Reward-Weighted Regression Converges to a Global Optimum

Miroslav Štrupl,[1*] Francesco Faccio,[1*] Dylan R. Ashley,[1]
Rupesh Kumar Srivastava,[2] Jürgen Schmidhuber[1,2,3]

[1] The Swiss AI Lab IDSIA, Università della Svizzera italiana (USI) & SUPSI, Lugano, Switzerland
[2] NNAISENSE, Lugano, Switzerland
[3] King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
{struplm, francesco, dylan.ashley}@idsia.ch, rupesh@nnaisense.com, juergen@idsia.ch

Abstract

Reward-Weighted Regression (RWR) belongs to a family of widely known iterative Reinforcement Learning algorithms based on the Expectation-Maximization framework. In this family, learning at each iteration consists of sampling a batch of trajectories using the current policy and fitting a new policy to maximize a return-weighted log-likelihood of actions. Although RWR is known to yield monotonic improvement of the policy under certain circumstances, whether and under which conditions RWR converges to the optimal policy have remained open questions. In this paper, we provide for the first time a proof that RWR converges to a global optimum when no function approximation is used, in a general compact setting. Furthermore, for the simpler case with finite state and action spaces we prove R-linear convergence of the state-value function to the optimum.

1 Introduction

Reinforcement learning (RL) is a branch of artificial intelligence that considers learning agents interacting with an environment (Sutton and Barto 2018). RL has enjoyed several notable successes in recent years. These include both successes of special prominence within the artificial intelligence community—such as achieving the first superhuman performance in the ancient game of Go (Silver et al. 2016)—and successes of immediate real-world value—such as providing autonomous navigation of stratospheric balloons to provide internet access to remote locations (Bellemare et al. 2020).

One prominent family of algorithms that tackle the RL problem is the Reward-Weighted Regression (RWR) family (Peters and Schaal 2007). The RWR family is notable in that it naturally extends to continuous state and action spaces. The lack of this functionality in many methods serves as a strong limitation. This prevents them from tackling some of the more practically relevant RL problems—such as many robotics tasks (Plappert et al. 2018). Recently, RWR variants were able to learn high-dimensional continuous control tasks (Peng et al. 2019). RWR works by transforming the RL problem into a form solvable by well-studied expectation-maximization (EM) methods (Dempster, Laird, and Rubin 1977). EM methods are, in general, guaranteed to converge to a point whose gradient is zero with respect to the parameters. However, these points could be both local minima or saddle points (Wu 1983). These benefits and limitations transfer to the RL setting, where it has been shown that an EM-based return maximizer is guaranteed to yield monotonic improvements in the expected return (Dayan and Hinton 1997). However, it has been challenging to assess under which conditions—if any—RWR is guaranteed to converge to the optimal policy. This paper presents a breakthrough in this challenge.

The EM probabilistic framework requires that the reward obtained by the RL agent is strictly positive, such that it can be considered as an improper probability distribution. Several reward transformations have been proposed, e.g., Peters and Schaal (2007, 2008); Peng et al. (2019); Abdolmaleki et al. (2018b), frequently involving exponential transformations. In the past, it has been claimed that a positive, strictly increasing transformation $u_\tau(s)$ with $\int_0^\infty u_\tau(r)\,\mathrm{d}r = const$ would not alter the optimal solution for the MDP (Peters and Schaal 2007). Unfortunately, as we demonstrate in Appendix A, this is not the case. The consequence of this is that we cannot rely on those transformations if we want prove convergence. Therefore, we consider only linear transformation of the reward. A possible disadvantage of relying on linear transformations is that it is necessary to know a lower bound on the reward to construct such a transformation.

In this work, we provide the first proof of RWR's global convergence in a setting without function approximation or reward transformations[1]. The paper is structured as follows: Section 2 introduces the MDP setting and other preliminary material; Section 3 presents a closed-form update for RWR based on the state and action-value functions and Section 4 shows that the update induces monotonic improvement related to the variance of the action-value function with respect to the action sampled by the policy; **Section 5 proves global convergence of the algorithm in the general compact setting and convergence rates in the finite setting**; Section 6 illustrates experimentally that—for a simple MDP—the presented update scheme converges to the optimal policy; Section 7 discusses related work; and Section 8 concludes.

[*]Equal contribution. Correspondence to struplm@idsia.ch

[1]Note that—without loss of generality—we do assume here that a linear reward transformation is already provided, such that the reward is positive

2 Background

Here we consider a Markov Decision Process (MDP) (Stratonovich 1960; Puterman 2014) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p_T, R, \gamma, \mu_0)$. We assume that the state and action spaces $\mathcal{S} \subset \mathbb{R}^{n_S}$, $\mathcal{A} \subset \mathbb{R}^{n_A}$ are compact sub-spaces [2] (equipped with subspace topology), with measurable structure given by measure spaces $(\mathcal{S}, \mathcal{B}(\mathcal{S}), \mu_S)$, $(\mathcal{A}, \mathcal{B}(\mathcal{A}), \mu_A)$ where $\mathcal{B}(\cdot)$ denotes the Borel σ-algebra after completion, and reference measures μ_S, μ_A are assumed to be finite and strictly positive on $\mathcal{S}, \mathcal{A}$ respectively. The distributions of state (action) random variables (except in Section 5 where greedy policies are used) are assumed to be dominated by μ_S (μ_A), thus having a density with respect to μ_S (μ_A). Therefore, we reserve symbols $\mathrm{d}s, \mathrm{d}a$ in integral expression not to integration with respect to Lebesgue measure, as usual, but to integration with respect to μ_S and μ_A respectively, e.g. $\int_{\mathcal{S}}(\cdot)\mathrm{d}s := \int_{\mathcal{S}}(\cdot)\mathrm{d}\mu_S(s)$. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to \mathbb{R}^+$ a $\mathcal{F}$ measurable function (density). We denote by $f \cdot \mu$ the measure which assigns to every set $B \in \mathcal{F}$ a measure $f \cdot \mu(B) := \int_B f \mathrm{d}\mu$.

In the MDP framework, at each step, an agent observes a state $s \in \mathcal{S}$, chooses an action $a \in \mathcal{A}$, and subsequently transitions into state s' with probability density $p_T(s'|s, a)$ to receive a deterministic reward $R(s, a)$. The transition probability kernel is assumed to be continuous in total variation in $(s, a) \in \mathcal{S} \times \mathcal{A}$ (the product topology is assumed on $\mathcal{S} \times \mathcal{A}$), and thus the density $p_T(s'|s, a)$ is continuous (in $\|\cdot\|_1$ norm). $R(s, a)$ is assumed to be a continuous function on $\mathcal{S} \times \mathcal{A}$.

The agent starts from an initial state (chosen under a probability density $\mu_0(s)$) and is represented by a stochastic policy π: a probability kernel which provides the conditional probability distribution of performing action a in state s.[3] The policy is deterministic if, for each state s, there exists an action a such that $\pi(\{a\}|s) = 1$. The return R_t is defined as the cumulative discounted reward from time step t: $R_t = \sum_{k=0}^{\infty} \gamma^k R(s_{t+k+1}, a_{t+k+1})$ where $\gamma \in (0, 1)$ is a discount factor. We discuss the undiscounted case ($\gamma = 1$) in Appendix B, which covers the scenario with absorbing states.

The agent's performance is measured by the cumulative discounted expected reward (i.e., the expected return), defined as $J(\pi) = \mathbb{E}_\pi[R_0]$. The state-value function $V^\pi(s) = \mathbb{E}_\pi[R_t|s_t = s]$ of a policy π is defined as the expected return for being in a state s while following π. The maximization of the expected cumulative reward can be expressed in terms of the state-value function by integrating it over the state space $\mathcal{S}$: $J(\pi) = \int_{\mathcal{S}} \mu_0(s) V^\pi(s) \mathrm{d}s$. The action-value function $Q^\pi(s, a)$—defined as the expected return for performing action a in state s and following a policy π—is $Q^\pi(s, a) = \mathbb{E}_\pi[R_t|s_t = s, a_t = a]$. State and action value functions are related by $V^\pi(s) = \int_{\mathcal{A}} \pi(a|s) Q^\pi(s, a) \mathrm{d}a$. We define as $d^\pi(s')$ the discounted weighting of states encountered starting at $s_0 \sim \mu_0(s)$ and following the policy π: $d^\pi(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} \mu_0(s) p_{s_t|s_0,\pi}(s'|s) \mathrm{d}s$, where $p_{s_t|s_0,\pi}(s'|s)$ is the probability density of transitioning to s' after t time steps, starting from s and following policy π. We assume that the reward function $R(s, a)$ is strictly positive[4], so that state and action value functions are also bounded $V^\pi(s) \leq \frac{1}{1-\gamma}||R||_\infty = B_V < +\infty$. We define the operator[5] $W : L_\infty(\mathcal{S}) \to C(\mathcal{S} \times \mathcal{A})$ as $[W(V)](s, a) := R(s, a) + \gamma \int_{\mathcal{S}} V(s') p_T(s'|s, a) \mathrm{d}s'$ and the Bellman's optimality operator $T : L_\infty(\mathcal{S} \times \mathcal{A}) \to C(\mathcal{S} \times \mathcal{A})$ as $[T(Q)](s, a) := R(s, a) + \gamma \int_{\mathcal{S}} \max_{a'} Q(s', a') p_T(s'|s, a) \mathrm{d}s'$. An action-value function Q^π is optimal if it is the unique fixed point for T. If Q^π is optimal, then π is an optimal policy.

3 Reward-Weighted Regression

Reward-Weighted Regression (RWR, see (Dayan and Hinton 1997),(Peters and Schaal 2007),(Peng et al. 2019)) is an iterative algorithm which consists of two main steps. First, a batch of episodes is generated using the current policy π_n (all policies in this section are given as conditional densities with respect to μ_A). Then, a new policy is fitted to (using supervised learning under maximum likelihood criterion) a sample representation of the conditional distribution of an action given a state, weighted by the return. The RWR optimization problem is:

$$\pi_{n+1} = \arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{s \sim d^{\pi_n}(\cdot), a \sim \pi_n(\cdot|s)} \Big[\mathop{\mathbb{E}}_{R_t \sim p(\cdot|s_t = s, a_t = a, \pi_n)} [R_t \log \pi(a|s)] \Big], \quad (1)$$

where Π is the set of all conditional probability densities (meant with respect to μ_A)[6]. Notice that π_{n+1} is defined correctly as its expression does not depend on t. This is equivalent to the following:

$$\pi_{n+1} = \arg\max_{\pi \in \Pi} \mathop{\mathbb{E}}_{s \sim d^{\pi_n}(\cdot), a \sim \pi_n(\cdot|s)} [Q^{\pi_n}(s, a) \log \pi(a|s)]. \quad (2)$$

We start by deriving a closed form solution to the optimization problem:

Theorem 3.1. *Let π_0 be an initial policy and let $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}\ R(s, a) > 0$. At each iteration $n > 0$, the solution of the RWR optimization problem is:*

$$\pi_{n+1}(a|s) = \frac{Q^{\pi_n}(s, a) \pi_n(a|s)}{V^{\pi_n}(s)}. \quad (3)$$

Proof.

$$\pi_{n+1} = \arg\max_{\pi \in \Pi} \int_{\mathcal{S}} d^{\pi_n}(s) \times \int_{\mathcal{A}} \pi_n(a|s) Q^{\pi_n}(s, a) \log \pi(a|s) \mathrm{d}a \mathrm{d}s.$$

[2] This allows for state and action vectors that have discrete, continuous, or mixed components.

[3] In Sections 3 and 4, a policy is given through its conditional density with respect to μ_A. We also refer to this density as a policy.

[4] It is enough to assume that the reward is bounded, so it can be linearly mapped to a positive value.

[5] W maps to continuous functions since $R(s, a)$ is continuous and continuity of the integral follows from continuity of p_T in $\|\cdot\|_1$ norm and boundedness of V.

[6] We can restrict to talk about probability kernels dominated by μ_A instead of all probability kernels thanks to Lebesgue decomposition.

Define $\hat{f}(s,a) := d^{\pi_n}(s)\pi_n(a|s)Q^{\pi_n}(s,a)$. $\hat{f}(s,a)$ can be normalized such that it becomes a density that we fit by π_{n+1}:

$$f(s,a) = \frac{\hat{f}(s,a)}{\int_{\mathcal{S}}\int_{\mathcal{A}}\hat{f}(s,a)\,\mathrm{d}a\,\mathrm{d}s} = \frac{d^{\pi_n}(s)\pi_n(a|s)Q^{\pi_n}(s,a)}{\int_{\mathcal{S}}\int_{\mathcal{A}}d^{\pi_n}(s)\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a\,\mathrm{d}s}.$$

For the function to be maximized we have:

$$\int_{\mathcal{S}}\int_{\mathcal{A}} f(s,a)\log\pi(a|s)\,\mathrm{d}a\,\mathrm{d}s = \int_{\mathcal{S}} f(s)\int_{\mathcal{A}} f(a|s)\log\pi(a|s)\,\mathrm{d}a\,\mathrm{d}s \le \int_{\mathcal{S}} f(s)\int_{\mathcal{A}} f(a|s)\log f(a|s)\,\mathrm{d}a\,\mathrm{d}s,$$

where the last inequality holds for any policy π, since $\forall s \in \mathcal{S}$ we have that $\int_{\mathcal{A}} f(a|s)\log\pi(a|s)\,\mathrm{d}a \le \int_{\mathcal{A}} f(a|s)\log f(a|s)\,\mathrm{d}a$, as $f(a|s)$ is the maximum likelihood fit. Note that for all states $s \in \mathcal{S}$ such that $d^{\pi_n}(s) = 0$, we have that $f(s,a) = 0$. Therefore, for such states, the policy will not contribute to the objective and can be defined arbitrarily. Now, assume $d^{\pi_n}(s) > 0$. The objective function achieves a maximum when the two distributions are equal:

$$\pi_{n+1}(a|s) = f(a|s) = \frac{f(s,a)}{f(s)} = \frac{f(s,a)}{\int_{\mathcal{A}} f(s,a)\,\mathrm{d}a} = \frac{d^{\pi_n}(s)\pi_n(a|s)Q^{\pi_n}(s,a)}{\int_{\mathcal{S}}\int_{\mathcal{A}} d^{\pi_n}(s)\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a\,\mathrm{d}s} \cdot \frac{\int_{\mathcal{S}}\int_{\mathcal{A}} d^{\pi_n}(s)\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a\,\mathrm{d}s}{\int_{\mathcal{A}} d^{\pi_n}(s)\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a} = \frac{\pi_n(a|s)Q^{\pi_n}(s,a)}{\int_{\mathcal{A}}\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a} = \frac{Q^{\pi_n}(s,a)\pi_n(a|s)}{V^{\pi_n}(s)}.$$

We can now set $\pi_{n+1}(a|s) = \frac{Q^{\pi_n}(s,a)\pi_n(a|s)}{V^{\pi_n}(s)}$ also for all s such that $d^{\pi_n}(s) = 0$, which completes the proof. Note that $V^{\pi_n}(s)$ is positive thanks to the assumption of positive rewards. Similarly, the denominator $\int_{\mathcal{S}}\int_{\mathcal{A}}\hat{f}(s,a)\,\mathrm{d}a\,\mathrm{d}s = \int_{\mathcal{S}} d^{\pi_n}(s)V^{\pi_n}(s)\,\mathrm{d}s > 0$ is positive. □

When function approximation is used for policy π, the term $f(s)$ weighs the mismatch between $\pi(a|s)$ and $f(a|s)$. Indeed, we have $f(s) \propto d^{\pi}(s)V^{\pi}(s)$, suggesting that the error occurring with function approximation would be weighted more for states visited often and with a bigger value. In our setting, however, the two terms are equal since no function approximation is used.

Theorem 3.1 provides us with an interpretation on how the RWR update rule works: at each iteration, given a state s, the probability over an action a produced by policy π_n will be weighted by the expected return obtained from state s, choosing action a and following π_n. This result will be then normalized by $V^{\pi_n}(s)$, providing a new policy π_{n+1}. Alternatively, we can interpret this new policy as the fraction of return obtained by policy π_n from state s, after choosing action a with probability $\pi_n(\cdot|s)$. Intuitively, assigning more weight to actions which lead to better return should improve the policy. We prove this in the next section.

4 Monotonic Improvement Theorem

Here we prove that the update defined in Theorem 3.1 leads to monotonic improvement.[7]

Theorem 4.1. *Fix $n > 0$ and let $\pi_0 \in \Pi$ be a policy*[8]*. Assume $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$, $R(s,a) > 0$. Define the operator $B : \Pi \to \Pi$ such that $B(\pi) := \frac{Q^{\pi}\pi}{V^{\pi}}$ for $\pi \in \Pi$. Thus $\pi_{n+1} = B(\pi_n)$, i.e. $\forall s \in \mathcal{S}, \forall a \in \mathcal{A} : \pi_{n+1}(a|s) = (B\pi_n)(a|s) = \frac{Q^{\pi_n}(s,a)\pi_n(a|s)}{V^{\pi_n}(s)}$. Then $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$ we have that $V^{\pi_{n+1}}(s) \ge V^{\pi_n}(s)$ and $Q^{\pi_{n+1}}(s,a) \ge Q^{\pi_n}(s,a)$. Moreover, if for some $s \in \mathcal{S}$ holds $\mathrm{Var}_{a\sim\pi_n(a|s)}[Q^{\pi_n}(s,a)] > 0$ then the first inequality above is strict, i.e. $V^{\pi_{n+1}}(s) > V^{\pi_n}(s)$.*

Proof. We start by defining a function $V^{\pi_{n+1},\pi_n}(s)$ as the expected return for using policy π_{n+1} in state s and then following policy π_n: $V^{\pi_{n+1},\pi_n}(s) := \int_{\mathcal{A}}\pi_{n+1}(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a$. By showing that $\forall s \in \mathcal{S}$, $V^{\pi_{n+1},\pi_n}(s) \ge V^{\pi_n}(s)$, we get that $\forall s \in \mathcal{S}$, $V^{\pi_{n+1}}(s) \ge V^{\pi_n}(s)$. [9]

Now, let s be fixed:

$$\begin{aligned}
&V^{\pi_{n+1},\pi_n}(s) \ge V^{\pi_n}(s)\\
\Leftrightarrow &\int_{\mathcal{A}}\pi_{n+1}(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a \ge \int_{\mathcal{A}}\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a\\
\Leftrightarrow &\int_{\mathcal{A}}\frac{\pi_n(a|s)Q^{\pi_n}(s,a)^2}{V^{\pi_n}(s)}\,\mathrm{d}a \ge \int_{\mathcal{A}}\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a\\
\Leftrightarrow &\int_{\mathcal{A}}\pi(a|s)Q^{\pi_n}(s,a)^2\,\mathrm{d}a \ge \Big(\int_{\mathcal{A}}\pi_n(a|s)Q^{\pi_n}(s,a)\,\mathrm{d}a\Big)^2\\
\Leftrightarrow &\mathop{\mathbb{E}}_{a\sim\pi_n(a|s)}[Q^{\pi_n}(s,a)^2] \ge \mathop{\mathbb{E}}_{a\sim\pi_n(a|s)}[Q^{\pi_n}(s,a)]^2\\
\Leftrightarrow &\mathrm{Var}_{a\sim\pi_n(a|s)}[Q^{\pi_n}(s,a)] \ge 0,
\end{aligned}$$

which always holds. Finally, $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$:

$$\begin{aligned}
Q^{\pi_{n+1}}(s,a) &= R(s,a) + \gamma\int_{\mathcal{S}} p_T(s'|s,a)V^{\pi_{n+1}}(s')\,\mathrm{d}s'\\
&\ge R(s,a) + \gamma\int_{\mathcal{S}} p_T(s'|s,a)V^{\pi_n}(s')\,\mathrm{d}s'\\
&= Q^{\pi_n}(s,a).
\end{aligned}$$

□

Theorem 4.1 provides a relationship between the improvement in the state-value function and the variance of

[7] The case where the MDP has non-negative rewards and the undiscounted case are more complex and treated in Appendix B.

[8] Also in this section all policies are given as conditional densities with respect to μ_A.

[9] The argument is the same as given in (Puterman 2014), see section on Monotonic Policy Improvement.

the action-value function with respect to the actions sampled. Note that if at a certain point the policy becomes deterministic—or it becomes the greedy policy of its action-value function (i.e. the optimal policy)—, then the operator B will map the policy to itself and there will be no improvement.

5 Convergence Results

Weak convergence in topological factor

It is worth discussing what type of convergence we can achieve by iterating the B-operator $\pi_n := B(\pi_{n-1})$, where π_n are probability densities with respect to a fixed reference measure μ_A.

Consider first the classic "continuous" variable case, where μ_A is the Lebesgue measure and fix $s \in S$. Optimal policies are known to be greedy on the optimal action-value function $Q^*(s,a)$. That is, they concentrate all mass on $\arg\max_a Q^*(s,a)$. If $\arg\max_a Q^*(s,a)$ consists of just a single point $\{a^*\}$, then the optimal policy (measure), $\pi^*(\cdot|s)$ for s, concentrates all its mass in $\{a^*\}$. This means that the optimal policy does not have a density with respect to the Lebesgue measure. Furthermore $(\pi_n(\cdot|s)\cdot\mu_A)(\{a^*\}) = \int_{\{a^*\}} \pi_n(a|s)\mathrm{d}\mu_A(a) = 0$, while $\pi^*(\{a^*\}|s) = 1$. However, we still want to show that the measures $\pi_n(\cdot|s)\cdot\mu_A$ get concentrated in the neighbourhood of a^* and that this neighbourhood gets tinier as n increases. We will use the concept of weak convergence to prove this.

Another problem arises when considering the above: since $\arg\max_a Q^*(s,a)$ can consist of multiple points, the set of optimal policies is $\mathcal{P}(\arg\max_a Q^*(s,a))$, where $\mathcal{P}(F) := \{\mu : \mu \text{ is a probability measure on } \mathcal{B}(\mathcal{A}), \mu(F) = 1\}$ for a $F \in \mathcal{B}(\mathcal{A})$. We want to prove convergence even when the sequence of policies π_n oscillates near $\mathcal{P}(\arg\max_a Q^*(s,a))$. A way of coping with this is to make $\arg\max_a Q^*(s,a)$ a single point through topological factorisation, to obtain the limit by working in a quotient space. The notion of convergence we will be using is described in the following definition.

Definition 1. *(Weak convergence of measures in metric space relative to a compact set) Let (X,d) be a metric space, $F \subset X$ a compact subset, $\mathcal{B}(X)$ its Borel σ-algebra. Denote $(\tilde{X},\tilde{d})$ a metric space resulting as a topological quotient with respect to F and ν the quotient map $\nu : X \to \tilde{X}$ (see Lemma C.2 for details). A sequence of probability measures P_n is said to converge weakly relative to F to a measure P denoted*

$$P_n \to^{w(F)} P,$$

if and only if the image measures of P_n under ν converge weakly to the image measure of P under ν:

$$\nu P_n \to^{w} \nu P.$$

Note that the limit is meant to be unique just in quotient space, thus if P is a weak limit (relative to F) of a sequence (P_n), then also all measures P' for which $\nu P' = \nu P$ are relatively weak limits, i.e. $P'|_{\mathcal{B}(X)\cap F^c} = P|_{\mathcal{B}(X)\cap F^c}$. Thus, they can differ on $\mathcal{B}(X)\cap F$. While the total mass assigned to F must be the same for P and P', the distribution of masses inside F may differ.

Main results

Consider for all $n > 0$ the sequence generated by $\pi_n := B(\pi_{n-1})$. For convenience, for all $n \geq 0$, we define $Q_n := Q_{\pi_n}, \quad V_n := V_{\pi_n}$. First we note that, since the reward is bounded, the monotonic sequences of value functions converge point-wise to a limit:

$$(\forall s \in \mathcal{S}) : V_n(s) \nearrow V_L(s) \leq B_V < +\infty$$
$$(\forall s \in \mathcal{S}, a \in \mathcal{A}) : Q_n(s,a) \nearrow Q_L(s,a) \leq B_V < +\infty,$$

where $B_V = \frac{1}{1-\gamma}||R||_\infty$. Further $\forall n\, Q_n$ is continuous since $Q_n = W(V_n)$ and W maps all bounded functions to continuous functions.

The convergence proof proceeds in four steps:

1. First we show in Lemma 5.1 that Q_L can be expressed in terms of V_L through W operator. This helps when showing that Q_n converges uniformly to Q_L.
2. Then we demonstrate in Lemma 5.2 that $\forall s \in \mathcal{S}$ the sequence of policy measures $\pi_n(\cdot|s)\cdot\mu_A$ converges weakly relative to the set $M(s) := \arg\max_a Q_L(s,a)$ to a measure that assigns all probability mass to greedy actions of $Q_L(\cdot,s)$, i.e. $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s) \in \mathcal{P}(M(s))$. However we are interested just in those π_L which are kernels, i.e. $\pi_L \in \Pi_L := \{\pi'_L : \pi'_L$ is a probability kernel from $(\mathcal{S},\mathcal{B}(\mathcal{S}))$ to $(\mathcal{A},\mathcal{B}(\mathcal{A}))$, $\forall s \in \mathcal{S},\ \pi'_L(.|s) \in \mathcal{P}(M(s))\}$ — the set of all greedy policies on Q_L.
3. At this point we do not know yet if Q_L and V_L are the value functions of π_L. We prove this in Lemma 5.3 (together with previous Lemmas) by showing that they are fixed points of the Bellman operator.
4. Finally, we state the main results in Theorem 5.1. Since V_L and Q_L are value functions for π_L and π_L is greedy with respect to Q_L, then Q_L is the unique fixed point of the Bellman's optimality operator:

$$Q_L(s,a) = [T(Q)](s,a) = \\ = R(s,a) + \gamma \int_{\mathcal{S}} \max_{a'} Q(s',a') p_T(s'|s,a)\mathrm{d}s'.$$

Therefore Q_L and V_L are optimal value functions and π_L is an optimal policy for the MDP.

Lemma 5.1. *The following holds:*
1. *$Q_L = W(V_L)$,*
2. *Q_L is continuous,*
3. *Q_n converges to Q_L uniformly.*

Proof. **1.** Fix $(s,a) \in \mathcal{S}\times\mathcal{A}$. We aim to show $Q_L(s,a) - [W(V_L)](s,a) = 0$. Since $Q_n = W(V_n)$, we can write:

$$\begin{aligned} Q_L(s,a) - [W(V_L)](s,a) &= \\ &= Q_L(s,a) - Q_n(s,a) \\ &\quad - [W(V_L)](s,a) + [W(V_n)](s,a) \\ &\leq |Q_L(s,a) - Q_n(s,a)| \\ &\quad + |[W(V_L)](s,a) - [W(V_n)](s,a)|. \end{aligned}$$

The first part can be made arbitrarily small as $Q_n(s,a) \to Q_L(s,a)$. Consider the second part and fix $\epsilon > 0$. Since

$V_n \to V_L$ point-wise, from Severini-Egorov's theorem (Severini 1910) there exists $S_\epsilon \subset S$ with $(p_T(\cdot|s,a)\cdot\mu_S)(S_\epsilon^c) < \epsilon$ such that $\|V_n - V_L\|_\infty \to 0$ on S_ϵ. Thus there exists n_0 such that $\|V_n - V_L\|_\infty < \epsilon$ for all $n > n_0$. Now let us rewrite the second part for $n > n_0$:

$$\begin{aligned}
&|[W(V_L)](s,a) - [W(V_n)](s,a)| \\
&\quad\le \int_{\mathcal{S}} |V_L(s') - V_n(s')| p_T(s'|s,a)\mathrm{d}\mu_S(s') \\
&\quad= \int_{S_\epsilon} |V_L(s') - V_n(s')| p_T(s'|s,a)\mathrm{d}\mu_S(s') \\
&\qquad+ \int_{S_\epsilon^c} |V_L(s') - V_n(s')| p_T(s'|s,a)\mathrm{d}\mu_S(s') \\
&\quad\le \|V_L - V_n\|_\infty + B_V \int_{S_\epsilon^c} p_T(s'|s,a)\mathrm{d}\mu_S(s') \\
&\quad\le \epsilon + B_V\epsilon,
\end{aligned}$$

which can be made arbitrarily small.

2. Q_L is continuous because W maps all bounded measurable functions to continuous functions.

3. Since Q_n and Q_L are continuous functions in a compact space and Q_n is a monotonically increasing sequence that converges point-wise to Q_L, we can apply Dini's theorem (see Th. 7.13 on page 150 in (Rudin 1976)) which ensures uniform convergence of Q_n to Q_L. □

Lemma 5.2. *Let π_n be a sequence generated by $\pi_n := B(\pi_{n-1})$. Let π_0 be continuous in actions and $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$, $\pi_0(a|s) > 0$. Define $M(s) := \arg\max Q_L(\cdot|s)$. Then $\forall \pi_L \in \Pi_L \neq \emptyset$, $\forall s \in \mathcal{S}$, we have $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s) (\in \mathcal{P}(M(s)))$.*

Proof. First notice that the set Π_L is nonempty[10]. Fix $\pi_L \in \Pi_L$ and $s \in \mathcal{S}$. In order to prove that $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s)$, we will use a characterization of relative weak convergence that follows from an adaptation of the Portmanteau Lemma (Billingsley 2013) (see Appendix C.3). In particular, it is enough to show that for all open sets $U \subset \mathcal{A}$ such that $U \cap M(s) = \emptyset$ or such that $M(s) \subset U$, we have that $\liminf_n(\pi_n(\cdot|s)\cdot\mu_A)U \ge \pi_L(\cdot|s)U$.

The case $U \cap M(s) = \emptyset$ is trivial since $\pi_L(\cdot|s)(U) = 0$. For the remaining case $M(s) \subset U$ it holds $\pi_L(\cdot|s)(U) = 1$. Thus we have to prove $\liminf_n(\pi_n(\cdot|s)\cdot\mu_A)U = 1$. If we are able to construct an open set $D \subset U$ such that $(\pi_n(\cdot|s)\cdot\mu_A)(D) \to 1$ for $n \to \infty$, then we will get that $\liminf_n(\pi_n(\cdot|s)\cdot\mu_A)U \ge 1$, satisfying the condition for relative weak convergence of $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s)$.

The remainder of the proof will focus on constructing such a set. Fix $a^* \in M(s)$ and $0 < \epsilon < 1/3$. Define a continuous map $\lambda : \mathcal{A} \to \mathbb{R}^+$ and closed sets A_ϵ and B_ϵ:

$$\begin{aligned}
\lambda(a) &:= \frac{Q_L(a)}{Q_L(a^*)}, \\
A_\epsilon &:= \{a \in \mathcal{A} | \lambda(a) \le 1 - 2\epsilon\}, \\
B_\epsilon &:= \{a \in \mathcal{A} | \lambda(a) \ge 1 - \epsilon\},
\end{aligned}$$

where continuity of the map stems from $Q_L(a^*) > 0$ and continuity of Q_L (Lemma 5.1). We will prove that the candidate set is $D = A_\epsilon^c$. In particular, we must prove that $A_\epsilon^c \subset U$ and that $(\pi_n(\cdot|s)\cdot\mu_A)(A_\epsilon) \to 0$. Using Lemma C.1 (Appendix) on function λ, we can choose $\epsilon > 0$ such that $A_\epsilon^c \subset U$, satisfying the first condition. We are left to prove that $(\pi_n(\cdot|s)\cdot\mu_A)(A_\epsilon) \to 0$.

Assume $A_\epsilon \neq \emptyset$ (otherwise the condition is proven): for all $a \in A_\epsilon$ and $b \in B_\epsilon$ it holds:

$$\begin{aligned}
\frac{Q_L(a)}{Q_L(b)} = \frac{\frac{Q_L(a)}{Q_L(a^*)}}{\frac{Q_L(b)}{Q_L(a^*)}} &\le \frac{Q_L(a)}{Q_L(a^*)(1-\epsilon)} \\
&\le \frac{1-2\epsilon}{1-\epsilon} = 1 - \frac{\epsilon}{1-\epsilon} =: \alpha_1 < 1.
\end{aligned}$$

For Lemma 5.1 Q_n converges uniformly to Q_L. Therefore we can fix $n_0 > 0$ such that $\|Q_n - Q_L\|_\infty < \epsilon'$ for all $n \ge n_0$, where we define $\epsilon' := 0.1 \times Q_L(a^*)(1-\epsilon)(1-\alpha_1)$. Now we can proceed by bounding Q_n ratio from above. For all $n \ge n_0$, $a \in A_\epsilon$ and $b \in B_\epsilon$:

$$\begin{aligned}
\frac{Q_n(a)}{Q_n(b)} &\le \frac{Q_L(a)}{Q_L(b) - \epsilon'} \le \frac{Q_L(a)}{Q_L(a^*)(1-\epsilon) - \epsilon'} \\
&= \frac{Q_L(a)}{Q_L(a^*)(1-\epsilon)(1 - 0.1(1-\alpha_1))} \\
&= \frac{\alpha_1}{(0.9 + 0.1\alpha_1)} =: \alpha < 1.
\end{aligned}$$

Finally, we can bound the policy ratio. For all $n \ge n_0$, $a \in A_\epsilon$, $b \in B_\epsilon$:

$$\frac{\pi_n(a|s)}{\pi_n(b|s)} = \frac{\pi_0(a|s)}{\pi_0(b|s)} \prod_{i=0}^{n} \frac{Q_i(s,a)}{Q_i(s,b)} \le \alpha^n c(a,b),$$

where

$$c(a,b) := \alpha^{-n_0} \frac{\pi_0(a|s)}{\pi_0(b|s)} \prod_{i=0}^{n_0} \frac{Q_i(s,a)}{Q_i(s,b)}.$$

The function $c : A_\epsilon \times B_\epsilon \to \mathbb{R}^+$ is continuous as π_0, Q_i are continuous (and denominators are non-zero due to $\pi_0(b|s) > 0$ and $Q_i(s,b) > 0$). Since $A_\epsilon \times B_\epsilon$ is a compact set, there exists c_m such that $c \le c_m$. Thus we have that for all $n > n_0$:

$$\pi_n(a|s) \le \alpha^n c_m \pi_n(b|s).$$

Integrating with respect to a over A_ϵ and then with respect to b over B_ϵ (using reference measure μ_A in both cases) we obtain:

$$\begin{aligned}
&(\pi_n(\cdot|s)\cdot\mu_A)(A_\epsilon) \times (\mu_A B_\epsilon) \\
&\qquad\le \alpha^n c_m (\pi_n(\cdot|s)\cdot\mu_A)(B_\epsilon) \times (\mu_A A_\epsilon).
\end{aligned}$$

[10] The argument goes as follows: $H := \cup_{s\in S}\{s\} \times M(s)$ is a closed set, then $f(s) := \sup M(s)$ is upper semi-continuous and therefore measurable. Then graph of f is measurable so we can define a probability kernel $\pi_L(B|s) := \mathbf{1}_B(f(s))$ for all B measurable.

Rearranging terms, we have:

$$(\pi_n(\cdot|s)\cdot\mu_A)(A_\epsilon) \leq \alpha^n\left[c_m\frac{\mu_A A_\epsilon}{\mu_A B_\epsilon}(\pi_n(\cdot|s)\cdot\mu_A)B_\epsilon\right]\to 0, n\to\infty,$$

since the nominator in brackets is composed by finite measures of sets, thus finite numbers, while the denominator $\mu_A B_\epsilon > 0$. Indeed, define the open set $C := \{a \in \mathcal{A}|\lambda(a) > 1-\epsilon\} \subset B_\epsilon$. Then $\mu_A(B_\epsilon) \geq \mu_A(C) > 0$ (μ_A is strictly positive). To conclude, we have proven that for arbitrarily small $\epsilon > 0$, the term $(\pi_n(\cdot|s)\cdot\mu_A)(A_\epsilon)$ tends to 0, satisfying the condition for relative weak convergence of $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s)$. □

Lemma 5.3. *Assume that, for each $s \in \mathcal{S}$, for each $\pi_L \in \Pi_L$, we have that $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s)(\in \mathcal{P}(M(s)))$. Then this holds:*

$$V_L(s) = \int_{\mathcal{A}} Q_L(s,a)\,\mathrm{d}\pi_L(a|s). \tag{4}$$

Proof. Fix $s \in \mathcal{S}$ and $\pi_L \in \Pi_L$. We aim to show $V_L(s) - \int_{\mathcal{A}} Q_L(s,a)\,\mathrm{d}\pi_L(a|s) = 0$. Since $V_n(s) - \int_{\mathcal{A}} Q_n(s,a)\pi_n(a|s)\,\mathrm{d}\mu_A(a) = 0$, we have:

$$\begin{aligned}
&\Big|V_L(s) - \int_{\mathcal{A}} Q_L(s,a)\,\mathrm{d}\pi_L(a|s)\Big| \\
&\quad= \Big|V_L(s) - V_n(s) - \int_{\mathcal{A}} Q_L(s,a)\,\mathrm{d}\pi_L(a|s) \\
&\qquad + \int_{\mathcal{A}} Q_n(s,a)\pi_n(a|s)\,\mathrm{d}\mu_A(a)\Big| \\
&\quad\leq \Big|V_L(s) - V_n(s)\Big| + \Big|\int_{\mathcal{A}} Q_L(s,a)\,\mathrm{d}\pi_L(a|s) \\
&\qquad - \int_{\mathcal{A}} Q_n(s,a)\pi_n(a|s)\,\mathrm{d}\mu_A(a)\Big|.
\end{aligned}$$

The first part can be made arbitrarily small due to $V_n(s) \to V_L(s)$. For the second part:

$$\begin{aligned}
&\Big|\int_{\mathcal{A}} Q_L(s,a)\,\mathrm{d}\pi_L(a|s) - \int_{\mathcal{A}} Q_n(s,a)\pi_n(a|s)\,\mathrm{d}\mu_A(a)\Big| \\
&= \Big|\int_{\mathcal{A}} Q_L(s,a)\mathrm{d}\pi_L(a|s) - \int_{\mathcal{A}} Q_L(s,a)\pi_n(a|s)\mathrm{d}\mu_A(a) \\
&\quad + \int_{\mathcal{A}} Q_L(s,a)\pi_n(a|s)\mathrm{d}\mu_A(a) \\
&\quad - \int_{\mathcal{A}} Q_n(s,a)\pi_n(a|s)\,\mathrm{d}\mu_A(a)\Big| \\
&\leq \Big|\int_{\mathcal{A}} Q_L(s,a)\mathrm{d}\pi_L(a|s) - \int_{\mathcal{A}} Q_L(s,a)\pi_n(a|s)\mathrm{d}\mu_A(a)\Big| \\
&\quad + \int_{\mathcal{A}} |Q_L(s,a) - Q_n(s,a)|\pi_n(a|s)\mathrm{d}\mu_A(a),
\end{aligned}$$

where the first term tends to zero since $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s)$ and Q_L is continuous and constant on $M(s)$, satisfying the conditions of the adapted Portmanteau Lemma (Billingsley 2013) (see Appendix C.3). The second term can be arbitrarily small since Lemma 5.1 ensures uniform convergence of Q_n to Q_L. □

Theorem 5.1. *Let π_n be a sequence generated by $\pi_n := B(\pi_{n-1})$. Let π_0 be such that $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$ $\pi_0(a|s) > 0$ and continuous in actions. Then $\forall s \in \mathcal{S}$ $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s)$, where $\pi_L \in \Pi_L$ is an optimal policy for the MDP. Moreover, $\lim_{n\to\infty} V_n = V_L$, $\lim_{n\to\infty} Q_n = Q_L$ are the optimal state and action value functions.*

Proof. Fix $\pi_L \in \Pi_L$ (we have already shown that $\Pi_L \neq \emptyset$). Due to Lemma 5.2, we know that for all $s \in \mathcal{S}$, $\pi_L(\cdot|s)$ is the relative weak limit $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s)$ and further we know that π_L is greedy on $Q_L(s,a)$ (from definition of Π_L). Moreover, thanks to Lemmas 5.3 and 5.1, $V_L(s)$ and $Q_L(s,a)$ are the state and action value functions of π_L because they are fixed points of the Bellman operator. Since $\pi_L(\cdot|s) \in \mathcal{P}(\arg\max_a Q_L(s,a))$, $V_L(s)$ and $Q_L(s,a)$ are also the unique fixed points of Bellman's optimality operator, hence V_L, Q_L are optimal value functions and π_L is an optimal policy. □

This result has several implications. First, it provides a solid theoretical ground for both previous and future works that are based on RWR (Dayan and Hinton 1997; Peters and Schaal 2007; Peng et al. 2019) and lends us some additional understanding regarding the properties of similar algorithms (e.g., (Abdolmaleki et al. 2018b)). It should also be stressed that the results presented herein are for compact state and action spaces: traits of some key domains such as robotic control. In addition to the above, one should also note that the upper bound on $(\pi_n(\cdot|s)\cdot\mu_A)(A_\epsilon)$ constructed in lemma 5.2 can be used to study convergence orders and convergence rates of RWR. The following corollary, for example, proves R-linear convergence for the special case of finite state and action spaces:

Corollary 5.1. *Under the assumptions of lemma 5.2, if $\mathcal{S}$ and $\mathcal{A}$ are finite, then $\|V^* - V_n\|_\infty = O(\alpha_m^n)$, where $0 \leq \alpha_m < 1$, $\alpha_m := \frac{2\lambda_m}{0.9+1.1\lambda_m}$, and $\lambda_m := \max_{s\in\mathcal{S}}\max_{a\in\mathcal{A}\setminus M(s)}\frac{Q^*(s,a)}{V^*(s)}$.*

A proof of the above is included in the appendix D. We observe that in the finite case, $\|V^* - V_n\|_\infty$ converges to 0 R-linearly (i.e., $\|V^* - V_n\|_\infty$ is bounded by a Q-linearly converging sequence α_m^n). We provide an example of a finite MDP in lemma D.1 which exhibits linear convergence rate, showing that the upper bound from the corollary 5.1 is asymptotically tight in regards to the convergence order. Therefore it is not possible to achieve an order of convergence better than linear. Furthermore, the example in lemma D.2 shows that, in the continuous case, the convergence order could be sub-linear. Appendix E discusses the motivation of our approach.

6 Demonstration of RWR Convergence

To illustrate that the update scheme of Theorem 3.1 converges to the optimal policy, we test it on a simple environment that meets the assumptions of the Theorem. In particular, we ensure that rewards are positive and that there is no function approximations for value functions and policies. In order to meet these criteria, we use the modified four-room gridworld domain (Sutton, Precup, and Singh 1999) shown

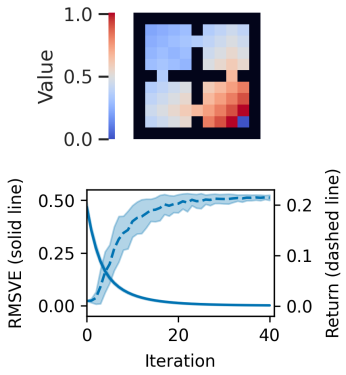

Figure 1: (Top) the value of states under the optimal policy in the four-room gridworld domain. (Bottom) the root-mean-squared value error of reward-weighted regression in the four-room gridworld domain—compared to the optimal policy—and the return obtained by running the learned policy of reward-weighted regression. All lines are averages of 100 runs under different uniform random initial policies. Shading shows standard deviation.

on the left of Figure 1. Here the agent starts in the upper left corner and must navigate to the bottom right corner (i.e., the goal state). In non-goal states actions are restricted to moving one square at each step in any of the four cardinal directions. If the agent tries to move into a square containing a wall, it will remain in place. In the goal state, all actions lead to the agent remaining in place. The agent receives a reward of 1 when transitioning from a non-goal state to the goal state and a reward of 0.001 otherwise. The discount-rate is 0.9 at each step. At each iteration, we use Bellman's updates to obtain a reliable estimate of Q_n and V_n, before updating π_n using the operator in Theorem 3.1.

The bottom left of Figure 1 shows the root-mean-squared value error (RMSVE) of the learned policy at each iteration as compared to the optimal policy, while the bottom right shows the return obtained by the learned policy at each iteration. Smooth convergence can be observed under reward-weighted regression. The source code for this experiment is available at https://github.com/dylanashley/reward-weighted-regression.

7 Related Work

The principle behind expectation-maximization was first applied to artificial neural networks by Von der Malsburg (1973). The reward-weighted regression (RWR) algorithm, though, originated in the work of Peters and Schaal (2007) which sought to bring earlier work of Dayan and Hinton (1997) to the domain of operational space control and reinforcement learning. However, Peters and Schaal (2007) only considered the immediate-reward reinforcement learning (RL) setting. This was later extended to the episodic setting separately by Wierstra et al. (2008a) and then by Kober and Peters (2011). Wierstra et al. (2008a) went even further and also extended RWR to partially observable Markov decision processes, and Kober and Peters (2011) applied it to motor learning in robotics. Separately, Wierstra et al. (2008b) extended RWR to perform fitness maximization for evolutionary methods. Hachiya, Peters, and Sugiyama (2009, 2011) later found a way of reusing old samples to improve RWR's sample complexity. Much later, Peng et al. (2019) modified RWR to produce an algorithm for off-policy RL, using deep neural networks as function approximators.

Other methods based on principles similar to RWR have been proposed. Neumann and Peters (2008), for example, proposed a more efficient version of the well-known fitted Q-iteration algorithm (Riedmiller 2005; Ernst, Geurts, and Wehenkel 2005; Antos, Munos, and Szepesvári 2007) by using what they refer to as *advantaged-weighted regression*—which itself is based on the RWR principle. Ueno et al. (2012) later proposed *weighted likelihood policy search* and showed that their method both has guaranteed monotonic increases in the expected reward. Osa and Sugiyama (2018) subsequently proposed a hierarchical RL method called *hierarchical policy search via return-weighted density estimation* and showed that it is closely related to the episodic version of RWR by (Kober and Peters 2011).

Notably, all of the aforementioned works, as well as a number of other proposed similar RL methods (e.g., Peters, Mülling, and Altun (2010), Neumann (2011), Abdolmaleki et al. (2018b), Abdolmaleki et al. (2018a)), are based on the expectation-maximization framework of Dempster, Laird, and Rubin (1977) and are thus known to have monotonic improvements of the policy in the RL setting under certain conditions. However, it has remained an open question under which conditions convergence to the optimal is guaranteed.

8 Conclusion and Future Work

We provided the first global convergence proof for Reward-Weighted Regression (RWR) in absence of reward transformation and function approximation. The convergence achieved is linear when using finite state and action spaces and can be sub-linear in the continuous case. We also highlighted problems that may arise under nonlinear reward transformations, potentially resulting in changes to the optimal policy. In real-world problems, access to true value functions may be unrealistic. Future work will study RWR's convergence under function approximation. In such a case, the best scenario that one can expect is to achieve convergence to a local optimum. One possible approach is to follow a procedure similar to standard policy gradients (Sutton et al. 1999) and derive a class of value function approximators that is compatible with the RWR objective. It might be possible then to prove local convergence under value function approximation using stochastic approximation techniques (Borkar 2008; Sutton, Maei, and Szepesvári 2009; Sutton et al. 2009). This would require casting the value function and policy updates in a system of equations and studying the convergence of the corresponding ODE under specific assumptions. Our RWR is on-policy, using only recent data to update the current policy. Future work will also study convergence in challenging off-policy settings (using all past data), which require corrections of the mismatch between state-distributions, typically through a mechanism like Importance Sampling.

Acknowledgements

We would like to thank Sjoerd van Steenkiste and František Žák for their insightful comments. This work was supported by the European Research Council (ERC, Advanced Grant Number 742870), the Swiss National Supercomputing Centre (CSCS, Project s1090), and by the Swiss National Science Foundation (Grant Number 200021_192356, Project NEUSYM). We also thank both the NVIDIA Corporation for donating a DGX-1 as part of the Pioneers of AI Research Award and IBM for donating a Minsky machine.

References

Abdolmaleki, A.; Springenberg, J. T.; Degrave, J.; Bohez, S.; Tassa, Y.; Belov, D.; Heess, N.; and Riedmiller, M. 2018a. Relative Entropy Regularized Policy Iteration. arXiv:1812.02256.

Abdolmaleki, A.; Springenberg, J. T.; Tassa, Y.; Munos, R.; Heess, N.; and Riedmiller, M. A. 2018b. Maximum a Posteriori Policy Optimisation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Antos, A.; Munos, R.; and Szepesvári, C. 2007. Fitted Q-iteration in continuous action-space MDPs. In Platt, J. C.; Koller, D.; Singer, Y.; and Roweis, S. T., eds., *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 9–16. Curran Associates, Inc.

Bellemare, M. G.; Candido, S.; Castro, P. S.; Gong, J.; Machado, M. C.; Moitra, S.; Ponda, S. S.; and Wang, Z. 2020. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836): 77–82.

Billingsley, P. 2013. *Convergence of Probability Measures*. John Wiley & Sons.

Borkar, V. S. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48. Hindustan Book Agency.

Dayan, P.; and Hinton, G. E. 1997. Using Expectation-Maximization for Reinforcement Learning. *Neural Comput.*, 9(2): 271–278.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22.

Ernst, D.; Geurts, P.; and Wehenkel, L. 2005. Tree-Based Batch Mode Reinforcement Learning. *J. Mach. Learn. Res.*, 6: 503–556.

Hachiya, H.; Peters, J.; and Sugiyama, M. 2009. Efficient Sample Reuse in EM-Based Policy Search. In Buntine, W. L.; Grobelnik, M.; Mladenic, D.; and Shawe-Taylor, J., eds., *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I*, volume 5781 of *Lecture Notes in Computer Science*, 469–484. Springer.

Hachiya, H.; Peters, J.; and Sugiyama, M. 2011. Reward-Weighted Regression with Sample Reuse for Direct Policy Search in Reinforcement Learning. *Neural Comput.*, 23(11): 2798–2832.

Kober, J.; and Peters, J. 2011. Policy search for motor primitives in robotics. *Mach. Learn.*, 84(1-2): 171–203.

Munkres, J. 2000. *Topology*. Prentice Hall, Incorporated.

Neumann, G. 2011. Variational Inference for Policy Search in changing situations. In Getoor, L.; and Scheffer, T., eds., *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 817–824. Omnipress.

Neumann, G.; and Peters, J. 2008. Fitted Q-iteration by Advantage Weighted Regression. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, 1177–1184. Curran Associates, Inc.

Osa, T.; and Sugiyama, M. 2018. Hierarchical Policy Search via Return-Weighted Density Estimation. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 3860–3867. AAAI Press.

Peng, X. B.; Kumar, A.; Zhang, G.; and Levine, S. 2019. Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. arXiv:1910.00177.

Peters, J.; Mülling, K.; and Altun, Y. 2010. Relative Entropy Policy Search. In Fox, M.; and Poole, D., eds., *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press.

Peters, J.; and Schaal, S. 2007. Reinforcement Learning by Reward-Weighted Regression for Operational Space Control. In Ghahramani, Z., ed., *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, 745–750. ACM.

Peters, J.; and Schaal, S. 2008. Learning to Control in Operational Space. *The International Journal of Robotics Research*, 27(2): 197–212.

Plappert, M.; Andrychowicz, M.; Ray, A.; McGrew, B.; Baker, B.; Powell, G.; Schneider, J.; Tobin, J.; Chociej, M.; Welinder, P.; Kumar, V.; and Zaremba, W. 2018. Multi-Goal Reinforcement Learning: Challenging Robotics Environments and Request for Research. arXiv:1802.09464.

Pollard, D. 2001. *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Riedmiller, M. A. 2005. Neural Fitted Q Iteration - First Experiences with a Data Efficient Neural Reinforcement Learning Method. In Gama, J.; Camacho, R.; Brazdil, P.; Jorge, A.; and Torgo, L., eds., *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, volume 3720 of *Lecture Notes in Computer Science*, 317–328. Springer.

Rudin, W. 1976. *Principles of Mathematical Analysis*. McGraw-hill New York, 3d ed. edition.

Severini, C. 1910. Sulle successioni di funzioni ortogonali [On Sequences of Orthogonal Functions]. *Atti dell'Accademia Gioenia, serie 5a (in Italian), 3 (5): Memoria XIII, 1-7, JFM 41.0475.04. Published by the Accademia Gioenia in Catania*.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T. P.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nat.*, 529(7587): 484–489.

Stratonovich, R. 1960. Conditional Markov processes. *Theory of Probability And Its Applications*, 5(2): 156–178.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. USA: A Bradford Book. ISBN 0262039249, 9780262039246.

Sutton, R. S.; Maei, H. R.; Precup, D.; Bhatnagar, S.; Silver, D.; Szepesvári, C.; and Wiewiora, E. 2009. Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 993–1000. New York, NY, USA: Association for Computing Machinery. ISBN 9781605585161.

Sutton, R. S.; Maei, H. R.; and Szepesvári, C. 2009. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, 1609–1616.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, 1057–1063. Cambridge, MA, USA: MIT Press.

Sutton, R. S.; Precup, D.; and Singh, S. P. 1999. Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artif. Intell.*, 112(1-2): 181–211.

Ueno, T.; Hayashi, K.; Washio, T.; and Kawahara, Y. 2012. Weighted Likelihood Policy Search with Model Selection. In Bartlett, P. L.; Pereira, F. C. N.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 2366–2374.

Von der Malsburg, C. 1973. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2): 85–100.

Wierstra, D.; Schaul, T.; Peters, J.; and Schmidhuber, J. 2008a. Episodic Reinforcement Learning by Logistic Reward-Weighted Regression. In Kurková, V.; Neruda, R.; and Koutník, J., eds., *Artificial Neural Networks - ICANN 2008 , 18th International Conference, Prague, Czech Republic, September 3-6, 2008, Proceedings, Part I*, volume 5163 of *Lecture Notes in Computer Science*, 407–416. Springer.

Wierstra, D.; Schaul, T.; Peters, J.; and Schmidhuber, J. 2008b. Fitness Expectation Maximization. In Rudolph, G.; Jansen, T.; Lucas, S. M.; Poloni, C.; and Beume, N., eds., *Parallel Problem Solving from Nature - PPSN X, 10th International Conference Dortmund, Germany, September 13-17, 2008, Proceedings*, volume 5199 of *Lecture Notes in Computer Science*, 337–346. Springer.

Wu, C. J. 1983. On the Convergence Properties of the EM Algorithm. *The Annals of statistics*, 11(1): 95–103.