Off-Policy Temporal Difference Learning for Perturbed Markov Decision Processes

Forootani, Ali; Iervolino, Raffaele; Tipaldi, Massimo; Khosravi, Mohammad

DOI
10.1109/LCSYS.2025.3547629

Publication date
2025

Document Version
Final published version

Published in
IEEE Control Systems Letters

Off-Policy Temporal Difference Learning for Perturbed Markov Decision Processes

Ali Forootani, *Senior Member, IEEE*, Raffaele Iervolino, *Senior Member, IEEE*, Massimo Tipaldi, and Mohammad Khosravi, *Member, IEEE*

***Abstract*—Dynamic Programming suffers from the curse of dimensionality due to large state and action spaces, a challenge further compounded by uncertainties in the environment. To mitigate these issue, we explore an off-policy based Temporal Difference Approximate Dynamic Programming approach that preserves contraction mapping when projecting the problem into a subspace of selected features, accounting for the probability distribution of the perturbed transition probability matrix. We further demonstrate how this Approximate Dynamic Programming approach can be implemented as a particular variant of the Temporal Difference learning algorithm, adapted for handling perturbations. To validate our theoretical findings, we provide a numerical example using a Markov Decision Process corresponding to a resource allocation problem.**

***Index Terms*—Reinforcement learning, Markov decision processes, temporal difference learning, perturbed probability transition matrix.**

I. Introduction

IN LARGE-SCALE Markov Decision Processes (MDPs), Dynamic Programming (DP) faces the curse of dimensionality as state and action spaces grow exponentially, making computation infeasible [1]. To address this, Approximate Dynamic Programming (ADP) methods leverage function approximations and simulations to reduce complexity [2], [3]. ADP techniques such as Temporal Difference (TD) further enhance scalability [4], [5]. On-policy and off-policy TD methods estimate cost functions using three main approaches: (i) gradient-based solutions, which apply stochastic gradient descent to minimize prediction error incrementally [6]; (ii) least-squares minimization, which reduces the least-squares error between estimated and true returns [1], [7], [8]; and (iii) probabilistic approaches, which use Bayesian methods to model uncertainty in value estimates [9], [10]. Off-policy TD, a key ADP technique, estimates cost functions from data generated by policies different from the one being optimized [1], enabling greater exploration and improved policies [11]. Exploration strategies also play a crucial role in Reinforcement Learning (RL), as seen in Deep Q-Learning [12], while Proximal Policy Optimization (PPO) [13] improves the exploration-exploitation trade-off using a clipped surrogate objective, making it robust for both continuous and discrete action spaces. MDP uses simulations to approximate optimal cost functions for effective policies, but model inaccuracies and dynamic environmental changes can degrade performance. Understanding how deviations between the model and reality impact policy effectiveness is crucial, particularly in the context of *perturbed transition probability matrices*, leading to *perturbed* MDPs. Perturbation analysis in RL examines how environmental or model parameter changes affect policy performance and value functions.

This letter proves that sufficiently small perturbation in the environment (transition probability matrices) will lead to restricted bound between the optimal cost-to-go function and non-optimal one. In particular, we examine how transition probability matrix perturbations can be account as off-policy TD learning, specifically Q-learning, by investigating the impact of weighted combinations of exploration and optimal actions on efficient learning. In addition, we analyze the variations of cost functions in the context of policy perturbations and provide the corresponding upper bounds for discounted MDP s. More precisely, let the transition matrices $\mathcal{P}^*$ and $\mathcal{Q}$ correspond to the optimal policy π^* and an exploratory policy $\pi_{\mathcal{Q}}$, respectively. Also, let the perturbed transition matrix $\bar{\mathcal{P}}$ be defined as a weighted combination of these two matrices, i.e., for $\bar{\mathcal{P}}$, we have

$$\bar{\mathcal{P}} = (I - \mathcal{A})\mathcal{Q} + \mathcal{A}\mathcal{P}^*, \tag{1}$$

where I denotes the identity matrix and $\mathcal{A}$ is a diagonal matrix with $\alpha \in]0, 1[$ on its diagonal, i.e., $\mathcal{A} = \alpha I$ and α is the discounted factor, specifying the trade-off between the use of the optimal policy π^* and the exploratory policy $\pi_{\mathcal{Q}}$. Our framework explicitly links the perturbation of transition matrices to the exploration-exploitation trade-off through a

Received 2 December 2024; revised 30 January 2025; accepted 22 February 2025. Date of publication 4 March 2025; date of current version 20 March 2025. Recommended by Senior Editor C. Briat. *(Corresponding author: Ali Forootani.)*

Ali Forootani is with the Hamilton Institute, Helmholtz Center for Environmental Research-UFZ, 04318 Leipzig, Germany (e-mail: aliforootani@ieee.org).

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

Massimo Tipaldi is with the Department of Electrical and Information Engineering, Polytechnic University of Bari, 70126 Bari, Italy (e-mail: massimo.tipaldi@poliba.it).

Mohammad Khosravi is with the Delft Center for Systems and Control, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: mohammad.khosravi@tudelft.nl).

Digital Object Identifier 10.1109/LCSYS.2025.3547629

convex combination of $\mathcal{Q}$ and $\mathcal{P}^*$. While similar trade-offs exist in algorithms like PPO and TRPO [13], [14], the novelty lies in our matrix-based formulation and the explicit derivation of convergence guarantees under perturbations, which is less explored in MDP research. We characterize the performance resulting from the utilization of suboptimal policies associated with $\bar{\mathcal{P}}$. In particular, we discuss the projection on the subspace of specifically selected features based on such perturbations and the convergence properties for off-policy TD methods. Furthermore, following the work of [15], we ensure that the perturbed cost function remains bounded over an infinite time horizon, provided that for any pairs of stationary policy (including $\pi = \pi^*$) it is $\|\mathcal{P} - \bar{\mathcal{P}}\|_\infty \leq 1 - \alpha$.

Non-stationary MDP s have been widely studied in RL literature, particularly in scenarios where transition dynamics or rewards evolve over time. Prior works [16], [17], [18], [19], [20] address arbitrary or adversarial transitions, often requiring explicit tracking mechanisms. In contrast, our formulation introduces a structured perturbation model, representing the perturbed transition matrix as a convex combination of optimal and exploratory transitions (1). This principled approach enables smooth interpolation between policies, aligning with exploration-exploitation trade-offs in optimistic RL methods [21], [22], [23]. Unlike prior works relying on tabular settings or parametric drift [24], [25], our framework generalizes to off-policy learning scenarios by explicitly linking policy perturbations to transition matrix perturbations, offering a novel perspective on policy optimization under controlled non-stationarity.

This letter is organized as follows. Section II provides some preliminaries on DP and outlines the cost function approximation problem. The on-policy ADP approach is discussed in Section III. Section IV analyzes the perturbed transition probability matrices and their impacts on cost functions. Section V explores the connection between perturbed transition probability matrices and Q-learning. In Section VI, the proposed off-policy TD approach is assessed using a resource allocation problem. Section VII concludes this letter.

II. Preliminaries

Let us consider an infinite-horizon DP problem for an MDP. Such a problem often is solved by employing Bellman's principle of optimality recursively backwards in time [11]. In this regard, a shorthand notation is introduced for Bellman operator as $\mathcal{F}^* : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{X}|}$ for any cost function vector $J \in \mathbb{R}^{|\mathcal{X}|}$ which can be expressed as[1]:

$$\left(\mathcal{F}^* J\right)(x) = \min_{u \in \mathcal{U}} \left[\mathcal{R}(x) + \alpha \sum_{x' \in \mathcal{X}} \mathcal{P}_{xx'}(u) J\left(x'\right) \right], \tag{2}$$

where $\mathcal{X}$ is the MDP state space with cardinality $|\mathcal{X}|$, with x and x' being its two generic elements. Moreover, $u \in \mathcal{U}$ is a generic element of finite set of control actions $\mathcal{U}$, $\mathcal{R} \in \mathbb{R}^{|\mathcal{X}|}$ the vector of instant cost with element $\mathcal{R}(x)$,[2] $\mathcal{P} : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \to [0, 1]$ the state transition probability matrix, with generic element $\mathcal{P}_{xx'}(u)$, and $\alpha \in]0, 1[$ the discount factor. In matrix form, (2) simplifies to: $\mathcal{F}^* J = \mathcal{R} + \alpha \mathcal{P}^* J$. As noted in [11], this shorthand notation is common for simplifying complex expressions. Applying the sequence of optimal decision functions $\{\mu^*, \mu^*, \dots\}$, where $\mu^* : \mathcal{X} \to \mathcal{U}$, as time horizon goes to infinity results in optimal stationary policy π^* with the associate steady state probability distribution $\epsilon^* \in \mathbb{R}^{|\mathcal{X}|}$(we denote by ϵ^*_x an element of this vector corresponding to the state x). The Bellman equation for the optimal cost function satisfies: $J^* = \mathcal{R} + \alpha \mathcal{P}^* J^*$, where $J^* \in \mathbb{R}^{|\mathcal{X}|}$ is the vector of optimal costs, and $\mathcal{P}^* \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is the optimal transition probability matrix. Any non-optimal stationary policy is shown by π with its associated shorthand notation $\mathcal{F}_\pi$ and associated cost function J_π [1]. The cost function $J^*(x)$ is usually approximated by a parametric model $\tilde{J}^*(x, r)$, where $r \in \mathbb{R}^\psi$ is the parameter vector to be optimized, and $\psi \in \mathbb{N}^+$ is a selected number of features. This approximation, $\tilde{J}^* : \mathcal{X} \times \mathbb{R}^\psi \to \mathbb{R}^{|\mathcal{X}|}$, uses a low-dimensional linear function $J^* \approx \Phi r^*$, where $\Phi \in \mathbb{R}^{|\mathcal{X}| \times \psi}$ is a feature matrix with linearly independent columns which we show by $\phi(x)$ its row corresponding to state x, and $r^* \in \mathbb{R}^\psi$ is the parameter vector to compute. Substituting the approximation into the Bellman equation yields the projected equation: $\Phi r^* = \Pi(\mathcal{R} + \alpha \mathcal{P}^* \Phi r^*)$, where $\Pi = \Phi(\Phi^\top \Theta^* \Phi)^{-1} \Phi^\top \Theta^*$ is the projection operator, and Θ^* is the diagonal matrix with ϵ^* on the diagonal and zero elsewhere [26]. Here, Φr^* approximates J^*, making the equation solvable. This approach relies on three key assumptions for any stationary policy π (including the optimal π^*) and related cost function J_π: (i) the corresponding finite Markov chain is regular, with the stochastic matrix $\mathcal{P}$ having a unique steady-state probability distribution ϵ, with strictly positive elements ϵ_x, i.e., $\epsilon_x > 0, \forall x \in \mathcal{X}$; (ii) the feature matrix Φ is full rank with rank ψ; (iii) the feature matrix Φ captures the key characteristics of the cost function J_π, allowing Φr to closely approximate it. In this context, the dimensionality of the feature space (ψ) plays a crucial role in determining computational feasibility.

III. ADP Methods Based on On-Policy Framework

Before delving into the main results of this letter, we present introductory material on the on-policy ADP approach to prepare for the discussion of off-policy methods. In a more general framework, TD methods can be classified into on-policy and off-policy approaches [27]. In on-policy learning, the goal is to learn the cost function $\tilde{J}_\pi(x)$, which represents the approximated expected long-term discounted costs from state x when following a target policy π. The learning process occurs while the agent actively follows the same policy π. On-policy methods are particularly effective in ensuring good convergence, especially when using linear function approximations. To find the optimal parameter vector r^* in the on-policy case for the optimal policy π^*, we solve the following optimization problem:

$$r^* = \arg\min_r \left\| \Phi r - \left(\mathcal{R} + \alpha \mathcal{P}^* \Phi r\right) \right\|^2_{\epsilon^*}, \tag{3}$$

[1]This approach can be applied similarly for value function optimization by replacing min function by max function. See the example in Section VI.

[2]We omit the dependence of $\mathcal{R}$ on a specific control action.

which minimizes the square of the weighted Euclidean norm[3,4] of the error between Φr and the projected Bellman equation, where the weight ϵ^* is the steady-state probability distribution associated to π^*. By differentiating (3) and setting the gradient to zero, we obtain: $\Phi^\top\Theta^*(\Phi r^* - \mathcal{R} - \alpha\mathcal{P}^*\Phi r^*) = 0$, where Θ^* is a diagonal matrix with distribution ϵ^* along its diagonal. Solving this equation gives: $r^* = (\Phi^\top\Theta^*\Phi - \alpha\Phi^\top\Theta^*\mathcal{P}^*\Phi)^{-1}(\Phi^\top\Theta^*)\mathcal{R}$. In this formulation, the term $(\Phi^T\Theta^*\Phi - \alpha\Phi^\top\Theta^*\mathcal{P}^*\Phi)$ represents the matrix involved in solving for r^*, incorporating both the feature matrix and the transition dynamics. This approach provides a way to approximate the optimal cost function in infinite-horizon problems using a lower-dimensional representation, leveraging the structure of the Bellman equation and the projection framework to find an effective solution. Defining $\mathcal{Z}^* = \Phi^\top\Theta^*(I-\alpha\mathcal{P}^*)\Phi$, and $d^* = \Phi^\top\Theta^*\mathcal{R}$, then, the optimal parameter vector is: $r^* = \mathcal{Z}^{*-1}d^*$. For large state spaces, computing r^* directly is infeasible, so iterative methods are employed. The update rule can be derived as follows: $r^*_{k+1} = \arg\min_r \|\Phi r - (\mathcal{R} + \alpha\mathcal{P}^*\Phi r_k)\|^2_{\epsilon^*}$. Taking the gradient and setting it to zero: $(\Phi^\top\Theta^*\Phi)r^*_{k+1} = \Phi^\top\Theta^*(\mathcal{R}+\alpha\mathcal{P}^*\Phi r^*_k)$, and rearranging and simplifying:

$$r^*_{k+1} = r^*_k - \left(\Phi^\top\Theta^*\Phi\right)^{-1}\left(\mathcal{Z}^* r^*_k - d^*\right). \tag{4}$$

This iterative update formula is related to Least Squares Temporal Difference (LSTD) methods [11]. Note that (4) is applicable to any generic target policy π when considering the associated steady state probability distribution ϵ (whose elements will be the diagonal elements of a diagonal matrix Θ). In off-policy learning, the objective remains the same: to estimate the cost function $J_\pi(x)$ for the target policy π. However, the actions taken during the learning process follow a different behavior policy $\bar{\pi}$. Even though the agent follows $\bar{\pi}$ during learning, the focus is still on accurately estimating the cost function for the target policy π (see [28] and reference therein). The mechanism to compute iteratively the parameter vector is analogous to the one in (4) and is based on perturbation analysis, which is discussed in the next section.

IV. Perturbation Analysis of Stochastic Matrices and Cost Function Bounds

In order to apply off-policy methods, which require perturbation of the transition probability matrices (for exploration purposes), it is important to understand how perturbations in transition dynamics affect cost functions. This section establishes bounds for the differences between cost functions from perturbed and unperturbed dynamics, using properties of stochastic matrices and positive definiteness. These results provide insights into the resilience of MDP s under uncertain transition probabilities, aiding ADP methods in developing robust decision strategies. Using the results reported in [29, Lemma 6.3.1], the following Lemma 1 and Remark 1 guarantee the convergence of on-policy TD approaches and accordingly recursive iteration (4). This result will be extended to show the convergence property in the case of off-policy TD approach as well.

Lemma 1: For any stochastic matrix $\mathcal{P}$ corresponding to irreducible and regular Markov chain with associated stationary probability distribution ϵ, whose elements are arranged along the diagonal of a diagonal matrix Θ, the matrix $\Theta(I - \alpha\mathcal{P})$ is positive definite, where $\alpha \in]0, 1[$.

Proof: To prove that $\Theta(I - \alpha\mathcal{P})$ is positive definite, we want to show that for any non-zero vector $J \in \mathbb{R}^{|\mathcal{X}|}$: $J^\top\Theta(I - \alpha\mathcal{P})J > 0$. We have: $J^\top\Theta J = \sum_x \epsilon_x J_x^2$ and: $J^\top\Theta\mathcal{P}J = \sum_x \epsilon_x J_x \sum_{x'} \mathcal{P}_{xx'}J_{x'}$, where the term $\sum_{x'} \mathcal{P}_{xx'}J_{x'}$ is the expected value of $J_{x'}$ given x, denoted by $\mathbb{E}[J_{x'}|x]$. Since ϵ is a stationary distribution we can rewrite: $\sum_x \epsilon_x J_x \sum_{x'} \mathcal{P}_{xx'}J_{x'} = \mathbb{E}_\epsilon[J_x \cdot \mathbb{E}[J_{x'}|x]]$. Applying the Cauchy-Schwarz inequality to the expectation $\mathbb{E}_\epsilon[J_x \cdot \mathbb{E}[J_{x'}|x]]$, we get: $\mathbb{E}_\epsilon[J_x \cdot \mathbb{E}[J_{x'}|x]] \leq \sqrt{\mathbb{E}_\epsilon[J_x^2]} \cdot \sqrt{\mathbb{E}_\epsilon[(\mathbb{E}[J_{x'}|x])^2]}$. By Jensen's inequality for convex functions, we also have: $\mathbb{E}[(\mathbb{E}[J_{x'}|x])^2] \leq \mathbb{E}[J_{x'}^2]$. Thus, $\sqrt{\mathbb{E}_\epsilon[(\mathbb{E}[J_{x'}|x])^2]} \leq \sqrt{\mathbb{E}_\epsilon[J_{x'}^2]}$. Therefore, $\mathbb{E}_\epsilon[J_x \cdot \mathbb{E}[J_{x'}|x]] \leq \mathbb{E}_\epsilon[J_x^2]$. Or, equivalently, $\sum_x \epsilon_x J_x \sum_{x'} \mathcal{P}_{xx'}J_{x'} \leq \sum_x \epsilon_x J_x^2$. Being $\alpha \in]0, 1[$, this gives us the desired result: $\sum_x \epsilon_x J_x^2 > \alpha \sum_x \sum_{x'} \epsilon_x \mathcal{P}_{xx'} J_x J_{x'}$. ∎

Remark 1: From the results of Lemma 1 and full rank assumption of matrix Φ we know $\mathcal{Z} = \Phi^\top\Theta(I - \alpha\mathcal{P})\Phi$ is positive definite. Moreover $(\Phi^\top\Theta\Phi)$ is symmetric positive definite, therefore invertible (see [30, Lemma 5.5]). From [29, Proposition 6.3.3 and Lemma 6.3.2] the recursive iteration (4) directly converges to the solution of the projected equation since the matrix $I - (\Phi^\top\Theta\Phi)^{-1}\mathcal{Z}$ has eigenvalues strictly within the unit circle (see also [30, Th. 5.6]).

Remark 1 implies that necessary and sufficient condition for the convergence of any on-policy TD algorithm depends on the positive definiteness of matrix $\mathcal{Z}$, and hence of $\Theta(I-\alpha\mathcal{P})$. This concept can be extended to the case of off-policy TD algorithm. In the next Lemma, we investigate such property for the case of perturbed transition probability matrix $\bar{\mathcal{P}}$.

Lemma 2: For any stochastic matrix $\mathcal{P}$ corresponding to an irreducible and regular Markov chain, consider a stationary probability distribution $\bar{\epsilon}$, whose elements are arranged along the diagonal of a diagonal matrix $\bar{\Theta}$, associated with the perturbed matrix $\bar{\mathcal{P}} = (I - \mathcal{A})\mathcal{Q} + \mathcal{A}\mathcal{P}$, being $\mathcal{A}$ a diagonal matrix with $\alpha \in]0, 1[$ on its diagonal, and $\mathcal{Q}$ another stochastic matrix. Then, the matrix $\bar{\Theta}(I - \alpha\mathcal{P})$ is positive definite.

Proof: We need to show that for any non-zero vector $J \in \mathbb{R}^{|\mathcal{X}|}$, $J^\top\bar{\Theta}(I - \alpha\mathcal{P})J > 0$. To this end, consider the matrix $\bar{\Theta}(I - \alpha\bar{\mathcal{P}})$, and apply the results of Lemma 1. Following the same proof steps, we can state that:

$$\sum_x \bar{\epsilon}_x J_x^2 \geq \sum_{x,x'} \bar{\epsilon}_x \bar{\mathcal{P}}_{xx'} J_x J_{x'}, \tag{5}$$

and also $\sum_x \bar{\epsilon}_x J_x^2 - \alpha \sum_x \bar{\epsilon}_x \sum_{x'} \frac{\bar{\mathcal{P}}_{xx'}}{\alpha} J_x J_{x'} \geq 0$. Since $\bar{\mathcal{P}}_{xx'} = \alpha\mathcal{P}_{xx'} + (1-\alpha)\mathcal{Q}_{xx'}$, and being $\mathcal{Q}$ another stochastic matrix, by substitution in (5), it is: $\sum_x \bar{\epsilon}_x J_x^2 - \alpha \sum_x \bar{\epsilon}_x \sum_{x'} \mathcal{P}_{xx'} J_x J_{x'} > 0$, which completes the proof. ∎

[3]Weighted Euclidean norm on $\mathbb{R}^{|\mathcal{X}|}$ for any vector $J \in \mathbb{R}^{|\mathcal{X}|}$ is defined as $\|J\|_{\epsilon^*} = \sqrt{\sum_{\forall x\in\mathcal{X}} \epsilon^*_x (J(x))^2}$.

[4]In this letter, we make use of standard definitions of norms for matrices and vectors, e.g., $\|\mathcal{P}\|_\infty$ or $\|J\|_\infty$.

Now we are ready to consider the impact of perturbation introduced by the matrix $\bar{\mathcal{P}}$, defined as $\bar{\mathcal{P}} = (I - \mathcal{A})\mathcal{Q} + \mathcal{A}\mathcal{P}^*$, on the projected Bellman equation. This perturbation can be incorporated into the projected Bellman equation by considering the perturbed transition matrix $\bar{\mathcal{P}}$ instead of $\mathcal{P}^*$. The projected Bellman equation with perturbation is: $\Phi\bar{r} = \bar{\Pi}(\mathcal{R} + \alpha\mathcal{P}^*\Phi\bar{r})$, where $\bar{\Pi} = \Phi(\Phi^\top\bar{\Theta}\Phi)^{-1}\Phi^\top\bar{\Theta}$ is the projection operator. Substituting $\bar{\Pi}$ and rearranging, we get: $(\Phi^\top\bar{\Theta}\Phi - \alpha\Phi^\top\bar{\Theta}\mathcal{P}^*\Phi)\bar{r} = \Phi^\top\bar{\Theta}\mathcal{R}$. Defining $\bar{\mathcal{Z}} = \Phi^\top\bar{\Theta}(I - \alpha\mathcal{P}^*)\Phi, \bar{d} = \Phi^\top\bar{\Theta}\mathcal{R}$, so the solution for $\bar{r}$ is: $\bar{r} = \bar{\mathcal{Z}}^{-1}\bar{d}$. This formulation allows us to solve for the parameter vector $\bar{r}$ that approximates the cost function under the perturbed policy, providing a way to assess the impact of deviations from the optimal policy within the framework of ADP. Similar to previous case, for large state spaces, solving directly for $\bar{r}$ is impractical. Instead, iterative methods like TD are used. The iterative update formula in the off-policy case is:

$$\bar{r}_{k+1} = \bar{r}_k - \left(\Phi^\top\bar{\Theta}\Phi\right)^{-1}\left(\bar{\mathcal{Z}}r_k - \bar{d}\right), \tag{6}$$

whose convergence property can be derived from the following theorem.

Theorem 1: Consider a Markov Decision Process (MDP) with state space $\mathcal{X}$, action space $\mathcal{U}$, reward vector $\mathcal{R}$, and optimal transition probability matrix $\mathcal{P}^*$. Let the perturbed transition probability matrix be defined as: $\bar{\mathcal{P}} = (I - \mathcal{A})\mathcal{Q} + \mathcal{A}\mathcal{P}^*$, where $\mathcal{A}$ is a diagonal matrix with the discount factor $\alpha \in]0, 1[$ along its diagonal, and $\mathcal{Q}$ is a stochastic matrix associated with an exploratory policy. Define: $\bar{\mathcal{Z}} = \Phi^\top\bar{\Theta}(I - \alpha\mathcal{P}^*)\Phi$, $\bar{d} = \Phi^\top\bar{\Theta}\mathcal{R}$, where $\Phi \in \mathbb{R}^{|\mathcal{X}|\times m}$ is a feature matrix, and $\bar{\Theta}$ is a diagonal matrix with the stationary probability distribution $\bar{\epsilon}$ on its diagonal. The recursive iteration for the parameter vector $\bar{r}_k \in \mathbb{R}^\psi$ is given by: $\bar{r}_{k+1} = \bar{r}_k - (\Phi^\top\bar{\Theta}\Phi)^{-1}(\bar{\mathcal{Z}}\bar{r}_k - \bar{d})$. Then, the iterative update $\bar{r}_k$ converges to the solution $\bar{r}$ of the projected Bellman equation: $\Phi\bar{r} = \bar{\Pi}(\mathcal{R} + \alpha\mathcal{P}^*\Phi\bar{r})$, where $\bar{\Pi} = \Phi(\Phi^\top\bar{\Theta}\Phi)^{-1}\Phi^\top\bar{\Theta}$.

Proof: From the results of Lemma 2 for the case $\mathcal{P} = \mathcal{P}^*$, we know $\bar{\mathcal{Z}} = \Phi^\top\bar{\Theta}(I - \alpha\mathcal{P}^*)\Phi$ is positive definite. Moreover $(\Phi^\top\bar{\Theta}\Phi)$ is symmetric positive definite, therefore invertible. Since the convergence of (6) is directly related to the positive definiteness of the term $\bar{\Theta}(I - \alpha\mathcal{P}^*)$, then we can guarantee the convergence of projected based ADP with perturbed transition matrices. ∎

Remark 2: The recursive iteration (6) can be computed using the Monte Carlo simulation mechanism. This procedure involves producing two sequence of state visits via off-policy TD methods, where the sequence $\{\bar{x}(0), \bar{x}(1), \bar{x}(2), \ldots\}$ is generated using the transition matrix $\bar{\mathcal{P}}$ (behavior policy) or a steady-state distribution $\bar{\epsilon}$, and we also generate an additional sequence of independent transitions $\{(\bar{x}(0), x(0)), (\bar{x}(1), x(1)), \ldots\}$ according to an original transition matrix $\mathcal{P}$ not necessarily optimal (target policy).

An important metric in analyzing perturbed MDPs is the upper bound for the difference between cost functions derived from perturbed and unperturbed transition probabilities matrices. This ensures consistent policy performance under non-stationary or perturbed conditions in MDPs and helps manage approximation errors in value iteration and ADP methods, thereby maintaining optimal policy quality [13].

The next theorem provides an upper bound for the difference between optimal cost function and its perturbed counterpart. This result is particularly useful when the optimal policy is not known and we want to obtain a bound on the norm of the error between the estimated cost function and optimal one.

Theorem 2: Consider an MDP with $\mathcal{X}, \mathcal{U}, \mathcal{R}$, and two different transition probability matrices, $\mathcal{P}^*$ and $\bar{\mathcal{P}}$, with associated cost functions J^* and $\bar{J}$, respectively. Let $\bar{\mathcal{P}} = (I - \mathcal{A})\mathcal{Q} + \mathcal{A}\mathcal{P}^*$, where $\mathcal{A}$ is a diagonal matrix with $\alpha \in]0, 1[$ on the diagonal, and $\mathcal{Q}$ is a stochastic matrix. If $\|\mathcal{P}^* - \bar{\mathcal{P}}\|_\infty \leq (1-\alpha)$, the difference between J^* and $\bar{J}$ is bounded by: $\|J^* - \bar{J}\|_\infty \leq \frac{\alpha\|\mathcal{R}\|_\infty}{1-\alpha}$.

Proof: Using the Bellman optimality equations and the fixed point mapping of the DP for J^* and $\bar{J}$ [11], we can write the following: $|J^*(x) - \bar{J}(x)| = |\mathcal{R}(x) + \alpha\sum_{x'\in\mathcal{X}}\mathcal{P}^*_{xx'}J^*(x') - \mathcal{R}(x) - \alpha\sum_{x'\in\mathcal{X}}\bar{\mathcal{P}}_{xx'}\bar{J}(x')|$. This simplifies to:

$$\left|J^*(x) - \bar{J}(x)\right| = \alpha\left|\sum_{x'\in\mathcal{X}}\mathcal{P}^*_{xx'}J^*\left(x'\right) - \sum_{x'\in\mathcal{X}}\bar{\mathcal{P}}_{xx'}\bar{J}\left(x'\right)\right|. \tag{7}$$

We now decompose the sum:

$$\sum_{x'\in\mathcal{X}}\mathcal{P}^*_{xx'}J^*\left(x'\right) - \sum_{x'\in\mathcal{X}}\bar{\mathcal{P}}_{xx'}\bar{J}(x) = \sum_{x'\in\mathcal{X}}\left(\mathcal{P}^*_{xx'}J^*\left(x'\right) - \bar{\mathcal{P}}_{xx'}J^*\left(x'\right)\right) + \sum_{x'\in\mathcal{X}}\bar{\mathcal{P}}_{xx'}\left(J^*\left(x'\right) - \bar{J}\left(x'\right)\right). \tag{8}$$

Substituting (8) into (7) we have:

$$\left|J^*(x) - \bar{J}(x)\right| = \alpha\left|\sum_{x'\in\mathcal{X}}\left(\mathcal{P}^*_{xx'}J^*\left(x'\right) - \bar{\mathcal{P}}_{xx'}J^*\left(x'\right)\right) + \sum_{x'\in\mathcal{X}}\bar{\mathcal{P}}_{xx'}\left(J^*\left(x'\right) - \bar{J}\left(x'\right)\right)\right|,$$

Using triangle inequality we get:

$$\left|J^*(x) - \bar{J}(x)\right| \leq \alpha\sum_{x'\in\mathcal{X}}\left|\mathcal{P}^*_{xx'} - \bar{\mathcal{P}}_{xx'}\right|\left|J^*\left(x'\right)\right| + \alpha\sum_{x'\in\mathcal{X}}\left|\bar{\mathcal{P}}_{xx'}\right|\left|J^*\left(x'\right) - \bar{J}\left(x'\right)\right|, \tag{9}$$

where the first term contains the difference between the transition matrices $\mathcal{P}^*$ and $\bar{\mathcal{P}}$, while the second term accounts for the difference in cost functions. Using the assumption that $\|\mathcal{P}^* - \bar{\mathcal{P}}\|_\infty < (1-\alpha)$, the properties of infinite norm, and the fact that (9) holds for any norm, we bound the first term: $\|J^* - \bar{J}\|_\infty \leq \alpha(1-\alpha)\|J^*\|_\infty + \alpha\sum_{x'\in\mathcal{X}}|\bar{\mathcal{P}}_{xx'}||J^*(x') - \bar{J}(x')|$. Next, we use the fact that $\sum_{x'\in\mathcal{X}}\bar{\mathcal{P}}_{xx'} = 1$, since $\bar{\mathcal{P}}$ is a stochastic matrix. Therefore, we have: $\|J^* - \bar{J}\|_\infty \leq \alpha(1-\alpha)\|J^*\|_\infty + \alpha\|J^* - \bar{J}\|_\infty$, Rearranging the inequality to isolate $\|J^* - \bar{J}\|_\infty$, we get: $\|J^* - \bar{J}\|_\infty \leq \frac{\alpha(1-\alpha)\|J^*\|_\infty}{1-\alpha}$. Since we know from the Bellman equation that $\|J^*\|_\infty \leq \frac{\|\mathcal{R}\|_\infty}{1-\alpha}$, by substituting this bound, we obtain: $\|J^* - \bar{J}\|_\infty \leq \frac{\alpha\|\mathcal{R}\|_\infty}{1-\alpha}$, which completes the proof. ∎

Remark 3: The norm condition $\|\mathcal{P}^* - \bar{\mathcal{P}}\|_\infty \leq (1-\alpha)$ in Theorem 2 is satisfied whenever it is $\|\mathcal{P}^* - \mathcal{Q}\|_\infty \leq 1$. Indeed,

Algorithm 1 Off-Policy TD Learning With $\bar{\mathcal{P}}$

Initialize parameter vector $\bar{r}_0$, discount factor α, feature representation $\phi(x)$ for each state x, and behavior policy transition matrix $\bar{\mathcal{P}}$
for each episode **do**
  for each time step k **do**
    Generate the state $\bar{x}(k)$ using $\bar{\mathcal{P}}$
    Generate the state $x(k)$ from the transition pair $\big(\bar{x}(k), x(k)\big)$ using $\mathcal{P}$
    Observe the immediate cost $\mathcal{R}(x(k))$ associated with the target policy
    Compute the TD error:

$$\delta_k = \mathcal{R}(x(k)) + \alpha\phi(x(k))^\top \bar{r} - \phi(\bar{x}(k))^\top \bar{r}$$

    Update parameter vector: $\bar{r} \leftarrow \bar{r} + \delta_k \phi(x_k)$
  end for
end for

the deviation from the optimal policy introduces a perturbation matrix $\mathcal{P}^* - \bar{\mathcal{P}} = (I - \mathcal{A})(\mathcal{P}^* - \mathcal{Q})$, whose norm is bounded as follows: $\|\mathcal{P}^* - \bar{\mathcal{P}}\|_\infty \leq (1-\alpha)\|\mathcal{P}^* - \mathcal{Q}\|_\infty$.

Algorithm 1 summarizes the off-policy implementation using TD learning and perturbed transition matrix $\bar{\mathcal{P}}$ for the general target transition probability matrix $\mathcal{P}$.

V. Analogy Between the Perturbed Transition Matrix and Q-Learning

In this section, we discuss how the analysis on perturbed transition matrix $\bar{\mathcal{P}}$ relates to Q-learning, where the optimal policy π^* is combined with an exploratory policy, say $\pi_{\mathcal{Q}}$. The mixture of these two policies can be expressed as a weighted combination, similar to the exploration-exploitation trade-off in Q-learning. In Q-learning, the policy is a mixture of an *exploratory policy* and an *optimal policy*, in the sense that the action-selection strategy in Q-learning combines exploitation of the optimal policy π^* and exploration using a stochastic policy $\pi_{\mathcal{Q}}$. This is mathematically described by the policy mixture [27]: $\pi_\xi(u|x) = \xi \cdot \pi_{\mathcal{Q}}(u|x) + (1-\xi)\cdot\pi^*(u|x)$, where $\pi^*(u|x)$ is the optimal policy that chooses action u in state x based on the optimal Q-function $\mathrm{Q}^*(x,u)$, $\pi_{\mathcal{Q}}(u|x)$ is an exploratory policy, and $\xi \in]0,1[$ is the exploration parameter controlling the balance between exploration and exploitation. Under this mixed policy π_ξ, the transition dynamics can be expressed as a combination of the transition matrices corresponding to π^* and $\pi_{\mathcal{Q}}$: $P_{\pi_\xi} = \xi P_{\pi_{\mathcal{Q}}} + (1-\xi)P_{\pi^*}$. We now observe that the perturbed transition matrix $\bar{\mathcal{P}}$ can be seen as a similar mixture of policies, with α playing the role of $1-\xi$ in Q-learning: $\bar{\mathcal{P}} = (I-\mathcal{A})\mathcal{Q} + \mathcal{A}\mathcal{P}^*$. This formulation shows that $\bar{\mathcal{P}}$ models the transition dynamics of a perturbed policy $\bar{\pi}$, which can be described as a probabilistic mixture of the optimal policy π^* and the exploratory policy $\pi_{\mathcal{Q}}$: $\bar{\pi}(u|x) = (1-\alpha)\pi_{\mathcal{Q}}(u|x) + \alpha\pi^*(u|x)$.[5] Thus, α acts similarly to $1-\xi$, determining how much the system follows the optimal policy $\mathcal{P}^*$ versus exploring the exploratory transitions defined by $\mathcal{Q}$. Since $\bar{\mathcal{P}}$ is a weighted combination of the optimal transition matrix $\mathcal{P}^*$ and the exploratory matrix $\mathcal{Q}$, the deviation from the optimal transition matrix is given by: $\|\mathcal{P}^* - \bar{\mathcal{P}}\|_\infty = (1-\alpha)\|\mathcal{P}^* - \mathcal{Q}\|_\infty$. Assuming that $\|\mathcal{P}^* - \mathcal{Q}\|_\infty \leq 1$, this deviation is bounded by: $\|\mathcal{P}^* - \bar{\mathcal{P}}\|_\infty \leq 1-\alpha$. This expression is analogous to the temporal difference error in Q-learning, where α controls the trade-off between exploration and exploitation [27]. As α decreases, the deviation from the optimal transition dynamics increases, reflecting a greater degree of exploration. This trade-off allows the system to balance between following the optimal policy π^* and exploring alternative policies through $\pi_{\mathcal{Q}}$. The deviation from the optimal transition matrix $\mathcal{P}^*$ is controlled by α, similar to the exploration parameter in Q-learning. While the principles of perturbation in Q-learning and the presented approach share similarities, they are *not interchangeable* in general. Q-learning excels in model-free and discrete setups, while our approach is better suited for structured, large-scale, or robust policy optimization tasks. These differences make them complementary tools in the RL framework.

[5]The policy $\bar{\pi}$ deviates from π^* to $\pi_{\mathcal{Q}}$ with probability $1-\alpha$.

VI. Numerical Simulations

In this section, we compute an estimation of the optimal value function J^* through $\bar{J}$, and hence the parameter vector $\bar{r}$, utilizing the off-policy TD approach discussed earlier for an MDP related to a resource allocation problem [1], [2]. In summary, the challenge of managing resource pricing for dynamic allocation is tackled by utilizing parallel stochastic Markov chains. In this model, customers request a resource with a probability of λ_i and release it with a probability of μ_i, both of which correspond to the price c_i. The decision-maker's objective is to maximize profit over an infinite time horizon, which is analogous to minimizing costs in the dual formulation of the problem. Denoting by $x_i(k)$ the state of Markov chain (here number of customers) associated to c_i, we can compute the value function as follows:

$$J^*(x) = \max_{u(k)\in\mathcal{U}} \lim_{T\to\infty} \mathbb{E}\left[\sum_{k=0}^{T} \alpha^k \sum_{i=1}^{m} c_i x_i(k)\right],$$

where $\mathcal{U} = \{c_1, c_2, \ldots, c_m\}$. This resource allocation problem exhibits the curse of dimensionality as the number of available resources N and the possible price choice m increase (see [1] for details). Consider a resource allocation problem with $m = 4$ price levels and $N = 20$ resources. In this case, the cardinality of the state space is $|\mathcal{X}| = 10626$, highlighting the curse of dimensionality in DP problems. Consider a discount factor of $\alpha = 0.9$, with prices $c = [0.8, 0.9, 1, 1.1]$, arrival probabilities $\lambda = [0.090596, 0.048632, 0.015657, 0.005088]$, and service probabilities $\mu = [0.483723, 0.444019, 0.024843, 0.335103]$. These probabilities satisfy the condition: $\max_{c_i,c_j\in\mathcal{U}} \sum_{x'} |\mathcal{P}_{xx'}(c_i) - \mathcal{P}_{xx'}(c_j)| < (1-\alpha)$, for all $i,j = 1,\ldots,m$, ensuring stability across all pairs of stationary policies (including the optimal one). We consider 5 features for each state x of MDP, as $\phi_1(x) = 1, \phi_i(x) = x_i,\ i = 1,\ldots,\ m$, hence the parameter vector $r \in \mathbb{R}^5$. We ran 100 Monte Carlo simulations each starting from an arbitrary initial state with the length of 50000 iterations. In particular for Monte Carlo simulations, we employ the procedure explained in Remark 2 and Algorithm 1. The results of these simulations are presented in Fig. 1,

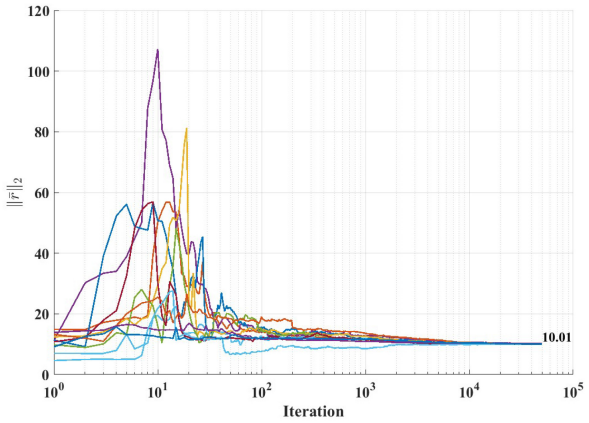

Fig. 1. The behavior of parameter vector $\bar{r}$ through off-policy TD approach for the resource allocation with $m = 4$ prices, $N = 20$ resources and $|\mathcal{X}| = 10626$.

where, for simplicity, we have depicted the behavior of $\|\bar{r}_k\|_2$ of parameter vectors across 10 experiments, showing the convergence of the proposed off-policy TD learning (see Lemma 2 and Remark 2). By averaging the parameter vectors resulted from these experiments, we have $\bar{r} = [0.0212,\ 1.9179,\ 2.2853,\ 8.9394,\ 3.3748]^\top$. After computing $\bar{r}$, one can use the results of the Theorem 2 to approximate J^* for a given state x. In Fig. 1, for the horizontal axis we employ logarithmic scale since the alteration rates of the curves in linear scale are not adequately detectable through iterations. Due to space limitation, further numerical simulations showing the applicability of Algorithm 1 to different target policies can be found in [31].

VII. Conclusion

In this letter, we analyzed the impact of perturbations on optimal policy in DP problems and computed the resulting cost function's deviation from its optimal value. In particular, we handled perturbations in state transition probability matrices using an off-policy TD projection-based approach, which also allowed to address the curse of dimensionality in large-scale MDPs. We also provided the necessary and sufficient conditions for the convergence of the associated Monte Carlo based simulations algorithm. Finally, we validated the presented theoretical results via a suitable numerical example.

References

[1] 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, Sep. 2020, Art. no. 100106.

[2] A. Forootani, M. Tipaldi, R. Iervolino, and S. Dey, "Enhanced exploration least-squares methods for optimal stopping problems," *IEEE Control Syst. Lett.*, vol. 6, pp. 271–276, 2021.

[3] R. Iervolino, M. Tipaldi, and A. Forootani, "A Lyapunov-based version of the value iteration algorithm formulated as a discrete-time switched affine system," *Int. J. Control*, vol. 96, no. 3, pp. 577–592, 2023.

[4] Z. Lin et al., "Policy iteration based approximate dynamic programming toward autonomous driving in constrained dynamic environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 5003–5013, May 2023.

[5] A. Forootani, R. Iervolino, M. Tipaldi, and S. Baccari, "A kernel-based approximate dynamic programming approach: Theory and application," *Automatica*, vol. 162, Apr. 2024, Art. no. 111517.

[6] R. S. Sutton, C. Szepesvári, and H. R. Maei, "A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation," in *Proc. Conf. Neural Inf. Process. Syst.*, 2008, pp. 1609–1616.

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

[8] D. Choi and B. Van Roy, "A generalized Kalman filter for fixed point approximation and efficient temporal-difference learning," *Discrete Event Dyn. Syst.*, vol. 16, no. 2, pp. 207–239, 2006.

[9] M. Geist and O. Pietquin, "Kalman temporal differences," *J. Artif. Intell. Res.*, vol. 39, pp. 483–532, Oct. 2010.

[10] Y. Engel, S. Mannor, and R. Meir, "Reinforcement learning with Gaussian processes," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 201–208.

[11] D. Bertsekas, *Reinforcement Learning and Optimal Control*, vol. 1. Nashua, NH, USA: Athena Sci., 2019.

[12] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[14] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.

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

[16] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, no. 51, pp. 1563–1600, 2010.

[17] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism," 2020, *arXiv:2006.14389*.

[18] C.-Y. Wei and H. Luo, "Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach," in *Proc. Conf. Learn. Theory*, 2021, pp. 4300–4354.

[19] O. D. Domingues, P. Ménard, M. Pirotta, E. Kaufmann, and M. Valko, "A kernel-based approach to non-stationary reinforcement learning in metric spaces," 2020, *arXiv:2007.05078*.

[20] H. Zhong, Z. Yang, Z. Wang, and C. Szepesvári, "Optimistic policy optimization is provably efficient in non-stationary MDPs," 2021, *arXiv:2110.08984*.

[21] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," 2019, *arXiv:1907.05388*.

[22] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. F. Yang, "Model-based reinforcement learning with value-targeted regression," 2020, *arXiv:2006.01107*.

[23] D. Zhou, J. He, and Q. Gu, "Provably efficient reinforcement learning for discounted MDPs with feature mapping," 2020, *arXiv:2006.13165*.

[24] Y. Fei, Z. Yang, Z. Wang, and Q. Xie, "Dynamic regret of policy optimization in non-stationary environments," 2020, *arXiv:2007.00148*.

[25] H. Zhou, J. Chen, L. R. Varshney, and A. Jagmohan, "Nonstationary reinforcement learning with linear function approximation," 2020, *arXiv:2010.04244*.

[26] D. Bertsekas, "Multiagent reinforcement learning: Rollout and policy iteration," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 2, pp. 249–272, Feb. 2021.

[27] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[28] R. S. Sutton, A. R. Mahmood, and M. White, "An emphatic approach to the problem of off-policy temporal-difference learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2603–2631, 2016.

[29] D. P. Bertsekas, *Dynamic Programming and Optimal Control 3rd Edition, Volume II*. Belmont, MA, USA: Athena Sci., 2011.

[30] A. Forootani, R. Iervolino, M. Tipaldi, and S. Dey, "Transmission scheduling for multi-process multi-sensor remote estimation via approximate dynamic programming," *Automatica*, vol. 136, Feb. 2022, Art. no. 110061.

[31] A. Forootani, R. Iervolino, M. Tipaldi, and M. Khosravi, "Off-policy temporal difference learning for perturbed Markov decision processes: Theoretical insights and extensive simulations," 2025, *arXiv:2502.18415*.