Robust Temporal Difference Learning for Critical Domains

Richard Klima
University of Liverpool
Liverpool, United Kingdom
richard.klima@liverpool.ac.uk

Daan Bloembergen
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
d.bloembergen@cwi.nl

Michael Kaisers
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
m.kaisers@cwi.nl

Karl Tuyls
University of Liverpool
Liverpool, United Kingdom
karltuyls@google.com

ABSTRACT

We present a new Q-function operator for temporal difference (TD) learning methods that explicitly encodes robustness against significant rare events (SRE) in critical domains. The operator, which we call the κ-operator, allows to learn a robust policy in a model-based fashion without actually observing the SRE. We introduce single- and multi-agent robust TD methods using the operator κ. We prove convergence of the operator to the optimal robust Q-function with respect to the model using the theory of Generalized Markov Decision Processes. In addition we prove convergence to the optimal Q-function of the original MDP given that the probability of SREs vanishes. Empirical evaluations demonstrate the superior performance of κ-based TD methods both in the early learning phase as well as in the final converged stage. In addition we show robustness of the proposed method to small model errors, as well as its applicability in a multi-agent context.

KEYWORDS

reinforcement learning; robust learning; multi-agent learning

ACM Reference Format:
Richard Klima, Daan Bloembergen, Michael Kaisers, and Karl Tuyls. 2019. Robust Temporal Difference Learning for Critical Domains. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

Many critical systems exhibit global system dynamics that are highly sensitive to the local performance of individual components. This holds for example for (air) traffic and transport networks, communication networks, security systems, and (smart) power grids [5, 13, 16, 24]. In each case, the failure of or malicious attack on a small set of nodes may lead to knock-on effects that can potentially destabilise the whole system. Innovations in critical systems may introduce additional vulnerabilities to such attacks: e.g., in smart grids communication channels are needed for distributed intelligent energy management strategies, while simultaneously forming a potential target that could compromise safety [34]. Our research is motivated precisely by the need for safety in these critical systems, which can be achieved by building in robustness against rare but significant deviations caused by one or more system components failing or being compromised in an attack.

In this article we present a new approach for learning policies in such systems that are robust against a chosen scenario of potential attacks or failures. We accomplish this by introducing a new Q-function operator, which we call the κ-operator, that encodes robustness into the bootstrapping update of traditional temporal difference (TD) learning methods. In particular, we design the operator to encode the possibility of significant rare events (SREs) without requiring the learning agent to observe such events in training. Although the κ-operator is model-based with respect to these SREs, it can be combined with any TD method and can thus still be model-free with respect to the environment dynamics.

We prove convergence of our methods to the optimal robust Q-function with respect to the model using the theory of Generalized Markov Decision Processes. In addition we prove convergence to the optimal Q-function of the original MDP given that the probability of SREs vanishes. Empirical evaluations demonstrate the superior performance of κ-based TD methods both in the early learning phase as well as in the final converged stage. In addition we show robustness of the proposed method to small model errors, as well as its applicability to multi-agent joint-action learning.

This articles proceeds with related work in Section 2 and background concepts in Section 3. We formally introduce the new TD operator κ in Section 4 and subsequently prove convergence. Section 6 provides empirical results, and Section 7 concludes.

2 RELATED WORK

The aim to find robust policies is relevant to multiple research areas, including security games, robust control/learning, safe reinforcement learning and multi-agent reinforcement learning.

The domain of **security games** has expanded in recent years with many real-world applications in critical domains [20, 23], where the main approach has been computing exact solutions and deriving strong theoretical guarantees, mostly using equilibria concepts such as Nash and Stackelberg equilibria [14, 17]. In contrast, we base our approach on *reinforcement learning from interactions* with the environment, thus we do not need to know the system model; such an approach to security games has been studied less, exceptions being for example Ruan et al. [22] and Klima et al. [12] who use reinforcement learning in the context of patrolling and illegal rhino poaching problems, respectively. Security games often assume frequent adversarial attack, whereas our work focuses

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

on occasional loss of control over the system, which can represent e.g. failures or adversarial attack. Moreover, our work adopts the information asymmetry assumption often used in Stackelberg Security games [14], providing the model of attack types for the *leader*, and allowing leader-strategy-informed best response strategies by attackers. Similar to security games, control theory starts with a model of the system to be controlled (the *plant*), and for the purpose of **robust control** assumes a set of possible plants as an explicit model of uncertainty, seeking to design a policy that stabilises all these plants [35]. A slightly weaker assumption is made in related work that assumes control over the number of observations for *significant rare events* (SREs), performing updates by sampling from the model [3]. In contrast, our work assumes that the model of this system is not known a priori, and a policy needs to be *learned by interacting* with it. While early work on **robust reinforcement learning** focused on learning within parameterised acceptable policies [26], later work transferred the objective of maximising tolerable disturbances from control theory to reinforcement learning [18]. Our work is similar to the therein defined *Actor-disturber-critic*, but we replace its model of minimax simultaneous actions with stochastic transitions between multiple controllers (one being in control at any time) with arbitrary objectives for each controller. In relation to the taxonomy of **safe reinforcement learning** of Garcıa and Fernández [8] our method falls in between *Worst-Case Criterion under Parameter Uncertainty* and *Risk-Sensitive Reinforcement Learning Based on the Weighted Sum of Return and Risk*, depending on the chosen alternate controller objectives. Our Q(κ) method is comparable to the β-pessimistic Q-learning method of Gaskett [9], however, we propose a more general κ operator of which Q(κ) is only an example. Finally, our approach has commonalities with the **multi-agent reinforcement learning** algorithm Minimax-Q [15] for zero-sum games, which assumes minimisation over the opponent action space. However, in contrast, we define an attack to minimise over our own action space, and thus learn (but not enact) simultaneously our optimal policy and the (rare) attacks it is susceptible to. We further cover not only minimising adversaries but also random failures or any other policy encoding other adversaries' agendas (see Section 4.1).

3 BACKGROUND

This work belongs to the field of reinforcement learning (RL) [29], and makes use of the core concept of a *Markov decision process (MDP)*. An MDP is formally defined by a tuple (S, A, R, P), where S is a finite set of states, A is a finite set of actions, $R(s,a) \rightarrow r \in \mathbb{R}$ is the reward function for a given state $s \in S$ and an action $a \in A$ and $P(s'|s,a)$ is the transition function giving a probability of reaching state s' after taking action a in state s. In this work we also consider a multi-agent setting, which uses the formalism of the *stochastic game*, which generalizes the MDP to multiple agents and is defined by a tuple $(n, S, A_1 \dots A_n, R_1 \dots R_n, P)$, where n is the number of agents and A_i is the action space of agent i. The joint action space is $A = A_1 \cup \dots \cup A_n$, and a joint action is $\boldsymbol{a} = (a_1, a_2, \dots, a_n)$.[1] $R_i(s, a_i, \boldsymbol{a_{-i}}) \rightarrow r_i$ is the reward function of agent i for given state s and joint action $\boldsymbol{a}$, and $P(s'|s, \boldsymbol{a})$ is the state transition function.

[1]We use the common shorthand $\boldsymbol{a_{-i}}$ to denote the joint action of all agents except agent i, i.e., $\boldsymbol{a_{-i}} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$.

The main goal of RL is finding an optimal policy for given MDP. A common method is *temporal difference (TD)* learning, which estimates the value of a state by bootstrapping from the value-estimates of successor states using Bellman-style equations. TD methods work by updating their state-value estimates in order to reduce the *TD error*, which describes the difference between the current estimate of the (state-)value and a new sample obtained from interacting with the environment. In this work we focus on modifying the update *target* of this TD error, which has the standard form of $r + \gamma V(s')$, where γ is the discount factor and $V(s')$ is the current estimate of the next state's value. In *on-policy* methods such as SARSA the target is induced by the actual (behaviour) policy being followed, while *off-policy* methods use an alternative operator (e.g., greedy maximization as in Q-learning). We refer the reader to Sutton and Barto [29] for an overview of RL.

4 THE ROBUST TD OPERATOR κ

Before we formally define our robust TD operator κ, we give an intuitive example. Suppose a Q-learning agent needs to learn a robust policy against a potential malicious adversary who could, with some probability $\varkappa$, take over control in the next state.[2] The value of the next state s_{t+1} thus depends on who is in control: if the agent is in control, she can choose an optimal action that maximizes expected return; or if the adversary is in control he might, in the worst case, aim to minimize the expected return. This can be captured by the following modified TD error

$$\delta_t = r_{t+1} + \gamma \left((1-\varkappa) \max_a Q(s_{t+1}, a) + \varkappa \min_a Q(s_{t+1}, a) \right) - Q(s_t, a_t),$$

where we assume that the agent has knowledge of (or can estimate) the probability $\varkappa$.[3]

In the following we first present a formal, general model of the operator κ, by modifying the target in the standard Bellman style value function. We then present practical implementations of TD(κ) methods that use this operator, for both single- and multi-agent settings, based on the classical on- and off-policy TD learning algorithms (Expected) SARSA and Q-learning.

4.1 Formal Model

We consider a set of m possible *control policies* $C = \{\sigma_1, \dots, \sigma_m\}$. At each time step, one of these policies is in control (and thus decides on the next action) with some probability $p(\sigma_i|s)$ that may depend on the state s. The set C and probability function $p(\cdot)$ are assumed to be (approximately) known by the agent. In our new TD methods, the value of the next state s' then becomes a function of both the state and the function $p(\cdot)$, which we capture in our proposed operator κ as $V^{\kappa}(s')$. Note that the set C includes the focal policy π that we seek to optimise in face of (possibly adversarial) alternative controllers. Such external control policies can represent for example a malicious attacker, aiming to minimize the expected return, or any arbitrary

[2]We use the symbol κ to denote the proposed TD operator and the symbol $\varkappa$ for the parameter denoting the probability of attack.

[3]Note that while a token of control could be included in the state (doubling its size), our approach instead directly applies model-based bootstrap updates. This makes it explicit that the robustness target is a chosen parameter of the operator, and allows to learn robust strategies before observing SREs or when learning during the SREs is not possible. This also highlights the difference between state transition probabilities, which are part of the environment and thus external to the agent, and the expected probability of SREs given by $\varkappa$ which are part of the agent's internal model.

dynamics, such as random failures (e.g. represented by a uniformly random policy). Based on a prior assumption about the nature of σ we want to optimise the focal policy π without necessarily observing actual attacks or failures. This means learning our robust policy π right from the start.

We define σ in terms of our own Q-value function, for example an attacker that is minimising our expected return. Thus we need to learn only one Q-value function Q^π. This is similar to the standard assumption in Stackelberg games that the attacker is able to fully observe our past actions and thus can enact the informed best response. We define the Q-value function update for our policy π based on standard Bellman equation and given the operator κ as

$$Q^\pi(s,a) \leftarrow Q^\pi(s,a) + \alpha\Big[\underbrace{r + \gamma V^\kappa(s')}_{\text{target}} - Q^\pi(s,a)\Big]. \tag{1}$$

Note that where in the standard Bellman equation we would have $V^\pi(s) = \sum_a \pi(s,a)Q^\pi(s,a)$, in our case we have

$$V^\kappa(s) = \sum_{\sigma \in C} p(\sigma|s) \sum_a \sigma(s,a) Q^\pi(s,a), \tag{2}$$

computed as a weighted sum over all possible control policies $\sigma \in C$. Note that we can learn Q^π without actually experiencing any attack or malfunction, based only on prior assumptions about the possible control policies as captured by the operator κ. We refer to this target modification as the operator κ because it closely resembles the Bellman optimality operator $\mathcal{T}^\star$, which is defined as $\mathcal{T}^\star V(s) = \max_a \left[R(s,a) + \sum_{s'} P(s'|s,a)\gamma V(s')\right]$. Thus, we can then formally define the κ optimality operator $\mathcal{T}_\kappa^\star$ by substituting the value function $V(\cdot)$ with $V^\kappa(\cdot)$.

In the following we present several κ-versions of classical TD methods. For simplicity we assume a scenario in which we have only a single adversarial external policy σ that aims to minimize our value, and thus $C = \{\pi, \sigma\}$. Note however that our model is general, and would work for any C and $p(\cdot)$.

4.2 Examples of TD(κ) Methods

We first present single-agent κ-based learning methods by building on the standard TD methods Q-learning and Expected SARSA. Then we present two-agent joint-action learning approaches. Although a generalization to n agents is relatively straightforward, we choose to focus solely on the single- and two-agent case in this paper for clarity of exposition. In each case, we consider the setting in which either the focal agent, with policy π, is in control, or the external adversary with policy σ aiming to minimize return. We further simplify the model by making the control policy probability function $p(\cdot)$ state-independent, reducing it to a probability vector.

4.2.1 Single-Agent Methods. Before we present the algorithms, it is important to note that we need to distinguish the *target* and *behaviour* policies. The κ-operator is defined on the target (see Eq. (1)), while the behaviour policy is used only for selecting actions. We assume an ϵ-greedy behaviour policy throughout.

In **off-policy Q(κ)**, the target policy is the greedy policy $\pi(s) = \arg\max_a Q(s,a)$ that maximizes expected return. The adversarial policy on the other hand aims to minimize the return, i.e., $\sigma(s) = \arg\min_a Q(s,a)$. Assuming a probability of attack of $\varkappa$ as before, we have $p(\pi) = (1-\varkappa)$ and $p(\sigma) = \varkappa$. Thus, Eq. (2) becomes

$$V^\kappa(s) = (1-\varkappa)\max_a Q(s,a) + \varkappa \min_a Q(s,a).$$

For **on-policy Expected SARSA(κ)** the target is the (expectation over the) focal policy π, while the adversarial policy σ remains the same as before. Thus, we have

$$\begin{aligned} V^\kappa(s) &= (1-\varkappa)\mathbb{E}_{a\sim\pi}\left[Q(s,a)\right] + \varkappa \min_a Q(s,a) \\ &= (1-\varkappa)\sum_a \pi(a|s)Q(s,a) + \varkappa \min_a Q(s,a). \end{aligned}$$

4.2.2 Multi-Agent Methods. We move from a single-agent setting to a scenario in which multiple agents interact. For sake of exposition we only present a two-agent case with different action spaces, A_1 and A_2, but an identical reward function and thus a shared joint action Q-value function $Q : S \times A_1 \times A_2 \to \mathbb{R}$. Moreover, we assume full communication during the learning phase, allowing the agents to take each other's policies into account when selecting the next action.[4] Our algorithms are therefore based on the joint-action learning (JAL) paradigm [4]. We further assume that only one agent can be attacked at each time step.[5] For **multi-agent Q(κ)** we can write Eq. (2) for each individual agent as

$$\begin{aligned} V^\kappa(s) = (1-\varkappa)\max_{A_1}\max_{A_2} Q(s,\langle a_1,a_2\rangle) + \frac{\varkappa}{2}\min_{A_1}\max_{A_2} Q(s,\langle a_1,a_2\rangle) \\ + \frac{\varkappa}{2}\min_{A_2}\max_{A_1} Q(s,\langle a_1,a_2\rangle) \end{aligned}$$

with $a_1 \in A_1$ and $a_2 \in A_2$, representing the scenario in which no attack happens with probability $(1-\varkappa)$, and each agent is attacked individually with probability $\varkappa/2$.[6] Analogously, we can define Eq. (2) for **multi-agent Expected SARSA(κ)** as

$$\begin{aligned} V^\kappa(s) = (1-\varkappa) \quad & \mathbb{E}_{a_1\sim\pi_1, a_2\sim\pi_2}\left[Q(s,\langle a_1,a_2\rangle)\right] \\ + \quad \frac{\varkappa}{2} \quad & \min_{A_1} \mathbb{E}_{a_2\sim\pi_2}\left[Q(s,\langle a_1,a_2\rangle)\right] \\ + \quad \frac{\varkappa}{2} \quad & \min_{A_2} \mathbb{E}_{a_1\sim\pi_1}\left[Q(s,\langle a_1,a_2\rangle)\right] \end{aligned}$$

where we now compute an expectation over the actual policy of the agents that are not attacked, while the attacker is still minimizing.

5 THEORETICAL ANALYSIS

In this section we analyze theoretical properties of the proposed κ-methods. We start by relating the different algorithms to each other in the limit of their respective parameters. Then we proceed to show convergence of both Q(κ) and Expected SARSA(κ) to two different fixed points: (i) to the optimal value function $Q^\star$ of the original MDP in the limit where $\varkappa \to 0$; and (ii) to the optimal robust value function $Q_\kappa^\star$ of the MDP that is *generalized* w.r.t. the operator κ for constant parameter $\varkappa$. Note that *optimality* in this sense is purely induced by the relevant operator. In (i) this is the standard Bellman optimality which maximizes the expected discounted return of the MDP. However, in (ii) we derive optimality in the context of *Generalized MDPs* [31], where optimal simply means the fixed point of a given operator, which can take many forms.

[4] A common practice in cooperative multi-agent learning settings, see e.g., [7, 28].
[5] Although relaxing this assumption is straightforward, we opt to keep it for clarity.
[6] Note the order of the min max, which follows the Stackelberg assumption of an all-knowing attacker who moves last.

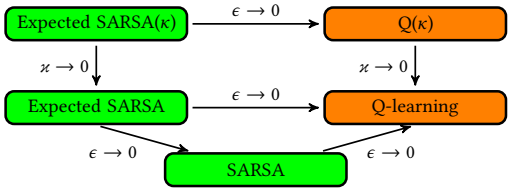

Figure 1: The relationship between the learning targets of different algorithms in the limits of their parameters. On-policy methods are in green, off-policy methods in orange.

Before proceeding with the convergence proofs, Figure 1 summarizes some relationships between the algorithms in terms of their targets, in the limit of their respective parameters: As is known, Expected SARSA, SARSA, and Q-learning become identical in the limit of a greedy policy [29, 33]. Furthermore, the update targets of our κ-methods approach the update targets of the standard TD methods on which they are based as $\varkappa \to 0$. Finally, Expected SARSA(κ) and Q(κ) share the same relationship as their original versions, and thus Expected SARSA(κ) approaches Q(κ) as $\epsilon \to 0$. Note that the algorithms' equivalence in the limit does not hold in the transient phase of the learning process, and hence in practice they may converge on different paths and to different policies that share the same value function. For a comprehensive understanding of the algorithms introduced in Section 4.2, the following sections provide proofs for both convergence of κ methods for $\varkappa \to 0$, as well as their convergence when $\varkappa$ stays constant.[7]

5.1 Convergence to the Optimal $Q^\star$

There exist several proofs of convergence for the temporal difference algorithms Q-learning [11, 32], SARSA [25], and Expected SARSA [33]. Each of these proofs hinges on linking the studied algorithm to a stochastic process, and then using convergence results from stochastic approximation theory [6, 21]. These proofs are based on the following lemma, presented as Theorem 1 in Jaakkola et al. [11] and as Lemma 1 in Singh et al. [25]. These differ in the third condition, which describes the contraction mapping of the operator. The contraction property used for the Q-learning proof [11] has the form $||\mathbb{E}\{F_t(\cdot)|P_t\}|| \leq \gamma||\Delta_t||$, where $\gamma \in [0, 1)$. We show the lemma as it was used for the SARSA proof provided by Singh et al. [25], who show that the contraction property does not need to be strict; strict contraction is required to hold only asymptotically.

Lemma 5.1. *Consider a stochastic process $(\alpha_t, \Delta_t, F_t), t \geq 0$, where $\alpha_t, \Delta_t, F_t : X \to \mathbb{R}$ satisfy the equations*

$$\Delta_{t+1}(x) = \big(1 - \alpha_t(x)\big)\Delta_t(x) + \alpha_t(x)F_t(x),\ x \in X,\ t = 0, 1, 2, \ldots$$

Let P_t be a sequence of increasing σ-fields such that α_0 and Δ_0 are P_0-measurable and α_t, Δ_t and F_{t-1} are P_t-measurable, $t = 1, 2, \ldots$. Then, Δ_t converges to zero with probability one (w.p.1) under the following assumptions:

(1) *the set X is finite,*
(2) $0 \leq \alpha_t(x_t) \leq 1$, $\sum_t \alpha_t(x_t) = \infty$, $\sum_t \alpha_t^2(x_t) < \infty$ *w.p.1,*
(3) $||\mathbb{E}\{F_t(\cdot)|P_t\}|| \leq \gamma||\Delta_t|| + c_t$, *where $\gamma \in [0, 1)$ and c_t converges to zero w.p.1,*
(4) $Var\{F_t(x_t)|P_t\} \leq K(1 + ||\Delta_t||)^2$, *where K is some constant,*

where $||\cdot||$ denotes a maximum norm.

The proof continues by relating Lemma 5.1 to the temporal difference algorithm, following the same reasoning as Van Seijen et al. [33] in their convergence proof for Expected SARSA. We define $X = S \times A$, $P_t = \{Q_0, s_0, a_0, r_0, \alpha_0, s_1, a_1, \ldots, s_t, a_t\}$, $x_t = (s_t, a_t)$, which represents the past at step t and $\alpha_t(x_t) = \alpha_t(s_t, a_t)$ is a learning rate for state s_t and action a_t. To show the convergence of Q to the optimal fixed point $Q^\star$ we set $\Delta_t(x_t) = Q_t(s_t, a_t) - Q^\star(s_t, a_t)$, therefore when Δ_t converges to zero, then the Q values converge to $Q^\star$. The maximum norm $||\cdot||$ can be expressed as maximizing over states and actions as $||\Delta_t|| = \max_s \max_a |Q_t(s, a) - Q^\star(s, a)|$.

We follow the reasoning of Theorem 1 from Van Seijen et al. [33], where we repeat the conditions (1), (2) and (4) and modify the condition **(3)** for the κ methods as:

Theorem 5.2. *Q(κ) and Expected SARSA(κ) as defined in Section 4.2.1 using the respective value function V^κ, defined by*

$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t(s_t, a_t))Q_t(s_t, a_t) + \alpha_t(s_t, a_t)[r_t + \gamma V_t^\kappa(s_{t+1})]$$

converge to the optimal Q function $Q^\star(s, a)$ if:

(1) *the state space S and action space A are finite,*
(2) $\alpha_t(s_t, a_t) \in (0, 1)$, $\sum_t \alpha_t(s_t, a_t) = \infty$ *and* $\sum_t \alpha_t^2(s_t, a_t) < \infty$ *w.p.1,*
(3) *$\varkappa$ converges to zero w.p.1,*
(3*a*) *for Expected SARSA(κ) the policy is greedy in the limit with infinite exploration (GLIE assumption),*
(4) *the reward function is bounded.*

Proof. Convergence of Q(κ): To prove convergence of Q(κ) we have to show that the conditions from Lemma 5.1 hold. Conditions (1), (2) and (4) of Theorem 5.2 correspond to conditions (1), (2) and (4) of Lemma 5.1 [33]. We now need to show that the contraction property holds as well, using condition (3) of Theorem 5.2. Adapting the proof of Van Seijen et al. [33], we set $F_t(x) = F_t(s, a) = r_t(s, a) + \gamma V_t^\kappa(s') - Q^\star(s, a)$ to show that $F_t(s, a)$ is a contraction mapping, i.e., condition (3) in Lemma 5.1. For Q(κ) we write:

$$F_t = r_t + \gamma\big((1-\varkappa)\max_a Q_t(s_{t+1}, a) + \varkappa \min_a Q_t(s_{t+1}, a)\big) - Q^\star(s_t, a_t).$$

We want to show that $||\mathbb{E}\{F_t\}|| \leq \gamma||\Delta_t|| + c_t$ to prove the convergence of Q(κ) to the optimal value $Q^\star$.

$$\begin{aligned}
&||\mathbb{E}\{F_t\}|| \\
&= ||\mathbb{E}\{r_t + \gamma\big((1-\varkappa)\max_a Q_t(s_{t+1}, a) + \varkappa \min_a Q_t(s_{t+1}, a)\big) - Q^\star(s_t, a_t)\}|| \\
&\leq ||\mathbb{E}\{r_t + \gamma \max_a Q_t(s_{t+1}, a) - Q^\star(s_t, a_t)\}|| + \\
&\quad \gamma||\mathbb{E}\{\varkappa \min_a Q_t(s_{t+1}, a) - \varkappa \max_a Q_t(s_{t+1}, a)\}|| \\
&\leq \gamma \max_s |\max_a Q_t(s, a) - \max_a Q^\star(s, a)| + \\
&\quad \gamma \max_s |\varkappa \min_a Q_t(s, a) - \varkappa \max_a Q_t(s, a)| \\
&\leq \gamma||\Delta_t|| + \gamma\varkappa \max_s |\min_a Q_t(s, a) - \max_a Q_t(s, a)|,
\end{aligned}$$

where the first inequality follows from standard algebra and the fact that splitting the maximum norm yields at least as large a number, the second inequality follows from the definition of $Q^\star$ and the maximal difference in values over all states being at least

[7] While we focus on the adversarial targets considered in Section 4.2, a previous proof of convergence under persistent exploration [31] can be interpreted as a model of random failures with fixed kappa.

as large as a difference between values given in state s_{t+1}, and the third inequality follows from the definition of $||\Delta_t||$ above.[8] We can see that if we set $c_t = \gamma\varkappa \max_s |\min_a Q_t(s,a) - \max_a Q_t(s,a)|$, then for $\varkappa \to 0$ we get c_t converging to zero w.p.1, thus proving convergence of Q(κ). ■

Proof. Convergence of Expected SARSA(κ): Similarly as in the proof of Q(κ) we need to show that the contraction property holds as well, this time using conditions (3) and (3a) of Theorem 5.2. We first define:

$$F_t = r_t + \gamma\big((1-\varkappa)\sum_a \pi_t(a|s_{t+1})Q_t(s_{t+1},a) + \varkappa\min_a Q_t(s_{t+1},a)\big) - Q^\star(s_t,a_t)$$

and then show the following:

$$\begin{aligned}
&||\mathbb{E}\{F_t\}||\\
&= ||\mathbb{E}\{r_t + \gamma\big((1-\varkappa)\sum_a \pi_t(a|s_{t+1})Q_t(s_{t+1},a) + \varkappa\min_a Q_t(s_{t+1},a)\big) - Q^\star(s_t,a_t)\}||\\
&\le ||\mathbb{E}\{r_t + \gamma\max_a Q_t(s_{t+1},a) - Q^\star(s_t,a_t)\}||+\\
&\quad \gamma||\mathbb{E}\{(1-\varkappa)\sum_a \pi_t(a|s_{t+1})Q_t(s_{t+1},a) + \varkappa\min_a Q_t(s_{t+1},a) - \max_a Q_t(s_{t+1},a)\}||\\
&\le \gamma\max_s |\max_a Q_t(s,a) - \max_a Q^\star(s,a)|+\\
&\quad \gamma\max_s |(1-\varkappa)\sum_a \pi_t(a|s)Q_t(s,a) + \varkappa\min_a Q_t(s,a) - \max_a Q_t(s,a)|,
\end{aligned}$$

where the inequalities use the same operations as above in the proof of Q(κ). If we set $c_t = \gamma\max_s |(1-\varkappa)\sum_a \pi_t(a|s)Q_t(s,a) + \varkappa\min_a Q_t(s,a) - \max_a Q_t(s,a)|$ and assume that the policy is greedy in the limit with infinite exploration (GLIE assumption) and parameter $\varkappa \to 0$ w.p.1 (conditions (3) and (3a)), it follows that c_t converges to zero w.p.1, thereby proving that Expected SARSA(κ) converges to optimal fixed point $Q^\star$. ■

5.2 Convergence to the Robust $Q^\star_\kappa$

In this section we show convergence to the robust value function $Q^\star_\kappa$ which is optimal w.r.t. the operator κ. The main difference with the proof of Theorem 5.2 is that here we do not require $\varkappa \to 0$ but instead assume it remains constant over time. We base our reasoning on the theory of *Generalized MDPs* [31]. A Generalized MDP is defined using operator-based notation as

$$\Big(\bigotimes\bigoplus(R+\gamma V)\Big)(s) = \max_a \sum_{s'} P(s'|s,a)\big(R(s,a) + \gamma V(s')\big),$$

where the operator $\bigotimes$ defines how an optimal agent chooses her actions (in the classic Bellman equation this denotes maximization) and operator $\bigoplus$ defines how the value of the current state is updated by the value of the next state (in the classic Bellman equation this denotes a probability weighted average over the transition function). These operators can be chosen to model various different scenarios. The generalized Bellman equation can now be written as $V^\star = \bigotimes\bigoplus(R+\gamma V^\star)$. The main result of Szepesvari and Littman [31] is that if $\bigotimes$ and $\bigoplus$ are non-expansions, then there is a unique optimal solution to which the generalized Bellman equation converges, given certain assumptions. For $0 \le \gamma < 1$ and non-expansion properties of $\bigotimes$ and $\bigoplus$ we get a contraction mapping of the Bellman operator $\mathcal{T}$ defined as $\mathcal{T}V = \bigotimes\bigoplus(R+\gamma V)$. Then, the operator $\mathcal{T}$ has a unique fixed point by the Banach fixed-point theorem [27].

Building on the stochastic approximation theory results (as we also used in the Section 5.1), Szepesvari and Littman [31] show the following:

Lemma 5.3. *Generalized Q-learning with operator $\bigotimes$ using Bellman operator $\mathcal{T}_t(Q',Q)(s,a)$ =*

$$\begin{cases}\big(1-\alpha_t(s,a)\big)Q'(s,a) + \alpha_t(s,a)\big(r_t + \gamma(\bigotimes Q)(s'_t)\big) & \text{if } s = s_t, a = a_t\\ Q'(s,a) & \text{otherwise}\end{cases}$$

converges to the optimal Q function w.p.1, if

1. *s'_t is randomly selected according to the probability distribution defined by $P(s_t,a_t,\cdot)$,*
2. *$\alpha_t(s_t,a_t) \in (0,1)$, $\sum_t \alpha_t(s_t,a_t) = \infty$ and $\sum_t \alpha_t^2(s_t,a_t) < \infty$ w.p.1,*
3. *$\bigotimes$ is a non-expansion,*
4. *the reward function is bounded.*

We base our convergence proofs for Q(κ) and Expected SARSA(κ) on the insights of Szepesvari and Littman [31] given in Lemma 5.3.

Theorem 5.4. *Q(κ) and Expected SARSA(κ) as defined in Section 4.2.1 converge to the robust Q function $Q^\star_\kappa$ for any fixed $\varkappa$.*

Proof. Convergence of Q(κ) to $Q^\star_\kappa$: To prove convergence of Q(κ) we follow the proof of Generalized Q-learning in Lemma 5.3. The only condition we need to guarantee is the non-expansion property of the operator in the value function update, which for Q(κ) is a weighted average of the operators *min* and *max*. We write the operator $\bigotimes$ for Q(κ) as $\bigotimes^\kappa$ and define it as

$$(\bigotimes{}^\kappa Q)(s,a) = (1-\varkappa)\max_a Q(s,a) + \varkappa\min_a Q(s,a).$$

In Appendix B of Szepesvari and Littman [31], Theorem 9 states that any linear combination of non-expansion operators is also a non-expansion operator. Moreover Theorem 8 states that the summary operators max and min are also non-expansions. Therefore, $\bigotimes^\kappa$ is a non-expansion as well, thus proving the convergence of Q(κ) to the robust fixed point $Q^\star_\kappa$ induced by the operator κ. ■

Proof. Convergence of Expected SARSA(κ) to $Q^\star_\kappa$: We base our convergence proof of Expected SARSA(κ) again on the work of Szepesvari and Littman [31], this time on their insights regarding persistent exploration (Section 4.5 in their paper). They show that Generalized Q-learning with ϵ-greedy action selection converges, for a fixed ϵ, in the Generalized MDP. Following similar reasoning, we define the operator $\bigotimes$ for Expected SARSA(κ) with fixed ϵ as

$$(\bigotimes{}^\kappa Q)(s,a) = (1-\varkappa)\Big(\epsilon\tfrac{1}{|A|}\textstyle\sum_a Q(s,a) + (1-\epsilon)\max_a Q(s,a)\Big) + \varkappa\min_a Q(s,a).$$

Again, from repeated application of Theorems 8 and 9 in Appendix B of Szepesvari and Littman [31] it follows that $\bigotimes^\kappa$ is a non-expansion as well. Therefore, by Lemma 5.3, Expected SARSA(κ) converges to $Q^\star_\kappa$ for fixed exploration ϵ. ■

It remains an open question whether Expected SARSA(κ) also converges for decreasing ϵ, e.g., under the GLIE assumption, even though we conjecture that it might.

[8] Recall that we set out in this section to show convergence to the same optimal Q-value as classical Q-learning $Q^\star(s_t,a) = r_t + \gamma\max_{a'} Q^\star(s_{t+1},a')$, even if we do so by our new operator.

5.3 Convergence in the Multi-Agent Case

We now prove convergence of the cooperative multi-agent variant of the κ methods presented in Section 4.2.2. This proof builds on the theory of Generalised MDPs, similar to the proofs presented in Section 5.2. Therefore this proof also assumes a fixed probability of attack $\varkappa$. In addition, we make use of the assumption that agents can communicate freely in the learning phase, and thus receive identical information and can build a common joint-action Q-table.

Theorem 5.5. *Multi-agent Q(κ) and Expected SARSA(κ) as defined in Section 4.2.2 converge to the robust Q function $Q^\star_\kappa$ for any fixed $\varkappa$.*

Proof. The $\bigotimes^\kappa$ operator for our multi-agent versions of Q(κ) and Expected SARSA(κ) consists of a nested combination of different components, in particular $\max_a Q(s,a)$, $\min_a Q(s,a)$, and $\sum_a \pi^\epsilon(s,a)Q(s,a)$ where π^ϵ is the ϵ-greedy policy. By Theorem 8 of Szepesvari and Littman [31], max and min are non-expansions. By Theorem 9 of [31], linear combinations of non-expansion operators are also non-expansion operators. Finally, by Theorem 10 of [31], products of non-expansion operators are also non-expansion operators. Therefore, also max max, max min, and min max are non-expansion operators, as are linear combinations of those compounds. Similarly, $\sum_a \pi^\epsilon(s,a)Q(s,a)$ for fixed ϵ can be written as a linear combination of summary operators, which by Theorems 8 and 9 of Szepesvari and Littman [31] is a non-expansion. Therefore, the $\bigotimes^\kappa$ operator used in both multi-agent Q(κ) and Expected SARSA(κ) is a non-expansion. Thus, by Lemma 5.3, Q(κ) and Expected SARSA(κ) converge to $Q^\star_\kappa$ for fixed κ, and in the case of Expected SARSA(κ), for fixed ϵ. ∎

6 EXPERIMENTS AND RESULTS

In this section we evaluate temporal difference methods with the proposed operator κ; off-policy type of learning Q(κ) and on-policy type of learning Expected SARSA(κ). We experiment with a classic cliff walking scenario for the single-agent case and a multi-agent puddle world scenario. Both these domains contain some critical states, a cliff and a puddle respectively, which render very high negative reward for the agent(s) in case of stepping into them. These critical states represent the significant rare events (SREs). We compare our methods to classic temporal difference methods like SARSA, Q-learning and Expected SARSA. In all the experiments we consider an undiscounted (γ=1), episodic scenario.

Cliff Walking: single-agent. The Cliff Walking experiment as shown in Figure 2 is a classic scenario proposed in Sutton and Barto [29] and used frequently ever since (e.g., [33]). The agent needs to get from the start state [S] to the goal state [G], while avoiding stepping into the cliff, otherwise rendering a reward of -100 and sending him back to the start. For every move which does not lead into the cliff the agent receives a reward of -1.

Puddle World: multi-agent. The Puddle World environment is a grid world with puddles which need to be avoided by the joint-action learning agents. The two agents jointly control the movement of a single robot in the Puddle World, each controlling either direction ⟨up, down⟩ or ⟨left, right⟩. Agent 1 can take the actions {*stay, move down, move up*} and agent 2 can choose {*stay, move left, move right, move right by 2*}, thus their action spaces are different, further

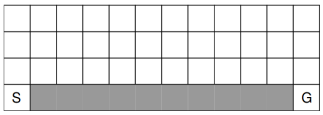

Figure 2: The Cliff Walking: The agent needs to get from the start [S] to the goal [G], avoiding the cliff (grey tiles).

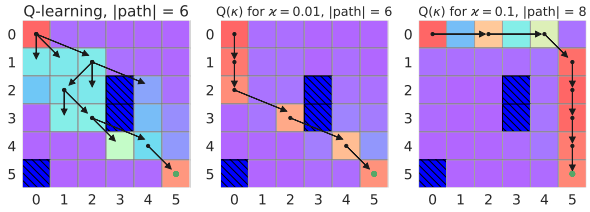

Figure 3: The Puddle World: Q(κ) learns a safer path with increasing $\varkappa$. Puddles are dark blue, the arrows show the optimal actions on the learned path, and the heatmap shows the number of visits to each state (, blue is none).

complicating the learning process compared to the single-agent scenario. The joint action is the combination of the two selected actions. We assume a reward of -1 for every move and -100 for stepping into a puddle (returning to the start node). The agents have to move together from the start node at the top left corner to the goal at the bottom right corner. Figure 3 shows the policy learned by our proposed algorithm Q(κ) for the two joint-learning agents. Note how a safer path (longer, avoiding the puddles) is learned with increasing parameter $\varkappa$ (i.e., higher probability of SREs). For $\varkappa = 0$ our algorithm degenerates to Q-learning (left panel).

6.1 Performance

We replicate the experiment of Van Seijen et al. [33] on the Cliff Walking domain, in which we compare our κ methods with Q-learning, SARSA and Expected SARSA, and perform a similar experiment on the Puddle World domain. In line with [33] we show (i) early performance, which is the average return over the first 100 training episodes, and (ii) converged performance, which is the average return over 100,000 episodes. Figure 4 shows the results for three different settings of both scenarios: (i) a deterministic environment, where each action chosen by the policy is executed with certainty; (ii) an environment with 10% stochasticity, in which a random action is taken with 10% of the time; and (iii) an environment with 10% probability of attack, in which an adversarial action is taken 10% of the time. As before, we define an attack as an action that minimizes the Q-value in the given state. The stochastic environment can be seen as modelling random failures.

The early performance experiments are averaged over 300 trials and the converged performance experiments are averaged over 10 trials. We also show the 95% confidence intervals on all results. We fix the exploration rate to $\epsilon = 0.1$; for the κ methods we set $\varkappa = 0.1$ (later in this section we also experiment with different settings of $\varkappa$). Note that the y-axis, showing the average return, is the same in each row for easy comparison. The x-axis shows different learning rates α. We can see how the average return decreases with more

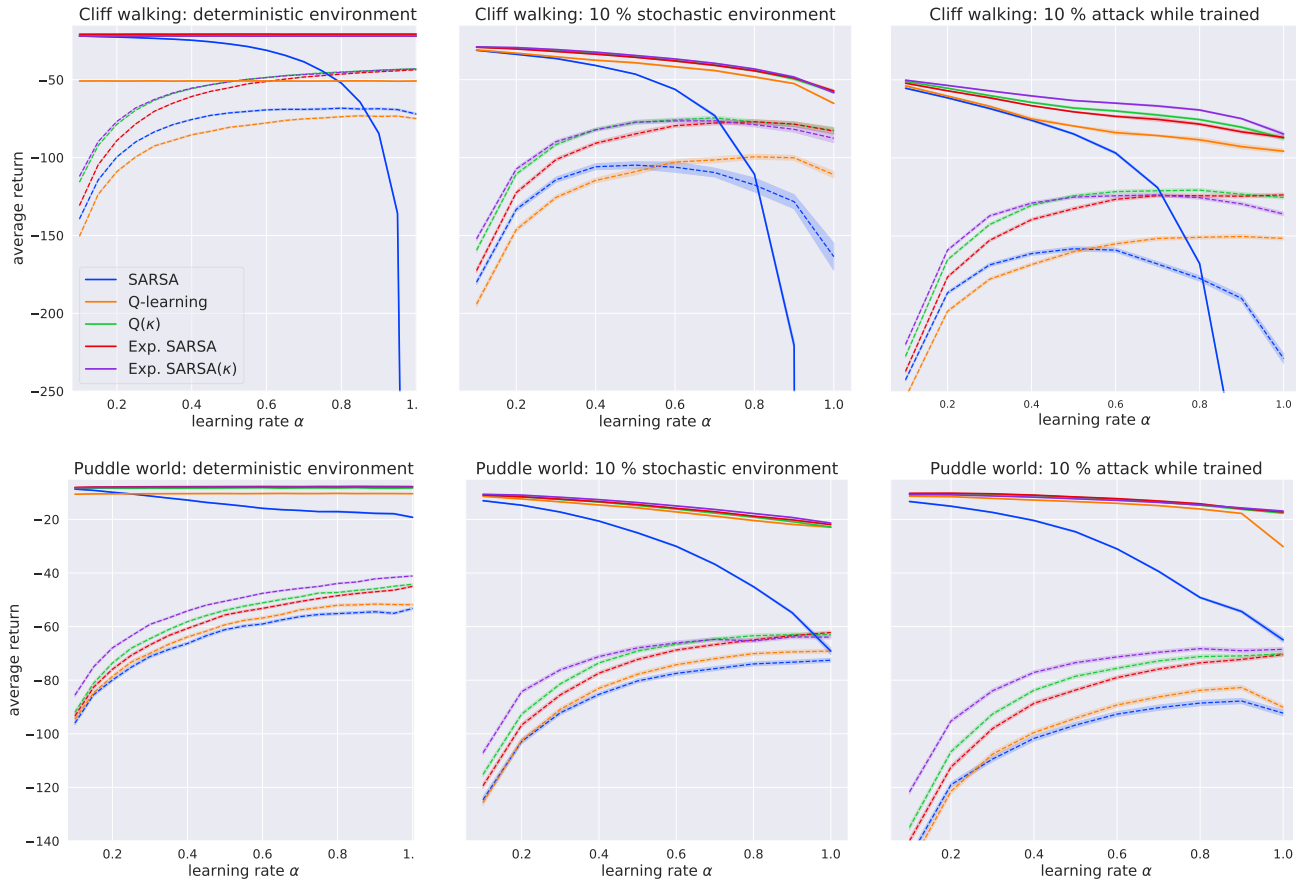

Figure 4: The Cliff Walking (single-agent) in first row and the Puddle World (multi-agent) in second row. Deterministic environment (first column), 10 % stochastic environment (second column) and 10 % attack while training (third column). ϵ-greedy policy with fixed $\epsilon = 0.1$. Early performance - dashed lines (100 episodes), converged performance - solid lines (100, 000 episodes).

complex scenarios, from deterministic, over to stochastic, to one with attacks. The κ methods are superior to the other baselines in the early performance experiments, especially in the attack case, which is the scenario the κ methods are designed for. In the converged performance experiments the κ methods beat Q-learning and SARSA and performs at least as well as Expected SARSA.

6.2 Different Levels of Probability of Attack

In this section we investigate how the methods behave under different levels of attack, defined by the probability of attack per state. We consider an attack on trained (converged) methods, thus we first train each method for 100, 000 episodes (in deterministic environment) and then we test it on 50, 000 trials with given probability of attack per state. We average the results over 10 trials and provide 95% confidence intervals. Note, that this is a different methodology of testing the methods against an adversarial attack compared to the experiments in Figure 4, where we considered attacks during training. This experiment shows the strength of the κ methods for different levels of attacks. We assume the probability of attack to be known here and thus we set the parameter $\varkappa$ to be equal to that probability, which is the meaning of the parameter $\varkappa$ as described before. In other words, parameter $\varkappa$ prescribes how much safely we want to act. We consider very rare attacks (0.001 probability of attack in each state) to more frequent attacks (0.2 probability of attack in each state) as shown in Figure 5. For better visualisation we use logarithmic axes. We train all the methods with fixed exploration rate $\epsilon = 0.1$ and learning rate $\alpha = 0.1$, note that the methods (except SARSA) converge to the same result for different learning rates as shown in left panel of Figure 4. SARSA is very unstable for different learning rates (demonstrated by wide confidence intervals), learns different paths for different α and does not converge fast enough or not at all, which can be partly explained by its higher variance [33]. We test the different levels of probability of attack on the Cliff Walking experiment in the left panel of Figure 5, where we can see that the κ methods compare favourably to the other baselines, however in some parts they give similar performance as Expected SARSA or SARSA. The Cliff Walking experiment has a limited expressiveness for testing the methods due to a limited number of possible safe paths with low costs (see Figure 2), which is the reason for the κ methods to show only similar performance compared to the

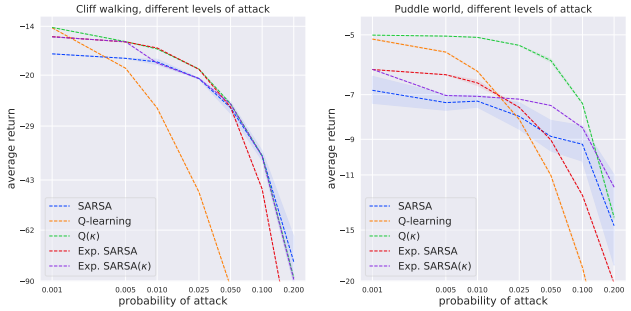

Figure 5: Varying probability of attack: Cliff Walking (left), Puddle World (right), trained 100k, test 50k, $\alpha = 0.1$, $\epsilon = 0.1$.

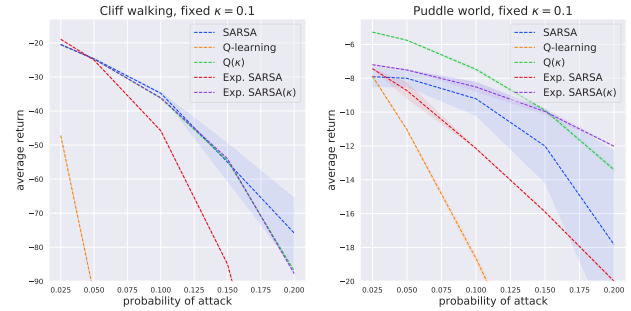

Figure 6: Robustness analysis: Cliff Walking (left), Puddle World (right), trained 100k, test 50k, $\alpha = 0.1$, $\epsilon = 0.1$, $\varkappa = 0.1$.

baselines, not reaching their full potential. However, the Puddle World is more expressive, because there are several possible paths differing in level of safety and cost. The bigger solution space of the Puddle World is also induced by the two cooperating agents, each having their own action space. Therefore, on the right panel of Figure 5 we show the Puddle World experiment for different levels of probability of attack. Here, we can clearly see the κ methods outperform the baselines, especially Q(κ) is superior over the whole range of considered probabilities of attack. Note that Q(κ) learns a safer path even for very rare attacks (0.001), which is also shown in Figure 3, where Q(κ) learns a path with the same cost (distance) compared to Q-learning, but further to the puddles.

6.3 Robustness Analysis

We now test the robustness of the proposed algorithms to an incorrect attack model, meaning that the value of $\varkappa$ in Q(κ) and Expected SARSA(κ) no longer matches the actual probability of attack (in our previous experiments $\varkappa$ matched the actual probability of attack precisely). Figure 6 shows the performance of our algorithms for a range of actual attack probabilities (y-axis) while learning using a fixed parameter $\varkappa = 0.1$. To better highlight the robustness of our methods we choose a range of relatively high actual probabilities of attack around the fixed value of $\varkappa = 0.1$ (note that we no longer use a logarithmic scale). One can see that even when $\varkappa$ is not equal to the actual probability of attack the proposed κ algorithms still outperform the baselines in most cases. In the Cliff Walking experiment (Figure 6 left) the κ methods perform similar to SARSA, however SARSA is quite unstable, as discussed before and as one can see by the width of the confidence interval. The Puddle World experiment (Figure 6 right) demonstrates the superior performance of the κ methods, which beat all the baselines even for the fixed parameter $\varkappa$. These results show that even when we do not know the probability of attack accurately we can learn a more robust strategy using the κ methods.

7 DISCUSSION AND CONCLUSION

We presented a new operator κ for temporal difference learning, which improves robustness of the learning process against potential attacks or perturbations in control. We proved convergence of Q(κ) and Expected SARSA(κ) to (i) the optimal value function $Q^\star$ of the original MDP in the limit where $\varkappa \to 0$; and (ii) the optimal robust value function $Q^\star_\kappa$ of the MDP that is generalized w.r.t. κ for constant parameter $\varkappa$, in both single- and multi-agent versions of the methods. Our complementary empirical results demonstrated that the proposed κ-methods indeed provide robustness against a chosen scenario of potential attacks and failures. Although our method assumes that a model of such attacks and failures is known to the agent, we further demonstrated that our methods are robust against small model errors. Moreover, we have shown that even in absence of attacks or failures, our method learns a policy that is robust in general against environment stochasticity, in particular in the early stages of learning.

There are several interesting directions for future work. The control space can be extended, allowing for more agents being attacked or malfunctioning with different intensity, or with control transitions depending on additional variables other than the state. Furthermore, the target of adversarial policies could be learned from experience using ideas from opponent modelling (e.g. DPIQN [10]). Our proposed operator κ can potentially be combined with some recent state-of-the-art reinforcement learning methods. For example, the operator could be combined with the multi-step Retrace(λ) [19] algorithm, potentially speeding up convergence. Mixed multi-step updates could be introduced by combination with Q(σ) [1], where the parameter σ can also be state-dependent similarly to the control transitions in our model, allowing to learn robust policies against e.g. multi-step attacks. Another interesting extension along this line would be to model the control transition similar to the options framework [2, 30], in which case the alternate control policies could be seen as "malicious" options over which the agent has no control, with potentially complex initiation sets and termination conditions. Such extensions would further increase the flexibility of our proposed operator and narrow the reality gap, making it applicable to a wide range of real-world scenarios.

ACKNOWLEDGMENTS

This project has received funding in the framework of the joint programming initiative ERA-Net Smart Energy Systems' focus initiative Smart Grids Plus, with support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 646039. We are indebted to the anonymous reviewers of AAMAS 2019 for their valuable feedback.

REFERENCES

[1] Kristopher De Asis, J. Hernandez-Garcia, G. Holland, and Richard S. Sutton. 2018. Multi-Step Reinforcement Learning: A Unifying Algorithm. In *AAAI Conference on Artificial Intelligence*.

[2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The Option-Critic Architecture. In *Proceedings of Association for the Advancement of Artificial Intelligence Conference (AAAI)*. 1726–1734.

[3] Kamil Andrzej Ciosek and Shimon Whiteson. 2017. OFFER: Off-Environment Reinforcement Learning. In *Proceedings of Association for the Advancement of Artificial Intelligence Conference (AAAI)*.

[4] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* (1998), 746–752.

[5] Flaviu Cristian, Bob Dancey, and Jon Dehn. 1996. Fault-tolerance in air traffic control systems. *ACM Transactions on Computer Systems (TOCS)* 14, 3 (1996), 265–286.

[6] Aryeh Dvoretzky. 1956. On Stochastic Approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, 39–55.

[7] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[8] Javier Garcıa and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16, 1 (2015), 1437–1480.

[9] Chris Gaskett. 2003. Reinforcement learning under circumstances beyond its control. In *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation*.

[10] Zhang-Wei Hong, Shih-Yang Su, Tzu-Yun Shann, Yi-Hsiang Chang, and Chun-Yi Lee. 2018. A deep policy inference q-network for multi-agent systems. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1388–1396.

[11] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. 1994. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*. 703–710.

[12] Richard Klima, Karl Tuyls, and Frans Oliehoek. 2016. Markov Security Games: Learning in Spatial Security Problems. *NIPS Workshop on Learning, Inference and Control of Multi-Agent Systems* (2016), 1–8.

[13] John C Knight. 2002. Safety critical systems: challenges and directions. In *Proceedings of the 24th International Conference on Software Engineering*. ACM, 547–550.

[14] Dmytro Korzhyk, Zhengyu Yin, Christopher Kiekintveld, Vincent Conitzer, and Milind Tambe. 2011. Stackelberg vs. Nash in Security Games: An Extended Investigation of Interchangeability, Equivalence, and Uniqueness. *Journal of Artificial Intelligence Research* 41 (2011), 297–327.

[15] Michael L. Littman. 1994. *Markov games as a framework for multi-agent reinforcement learning*. Technical Report. Brown University. 157–163 pages.

[16] Jing Liu, Yang Xiao, Shuhui Li, Wei Liang, and CL Philip Chen. 2012. Cyber security and privacy issues in smart grids. *IEEE Communications Surveys & Tutorials* 14, 4 (2012), 981–997.

[17] Jian Lou, Andrew M Smith, and Yevgeniy Vorobeychik. 2017. Multidefender security games. *IEEE Intelligent Systems* 32, 1 (2017), 50–60.

[18] Jun Morimoto and Kenji Doya. 2005. Robust reinforcement learning. *Neural computation* 17, 2 (2005), 335–359.

[19] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. 2016. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*. 1054–1062.

[20] James Pita, Manish Jain, Janusz Marecki, Fernando Ordonez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. 2008. Deployed ARMOR Protection: The Application of a Game Theoretic Model for Security at the Los Angeles International Airport. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Vol. 3. 1805–1812.

[21] Herbert Robbins and Sutton Monro. 1951. A Stochastic Approximation Method. *The Annals of Mathematical Statistics* 22, 3 (1951), 400–407.

[22] Sui Ruan, Candra Meirina, Feili Yu, Krishna R Pattipati, and Robert L Popp. 2005. *Patrolling in a stochastic environment*. Technical Report. Electrical and Computer Engineering Department, University of Connecticut, Storrs.

[23] Eric Shieh, Bo An, Rong Yang, Milind Tambe, Craig Baldwin, Joseph DiRenzo, Ben Maule, and Garrett Meyer. 2012. PROTECT: A Deployed Game Theoretic System to Protect the Ports of the United States. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* 1 (2012), 13–20.

[24] Martin L Shooman. 2003. *Reliability of computer systems and networks: fault tolerance, analysis, and design*. John Wiley & Sons.

[25] Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. 2000. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning* 38, 3 (2000), 287–308.

[26] Satinder P Singh, Andrew G Barto, Roderic Grupen, and Christopher Connolly. 1994. Robust reinforcement learning in motion planning. In *Advances in Neural Information Processing Systems (NIPS)*. 655–662.

[27] D. R. Smart. 1974. *Fixed point theorems*. Cambridge University Press, Cambridge.

[28] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and Thore Grapel. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296* (2017).

[29] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press, Cambridge, MA.

[30] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112, 1-2 (1999), 181–211.

[31] Csaba Szepesvari and Michael L Littman. 1997. *Generalized Markov Decision Processes: Dynamic-programming and Reinforcement-learning Algorithms*. Technical Report. Brown University.

[32] John N Tsitsiklis. 1994. Asynchronous stochastic approximation and Q-learning. *Machine Learning* 16, 3 (1994), 185–202.

[33] Harm van Seijen, Hado van Hasselt, Shimon Whiteson, and Marco Wiering. 2009. A theoretical and empirical analysis of Expected Sarsa. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, ADPRL 2009*. 177–184.

[34] Ye Yan, Yi Qian, Hamid Sharif, and David Tipper. 2013. A survey on smart grid communication infrastructures: Motivations, requirements and challenges. *IEEE Communications Surveys & Tutorials* 15, 1 (2013), 5–20.

[35] Kemin Zhou and John Comstock Doyle. 1998. *Essentials of robust control*. Vol. 104. Prentice hall, Upper Saddle River, NJ.