Adversarial Attacks on Computation of the Modified Policy Iteration Method

Ali Yekkehkhany, Han Feng, and Javad Lavaei

***Abstract*— Adversarial attacks on Markov decision processes (MDPs) and reinforcement learning (RL) have been studied in the literature in the context of robust learning and adversarial game theory. In this paper, we introduce a new notion of adversarial attacks on MDP and RL computation that is motivated by the emergence of edge computing. The large-scale computation of MDP and RL models in the form of value/policy iteration and Q-learning is being offloaded from agents to distributed servers, giving rise to edge reinforcement learning. By the inherently distributed nature of edge RL, the MDP/RL computation can be prone to adversarial attacks in different forms. We analyze a probabilistic model of adversarial attacks on the computation of the modified policy iteration method in which the principal contraction property of the Bellman operator is undermined with a certain probability in iterations of the policy evaluation step of the aforementioned method. This can result in luring the agent to search among sub-optimal policies without improving the true values of policies. We prove that under certain conditions, the attacked modified policy iteration method can still converge to the vicinity of the optimal policy with high probability if the number of policy evaluation iterations is larger than a threshold that is logarithmic in the inverse of a desired precision. We also provide an upper bound on the number of iterations needed for the attacked modified policy iteration method to terminate, which holds with an associated confidence level.**

***Index Terms*— Adversarial Reinforcement Learning; Markov Decision Process; Contraction-expansion Mapping.**

I. Introduction and Related Work

Markov Decision Processes and Reinforcement Learning frameworks are adopted in a myriad of applications spanning autonomous vehicles [1], [2], healthcare [3], [4], finance [5], [6], energy [7], and cybersecurity [8]–[11]. Given the widespread success of such models in the aforementioned and akin applications, MDP and RL frameworks are potential targets for adversarial attacks with the goal of compromising their performance, which can have catastrophic consequences [12]–[15]. Consequently, adversarial attacks on MDP and RL frameworks in the form of systematic injected perturbations/disturbances by a destabilizing adversary have been the focus of numerous researchers in recent years [16]–[20]. A classical approach to overcoming the effect of adversarial disturbances is to train the MDP and RL models on a set of randomized environments, such as adding noise to state observations [21]–[23]. In a related line of research, the work [24] has used Bayesian optimisation and Bayesian quadrature to make RL robust to presence of rare events. Another popular approach is to formulate training as a two-player game between the agent and the adversary, which is referred to as robust adversarial reinforcement learning [25], [26]. In this paper, we focus on another type of adversarial attack on the computation of MDP and RL models.

By the emergence of cloud, edge, and fog computing, the large-scale computation of MDP and RL models is offloaded from agents to distributed servers, giving rise to edge reinforcement learning [27]–[30]. The computation of these models can be in the form of value iteration, policy iteration, Q-learning, and their variants, which can become vulnerable to adversarial attacks when deployed on the edge. For details on adversarial attacks on edge-deployed computing, the reader can refer to [31]–[33] and the references therein. The convergence of the value/policy iteration and Q-learning methods relies heavily on the contraction property of the Bellman operator. As a result, a natural malevolent attack would be to contaminate the RL computation such that the contraction property of the Bellman operator is undermined. The cause of such disturbances can be a malicious adversary or approximation errors utilized to deal with the intensive RL computation.

The main sources of computation errors can be approximation errors in computing expectation and maximization, as well as utilizing parametric feature-based approximation methods in MDP and RL models [34]. In particular, computing the expectation can be costly, so certainty equivalence, Monte Carlo tree search, and adaptive simulation are utilized to circumvent the issue [35]–[38]. MDP and RL models also deal with maximizing over spaces with a possibly large number of elements, so discretization of the space, linear and nonlinear programming techniques are made use of in order to approximate the optimization of interest. Furthermore, if the number of MDP and RL states and actions are relatively large, tabular methods are substituted with approximation methods, such as neural network architectures [39]–[42]. All the above approximation errors can aggregate and act as an adversary in computation of MDP and RL models.

In this paper, we study a probabilistic model of adversarial attacks on the computation of the modified policy iteration method. The modified policy iteration method consists of a policy evaluation step and a policy improvement step. In the policy evaluation step, the Bellman operator is applied to the value function for a finite number of times. Due to the contraction property of the Bellman operator, the distance between the value function and the true value function of the policy of interest contracts by at least a known factor in each iteration. In the presence of an adversary though, the policy

This work was supported by grants from ARO, AFOSR, ONR and NSF. The authors are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley.
Email: {aliyek, han_feng, lavaei}@berkeley.edu

evaluation step is contaminated with a certain probability in each iteration such that the contraction of the Bellman operator is reversed to an expansion up to a constant. This can result in luring the agent to alternate between sub-optimal policies without improving the true values of policies. We prove that for adversarial expansions up to a particular factor, the attacked modified policy iteration method can still converge to the vicinity of the optimal policy with high probability if the number of policy evaluation iterations is larger than a threshold that is logarithmic in the inverse of a desired precision. We also provide an upper bound on the number of iterations needed for the attacked modified policy iteration method to terminate, which holds with an associated confidence level. This paper is related to our recent work [43] that studies the contraction-expansion attack for the value iteration method, rather than the policy iteration method that turns out to be a more challenging problem.

The rest of the paper is outlined as follows. In Section II, the Markov decision process is described and some preliminaries on the policy iteration method and its modified version are provided. The adversarial attack on the modified policy iteration method is formally introduced in Section III, followed by theoretical results on the convergence of the modified policy iteration method in the presence of an adversary. Concluding remarks are given in Section V.

II. Problem Statement and Preliminaries

Consider a Markov Decision Process (MDP) with the state space $\mathcal{S}$ consisting of a finite number of states, the action space $\mathcal{A}$ with a finite number of actions, and the immediate reward function $r(s_t, a_t, w_t)$, where $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ are the state of the system and the taken action at time $t \in \{0, 1, 2, \dots\}$, respectively, and the randomness in the system is modeled by the sequence of independent and identically distributed random variables w_t for $t \in \{0, 1, 2, \dots\}$. The absolute value of the reward function is assumed to be upper bounded by $R > 0$. Furthermore, the time-invariant state transition function h determines the evolution of the system as $s_{t+1} = h(s_t, a_t, w_t)$.

Consider a deterministic policy space $\mathcal{P}$. Given a policy $\mu : \mathcal{S} \to \mathcal{A}$ in $\mathcal{P}$ mapping states to actions in a deterministic manner together with a discount factor $q \in (0, 1)$, the real-valued value function $V^\mu : \mathcal{S} \to \mathbb{R}$ is defined as the expected discounted sum of rewards over an infinite horizon, i.e.,

$$V^\mu(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} q^t \cdot r(s_t, \mu(s_t), w_t) \middle| s_0 = s\right], \quad \forall s \in \mathcal{S} \tag{1}$$

where the expectation is taken over w_t for $t \in \{0, 1, 2, \dots\}$. Then, the optimal value function V^* is given by

$$V^*(s) = \max_{\mu} V^\mu(s), \quad \forall s \in \mathcal{S}. \tag{2}$$

The objective is to find an optimal policy μ^* in the sense that $V^{\mu^*}(s) = V^*(s)$ for all $s \in \mathcal{S}$. It is straightforward to verify that given the optimal value function V^*, we have

$$\mu^*(s) = \arg\max_{a \in \mathcal{A}} \mathbb{E}\big[r(s, a, w) + q \cdot V^*(h(s, a, w))\big], \ \forall s \in \mathcal{S}, \tag{3}$$

where the expectation is taken over the random variable w that has the same distribution as w_t for some $t \in \{0, 1, \dots\}$. Equations (1), (2), and (3) lead to the Bellman equation

$$V^*(s) = \max_{a \in \mathcal{A}} \mathbb{E}\big[r(s, a, w) + q \cdot V^*(h(s, a, w))\big] \quad \forall s \in \mathcal{S}. \tag{4}$$

Utilizing the Banach fixed-point theorem [44], the Bellman equation has a unique solution since the Bellman operator $\mathcal{T}(\cdot)$ defined as $(\mathcal{T}(V))(s) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r(s, a, w) + q \cdot V(h(s, a, w))\right]$ for all $s \in \mathcal{S}$ is a contraction mapping with respect to the infinity norm, $\|\cdot\|_\infty$, so it has a unique fixed point.

There are various methods for finding the optimal policy such as the value iteration method, the policy iteration method, Q-learning, multi-step look-ahead, and their variants. The focus of this work is on the policy iteration method, which will be described below after presenting some preliminaries. Given a policy μ, Equation (1) provides a system of linear equations in terms of $V^\mu(s)$ for all $s \in \mathcal{S}$. To make this clear, we can rewrite (1) as

$$V^\mu(s) = r_\mu(s) + q \sum_{s' \in \mathcal{S}} p\left(s' | s, \mu(s)\right) \cdot V^\mu(s'), \quad \forall s \in \mathcal{S}, \tag{5}$$

where given that the policy μ is employed at the initial state s, the term $r_\mu(s)$ is the expected immediate reward and $p\left(s' | s, \mu(s)\right)$ is the probability that the next state is s'. The terms $p\left(s' | s, \mu(s)\right)$ for all $s, s' \in \mathcal{S}$ are derived from the transition function h. By appropriately defining the matrix P_μ using $p\left(s' | s, \mu(s)\right)$ for all $s, s' \in \mathcal{S}$, Equation (5) in vector notation is given by

$$\boldsymbol{V}^\mu = \boldsymbol{r}_\mu + q \cdot P_\mu \boldsymbol{V}^\mu, \tag{6}$$

where $\boldsymbol{V}^\mu$ and $\boldsymbol{r}_\mu$ denote the vectors associated with the sets $\{V^\mu(s) | s \in \mathcal{S}\}$ and $\{r_\mu(s) | s \in \mathcal{S}\}$, respectively. The policy iteration method starts with an arbitrary policy μ_0 and alternates between two steps, namely the policy evaluation step and the policy improvement step, as follows for $n \in \{0, 1, 2, \dots\}$:

- Policy Evaluation: Using (6), the value function associated with policy μ_n at iteration n is evaluated as
$$\boldsymbol{V}^{\mu_n} = (I - q \cdot P_{\mu_n})^{-1} \boldsymbol{r}_{\mu_n}, \tag{7}$$
where I is the $\text{card}(\mathcal{S}) \times \text{card}(\mathcal{S})$ identity matrix with $\text{card}(\cdot)$ denoting the cardinality of the input space.
- Policy Improvement: Given the value function $\boldsymbol{V}^{\mu_n}$ in the policy evaluation step, the policy μ_n can potentially be improved to another policy μ_{n+1} as
$$\mu_{n+1}(s) = \arg\max_{a \in \mathcal{A}} \big(r_a(s) + q(P_a \boldsymbol{V}^{\mu_n})(s)\big), \quad \forall s \in \mathcal{S}, \tag{8}$$
where it is guaranteed that $\boldsymbol{V}^{\mu_{n+1}} \geq \boldsymbol{V}^{\mu_n}$ element-wise. If $\boldsymbol{V}^{\mu_{n+1}} = \boldsymbol{V}^{\mu_n}$, it implies that $\boldsymbol{V}^{\mu_n}$ is the unique solution of the Bellman equation in (4), so $\mu^* = \mu_n$ and the iterations stop; otherwise, the policy evaluation and improvement steps repeat.

Note that the policy evaluation step of the policy iteration method is computationally costly since it involves calculation

of the inverse of a matrix. In particular, the computation cost of the policy evaluation step is as high as $\mathcal{O}(\text{card}(\mathcal{S})^3)$. In order to reduce the cost per iteration, a modified/optimistic version of the policy iteration method is proposed in the literature. Starting with an arbitrary policy μ_0 and value function $\boldsymbol{V}_0^{\mu_0}$, the two steps of the modified policy iteration method are as follows for $n \in \{0, 1, 2, \dots\}$:

- Modified Policy Evaluation: A partial policy evaluation is performed for policy μ_n by using the simplified value iteration (which does not involve matrix inversion or exact computation of the value function):

$$\boldsymbol{V}_{k+1}^{\mu_n} = \mathcal{B}^{\mu_n}(\boldsymbol{V}_k^{\mu_n}) = \boldsymbol{r}_{\mu_n} + qP_{\mu_n}\boldsymbol{V}_k^{\mu_n} \tag{9}$$

  for $k \in \{0, 1, \dots, K-1\}$, where K is the number of times that the Bellman backup operator $\mathcal{B}^{\mu_n}(\boldsymbol{V}) = \boldsymbol{r}_{\mu_n} + qP_{\mu_n}\boldsymbol{V}$ is applied. If K goes to infinity, $\boldsymbol{V}_k^{\mu_n}$ converges to the unique fixed point of the operator, $\boldsymbol{V}^{\mu_n} = \mathcal{B}^{\mu_n}(\boldsymbol{V}^{\mu_n})$, by the Banach fixed-point theorem. In practice, K is chosen as a relatively small number to reduce the computation cost of policy evaluation.
- Modified Policy Improvement: Given $\boldsymbol{V}_K^{\mu_n}$, the policy μ_n is updated as

$$\mu_{n+1}(s) = \arg\max_{a\in\mathcal{A}} \left(r_a(s) + q(P_a\boldsymbol{V}_K^{\mu_n})(s)\right), \quad \forall s \in \mathcal{S} \tag{10}$$

  and the value function $\boldsymbol{V}_0^{\mu_{n+1}}$ is set correspondingly as

$$\boldsymbol{V}_0^{\mu_{n+1}} = \max_{a\in\mathcal{A}} \left(r_a(s) + q(P_a\boldsymbol{V}_K^{\mu_n})(s)\right), \quad \forall s \in \mathcal{S}. \tag{11}$$

  If $\|\boldsymbol{V}_0^{\mu_{n+1}} - \boldsymbol{V}_K^{\mu_n}\|_\infty < \epsilon$ for a pre-selected precision $\epsilon > 0$, the modified policy iteration is terminated; otherwise, the modified policy evaluation and improvement steps repeat.

From [45], after the termination of the modified policy iteration, we have $\|\boldsymbol{V}_0^{\mu_{n+1}} - \boldsymbol{V}^*\|_\infty < \frac{\epsilon}{1-q}$ and $\|\boldsymbol{V}^{\mu_{n+1}} - \boldsymbol{V}^*\|_\infty < \frac{2\epsilon}{1-q}$. In the next section, a model for adversarial attacks on the computation of the simplified value iteration step is presented and analyzed.

III. Adversarial Attack Model and Analysis for Modified Policy Iteration Method

The computation of the modified policy evaluation step is expensive for large-scale systems. As a result, the workload is offloaded to the edge and clouds, giving rise to edge reinforcement learning. The distributed nature of edge reinforcement learning brings a host of new adversarial attacks on the aforementioned computation. Note that the principal component of the simplified value iteration method that guarantees improvements in all K iterations, in the sense that the updated value function becomes closer to the true value function of the policy of interest by a factor q, is the contraction property of the Bellman backup operator $\mathcal{B}^{\mu_n}$. In particular, for all $k \in \{0, 1, \dots, K-1\}$, we have

$$\begin{aligned}\left\|\mathcal{B}^{\mu_n}(\boldsymbol{V}_k^{\mu_n}) - \boldsymbol{V}^{\mu_n}\right\|_\infty &= \left\|\boldsymbol{V}_{k+1}^{\mu_n} - \boldsymbol{V}^{\mu_n}\right\|_\infty \\ &\le q \cdot \left\|\boldsymbol{V}_k^{\mu_n} - \boldsymbol{V}^{\mu_n}\right\|_\infty.\end{aligned} \tag{12}$$

The existing theoretical convergence results on policy/value iteration methods are mainly based on the contraction property of the underlying mappings. As a result, an adversary may attempt to undermine this essential contraction property of the Bellman backup operator by contaminating the computation of the simplified value iteration such that an expansion up to a factor $Q \ge 1$ occurs with probability (w.p.) $1-p$ in the K iterations of the modified policy evaluation step independently from each other, where $p \in (0, 1]$. In other words, (9) is modified as

$$\begin{aligned}\overline{\boldsymbol{V}}_{k+1}^{\mu_n} &= \overline{\mathcal{B}}^{\mu_n}(\overline{\boldsymbol{V}}_k^{\mu_n}) \\ &= \begin{cases} \boldsymbol{r}_{\mu_n} + qP_{\mu_n}\overline{\boldsymbol{V}}_k^{\mu_n}, & \text{w.p.} \quad p \\ V : \le Q\cdot\|\overline{\boldsymbol{V}}_k^{\mu_n} - \boldsymbol{V}^{\mu_n}\|_\infty, \\ \quad \text{and } \|V\|_\infty \le \frac{R}{1-q} & \text{otherwise} \end{cases}\end{aligned} \tag{13}$$

for $k \in \{0, 1, \dots, K-1\}$, where $\{\overline{\boldsymbol{V}}_k^{\mu_n}\}_{k=0}^K$ is the compromised stochastic sequence of value functions due to the adversary and $\overline{\mathcal{B}}^{\mu_n}(\cdot)$ is the compromised Bellman backup operator. Note that $|r(s,a,w)| \le R$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, where $|\cdot|$ is the absolute value function, and therefore $\|\boldsymbol{V}^\mu\|_\infty \le \frac{R}{1-q}$ for all $\mu \in \mathcal{P}$. Correspondingly, the adversary causes an expansion up to a factor Q in (13) such that $\|\overline{\boldsymbol{V}}_{k+1}^{\mu_n}\|_\infty \le \frac{R}{1-q}$ is satisfied for all $\mu_n \in \mathcal{P}$ and $k \in \{0, 1, \dots, K\}$ so that the attack remains indistinguishable. Equation (13) results in

$$\begin{aligned}&\left\|\overline{\mathcal{B}}^{\mu_n}(\overline{\boldsymbol{V}}_k^{\mu_n}) - \boldsymbol{V}^{\mu_n}\right\|_\infty = \left\|\overline{\boldsymbol{V}}_{k+1}^{\mu_n} - \boldsymbol{V}^{\mu_n}\right\|_\infty \\ &\le \begin{cases} q\cdot\|\overline{\boldsymbol{V}}_k^{\mu_n} - \boldsymbol{V}^{\mu_n}\|_\infty, & \text{w.p. } p \\ Q\cdot\|\overline{\boldsymbol{V}}_k^{\mu_n} - \boldsymbol{V}^{\mu_n}\|_\infty, & \text{otherwise} \end{cases},\end{aligned} \tag{14}$$

for $k \in \{0, 1, \dots, K-1\}$. We name the compromised operator $\overline{\mathcal{B}}^{\mu_n}(\cdot)$ a probabilistic contraction-expansion mapping.

It is not known whether the modified policy iteration method converges to a close vicinity of the value function of the optimal policy under an adversarial attack. In particular, even with infinite iterations of the simplified value iteration, the Banach fixed-point theorem cannot be utilized to guarantee convergence of the compromised sequence $\{\overline{\boldsymbol{V}}_k^{\mu_n}\}$ to $\{\boldsymbol{V}^{\mu_n}\}$ since the compromised operator $\overline{\mathcal{B}}^{\mu_n}(\cdot)$ is not a contraction mapping. In this work, we consider a slightly modified variant of the policy iteration method subject to an adversary for which we develop theoretical results. Starting with an arbitrary policy $\overline{\mu}_0$ and a value function $\overline{\boldsymbol{V}}_0^{\overline{\mu}_0}$, the two steps of the modified policy iteration in the presence of an adversary are as follows for $n \in \{0, 1, 2, \dots\}$:

- Modified Policy Evaluation in the Presence of an Adversary: The partial policy evaluation is contaminated by an adversary such that the contaminated stochastic value function sequence $\{\overline{\boldsymbol{V}}_k^{\overline{\mu}_n}\}_{k=1}^K$ is generated according to (13) with the property in (14).
- Modified Policy Improvement: Given $\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}$, the policy $\overline{\mu}_{n+1}$ is derived in the same way as (10) and the value function $\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}}$ is set accordingly as in (11). If

$\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}} - \overline{\boldsymbol{V}}_K^{\overline{\mu}_n}\|_\infty < \epsilon$, the loop is terminated; otherwise, the modified policy evaluation and improvement steps repeat.

In the following, we present two theorems characterizing the condition under which the modified policy iteration method converges to the vicinity of the optimal policy in the presence of an adversary with an associated confidence level. We also provide an upper bound on the number of iterations needed by the contaminated modified policy iteration method to guarantee a user-defined confidence level. Prior to developing the technical results, we define a parameter δ that captures the inherent difficulty of the Markov decision process of interest. Let

$$\delta = \min_{\mu\in\mathcal{P}, s\in\mathcal{S}} \Big(\max_{a\in\mathcal{A}} \left(r_a(s) + q(P_a \boldsymbol{V}^\mu)(s)\right) - \max_{a\in\mathcal{A}\setminus\mathcal{A}^\mu(s)} \left(r_a(s) + q(P_a \boldsymbol{V}^\mu)(s)\right) \Big), \tag{15}$$

where $\mathcal{A}^\mu(s) = \arg\max_{a\in\mathcal{A}} \left(r_a(s) + q(P_a \boldsymbol{V}^\mu)(s)\right)$ for all $s \in \mathcal{S}$.

Theorem 1: Consider the Markov Decision Process $(\mathcal{S}, \mathcal{A}, r, p, q)$ and the contaminated policy iteration method whose policy evaluation step is according to the probabilistic contraction-expansion mapping in (13) with the expansion factor Q such that $p\cdot\log(q) + (1-p)\cdot\log(Q) < 0$. Let K, the number of iterations in the contaminated modified policy evaluation step, satisfy

$$K \geq \max\left\{ \frac{\log\left(\frac{1}{a}\right)\cdot\left(\log\left(\frac{Q}{q}\right)\right)^2}{2\big(L + p\cdot\log(q) + (1-p)\cdot\log(Q)\big)^2}, \frac{\log\left(\frac{2R}{(1-q)\cdot\min\{\frac{\delta}{2q}, \frac{\epsilon}{1+q}\}}\right)}{L} \right\}, \tag{16}$$

where $L \in \big(0, -p\cdot\log(q) - (1-p)\cdot\log(Q)\big)$, $a \in (0,1]$, and δ is defined in (15). Let n denote the iteration at which the contaminated policy iteration method terminates and $\overline{\mu}_{n+1}$ denote the associated policy. Then, $\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}} - \boldsymbol{V}^*\|_\infty < \frac{2\epsilon}{1-q}$ and $\|\boldsymbol{V}^{\overline{\mu}_{n+1}} - \boldsymbol{V}^*\|_\infty < \frac{4\epsilon}{1-q}$ with the confidence level $(1-a)$.

Proof: Given the definition of the probabilistic contraction-expansion mapping $\overline{\mathcal{B}}^{\overline{\mu}_n}(\cdot)$ in (13) and its property in (14), we have

$$\begin{aligned} &\left\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty = \left\|\overline{\mathcal{B}}^{\overline{\mu}_n}(\overline{\boldsymbol{V}}_{K-1}^{\overline{\mu}_n}) - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \\ \leq& B_K \cdot \left\|\overline{\boldsymbol{V}}_{K-1}^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty = \left\|\overline{\mathcal{B}}^{\overline{\mu}_n}(\overline{\boldsymbol{V}}_{K-1}^{\overline{\mu}_n}) - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \\ \leq& B_K \cdot B_{K-1} \cdot \left\|\overline{\boldsymbol{V}}_{K-2}^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \\ \leq& \prod_{i=1}^K B_i \cdot \left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty, \end{aligned} \tag{17}$$

where the independent and identically distributed random variables B_i for $i \in \{1, 2, \ldots, K\}$ have the distribution

$$B_i = \begin{cases} q, & \text{w.p.} \quad p \\ Q, & \text{otherwise} \end{cases}. \tag{18}$$

Taking logarithm of both sides of (17) leads to

$$\begin{aligned} &\log\left(\left\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty\right) \\ &\leq \sum_{i=1}^K \log(B_i) + \log\left(\left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty\right) \\ &\leq K\cdot\overline{S}_K + \log\left(\left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty\right), \end{aligned} \tag{19}$$

where the random variable $\overline{S}_K$ is defined as $\overline{S}_K = \frac{\sum_{i=1}^K \log(B_i)}{K}$. Note that random variables $\log(B_i)$ for $i \in \{1, 2, \ldots, K\}$ are strictly restricted to the interval $[\log(q), \log(Q)]$ and using the law of the unconscious statistician (LOTUS), we have $\mathbb{E}[\log(B_i)] = p\cdot\log(q) + (1-p)\cdot\log(Q)$. As a result, it results from Hoeffding's inequality that

$$\begin{aligned} &\mathbb{P}\Big\{\overline{S}_K - p\cdot\log(q) - (1-p)\cdot\log(Q) < \overline{L}\Big\} \\ &> 1 - \exp\left(-\frac{2K\overline{L}^2}{(\log(Q) - \log(q))^2}\right), \end{aligned} \tag{20}$$

where $\overline{L} > 0$ and $\mathbb{P}\{\cdot\}$ takes the probability of the input events. Combining (19) and (20) gives rise to

$$\begin{aligned} &\mathbb{P}\Bigg\{ \log\left(\left\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty\right) \\ &< -KL + \log\left(\left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty\right) \Bigg\} \\ &> 1 - \exp\left(-\frac{2K\big(L + p\cdot\log(q) + (1-p)\cdot\log(Q)\big)^2}{(\log(Q) - \log(q))^2}\right), \end{aligned} \tag{21}$$

where $L = -\overline{L} - p\cdot\log(q) - (1-p)\cdot\log(Q)$. Taking exponential of both sides of the inequality inside the probability in Equation (21) results in

$$\begin{aligned} &\mathbb{P}\Big\{\left\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty < \left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \cdot \exp(-KL)\Big\} \\ &> 1 - \exp\left(-\frac{2K\big(L + p\cdot\log(q) + (1-p)\cdot\log(Q)\big)^2}{(\log(Q) - \log(q))^2}\right). \end{aligned} \tag{22}$$

As a result, for δ defined in (15) and $a \in (0,1]$, we have

$$\mathbb{P}\Big\{\left\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty < \min\left\{\frac{\delta}{2q}, \frac{\epsilon}{1+q}\right\}\Big\} > 1 - a \tag{23}$$

if K satisfies the two inequalities

$$\exp\left(-\frac{2K\left(L + p\cdot\log(q) + (1-p)\cdot\log(Q)\right)^2}{(\log(Q) - \log(q))^2}\right) \leq a, \tag{24a}$$

$$\left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \cdot \exp(-KL) \leq \min\left\{\tfrac{\delta}{2q}, \tfrac{\epsilon}{1+q}\right\}. \tag{24b}$$

Assume that $\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n} - \boldsymbol{V}^{\overline{\mu}_n}\|_\infty \neq 0$; otherwise, $\overline{\boldsymbol{V}}_K^{\overline{\mu}_n} = \overline{\boldsymbol{V}}_0^{\overline{\mu}_n} = \boldsymbol{V}^{\overline{\mu}_n}$ is used for policy improvement, resulting in the policy improvement step being unaffected by the adversary. If $p\cdot\log(q) + (1-p)\cdot\log(Q) < 0$ and $L \in \big(0, -p\cdot\log(q) - (1-p)\cdot\log(Q)\big)$, then the two inequalities

in (24a) and (24b) are satisfied when

$$K \geq \max\left\{ \frac{\log\left(\frac{1}{a}\right)\cdot\left(\log\left(\frac{Q}{q}\right)\right)^2}{2\big(L+p\cdot\log(q)+(1-p)\cdot\log(Q)\big)^2}, \frac{\log\left(\frac{\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\|_\infty}{\min\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\}}\right)}{L}\right\}. \tag{25}$$

Given that $\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\|_\infty \leq \|\overline{\boldsymbol{V}}_0^{\overline{\mu}_n}\|_\infty + \|\boldsymbol{V}^{\overline{\mu}_n}\|_\infty \leq \frac{2R}{1-q}$, the second argument of the max function in (25) can be written as $\frac{\log\left(2R/((1-q)\cdot\min\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\})\right)}{L}$. Consequently, if the simplified value iteration method in the policy evaluation step is performed for the number of times determined in (25), we obtain the following relations with the associated confidence level $1-a$ for all $s\in\mathcal{S}$:

$$\begin{aligned}
&\max_{a\in\mathcal{A}}\left|\big(r_a(s)+q(P_a\overline{\boldsymbol{V}}_K^{\overline{\mu}_n})(s)\big)-\big(r_a(s)+q(P_a\boldsymbol{V}^{\overline{\mu}_n})(s)\big)\right| \\
&=\max_{a\in\mathcal{A}}\left|q\big((P_a\overline{\boldsymbol{V}}_K^{\overline{\mu}_n})(s)-(P_a\boldsymbol{V}^{\overline{\mu}_n})(s)\big)\right| \\
&\leq q\cdot\left\|P_a\big(\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\big)\right\|_\infty \leq q\cdot\left\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \\
&\overset{(a)}{\leq} q\cdot\min\left\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\right\}\leq\frac{\delta}{2},
\end{aligned} \tag{26}$$

where (a) holds true with the associated confidence level given in (23). As a result, leveraging the assumption in (15), the policy improvement step is unaffected by the adversary with an associated confidence level if K satisfies the inequality in (25).

If the condition $\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}}-\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}\|_\infty<\epsilon$ is satisfied at iteration n, with the confidence level $(1-a)$, we have

$$\begin{aligned}
&\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}}-\boldsymbol{V}^{\overline{\mu}_n}\|_\infty \\
&\quad=\left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}}-\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}+\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \\
&\quad\overset{(a)}{\leq}\left\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}}-\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}\|_\infty+\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\right\|_\infty \\
&\quad\overset{(b)}{\leq}\epsilon+\min\left\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\right\}<2\epsilon,
\end{aligned} \tag{27}$$

where (a) follows from the triangular inequality and (b) is due to the termination condition of the modified policy iteration and (23). In light of [45], Since $\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}}-\boldsymbol{V}^{\overline{\mu}_n}\|_\infty<2\epsilon$ with probability at least $(1-a)$, we have $\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n+1}}-\boldsymbol{V}^*\|_\infty<\frac{2\epsilon}{1-q}$ and $\|\boldsymbol{V}^{\overline{\mu}_{n+1}}-\boldsymbol{V}^*\|_\infty<\frac{4\epsilon}{1-q}$ with the same confidence level.

■

Theorem 2: Given a natural number N and under the assumptions of Theorem 1, the policy iteration method in the absence of an adversary terminates in at most N iterations over all the choices of μ_0, then the number of iterations of the contaminated modified policy iteration N_a satisfies

$$\begin{aligned}
&\mathbb{P}\left\{N_a<\frac{(1-a)^{-N}-1}{a}+k\right\} \\
&\geq 1-\frac{a\cdot(1-a)^{-N}\cdot N^3+\frac{2\left((1-a)^{-N}-1\right)^2}{a^2}}{k^2},
\end{aligned} \tag{28}$$

where k is a natural number.

Proof: In Theorem 1, it is proved that upon the termination of the modified policy iteration method in the presence of an adversary, the obtained value function is in the vicinity of the optimal value function if the number of iterations of the policy evaluation is large enough. In the following, we present a confidence interval for the number of iterations in the modified policy iteration method with an associated confidence level in the presence of the adversary. Assume that the maximum number of policy improvement steps in the policy iteration method in the absence of the adversary is N over all the choices of μ_0. On the other hand, given (16) in the presence of the adversary, we have $\mathbb{P}\big\{\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\|_\infty<\min\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\}\big\}>1-a$. Due to (15), this inequality implies that iteration n of the modified policy iteration in the presence of the adversary results in a policy improvement that is the same as the policy improvement applied to $\overline{\mu}_n$ in the absence of the adversary with the associated confidence level. As a result, if the event $\big\{\|\overline{\boldsymbol{V}}_K^{\overline{\mu}_n}-\boldsymbol{V}^{\overline{\mu}_n}\|_\infty<\min\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\}\big\}$ occurs in N consecutive iterations, $n\in\{n_a-N+1,n_a-N+2,\dots,n_a\}$, the modified policy iteration in the presence of the adversary ends up with the optimal policy $\overline{\mu}_{n_a}=\mu^*$. Furthermore,

$$\begin{aligned}
&\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n_a+1}}-\overline{\boldsymbol{V}}_K^{\overline{\mu}_{n_a}}\|_\infty \\
&\quad\leq\|\overline{\boldsymbol{V}}_0^{\overline{\mu}_{n_a+1}}-V^*\|_\infty+\|V^*-\overline{\boldsymbol{V}}_K^{\overline{\mu}_{n_a}}\|_\infty \\
&\quad<q\cdot\min\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\}+\min\{\frac{\delta}{2q},\frac{\epsilon}{1+q}\}<\epsilon.
\end{aligned} \tag{29}$$

As a result, the termination condition of the modified policy iteration in the presence of the adversary is satisfied if the aforementioned event occurs in N consecutive iterations.

In order to find an upper bound on the expectation and variance of the number of iterations for the modified policy iteration in the presence of an adversary, consider a Bernoulli process $\{X_1,X_2,X_3,\dots\}$ with $\mathbb{E}[X_i]=1-a$ for all $i\in\{1,2,\dots\}$. Define the random variable N_a as the first time i such that $X_i=X_{i-1}=\cdots=X_{i-N+1}=1$. Then, the expectation and variance of the number of iterations for modified policy iteration in the presence of an adversary are upper bounded by $\mathbb{E}[N_a]$ and $\mathbb{E}[N_a^2]-N^2$, respectively. Note that the set of sequences of possible outcomes for the Bernoulli process is partitioned by events $E_1,E_2,\dots,E_{N+1}$, where $E_1=\{X_1=0\}$, $E_2=\{X_1=1,X_2=0\}$, $E_3=\{X_1=1,X_2=1,X_3=0\}$, …, $E_N=\{X_1=1,\dots,X_{N-1}=1,X_N=0\}$, and $E_{N+1}=\{X_1=1,\dots,X_{N-1}=1,X_N=1\}$. Using the law of total probability, we obtain that

$$\begin{aligned}
\mathbb{E}[N_a]=&\,\mathbb{E}[N_a|E_1]\cdot\mathbb{P}\{E_1\}+\mathbb{E}[N_a|E_2]\cdot\mathbb{P}\{E_2\} \\
&+\cdots+\mathbb{E}[N_a|E_N]\cdot\mathbb{P}\{E_N\}+N\cdot\mathbb{P}\{E_{N+1}\} \\
\overset{(a)}{=}&\,\mathbb{E}[N_a^1+1]\cdot\mathbb{P}\{E_1\}+\mathbb{E}[N_a^2+2]\cdot\mathbb{P}\{E_2\} \\
&+\cdots+\mathbb{E}[N_a^N+N|E_N]\cdot\mathbb{P}\{E_N\}+N\cdot\mathbb{P}\{E_{N+1}\} \\
=&\,a\cdot(\mathbb{E}[N_a]+1)+a\cdot(1-a)\cdot(\mathbb{E}[N_a]+2)+\cdots+ \\
&\,a\cdot(1-a)^{N-1}\cdot(\mathbb{E}[N_a]+N)+(1-a)^N\cdot N
\end{aligned}$$

$$\overset{(b)}{=} \left(1-(1-a)^N\right)\cdot \mathbb{E}[N_a] + \frac{1-(1-a)^N}{a}, \tag{30}$$

where (a) holds because conditioned on E_i occurring for $i \in \{1,\dots,N\}$, the first i components of the Bernoulli process do not contribute to the observation of N consecutive ones and the excess number of trials to observe N consecutive ones, denoted by N_a^i, has the same distribution as N_a, and (b) follows from the geometric series summation and differentiation of the geometric series summation formulas. By solving (30) for $\mathbb{E}[N_a]$, we have

$$\mathbb{E}[N_a] = \frac{(1-a)^{-N}-1}{a}. \tag{31}$$

The second moment of N_a can also be computed as

$$\begin{aligned}\mathbb{E}[N_a^2] =\ & a\cdot\mathbb{E}[(N_a+1)^2] + a\cdot(1-a)\cdot\mathbb{E}[(N_a+2)^2] \\ & + \cdots + a\cdot(1-a)^{N-1}\cdot\mathbb{E}[(N_a+N)^2] + (1-a)^N\cdot N^2.\end{aligned} \tag{32}$$

The second moment of N_a can be derived by solving (32), and then be used to upper bound $\mathrm{Var}\,(N_a)$ as

$$\begin{aligned}\mathrm{Var}\,(N_a) =& N^2 + a\cdot(1-a)^{-N}\cdot\sum_{i=1}^{N}\left(i^2\cdot(1-a)^{i-1}\right) \\ & + 2a\cdot\mathbb{E}[N_a]\cdot\sum_{i=1}^{N}\left(i\cdot(1-a)^{i-N-1}\right) - \left(\mathbb{E}[N_a]\right)^2 \\ \le\ & a\cdot(1-a)^{-N}\cdot N^3 + \frac{2\left((1-a)^{-N}-1\right)^2}{a^2}.\end{aligned} \tag{33}$$

Using Chebyshev's inequality together with (31) and (33), we have

$$\begin{aligned}&\mathbb{P}\left\{N_a < \frac{(1-a)^{-N}-1}{a} + k\right\} \\ &\quad \ge 1 - \frac{a\cdot(1-a)^{-N}\cdot N^3 + \frac{2\left((1-a)^{-N}-1\right)^2}{a^2}}{k^2}.\end{aligned} \tag{34}$$

■

IV. Numerical Example

To illustrates our results on robustness against adversarial attacks, consider the MDP illustrated in Figure 1. The set of states includes the feasible positions in the grid. The agent can take any of the four actions Up, Down, Right, and Left in each of the non-terminal states. By taking an action, the agent moves one block toward the desired action 90% of the time, or moves one block to one of the remaining directions uniformly at random 10% of the time. The agent bounces back to its original state before taking an action if movement in the direction described above is not possible due to the walls marked with diagonal strips or exiting the environment. The agent is incurred a cost of 0.02 by each move, and there are two terminal states in which the agent receives an immediate reward of $+1$ or -1. Therefore, the maximum reward is bounded by $R=1$. We consider the special case where the number of immediate blocks is $s=8$. For the parameters of the problem, we set $p=0.9$, $q=0.9$, and $Q=1.5$. Those parameters satisfy the relation $(1-p)\log(Q) + p\log(q) \approx -0.054 < 0$. To implement the compromised Bellman operator in (13), we add uniform noise to the value functions with probability $1-p$. The noises are re-scaled so that the inequality in (13) is satisfied. We let the number of policy evaluation steps be $K=4$.

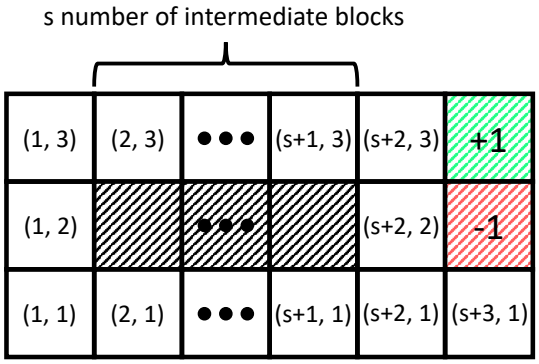

Fig. 1. Settings of the MDP Problem.

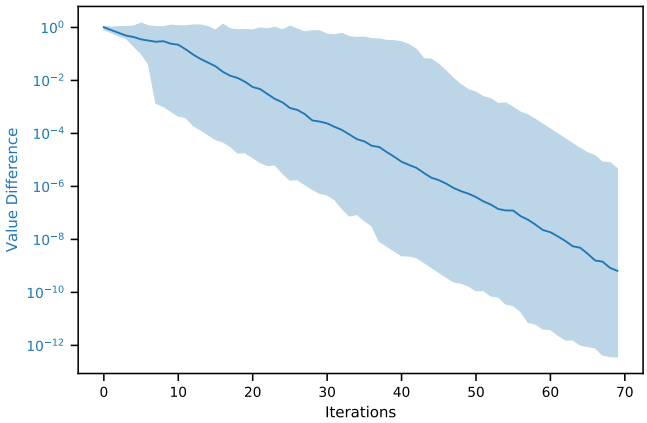

Fig. 2. Convergence of value function when there are 8 intermediate blocks. The difference in value functions is measured in $\|\cdot\|_\infty$. The solid line shows the median and the 90% confidence region is shaded.

Figure 2 and Figure 3 plot the difference of value function and policy against the number of iterations for the case with 8 intermediate blocks. The difference between the value function and the optimal value function is measured in $\|\cdot\|_\infty$, which is the maximum difference in value function across all states. The difference in policy is measured with $\|\cdot\|_0$, which counts the difference of two policies' actions across all states. The experiment is repeated 100 times. The solid line shows the median (both for value function and policy) and the 90% confidence region is shaded. The figure shows that the policies often converge quickly, even though the errors persist in the value function due to the compromised Bellman operator.

Figure 4 and Figure 5 illustrate the number of iterations necessary for the value function and the policy to converge. The experiment is repeated 100 times. The solid line shows the median and the 90% confidence region is shaded. The existence of adversarial attacks can dramatically change the number of iterations necessary for convergence, and

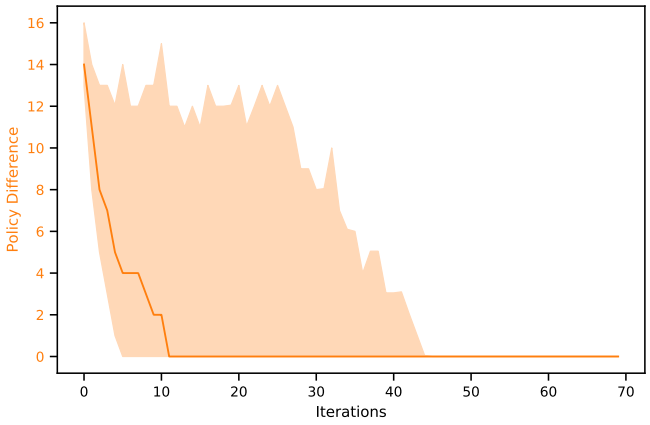

Fig. 3. Convergence of policy when there are 8 intermediate blocks. The difference in policy is measured with $\|\cdot\|_0$, which counts the difference of two policies' actions across all states. The solid line shows the median and the 90% confidence region is shaded.

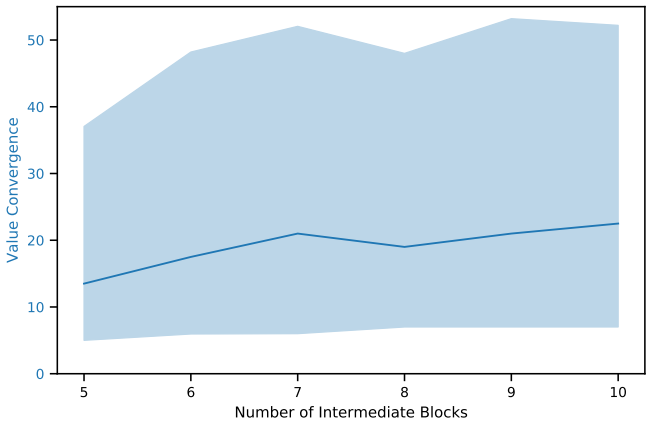

Fig. 4. The number of iterations for the value function to reach $\epsilon = 0.01$ neighborhood of the optimal value function. The difference between the value function and the optimal value function is measured in $\|\cdot\|_\infty$. The solid line shows the median and the 90% confidence region is shaded.

convergence can be slowed further as the number of states increases.

V. Conclusion and Future Work

Motivated by the emerging challenges of edge computing, this paper studies the adversarial attack on the policy evaluation steps of the modified policy iteration algorithm. The attack is modeled by undermining the contraction property of the Bellman operator. As shown in our numerical experiments, the attack can render the convergence of policy evaluation highly uncertain. We prove convergence to the vicinity of the optimal policy with high probability with a suitable number of modified policy evaluation iterations. Under a pre-specified confidence level, we provide an upper bound on the number of iterations needed for the attacked modified policy iteration method to terminate. Future work includes extending the attack model to model-free reinforcement learning algorithms.

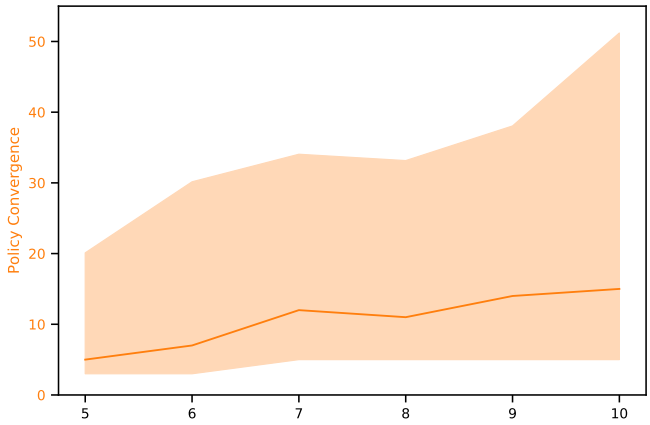

Fig. 5. The number of iterations for the policy to converge to the optimal policy. The difference in policy is measured with $\|\cdot\|_0$, which counts the difference of two policies' actions across all states. The solid line shows the median and the 90% confidence region is shaded.

References

[1] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electronic Imaging*, vol. 2017, no. 19, pp. 70–76, 2017.

[2] N. Li, A. Girard, and I. Kolmanovsky, "Stochastic predictive control for partially observable markov decision processes with time-joint chance constraints and application to autonomous vehicle control," *Journal of Dynamic Systems, Measurement, and Control*, vol. 141, no. 7, 2019.

[3] J. Patrick and M. A. Begen, "Markov decision processes and its applications in healthcare," *Handbook of healthcare delivery systems. CRC, Boca Raton*, 2011.

[4] O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi, "Guidelines for reinforcement learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 16–18, 2019.

[5] N. Bäuerle and U. Rieder, *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011.

[6] P. N. Kolm and G. Ritter, "Modern perspectives on reinforcement learning in finance," *The Journal of Machine Learning in Finance*, vol. 1, no. 1, 2020.

[7] M. Jin and J. Lavaei, "Stability-certified reinforcement learning: A control-theoretic perspective," *IEEE Access*, 2020.

[8] R. Ganesan, S. Jajodia, A. Shah, and H. Cam, "Dynamic scheduling of cybersecurity analysts for minimizing risk using reinforcement learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, pp. 1–21, 2016.

[9] C. Li and M. Qiu, *Reinforcement Learning for Cyber-Physical Systems with Cybersecurity Case Studies*. CRC Press, 2019.

[10] Y. Hao, M. Wang, and J. H. Chow, "Likelihood analysis of cyber data attacks to power systems with markov decision processes," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3191–3202, 2016.

[11] L. L. Njilla, C. A. Kamhoua, K. A. Kwiat, P. Hurley, and N. Pissinou, "Cyber security resource allocation: a markov decision process approach," in *2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)*. IEEE, 2017, pp. 49–52.

[12] X. Pan, D. Seita, Y. Gao, and J. Canny, "Risk averse robust adversarial reinforcement learning," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8522–8528.

[13] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.

[14] A. Javanmard and M. Soltanolkotabi, "Precise statistical analysis of classification accuracies for adversarial training," *arXiv preprint arXiv:2010.11213*, 2020.

[15] T. M. Moldovan and P. Abbeel, "Safe exploration in Markov decision processes," in *International Conference on Machine Learning*, 2017.

[16] L. Pinto, J. Davidson, and A. Gupta, "Supervision via competition: Robot adversaries for learning tasks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1601–1608.

[17] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[18] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," *arXiv preprint arXiv:1703.06748*, 2017.

[19] R. Elderman, L. J. Pater, A. S. Thie, M. M. Drugan, and M. Wiering, "Adversarial reinforcement learning in a cyber security simulation." in *ICAART (2)*, 2017, pp. 559–566.

[20] T. Spooner and R. Savani, "Robust market making via adversarial reinforcement learning," *arXiv preprint arXiv:2003.01820*, 2020.

[21] A. Mandlekar, Y. Zhu, A. Garg, L. Fei-Fei, and S. Savarese, "Adversarially robust policy learning: Active construction of physically-plausible perturbations," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3932–3939.

[22] A. Pattanaik, Z. Tang, S. Liu, G. Bommannan, and G. Chowdhary, "Robust deep reinforcement learning with adversarial attacks," *arXiv preprint arXiv:1712.03632*, 2017.

[23] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[24] S. Paul, K. Chatzilygeroudis, K. Ciosek, J.-B. Mouret, M. Osborne, and S. Whiteson, "Alternating optimisation and quadrature for robust control," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[25] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2817–2826.

[26] X. Ma, K. Driggs-Campbell, and M. J. Kochenderfer, "Improved robustness and safety for autonomous vehicle control with adversarial reinforcement learning," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1665–1671.

[27] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[28] H. Li, K. Ota, and M. Dong, "Learning IoT in edge: Deep learning for the internet of things with edge computing," *IEEE network*, vol. 32, no. 1, pp. 96–101, 2018.

[29] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[30] J. Na, H. Zhang, X. Deng, B. Zhang, and Z. Ye, "Accelerate personalized IoT service provision by cloud-aided edge reinforcement learning: A case study on smart lighting," in *International Conference on Service-Oriented Computing*. Springer, 2020, pp. 69–84.

[31] M. Isakov, V. Gadepally, K. M. Gettings, and M. A. Kinsy, "Survey of attacks and defenses on edge-deployed neural networks," in *2019 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2019, pp. 1–8.

[32] M. S. Ansari, S. H. Alsamhi, Y. Qiao, Y. Ye, and B. Lee, "Security of distributed intelligence in edge computing: Threats and countermeasures," in *The Cloud-to-Thing Continuum*. Palgrave Macmillan, Cham, 2020, pp. 95–122.

[33] Y. Xiao, Y. Jia, C. Liu, X. Cheng, J. Yu, and W. Lv, "Edge computing security: State of the art and challenges," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1608–1631, 2019.

[34] P. B. Dimitri, *Dynamic programming and optimal control.* Athena Scientific, 2017.

[35] H. S. Chang, J. Hu, M. C. Fu, and S. I. Marcus, *Simulation-based algorithms for Markov decision processes*. Springer Science & Business Media, 2013.

[36] R. Coulom, "Efficient selectivity and backup operators in monte-carlo tree search," in *International conference on computers and games*. Springer, 2006, pp. 72–83.

[37] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of monte carlo tree search methods," *IEEE Transactions on Computational Intelligence and AI in games*, vol. 4, no. 1, pp. 1–43, 2012.

[38] M. C. Fu, "Markov decision processes, alphago, and monte carlo tree search: Back to the future," in *Leading Developments from INFORMS Communities*. INFORMS, 2017, pp. 68–88.

[39] B. Van Roy, "Learning and value function approximation in complex decision processes," Ph.D. dissertation, Massachusetts Institute of Technology, 1998.

[40] J. N. Tsitsiklis and B. Van Roy, "Feature-based methods for large scale dynamic programming," *Machine Learning*, vol. 22, no. 1, pp. 59–94, 1996.

[41] B. Van Roy, "Performance loss bounds for approximate value iteration with state aggregation," *Mathematics of Operations Research*, vol. 31, no. 2, pp. 234–244, 2006.

[42] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement learning and dynamic programming using function approximators*. CRC press, 2010, vol. 39.

[43] A. Yekkehkhany, H. Feng, and J. Lavaei, "A hitting time analysis for stochastic time-varying functions with applications to adversarial attacks on computation of Markov decision processes," 2020. [Online]. Available: https://lavaei.ieor.berkeley.edu/Hitting_Time_2020_2.pdf

[44] S. Banach, "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales," *Fund. math*, vol. 3, no. 1, pp. 133–181, 1922.

[45] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.