Reinforcement learning in POMDPS without resets," in *Proc. 19th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2005, pp. 660–665.

[5] D. P. de Farias and N. Megiddo, "Combining expert advice in reactive environments," *J. ACM*, vol. 53, no. 5, pp. 762–799, 2006.

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

[7] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 10, pp. 2124–2147, Oct. 1998.
[8] P. Jacquet, W. Szpankowski, and I. Apostol, "A universal predictor based on pattern matching," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1462–1472, Jun. 2002.
[9] J. S. Vitter and P. Krishnan, "Optimal prefetching via data compression," *J. ACM*, vol. 43, no. 5, pp. 771–793, 1996.
[10] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA: Athena Scientific, 2006, vol. 2.
[11] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. 27, no. 3, pp. 199–207, Mar. 1981.
[12] E. Martinian, RoShamBo and Data Compression 2000 [Online]. Available: http://www.csua.berkeley.edu/~emin/writings/lz_rps/index.html
[13] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer, 1998.
[14] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inf. Theory*, vol. 38, no. 7, pp. 1258–1270, Jul. 1992.
[15] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 5, pp. 653–664, May 1995.
[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.

Vivek F. Farias received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 2007.

He is the Robert N. Noyce Career Development Assistant Professor of Management at the Massachusetts Institute of Technology (MIT), Cambridge.

Dr. Farias is a recipient of an IEEE Region 6 Undergraduate Student Paper Prize (2002), a Stanford School of Engineering Fellowship (2002), an INFORMS MSOM Student Paper Prize (2006), and an MIT Solomon Buchsbaum Award (2008).

Ciamac C. Moallemi (M'07) received the S.B. degrees in electrical engineering and computer science and in mathematics from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996. He received the Certificate of Advanced Study in Mathematics, with distinction, in 1997, from the University of Cambridge, U.K. He received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 2007.

He is an Assistant Professor with the Graduate School of Business of Columbia University, New York, NY, where he has been since 2007.

Dr. Moallemi is a member of INFORMS. He is the recipient of a British Marshall Scholarship (1996) and a Benchmark Stanford Graduate Fellowship (2003).

Benjamin Van Roy (M'99–SM'07) received the S.B. degree in 1993 in computer science and engineering and the S.M. and Ph.D. degrees in 1995 and 1998, respectively, in electrical engineering and computer science, all from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor of Management Science and Engineering, Electrical Engineering, and, by courtesy, Computer Science, with Stanford University, Stanford, CA. He has held visiting positions as the Wolfgang and Helga Gaul Visiting Professor with the University of Karlsruhe and as the Chin Sophonpanich Foundation Professor of Banking and Finance at Chulalongkorn University.

Dr. Van Roy is a member of INFORMS. He has served on the editorial boards of *Discrete Event Dynamic Systems*, *Machine Learning*, *Mathematics of Operations Research*, and *Operations Research*. He has been a recipient of the MIT George C. Newton Undergraduate Laboratory Project Award (1993), the MIT Morris J. Levin Memorial Master's Thesis Award (1995), the MIT George M. Sprowls Doctoral Dissertation Award (1998), the NSF CAREER Award (2000), and the Stanford Tau Beta Pi Award for Excellence in Undergraduate Teaching (2003). He has been a Frederick E. Terman Fellow and a David Morgenthaler II Faculty Scholar.

Tsachy Weissman (S'99–M'02–SM'07) received the undergraduate and graduate degrees from the Department of Electrical Engineering at the Technion, Israel.

Following his graduation, he has held a faculty position with the Technion, and Postdoctoral appointments with the Statistics Department, Stanford University, Stanford, CA, and with Hewlett-Packard Laboratories. Since summer 2003, he has been on the faculty of the Department of Electrical Engineering, Stanford University. Since summer 2007, he has also been with the Department of Electrical Engineering, Technion, from which he is currently on leave. His research interests span information theory and its applications, and statistical signal processing. He is inventor or coinventor of several patents in these areas and involved in a number of high-tech companies as a researcher or member of the Technical Board.

Prof. Weissman has received the NSF CAREER Award, a Horev fellowship for leaders in Science and Technology, and the Henry Taub prize for excellence in research. He is a Robert N. Noyce Faculty Scholar of the School of Engineering at Stanford, and a recipient of the 2006 IEEE joint IT/COM societies Best Paper Award.