

# Blackwell Online Learning for Markov Decision Processes

Tao Li, Guanze Peng and Quanyan Zhu  
Department of Electrical and Computer Engineering  
New York University  
Brooklyn, USA  
{tl2636, gp1363, quanyan.zhu}@nyu.edu

**Abstract**—This work provides a novel interpretation of Markov Decision Processes (MDP) from the online optimization viewpoint. In such an online optimization context, the policy of the MDP is viewed as the decision variable while the corresponding value function is treated as payoff feedback from the environment. Based on this interpretation, we construct a Blackwell game induced by MDP, which bridges the gap among regret minimization, Blackwell approachability theory, and learning theory for MDP. Specifically, Based on the approachability theory, we propose 1) Blackwell value iteration for offline planning and 2) Blackwell Q-learning for online learning in MDP, both of which are shown to converge to the optimal solution. Our theoretical guarantees are corroborated by numerical experiments.

**Index Terms**—Blackwell approachability, no-regret learning, reinforcement learning, online optimization

## I. INTRODUCTION

Sequential decision making under uncertainty lies in the heart of many real world problems, ranging from portfolio management to robotic control. Many models and methods have been proposed to study the dynamic decision-making process, among which Markov decision process (MDP) [1] and online optimization [2], together with their variants, are most popular and well studied ones. Based on Bellman principle [3], solving an MDP problem relies on the backward induction using dynamic programming, where the future is considered when making decisions. For online settings, where the transition kernel and/or the reward function are unknown, reinforcement learning (RL) algorithms [4] come into play, which more or less are based on dynamic programming idea. Combined with linear or nonlinear function approximators [5], [6], [7], dynamic programming-based RL methods such as Q-learning [8], actor-critic [9] have brought about many empirical successes.

On the other hand, online optimization operates in a forward fashion, where decisions are made based on history, and no information regarding the future is revealed during the process. In online optimization, the solution concept rests on the optimality in hindsight, widely referred to as no-regret property, as we shall introduce in the background. Closely

related to the no-regret idea, Blackwell approachability [10] gives a geometric interpretation of how regret vanishes over time in online decision making. As pointed out in [11], [12], approachability and no regret are equivalent, and we can develop no-regret algorithms based on the geometric intuition of approachability. As we will show in this paper, such connection can be made explicitly by considering a Blackwell game with vector-valued payoffs measuring the regret.

Instead of studying MDP from an online optimization perspective, most prior works focus on online version of MDP, where transition and/or reward are time-varying, referred to as online MDP [13], [14] or non-stationary RL [15], [16]. Under this setting, the no-regret idea plays an important role, and we see no difficulty in extending our approachability framework to these problems as we adopt the online learning viewpoint. Related to our work, [17] also applies the no-regret idea to MDP problems, which provides theoretical guarantees for offline settings. As shown in the paper, the convergence of the proposed method relies on no-absolute-regret algorithms, such as follow-the-regularized-leader (FTRL) with the linear cost. We argue that such no-regret method is a special case of our Blackwell approachability based framework.

In this paper, we take a step toward understanding MDP from the perspective of online optimization. We construct an auxiliary Blackwell game for MDP, so that we can leverage online optimization methods based on regret minimization. Our main contributions include: 1) we give a no-regret value iteration algorithm, based on Blackwell approachability, which we term Blackwell value iteration. We show that this method provides an asymptotic convergence guarantee as classical value iteration in discounted MDP; 2) We extend this idea to RL domain with unknown transition and reward, which accounts for online learning problems. Similar to Q-learning, our proposed method, Blackwell Q-learning, does not require any prior information nor any access to state sampling distribution. Hence, instead of an asynchronous version of value iteration [18], our Blackwell Q-learning is indeed a RL algorithm based approachability idea. To the best of our knowledge, this is the first work that interprets an MDP as a Blackwell game, which leads to provably convergent learning algorithms.

The rest of the paper is organized as follows. We first introduce some preliminaries, including Blackwell approach-

This research is partially supported by awards ECCS-1847056, CNS-1544782, CNS-2027884, and SES-1541164 from National Science of Foundation (NSF), grant W911NF-19-1-0041 from Army Research Office (ARO), and grant DE-NE0008986 from the Department of Energy (DOE).

ability and no-regret idea in Section II. We then move to our proposed methods based on Blackwell approachability in Section III, where we give both value iteration like and Q-learning like algorithms for both offline planning and online learning problems. Our theoretical analysis is supported by numerical examples presented in Section IV. Finally, Section V concludes the paper. Due to the limit of space, we suppress all proofs in the paper and they can be found in the supplementary [19].

## II. BACKGROUND

### A. Markov Decision Process

An infinite-horizon discounted MDP can be characterized by a tuple,  $\langle \mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma \rangle$ , where  $\mathcal{S}$  is the finite state set;  $\mathcal{A}$  is the finite action set;  $\mathbb{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  is the transition probability, and  $\Delta(\mathcal{S}) \subset \mathbb{R}^{|\mathcal{S}|}$  denotes the simplex over  $\mathcal{S}$ ;  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is the reward function; and  $\gamma \in (0, 1)$  is the discounting factor.

For a given policy  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ , the total expected reward starting from an initial state  $s \in \mathcal{S}$  is defined as  $V^\pi(s) = \mathbb{E}_{\mathbb{P}, \pi} [\sum_{k=1}^{\infty} \gamma^k r(s_k, a_k)]$ . If we denote  $\pi(s, a)$  the probability of choosing  $a$  at state  $s$ , then with Bellman principle [3],  $V^\pi$  can also be written as

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \left[ r(s, a) + \gamma \sum_{s' \sim \mathbb{P}(s, a)} V^\pi(s') \right],$$

where we denote  $Q^\pi(s, a) = r(s, a) + \gamma \sum_{s' \sim \mathbb{P}(s, a)} V^\pi(s')$ , known as the Q function or Q table. The goal is to find an optimal policy  $\pi^*$  such that  $V^{\pi^*}(s) \geq V^\pi(s)$  for all  $s \in \mathcal{S}$ .

Since we focus on finite-dimension problems throughout this paper, all functions introduced above are of finite dimensions. To better present our work, we use the following notations. For  $Q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ ,  $Q(s) := [Q(s, a)]_{a \in \mathcal{A}}$  denotes the vector in  $\mathbb{R}^{|\mathcal{A}|}$ . Similarly, for  $\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ ,  $\pi(s)$  denotes the a vector in  $\Delta(\mathcal{A})$ . Finally, we assume that for every  $s \in \mathcal{S}$  there exists an action  $a \in \mathcal{A}$  such that the Markov chain is aperiodic and irreducible, which is a common assumption in reinforcement learning [4].

### B. Blackwell Approachability

Blackwell approachability theory [10] was developed for studying repeated game play between two players with vector-valued payoffs. In such a repeated game, which we refer to as *Blackwell game*, at the  $k$ -th round, both Player 1 and Player 2 select their actions  $x_k \in \mathcal{X}$  and  $y_k \in \mathcal{Y}$ , and then Player 1 incurs the vector-valued payoff given by  $u(x_k, y_k) \in \mathbb{R}^m$ , where  $u : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^m$  is a bi-affine function. We assume that action sets  $\mathcal{X}, \mathcal{Y}$  are compact and convex. The objective of Player 1 is to guarantee that the average payoff converges to a desired closed convex target set  $\mathcal{D} \in \mathbb{R}^m$ . We let  $d(z, \mathcal{D}) := \inf_{d \in \mathcal{D}} \|z - d\|$  denote the distance between a point  $z \in \mathbb{R}^m$  and the set  $\mathcal{D}$  under norm  $\|\cdot\|$ . If we consider the Blackwell game  $\langle \mathcal{X}, \mathcal{Y}, u, \mathcal{D} \rangle$ , then we can define an approachable set for Player 1 as follows.

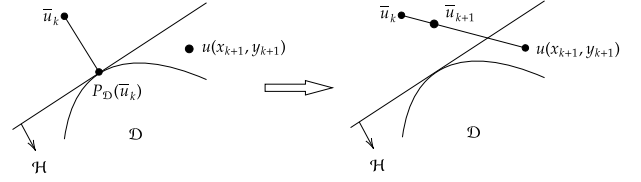


Fig. 1. Blackwell strategy ensures that the next iterate  $u(x_{k+1}, y_{k+1})$  always falls within the halfspace  $\mathcal{H}$ , no matter what  $y_{k+1}$  is.

**Definition 1** (Approachable Set [10]). A set  $\mathcal{D}$  is said to be approachable for Player 1, if there exists an algorithm  $\sigma_k(\cdot) : \mathcal{X}^k \times \mathcal{Y}^k \rightarrow \mathcal{X}$  which chooses an action at each round based on the history of play:  $x_k = \sigma_k(x_{1:k}, y_{1:k})$ , such that for any sequence of  $\{y_k\}_{k=1}^K$ ,  $\lim_{K \rightarrow \infty} d(\frac{1}{K} \sum_{k=1}^K u(x_k, y_k), \mathcal{D}) = 0$ .

A key concept in Blackwell approachability theory is the approachable halfspace, defined as below.

**Definition 2** (Approachable Halfspace [10]). A halfspace  $\mathcal{H} : \{z \in \mathbb{R}^m | a^\top z \leq b\}$  for some  $a \in \mathbb{R}^m, b \in \mathbb{R}$  is approachable for Player 1 if there exists  $x^* \in \mathcal{X}$  such that for all  $y \in \mathcal{Y}$ ,  $u(x^*, y) \in \mathcal{H}$ .

Blackwell's approachability theorem states that  $\mathcal{D}$  is approachable if and only if all halfspaces  $\mathcal{H}$  that contain  $\mathcal{D}$  are approachable. Based on this theorem, we can construct a Blackwell strategy that guarantees the approachability, as shown in [10].

We denote the average payoff up to time  $k$  by  $\bar{u}_k := \sum_{i=1}^k u(x_i, y_i)/k$  and the projection operator regarding the set  $\mathcal{D}$  by  $P_{\mathcal{D}}(z) := \{d \in \mathcal{D} : \|z - d\| = d(z, \mathcal{D})\}$ . Since  $\mathcal{D}$  is a convex set,  $P_{\mathcal{D}}(z)$  returns a singleton. If  $\mathcal{D}$  is approachable, then the halfspace  $\mathcal{H}$  defined by  $\mathcal{H} := \{z : \langle z, \bar{u}_k - P_{\mathcal{D}}(\bar{u}_k) \rangle \leq 0\}$  is approachable. Therefore, there exists  $x^* \in \mathcal{X}$  such that for all  $y$ ,  $u(x^*, y) \in \mathcal{H}$ , and hence, if we let  $x_{k+1} = x^*$ ,  $u(x_{k+1}, y_{k+1})$  falls into the same halfspace as the set  $\mathcal{D}$  does. By doing so, we make  $\bar{u}_{k+1}$  closer to the set, as shown in Fig. 1, and repeating the same procedure at each round, the average payoff converges to  $\mathcal{D}$ .

To better present our idea of leveraging Blackwell approachability and no-regret idea to solve RL problems, we first consider an example of online learning, which is a repeated play between an agent and the nature. We show in the following that this online learning problem can be solved using Blackwell approachability theory. In the online learning process, at each time  $k$ , the player chooses an action from the simplex in  $\mathbb{R}^m$ , that is  $x_k \in \Delta^m$ , while the nature chooses a payoff vector  $y_k \in \mathbb{R}^m$ . The performance of the agent's action  $x_k$  is evaluated according to the revealed payoff  $\langle x_k, y_k \rangle$ , the inner product of  $x_k$  and  $y_k$ . Then, the regret for not having played action  $e_i \in \Delta^m$ , the unit vector, at time  $k$  is given by  $y_k(i) - \langle x_k, y_k \rangle$ , measuring the difference of counterfactual outcomes of  $e_i$  and the received payoff, where  $y_k(i)$  is the  $i$ -th element of the vector  $y_k$ . The objective of the agent is to

have a sequence  $\{x_k\}_k$  that achieves the best possible result:

$$\lim_{K \rightarrow \infty} \frac{1}{K} \max_{x \in \Delta^m} \sum_{k=1}^K (\langle x, y_k \rangle - \langle x_k, y_k \rangle) = 0, \quad (1)$$

showing that the sequence yields the same average performance as the best action in hindsight, and a sequence is said to achieve no regret if it satisfies (1). One way to construct such no-regret sequences is to leverage Blackwell approachability as we show in the following.

We consider a Blackwell game between the agent and the nature  $\langle \Delta^m, \mathbb{R}^m, u, \mathbb{R}^m \rangle$ , where  $u : \Delta^m \times \mathbb{R}^m \rightarrow \mathbb{R}^m$  is

$$\begin{aligned} (x_k, y_k) &\mapsto y_k - \langle x_k, y_k \rangle \mathbf{1}_m \\ &= (y_k(1) - \langle x_k, y_k \rangle, \dots, y_k(m) - \langle x_k, y_k \rangle), \end{aligned}$$

where  $\mathbf{1}_m \in \mathbb{R}^m$  is an all-ones vector. We note that  $u$  measures the regret incurred at time  $k$ . If we adopt Blackwell strategy, we aim to find  $x \in \Delta^m$  such that for all  $y$ ,

$$\langle u(x, y), \bar{u}_k - P_{\mathbb{R}^m_-}(\bar{u}_k) \rangle = \langle u(x, y), [\bar{u}_k]^+ \rangle \leq 0,$$

where  $([u]^+)_i := \max\{u_i, 0\}$ . If we let

$$\mathcal{RM}(x_{1:k}, y_{1:k}) = \begin{cases} [\bar{u}_k]^+ / \|\bar{u}_k\|_1, & \text{if } [\bar{u}_k]^+ \neq 0 \\ \text{any point in } \Delta^m, & \text{otherwise} \end{cases} \quad (\text{RM})$$

then we obtain that, for  $x_{k+1} = \mathcal{RM}(x_{1:k}, y_{1:k})$ ,

$$\begin{aligned} \langle u(x_{k+1}, y), [\bar{u}_k]^+ \rangle &= \langle y - \langle x_{k+1}, y \rangle \mathbf{1}_m, [\bar{u}_k]^+ \rangle \\ &= \langle y, [\bar{u}_k]^+ \rangle - \|\bar{u}_k\|_1 \langle x_{k+1}, y \rangle = 0, \end{aligned}$$

showing that (RM) is indeed a Blackwell strategy. Intuitively, this strategy outputs the next action  $x_{k+1}$  that is proportional to current cumulative positive regret  $[\bar{u}_k]^+$ : actions with larger regret shall be played more frequently, as they bring up better payoffs. Hence, it is also referred to as regret matching (RM), and has been studied in various contexts, including game theory [20], [21] and online optimization [11].

### III. BLACKWELL Q-LEARNING

In this section, we present how to incorporate Blackwell approachability framework into MDP problems through the Blackwell game we introduced above.

#### A. Blackwell Value Iteration: Offline Planing

We first address the planing problem in the infinite-horizon discounted MDP, where the transition kernel and the reward function is known. From the dynamic programming perspective, to solve such a MDP problem, we either resort to value iteration or policy iteration, both of which rely on Bellman optimality principle [3]. In this subsection, we show that under stationary environments, online learning methods, operating in a forward fashion, also guarantee the optimality of the solutions.

As for implementation, we first initialize a Q value table,  $Q_1$ , and an initial policy  $\pi_1(s)$  for each  $s \in \mathcal{S}$ . We run  $|\mathcal{S}|$  copies of the algorithm; i.e., one for each state  $s \in \mathcal{S}$ , and the rewards  $r(s, a)$  are revealed for all  $a \in \mathcal{A}$ , which further

translates to the payoffs in the online learning problem induced by the MDP. In this algorithm, for each state  $s \in \mathcal{S}$ , we view the policy  $\pi_k(s) \in \Delta(\mathcal{A})$  as the decision variable and  $Q_k(s) = (Q_k(s, a))_{a \in \mathcal{A}}$  as the payoff vectors, which is obtained by

$$Q_k(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s, a), a' \sim \pi_{k-1}} [Q_{k-1}(s', a')]. \quad (2)$$

In this case, the payoff of the decision  $\pi_k(s)$  is given by  $\langle \pi_k(s), Q_k(s) \rangle$ .

Similar to the Blackwell game in Section II, we can also construct an approachability game for the MDP. For the decision  $\pi_k(s)$  and the feedback  $Q_k(s)$ , we define cost  $\mathcal{R} : \Delta(\mathcal{A}) \times \mathbb{R}^{|\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{A}|}$  as

$$(\pi_k(s), Q_k(s)) \mapsto Q_k(s) - \langle \pi_k(s), Q_k(s) \rangle \mathbf{1}_{|\mathcal{A}|}.$$

We note that for a given  $\pi, Q$ , the  $i$ -th entry of  $\mathcal{R}(\pi(s), Q(s))$  measures the regret for not having played  $e_{a_i} \in \Delta(\mathcal{A})$ , i.e., simply choosing  $a_i$ . Intuitively, larger  $\mathcal{R}_i$  implies that  $e_{a_i}$  could have given a better payoff, had it been implemented.

It is not surprising that when using the Blackwell strategy such as (RM):  $\pi_{k+1} = \mathcal{RM}(\pi_{1:k}, Q_{1:k})$ , we drive the averaged regret  $\bar{\mathcal{R}}_n = \frac{1}{n} \sum_{k=1}^n \mathcal{R}(\pi_k(s), Q_k(s))$  to the non-positive orthant  $\mathbb{R}^{|\mathcal{A}|}_-$  in the limit, which implies that no action can produce a positive regret and the limiting point achieves optimality. Since we consider the average regret under the Blackwell framework, our convergence result in proposition 1 is also about the average Q tables.

**Proposition 1.** *Let  $\bar{Q}_n = 1/n \sum_{k=1}^n Q_k$  and  $Q^*$  be the Q table under the optimal policy, then  $\lim_{n \rightarrow \infty} \bar{Q}_n = Q^*$ .*

Similar result has also been shown in [17], where FTRL is utilized. Compared with their approach, our Blackwell approachability-based method is in fact more generic. We argue that for the linear cost considered in that paper i.e.,  $\langle \pi(s), Q(s) \rangle$ , it can be shown that FTRL is equivalent to RM proposed here. The details are included in the supplementary, which is mainly based on the connection between Blackwell approachability and online linear optimization studied in [12]. On the other hand, though RM is probably the most natural Blackwell strategy, it is definitely not the only one. In the supplementary, we claim that various online linear optimizers, including online gradient descent and mirror descent algorithms, all can be leveraged to construct Blackwell strategies, offering much freedom in designing algorithms.

#### B. Blackwell Q-learning: Online Learning

Though intuitive and provably convergent, Blackwell value iteration only applies to the offline setting where full information regarding the MDP is known. However, in standard RL problems, the agent is required to find the optimal policy without any access to the transition probability and the reward function. Hence, we need an online version of this approachability-based algorithm.

As pointed out in [17], developing such online learning schemes is not straightforward, and there are two major challenges. The first one is about rewards revelation: in the

Blackwell value iteration, we require that at each iteration, rewards  $r(s, a)$  for all actions  $a \in \mathcal{A}$  at a certain state  $s \in \mathcal{S}$  are revealed for updating the Q table according to (2). However, in RL, since the reward function is unknown, we only have access to the feedback corresponding to the actual action executed at each iteration. Apparently, such a bandit feedback cannot produce the regret vector  $\bar{\mathcal{R}}_k$ . One workaround proposed in [17] is to use importance sampling technique in multi-arm bandit problems [22]. The update rule becomes  $Q_{k+1}(s, a) = (r(s, a) + \gamma \mathbb{E}_{\pi_k}[Q_k(s', a')]) / \pi_k(s, a)$  if  $a$  is an action sampled from  $\pi_k(s)$  for state  $s$ , and  $Q_{k+1}(s, a) = 0$  otherwise. Unfortunately, as in bandit problems, the incorporation of importance sampling can only ensure expectation convergence, a weaker guarantee than almost sure convergence in Q-learning, which is less desirable in practice.

The other challenge is that the asynchronous update in online learning is more involved than the synchronous one. In Blackwell value iteration, every state is updated at every iteration, whereas in an online setting, this synchronous update is impossible, which introduces additional complexity to the convergence analysis. Different from the synchronous update, in the online setting, it is highly likely that some states are visited more frequently than others. One straightforward remedy is to require all states to be visited with the same frequency, i.e., the state to be updated at each iteration is chosen uniformly from  $\mathcal{S}$ . With this additional condition, the asynchronous version of value iteration still guarantees the convergence as shown in [17], though it still falls within the realm of offline planning as the state transition, instead of fixed, is influenced by the chosen actions in online settings.

In this subsection, we propose an online learning scheme based on the Blackwell approachability. We address these issues by the two-time scale asynchronous stochastic approximation [23]. By leveraging the Lyapunov stability theory of differential inclusion developed in [24], [25], we show that adopting the Blackwell strategy in the asynchronous update gives a provably convergent online learning scheme for tackling RL problems.

As we have discussed above, in the online setting, the state transition is influenced by executed actions, sampled from the current policy, which further influences the update of the Q table. This coupled dynamics of the policy update and the Q table update makes it difficult to directly extend value iteration to online learning. One way to decouple the two dynamics is to adjust the timescales of the two. Simply put, we update the Q table in a faster timescale while the policy in a slower one, where the faster timescale sees the slower as quasi-static, while the slower timescale sees the faster as equilibrated. Specifically, we consider the following learning scheme based on the regret matching we have introduced.

Let  $s_{k+1} \in \mathcal{S}$  and  $a_{k+1} \in \mathcal{A}$  be the state and the action visited at time  $k + 1$ , and the agent receives a noised reward  $R_{k+1}$  from the environment. We assume that  $R_{k+1}$  is unbiased in the sense that for the  $\sigma$  fields  $\mathcal{F}_{k+1} = \sigma\{a_{1:k+1}, s_{1:k+1}, R_{1:k+1}\}$ ,  $\mathbb{E}[R_{k+1} | \mathcal{F}_{k+1}] = r(s_{k+1}, a_{k+1})$ . Since states and actions are visited asyn-

chronously, the agent need the asynchronous counters  $\phi_k(s, a) := \sum_{i=1}^k \mathbb{1}_{\{(s_i, a_i)=(s, a)\}}$ ,  $\psi_k(s) := \sum_{i=1}^k \mathbb{1}_{\{s_i=s\}}$ , together with the step sizes  $\{\alpha(k)\}_{k \in \mathbb{N}}$ ,  $\{\beta(k)\}_{k \in \mathbb{N}}$ , to determine the learning rates. Based on all the above, the agent estimates the Q function by (3), and then updates its policy by (4). Finally, the agent chooses an action based on regret matching idea in (RM), i.e, sampling an action from the probability proportional to  $[\mathcal{R}(\pi_k(s), Q_k(s))]^+$ , which we denote by  $\mathcal{RM}(\pi_k(s), Q_k(s))$  with abuse of notations. We summarize the scheme in the following: for every  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$ ,

$$Q_{k+1}(s, a) = Q_k(s, a) + \alpha(\phi_{k+1}(s, a)) \cdot \mathbb{1}_{\{(s, a)=(s_{k+1}, a_{k+1})\}} [R_{k+1} + \gamma V_k(s_{k+2}) - Q_k(s, a)], \quad (3)$$

$$\pi_{k+1}(s) = \pi_k(s) + \beta(\psi_{k+1}(s)) \mathbb{1}_{\{s=s_{k+1}\}} [e_{a_{k+1}} - \pi_k(s)], \quad (4)$$

$$a_{k+1} \sim \mathcal{RM}(\pi_k(s), Q_k(s)), \quad (5)$$

where  $V_k(s) = \sum_{a \in \mathcal{A}} \pi_k(s, a) Q_k(s, a)$ , and  $e_{a_{k+1}}$  is the unit vector in  $\mathbb{R}^{|\mathcal{A}|}$ . We note that different from Blackwell value iteration, we here do not rely on all historical  $\pi_k$  and  $Q_k$  as our stochastic approximation schemes (3) and (4) already return averaged results.

It is clear that (3) and (4) are coupled as  $V_k$  involves both  $\pi_k$  and  $Q_k$ . Technically speaking, in order to analyze the limiting behavior of the coupled dynamics, we must “decouple” them, and one possible approach as proposed in [23], [9] is to adjust the timescales. Specifically, we require that  $\beta(k) = o(\alpha(k))$ , meaning that the Q update (3) operates at a faster timescale than the policy one (4). Intuitively speaking, when synchronous update (2) becomes impossible in the online setting, in order to produce a feedback  $Q_k$  that can approximately evaluate the current policy  $\pi_k$ , we must wait until  $Q_k$  stabilizes before we update the policy. By running the policy update at a slow timescale, Q updates see  $\pi_k$  as quasi-static, and hence (3) can be viewed as expected SARSA [26], while policy update sees  $Q_k$  as stabilized, serving as an approximation to  $Q^{\pi_k}$ . In the subsequent, we show that the two timescale stochastic approximation indeed converges.

*1) Convergence of the fast timescale:* Following the multiple timescale stochastic approximation framework developed in [27], [23], we first show that under a fixed policy, (3) converges. Let  $\mathcal{T} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  be an operator, whose  $(s, a)$  entry is given by  $\mathcal{T}_{(s, a)}(\pi, Q) := r(s, a) + \gamma \sum_{a \in \mathcal{A}} \pi(s, a) Q(s, a)$ , and then we define a vector  $\Gamma_{k+1} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ , whose  $(s, a)$  entry is

$$\Gamma_{k+1}(s, a) = \mathbb{1}_{\{s=s_{k+1}, a=a_{k+1}\}} [R_{k+1} + \gamma V_k(s_{k+2}) - \mathcal{T}_{(s_{k+1}, a_{k+1})}(\pi_k, Q_k)].$$

Note that (3) is equivalent to

$$Q_{k+1}(s, a) - Q_k(s, a) = \alpha(\phi_{k+1}(s, a)) \mathbb{1}_{\{(s, a)=(s_{k+1}, a_{k+1})\}} \cdot [\mathcal{T}_{(s, a)}(\pi_k, Q_k) - Q_k(s, a) + \Gamma_{k+1}(s, a)].$$



We further define the asynchronous step sizes  $\bar{\alpha}_k$  and the relative step sizes  $\mu_k(s, a)$  as

$$\bar{\alpha}_k := \max_{(s,a)} \alpha(\phi_k(s, a)), \quad \mu_k(s, a) := \alpha(\phi_k(s, a)) / \bar{\alpha}_k.$$

By letting  $M_k$  be the  $|S||A| \times |S||A|$  diagonal matrix whose  $(s, a)$  entry is given by  $\mu_k(s, a)$ , we can rewrite the asynchronous update (3) as

$$Q_{k+1} - Q_k = \bar{\alpha}_{k+1} M_{k+1} [\mathcal{T}(\pi_k, Q_k) - Q_k + \Gamma_{k+1}].$$

Given a sequence  $\{Q_k\}_{k \in \mathbb{N}_+}$  produced by (3), based on the linear interpolation of the sequence, we can construct a continuous time process  $\{\bar{Q}_t\}_{t \in \mathbb{R}_+}$ , where the subscript  $t$  denotes the time variable. Let  $\tau_0 = 0$  and  $\tau_k = \sum_{i=1}^k \bar{\alpha}_i$ , we define

$$\bar{Q}_t := Q_k + (t - \tau_k) \frac{Q_{k+1} - Q_k}{\tau_{k+1} - \tau_k}, \quad t \in [\tau_k, \tau_{k+1}). \quad (6)$$

This continuous time process enjoys the same asymptotic properties as  $\{Q_k\}_{k \in \mathbb{N}_+}$  does, and hence, the convergence result of (3) can be studied by analyzing  $\bar{Q}_t$ . As shown in the following proposition,  $\bar{Q}_t$  is an asymptotic pseudo-trajectory of a differential inclusion (see [24, Proposition 1.3]), and its limiting behavior can be studied by Lyapunov stability theory.

**Proposition 2.** [27, Theorem 3.1] *There exists  $0 < \eta < 1$  such that almost surely, the linear interpolation  $\{\bar{Q}_t(s)\}_{t \in \mathbb{R}_+}$  in (6) is an asymptotic pseudo-trajectory to the differential inclusion,*

$$d\bar{Q}_t(s)/dt \in \Omega^\eta[\mathcal{T}_s(\pi_t, \bar{Q}_t) - \bar{Q}_t], \quad (7)$$

where

$$\Omega^\eta := \{\text{diag}(\{\omega(s)\}_{s \in \mathcal{S}}) : \omega(s) \in [\eta, 1], \forall s \in \mathcal{S}\}. \quad (8)$$

**Proposition 3.** *When  $\pi$  is fixed,  $Q^\pi$  is the unique global attractor of (7).*

2) *Convergence of the slow timescale:* From the discussions above, the sequence of Q tables  $\{Q_k\}_k$  converges to the  $Q^\pi$  when  $\pi$  is fixed, and it is Lipschitz continuous in  $\pi$  [27]. Therefore, we can study the limiting behavior of (4) by analyzing its continuous counterpart in (9). We here replace the Q table in (5) with the attractor  $Q^{\pi_t}$  under the current policy  $\pi_t$ , as the slow timescale views the Q updates as stabilized.

$$d\pi_t(s)/dt \in \Omega^{\eta'}[\mathcal{RM}(\pi_t, Q^{\pi_t}) - \pi_t(s)], \quad (9)$$

where  $\Omega^{\eta'}$  is defined similarly as (8). It remains to show that the differential inclusion (9) has a global attractor, which we prove by standard Lyapunov argument. Moreover, we shall present that such a global attractor is indeed the set of optimal policy. In the following lemma, we identify the Lyapunov function associated with (9)

**Lemma 1.** *For every  $s \in \mathcal{S}$  and any fixed  $\omega(s)$  in  $\Omega^{\eta'}$ , let  $d\pi_t(s)/dt = \omega(s)[\hat{\pi}_t(s) - \pi_t(s)]$ ,  $\hat{\pi}_t(s) = \mathcal{RM}(\pi_t(s), Q^{\pi_t}(s))$ , then  $\langle \nabla_{\pi_t} V^{\pi_t}(s), d\pi_t(s)/dt \rangle \geq 0$ .*

With this lemma, we now construct the Lyapunov function for (9), which further leads to the global convergence of the

algorithm. First, given  $\pi^*$ , an optimal policy, we define  $L(\pi) = \sum_{s \in \mathcal{S}} [V^{\pi^*}(s) - V^\pi(s)]$ . Apparently  $L(\pi)$  is a positive semi-definite function, since the optimality gives  $V^{\pi^*}(s) - V^\pi(s) \geq 0$  for all  $s \in \mathcal{S}$ , and  $L(\pi) = 0$  only if  $\pi$  is an optimal policy. Then with lemma 1, for any  $t > 0$ , we have

$$\langle \nabla_{\pi_t} L(\pi_t), d\pi_t/dt \rangle = - \sum_{s \in \mathcal{S}} \langle \nabla_{\pi_t} V^{\pi_t}(s), d\pi_t(s)/dt \rangle \leq 0.$$

This implies that  $L(\pi)$  is a Lyapunov function for the differential inclusion (9), with a global attractor  $\Pi = \{\pi : \pi \text{ is an optimal strategy}\}$ , showing that  $\pi_t$  given by (9) converges almost surely to the attractor. Therefore, from the convergence result of the continuous dynamics, we claim the convergence of the coupled dynamics (3), (4).

**Proposition 4.** *The sequence  $\{Q_k, \pi_k\}_k$  given by the coupled recursive scheme (3) and (4) converges almost surely to  $(Q^{\pi^*}, \pi^*)$ , where  $\pi^*$  is an optimal policy and  $Q^{\pi^*}$  is the associate optimal Q function.*

#### IV. NUMERICAL EXPERIMENTS

In this section, we present experimental results when applying our Blackwell Q-learning to MDP problems. Since our proposed method resembles expected SARSA [26], we consider cliff walking task in that paper, where the agent has to find its way from the start to the goal in a grid world. The agent can take any of four-movement actions: up, down, left, and right, each of which moves the agent one square in the corresponding direction. Each step results in a reward of -1, except when the agent steps into the cliff area, which results in a reward of -100 and an immediate return to the start state. The episode ends upon reaching the goal state.

We evaluate the performance of Q-learning, SARSA, expected SARSA, and our Blackwell Q-learning. It is noted that our Blackwell Q-learning does not need any hyperparameter for encouraging exploration, as (RM) always retains some probabilities for actions that yield positive regret. Hence, our method is less aggressive in terms of exploitation, compared with the others. In our experiments, we adopt  $\epsilon$ -greedy policy for the first three, where  $\epsilon = 0.1$  for Q-learning and SARSA, and for expected SARSA, we run the algorithm with two different exploration rates  $\epsilon = 0.1, 0.5$ . We test these algorithms with 2000 episodes and we average the results over 200 independent runs. The numerical results are shown in Fig. 2. At first glance, both expected SARSA with  $\epsilon = 0.1$  and our Blackwell Q-learning give the best performance in the end, though the expected SARSA converges faster, due to the greedy policy. However, we note that the success of the expected SARSA relies on a carefully crafted exploration rate. If we set  $\epsilon = 0.5$ , then the performance is even worse than that of SARSA. This observation highlights one merit of Blackwell Q-learning: it is hyperparameter free for exploration.

Though in our experiments, Blackwell Q-learning seems not to outperform expected SARSA in terms of the convergence rate, because of the difference in action selection, we argue that such conservative action selection is actually more desired

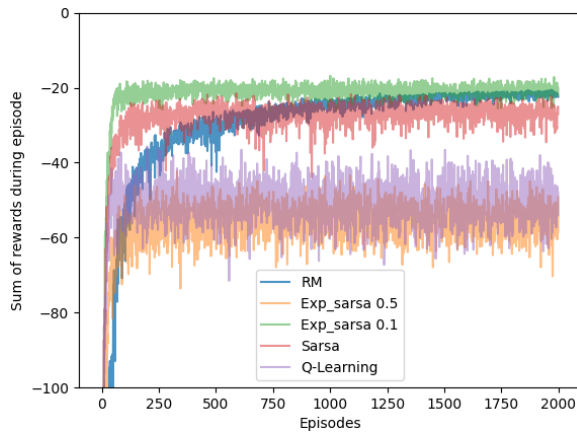


Fig. 2. Comparison between different learning methods in cliff walking experiments: RM, Expected SARSA, SARSA and  $Q$ -learning.

for online learning problem, where the environment is non-stationary. One prominent example is learning in games [28], where the payoff is jointly determined by all players' actions. In this case, if one player only seeks the best response based on his own  $Q$  function, he may not achieve any equilibrium in the end, as observed in [29]. Due to the limited space, we fully develop our arguments in the RM supplementary.

## V. CONCLUSION

We have introduced a novel approach for tackling MDP problems based on the Blackwell approachability theory. By constructing an auxiliary Blackwell game, we use its geometric interpretation to solve MDP problems by deriving no-regret learning from the Blackwell strategy, which provides an alternative to dynamic programming for MDP. Specifically, we have discussed one simple Blackwell strategy, regret matching, and how it can be incorporated into both offline planning methods (e.g., Blackwell value iteration) and online learning schemes (e.g. Blackwell  $Q$ -learning). Both are provably convergent. Related numerical results have been used to corroborate our results. As for future work, we would like to extend our Blackwell approachability-based idea to online (adversarial) MDP [14] and multi-agent systems, where the environment is not stationary from any player's perspective, hence imposing difficulties on applying dynamic programming.

## REFERENCES

- [1] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1st ed., 1994.
- [2] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.
- [3] R. Bellman, "The theory of dynamic programming," *Bulletin of the American Mathematical Society*, vol. 60, no. 6, pp. 503–515, 1954.
- [4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, Cambridge, MA: MIT press, 2018.
- [5] J. N. Tsitsiklis and B. V. Roy, "Analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.
- [6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. v. d. Drissi, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–9, 2016.
- [7] T. Li and Q. Zhu, "On Convergence Rate of Adaptive Multiscale Value Function Approximation for Reinforcement Learning," *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2019.
- [8] P. Dayan and C. J. Watkins, "Q-Learning," *Machine Learning*, vol. 8, no. 3–4, p. 279 292, 1992.
- [9] V. R. Konda and V. S. Borkar, "Actor-Critic-Type Learning Algorithms for Markov Decision Processes," *SIAM Journal on Control and Optimization*, vol. 38, no. 1, pp. 94–123, 1999.
- [10] D. Blackwell, "An analog of the minimax theorem for vector payoffs," *Pacific Journal of Mathematics*, vol. 6, no. 1, pp. 1–8, 1956. approachability theorem.
- [11] V. Perchet, "Approachability, regret and calibration: Implications and equivalences," *Journal of Dynamics & Games*, vol. 1, no. 2, pp. 181–254, 2014.
- [12] J. Abernethy, P. L. Bartlett, and E. Hazan, "Blackwell Approachability and No-Regret Learning are Equivalent," in *Proceedings of the 24th Annual Conference on Learning Theory*, vol. 19 of *Proceedings of Machine Learning Research*, pp. 27–46, PMLR, 2011.
- [13] J. Y. Yu, S. Mannor, and N. Shimkin, "Markov Decision Processes with Arbitrary Reward Processes," *Mathematics of Operations Research*, vol. 34, no. 3, pp. 737–757, 2009.
- [14] E. Even-Dar, S. M. Kakade, and Y. Mansour, "Online Markov Decision Processes," *Mathematics of Operations Research*, vol. 34, no. 3, pp. 726–736, 2009.
- [15] E. Lecarpentier and E. Rachelson, "Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 32, pp. 7216–7225, 2019.
- [16] B. C. da Silva, E. W. Basso, A. L. C. Bazzan, and P. M. Engel, "Dealing with non-stationary environments using context detection," in *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, p. 217–224, 2006.
- [17] I. A. Kash, M. Sullins, and K. Hofmann, "Combining No-Regret and Q-Learning," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20*, p. 593–601, 2020.
- [18] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming," Optimization and neural computation series, 1996.
- [19] T. Li, G. Peng, and Q. Zhu, "Blackwell online learning for markov decision processes," *arXiv preprint arXiv:2012.14043*, 2020.
- [20] S. Hart and A. Mas-Colell, "A Simple Adaptive Procedure Leading to Correlated Equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000. regret matching.
- [21] S. Hart and A. Mas-Colell, "Regret-based continuous-time dynamics," *Games and Economic Behavior*, vol. 45, no. 2, pp. 375–394, 2003.
- [22] A. Slivkins, "Introduction to Multi-Armed Bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1–2, pp. 1–286, 2019.
- [23] V. S. Borkar, *Stochastic Approximation, A Dynamical Systems Viewpoint*, vol. 48 of *Springer*. Springer, 2008.
- [24] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic Approximations and Differential Inclusions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.
- [25] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic Approximations and Differential Inclusions, Part II: Applications," *Mathematics of OR*, vol. 31, no. 4, pp. 673–695, 2006.
- [26] H. v. Seijen, H. v. Hasselt, S. Whiteson, and M. Wiering, "A Theoretical and Empirical Analysis of Expected Sarsa," *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pp. 177–184, 2009.
- [27] S. Perkins and D. S. Leslie, "Asynchronous Stochastic Approximation with Differential Inclusions," *Stochastic Systems*, vol. 2, no. 2, pp. 409–446, 2012.
- [28] D. Fudenberg, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.
- [29] M. Zinkevich, A. Greenwald, and M. L. Littman, "Cyclic Equilibria in Markov Games," in *Advances in Neural Information Processing Systems 18*, pp. 1641–1648, MIT Press, 2006.