Least Squares Monte Carlo Approach to Convex Control Problems

René Carmona
Bendheim Center for Finance
Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544.
E-mail rcarmona@princeton.edu.

J. Hinz
National University of Singapore,
Department of Mathematics,
10 Lower Kent Ridge Road, Singapore 119076
E-mail mathj@nus.edu.sg

Optimal control problems with convex functions are ubiquitous in applications of stochastic optimization. However, when applied in this context, the classical least squares Monte Carlo methodology makes no attempt to take advantage of this special structure: Given the convexity of value functions, it seems reasonable to search for the best least-squares fit among the elements of a cone of convex functions, rather than among a linear space of feature functions as required by the classical least squares Monte Carlo approach. In the present work, we build on this idea by introducing an appropriate modification to the classical method. We show that computation time and accuracy can be improved by using projections on convex cones instead of projections onto linear spaces.

Partially supported by NSF - DMS 0806591

Index Terms

Markov Decision, Approximate dynamic programming, Least Squares Monte Carlo

I. INTRODUCTION

When making decisions under uncertainty, the major difficulty is to determine how to update estimates and decisions in order to achieve optimality over a given time period. Most often, these kind of questions are framed within the realm of *Markov decision theory* which can be viewed as discrete-time *optimal stochastic control*. Following Bellman's principle, the instantaneous reward of re-positioning and the expected future revenue must be balanced against each other in order to find an optimal action at any given time.

The theoretical underpinnings of Markov decision theory are now well-understood. Rigorous mathematical treatments are available in textbook form. See for instance in [3], [13], [16] and [24]. We often refer to the recent book [1] which provides a well balanced introduction to the theory with applications in finance. However, practical applications remain persistently challenging despite the rich arsenal of theoretical tools available nowadays. Indeed more often than not, the complexity of typical real-world implementations goes beyond what is computationally feasible, and a great variety of ad-hoc methods were developed to serve different purposes. In this context, *approximate dynamic programming* grew from attempts at providing simultaneously practically implementable heuristics and theoretical insights on the reasons why these heuristics perform well in practice. The book [23] provides a modern survey on the state-of-art in this domain and on current challenges from an industrial perspective.

In order to control a large system, a practical solution to the high dimensionality of the state space is a finite discretization of the latter. Alternatively, one can rely on an efficient approximation of functions on this space. In this spirit, the least squares Monte Carlo approach suggests to approximate functions on the state space by linear combinations of a set of basic *feature functions*. Motivated by financial applications, most importantly the pricing of American options on large baskets of underlying interests, the least squares Monte Carlo method has attracted the attention of most quants and financial engineers over the last decade. Following earlier works [7], [26], [27], the contribution of Longstaff and Schwartz [18] has enjoyed an unprecedented popularity and became the source of subsequent research focused on its theoretical

justification. For instance, convergence issues are addressed in [9] and later generalized in [25], [10] and [11], extensions to multiple exercise rights were considered in [6], and most recently studied in [2] where the connections to statistical learning theory and the theory of empirical processes is emphasized. We also mention some related work on function-based methods in financial mathematics: [19] addresses variance reduction techniques, [8] examines opportunities for parallelization, and [17] discusses efficiency of weak Taylor schemes, whereas [5] gives a survey of basic feature functions methods for the valuation of options. For an overview of the applications of Monte Carlo methods in financial engineering we refer the interested reader to Glasserman's book [14] and to the literature cited therein.

Beyond financial applications, value function approximation methods have been used to capture the local behavior of value functions, and advanced regression methods, e.g. kernel methods [20], [21], local polynomial regression [12], and neural networks [4], have been brought to bear with this goal. In the particular case of *partially observable Markov decision processes*, several specific approaches been suggested [22]. The survey [15] gives an overview of these methods and describes applications to autonomous robot navigation.

One of the main advantage of the least-squares approach is that it reduces computations to simple linear algebraic operations in low dimension. However, combined with the successive iterations required by the implementation of the dynamic programming principle, instability and divergence are frequently observed. The thrust of this paper is to provide a modification, which when applicable, stabilizes the iterative process. Our procedure requires the convexity of the value function when a conditionally deterministic component is fixed. This assumption is satisfied in most financial applications. Furthermore, as a by product of our analysis, we suggest a methodology for an adaptive choice of a dictionary of basic feature functions, method which appears to be helpful when tackling high-dimensional control problems.

The paper is organized as follows. In Section II, we recall the standard set-up of infinite horizon Markov decision processes. In the following Section III, we tailor this general framework to the case of stochastic systems with a conditionally deterministic component. For tractability reasons, and motivated by practical applications, we choose this conditionally deterministic component to take only a finite number of values, while the remaining components which can be viewed as environmental variables or random factors, have a standard controlled Markovian evolution. Due to such a split of the set of state variables, we are able to take full advantage of the convexity of

the value function in the environment component which typically runs through a high-dimensional space. These models offer great flexibility in capturing the special features of many practical applications. In particular, they can be used to analyze optimal stopping and optimal switching problems as well as a large number of problems of control of *partially observable Markov decision processes*. In Section V, we introduce and analyze a modification of the least squares method taking advantage of the convexity of the value function in some variables. The last Section VII reports on implementations on specific examples and provides illustrations of the performance of the method.

II. MARKOV DECISION THEORY

We review the classical framework of infinite-horizon Markov decision theory following [1]. The state of the system varies in a measurable space $(E, \mathcal{E})$ and is affected by actions from a set A of possible actions. A mapping from E into A is called a decision rule. For each $a \in A$, we assume that $K_a(x, dx')$ is a stochastic kernel on $(E, \mathcal{E})$ and we consider a fixed sequence $(X_t)_{t\in\mathbb{N}}$ of random variables. One can think of them as the coordinate projections on the product of an infinite number of copies of $(E, \mathcal{E})$. In any case, for each initial point $x \in E$ and each sequence $\pi = (\pi_t)_{t\in\mathbb{N}}$ of decision rules, we consider the probability measure $\mathbb{P}_x^\pi$ for which $\mathbb{P}_x^\pi(X_0 = x) = 1$ and

$$\mathbb{P}_x^\pi(X_{t+1} \in B \,|\, X_0, \ldots, X_t) = K_{\pi_t(X_t)}(X_t, B) \tag{1}$$

for each measurable $B \in \mathcal{E}$ and $t \in \mathbb{N}$. So at time $t \in \mathbb{N}$ when the system is in state $X_t = x$, the action $a = \pi_t(X_t)$ is used to pick the transition probability $K_{,a=\pi_t(X_t)}$ giving the random evolution of the state from X_t to a random variable X_{t+1} with distribution $K_{\pi_t(X_t)}(X_t, \cdot)$. For the sake of notational convenience, we introduce a special notation, say T_a, for the one-step transition operator associated to the transition kernel K_a when the action $a \in A$ is chosen. In other words, for each action $a \in A$ we define the operator T_a on functions f by

$$(T_a f)(x) = \int_X f(x') K_a(x, dx') \qquad x \in E,\ t \in \mathbb{N} \tag{2}$$

whenever the above integrals are well-defined. We also assume that we are given a function $R : E \times A \mapsto \mathbb{R}$, called the *one-step reward function* as $R(x, a)$ stands for the reward for applying at any given time, action $a \in A$ in state $x \in E$. Our goal is to maximize the expected

discounted infinite-horizon total reward, in other words to find the argument $\pi^* = (\pi_t^*)_{t\in\mathbb{N}}$ of the maximization problem

$$\pi^* = \arg\sup_{\pi\in\mathcal{A}} \mathbb{E}_x^\pi \left(\sum_{t=0}^{\infty} \gamma^t R(X_t, \pi_t(X_t)) \right) \tag{3}$$

where $\gamma \in (0,1)$ is a fixed discount factor, $\mathcal{A}$ the set of admissible sequences of decision rules $\pi = (\pi_t)_{t\in\mathbb{N}}$, and $\mathbb{E}_x^\pi$ denotes expectation over the controlled Markov chain defined by (1). The maximization (3) is well-defined under the *integrability assumption*

$$\sup_{\pi\in\mathcal{A}} \mathbb{E}_x^\pi (\sum_{t=0}^{\infty} \gamma^t R(X_t, \pi_t(X_t))^+) < \infty, \qquad x \in E, \tag{4}$$

which is clearly satisfied if for example the function R is bounded. Furthermore, by introducing the *convergence assumption*

$$\lim_{n\to\infty} \sup_{\pi\in\mathcal{A}} \mathbb{E}_x^\pi (\sum_{t=n}^{\infty} \gamma^t R(X_t, \pi_t(X_t))^+) = 0, \qquad x \in E, \tag{5}$$

which is also satisfied when R is bounded, the finite-horizon problems and their limit

$$V^{(n)}(x) = \sup_{\pi\in\mathcal{A}} \mathbb{E}_x^\pi (\sum_{t=0}^{n-1} \gamma^t R(X_t, \pi_t(X_t))), \qquad V^*(x) = \lim_{n\to\infty} V^{(n)}(x), \qquad n \in \mathbb{N},\ x \in E \tag{6}$$

are well defined and the total reward maximization (3) becomes tractable. Namely, under additional (rather technical) *structure assumptions* (see [1], p. 199) it holds that the iteration

$$V^{(n+1)}(x) = \max_{a\in A} \left(R(x,a) + \gamma \int_E V^{(n)}(x') K_a(x, dx') \right), \qquad x \in E, \qquad V^0 = 0 \tag{7}$$

converges toward the solution of the optimality equation

$$V^*(x) = \max_{a\in A} \left(R(x,a) + \gamma \int_E V^*(x') K_a(x, dx') \right), \quad x \in E, \tag{8}$$

and that the decision rule

$$\pi_0^*(x) = \mathrm{argmax}_{a\in A} \left(R(x,a) + \gamma \int_E V^*(x') K_a(x, dx') \right), \quad x \in E$$

yields the time-independent policy $\pi^* = (\pi_t^*)_{t\in\mathbb{N}}$ defined by $\pi_t^* = \pi_0^*$ for all $t \in \mathbb{N}$ which is optimal since

$$V^*(x) = \mathbb{E}_x^\pi (\sum_{t=0}^{\infty} \gamma^t R(X_t, \pi_0^*(X_t))), \quad x \in E.$$

In terms of the operators $(T_a)_{a\in A}$, the optimal policy is given by

$$\pi^*(x) \;\; = \;\; \mathrm{argmax}_{a\in A} \left(R(x,a) + \gamma [T_a V^*](x) \right) \tag{9}$$

for all $x \in E$ where $V^*(x)$ is the limit of $(V^{(n)}(x))_{n\in\mathbb{N}}$ defined inductively by

$$V^{(n+1)}(x) = \max_{a\in A} \left(R(x,a) + \gamma[T_a V^{(n)}](x)\right), \qquad n \in \mathbb{N}, \;\; V^0 = 0. \tag{10}$$

III. Partially Deterministic Markov Systems

For the reminder of this work, we concentrate on Markov decision problems which satisfy the integrability, convergence and structure assumptions recalled above, and we search for numerical algorithms providing approximations for the computation of the value function and the optimal policy maximizing it.

We focus on Markov decision models whose state evolutions have a conditionally deterministic component. To be more specific, we assume that the state space $E = P \times Z$ is the product of of a compact metric space P and a measurable space $(Z, \mathcal{Z})$. In most applications, P is an interval or a finite set, and Z is a subset of a Eucledian space $\mathbb{R}^d$. We further assume that the evolution of the first component is conditionally deterministic given the values of the remaining components in Z. As explained in the introduction, we also assume that the time evolution of the state component in Z, which we view as the *environment components*, is given by a controlled Markov process $(Z_t)_{t\in\mathbb{N}}$ governed by a family $(k_a(z,dz'))_{a\in A}$ of transition probability kernels on Z parameterized by a compact metric set A giving the set of actions which can be taken by the controller of the system. We assume that the changes in the conditionally deterministic evolution are given by a continuous function:

$$\alpha : P \times A \ni (p,a) \hookrightarrow \alpha(p,a) \in P.$$

$\alpha(p,a) \in P$ is the new value of the first component of the state if the previous value is p and the action $a \in A$ was taken by the controller. In other words, the transition kernel K giving the evolution of the full state $X_t = (P_t, Z_t)$ is of the form:

$$K_a((p,z), d(p',z')) = \delta_{\alpha(p,a)}(dp')\, k_a(z, dz'),$$

where we denote by δ_p the Dirac unit mass at p. For such Markov decision process, the optimal policy is given by $\pi^* = (\pi_0^*)_{t\in\mathbb{N}}$, where for each state $(p,z) \in P \times Z$ the optimal action at any time is a maximizer

$$\pi_0^*(p,z) \;\; = \;\; \text{argmax}_{a\in A} \left(R(p,z,a) + \gamma \int_Z V^*(\alpha(p,a),z') k(z,dz',a) \right) \tag{11}$$

which is obtained from the dynamic programming principle as the pointwise limit

$$V^* = \lim_{n\to\infty} V^{(n)}$$

of iterated value functions defined inductively by

$$V^{(n+1)}(p,z) = \max_{a\in A}\left(R(p,z,a) + \gamma\int_Z V^{(n)}(\alpha(p,a),z')k(z,dz',a)\right), \qquad V^0 = 0. \tag{12}$$

As before, it is convenient to rewrite this equation in terms of the operator τ_a associated to the transition kernel k_a and defined on bounded measurable functions on $(Z,\mathcal{Z})$ by:

$$(\tau_a\varphi)(z) = \int_Z \varphi(z')k_a(z,dz') \qquad z\in Z,\ t\in\mathbb{N} \tag{13}$$

and the iterated value functions satisfy

$$V^{(n+1)}(p,z) = \sup_{a\in A}\left(R(p,z,a) + \gamma[\tau_a V^{(n)}(\alpha(p,z),\,\cdot\,)](z)\right), \qquad p\in P,\, z\in Z,\, n\in\mathbb{N},\ \ V^0 = 0. \tag{14}$$

Before we proceed, we describe an example of a great practical importance for which the above set up is natural. This example is an abstraction of a gas storage facility management model used in [**?**]. The solution of such a stochastic control problem is also used for valuation and hedging purposes within risk management of large portfolio of energy financial and physical assets.

Example. Consider a storage management problem, where the level of the commodity stored in the facility needs to be controlled over time. In this application, P is the set of possible levels of the commodity in the storage facility. Given storage costs and random price fluctuations, the controller has to decide when to purchase the commodity and add it to the storage, or withdraw from storage and sell it at the market price. A is the set of actions which can be taken in order to change the level in the storage facility. The action a yields a transition from the previous storage level p to the new level $\alpha(p,a)$. In the simplest form of this example, $(Z_t)_{t\in\mathbb{N}}$ describes the Markovian evolution of the market (spot) price of the underlying commodity. More generally, the state Z_t at time t could be multivariate, in which case one of the components of Z_t is usually the market (spot) price of the commodity at time t. The other components may be latent variables representing the current market conditions, stochastic factors which can also be needed to ensure the Markov property of the dynamics. In this example, the value $R(p,z,a)$ describes the cash flow associated with the decision a to buy or sell commodity at time t. Note that the value

$R(p, z, a)$ depends not only on the action a and on the market price through the corresponding price-component of z, but may also depend on the current inventory level p. For instance, in the case of gas storage, the injection/withdrawal rates and costs depend upon the storage level through the physical laws restricting gas pressure.

For the purpose of illustration we consider the simple case in which the storage level can either be full, half-full, or empty, and the agent must decide at which price and time to buy the commodity and fill the storage facility, or sell the commodity he has in storage to the market. This problem could also be viewed as an instance of a so-called *optimal switching* problem. For the sake of definiteness we set $P = \{1, 2, 3\}$ where 1 stands for "empty", 2 for "half-full" and 3 for "full", and $A = \{1, 2, 3\}$ where 1 stands for "withdraw and sell", 2 for "store and do nothing", and 3 for "inject and buy", with changes in the first component given by the function α whose values we give in the form of a table

		a		
		inject	store	withdraw
	empty	half-full	empty	empty
p	halpf-full	full	half-full	empty
	full	full	full	half-full

or equivalently, in the form of a matrix:

$$(\alpha(i,j))_{i,j=1}^{3} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 3 \end{bmatrix} \quad \text{for } (p,a) \in P \times A. \tag{15}$$

Clearly, P and A do not need to have the same numbers of elements, i.e. the number of levels of the storage facility does not have to be equal to the number of actions the operator can take. In this three-level storage example, we may assume that the commodity market price evolution is described by the last component $(Z_t^{(d)})_{t\in\mathbb{N}}$ of the process $(Z_t = (Z_t^1, \ldots, Z_t^d))_{t\geq 0}$. Hence for each $t \geq 0$, the instantaneous reward is given by the following affine linear functions in the variable $z = (z^{(1)}, \ldots, z^{(d)}) \in \mathbb{R}^d$

$$R(p, (z^{(1)}, \ldots, z^{(d)}), a) = -c|p - \alpha(p,a)| + (p - \alpha(p,a))z^{(d)}, \tag{16}$$

thereby, the constant $c > 0$ represents proportional transaction costs.

IV. Existence by Fixed Point Arguments

In this section, we use fixed point arguments to prove existence (and in some cases uniqueness) for the optimality equation (8). For our first theoretical result, we assume that Z is a locally compact metric space with Borel σ-field $\mathcal{Z}$, and we denote by $C(P \times Z)$ the real separable Banach space of real valued bounded uniformly continuous functions on $P \times Z$ equipped with the supremum norm.

Theorem 1. *If we assume that the reward function R is bounded and uniformly continuous in its three variables, and that for any non-negative continuous function φ on Z we have:*

$$\sup_{z\in Z}\sup_{a\in A}[\tau_a\varphi](z) \leq \sup_{z\in Z}\varphi(z) \tag{17}$$

then the map $V \hookrightarrow \phi(V)$ defined by

$$[\phi(V)](p,z) = \sup_{a\in A}\left[R(p,z,a) + \gamma[k_a V(\alpha(p,z),\,\cdot\,)](z)\right], \qquad p \in P,\, z \in Z$$

is a strict contraction on $C(P \times Z)$, and the optimality equation has a unique fixed point in this space.

Note that assumption (17) is clearly satisfied when the transition kernel k_a is independent of $a \in A$. This is the case in the application to commodity storage management and valuation used as a motivating example for our analysis.

Proof: If V_1 and V_2 are two elements of $C(P \times Z)$, then

$$\begin{aligned}
\|\phi(V_2) - \phi(V_1)\| &= \sup_{(p,z)\in P\times Z} |\phi(V_2)(p,z) - \phi(V_1)(p,z)| \\
&= \sup_{(p,z)\in P\times Z} \Big| \sup_{a\in A}\left[R(p,z,a) + \gamma[\tau_a V_1(\alpha(p,a),\,\cdot\,)](z)\right] \\
&\qquad\qquad - \sup_{a\in A}\left[R(p,z,a) + \gamma[\tau_a V_2(\alpha(p,a),\,\cdot\,)](z)\right] \Big| \\
&\leq \gamma \sup_{(p,z)\in P\times Z} \sup_{a\in A} |[\tau_a(V_2 - V_1)(\alpha(p,a),\,\cdot\,)](z)| \\
&\leq \gamma \sup_{z\in Z}\sup_{a\in A}[\tau_a(\sup_{p\in P}|(V_2 - V_1)(p,\,\cdot\,)|](z) \\
&\leq \gamma \|V_2 - V_1\|
\end{aligned}$$

where we used assumption (17) with $\varphi(z) = \sup_{p\in P}|(V_2 - V_1)(p,z)|$. This concludes the proof since $\gamma < 1$. □

Remark. A similar strict contraction result can be proven in Lebesgue spaces $L^p(Z, \mathcal{Z}, \mu; C(P))$ of $C(P)$-valued functions where $C(P)$ denotes the real separable Banach space of real valued continuous functions on the compact metric space P. Indeed, if we assume that μ is a probability measure on $(Z, \mathcal{Z})$ with respect to which all the probability measures $k_a(z, \cdot)$ are absolutely continuous, then under the condition

$$\left\| \sup_{a \in A} \tau_a \varphi \right\|_{L^1(\mu)} \leq \|\varphi\|_{L^1(\mu)}$$

for all non-negative functions φ, then ϕ is a strict contraction in $L^1(Z, \mathcal{Z}, \mu; C(P))$ and the optimality equation has a unique fixed point in this space.

As explained earlier, Theorem 1 is useful when the transition kernel of the random component Z_t of the state does not depend upon the action a. It is still possible to prove existence of a fixed point and a solution of the optimality equation when this condition is not satisfied. The following theoretical result implies existence (though not necessarily uniqueness) when the set A of actions is finite. As we already saw, this case is typical in practical application.

We assume that the function α is continuous and as before, that the reward function R is bounded and uniformly continuous in its three variables, and we denote by $\overline{R}$ the function

$$\overline{R}(p, a) = \sup_{z \in Z} R(p, z, a), \qquad p \in P,\ a \in A. \tag{18}$$

Classical results on infinite horizon, discrete time, deterministic optimal control problems guarantee the existence of a function $\overline{V}^* \in C(P)$ solving

$$\overline{V}^*(p) = \sup_{a \in A} [\overline{R}(p, a) + \gamma \overline{V}^*(\alpha(p, a)), \qquad p \in P. \tag{19}$$

We now denote by B the set:

$$B = \{V \in C(P \times Z);\ \forall p \in P,\ \forall z \in Z,\ V(p, z) \leq \overline{V}^*(p)\}. \tag{20}$$

We can now state and prove our second existence result:

Theorem 2. *Using the notation above, we assume that*

- *there exist a finite number N of continuous functions $\alpha_i : A \ni a \hookrightarrow \alpha_i(a) \in P$ such that for each $p \in P$ there exists $i \in \{1, \cdots, N\}$ such that $\alpha(p, a) = \alpha_i(a)$ for all $a \in A$;*
- *the reward function R is bounded and uniformly continuous in its three variables:*
- *and that for any non-negative continuous function φ on Z we have:*

- *the map $Z \ni z \hookrightarrow k_a(z, \cdot)$ into the space of probability measures on $(Z, \mathcal{Z})$ is uniformly continuous in variation norm, uniformly in $a \in A$.*

Under these conditions, the optimality equation has a solution in B.

Notice that the third assumption is satisfied when the set of admissible actions $a \in A$ is finite, while the first assumption is satisfied when P is finite as well.

Proof: Clearly, the set B is a closed convex subset of the Banach space $C(P \times Z)$. Moreover, ϕ maps B into itself. Indeed, if $V \in B$, then for every $p \in P$ and $z \in Z$ we have:

$$\begin{aligned} \phi(V)(p,z) &= \sup_{a\in A}[R(p,z,a) + \gamma \int_Z V(\alpha(p,a), z')k_a(z, dz')] \\ &\leq \sup_{a\in A}[\overline{R}(p,a) + \gamma \overline{V}^*(\alpha(p,a)) \\ &\leq \overline{V}^*(p) \end{aligned}$$

where we used the definition of $\overline{R}$ and the fact that $V \in B$ to obtain the first inequality, and the definition of $\overline{V}^*$ to derive the last inequality. This proves that $\phi(V) \in B$. So the proof will be complete if we can apply Schauder's fixed point theorem by proving that the image $\phi(B)$ is relatively compact in $C(P \times Z)$. Because of the uniform bound in the very definition of B, a form of the classical Arzela-Ascoli theorem implies that it is enough to check uniform equi-continuity of $\phi(B)$. So let $V \in B$ and (p_1, z_1) and (p_2, z_2) in $P \times Z$. Then

$$\begin{aligned} |\phi(V)(p_1,z_1) - \phi(V)(p_2,z_2)| &= \sup_{a\in A} |R(p_1,z_1,a) - R(p_2,z_2,a)| \\ &\quad + \gamma \sup_{a\in A} |[\tau_a V(\alpha(p_1,a), \cdot)](z_1) - [\tau_a V(\alpha(p_2,a), \cdot)](z_1)| \\ &\quad + \gamma \sup_{a\in A} |[\tau_a V(\alpha(p_2,a), \cdot)](z_1) - [\tau_a V(\alpha(p_2,a), \cdot)](z_2)| \\ &= (i) + (ii) + (iii). \end{aligned}$$

Given $\epsilon > 0$, by uniform continuity of the function R and the compactness of the set A, we can find $\delta_1 > 0$ so that $(i) \leq \epsilon/2$ as long as $d(p_1, p_2) < \delta_1$ and $d(z_1, z_2) < \delta_1$. Furthermore, the assumption on the fact that we only have a finite number of functions $\alpha(p, \cdot)$ implies, by compactness of P, that there exists $\delta_2 > 0$ such that $\alpha(p_1, a) = \alpha(p_2, a)$ for all $a \in A$ whenever $d(p_1, p_2) < \delta_2$. This implies that $(ii) = 0$ whenever $d(p_1, p_2) < \delta_2$. Finally, we notice that

$$(iii) \leq (\sup_{p\in P} |\overline{V}^*(p)|) \sup_{a\in A} \|k_a(z_1, \cdot) - k_a(z_2, \cdot)\|$$

which can be made smaller than $\epsilon/2$ provided that $d(z_1, z_2) < \delta_3$ for some $\delta_3 > 0$ whose existence is guaranteed by the uniform continuity of $Z \ni z \hookrightarrow k_a(z, \cdot)$ in variation norm, uniformly in $a \in A$. This concludes the proof of equicontinuity of $\phi(B)$. □

V. LEAST-SQUARES APPROXIMATIONS

The purpose of this section is to prepare the ground for the practical implementation of the solution of partially deterministic Markov decision problems introduced above. For the purpose of numerical computations, the transition operators T_a and τ_a need to be approximated. However, the convergence of the value function iterations defined in (10) is sensitive to the properties of these operators. Replacing T_a or τ_a by a numerical approximation may jeopardize the very existence of the fixed point and the limit of the sequence defined inductively in (10). Indeed problems do occur when T_a is approximated naively, e,g, as suggested by the classical least squares Monte Carlo method.

In this approach, we first fix a dictionary $\{\psi_k\}_{k=1}^m$ of feature functions, and then a sample $z'_1, \cdots, z'_n$ of n elements in Z. One can think of these as a sample of realizations of independent identically distributed random variables with common distribution μ on the measurable space $(Z, \mathcal{Z})$. Then for each possible action $a \in A$, we consider a sample $z''_1(a), \cdots, z''_n(a)$ of realizations of independent random variables with distributions $k_a(z'_1, dz''), \cdots, k_a(z'_n, dz'')$ respectively. For each $a \in A$, we denote by $\mathcal{S}(a)$ the sample $(z'_1, z''_1(a)), \cdots, (z'_n, z''_n(a))$ of n independent couples $(z', z") \in Z \times Z$. Once the dictionary $\{\psi_k\}_{k=1}^m$ and the samples $\mathcal{S}(a)$ are fixed, we use the approximate transition operator $\tilde{\tau}_a$ defined on functions $\varphi : Z \to \mathbb{R}$ by

$$\tilde{T}_a\varphi = \sum_{k=1}^m \tilde{\lambda}_k \psi_k,$$

where the coefficients $(\tilde{\lambda}_k)_{k=1}^m \in \mathbb{R}^m$ are chosen in order to minimize the sum of squared errors

$$\sum_{(z',z'')\in\mathcal{S}(a)} |\varphi(z'') - \sum_{k=1}^m \lambda_k \psi_k(z')|^2$$

over all the possible choices of $(\lambda_k)_{k=1}^m \in \mathbb{R}^m$. The feature functions ψ_k forming the dictionary are often called *basis functions* even though they may not be linearly independent. In fact, in many practical applications, they happen to be strongly dependent, their choice being dictated by the desire to take advantage of the redundancy in their features. Theoretically, the operator $\tilde{\tau}_a$ approximates τ_a if if the number m of functions in the dictionary and the size n of the set

of Monte Carlo samples are chosen sufficiently large. In [25] and [2], conditions are given on the relative growths of these two sizes which ensure the convergence of least squares Monte-Carlo transition operator approximations to the true Markov transition. Despite the existence of these theoretical results, one encounters two major problems in practical applications of the least-squares Monte-Carlo method:

1) an appropriate choice of the dictionary of basis functions turns out to be difficult, particularly for high-dimensional state spaces;
2) increasing the size of the dictionary may cause oscillations in $\tilde{\tau}_a\varphi$ if the sample size is too small.

From now on, we shall deal with specific partially deterministic Markov decision problems for which the functions

$$z \hookrightarrow V^*(p, z) \tag{21}$$

are convex for each $p \in P$. More precisely, we assume that the following conditions which are stronger than (21) are satisfied:

H1 The reward functions $z \mapsto R(p, z, a)$ are convex for all $p \in P, a \in A$;

H2 The transitions preserve convexity, i.e. for each $a \in A$, $T_a f$ is convex whenever f is.

Note that (H1) and (H2) imply (21), because $V^*(p, \cdot)$ is the pointwise limit of $(V^{(n)}(p, \cdot))_{n\in\mathbb{N}}$ which are convex, since at each iteration in (12), T_a preserves convexity, followed by a convexity-preserving maximization over A. In our storage example, condition (H1) holds due to the special form (16) of the reward. To ensure (H2), specific assumptions on the transition kernel of the Markov processes $(Z_t)_{t\in\mathbb{N}}$ must be made.

Remark The reader may be confused [1] by the mix of maximization and convexity, expecting the minimization of a convex function (or the maximization of a concave function) instead. Indeed, determining maxima of convex functions is computationally involving. Furthermore, standard applications in the portfolio optimization typically deal with maximization of concave functions. Note the situation considered here is completely different, since we do not maximize in the variable $z \in \mathcal{Z}$, in which the value function is convex. The maximization in (12) can be

[1] We are grateful to the anonymous referee for pointing out this question.

interpreted for each fixed $p \in \mathcal{P}$ as maximum over a *finite number* of convex functions on $\mathcal{Z}$, indexed by $a \in A$.

Control problems satisfying (H1) and (H2) frequently appear in applications, ranging from finance to autonomous robot navigation (see [15]). In particular, an important class of the so-called *partially observable Markov decision processes* is captured by this problem type.

In this work, we develop a simple method to overcome both difficulties under the standing convexity assumption (21) of the iterated value functions. To be more specific, knowing the convexity of the targeted function $\tau_a\varphi$, it appears natural to search for a best least-squares fit among a *cone of convex functions* rather than among a linear space. Following this idea, we pre-specify a cone spanned by non-negative linear combinations of convex basis functions and determine the least squares projection on this cone. With this, we address the second bullet point, adding significant stability with respect to increase in the size of the dictionary, since due to the convexity restriction, no oscillation occurs. Concerning the first bullet point, we suggest an adaptive method to choose the basis functions of the dictionary. Based on the analysis of practical examples for which the functions (21) are convex, we conclude that the most efficient choice is to select a basis dynamically, depending on and adjusted to the function φ being projected. As we shall see later, our choice for the cone spanned by the dictionary $(\psi_j)_{j=1}^m$ appears naturally in the spirit of projection pursuit regression as we use compositions of the function φ with affine functionals, namely $\psi_j(z) = \varphi(l_j(z))$ for $z \in Z$ for a particular choice of affine linear functionals $l_j : Z \to Z$ whose type is derived from an analysis of the transition operator τ_a.

VI. Fixed Point Construction

As in the classical least squares method, we consider for each $a \in A$ a finite set of Monte Carlo samples, say $\mathcal{S}(a) \subset Z \times Z$, which consists of pairs $(z', z'') \in Z \times Z$ of independent realizations of a couple $(Z', Z''(a))$ of random variables in $Z \times Z$ satisfying

- The first components z' form a sample (denoted by the letter Ξ) from the distribution of the random variable Z';
- For each $a \in A$, $\mathbb{P}(Z''(a) \in dz'' \,|\, Z' = z') = k(z', dz'', a)$, for all $z' \in Z'$.

We propose to overcome the problems articulated in the bullet points above by using a modification of the Monte-Carlo transition operator $\tilde{\tau}_a$, which we call *approximative transition*, and

denote $\breve{\tau}^b(a)$. For each $a \in A$, this operator is defined by

$$\breve{\tau}^b(a) f = \sum_{j=1}^{m} \breve{\lambda}_j(a) \psi_j, \tag{22}$$

where the coefficients $(\breve{\lambda}_j(a))_{j=1}^m \subset [0, \infty)$ solve the constrained optimization problem:

$$\inf_{\substack{\lambda_j \geq 0, j=1,\cdots,m, \\ \max_{z' \in \Xi} \sum_{j=1}^m \lambda_j \psi_j(z') \leq \max_{z' \in \Xi} f(z')}} \sum_{(z', z'') \in \mathcal{S}(a)} |f(z'') - \sum_{j=1}^{m} \lambda_j \psi_j(z')|^2. \tag{23}$$

The boundedness constraint is crucial for the existence of a fixed point $\breve{V}^*$ of the functional equation:

$$\breve{V}^*(p, z) = \max_{a \in A} \Big(R(p, z, a) + \gamma \breve{\tau}^b(a) \breve{V}^*(\alpha(p, a), \cdot)(z) \Big) \qquad (p, z) \in P \times \Xi. \tag{24}$$

See Theorem 3 below. However, this boundedness condition is not sufficient and additional assumptions are required in order to ensure that a solution $\breve{V}^*$ (24) exists. We suppose that the realizations of all rewards on Ξ are non-negative

$$R(a, z, p) \geq 0 \quad a \in A, z \in \Xi, p \in P \tag{25}$$

and introduce their maximum

$$R(a, z, p) \leq \bar{R}(a, p) := \max_{z \in \Xi} R(a, z, p) \quad a \in A, z \in \Xi, p \in P. \tag{26}$$

Further, we assume that the feature functions of the dictionary are chosen such that

$$\{(\psi_j(z))_{z \in \Xi} \, : \, j = 1, \ldots, m\} \text{ are linearly independent for each } a \in A. \tag{27}$$

and that

$$\text{all basis functions } (\psi_j)_{j=1}^m \text{ are strictly positive} \tag{28}$$

Remark All three assumptions are not restrictive. Namely, given reward functions, we always can address a related problem, whose rewards satisfy (25) and (26) and are given by

$$R_+(p, z, a) = R(p, z, a) - \min\{R(p, z, a) \, |, p \in \mathcal{P}, z \in \Xi, a \in \mathcal{A}\}. \tag{29}$$

for all $z \in \mathcal{Z}$, $p \in \mathcal{P}$, $a \in \mathcal{A}$. A straight-forward verification shows that by adding the same constant to each reward function, one obtains another Markov decision problem which possess exactly the same optimal policy. That is, we always can transform the reward R of a given problem by (29) to ensure (25). Considering (27), the independence feature functions as functions

on $\mathcal{Z}$ is a natural, since otherwise one can omit some feature elements. Their linear independence as functions on the sample Ξ is fulfilled if Ξ is sufficiently rich, which is typically the case in applications. Finally, given non-negative rewards, all value functions $(V^n)_{n\geq 0}$ are also non-negative. Having in mind their convexity, we agree that it is reasonable to chose our cone feature as in (**??**).

Theorem 3. *Suppose that (25), (27), and (28) hold. Given $(\breve{T}^b(a))_{a\in A}$ as defined by (22) and (23), there exists a solution $\breve{V}^*$ to (24).*

The proof of this theorem requires auxiliary results. First, we examine how the right hand side of (24) acts at the level of coefficients. Given a collection $\beta = ((\beta_j(p,a))_{j=1}^m)_{(p,a)\in P\times A}$ of non-negative coefficients, introduce

$$W^\beta(p,z) = \max_{a\in A}\left(R(p,z,a) + \gamma\sum_{j=1}^m \beta_j(\alpha(p,a),a)\psi_j(z)\right) \quad \text{for all } (p,z)\in P\times\Xi. \tag{30}$$

Expressing for each $a\in A$ and $p\in P$ the function $\breve{T}^b(a)W^\beta(p,\cdot)$ by basis $(\psi_j)_{j=1}^m$ yields new non-negative coefficients $I\beta = (((I\beta)_j(p,a))_{j=1}^m)_{(p,a)\in P\times A}$, which are determined by

$$(\breve{T}^b(a)W^\beta(p,\cdot))(z) = \sum_{j=1}^m (I\beta)_j(p,a)\psi_j(z) \quad \text{for all } p\in P, a\in A,\ z\in Z. \tag{31}$$

Lemma 1. *The mapping*

$$I : ([0,\infty[^m)^{P\times A} \to ([0,\infty[^m)^{P\times A} \qquad \beta\mapsto I\beta$$

defined by (30) and (31) is continuous.

Proof: Write I as a concatenation of the following mappings

$$\beta \mapsto (W^\beta(p,\cdot))_{p\in P} \mapsto (\breve{T}^b(a)W^\beta(p,\cdot))_{(p,a)\in P\times A} \mapsto I\beta. \tag{32}$$

The first function maps continuously $([0,\infty[^m)^{P\times A}$ into the finite dimensional space of vectors (indexed by P) of functions on Ξ. The second mapping is continuous, being a componentwise application of the restricted cone projections $(\breve{T}^b(a))_{a\in A}$. The last mapping acts as a componentwise coefficient representation with respect to the linearly independent vectors (27) and is also continuous. Note that all vector spaces in the chain (32) are finite-dimensional, hence the continuity can be understood with respect to the usual Euclidean norm.

Next, we consider an associated Markov decision problem, where the state space is P and the deterministic state transition is defined by the actions A as follows: If $a \in A$ is applied in the state $p \in P$, then the system jumps with certainty from $p \in P$ to $\alpha(a,p) \in P$ and yields an instantaneous reward $\bar{R}(a,p)$ from (26). Applying standard results from Markov decision theory, we obtain the value function $\bar{V}^* = (\bar{V}^*(p))_{p\in P}$ of this Markov decision problem, which satisfies

$$\max_{a\in A}\left(\bar{R}(p,a) + \gamma \bar{V}^*(\alpha(p,a))\right) = \bar{V}^*(p) \quad \text{for all } p \in P. \tag{33}$$

With $\bar{V}^*$, we define the set $\mathcal{B} \subset ([0,\infty[^m)^{P\times A}$ of non-negative coefficients

$$\mathcal{B} = \{((\beta_j(p,a))_{j=1}^m)_{(p,a)\in P\times A} \,:\, \sum_{j=1}^m \beta_j(p,a)\psi_j(z) \le \bar{V}^*(p) \text{ for } (p,a) \in P\times A,\, z \in \Xi\} \tag{34}$$

which is compact (due to (28)) and convex. The next lemma shows that $\mathcal{B}$ is invariant under I.

Lemma 2. *The mapping I satisfies $I\mathcal{B} \subset \mathcal{B}$.*

Proof: For each $\beta \in \mathcal{B}$, the non-negativity $I\beta \in ([0,\infty[^m)^{P\times A}$ of coefficients is ensured by (31) and by the definition of $(\breve{T}^b(a))_{a\in A}$. Hence, it remains to show

$$\sum_{j=1}^m (I\beta)_j(p,a)\psi_j(z) \le \bar{V}^*(p) \text{ for each } z \in \Xi \text{ and } p \in P, a \in A. \tag{35}$$

Given $\beta \in \mathcal{B}$ and $a \in A$, by definition (34) it holds

$$\sum_{j=1}^m \beta_j(p,a)\psi_j(z) \le \bar{V}^*(p), \text{ for each } z \in \Xi \text{ and } p \in P,\, a \in A. \tag{36}$$

Further, the estimate by (**??**)

$$\sum_{j=1}^m (I\beta)_j(p,a)\psi_j(z) = \breve{T}^b(a)W^\beta(p,z) \le W^\beta(p,z) \quad \text{for } p \in P,\, a \in A \text{ and } z \in \Xi. \tag{37}$$

Now maximize $W^\beta(p,z)$ on the right hand side of (37) in $z \in \Xi$ to obtain the assertion (35):

$$\begin{aligned}
\sum_{j=1}^m (I\beta)_j(p,a)\psi_j(z) &\le W^\beta(p,z) = \max_{a\in A}\left(R(p,z,a) + \gamma\sum_{j=1}^m \beta_j(\alpha(p,a),a)\psi_j(z)\right) \\
&\le \max_{a\in A}\max_{z\in\Xi} R(p,z,a) + \gamma \max_{a\in A}\max_{z\in\Xi}\sum_{j=1}^m \beta_j(\alpha(p,a),a)\psi_j(z) \\
&\le \max_{a\in A}\left(\bar{R}(p,a) + \gamma\bar{V}^*(\alpha(p,a))\right) = \bar{V}^*(p).
\end{aligned}$$

Here, the last inequality is due to (36) and to (26).

Now, we gathered all auxiliary results and enter the proof of the Proposition 3.

Proof: The results of the above lemmata show that the continuous function I maps a compact and convex set $\mathcal{B}$ to itself. Thus, the existence of the fixed point

$$\beta^* \in \mathcal{B} \text{ with } I\beta^* = \beta^*$$

follows from the Brower's fixed point theorem. Given such fixed point β^*, we define

$$W^{\beta^*}(p,z) = \max_{a\in A}\left(R(p,z,a) + \gamma\sum_{j=1}^{m}\beta_j^*(\alpha(p,a),a)\psi_j(z)\right) \quad \text{for all } (p,z)\in P\times\Xi.$$

Form $\beta^* = I\beta^*$ we conclude that for all $(p,z)\in P\times\Xi$ holds

$$\begin{aligned} W^{\beta^*}(p,z) &= \max_{a\in A}\left(R(p,z,a) + \gamma\sum_{j=1}^{m}(I\beta^*)_j(\alpha(p,a),a)\psi_j(z)\right) \\ &= \max_{a\in A}\left(R(p,z,a) + \gamma\breve{T}^b(a)W^{\beta^*}(\alpha(p,a),\cdot)(z)\right), \end{aligned}$$

where the last equality is obtained from (31). In other words, $\breve{V}^* := W^{\beta^*}$ solves (24), as required in the assertion of the Proposition 3.

Remark Brower's fixed point theorem yields the existence of a fixed point (24). Although convergence of value iteration can not be concluded under present assumptions, we believe that the above results are instructive, showing a reasonable way to address *practical solution* of an important class high-dimensional control problems by least-squares Monte-Carlo methods. Our study shows that a regularization of the control problem yields desirable results. We suggest to change the rewards to positive values and to mimic the most crucial properties of the true Markov transition, such as positivity $T_a f \geq 0$ if $f \geq 0$ (using projections on a cone of positive functions), contractility $\sup_{z\in Z} T_a f(z) \leq \sup_{z\in Z} f(z)$ (obtained due to (**??**)), and convexity preservation, valid in our context (by construction, as projection on cone of convex functions). Due to this regularization, the authors have observed a stable convergence of value iterations in all numerical experiments, studied so far.

VII. A NUMERICAL ILLUSTRATION WITH ADAPTIVE FEATURE SELECTION

We now illustrate the properties of our methodology with the analysis of the storage facility management problem introduced in Section V. We assume that the commodity price follows an auto-regression of order two, which we realize as a Markovian evolution on the state space

$Z = \mathbb{R}^2$. This assumption on the dimension of the model is made for illustration purposes only. The generalization to higher order auto-regressions is straight-forward.

In order to improve computation efficiency, we implement an adaptive feature selection according to which the dictionary of feature basis functions is chosen at each iteration depending upon the function on which the transition operator is computed. Our method is applicable in the case of linear state price price dynamics. We now describe the basic idea.

Let us assume that the d-dimensional Markov processes $(Z_t)_{t\in\mathbb{N}}$ giving the evolution of the commodity price, and possibly of other factors of the model, has the dynamics of a *linear state space model* recursively defined by $Z_{t+1} = AZ_t + W_{t+1}$ for $t \in \mathbb{N}$, starting from a given $Z_0 = z_0 \in Z \subset \mathbb{R}^d$, where $A \in \mathbb{M}_{d,d}$ is a $d \times d$ matrix and $(W_t)_{t\in\mathbb{N}}$ is a white noise sequence of $\mathbb{R}^d$-valued independent identically distributed mean-zero random variables. Such a model captures not only plain auto regressive models, but also time series models with seasonal and trend components. Note that in the present situation, the Markov transition operator $\tau = \tau_a$ does not depend upon the action $a \in \mathcal{A}$, so the assumption (17) of Theorem **??** is satisfied. This operator is given by

$$\begin{aligned}[\tau\varphi](z) &= \mathbb{E}(\varphi(Z_{t+1}) \,|\, Z_t = z) = \mathbb{E}(\varphi(AZ_t + W_{t+1}) \,|\, Z_t = z) \\ &= \mathbb{E}(\varphi(Az + W_{t+1})) = \int_{\mathbb{R}} \varphi(Az + w) P_W(dw) \end{aligned} \tag{38}$$

where P_W denotes the common distribution of the noise vectors W_t. From this expression we infer the convexity preservation property and we derive a reasonable candidate for the cone of feature basis functions. Indeed, for each $w \in \mathbb{R}^d$, the function $z \mapsto \varphi(Az + w)$ can be regarded as a *modification* of the original function $z \mapsto \varphi(z)$ obtained by composition $\varphi \circ l_w$ of φ with the affine linear function $l_w : Z \ni z \mapsto Az + w \in Z$ determined by its intercept $w \in \mathbb{R}^d$. Note that if φ is convex, then each modification $\varphi \circ l_w$ is also convex. Having in mind an approximation of the integral in (38) by a limit of sums, $\tau\varphi$ can be approximated by elements from the cone spanned by $\{\varphi \circ l_w \,:\, l_w : Z \to Z,\ w \in \mathbb{R}^d\}$. If φ is convex, this cone consists of convex functions, which yields also convexity of $\tau\varphi$. In order to mimic (38), we suggest to select our feature dictionary from the cone spanned by the functions $(\psi_j = \varphi \circ l_{w_j})_{j=1}^m$ for an appropriate set $(w_j)_{j=1}^n \in \mathbb{R}^d$ of intercepts which should be chosen in order to represent typical realizations of W_1.

In our numerical implementation, we assume that the commodity price follows an univariate

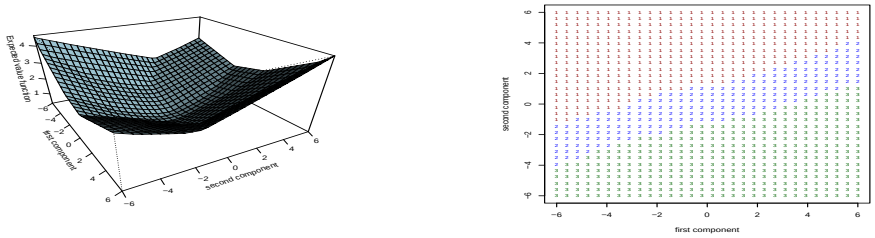

Fig. 1. For $p = 2$ the expected value function $z \mapsto \breve{T}\breve{V}^*(p, z)$ (left) and the optimal policy $z \mapsto \breve{\pi}^*(p, z)$ (right).

auto-regressive model of order $d = 2$ with coefficients 0.3 and 0.65, driven by a unit variance noise. We realize such a scalar process as the second component $(Z_t^{(2)})_{t\in\mathbb{N}}$ of the linear state space process $(Z_t)_{t\in\mathbb{N}}$ defined by the recursion

$$\begin{bmatrix} Z_{t+1}^{(1)} \\ Z_{t+1}^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0.65 & 0.3 \end{bmatrix} \begin{bmatrix} Z_t^{(1)} \\ Z_t^{(2)} \end{bmatrix} + \begin{bmatrix} 0 \\ W_{t+1}^{(2)} \end{bmatrix},$$

where $(W_t^{(2)})_{t\in\mathbb{N}}$ are independent identically standard normally distributed random variables. To demonstrate the performance of our method we take a very small Monte Carlo sample $\mathcal{S} = (Z_t(\omega), Z_{t+1}(\omega))_{t=1}^{500}$ constructed from of a single path $(Z_t(\omega))_{t=1}^{501}$ consisting of 501 observations only. We use the reward function (16) with transaction cost $c = 1$, and set the discount factor to $\lambda = 0.8$. Furthermore, we implemented the adaptive dictionary choice described above with with five intercepts $(w_j)_{j=1}^5$ equidistantly distributed from $w_1 = -1$ to $w_5 = 1$. With these data, we observed a stable convergence of the iterated value iterations even with only $N = 20$ steps. For the position $p = 2$ (storage facility half-full), the expected value function $z \mapsto \breve{V}^*(p, z)$ and the optimal policy $\breve{\pi}^*(p, z) = \max_{a\in\mathcal{A}} \Big(R(p, z, a) + \gamma\breve{T}(a)\breve{V}^*(\alpha(p, a), \cdot)(z)\Big)$ are plotted in Figure 1. To show the performance of this strategy, we applied the decision rule $\breve{\pi}^*$ to another test sample $(Z_t(\omega'))_{t=1}^{1000}$. Figure (2) shows the joint evolution of the price sample path $(Z^{(2)_t}(\omega'))_{t=1}^{1000}$, the resulting optimal storage levels $(\breve{p}_t^*)_{t=1}^{1000}$ (stepwise constant, oscillating between 1=empty and

3=full), and the running reward $(\sum_{u=1}^{t} R_u(\breve{p}_u^*, \breve{a}_u^*, Z_u))_{t=1}^{1000}$ (stepwise constant, increasing). Notice that in order to be able to include the three time evolutions on the same plot, we scaled all the processes down to the interval $[0, 1]$.

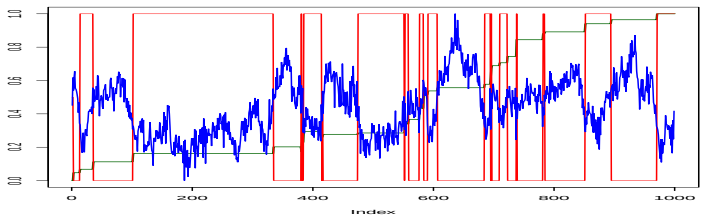

Fig. 2. Joint evolution of sample path, control policy and running cumulated reward.

VIII. Conclusion

Infinite horizon control problems with discounted reward are popular in applications. However, relevant real-world problems are still notoriously challenging even in this setting, due high-dimensionality of the state space. We demonstrate that the classical least squares Monte Carlo method can be improved when the value functions are convex. Although the convexity assumption may appear restrictive, it is satisfied in a large class of models. Furthermore, Monte-Carlo methods are generically less affected by the curse of dimensionality, we believe that our method can still be used when other techniques fail by reaching their computational limits.

References

[1] N. Bäuerle and U. Rieder. *Markov Decision Processes with Applications to Finance*. Springer, Heidelberg, 2011.

[2] A. Belomestny, N. Kolodko and J. Schoenmakers. A variance reduction technique for american option pricing. *SIAM J. Control Optim.*, 48(5):3562–3588, 2004.

[3] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2005.

[4] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[5] M. Broadie and J. B. Detemple. Option pricing: Valuation models and applications. *Management Science*, 50:1145–1177, 2004.

[6] R. Carmona and N. Touzi. Optimal multiple stopping and valuation of swing options. *Math. Finance*, 18:239Ű268, 2008.

[7] J. F. Carriere. Valuation of the early-exercise price for options using simulations and nonparametric regression. *Insurance: Mathematics and Economics*, 19:19–30, 1996.

[8] A. R. Choudhury, A. King, S Kumar, and Y. Sabharwal. Optimizations in financial engineering: The least-squares monte carlo method of longstaff and schwartz. Technical report, 22nd IEEE International Symposium on Parallel and Distributed Processing (IPDPS), 2008.

[9] E. Clément, D. Lamberton, and P. Protter. An analysis of the longstaff-schwartz algorithm for American option pricing. *Finance and Stochastics*, 6(4):449–471, 2002.

[10] D. Egloff. Monte carlo algorithms for optimal stopping and statistical learning. *Appl. Probab.*, 15:1396–1432, 2005.

[11] D. Egloff, M. Kohler, and N. Todorovic. A dynamic look-ahead monte carlo algorithm. *Appl. Appl. Probab.*, 17:1138–1171, 2007.

[12] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, 1996.

[13] E. A. Feinberg and A. Shwartz. *Handbook of Markov Decision Processes*. Kluwer Academic, 2002.

[14] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.

[15] M. Hauskrecht. Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

[16] Lasserre J.B. Hernández-Lerma, O. *Discrete-time Markov Control Processes*. Springer, New York, 1996.

[17] N. B. Liberati and E. Platen. On the efficiency of simplified weak Taylor schemes for Monte Carlo simulation in finance. *Lecture Notes in Computer Science*, 3039:771–778, 2004.

[18] F. Longstaff and E. Schwartz. Valuing American options by simulation: a simple least-squares approach. *Review of Financial Studies*, (14):113–147, 2001.

[19] N. Moreni. A variance reduction technique for american option pricing. *Physica A-statistical mechanichs and its applications*, 338:2292–295, 2004.

[20] D. Ormoneit and P. Glynn. Kernel-based reinforcement learning. *Machine Learning*, 49:161–178, 2002.

[21] D. Ormoneit and P. Glynn. Kernel-based reinforcement learning in average-cost problems. *IEEE Transactions in Automatic Control*, 47:1624–1636, 2002.

[22] Kaelbling L. P., Littman M. L., and A. R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[23] W. B. Powell. *Approximate dynamic programming: Solving the curses of dimensionality*. Wiley, 2007.

[24] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.

[25] L. Stentoft. Convergence of the least squares Monte Carlo approach to American option valuation. *Management Science*, 50(9):576–611, 2004.

[26] J. N. Tsitsiklis and B. Van Roy. Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, 2001.

[27] J.N. Tsitsiklis and B. Van Roy. Optimal stopping of Markov processes: Hilbert space, theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840–1851, 1999.