Enhanced Exploration Least-Squares Methods for Optimal Stopping Problems

Ali Forootani, *Member, IEEE*, Massimo Tipaldi, Raffaele Iervolino, *Senior Member, IEEE*, and Subhrakanti Dey, *Senior Member, IEEE*

***Abstract*—This letter presents an Approximate Dynamic Programming (ADP) least-squares based approach for solving optimal stopping problems with a large state space. By extending some previous work in the area of optimal stopping problems, it provides a framework for their formulation and resolution. The proposed method uses a combined on/off policy exploration mechanism, where states are generated by means of state transition probability distributions different from the ones dictated by the underlying Markov decision processes. The contraction mapping property of the associated projected Bellman operator is analysed as well as the convergence of the resulting algorithm.**

***Index Terms*—Optimal stopping problem, Markov decision process, approximate dynamic programming.**

I. Introduction

OPTIMAL stopping problems can be regarded as a specific class of Markov Decision Processes (MDPs), wherein the associated system dynamics evolve according to the underlying state transition probability distributions until a specific termination action is used. The theory of optimal stopping problems concerns the selection of a proper time at which to perform such termination action with the aim of minimising expected total costs [1]–[3]. Optimal stopping problems can be found in various fields, such as statistics and economics [1], [4]. The scientific literature presents optimal stopping problems in both discrete and continuous time, see [3], [5], [6] for their application to event-triggered control problems.

Optimal stopping problems can be written in the form of a Bellman equation, and therefore can be solved by using Dynamic Programming (DP) based approaches [1], [2]. In principle, exact DP algorithms (e.g., the Value Iteration) can be used to solve the related stochastic optimization problem and calculate the optimal policy, which gives a mapping between states and optimal decisions (or control actions) over the whole time horizon [7]. However, it is well known that such exact DP algorithms suffer from the so-called *curse of dimensionality*, which is due to the state space explosion of real-world applications, and from the availability of the system model [7], [8]. Such issues also occur for the case of optimal stopping problems [2]. It is then natural to consider Approximate Dynamic Programming (ADP) based approaches in order to compute a suitable approximation to the optimal cost function (the latter defined as the expected cumulative cost when starting from a specific state, and then applying the optimal policy [7]). As for the modeling of the system at hand, we assume either to have an estimate of the underlying state transition probability distributions or a representative system simulator able to generate states according to them [7].

This letter proposes an ADP least-squares projection based approach, and as such, it relies on cost function approximation (via a more compact parametric representation) in conjunction with Monte Carlo simulations (the latter to solve the optimality condition associated to the projected Bellman equation [7], [9]). As known from the literature, ADP least-squares projection approaches are convergent when states are sampled with the frequencies natural to the underlying Markov decision process, which means sampling according to its invariant probability distribution [10]–[12]. As for optimal stopping problems, a linear function approximation and an on-policy sampling mechanism (i.e., system states explored via the natural frequencies of the underlying Markov process) are used in [1], [2]. Sampling according to the natural frequencies of the system can have some drawbacks for the cost function approximation, since it can bias the Monte Carlo simulations, i.e., states that are less likely to occur by applying such invariant probability distributions can be disregarded [7], [12].

This letter proposes a novel combined on/off policy exploration based algorithm to solve optimal stopping problems. After outlining some background in Section II, the paper addresses the following aspects, which constitutes its main contributions: (i) the extension of the results presented in [1], [2] to solve optimal stopping problems via the above-mentioned on/off policy mechanism, see Sections III and IV; (ii) the analysis of the contraction mapping property of the optimal stopping problem Bellman operator w.r.t. some steady-state probability distributions different from the ones associated to the system natural frequencies, see Section IV; (iii) the combined on/off policy exploration based algorithm

Manuscript received December 31, 2020; revised March 3, 2021; accepted March 24, 2021. Date of publication March 30, 2021; date of current version June 24, 2021. Recommended by Senior Editor V. Ugrinovskii. *(Corresponding author: Ali Forootani.)*

Ali Forootani and Subhrakanti Dey are with Hamilton Institute, Maynooth University, W23 F2K8 Kildare, Ireland (e-mail: ali.forootani@mu.ie; subhra.dey@mu.ie).

Massimo Tipaldi is with the Department of Engineering, University of Sannio, 82100 Benevento, Italy (e-mail: mtipaldi@unisannio.it).

Raffaele Iervolino is with the Department of Electrical Engineering and Information Technology, University of Naples, 80125 Napoli, Italy (e-mail: rafierv@unina.it).

Digital Object Identifier 10.1109/LCSYS.2021.3069708

(derived from the Least-Squares Policy Evaluation (LSPE), see [13], [14]) along with its convergence analysis, see Section V. An illustrative example is provided in Section VI, where a comparison of the presented approach with the on-policy LSPE algorithm [1], [7] is shown. Section VII concludes the paper.

II. Optimal Stopping Problems

An optimal stopping problem can be defined as follows [2]. Let us consider a Markov chain with the state space $\Omega = \{1, \ldots, n\}$ and transition probabilities p_{ij}, where i, j are two generic states belonging to Ω. $P \in \mathbb{R}^{n\times n}$ is defined as the state transition probability matrix with the associated elements p_{ij}. For any given state i, two actions (or decisions) are foreseen: either to terminate and incur a positive cost $G(i)$, or to continue and incur a positive cost $g(i)$ (in a vector form, we have $G \in \mathbb{R}^n_+$ and $g \in \mathbb{R}^n_+$, where each element is denoted by $G(i)$ and $g(i)$, respectively). We consider the situation that there is no control affecting the actual transition from state i to j.

We define a stopping time τ to be a discrete random variable with non negative integer values and with respect to the natural filtration of the given stochastic process [1]. To each stopping time τ, we associate the following cost function $J^\tau(i) : \Omega \to \mathbb{R}_+$

$$J^\tau(i) = E\left[\sum_{t=0}^{\tau-1} \alpha^t g(i_t) + \alpha^\tau G(i_\tau) | i_0 = i\right], \tag{1}$$

where i_t denotes the state of the process at time t and $0 < \alpha < 1$ is the discount factor. The optimal stopping time τ^* satisfies the expression $J^{\tau^*}(i) = \min_\tau [J^\tau(i)]$, and is given by (see [1, Th. 1])

$$\tau^* = \min\{t : G(i_t) \leq J^*(i_t)\}. \tag{2}$$

In order to solve such minimization problem, we introduce the DP operator $\mathcal{T}$ for optimal stopping problems [1]

$$\mathcal{T}J = \min\{G, g + \alpha P J\}, \tag{3}$$

where $J \in \mathbb{R}^n_+$ is the vector with components $J(i)$, with $J(i) : \Omega \to \mathbb{R}_+$ being a generic cost function. By assuming that the Markov chain is irreducible, i.e., P has a unique steady-state probability vector $\xi = (\xi_1, \ldots, \xi_n)$ with positive components, the operator $\mathcal{T}$ becomes a contraction mapping with modulus α w.r.t. the weighted Euclidean norm $\|\cdot\|_\xi$, see Lemma 2 reported in [1]. In particular, for any pair of cost functions $J_1, J_2 \in \mathbb{R}^n_+$, we have

$$\left\|\mathcal{T}J_1 - \mathcal{T}J_2\right\|_\xi \leq \alpha \left\|J_1 - J_2\right\|_\xi. \tag{4}$$

Being a contraction mapping with modulus α, $\mathcal{T}$ has a unique fixed point J^*, which satisfies the Bellman equation $J^* = \mathcal{T}J^*$ [7]. Such fixed point J^* is equal to J^{τ^*} [1]. Hereinafter, for simplicity, we use J^* (and $J^*(i)$) instead of J^{τ^*} (and $J^{\tau^*}(i)$). The optimal cost function $J^*(i)$ is given by [1]

$$J^*(i) = \begin{cases} G(i), & \text{if} \quad G(i) \leq g(i) + \alpha \sum_{j=1}^n p_{ij} J^*(j) \\ g(i) + \alpha \sum_{j=1}^n p_{ij} J^*(j), & \text{otherwise.} \end{cases} \tag{5}$$

Note that $J^*(i) = G(i)$ for the states where the stopping decision is taken. In order to address the curse of dimensionality, we can replace the cost function $J(i)$ with a parametric approximation architecture by using a restricted set of selected m feature functions. The choice of such parametric approximation architecture is significant for the success of the approximation approach. One possibility is to use the linear feature-based approximation (for the case of the continuation decision),

$$\tilde{J}(i) = \sum_{l=1}^m r_l \phi_l(i), \tag{6}$$

where r_l is the $l-$th component of a parameter vector $r \in \mathbb{R}^m$, which has to be computed, and $\phi_l(i)$, $l \in \{1, \ldots, m\}$, are the given feature functions [7]. Thus, for each state i, the approximate value $\tilde{J}(i)$ can be written as the inner product $\phi(i)'r$, where $\phi(i) = [\phi_1(i), \ldots, \phi_m(i)]'$. In matrix form, it is $\tilde{J} = \Phi r$, where $\Phi \in \mathbb{R}^{n\times m}$ is the feature matrix whose rows are set to $\phi(i)'$. In this letter, we assume the columns of the feature matrix are linearly independent and $m \ll n$. In view of the next sections, the weighted least-squares projection operator is defined as $\Pi = \Phi(\Phi'\Xi\Phi)^{-1}\Phi'\Xi$, where $\Xi \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the components of the steady-state probability vector ξ on its diagonal. From the contraction property of $\mathcal{T}$ and the non-expansiveness of the projection operator Π, by implication the projected DP operator $\Pi\mathcal{T}$ is a contraction mapping of modulus α w.r.t. the weighted Euclidean norm $\|\cdot\|_\xi$ [1].

III. A Projected Enhanced Exploration Method for Markov Chains

This section analyses the contraction mapping property of the following two DP operators

$$\mathcal{F}J = g + \alpha P J, \tag{7}$$

$$\bar{\mathcal{F}}J = g + \alpha \bar{P} J, \tag{8}$$

where $\bar{P} \in \mathbb{R}^{n\times n}$ is an irreducible state transition probability matrix different from P with the unique steady-state distribution $\bar{\xi} = (\bar{\xi}_1, \ldots, \bar{\xi}_n)$. Note that $\mathcal{F}$ coincides with the optimal stopping problem DP operator when the decision to continue is taken and addresses its underlying Markov chain, while $\bar{\mathcal{F}}$ can be viewed as an enhanced exploration DP operator for such a Markov chain.

Exploration plays a relevant role in ADP based approaches [7], [15], [16], when a suitable approximation to the optimal cost function is computed by using an approximation architecture for cost functions in conjunction with Monte Carlo simulations. As for the latter, simulation trajectories can be generated according to the underlying irreducible Markov chain to preserve the contraction property of $\mathcal{F}$ w.r.t the weighted Euclidean norm $\|\cdot\|_\xi$ [7], [17]. In the literature this method is often called *on policy* approach.

However, sampling according to the probability distribution ξ can bias the Monte Carlo simulation by disregarding states that are less likely to occur under such probability distribution. For instance, this can affect the policy improvement step of the approximate Policy Iteration algorithm [7]. The same issue can occur when solving approximately optimal stopping problems. As shown later, the DP operator $\mathcal{F}$ determines the sampling mechanism for optimal stopping problems, since a new simulation trajectory has to be initialised whenever the stopping action is taken. Such sampling mechanism can be enriched by also using the matrix $\bar{P}$.

In this letter, we adopt an off-policy based mechanism for exploration enhancement. The irreducible matrix $\bar{P}$ is defined as [7]

$$\bar{P} = (I - \mathcal{B})P + \mathcal{B}Q, \tag{9}$$

where $\mathcal{B}$ is a diagonal matrix with diagonal components $\beta_i \in [0, 1)$ and Q is another transition probability matrix.[1] In this framework, at state i, the next state j is generated with probability $1 - \beta_i$ according to transition probabilities p_{ij}, and with probability β_i according to transition probabilities q_{ij}. Note that a computer program can be used to generate state transitions according to $\bar{P}$. To implement such off-policy based mechanism, pairs (i, j) with $q_{ij} > 0$ need not correspond to physically plausible transitions [7].

Since the underlying Markov chain is assumed to be irreducible, the DP operator $\mathcal{F}$ is a contraction mapping with modulus α w.r.t. the weighted Euclidean norm $\|\cdot\|_\xi$, and thus the corresponding Bellman equation has a fixed point [7]. The same applies to the DP operator $\bar{\mathcal{F}}$, but with the contraction mapping referred to $\|\cdot\|_{\bar{\xi}}$. Finding the fixed point of equations (7) and (8) for large scale DP problems can be impractical due to the curse of dimensionality. This calls for approximation, e.g., using the linear parametric approximation architecture along with Monte Carlo simulations.

As for the former, it is important to employ the weighted least-squares projection operator. Besides the projection operator Π (linked to the on-policy based exploration mechanism, thus determined by the matrix P), we introduce the following projection operator (which accounts for the off-policy based exploration mechanism, thus determined by the matrix $\bar{P}$)

$$\bar{\Pi} = \Phi(\Phi'\bar{\Xi}\Phi)^{-1}\Phi'\bar{\Xi}, \tag{10}$$

where $\bar{\Xi} \in \mathbb{R}^{n\times n}$ is the diagonal matrix, having $\bar{\xi}_i$ $(i = 1, \ldots, n)$ along its diagonal. Thanks to the non-expansiveness of the weighted least-squares projection operators Π and $\bar{\Pi}$, the projected DP operators $\Pi\mathcal{F}$ and $\bar{\Pi}\bar{\mathcal{F}}$ are contraction mappings w.r.t $\|\cdot\|_\xi$ and $\|\cdot\|_{\bar{\xi}}$, respectively. Thus, they have a fixed point satisfying the corresponding projected Bellman equation [7], i.e., there exists a unique parameter vector r^* which satisfies the following (on policy) projected Bellman equation

$$\Phi r = \Pi\mathcal{F}(\Phi r). \tag{11}$$

Likewise, there exists a unique parameter vector $\bar{r}^*$ satisfying the following projected Bellman equation

$$\Phi r = \bar{\Pi}\bar{\mathcal{F}}(\Phi r). \tag{12}$$

Hereafter, we analyse the contraction mapping property of the DP operator $\bar{\Pi}\mathcal{F}$ to guarantee the existence of the unique fixed point of the associated (off policy) projected equation

$$\Phi r = \bar{\Pi}\mathcal{F}(\Phi r), \tag{13}$$

where $\bar{\Pi}$ is still the projection w.r.t. the norm $\|\cdot\|_{\bar{\xi}}$ corresponding to the steady-state distribution $\bar{\xi}$ of $\bar{P}$. In Section IV, the results of this analysis will be applied to the optimal stopping problem to provide sufficient conditions for the contraction property of its DP operator $\mathcal{T}$, when using the off-policy exploration mechanism. By extending the result of [10, Lemma 1], the following lemma shows that the projected operator $\bar{\Pi}\mathcal{F}$ is a contraction mapping w.r.t. the norm $\|\cdot\|_{\bar{\xi}}$. As a result, the projected equation (13) has a unique fixed point. The following proof also provides the interval of values for β_i to guarantee the contraction property of the operators $\mathcal{F}$ and $\bar{\Pi}\mathcal{F}$.

Lemma 1: Assume that $\bar{P}$ is an irreducible state transition probability matrix and that $\bar{\xi}$ is its unique steady-state probability vector with positive components. Then, $\mathcal{F}$ and $\bar{\Pi}\mathcal{F}$ are contraction mappings with respect to $\|\cdot\|_{\bar{\xi}}$, and the associated modulus of contraction is at most equal to $\bar{\alpha}$, where

$$\bar{\alpha} = \alpha/\sqrt{1-\beta}, \quad \text{with} \quad \beta = \max_{i=1,\ldots,n} \beta_i. \tag{14}$$

Proof: For any $J \in \mathbb{R}^n_+$, we have

$$\begin{aligned}\left\|\alpha PJ\right\|^2_{\bar{\xi}} &= \sum_{i=1}^n \bar{\xi}_i \left(\sum_{j=1}^n \alpha p_{ij} J(j)\right)^2 = \alpha^2 \sum_{i=1}^n \bar{\xi}_i \left(\sum_{j=1}^n p_{ij} J(j)\right)^2 \\ &\le \alpha^2 \sum_{i=1}^n \bar{\xi}_i \sum_{j=1}^n p_{ij} J^2(j) \le \alpha^2 \sum_{i=1}^n \bar{\xi}_i \sum_{j=1}^n \frac{\bar{p}_{ij}}{1-\beta_i} J^2(j) \\ &\le \frac{\alpha^2}{1-\beta} \sum_{j=1}^n \sum_{i=1}^n \bar{\xi}_i \bar{p}_{ij} J^2(j) = \bar{\alpha}^2 \sum_{j=1}^n \bar{\xi}_j J^2(j) = \bar{\alpha}^2 \left\|J\right\|^2_{\bar{\xi}}, \end{aligned} \tag{15}$$

where the first inequality follows from the convexity of the quadratic function, the second inequality follows from the fact $(1 - \beta_i)p_{ij} \le \bar{p}_{ij}$ (see (9)), and the step before the last equality follows from the property of the steady-state probabilities $\sum_{i=1}^n \bar{\xi}_i \bar{p}_{ij} = \bar{\xi}_j$. By using the non-expansiveness of $\bar{\Pi}$, the definition $\mathcal{F}J = g + \alpha PJ$ and (15), we have $\|\bar{\Pi}\mathcal{F}J_1 - \bar{\Pi}\mathcal{F}J_2\|_{\bar{\xi}} \le \|\mathcal{F}J_1 - \mathcal{F}J_2\|_{\bar{\xi}} = \|\alpha P(J_1 - J_2)\|_{\bar{\xi}} \le \bar{\alpha}\|J_1 - J_2\|_{\bar{\xi}}$, for any cost function $J_1, J_2 \in \mathbb{R}^n_+$. Hence both $\mathcal{F}$ and $\bar{\Pi}\mathcal{F}$ are contractions of modulus $\bar{\alpha}$ with respect to $\|\cdot\|_{\bar{\xi}}$. ■

Note that the relation $\beta < 1-\alpha^2$ has to be fulfilled for $\bar{\Pi}\mathcal{F}$ to be a contraction mapping w.r.t. $0 < \bar{\alpha} < 1$.

Since $\bar{\Pi}\mathcal{F}$ is a contraction mapping w.r.t. $\bar{\xi}$ and by exploiting the assumption that Φ has full rank m, it is easy to prove the following lemma, which extends the [7, Proposition 6.3.1].

Lemma 2: Let the assumptions of Lemma 1 hold, and let the matrix Φ be of full rank m. Then, we have

$$\left\|J_{\mathcal{F}} - \Phi r^{*,\bar{\xi}}\right\|_{\bar{\xi}} \le 1/(\sqrt{1-\bar{\alpha}^2}) \left\|J_{\mathcal{F}} - \bar{\Pi}J_{\mathcal{F}}\right\|_{\bar{\xi}},$$

where $r^{*,\bar{\xi}}$ is the unique solution of the projected Bellman equation (13) and $J_{\mathcal{F}}$ is the fixed point of the mapping $\mathcal{F}$.

Since Φ is full rank, the unique fixed point of the operator $\bar{\Pi}\mathcal{F}$ can be represented by a unique parameter vector $r^{*,\bar{\xi}}$. By using (6), the high-dimensional original cost function can be represented via the lower-dimensional parameter vector r (with $m \ll n$). Since all the considered projected DP are contraction mappings, the associated Bellman equation can be solved by means of the Projected Value Iteration (PVI) algorithm [7]. As for the projected DP operator $\Pi\mathcal{F}$ (11), the parameter vector r^* can be iteratively computed by

$$r^*_{k+1} = \arg\min_{r\in\mathbb{R}^m} \sum_{i=1}^n \xi_i \left(\phi(i)'r - \sum_{j=1}^n g(i) + \alpha p_{ij}\phi(j)'r^*_k\right)^2, \tag{16}$$

where r^*_{k+1} is the approximate value of r^* computed at the iteration k of the PVI algorithm (note that the algorithm starts from $k = 0$ with an initial guess r_0). Since $\Pi\mathcal{F}$ is a

[1] In our case, $\bar{P}$ is irreducible if P is.

contraction mapping, it follows that the sequence $\{r_k^*\}$ converges to r^*. Solving each iteration step (16) of the PVI algorithm implies massive calculation since we have to compute the low-dimensional vector r_k^* by using high-dimensional calculations (note that two nested summations over n have to be performed in (16)). To solve this issue, we can use a Monte Carlo simulation-based implementation of the PVI iteration (16). The resulting algorithm is called Least-Squares Policy Evaluation (LSPE), see [7]. More specifically, as for the projected DP operator $\Pi\mathcal{F}$, the LSPE algorithm implies the generation of an infinitely long trajectory $(i_0, i_1, \dots)$ according to the state transition probability matrix P and the update of r_{k+1}^* after each state transition (i_k, i_{k+1}). In particular, the simulation-based PVI iteration step can be expressed as follows

$$r_{k+1}^* = \arg\min_{r\in\mathbb{R}^m} \sum_{t=0}^{k} \big(\phi(i_t)'r - g(i_t) - \alpha\phi(i_{t+1})'r_k^*\big)^2. \quad (17)$$

By setting the gradient of (17) to 0, we have

$$r_{k+1}^* = \left(\sum_{t=0}^{k} \phi(i_t)\phi(i_t)'\right)^{-1} \left(\sum_{t=0}^{k} \phi(i_t)\big(g(i_t) + \alpha\phi(i_{t+1})'r_k^*\big)\right). \quad (18)$$

The convergence analysis of the LSPE method can be found in [14], [17]. In Section V, an LSPE based algorithm for optimal stopping problems is presented, where samples are collected by applying a combined on/off policy exploration approach. Before describing this algorithm, it is necessary to investigate the contraction mapping property of the optimal stopping problem operator $\mathcal{T}$ (and its projected operator $\Pi\mathcal{T}$) w.r.t. the enhanced exploration steady-state distribution $\bar{\xi}$.

IV. The Enhanced Exploration DP Operator for Optimal Stopping Problems

In this section, we analyse the contraction mapping property of the operator $\mathcal{T}$ w.r.t. the steady-state probability distribution $\bar{\xi}$ of the enhanced exploration transition probability matrix $\bar{P}$ given by (9) in order to verify the existence of the fixed point of the enhanced exploration projected Bellman equation

$$\Phi r = \bar{\Pi}\mathcal{T}(\Phi r). \quad (19)$$

By using the results provided by the previous section, the following theorem can be proved.

Theorem 1: Assume that the enhanced exploration matrix $\bar{P}$ is an irreducible state transition probability matrix. Then, the optimal stopping problem DP operator $\mathcal{T}$ is a contraction mapping with respect to $\|\cdot\|_{\bar{\xi}}$, with the modulus of contraction at most equal to $\bar{\alpha}$ (see (14)).

Proof: For any two cost functions $J_1, J_2 \in \mathbb{R}_+^n$ evaluated at a given state i and by applying the DP operator $\mathcal{T}$ (3), we can have the following three cases: 1) If we decide to continue for both cost functions, it is

$$\big|(\mathcal{T}J_1)(i) - (\mathcal{T}J_2)(i)\big| = \alpha\Big|\sum_{j=1}^{n} p_{ij}(J_1(j) - J_2(j))\Big| = \leq \alpha\sum_{j=1}^{n} p_{ij}\big|J_1(j) - J_2(j)\big|.$$

2) If we decide to stop for both cost functions, it is

$$\big|(\mathcal{T}J_1)(i) - (\mathcal{T}J_2)(i)\big| = G(i) - G(i) \leq \alpha\sum_{j=1}^{n} p_{ij}\big|J_1(j) - J_2(j)\big|.$$

3) If we decide to stop for J_1 and to continue for J_2 (or vice-versa), it is

$$\big|(\mathcal{T}J_1)(i) - (\mathcal{T}J_2)(i)\big| = \Big|G(i) - \left(g(i) + \alpha\sum_{j=1}^{n} p_{ij}J_2(j)\right)\Big|$$
$$= \leq \alpha\Big|\sum_{j=1}^{n} p_{ij}J_1(j) - \sum_{j=1}^{n} p_{ij}J_2(j)\Big| \leq \alpha\sum_{j=1}^{n} p_{ij}\big|J_1(j) - J_2(j)\big|,$$

where, as for the first inequality, we have exploited the definition of the optimal stopping problem operator (3) to replace $G(i)$ with $g(i) + \alpha\sum_{j=1}^{n} p_{ij}J_1(j)$.

Thus, for all the three cases, we can write in vector notation $|\mathcal{T}J_1 - \mathcal{T}J_2| \leq \alpha P|J_1 - J_2|$, where $|J_1 - J_2|$ denotes a vector whose components are the absolute values of the components of $J_1 - J_2$. By using the fact that, given any two vectors $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^n$ with components $X(i)$ and $Y(i)$ such that $|X(i)| \leq |Y(i)|$, $\forall i = 1, \dots, n$, it is $\|X\| \leq \|Y\|$,[2] then we can write

$$\Big\|\mathcal{T}J_1 - \mathcal{T}J_2\Big\|_{\bar{\xi}} \leq \alpha\Big\|P(J_1 - J_2)\Big\|_{\bar{\xi}} \leq \bar{\alpha}\Big\|J_1 - J_2\Big\|_{\bar{\xi}},$$

where the last inequality follows from the relation $\alpha\|PJ\|_{\bar{\xi}} \leq \bar{\alpha}\|J\|_{\bar{\xi}}$, see the proof of Lemma 1. ■

From Theorem 1 and Lemma 2, the next result follows.

Corollary 1: Since $\mathcal{T}$ is a contraction with respect to $\|\cdot\|_{\bar{\xi}}$ (with the modulus of contraction at most equal to $\bar{\alpha}$) and the projection operator $\bar{\Pi}$ is non-expansive, the mapping $\bar{\Pi}\mathcal{T}$ is also a contraction with respect to $\|\cdot\|_{\bar{\xi}}$ with the modulus of contraction $\bar{\alpha}$. Moreover, we have

$$\Big\|J^* - \Phi r^{*,\bar{\xi}}\Big\|_{\bar{\xi}} \leq 1/(\sqrt{1-\bar{\alpha}^2})\Big\|J^* - \bar{\Pi}J^*\Big\|_{\bar{\xi}}, \quad (20)$$

where J^* is given by (5) and $\Phi r^{*,\bar{\xi}} = \bar{\Pi}\mathcal{T}(\Phi r^{*,\bar{\xi}})$.

The relation (20) provides a bound for the expected error between the optimal cost function J^* and its feature-based approximate value $\tilde{J} = \Phi r^{*,\bar{\xi}}$. For simplicity we denote with the same notation $r^{*,\bar{\xi}}$ the fixed points of both the exploration enhanced projected Bellman operators $\bar{\Pi}\mathcal{T}$ and $\bar{\Pi}\mathcal{F}$ (such fixed points are generally different since the operator $\mathcal{F}$ does not apply the termination action). Moreover, no approximation onto the feature subspace is needed in case $J^*(i) = G(i)$, see (5). Since $J^*(i)$ cannot be computed for optimal stopping problems with a large state space, we simply set the approximate cost function to $G(i)$ whenever $\tilde{J}(i) = \phi(i)'r^{*,\bar{\xi}} \geq G(i)$.

V. The LSPE Based Algorithm for Solving Optimal Stopping Problems

This section presents the LSPE based algorithm for addressing optimal stopping problems. This algorithm solves approximately the enhanced exploration projected Bellman equation (19) by generating multi-trajectories Monte Carlo simulations with a combined on/off policy sampling mechanism.

[2]This property holds for the any weighted Euclidean norm $\|\cdot\|_{\bar{\xi}}$, see [1].

Being $\bar{\Pi}\mathcal{T}$ a contraction mapping, the associated Bellman equation (19) can be solved by means of the PVI algorithm [7]. So as in Section III, the fixed point $r^{*,\bar{\xi}}$ can be computed by applying iteratively

$$r_{k+1}^{*,\bar{\xi}} = \arg\min_{r\in\mathbb{R}^m} \sum_{i=1}^{n} \bar{\xi}_i \times \left(\phi(i)'r + -g(i) - \alpha \sum_{j=1}^{n} p_{ij} \min\left\{G(j), \phi(j)' r_k^{*,\bar{\xi}}\right\}\right)^2, \quad (21)$$

where $r_{k+1}^{*,\bar{\xi}}$ is the approximate value of $r^{*,\bar{\xi}}$ computed at the iteration k of the PVI algorithm. Note that (21) only shows the case when a decision to continue is taken. In case of stopping for a state i, its related term in the summation becomes $\phi(i)'r - G(i)$.

Thanks to the contraction mapping of the operator $\bar{\Pi}\mathcal{T}$, the sequence $\{r_k^{*,\bar{\xi}}\}$ converges to $r^{*,\bar{\xi}}$. However, solving each iteration step (21) of the PVI algorithm can imply massive calculations. To address this issue, we can apply an LSPE based algorithm to the (21). In particular, instead of producing a long trajectory, we can generate a sequence of states $\{i_0, i_1, \dots\}$ according to the enhanced exploration steady-state probability distribution $\bar{\xi}$, and a sequence of transitions $\{(i_0, j_0), (i_1, j_1), \dots\}$ with probabilities p_{ij}. As a result, we have a multi-trajectory algorithm, with the length of each trajectory set to 1. From the sequence of state transitions (i_t, j_t), we can formulate the following least square minimization problem at each step k

$$r_{k+1}^{*,\bar{\xi}} = \arg\min_{r\in\mathbb{R}^m} \sum_{t=0}^{k} \left(\phi(i_t)'r - g(i_t) - \alpha \min\left\{G(j_t), \phi(j_t)' r_k^{*,\bar{\xi}}\right\}\right)^2,$$

whose solution is

$$r_{k+1}^{*,\bar{\xi}} = \frac{\sum_{t=0}^{k} \phi(i_t)\left(g(i_t) + \alpha \min\left\{G(j_t), \phi(j_t)' r_k^{*,\bar{\xi}}\right\}\right)}{\sum_{t=0}^{k} \phi(i_t)\phi(i_t)'}. \quad (22)$$

Note that (21) and (22) use the same notation for the computed parameter vector. The update of the parameter vector $r_k^{*,\bar{\xi}}$ is skipped when the stopping decision has to be taken for the transition (i_k, j_k) since the approximate cost function for the sampled state i_k can be set to $G(i_k)$. In other words, the transition (i_k, j_k) is discarded, and the algorithm selects for the current iteration k another state sampled according to the enhanced exploration probability distribution $\bar{\xi}$.

A. Convergence Analysis

The convergence analysis of the proposed algorithm has been derived from [2] by extending its results to the off-policy sampling mechanism case. At each step k, the iteration formula (22) can be replaced by

$$r_{k+1}^{*,\bar{\xi}} = \frac{\sum_{i=1}^{n} \hat{\bar{\xi}}_{k,i}\phi(i)\left(g(i) + \alpha \hat{p}_{k,ij} \min\left\{G(j), \phi(j)' r_k^{*,\bar{\xi}}\right\}\right)}{\sum_{i=1}^{n} \hat{\bar{\xi}}_{k,i}\, \phi(i)\phi(i)'}, \quad (23)$$

where the following two elements are used

$$\hat{\bar{\xi}}_{k,i} = \frac{\sum_{t=0}^{k} \delta(i_t = i)}{k+1}, \quad \hat{p}_{k,ij} = \frac{\sum_{t=0}^{k} \delta(i_t = i, j_t = j)}{\sum_{t=0}^{k} \delta(i_t = i)}. \quad (24)$$

In a more compact form, we have $\Phi r_{k+1}^{*,\bar{\xi}} = \hat{\bar{\Pi}}_k \hat{\mathcal{T}}_k(\Phi r_k^{*,\bar{\xi}})$, where the mappings $\hat{\bar{\Pi}}_k$ and $\hat{\mathcal{T}}_k$ are simulation-based approximations to $\bar{\Pi}$ and $\mathcal{T}$, and can expressed as follows

$$\hat{\bar{\Pi}}_k = \Phi(\Phi'\hat{\bar{\Xi}}_k\Phi)^{-1}\Phi'\hat{\bar{\Xi}}_k, \quad \hat{\bar{\Xi}}_k = \mathrm{diag}(\dots, \hat{\bar{\xi}}_{k,i}, \dots), \quad (25)$$

$$\hat{\mathcal{T}}_k J = g + \alpha \hat{P}_k \min\{G, J\}, \quad \forall J \in \mathbb{R}_+^n. \quad (26)$$

Thanks to the ergodicity of the involved Markov chain, as we proceed in the simulation, we have $\hat{\bar{\xi}}_{k,i} \to \bar{\xi}_i$, $\hat{\bar{\xi}}_k \to \bar{\xi}$, $\hat{p}_{k,ij} \to p_{ij}$, see [13], [14]. Moreover, we have the following facts [2]:

1) For any $\epsilon > 0$ and a sample trajectory with converging sequences $\hat{\bar{\xi}}_k$, there exists a time $\bar{k}$ such that for all $k > \bar{k}$

$$1/(1+\epsilon) \le \hat{\bar{\xi}}_i/\bar{\xi}_i \le (1+\epsilon), \ \forall i. \quad (27)$$

2) When (27) holds, then for any $J \in \mathbb{R}_+^n$, for all $k > \bar{k}$ it is

$$\|J\|_{\bar{\xi}} \le (1+\epsilon)\|J\|_{\hat{\bar{\xi}}_k}. \quad (28)$$

By extending the proof of [2, Lemma 2] to the off-policy sampling mechanism, it is easy to prove the following lemma.

Lemma 3: Let $\hat{\bar{\alpha}} \in (\bar{\alpha}, 1)$. Then, with probability 1, $\hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k$ is a $\|\cdot\|_{\bar{\xi}}$ contraction with modulus $\hat{\bar{\alpha}}$ for all k sufficiently large.

In particular, by letting ϵ be such that $(1+\epsilon)^2\bar{\alpha} < \hat{\bar{\alpha}} < 1$, it is possible to see that $\hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k$ is a $\|\cdot\|_{\bar{\xi}}$ contraction mapping with modulus $\hat{\bar{\alpha}}$ for all k sufficiently large.

Theorem 2: The parameter vector $r_k^{*,\bar{\xi}}$ computed by (22) converges to $r^{*,\bar{\xi}}$ with probability 1.

Proof: We select $\bar{k}$ such that, for all $k \ge \bar{k}$, the contraction mapping property of Lemma 3 applies. For all such k, it is

$$\begin{aligned}\left\|\Phi r_{k+1}^{*,\bar{\xi}} - \Phi r^{*,\bar{\xi}}\right\|_{\bar{\xi}} &= \left\|\hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k(\Phi r_k^{*,\bar{\xi}}) - \bar{\Pi}\mathcal{T}(\Phi r^{*,\bar{\xi}})\right\|_{\bar{\xi}} \\ &= \left\|\hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k(\Phi r_k^{*,\bar{\xi}}) + \hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k(\Phi r^{*,\bar{\xi}}) - \hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k(\Phi r^{*,\bar{\xi}}) - \bar{\Pi}\mathcal{T}(\Phi r^{*,\bar{\xi}})\right\|_{\bar{\xi}} \\ &\le \hat{\bar{\alpha}}\left\|\Phi r_k^{*,\bar{\xi}} - \Phi r^{*,\bar{\xi}}\right\|_{\bar{\xi}} + \epsilon_k,\end{aligned}$$

where $\epsilon_k = \left\|\hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k(\Phi r^{*,\bar{\xi}}) - \bar{\Pi}\mathcal{T}(\Phi r^{*,\bar{\xi}})\right\|_{\bar{\xi}}$. Since $\left\|\hat{\bar{\Pi}}_k\hat{\mathcal{T}}_k(\Phi r^{*,\bar{\xi}}) - \bar{\Pi}\mathcal{T}(\Phi r^{*,\bar{\xi}})\right\|_{\bar{\xi}} \to 0$, we have $\epsilon_k \to 0$. Moreover, since $\hat{\bar{\alpha}} < 1$, we have that $\Phi r_k^{*,\bar{\xi}} \to \Phi r^{*,\bar{\xi}}$ (or equivalently, $r_k^{*,\bar{\xi}} \to r^{*,\bar{\xi}}$). ■

VI. An Illustrative Example

In this section, the resource allocation problem formulation presented in [15], [16] is tailored to model an urban parking lot management system as an optimal stopping problem. Let us consider a car factory storing its products into a parking lot before sending them to its car dealer group. The parking lot manager selects a price c_l, $l = 1, \dots, h$, among h possible choices. The car factory can hold the allocated spot to the necessary extent with the proposed price, and only one spot can be allocated at each time slot. The time slot duration is chosen so that, for each price c_l, at most one allocated spot can be released. By denoting with i^l the number of cars in the parking lot associated to the price c_l, we can define the generic state of the MDP associated to the parking lot as $i = [i^1, i^2, \dots, i^h]'$.

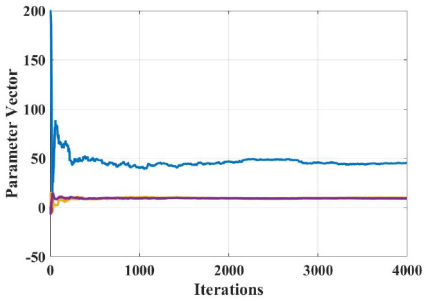

Fig. 1. Evolution of $r_k^{*,\bar{\xi}}$ components in on/off policy LSPE.

The car factory can either continue to use and request for a spot and incur a positive cost $g(i) = \sum_{l=1}^{h} c_l i^l$ or terminate and incur a positive cost $G(i)$. We assume the termination cost is the same for all states i, i.e., $G(i) = 120\$$. As for the linear feature-based approximation, we define the $m = 1 + h$ feature functions as $\phi_1(i) = 1, \phi_l(i) = i^{l-1}, \quad l = 2, \ldots, m$. With a parking capacity $N = 10$, a number of prices $h = 3$, the cardinality of the state space is $n = 286$ (see [16] for modeling details). We set the price vector $c = [0.9\ 1\ 1.1]'$, the discount factor $\alpha = 0.95$ and $\beta = 0.00353$ to satisfy the condition $\beta < 1 - \alpha^2$. By using the proposed on/off policy LSPE based algorithm, we can compute the approximate parameter vector $r_k^{*,\bar{\xi}}$. The convergence is reached after 4000 iterations (see Fig. 1), and the computed approximate parameter vector value is $r_{4000}^{*,\bar{\xi}} = [44.9\ 8.9\ 9.8\ 9.1]'$. This value can be used to generate the stopping time $\tilde{\tau}_{off} = \min\{t | G(i_t) \leq (\Phi r_{4000}^{*,\bar{\xi}})(i_t)\}$. In particular, we set 1000 experiments of 100 time slots in length to simulate our parking lot management system and computed $\tilde{\tau}_{off}$ associated to each experiment. By averaging such stopping times, we computed the approximate optimal stopping time $\tilde{\tau}_{off}^* = 24$.

To make a comparison, we applied the on-policy LSPE algorithm in [1] to solve the optimal stopping problem associated to the same parking lot management system. The components of the computed approximate parameter vector r_k^* converged after 10^5 iterations. The parameter vector is $r_{10^5}^* = [82\ 8.5\ 3\ 1.6]'$. We computed $\tilde{\tau}_{on}$ like the on/off policy algorithm case with the same experimental setup. By averaging the generated stopping times, we computed the approximate optimal stopping time $\tilde{\tau}_{on}^* = 20$. The proposed on/off policy LSPE algorithm manages to explore better the system state when computing the approximate parameter vector. Indeed, the approximate parameter vector $r_{10^5}^*$ computed by the on-policy LSPE is more biased towards the first component of the feature subspace and the values associated to the third and forth components are smaller. This affects the computed stopping times $\tilde{\tau}_{on}^*$ and $\tilde{\tau}_{off}^*$. As a result, the car factory manages to utilize better the parking lot resources thanks to the exploratory enhancement property of the proposed approach.

VII. Conclusion

This letter has proposed a combined on/off policy exploration based algorithm to solve optimal stopping problems with a large state space. By extending some results available in the literature, it has first provided a framework to formulate and solve optimal stopping problems. It has been proven that the associated Bellman operator is also a contraction mapping with respect to the steady-state probability distributions of enhanced exploration irreducible transition probability matrices. Thanks to this, it is possible to perform Monte Carlo simulations with a combined on/off policy sampling approach to solve approximately the corresponding projected Bellman equation, i.e., to compute an approximate parameter vector in the feature subspace. Finally, the convergence of the proposed on/off policy exploration based algorithm has been analysed. The main assumption for our approach to work is the availability of good features, which is challenging in general. As for future work, we plan to apply the proposed framework to event-triggered control problems for discrete-time systems and to extend it by integrating deep neural networks to learn proper features from real life optimal stopping training data.

References

[1] 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 Trans. Autom. Control*, vol. 44, no. 10, pp. 1840–1851, Oct. 1999.

[2] H. Yu and D. P. Bertsekas, "*Q*-learning algorithms for optimal stopping based on least squares," in *Proc. Eur. Control Conf.*, 2007, pp. 2368–2375.

[3] A. Goldenshluger and L. Mirkin, "On minimum-variance event-triggered control," *IEEE Control Syst. Lett.*, vol. 1, no. 1, pp. 32–37, Jul. 2017.

[4] M. D. Marcozzi, "On the approximation of optimal stopping problems with application to financial mathematics," *SIAM J. Sci. Comput.*, vol. 22, no. 5, pp. 1865–1884, 2001.

[5] M. T. Andrén, "Using radial basis functions to approximate the LQG-optimal event-based sampling policy," in *Proc. 18th Eur. Control Conf. (ECC)*, 2019, pp. 2832–2838.

[6] D. J. Antunes and M. H. I. Balaghi, "Consistent event-triggered control for discrete-time linear systems with partial state information," *IEEE Control Syst. Lett.*, vol. 4, no. 1, pp. 181–186, Jan. 2020.

[7] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. II (4rd Ed.)*. Belmont, MA, USA: Athena Sci., 2012.

[8] Y. Liu, E. K. P. Chong, A. Pezeshki, and Z. Zhang, "A general framework for bounding approximate dynamic programming schemes," *IEEE Control Syst. Lett.*, vol. 5, no. 2, pp. 463–468, Apr. 2021.

[9] M. Geist and O. Pietquin, "Algorithmic survey of parametric value function approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 845–867, Jun. 2013.

[10] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. Autom. Control*, vol. 42, no. 5, pp. 674–690, May 1997.

[11] H. Yu, "Least square temporal difference methods: An analysis under general conditions," *SIAM J. Control Optim.*, vol. 50, no. 6, pp. 3310–3343, 2012.

[12] A. Forootani, M. Tipaldi, M. G. Zarch, D. Liuzza, and L. Glielmo, "A least-squares temporal difference based method for solving resource allocation problems," *IFAC J. Syst. Control*, vol. 13, pp. 1–15, Sep. 2020.

[13] A. Nedic and D. P. Bertsekas, "Least squares policy evaluation algorithms with linear function approximation," *Discr. Event Dyn. Syst.*, vol. 13, pp. 79–110, Jan. 2003.

[14] H. Yu and D. P. Bertsekas, "Convergence results for some temporal difference methods based on least squares," *IEEE Trans. Autom. Control*, vol. 54, no. 7, pp. 1515–1531, Jul. 2009.

[15] A. Forootani, R. Iervolino, M. Tipaldi, and J. Neilson, "Approximate dynamic programming for stochastic resource allocation problems," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 4, pp. 975–990, Jul. 2020.

[16] A. Forootani, R. Iervolino, and M. Tipaldi, "Applying unweighted least-squares based techniques to stochastic dynamic programming: Theory and application," *IET Control Theory Appl.*, vol. 13, no. 15, pp. 2387–2398, 2019.

[17] D. P. Bertsekas, "Temporal difference methods for general projected equations," *IEEE Trans. Autom. Control*, vol. 56, no. 9, pp. 2128–2139, Sep. 2011.