

# EXTENSIONS OF PAC LEARNING FOR PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

RAHUL JAIN AND PRAVIN VARAIYA  
 EECS DEPARTMENT  
 UNIVERSITY OF CALIFORNIA, BERKELEY  
 {rjain,varaiya}@eecs.berkeley.edu

**Abstract**— We extend Valiant’s probably approximately correct (PAC) model of learning to Markov decision processes (MDPs). The work is related to the simulation-based estimation of value functions for discounted reward MDPs. We obtain uniform sample complexity results for MDP simulation. We also obtain uniform sample complexity results for the case when the states are partially observable and policies are non-stationary and have memory. The results can be extended to Markov games with a finite number of players.

## I. INTRODUCTION

Markov decision processes (MDPs) are used as models for many problems and solved using dynamic programming. But dynamic programming for MDPs is computationally intractable and known to be PSPACE-complete [11]. Moreover, DP methods require an accurate analytical model and suffer from the “curse of dimensionality” in large or infinite state space. In such cases, simulation-based methods become important. They fit well into the learning model introduced by Valiant [12] and generalized by others ([1], [4], [5], [6]).

Consider a bounded real-valued function  $f \in \mathcal{F}$  over some space  $\mathcal{X}$  with a probability measure  $P$  on it. Given samples  $S = \{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$ , in which the  $x_k$  are drawn independently from  $P$  (obtained by simulation), the goal is to learn the function  $f$ .  $\mathcal{F}$  is PAC (probably approximately correct)-learnable if there is an algorithm that maps  $S$  to  $h_{n,f} \in \mathcal{F}$  such that for any  $\epsilon > 0$ , the probability that the empirical error

$$\text{err}(f, h_{n,f}) := \mathbb{E}_P[|f - h_{n,f}|]$$

is greater than  $\epsilon$  goes to zero as  $n \rightarrow \infty$  (Note that  $h_{n,f}$  is a function of  $S$ ). In other words, for  $n$  large enough the probability that the error is larger than  $\epsilon$  is smaller than some given  $\delta > 0$ .

The class of functions  $\mathcal{F}$  has the *uniform convergence of empirical means (UCEM) property* if

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_P[f(X)] \right| \rightarrow 0$$

in probability. It is known that a class of bounded real-valued functions with the uniform convergence of empirical means property is not only PAC-learnable but PUAC (probably uniformly approximately correct)-learnable [13], i.e.,

$$\lim_{n \rightarrow \infty} Pr\{\sup_{f \in \mathcal{F}} \text{err}(f, h_{n,f}) > \epsilon\} = 0.$$

Research supported by DARPA contract F 30602-01-2-0548. We thank Peter Bartlett, Ram Rajagopal, Slobodan Simic and Tunc Simsek for many helpful discussions.

This has the interesting interpretation that if a family of functions is such that the mean value of each function can be determined with small error and high probability then the function itself can be accurately determined with high probability. This is discussed in length in [14].

The PAC learning framework thus addresses the fundamental question of system identifiability. Moreover, it provides the properties that a system identification algorithm should have. Thus, in this paper, we develop PAC learning for MDPs and games. While the PAC learning model has been generalized to the case when the inputs are Markovian ([1], [5]), it has not been extended to MDPs and games.

Consider an MDP on state space  $\mathcal{X}$  and action space  $A$ , with probability transition function  $P_a$  and space  $\Pi$  of policies with a policy  $\pi(x, a)$  specifying the probability of taking action  $a$  given current state  $x$ . Let  $r(x)$  be a real-valued bounded reward function and  $V(\pi)$  the discounted total reward (also called the value of the policy). Let a policy  $\pi$  in  $\Pi$  be fixed but unknown. Our learning problem is to identify with small error and high probability which policy the system is using. We are given the “black box” system simulator, using which we can obtain samples of the value function of the unknown policy  $\pi$ . The simulator takes as input the initial state and a ‘noise’ sequence (each term is from uniform  $[0,1]$ ), which is used to simulate the state sequence. The output is the value of the policy or the total discounted reward under that state sequence. We obtain  $n$  such input-output samples from which the learning algorithm determines the policy  $\mu$  that may have generated the samples. We denote the empirical mean of such outputs by  $\hat{V}(\pi)$ .

We define the policy space  $\Pi$  of an MDP to be PAC-learnable if given an  $\epsilon > 0$  and any policy  $\pi \in \Pi$ , there exists a learning algorithm such that the policy  $\mu \in \Pi$ , identified by it as the one most likely to have generated the samples, is such that the probability that  $|V(\pi) - V(\mu)|$  is greater than  $\epsilon$  goes to zero as the number of samples  $n$  goes to infinity. As discussed above, the UCEM property is sufficient for PAC learnability. Therefore, in this paper we shall address that problem. Moreover, we shall obtain uniform sample complexity bounds for value function estimation, i.e., the number of samples  $n(\epsilon, \delta)$  needed such that the probability that

$$\sup_{\pi \in \Pi} |\hat{V}(\pi) - V(\pi)| > \epsilon$$

is bounded by some given  $\delta > 0$ .

The PAC learning problem for MDPs is more difficult than the PAC learning problem for function spaces because we have to consider the combinatorial complexity of a function

space obtained by multiple iterations of functions that belong to a particular function space. Relating the combinatorial complexity of the original function space with that of the policy space is difficult.

The key contributions of this paper are in extending the PAC learning framework to discounted reward Markov decision processes and in obtaining (non-asymptotic) uniform sample complexity bounds for value function estimation. The results are extended to the case when the state is partially observable.

Since this work also has implications for simulation-based optimization of MDPs, it is pertinent to mention [3], [9], [10] is quite close in spirit to this paper but the results are not in the context of PAC learning. Moreover, as we point out later, the assumptions in that paper are very strong and not realistic.

## II. PRELIMINARIES

Consider an MDP  $M$ , with countable state space  $\mathcal{X}$ , action space  $A$ , transition probability function  $P_a(x, x')$ , initial state distribution  $\lambda$ , reward function  $r(x)$  bounded in  $[0, R]$  and discount factor  $\gamma$ . The value function is the discounted reward for a policy  $\pi$

$$V(\pi) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(x_t)\right],$$

where  $x_t$  is the state in the  $t$ th step under policy  $\pi$ . For the average rewards case the value function is

$$V(\pi) = \mathbb{E}\left[\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T r(x_t)\right].$$

Let  $\Pi_0$  denote the space of all *stationary stochastic policies*  $\{\pi(x, a) : a \in A, x \in \mathcal{X}, \sum_a \pi(x, a) = 1\}$  and let  $\Pi \subseteq \Pi_0$  be the subset of policies of interest. The MDP  $M$  under a fixed stationary policy  $\pi$  induces a Markov chain with transition probability function  $P_\pi(x, x') = \sum_a P_a(x, x')\pi(x, a)$ . The initial distribution on the Markov chains is the same as for the MDP and we take  $P_\pi$  to characterize the Markov chain. Denote  $\mathcal{P} := \{P_\pi : \pi \in \Pi\}$ .

We first relate the combinatorial dimension<sup>1</sup> of  $\Pi$  to that of  $\mathcal{P}$ . Recall some measures of combinatorial complexity.

Let  $\mathcal{F}$  be a set of binary-valued functions from  $\mathcal{X}$  to  $\{0, 1\}$ . We say that  $\mathcal{F}$  *shatters*  $\{x_1, \dots, x_n\}$  if the set  $\{(f(x_1), \dots, f(x_n)), f \in \mathcal{F}\}$  has cardinality  $2^n$ . The largest such  $n$  is the *VC-dim*( $\mathcal{F}$ ).

Let  $\mathcal{F}$  be a set of real-valued functions from  $\mathcal{X}$  to  $[0, 1]$ . We say that  $\mathcal{F}$  *P-shatters*  $\{x_1, \dots, x_n\}$  if there exists a witness vector  $c = (c_1, \dots, c_n)$  such that the set  $\{(\eta(f(x_1) - c_1), \dots, \eta(f(x_n) - c_n)), f \in \mathcal{F}\}$  has cardinality  $2^n$ , where  $\eta(\cdot)$  is the sign function. The largest such  $n$  is the *Pseudo-dim*( $\mathcal{F}$ ).

Let  $\lambda$  and  $\mu$  be probability measures on  $\mathcal{X}$  and  $A$  respectively. Given a set  $\mathcal{F}$  of real-valued functions on  $\mathcal{X}$ ,  $\rho$  a metric on  $\mathbb{R}$ , let  $d_{\rho(\lambda)}$  be the pseudo-metric on  $\mathcal{F}$  with respect to measure  $\lambda$ :

$$d_{\rho(\lambda)}(f, g) = \int \rho(f(x), g(x))\lambda(dx)$$

<sup>1</sup>Examples of combinatorial dimensions are VC-dimension and Pseudo-dimension. They are different from the algebraic dimension. Typically we consider a class of real-valued functions, which have infinite algebraic dimension but may have finite P-dimension. See [14], [2] for more details.

A subset  $\mathcal{G} \subseteq \mathcal{F}$  is an  $\epsilon$ -net for  $\mathcal{F}$  if  $\forall f \in \mathcal{F}, \exists g \in \mathcal{G}$  with  $d_{\rho(\lambda)}(f, g) \leq \epsilon$ . The size of the minimal  $\epsilon$ -net is the  $\epsilon$ -covering number, denoted by  $\mathcal{N}(\epsilon, \mathcal{F}, d_{\rho(\lambda)})$ . The  $\epsilon$ -capacity of  $\mathcal{F}$  under the  $\rho$  metric,  $\mathcal{C}(\epsilon, \mathcal{F}, \rho) = \sup_\lambda \mathcal{N}(\epsilon, \mathcal{F}, d_{\rho(\lambda)})$ . The *upper metric dimension* of  $\mathcal{F}$  under the  $\rho$  metric,  $\overline{\dim}(\mathcal{F}) = \limsup_{\epsilon \rightarrow 0} \frac{\log \mathcal{C}(\epsilon, \mathcal{F}, \rho)}{\log(1/\epsilon)}$ . The lower metric dimension is defined similarly with  $\inf$  replacing  $\sup$ . When both exist and are equal, it is called the *metric dimension*. (See [8], [2] for more on properties of covering numbers.)

We define the following  $L_1$  pseudo-metric on  $\Pi$ :

$$d_{L_1(\lambda \times \mu)}(\pi, \pi') := \sum_{x \in \mathcal{X}} \lambda(x) \sum_{a \in A} \mu(a) |\pi(x, a) - \pi'(x, a)|,$$

and the following total variation-type pseudo-metric on  $\mathcal{P}$ :

$$d_{TV(\lambda)}(P, P') := \sum_{x \in \mathcal{X}} \lambda(x) \sum_{y \in \mathcal{X}} |P(x, y) - P'(x, y)|.$$

Given a policy space  $\Pi$  and  $\mathcal{P}$ , the set of transition probabilities of the Markov chains it induces, we relate the covering numbers on the two spaces under the pseudo-metrics defined above.

*Lemma 1:* Suppose  $\Pi \subseteq \Pi_0$  and  $\mathcal{P}$  as defined above with  $\text{P-dim}(\Pi) = d$ . Assume that there is a probability measure  $\lambda$  on  $\mathcal{X}$  and a probability measure  $\mu$  on  $A$  such that  $\pi(x, a)/\mu(a) \leq K, \forall x \in \mathcal{X}, a \in A, \pi \in \Pi$ . Then,

$$\mathcal{N}(\epsilon, \mathcal{P}, d_{TV(\lambda)}) \leq \mathcal{N}(\epsilon, \Pi, d_{L_1(\lambda \times \mu)}) \leq \left(\frac{2eK}{\epsilon} \log \frac{2eK}{\epsilon}\right)^d.$$

i.e.,  $\overline{\dim}(\mathcal{P}) \leq d$ .

The above lemma is a special case of lemma 4, for which a proof is provided. It can also be found in [7].

## III. THE SIMULATION MODEL

We estimate the value  $V(\pi)$  of policy  $\pi \in \Pi$  from independent samples of the discounted rewards. The samples are generated by a simulation ‘engine’  $h$ . This is a deterministic function to which we feed a noise sequence  $\omega = (\omega_1, \omega_2, \dots)$  (with  $\omega_i$  i.i.d. and uniform  $[0, 1]$ ) and different initial states  $x_0^1, \dots, x_0^n$  (with  $x_0^i$  i.i.d. and distribution  $\lambda$ ). The engine generates a state sequence with the same distribution as the Markov chain corresponding to  $\pi$ . The estimate of  $V(\pi)$  is the average of the total discounted reward starting in different initial states.

Because simulation cannot be performed indefinitely, we truncate the simulation at some time  $T$ , after which the contribution to the total discounted reward falls below  $\epsilon/2$  for required estimation error bound  $\epsilon$ .  $T$  is the  $\epsilon/2$ -horizon time. The function  $h : \mathcal{X} \times A \times \Omega \rightarrow \mathcal{X}$  can also be written as  $h : \mathcal{X} \times \mathcal{X} \times A \times \Omega \rightarrow \{0, 1\}$ , that is  $h(x, x', a, \omega_i) = 1$  if from state  $x$ , with action  $a$ , and noise  $\omega_i$ , system transitions to state  $x'$ , and 0 otherwise.

Many simulation functions are possible. We will work with the following simple simulation model, for  $\mathcal{X} = \mathbb{N}$ , which can be extended for multi-dimensional state space.

*Definition 1 (Simple simulation model):* The simple simulation model  $h$  for a given MDP is given by

$$h(x, a, \omega) = \inf\{y \in \mathcal{X} : \omega \in [F_{a,x}(y-1), F_{a,x}(y)]\},$$

in which  $F_{a,x}(y) := \sum_{y' \leq y} P_a(x, y')$  is the c.d.f. corresponding to the transition probability function  $P_a(x, y)$ .

Similarly, with a slight abuse of notation, we define the *simple* simulation model  $h$  for the Markov chain  $P$  as

$$h(x, P, \omega) = \inf\{y \in \mathcal{X} : \omega \in [F_{P,x}(y-1), F_{P,x}(y)]\}$$

where  $F_{P,x}(y) := \sum_{y' \leq y} P(x, y')$ , is the c.d.f. corresponding to the transition probability function  $P(x, y)$ .

This is the simplest method of simulation. For example, to simulate a probability distribution on a discrete state space, we partition the unit interval such that the first subinterval has length equal to the mass on the first state, the second subinterval has length equal to the mass on the second state, and so on.

The state sequence  $\{x_t\}$  for policy  $\pi$  is obtained by

$$x_{t+1} = f_{P_\pi}(x_t, \omega_{t+1}) = h(x_t, P_\pi, \omega_{t+1}),$$

in which  $P_\pi$  is the transition probability function of the Markov chain induced by  $\pi$  and  $\omega_{t+1} \in \Omega$  is noise. The initial state  $x_0$  is drawn according to the given initial state distribution  $\lambda$ . The function  $f_{P_\pi} : \mathcal{X} \times \Omega \rightarrow \mathcal{X}$  is called the *simulation function* for the Markov chain transition probability function  $P_\pi$ . As before,  $\mathcal{P} = \{P_\pi : \pi \in \Pi\}$ . We denote the set of all simulation functions induced by  $\mathcal{P}$  by  $\mathcal{F} = \{f_P : P \in \mathcal{P}\}$ .

To every  $P \in \mathcal{P}$ , there corresponds a function  $f \in \mathcal{F}$ . Observe that  $f \in \mathcal{F}$  simulates  $P \in \mathcal{P}$  given by

$$P(x, y) = \mu\{\omega : f(x, \omega) = y\},$$

in which  $\mu$  is the Lebesgue measure on  $[0, 1]$ . Unless specified otherwise,  $\mathcal{F}$  will denote the set of simulation functions for the class  $\mathcal{P}$  under the simple simulation model.

We now show that the complexity of the space  $\mathcal{F}$  is the same as that of  $\mathcal{P}$  if  $\mathcal{P}$  is convex.

*Lemma 2:* Suppose  $\mathcal{P}$  is convex (being generated by a convex space of policies) with pseudo-dimension  $d$ . Let  $\mathcal{F}$  be the corresponding space of simple simulation functions induced by  $\mathcal{P}$ . Then,  $\text{P-dim}(\mathcal{F}) = d$ . Moreover, the algebraic dimension of  $\mathcal{P}$  is also  $d$ .

*Proof:* First note that there is a one-to-one map between the space of simple simulation functions  $\mathcal{F}$  and the space of cdf's  $\tilde{\mathcal{F}}$  corresponding to  $\mathcal{P}$ . ( $\tilde{\mathcal{F}} = \{\tilde{F} : \tilde{F}(x, y) = \sum_{y' \leq y} P(x, y'), P \in \mathcal{P}\}$ .) Moreover  $\mathcal{F}$  and  $\tilde{\mathcal{F}}$  have the same pseudo-dimension. This is because, for any  $F \in \mathcal{F}$ ,  $F(x, \omega) > y$  if and only if for the corresponding  $\tilde{F} \in \tilde{\mathcal{F}}$ ,  $\tilde{F}(x, y) < \omega$ . Thus,  $\tilde{\mathcal{F}}$  shatters  $\{(x_1, y_1), \dots, (x_d, y_d)\}$  with witness vector  $(\omega_1, \dots, \omega_d)$  if and only if  $\mathcal{F}$  shatters  $\{(x_1, \omega_1), \dots, (x_d, \omega_d)\}$  with witness vector  $(y_1, \dots, y_d)$ . So, in the following discussion, we shall treat them as the same space  $\mathcal{F}$ .

Because  $\mathcal{P}$  has pseudo-dimension  $d$ , there exists  $S = \{(x_1, y_1), \dots, (x_d, y_d)\}$  that is shattered by  $\mathcal{P}$  with some witness vector  $c = (c_1, \dots, c_d)$ . Consider the projection of the set  $\mathcal{P}$  on the  $S$  coordinates:  $\mathcal{P}|_S = \{(P(x_1, y_1), \dots, P(x_d, y_d)) : P \in \mathcal{P}\}$ . The definition of shattering implies that there is a  $d$ -dimensional hypercube contained in  $\mathcal{P}|_S$  with center  $c$ . Also note that  $\mathcal{P}|_S$  is convex and its algebraic dimension is  $d$ . To argue that the algebraic dimension of  $\mathcal{P}$  cannot be  $d + 1$ ,

suppose that it is. Then, it would contain  $d + 1$  coordinates such that the projection of  $\mathcal{P}$  along those coordinates contains a hypercube of dimension  $d + 1$ . Thus,  $\mathcal{P}$  would shatter  $d + 1$  points with the center of the hypercube being a witness vector. But that contradicts the assumption that the pseudo-dimension is  $d$ .

Thus, for convex spaces, algebraic dimension and its pseudo-dimension are the same.

Next,  $\mathcal{F}$  is obtained from  $\mathcal{P}$  by an invertible linear transformation, hence its algebraic dimension is also  $d$ . Thus, it has  $d$  coordinates  $S$  such that the projected space  $\mathcal{F}|_S$ , has algebraic dimension  $d$ . Moreover, it contains a hypercube of dimension  $d$ . Hence, its pseudo-dimension is at least  $d$ . Since the argument is reversible starting from space  $\mathcal{F}$  to space  $\mathcal{P}$ , it implies  $\text{P-dim}(\mathcal{P}) = \text{P-dim}(\mathcal{F})$ . ■

It is an open question whether the simple simulation model yields the space  $\mathcal{F}$  of functions that simulates a *non-convex* set  $\mathcal{P}$  with the least complexity. The question is important, because as we will see below the sample complexity depends on the pseudo-dimension of the space of simulation functions. Thus, if we pick the simulation model with a smaller pseudo-dimension, we get a better sample complexity result.

#### IV. DISCOUNTED-REWARD MDPS

Consider an MDP  $M$  with state space  $\mathcal{X}$ , action space  $A$ , transition probability function  $P_a(x, x')$ , initial state distribution  $\lambda$ , reward function  $r(x)$ , and discount factor  $\gamma < 1$ . The value function is the total discounted reward for a policy  $\pi$  in some set of stationary policies  $\Pi$ . Denote  $\Omega = [0, 1]$ .

We redefine  $\mathcal{F}$  to be the set of measurable functions from  $Y := \mathcal{X} \times \Omega^\infty$  onto itself which simulates  $\mathcal{P}$ , the transition probabilities induced by  $\Pi$  under the simple simulation model. However, each function only depends on the first component of the sequence  $\omega = (\omega_1, \omega_2, \dots)$ . Thus the results and discussion of the previous section still hold. Let  $\theta$  be the left-shift operator on  $\Omega^\infty$ ,  $\theta(\omega_1, \omega_2, \dots) = (\omega_2, \omega_3, \dots)$ . Thus, for a policy  $\pi$ , our simulation system is defined as

$$(x_{t+1}, \theta\omega) = f_{P_\pi}(x_t, \omega),$$

in which  $x_{t+1}$  is the next state starting from  $x_t$  and the simulator also outputs the shifted noise sequence  $\theta\omega$ . This definition of the simulation function is introduced so that we can define the iteration of simulation functions. Let  $\mathcal{F}^2 := \{f \circ f : Y \rightarrow Y, f \in \mathcal{F}\}$  and  $\mathcal{F}^t$  its generalization to  $t$  iterations. Let  $\mu$  be a probability measure on  $\Omega^\infty$  and  $\lambda$  the initial distribution on  $\mathcal{X}$ . Denote the product measure on  $Y$  by  $\mathbb{P} = \lambda \times \mu$ , and on  $Y^n$  by  $\mathbb{P}^n$ . Define the two pseudo-metrics on  $\mathcal{F}$ :

$$\rho_{\mathbb{P}}(f, g) = \sum_x \lambda(x) \mu\{\omega : f(x, \omega) \neq g(x, \omega)\},$$

and

$$d_{L_1(\mathbb{P})}(f, g) := \sum_x \lambda(x) \int |f(x, \omega) - g(x, \omega)| d\mu(\omega).$$

We present a key technical result which relates the covering number of the iterated functions  $\mathcal{F}^t$  under the  $\rho$  pseudo-metric with the covering number for  $\mathcal{F}$  under the  $L_1$  pseudo-metric.

*Lemma 3:* Let  $\lambda$  be the initial distribution on  $\mathcal{X}$  and let  $\lambda_f$  the (one-step) distribution given by  $\lambda_f(y) = \sum_x \lambda(x) \mu\{\omega : f(x, \omega) = (y, \theta\omega)\}$  for  $f \in \mathcal{F}$ . Suppose that

$$\sup_{f \in \mathcal{F}, y \in \mathcal{X}} \frac{\lambda_f(y)}{\lambda(y)} \leq K, \quad \text{for } K \geq 2. \quad (1)$$

Then,

$$\mathcal{N}(\epsilon, \mathcal{F}^t, \rho_{\mathbb{P}}) \leq \mathcal{N}(\epsilon/K^t, \mathcal{F}, d_{L_1}(\mathbb{P})).$$

The proof is technical and can be found in the appendix to [7]. The condition of the lemma essentially means that under distribution  $\lambda$  the change in the probability mass on any state under any policy after one transition is bounded.

The estimation procedure is this. Obtain  $n$  initial states  $x_0^{(1)}, \dots, x_0^{(n)}$  drawn iid. according to  $\lambda$ , and  $n$  trajectories  $\omega^{(1)}, \dots, \omega^{(n)} \in \Omega^\infty$  ( $\Omega = [0, 1]$ ) drawn according to  $\mu$ , the product measure on  $\Omega^\infty$  of uniform probability measures on  $\Omega$ . Denote the samples by  $S = \{(x_0^1, \omega^1), \dots, (x_0^n, \omega^n)\}$  with measure  $\mathbb{P}^n$ .

This is our first main result.

*Theorem 1:* Let  $(\mathcal{X}, \Gamma, \lambda)$  be the measurable state space. Let  $A$  be the action space and  $r(x)$  the real-valued reward function, with values in  $[0, R]$ . Let  $\Pi \subseteq \Pi_0$ , the space of stationary stochastic policies,  $\mathcal{P}$  be the space of Markov chain transition probabilities induced by  $\Pi$ , and  $\mathcal{F}$  the space of simulation functions of  $\mathcal{P}$  under the simple simulation model  $h$ . Suppose that  $\text{P-dim}(\mathcal{F}) \leq d$  and the initial state distribution  $\lambda$  is such that  $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \frac{\lambda_f(x)}{\lambda(x)} \leq K$  for some  $K \geq 2$ . Let  $\hat{V}(\pi)$  be the estimate of  $V(\pi)$  obtained by averaging the reward from  $n$  samples. Then, given any  $\epsilon, \delta > 0$ , and with probability at least  $1 - \delta$ ,

$$\sup_{\pi \in \Pi} |\hat{V}(\pi) - V(\pi)| < \epsilon$$

for

$$n \geq \frac{64R^2}{\alpha^2} \left( \log \frac{4}{\delta} + 2d \left( \log \frac{32eR}{\alpha} + T \log K \right) \right). \quad (2)$$

Here  $T$  is the  $\epsilon/2$ -horizon time and  $\alpha = \epsilon/2(T+1)$ .

*Proof:* Recall that under the simple simulation model,  $f_P(x, \omega) := h(x, P, \omega)$  and  $\mathcal{F} = \{f_P : \mathcal{X} \times \Omega^\infty \rightarrow \mathcal{X} \times \Omega^\infty, P \in \mathcal{P}\}$ . Define the function  $R_t(x_0, \omega) := r \circ f_P \circ \dots \circ f_P(x_0, \omega)$ , with  $f_P$  composed  $t$  times. Let  $\mathcal{R}_t := \{R_t : \mathcal{X} \times \Omega^\infty \rightarrow [0, R], P \in \mathcal{P}\}$ . Let  $V(\pi)$  be the expected discounted reward, and  $V^T(\pi)$  the expected discounted reward truncated upto  $T$  steps. Let  $\hat{V}_n^T(\pi) = \frac{1}{n} \sum_{i=1}^n [\sum_{t=0}^T \gamma^t R_t^\pi(x_0^i, \omega^i)]$ . Then,

$$\begin{aligned} |V(\pi) - \hat{V}_n^T(\pi)| &\leq |V(\pi) - V^T(\pi)| + |V^T(\pi) - \hat{V}_n^T(\pi)|, \\ &\leq \epsilon/2 + |V^T(\pi) - \hat{V}_n^T(\pi)| \\ &\leq \epsilon/2 + \sum_{t=0}^T \left| \frac{1}{n} \sum_{i=1}^n [R_t^\pi(x_0^i, \omega^i) - \mathbb{E}(R_t^\pi)] \right|. \end{aligned}$$

The expectation is with respect to the product measure  $P_\pi^t \times \lambda \times \mu$ . We show that with high probability, each term in the sum over  $t$  is bounded by  $\alpha = \epsilon/2(T+1)$ .

Note that

$$\int |r(f^t(x, \omega)) - r(g^t(x, \omega))| d\mu(\omega) d\lambda(x)$$

$$\leq R \cdot \sum_x \lambda(x) \mu\{\omega : f^t(x, \omega) \neq g^t(x, \omega)\},$$

which as in lemma 3 implies that

$$d_{L_1}(\mathbb{P})(r \circ f^t, r \circ g^t) \leq R \cdot K^T d_{L_1}(\mathbb{P})(f, g).$$

From theorem 3 in [6], lemma 3, and above inequality, we get

$$\begin{aligned} &\mathbb{P}^n [\sup_{R_t \in \mathcal{R}_t} |\frac{1}{n} \sum_{i=1}^n R_t(x_0^i, \omega^i) - \mathbb{E}(R_t)| > \alpha] \\ &\leq 4\mathcal{C}(\alpha/16, \mathcal{R}_t, d_{L_1}) \exp(-n\alpha^2/64R^2) \\ &\leq 4 \left( \frac{32eRK^T}{\alpha} \log \frac{32eRK^T}{\alpha} \right)^d \exp(-n\alpha^2/64R^2). \end{aligned}$$

This implies that for the estimation error to be bounded by  $\alpha$  with probability at least  $\delta$ , the number of samples needed is

$$n \geq \frac{64R^2}{\alpha^2} \left( \log \frac{4}{\delta} + 2d \left( \log \frac{32eR}{\alpha} + T \log K \right) \right). \quad \blacksquare$$

*Remarks.* 1. Theorem 1 implies that  $\sup_{\pi \in \Pi} |\hat{V}(\pi) - V(\pi)|$  converges to zero in probability, hence the policy space  $\Pi$  is PAC-learnable. Like in [10], the theorem assumes that the pseudo-dimension of the  $\mathcal{F}$  space is finite. Combined with lemma 2 we get the following corollary.

*Corollary 1:* Under assumption 1, when  $\mathcal{P}$  is convex with  $\text{P-dim}(\mathcal{P}) = d$ , Theorem 2 holds.

2. Our sample complexity is of the same order in terms of  $\delta, \epsilon, T, R$  and  $d$ , as the results of [10] though the two results are not directly comparable due to the different assumptions made. Unlike in [10], we do not require the simulation functions to be Lipschitz continuous. For a discrete state space, this is not a realistic assumption as the following examples show.

(1) Consider the Markov chain on  $\mathbb{N}$  such that the only transitions are from state 1 to state 2 with probability 1/2, to state 4 with probability 1/4,  $\dots$ , to state  $2^k$  with probability  $1/2^k$ , etc. Let  $\omega_1^k = \sum_{i=1}^k 2^{-i}$ , and  $\epsilon^k = 2^{-k-1}$ . Define the  $L_1$  metric  $\rho$  on  $\mathcal{X} \times \Omega$ ,  $\rho((x_1, \omega_1), (x_2, \omega_2)) = |x_1 - x_2| + |\omega_1 - \omega_2|$ . Then,

$$\rho(f(1, \omega_1^k - \epsilon^k/2), f(1, \omega_1^k + \epsilon^k/2)) \geq 2^k.$$

Thus,  $f$  is not Lipschitz continuous in  $\Omega$ .

(2) Consider the following Markov chain: State space  $\mathcal{X} = \mathbb{N}$  again endowed with the  $L_1$  metric. Transitions are deterministic: Transition from an even state  $n$  is to state  $2n$ , and from an odd state  $n+1$  is to  $3n$ . Then,  $\rho(f(n+1, \omega), f(n, \omega)) = n$  and so is not Lipschitz continuous in  $\mathcal{X}$ .

3. The sample complexity depends on the pseudo-dimension of the simulation functions space. Thus, if we simulate using a model with the least complexity, we can improve our sample complexity bound. However, the complexity of the space  $\mathcal{P}$  induced by a given policy space  $\Pi$  on an MDP  $M$  seems to us to be a more fundamental object than the space of simulation functions.

## V. PARTIAL OBSERVABILITY UNDER NONSTATIONARY POLICIES WITH MEMORY

We now consider the case when only partial observations are available. Let  $M$  be a partially observable MDP on state space  $\mathcal{X}$ , with actions in  $A$ , observations in  $\mathcal{Y}$  and reward function  $r : \mathcal{X} \rightarrow [0, R]$ . Let  $\lambda$  be the initial state probability measure on  $\mathcal{X}$ ,  $P_a(x, x')$  the state transition probability function, with the observations governed by the conditional probability measure  $\nu(y|x)$  which gives the probability of observing  $y \in \mathcal{Y}$  when the state is  $x \in \mathcal{X}$ . Let  $h_t$  denote the history  $(y_0, a_1, y_1, \dots, a_t, y_t)$ , i.e., the observations and actions upto time  $t$ .

Let  $\mathcal{H}_t$  denote  $\{h_t = (y_0, a_1, y_1, \dots, a_t, y_t) : a_s \in A, y_s \in \mathcal{Y}, 0 \leq s \leq t\}$ . Let  $\Pi$  be the set of policies where a  $\pi \in \Pi$  is such that  $\pi = (\pi_1, \pi_2, \dots)$  and  $\pi_t : \mathcal{H}_t \times A \rightarrow [0, 1]$  is a probability measure on  $A$  conditioned on  $h_t \in \mathcal{H}_t$ . Let  $\Pi_t$  denote the set of all policies  $\pi_t$  at time  $t$  with  $\pi \in \Pi$ . This gives rise to a conditional state transition function  $P_t(x, x'; h_t)$ , the probability of transition from state  $x$  to  $x'$  given history  $h_t$  upto time  $t$ . Thus, under policy  $\pi$ ,

$$P_t(x, x'; h_t) = \sum_a P_a(x, x') \pi_t(h_t, a).$$

Let  $\mathcal{P}_t$  denote the set of all  $P_{\pi_t}$  induced by the policies  $\pi_t$  with  $\pi \in \Pi$ . Then, defining the usual TV metric on  $\mathcal{P}_t$  and the usual  $L_1$  metric on  $\Pi_t$ , we get.

*Lemma 4:* Suppose  $\Pi_t$  and  $\mathcal{P}_t$  are as defined above with  $\text{P-dim}(\Pi_t) \leq d$ . Assume  $\lambda$  the initial state probability measure on  $\mathcal{X}$ ,  $\rho$  the probability measure on  $\mathcal{H}_t$  conditioned on  $x \in \mathcal{X}$ , and  $\mu$  a probability measure on  $A$  such that  $\pi_t(h_t, a)/\mu(a) \leq K, \forall h_t \in \mathcal{H}_t, a \in A, \pi_t \in \Pi_t$ . Then,

$$\begin{aligned} \mathcal{N}(\epsilon, \mathcal{P}_t, d_{TV}(\lambda \times \rho)) &\leq \mathcal{N}(\epsilon, \Pi_t, d_{L_1}(\mu \times \lambda \times \rho)) \\ &\leq \left( \frac{2eK}{\epsilon} \log \frac{2eK}{\epsilon} \right)^d, \end{aligned}$$

i.e.,  $\overline{\dim}(\mathcal{P}_t) \leq d$ .

*Proof:* We first show that the function  $\pi_t \mapsto P_{\pi_t}$  is uniformly Lipschitz continuous with constant 1. Suppose  $P_t = P_{\pi_t}$ , and  $P'_t = P_{\pi'_t}$ . Then,  $d_{TV}(\lambda \times \rho)(P_t, P'_t) :=$

$$\begin{aligned} &\sum_x \lambda(x) \sum_{x' \in \mathcal{X}} \left| \sum_{a \in A, h_t \in \mathcal{H}_t} P_a(x, x') (\pi_t(h_t, a) - \pi'_t(h_t, a)) \rho(h_t|x) \right| \\ &\leq \sum_x \sum_{h_t} \lambda(x) \rho(h_t|x) \sum_a \sum_{x'} P_a(x, x') |\pi_t(h_t, a) - \pi'_t(h_t, a)| \\ &\leq \sum_x \sum_{h_t} \lambda(x) \rho(h_t|x) \sum_a \mu(a) \left| \frac{\pi_t(h_t, a)}{\mu(a)} - \frac{\pi'_t(h_t, a)}{\mu(a)} \right| \end{aligned}$$

which is  $=: d_{L_1}(\mu \times \lambda \times \rho) \left( \frac{\pi_t}{\mu}, \frac{\pi'_t}{\mu} \right)$ . where the second inequality above follows by changing the order of the sums over  $a$  and  $x'$ . This leads to the desired conclusion.  $\blacksquare$

Let  $\mathcal{F}_t$  be the set of simulation functions of  $\mathcal{P}_t$  under the simple simulation model. Thus, an  $f_t \in \mathcal{F}_t$  for  $t \geq 2$  is defined on  $f_t : \mathcal{X} \times \mathcal{H}_{t-1} \times \Omega^\infty \rightarrow \mathcal{X} \times \mathcal{H}_t \times \Omega^\infty$  while  $f_1 \in \mathcal{F}_1$

shall be defined on  $f_1 : \mathcal{X} \times \Omega^\infty \rightarrow \mathcal{X} \times \mathcal{H}_1 \times \Omega^\infty$ . It is straightforward to verify that lemma 2 also extends.

*Lemma 5:* Suppose  $\mathcal{P}_t$  is convex (being generated by a convex space of policies  $\Pi$ ). Let  $\text{P-dim}(\mathcal{P}_t) = d$ . Let  $\mathcal{F}_t$  be the corresponding space of simple simulation functions induced by  $\mathcal{P}_t$ . Then,  $\text{P-dim}(\mathcal{F}_t) = d$ .

By  $\mathcal{F}^t$  we shall denote the set of functions  $f^t = f_t \circ \dots \circ f_1$  where  $f_s \in \mathcal{F}_s$  and they arise from a common policy  $\pi$ . Note that  $f^t : \mathcal{X} \times \Omega^\infty \rightarrow \mathcal{Z}_t \times \Omega^\infty$  where  $\mathcal{Z}_t = \mathcal{X} \times \mathcal{H}_t$ . We shall consider the following pseudo-metric on  $\mathcal{F}_t$  with respect to a measure  $\lambda_t$  on  $\mathcal{Z}_{t-1}$  for  $t \geq 2$  and measure  $\mu$  on  $\Omega^\infty$ ,

$$\rho_t(f_t, g_t) := \sum_{z \in \mathcal{Z}_{t-1}} \lambda_t(z) \mu\{\omega : f_t(z, \omega) \neq g_t(z, \omega)\}.$$

We shall take  $\rho_1$  as the pseudo-metric on  $\mathcal{F}^t$  w.r.t the product measure  $\lambda \times \mu$ . Let

$$\lambda_{f^t}(z) := \sum_{x \in \mathcal{Z}_t} \lambda(x) \mu\{\omega : f^t(x, \omega) = z\}$$

be a probability measure on  $\mathcal{Z}_t$ . We now state the extension of the technical lemma needed for the main theorem of this section.

*Lemma 6:* Let  $\lambda$  be a probability measure on  $\mathcal{X}$  and  $\lambda_{f^t}$  be the probability measure on  $\mathcal{Z}_t$  as defined above. Suppose that  $\text{P-dim}(\mathcal{F}_t) \leq d$  and there exists probability measures  $\lambda_t$  on  $\mathcal{Z}_t$  such that  $\sup_{f^t \in \mathcal{F}^t, z \in \mathcal{Z}_t} \frac{\lambda_{f^t}(z)}{\lambda_{t+1}(z)} \leq K$  for some  $K > 0$ . Then, for  $1 \leq t \leq T$ ,

$$\begin{aligned} \mathcal{N}(\epsilon, \mathcal{F}^t, \rho_1) &\leq \mathcal{N}\left(\frac{\epsilon}{Kt}, \mathcal{F}_t, \rho_t\right) \dots \mathcal{N}\left(\frac{\epsilon}{Kt}, \mathcal{F}_1, \rho_1\right) \\ &\leq \left( \frac{2eKt}{\epsilon} \log \frac{2eKt}{\epsilon} \right)^{dt}. \end{aligned}$$

The proof can be found in the appendix. We now obtain our sample complexity result.

*Theorem 2:* Let  $(\mathcal{X}, \Gamma, \lambda)$  be the measurable state space,  $A$  the action space,  $\mathcal{Y}$  the observation space,  $P_a(x, x')$  the state transition function and  $\nu(y|x)$  the conditional probability measure that determines the observations. Let  $r(x)$  the real-valued reward function bounded in  $[0, R]$ . Let  $\Pi$  be the set of stochastic policies (non-stationary and with memory in general),  $\mathcal{P}_t$  be the set of state transition functions induced by  $\Pi_t$ , and  $\mathcal{F}_t$  the set of simulation functions of  $\mathcal{P}_t$  under the simple simulation model. Suppose that  $\text{P-dim}(\mathcal{P}_t) \leq d$ . Let  $\lambda$  and  $\mu$  be probability measures on  $\mathcal{X}$  and  $A$  respectively and  $\lambda_{t+1}$  a probability measure on  $\mathcal{Z}_t$  such that

$$\sup_{f^t \in \mathcal{F}^t, z \in \mathcal{Z}_t} \frac{\lambda_{f^t}(z)}{\lambda_{t+1}(z)} \leq K$$

for some  $K > 0$ , where  $\lambda_{f^{t-1}}$  is as defined above. Let  $\hat{V}(\pi)$ , the estimate of  $V(\pi)$  obtained from  $n$  samples. Then, given any  $\epsilon, \delta > 0$ , and with probability at least  $1 - \delta$ ,

$$\sup_{\pi \in \Pi} |\hat{V}(\pi) - V(\pi)| < \epsilon$$

for  $n \geq \frac{64R^2}{\alpha^2} \left( \log \frac{4}{\delta} + 2dT \left( \log \frac{32eR}{\alpha} + \log KT \right) \right)$  where  $T$  is the  $\epsilon/2$  horizon time and  $\alpha = \epsilon/2(T+1)$ .

## VI. CONCLUSIONS

In this work, we have presented an extension of Valiant's probably approximately correct learning model to Markov decision processes. We obtained sample complexity results for the discounted reward case which are linear in  $T$ , the horizon time. Moreover, we do not assume any regularity conditions like Lipschitz continuity. We showed that there is a canonical simulation model under which the combinatorial complexity of the simulation functions is the same as that of the underlying space of induced Markov chains when the space is convex. The results are extended to when the states of the MDP are only partially observable with policies that are non-stationary with memory.

The results can be extended to dynamic Markov games and similar uniform sample results can be obtained. They are presented in [7].

Obtaining value function estimates with uniform sample complexity bounds, of course, is only a step towards finding the optimal policy in the given space. How such estimates will be used to find the optimal policy is an open problem for future work.

## APPENDIX

The proof of lemma 6 is along same lines as lemma 3 but the details are somewhat involved and presented below.

*Proof:* Consider any  $f^t, g^t \in \mathcal{F}^t$  and  $x \in \mathcal{X}$ . Then,

$$\begin{aligned}
& \mu\{\omega : f^t(x, \omega) \neq g^t(x, \omega)\} \\
&= \mu\{\omega : f^t(x, \omega) \neq g^t(x, \omega), f^{t-1}(x, \omega) = g^{t-1}(x, \omega)\} \\
&\quad + \mu\{\omega : f^t(x, \omega) \neq g^t(x, \omega), f^{t-1}(x, \omega) \neq g^{t-1}(x, \omega)\} \\
&= \mu\{\cup_{z \in \mathcal{Z}_{t-1}} (\omega : f^t(x, \omega) \neq g^t(x, \omega), \\
&\quad f^{t-1}(x, \omega) = g^{t-1}(x, \omega) = z)\} \\
&\quad + \mu\{\cup_{z \in \mathcal{Z}_{t-1}} (\omega : f^t(x, \omega) \neq g^t(x, \omega), \\
&\quad f^{t-1}(x, \omega) \neq g^{t-1}(x, \omega), f^{t-1}(x, \omega) = z)\} \\
&\leq \mu\{\cup_{z \in \mathcal{Z}_{t-1}} (\omega : f^t(x, \omega) \neq g^t(x, \omega), \\
&\quad f^{t-1}(x, \omega) = g^{t-1}(x, \omega) = z)\} \\
&\quad + \mu\{\cup_{z \in \mathcal{Z}_{t-1}} (\omega : f^{t-1}(x, \omega) \neq g^{t-1}(x, \omega), \\
&\quad f^{t-1}(x, \omega) = z)\} \\
&\leq \sum_{z \in \mathcal{Z}_{t-1}} \mu\{\omega : f_t(z, \omega) \neq g_t(z, \omega) | f^{t-1}(x, \omega) = z\} \\
&\quad \mu\{\omega : f^{t-1}(x, \omega) = z\} \\
&\quad + \mu\{\omega : f^{t-1}(x, \omega) \neq g^{t-1}(x, \omega)\}.
\end{aligned}$$

As before, the  $\omega$  part of both the functions is the same and hence is ignored. So,

$$\begin{aligned}
& \sum_{z \in \mathcal{Z}_{t-1}} \lambda_{f^{t-1}}(z) \mu\{\omega : f_t(z, \omega) \neq g_t(z, \omega)\} \\
&\leq K \cdot \sum_{z \in \mathcal{Z}_{t-1}} \lambda_t(z) \mu\{\omega : f_t(z, \omega) \neq g_t(z, \omega)\}.
\end{aligned}$$

This by induction implies

$$\rho_1(f^t, g^t) \leq K(\rho_t(f_t, g_t) + \dots + \rho_1(f_1, g_1))$$

This implies the first inequality. To argue the second, note that

$$\begin{aligned}
& \sum_z \lambda_t(z) \mu\{\omega : f_t(z, \omega) \neq g_t(z, \omega)\} \\
&\leq \sum_z \lambda_t(z) \int |f_t(z, \omega) - g_t(z, \omega)| d\mu(\omega).
\end{aligned}$$

■

## REFERENCES

- [1] D. ALDOUS AND U. VAZIRANI, "A Markovian extension of Valiant's learning model", *Information and computation*, 117(2), pp.181-186, 1995.
- [2] M. ANTHONY AND P. BARTLETT, *Neural network learning: theoretical foundations*, Cambridge university press, 1999.
- [3] J. BAXTAR AND P. L. BARTLETT, "Infinite-horizon policy-gradient estimation", *J. of A.I. Research*, 15, pp.319-350, 2001.
- [4] K. L. BUESCHER AND P. R. KUMAR, "Learning by canonical smooth estimation-Part I: Simultaneous estimation", *IEEE Trans. Automatic Control*, 41(4), pp.545-556, 1996.
- [5] D. GAMARNIK, "Extensions of the PAC framework to finite and Countable Markov Chains", *IEEE Trans. Information Theory*, 49(1), pp.338-345, 2003.
- [6] D. HAUSSLER, "Decision theoretic generalizations of the PAC model for neural nets and other learning applications", *Information and computation*, 100(1), pp.78-150, 1992.
- [7] R. JAIN AND P. VARAIYA, "PAC learning for Markov decision processes and dynamic games", submitted to ISIT, december 2003.
- [8] A. N. KOLMOGOROV AND V. M. TIHOMIROV, " $\epsilon$ -entropy and  $\epsilon$ -capacity of sets in functional spaces", *American Math. Soc. Translation Series 2*, 17, pp.277-364.
- [9] P. MARBACH AND J. N. TSITSIKLIS, "Approximate gradient methods in policy-space optimization of Markov reward processes", *J. Discrete Event Dynamical Systems*, Vol. 13, pp.111-148, 2003.
- [10] A. Y. NG AND M. I. JORDAN, "Pegasus: A policy search method for large MDPs and POMDPs", *Proc. UAI..*, 2000.
- [11] C. H. PAPADIMITRIOU AND J. N. TSITSIKLIS, "The complexity of Markov decision processes", *Mathematics of Operations Research*, 12(3), pp.441-450, 1987.
- [12] L. VALIANT, "A theory of the learnable", *Communications of the ACM*, 27(4), pp.1134-1142, 1984.
- [13] V. VAPNIK AND A. CHERVONENKIS, "The necessary and sufficient conditions for consistency in the empirical risk minimization method", *Pattern recognition and image analysis*, 1(3), pp.283-305, 1991
- [14] M. VIDYASAGAR, *Learning and generalization: With applications to neural networks*, second edition, Springer-Verlag, 2003.