Abstract—We generalize the PAC Learning framework for Markov Decision Processes developed in [18]. We consider the reward function to depend on both the state and the action. Both the state and action spaces can potentially be countably infinite. We obtain an estimate for the value function of a Markov decision process, which assigns to each policy its expected discounted reward. This expected reward can be estimated as the empirical average of the reward over many independent simulation runs. We derive bounds on the number of runs needed for the convergence of the empirical average to the expected reward uniformly for a class of policies, in terms of the V-C or pseudo dimension of the policy class. We then propose a framework to obtain an $\epsilon$-optimal policy from simulation. We provide sample complexity of such an approach.

I. INTRODUCTION

It is well-known that solving Markov decision processes using dynamic programming is computationally intractable. Thus various approximate dynamic programming techniques as well as simulation-based techniques have been developed. We propose an empirical process theory approach to simulation-based optimization of Markov decision processes.

The empirical process theory (EPT) [12] studies the uniform behavior of a class $G$ of measurable functions in the law of large numbers (as well as the central limit theorem [3]) regime. In particular, EPT studies the conditions for uniform convergence in terms of the V-C or pseudo dimension of the policy class. We then propose a framework to obtain an $\epsilon$-optimal policy from simulation. We provide sample complexity of such an approach.

We estimate $V$ from independent samples of the discounted reward by the empirical mean, $\hat{V}(\pi)$. We obtain the number of samples $n(\epsilon, \delta)$ (or sample complexity) needed so that the probability

$$\Pr\{\sup_{\pi \in \Pi} |\hat{V}(\pi) - V(\pi)| > \epsilon \} < \delta.$$  \hspace{1cm} (2)

Our approach is broadly inspired by [4], [16], [17] and influenced by [6]. Thus, we would like to reduce the problem in equation (2) to understanding the geometry of $\Pi$ in terms of its covering number. (If the covering number is finite, it is the minimal number of elements of a set needed to approximate any element in the set $\Pi$ with a given accuracy.) We first relate the covering numbers of the space of stationary stochastic policies and the space of Markov chains that they induce. We relate these to the space of simulation functions that simulate the policies when the policy space is convex. These results together yield the rate of convergence of the empirical estimate to the expected value for the discounted-reward MDPs. What makes the problem non-trivial is that obtaining empirical discounted reward from simulation involves an iteration of simulation functions. The geometry of the space of iterated simulation functions is much more complex than that of the original space.

The problem of uniform convergence of the empirical average to the value function for discounted MDPs was studied in [5], [10] in a machine learning context. While [5] only considered finite state and action spaces, [10] obtains the conditions for uniform convergence in terms of the simulation model rather than the geometric characteristics (such as covering numbers or the P-dimension) of the simulation function space as opposed to that of the more natural policy space. Large-deviations results for finite state and action spaces for the empirical state-action frequencies and general reward functions were obtained in [1], [8]. A different approach more akin to importance sampling is explored in [11].

While the problem of uniform estimation of the value function for discounted and average-reward partially observed MDPs is of interest in itself, we also present a framework for simulation-based optimal policy search. The simulation-based estimates such as those proposed in this paper, have been used in a gradient-based method for finding Nash equilibrium policies in a pursuit-evasion game problem [14] though the theoretical understanding is far from complete. Other simulation-based methods to find approximations to the optimal policy include [9], [2], [19].
II. PRELIMINARIES

Consider an MDP $M$, with state space $X$ and action space $A$, transition probability function $P_a(x,x')$, initial state distribution $\lambda$, reward function $r(x,a)$ (which depends on both state and action) with values in $[0,R]$, and discount factor $0 < \gamma < 1$. The value function for a policy $\pi$ is the expected discounted reward

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

in which $(x_t,a_t)$ is the state-action pair in the $t$th step under policy $\pi$. Let $\Pi_\lambda$ denote the space of all stationary stochastic policies $\{\pi(x,a) : a \in A, x \in X, \sum_a \pi(x,a) = 1\}$ and let $\Pi \subseteq \Pi_\lambda$ 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 $\lambda$, and we identify $P_\pi$ with the Markov chain. Denote $\mathcal{P} := \{P_\pi : \pi \in \Pi\}$.

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

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

A subset $G \subseteq \mathcal{F}$ is an $\epsilon$-net for $\mathcal{F}$ if $\forall f \in \mathcal{F}$, $\exists g \in G$ with $d_{p(\lambda)}(f,g) < \epsilon$. The size of the minimal $\epsilon$-net is the $\epsilon$-covering number, denoted $N(\epsilon, F, d_{p(\lambda)})$. The $\epsilon$-capacity of $\mathcal{F}$ under the $\rho$ metric is $C(\epsilon, F, \rho) = \sup_{\lambda} N(\epsilon, F, d_{p(\lambda)})$. Essentially, the $\epsilon$-net can be seen as a subset of functions that can $\epsilon$-approximate any function in $\mathcal{F}$. The covering number is a measure of the richness of the function class. The richer it is, the more approximating functions we will need for a given measure of approximation $\epsilon$. The capacity makes it independent of the underlying measure $\lambda$ on $X$. (See [6] for an elegant treatment of covering numbers.)

Let $\mathcal{F}$ be a set of real-valued functions from $X$ to $[0,1]$. Say that $\mathcal{F}$ $\rho$-shatters $\{x_1, \ldots, x_n\}$ if there exists a witness vector $c = (c_1, \ldots, c_n)$ such that the set $\{\eta(f(x_1) - c_1), \ldots, \eta(f(x_n) - c_n) \mid f \in \mathcal{F}\}$ has cardinality $2^n$; $\eta(\cdot)$ is the sign function. The largest such $n$ is the $P$-$\text{dim}(\mathcal{F})$. This is a generalization of VC-dim and for $[0,1]$-valued functions, the two definitions are equivalent.

Other combinatorial dimensions such as the fat-shattering dimension introduced in [7] yield both an upper and lower bound on the covering numbers but in this paper we will use the $P$-dim. Results using fat-shattering dimension can be established similarly.

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’ $(q_\pi, h)$. This is a deterministic function to which we feed two noise sequences $\nu = (\nu_1, \nu_2, \cdots)$ and $\omega = (\omega_1, \omega_2, \cdots)$ (with $\nu_i$ and $\omega_i$ i.i.d. and uniform $\Omega = [0,1]$) and different initial states and actions $(x_0^1, a_0^1), \cdots, (x_0^n, a_0^n)$ (with $(x_0^i, a_0^i)$ i.i.d. with distribution $\lambda$). (Note that we have put an initial distribution on initial actions but these play no role. The actions at $t=1$ are determined by the policies which are functions of $x_0$ and action at $t=1$, $a_1$.) The engine then generates a state and action sequence, with the state-sequence same as the Markov chain corresponding to $\pi$, $P_\pi(x,y) = \sum_a P_a(x,y)\pi(x,a)$. The estimate of $V(\pi)$ is the average of the total discounted reward starting with 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 $c/2$-horizon time. The function $h : X \times A \times \Omega \to X$ gives the next state $x'$ given current state is $x$, action taken is $a$, and noise $\omega_i$.

Many simulation functions are possible. We will work with the following simple simulation model, for $X = \mathbb{N}$ and $A = \{1 : N\}$, for $N$ finite.

Definition 1 (Simple simulation model): The simple simulation model $\{q_\pi, h\}$ for a given MDP with policy $\pi \in \Pi$ is given by

$$q_\pi(x, \nu) = \inf\{b \in \mathbb{N} : \nu \in [Q_{\pi,x} (b - 1), Q_{\pi,x} (b)]\},$$

where $Q_{\pi,x} (b) := \sum_{b' \leq b} \pi(x, b')$ and

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

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 $P_a(x, y)$.

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. It is a surprising fact that there are other simulation functions $h'$ that generate the same distribution, but which have a much larger complexity than $h$.

The state and action sequence $\{(x_t, a_t)\}_{t=0}^{\infty}$ for policy $\pi$ is obtained by

$$a_{t+1} = q_\pi(x_t, \nu_{t+1}),$$

$$x_{t+1} = f_x(x_t, \nu_{t+1}, \omega_{t+1}),$$

$$h(x_t, q_\pi(x_t, \nu_{t+1}), \omega_{t+1}),$$

where $\nu_t$ and $\omega_t \in \Omega$ are noises. The initial state-action pair $(x_0, a_0)$ is drawn according to the given initial state-action distribution $\lambda$ (with marginal distribution $\lambda$ over $X$).

Denote $z_t = (x_t, a_t) \in \mathbb{Z} := X \times A$ and $\xi_t = (\nu_t, \omega_t) \in \Omega_2 := \Omega \times \Omega$, then

$$s_\pi(z_t, \xi_t) = z_{t+1} = (q_\pi(x_t, \nu_{t+1}), f_x(x_t, \nu_{t+1}, \omega_{t+1})).$$

The function $s_\pi : \mathbb{Z} \times \Omega_2 \to \mathbb{Z}$ is called the simulation system for the policy $\pi$. We denote $\mathcal{Q} = \{q_\pi : \pi \in \Pi\}$,
$F = \{ f_{\pi} : \pi \in \Pi \}$ and the set of all simulation systems induced by $\Pi$ by $S = \{ s_{\pi} : \pi \in \Pi \} = Q \times F$.

Let $\mu$ denote the Lebesgue measure on $\Omega_2$. Then,

$$Pr\{ z_{t+1} = z' | z_t = z, \pi \} = \mu\{ \xi : s_{\pi}(z, \xi) = z' \}. $$

Unless specified otherwise, $S$ will denote the set of simulation functions for the policy space $\Pi$ under the simple simulation model.

The question now is how does the complexity of the space $Q$ compare with that of the policy space $\Pi$. We connect the two when $\Pi$ is convex.

**Lemma 1:** Suppose $\Pi$ is convex with P-dimension $^1 d$.

Let $Q$ be the corresponding space of simple simulation functions that simulate $\Pi$. Then, $P\text{-dim}(Q) = d$. Moreover, the algebraic dimension of $Q$ is also $d$.

The proof essentially follows the argument in the proof of Lemma 4.2 in [18] and will not be repeated here.

**IV. SAMPLE COMPLEXITY FOR DISCOUNTEDREWREWARD MDPs**

Consider an MDP $M$ with countably infinite state space $X = \mathbb{N}$, and finite action space $A = \{ 1 : N_{\lambda}\}$ (for some $N_{\lambda} \in \mathbb{N}$), transition probability function $P_{\alpha}(x, x')$, initial state distribution $\lambda$, reward function $r(x, a)$, and discount factor $\gamma < 1$. The value function is the total discounted reward for a policy $\pi$ in some set of stationary policies $\Pi$.

Let $\theta$ be the left-shift operator on $\Omega^\infty$, $\theta(\omega_1, \omega_2, \cdots) = (\omega_2, \omega_3, \cdots)$. We redefine $Q$ to be the set of measurable functions from $W := X \times A \times \Omega^\infty \times \Omega^\infty$ onto itself which simulates $\Pi$, the set of given stationary policies under the simple simulation model to generate the action sequence. The action-simulation function $q_{\pi}(x, a, \nu, \omega)$ may only depend on current state $x$ (though we define a more general function for notational simplicity), and $\nu_1$, the first component of the sequence $\nu = (\nu_1, \nu_2, \cdots)$ (with $\nu_i$ i.i.d. and uniform $[0,1]$).

Similarly, we redefine $F$ to be the set of measurable functions from $W := X \times A \times \Omega^\infty \times \Omega^\infty$ onto itself which simulates $\Pi$, the set of given stationary policies under the simple simulation model to generate the state sequence. Thus, state-simulation functions $f_{\pi}(x, a, \nu, \omega)$ may depend only on the current state $x$ (but again we define a more general function), and $\xi_1$, the first component of the sequence $\xi = (\xi_1, \xi_2, \cdots)$ (with $\xi_i = (\nu_i, \omega_i)$ each i.i.d. and uniform $[0,1]$).

As before, $S = Q \times F$, the set of simulation systems that simulate the policies $\Pi$, and generate the state-action sequence. Thus, the results and discussion of the previous section hold.

For a policy $\pi$, our simulation system

$$(z_{t+1}, \theta \xi) = s_{\pi}(z_t, \xi),$$

in which $z_{t+1}$ is the next state-action pair starting from the current state-action pair $z_t$ and the simulator also outputs the shifted noise sequences $\theta \xi$. This definition of the simulation function is introduced to facilitate the iteration of simulation functions. Let $S^2 := \{ s \circ s : W \times W \rightarrow W \times W, s \in S \}$ and $S^t$ its generalization to $t$ iterations.

Let $\mu$ be a probability measure on $\Omega_2^\infty$ and $\lambda$ the initial distribution on $\Omega$. Denote the product measure on $W$ by $P = \lambda \times \mu$, and on $W^n$ by $P^n$. Define the two pseudo-metrics on $S$:

$$\rho_P(s_1, s_2) = \sum_{z} \lambda(z) \mu\{ \xi : s_1(z, \xi) \neq s_2(z, \xi) \},$$

and

$$d_{L_1(P)}(s_1, s_2) := \sum_{z} \lambda(z) \int |s_1(z, \xi) - s_2(z, \xi)| d\mu(\xi).$$

The $\rho_P$ and $d_{L_1(P)}$ pseudo-metrics for the functions in $Q$ and $F$ are defined similarly.

We present a key technical result which relates the covering number of the iterated functions $S$ under the $P$ pseudo-metric with the covering number for $Q$ under the $L_1$ pseudo-metric.

**Lemma 2:** Let $\lambda$ be the initial distribution on $\Omega$ and let $\lambda_s$ the (one-step) distribution given by $\lambda_s(y) = \sum_z \lambda(z) \mu\{ \xi : s(z, \xi) = (y, \theta \xi) \}$ for $s \in S$. Suppose that

$$\mathcal{K} := \{ \sup_{s \in S, y \in \Omega} \lambda_s(y) / \lambda(y) \} < \infty. \quad (3)$$

Then,

$$\mathcal{N}(\epsilon, S^t, \rho_P) \leq \mathcal{N}(\epsilon/\mathcal{K}, Q, d_{L_1(P)}).$$

The proof of Lemma 2 is similar to that of Lemma 4 and is omitted. The condition of the lemma essentially means that under distribution $\lambda$ the change in the probability mass on any state-action pair under any policy after one transition is uniformly bounded.

The estimation procedure is this. Obtain $n$ initial state-action pairs $z_0^{(1)}, \cdots, z_0^{(n)}$ drawn i.i.d. according to $\lambda$, and $n$ trajectories $\xi^{(1)}, \cdots, \xi^{(n)} \in \Omega^\infty_2$ ($\Omega = [0, 1]$) drawn according to $\mu$. The product measure on $\Omega^\infty_2$ of uniform probability measures on $\Omega_2$. Denote the samples by $Z^n = \{ z_0^{(1)}, \xi^{(1)}, \cdots, z_0^{(n)}, \xi^{(n)} \}$ drawn with measure $P^n$.

This is our first main result.

**Theorem 1:** Let $(Z, \Gamma, \lambda)$ be the measurable state-action space and $r(x, a)$ the real-valued reward function, with values in $[0, R]$. Let $\Pi \subseteq \Pi_0$, the space of stationary stochastic policies, and $S$ the space of simple simulation systems of $\Pi$. Suppose that $P\text{-dim}(Q) \leq d$ and the initial state-action distribution $\lambda$ is such that $\mathcal{K} := \max\{ \sup_{s \in S, y \in Z} \lambda_s(y) / \lambda(y) \} < \infty$. Let $V_n(T)(\pi)$ be the estimate of $V(\pi)$ obtained by averaging the reward from $n$ samples, each $T$ steps long. Then, for $\epsilon, \delta > 0$,

$$\mathbb{P}^n\{ \sup_{\pi \in \Pi} \left| V_n(T)(\pi) - V(\pi) \right| > \epsilon \} < \delta$$

if

$$n \geq n_0(\epsilon, \delta) := \frac{32 R^2}{\alpha^2} \left( \log \frac{4}{\delta} + 2d(\log \frac{32 e R}{\alpha} + T \log K) \right). \quad (4)$$

---

1 We refer the reader to [18] for a preliminary discussion about P-dimension, $\epsilon$-covering numbers and $\epsilon$-capacity.
Here $T$ is the $\epsilon/2$-horizon time and $\alpha = \epsilon/2(T + 1)$.

Proof: Fix a policy $\pi$. Let $s_\pi$ be the corresponding simple simulation system that yields the state-action sequence given a noise sequence $\xi$. Define the function $R_t^\pi(z_0, \xi) := r \circ s_\pi \circ \cdots \circ s_\pi(z_0, \xi)$, with $s_\pi$ composed $t$ times. Let $R_t := \{R_t^\pi : \mathbb{Z} \times \OmegaZ \to [0, R], \pi \in \Pi\}$. Let $V(\pi)$ be the expected discounted reward, and $V^T(\pi)$ the expected discounted reward truncated up to $T$ steps. Let

$$V_n^T(\pi) = \frac{1}{n} \sum_{t=0}^{n} \left[ \sum_{\pi=0}^{T} R_t^\pi(x_0, \omega_0) \right],$$

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

Note that (after abusing the definition of the reward function a bit) we have

$$\int |r(s_1^\pi(z, \xi)) - r(s_2^\pi(z, \xi))| d\mu(\xi) d\lambda(z) \leq R \cdot \sum_{\xi} \lambda(\xi) \mu(\xi) = \sum_{\xi} \mu(\xi) \lambda(\xi) - \mathbb{E}(R_t^\pi),$$

which as in lemma 2 implies that

$$d_{L_1(\mathbb{P})}(r \circ s_1^\pi, r \circ s_2^\pi) \leq R \cdot K T d_{L_1(\mathbb{P})}(q_1, q_2).$$

From theorem 5.7 in [17] (also theorem 3 in [4]), lemma 2, and the inequality above, we get

$$\mathbb{P}^n \left[ \sup_{R_t \in \mathcal{R}_t} \left| \frac{1}{n} \sum_{i=1}^{n} R_t^\pi(z_0, \xi_i) - \mathbb{E}(R_t^\pi) \right| > \alpha \right] \leq 4c(\alpha/16, R_t, d_{L_1(\mathbb{P})}) \exp(-n\alpha^2/32R^2) \leq 4 \left( \frac{32\epsilon R K_T^2}{\alpha^2} \right) \frac{\log(32\epsilon R K_T^2)}{\alpha} \exp(-n\alpha^2/32R^2).$$

This implies the estimation error is bounded by $\alpha$ with probability at least $\delta$, if the number of samples is

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

V. PARTIALLY OBSERVABLE MDPs WITH GENERAL POLICIES

We now consider partially observed discounted-reward MDPs with general policies (non-stationary with memory). The setup is as before, except that the policy depends on observations $y \in \mathcal{Y}$, governed by the (conditional) probability $\tau(y|x)$ of observing $y \in \mathcal{Y}$ when the state is $x \in \mathcal{X}$. Let $h_t$ denote the history $(a_0, y_0, a_1, y_1, \cdots, a_t, y_t)$ of observations and actions before time $t$.

The results of section IV extend when the policies are non-stationary however there are many subtleties regarding the domain and range of simulation functions, and measures, and some details are different. Let $H_t := \{h_t = (a_0, y_0, a_1, y_1, \cdots, a_t, y_t) : a_s \in A, y_s \in \mathcal{Y}, 0 \leq s \leq t\}$ (we introduce $a_0$ for notational convenience but the next action and state might only depend on $y_0$). Let $\Pi$ be the set of policies $\pi = (\pi_1, \pi_2, \cdots)$, with $\pi_{t+1} : \mathcal{H}_t \times \mathcal{A} \to [0, 1]$ a probability measure on $\mathcal{A}$ conditioned on $h_t$ in $\mathcal{H}_t$. Let $\Pi_t$ denote the set of all policies $\pi_t$ at time $t$ with $\pi \in \Pi$. We can simulate a policy $\pi$ in the following manner:

$$a_{t+1} = q_{\pi, t+1}(h_t, \nu_{t+1})$$

$$x_{t+1} = f_{\pi, t+1}(x_t, h_t, \nu_{t+1}, \omega_{t+1}) := h(x_t, a_{t+1}, \omega_{t+1})$$

$$y_{t+1} = g_{\pi, t+1}(x_t, h_t, \nu_{t+1}, \omega_{t+1}, \zeta_{t+1}) := g(x_{t+1}, \zeta_{t+1})$$

where $\nu_t, \omega_t, \zeta_t$ are i.i.d. uniform $[0,1]$, and $q_{\pi, \nu}$, $h$ and $g$ are simulation functions under the simple simulation model (which simulate $\pi_{t+1}$, $P_\pi$ and $\tau$ respectively). The initial state-action pair $(x_0, a_0)$ is drawn according to the given initial state-action distribution $\lambda$.

Denote $z_t = (x_t, h_t) \in \mathcal{Z}_t := \mathcal{X} \times \mathcal{H}_t$ and $\xi_t = (\nu_t, \omega_t, \zeta_t) \in \Omega_3 := \mathcal{X} \times \mathcal{A} \times \mathcal{A}$, and $\xi = (\xi_1, \xi_2, \cdots)$ then

$$\pi_{t+1}(z_t, \xi_t) = (z_{t+1}, \theta \xi_t) = (q_{\pi, t+1}(h_t, \nu_{t+1}), f_{\pi, t+1}(x_t, h_t, \nu_{t+1}, \omega_{t+1}), g_{\pi, t+1}(x_t, h_t, \nu_{t+1}, \omega_{t+1}, \zeta_{t+1}), h_t, \theta \xi_t)$$

where $\theta$ is a left-shift operator. The function sequence $\pi = (\pi_1, \pi_2, \cdots)$ will be called the simulation system for a general policy $\pi$. The function sequences $q_{\pi, \nu} = (q_{\pi_1, \nu}, q_{\pi_2, \nu}, \cdots)$, $f_{\pi, \nu} = (f_{\pi_1, \nu}, f_{\pi_2, \nu}, \cdots)$ and $g_{\pi, \nu} = (g_{\pi_1, \nu}, g_{\pi_2, \nu}, \cdots)$ will be called action, state and observation simulation functions corresponding to policy $\pi$.

We denote $Q = \{q_{\pi, \nu} : \nu \in \Pi\}$, $F = \{f_{\pi, \nu} : \nu \in \Pi\}$, $G = \{g_{\pi, \nu} : \nu \in \Pi\}$ and the set of all simulation functions induced by $\Pi$ by $S = \{s_{\pi, \nu} : \pi \in \Pi\}$. Note that we can define $s_{\pi, \nu, \cdots \cdot \cdot \cdot}$ and $s_{\pi, \nu, \cdots \cdot \cdot \cdot \cdot \cdot \cdot \cdot}$, which we shall denote by $s_{\pi, \nu, \cdots \cdot \cdot \cdot}$.

We first connect the P-dimensions of $\Pi_t$ and $Q_t := \{q_{\pi, \nu} : \pi \in \Pi_t\}$. ($S_t$ and $F_t$ will be defined similarly).

Lemma 3: Suppose $\Pi_t$ is convex and P-dim ($\Pi_t$) = $d_t$. Then, P-dim ($Q_t$) = $d_t$.

The proof follows that of lemma 1 and is omitted.

Let $\mu$ be a probability measure on $\OmegaZ_3$ and $\lambda_t$ a measure on $\mathcal{Z}_t-1$. Denote the product measure on $\mathcal{W}_{t-1} = \OmegaZ_3 \times \mathcal{Z}_t-1$ by $\mathbb{P} = \lambda_t \times \mu$, and on $\mathbb{W}_t$ by $\mathbb{P}_t$. Define the two pseudo-metrics on $S_t$:

$$\rho_t(s_{1t}, s_{2t}) = \sum_{z \in \mathcal{Z}_{t-1}} \lambda_t(z) \mu(\xi : s_{1t}(z, \xi) \neq s_{2t}(z, \xi)),$$

and

$$d_{L_1(\mathbb{P}_t)}(s_{1t}, s_{2t}) := \sum_{z \in \mathcal{Z}_{t-1}} \lambda_t(z) \int |s_{1t}(z, \xi) - s_{2t}(z, \xi)| d\mu(\xi).$$

The $\rho_t$ and $d_{L_1(\mathbb{P}_t)}$ pseudo-metrics for the function spaces $Q_t$ and $F_t$ are defined similarly.
Define for \( s \in \mathcal{S} \) and \( z \in \mathbb{Z}_d \),
\[
\lambda_d(z) := \sum_{z' \in \mathbb{Z}_d} \lambda(z') \mu_\epsilon \{ x : s^t(x, \xi) = (z, \theta^t \xi) \}
\]  
be a probability measure on \( \mathbb{Z}_d \). We now state the extension of the technical lemma needed for the main theorem of this section.

**Lemma 4:** Let \( \lambda \) be a probability measure on \( \mathbb{Z}_0 \) and \( \lambda_d \) be the probability measure on \( \mathbb{Z}_d \) as defined above. Suppose that P-dim \( (Q_d) \leq d, \forall t \geq 1 \), and there exists probability measures \( \lambda_t \) on \( \mathbb{Z}_d \) such that \( K := \max \{ \sup \mu_\epsilon \sup_{z' \in \mathbb{Z}_d} \lambda_d(z') \} \) is finite. Then, for \( 1 \leq t \leq T \),
\[
N(\epsilon, S^t, \rho_1) \leq N(\frac{\epsilon}{Kt}, Q_t, \rho_1) \cdots N(\frac{\epsilon}{Kt}, Q_1, \rho_1) \leq \left( \frac{2eKt}{\epsilon} \log \frac{2eKt}{\epsilon} \right)^{dt}. \]
The proof can be found in the appendix. We now obtain our sample complexity result.

**Theorem 2:** Let \( (\mathbb{Z}, \Gamma, \lambda) \) be the measurable state-action space, \( \mathcal{Y} \) the observation space, \( P_a(x, x') \) the state transition function and \( r(y|x) \) the conditional probability measure that determines the observations. Let \( r(x, a) \) the real-valued reward function bounded in \([0, \bar{R}]\). Let \( \Pi \) be the set of stochastic policies (non-stationary and with memory in general), \( \mathcal{S} \) be the set of simple simulation systems that simulate \( \pi \in \Pi \). Suppose that P-dim \( (Q_\Pi) \leq d, \forall t \geq 1 \), and let the probability measures \( \lambda, \mu \) (on \( \mathbb{Z}_0 \) and \( \Omega^\infty \) respectively) and \( \lambda_{t+1} \) be such that \( K := \max \{ \sup \mu_\epsilon \sup_{x, z \in \mathbb{Z}_d} \lambda_d(z') \} \) is finite, where \( \lambda_d \) is as defined above. Let \( \hat{V}_n^T(\pi) \), the estimate of \( V(\pi) \) obtained from \( n \) samples with \( T \) steps. Then, given any \( \epsilon, \delta > 0 \), with probability at least \( 1 - \delta \),
\[
\sup_{\pi \in \Pi} | V^T_n(\pi) - V(\pi) | < \epsilon
\]
for \( n \geq \frac{32R^2}{\epsilon^2} \left( \log \frac{4}{\delta} + 2dT(\log 32eK + \log KT) \right) \) where \( T \) is the \( \epsilon/2 \) horizon time and \( \alpha = \epsilon/2(T + 1) \).

The above results can be extended to Multi-armed Markov bandits with discounted rewards. Due to space constraints, such results will be omitted.

**VI. SIMULATION-BASED OPTIMIZATION**

We propose a simulation-based optimization framework based on the empirical process theory for Markov decision processes developed above and in [18]. Given an MDP \( M \) with a convex and compact policy space \( \Pi \). Let \( \mathcal{Q} \) be the set of simple action-simulation functions that simulate policies \( \Pi \). Let \( \mathcal{Q}_d \) be an \( \epsilon \)-net for \( \mathcal{Q} \) under the \( d_{L_1} \) metric. Denote the set of policies in \( \Pi \) corresponding to its \( \epsilon \)-net by \( \hat{\Pi}_\epsilon \). We know from Lemma 1 that if the P-dimension of \( \Pi \) is some finite \( d \), the P-dimension of \( \mathcal{Q} \) is also \( d \), which implies that the \( \epsilon \)-net \( \hat{\Pi}_\epsilon \) is finite, and its cardinality is bounded by \( n_1(\epsilon, N_\lambda) = 2(x \log x)^d \), where \( x = \frac{2eKd}{\epsilon} \), where \( N_\lambda \) is the cardinality of the action set \( A \).

We pick each \( \pi \in \hat{\Pi}_\epsilon \) and simulate \( n \) sample-paths for \( T \) times steps, which is the \( \epsilon/2 \)-horizon time for a given \( \epsilon > 0 \). Thus, we obtain estimates \( \hat{V}_n^T(\pi) \). We pick
\[
\hat{\pi}^* \in \arg \sup_{\pi \in \hat{\Pi}_\epsilon} \hat{V}_n^T(\pi).
\]
The optimal policy is, of course,
\[
\pi^* \in \arg \sup_{\pi \in \Pi} V(\pi).
\]
We also define
\[
\hat{\pi} \in \arg \sup_{\pi \in \hat{\Pi}_\epsilon} \hat{V}_n^T(\pi).
\]
Define the regret of a policy as
\[
\varrho(\pi) := V(\pi^*) - V(\pi).
\]
Now, from results of section IV, we know that for a given \( \epsilon, \delta > 0 \), and with \( n \geq n_0(\epsilon/3, \delta) \), with probability at least \( 1 - \delta \)
\[
V(\pi) \leq V^T_n(\pi) + \epsilon/3, \forall \pi \in \Pi,
\]
and in particular for \( \pi = \pi^* \). Also,
\[
\hat{V}_n^T(\pi) \leq V(\pi) + \epsilon/3, \forall \pi \in \Pi,
\]
and in particular for \( \pi = \hat{\pi} \).

Consider any \( \pi_1, \pi_2 \in \Pi \) with simulation functions \( q_1, q_2 \in \mathcal{Q} \) and corresponding \( s_1, s_2 \in \mathcal{S} \). Then, from the proof of theorem 1, we know that
\[
d_{L_1}(\hat{V}_n^T(\pi_1), \hat{V}_n^T(\pi_2)) \leq R \cdot \sum_{i=0}^{T} \gamma^i d_{L_1}(s_1^i, s_2^i)
\]
\[
\leq R \cdot \sum_{i=0}^{T} \gamma^i K_i^t d_{L_1}(q_1, q_2).
\]

Then, if \( d_{L_1}(q_1, q_2) \leq \varepsilon \), then
\[
d_{L_1}(\hat{V}_n^T(\pi_1), \hat{V}_n^T(\pi_2)) \leq R \cdot \frac{1 - (\gamma K)^{T+1}}{1 - \gamma K} \leq \varepsilon/3
\]
for \( \gamma K < 1 \) and \( \varepsilon \leq \frac{\varepsilon(1 - \gamma K)}{3R(1 - (\gamma K)^{T+1})} \).

Now, note that if we take \( \pi_1 = \pi^* \), then by definition of an \( \epsilon \)-net, there exists a \( \pi_2 = \hat{\pi} \in \hat{\Pi}_\epsilon \) such that \( d_{L_1}(q_1, q_2) \leq \varepsilon \). Thus,
\[
\hat{V}_n^T(\pi^*) \leq \hat{V}_n^T(\hat{\pi}) + \epsilon/3 \leq \hat{V}_n^T(\hat{\pi}) + \epsilon/3. \tag{8}
\]
From equations (6), (7) and (8), we get that
\[
V(\pi^*) \leq V(\hat{\pi}^*) + \epsilon,
\]
i.e., \( \hat{\pi}^* \) is an \( \epsilon \)-optimal policy.

Formally, we have shown that

**Theorem 3:** Given an MDP with countably infinite state space and a finite action space (of cardinality \( N_q \)), with a convex, compact policy space \( \Pi \) with P-dim(\( \Pi \))= \( d \), an \( \epsilon, \delta > 0 \), and \( \gamma < 1/K \), if we estimate the value function for each policy in a given \( \epsilon \)-net for \( \mathcal{Q} \) (with \( \varepsilon < \frac{\epsilon(1 - \gamma K)}{3R(1 - (\gamma K)^{T+1})} \)) by doing \( n_0(\epsilon/3, \delta) \) simulation runs,
each for $T$ times steps, then the obtained policy $\hat{\pi}$ is $\epsilon$-optimal, in the sense that

$$\Pr^n\{\varrho(\pi) > \epsilon\} < \delta.$$ 

And moreover the sample complexity is given by

$$n_0(\epsilon/3, \delta) \cdot n_1(\epsilon, N_h) \sim O\left(\frac{1}{\epsilon^2} \log^2 \frac{1}{\epsilon} \log \frac{1}{\delta}\right),$$

a polynomial in $1/\epsilon$.

**APPENDIX**

Proof of Lemma 4.

The proof of Lemma 4 is similar but the details are somewhat more involved.

*Proof:* Consider any $s_1^t, s_2^t \in F^t$ and $z \in Z$. Then,

$$\mu(\xi : s_1^t(z, \xi) \neq s_2^t(z, \xi))$$

$$\leq \mu(\xi : s_1^t(z, \xi) \neq s_2^t(z, \xi), s_1^t(z, \xi) = s_1^t(z, \xi))$$

$$+ \mu(\xi : s_1^t(z, \xi) \neq s_2^t(z, \xi), s_1^t(z, \xi) = s_1^t(z, \xi))$$

for all $t \leq T$.

Multiplying both RHS and LHS of the above sequence of inequalities and summing over $z$, and observing again that

$$\mu(\xi : s_1(z, \theta^{t-1}) \neq s_2(z, \theta^{t-1}))$$

we get that the first part of RHS is

$$\leq K \cdot \sum_{z \in Z_{t-1}} \lambda_t(z) \mu(\xi : s_1(z, \xi) \neq s_2(z, \xi)).$$

This by induction implies

$$\rho_1(s_1^t, s_2^t) \leq K(\rho_t(s_1t, s_2t) + \cdots + \rho_1(s_{11}, s_{2t})),$$

which implies the first inequality. For the second inequality, note that the $\rho$ pseudo-metric and the $L_1$ pseudo-metric are related thus

$$\sum_\xi \lambda_t(z) \mu(\xi : s_1(z, \xi) \neq s_2(z, \xi))$$

where the first inequality follows because only the events $\{q_{1t} \neq q_{2t}, f_t = f_t, q_{1t} = g_{2t}\}$, and $\{q_{1t} \neq q_{2t}, f_t \neq f_{2t}\}$ have non-zero probability and non-zero $L_1$ distance. Both these events are contained in $\{q_{1t} \neq q_{2t}\}$. The second inequality follows because $\mu$ is a probability measure and $|q_{1t} - q_{2t}| \geq 1$ when $q_{1t} \neq q_{2t}$, and thus $\rho_1(s_{1t}, s_{2t}) \leq K \cdot d_f(z_{1t}, z_{2t})$ which proves the required assertion.

**REFERENCES**


