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]

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. For the latest iteration of this work, see `https://arxiv.org/abs/2107.09088`.

1. Introduction

Reinforcement learning (RL) is a branch of artificial intelligence that considers learning agents interacting with an environment (Sutton & 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 & Schaal, 2007). RWR works by transforming the RL problem into a form solvable by well-studied expectation-maximization (EM) methods (Dempster et al., 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 average reward (Dayan & 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 & Schaal (2007; 2008); Peng et al. (2019); Abdolmaleki et al. (2018b). Frequently these involve an exponential transformation. In the past, it has been claimed that a positive, strictly increasing transformation $u_\tau(s)$ with $\int_0^\infty u_\tau(r)\,\mathrm{d}r = \textit{const}$ would not alter the optimal solution for the MDP (Peters & Schaal, 2007). Unfortunately, as demonstrated in Appendix C, this is not the case. The consequence of this is that we cannot rely on those transformations if we want prove convergence. Therefore, we restrict ourselves here to 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**; 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.

[1]The Swiss AI Lab IDSIA/USI/SUPSI, Lugano, Ticino, Switzerland [2]NNAISENSE, Lugano, Ticino, Switzerland. Correspondence to: Miroslav Štrupl <miroslav.strupl@idsia.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[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. $\int_S (\cdot)\mathrm{d}s := \int_S (\cdot)\mathrm{d}\mu_S(s)$. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to \mathbb{R}^+$ measurable. We denote by $f \cdot \mu$ the measure arising from density f and reference measure μ.

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 $(s, a) \in \mathcal{S} \times \mathcal{A}$, and thus the density $p_T(s'|s, a)$ is continuous (in $\|\cdot\|_1$ norm) for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. $R(s, a)$ 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.

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$W : L_\infty(S) \to C(S \times A)$ as $[W(V)](s, a) := R(s, a) + \gamma \int_S V(s')p_T(s'|s, a)\mathrm{d}s'$ and the Bellman's optimality operator $T : L_\infty(S \times A) \to C(S \times A)$ as $[T(Q)](s, a) := R(s, a) + \gamma \int_S \max_{a'} Q(s', a')p_T(s'|s, a)\mathrm{d}s'$. 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) 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) a sample representation of π_n, weighted by the return. At each iteration, RWR's objective is to find policy π maximizing:

$$\mathop{\mathbb{E}}_{s\sim d^{\pi_n}(\cdot),a\sim\pi_n(\cdot|s)}\left[\mathop{\mathbb{E}}_{R_t\sim p(\cdot|s_t=s,a_t=a,\pi_n)}[R_t \log \pi(a|s)]\right], \tag{1}$$

where Π is the set of all conditional probability densities (meant with respect to μ_A)[5]. This is equivalent to the following:

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

We start by deriving a closed form solution to the optimization problem. Proof is in Appendix A.

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:*

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

4. Monotonic Improvement Theorem

Here we prove that the update defined in Theorem 3.1 leads to monotonic improvement. Proof is in Appendix A.

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

Theorem 4.1. *Fix $n > 0$ and let $\pi_0 \in \Pi$ be a policy*[6]*. Assume $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$, $R(s,a) > 0$. Define the operator $B : \Pi \to \Pi$ such that $\pi_{n+1} = B(\pi_n) = \frac{Q^{\pi_n}(s,a)\pi_n(a|s)}{V^{\pi_n}(s)}$. Then $\forall s \in \mathcal{S}$ we have that $V^{\pi_{n+1}}(s) \geq V^{\pi_n}(s)$ and $Q^{\pi_{n+1}}(s,a) \geq Q^{\pi_n}(s,a)$. Moreover, $\forall s \in \mathcal{S}$: $\mathrm{Var}_{a\sim\pi_n(a|s)}[Q^{\pi_n}(s,a)] > 0$ the inequalities above are strict.*

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

5.1. 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 $\mathcal{P}(F) := \{\mu : \mu \text{ is a probability measure on } \mathcal{B}(A), \mu(F) = 1\}$ for a $F \in \mathcal{B}(A)$. 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 B.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: [7] *$\nu P_n \to^w \nu P$.*

5.2. 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:

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

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 $\forall s \in S$ 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. $\pi_n(\cdot|s)\cdot\mu_A \to^{w(M(s))} \pi_L(\cdot|s) \in \mathcal{P}(M(s))$. Moreover $\pi_L \in \Pi_L := \{\pi'_L : \pi'_L$ is a probability kernel from $(\mathcal{S},\mathcal{B}(\mathcal{S}))$ to $(\mathcal{A},\mathcal{B}(\mathcal{A})), \forall s \in \mathcal{S},\ \pi'_L(.|s) \in \mathcal{P}(M(s))\}$.
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. Therefore Q_L and V_L are optimal value functions and π_L is an optimal policy for the MDP.

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

[7] 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.

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.*

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* $\pi_n(\cdot|s) \cdot \mu_A \to^{w(M(s))} \pi_L(\cdot|s) (\in \mathcal{P}(M(s)))$.

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:*

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

Proofs for the lemmas above can be found in Appendix A.

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 $s \in S$, $\pi_L(\cdot|s)$ 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. □

6. Experiments

To illustrate that the update scheme of Theorem 3.1 converges to the optimal policy, we test it on the modified[8] four-room gridworld domain (Sutton et al., 1999) shown 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.

[8]So as to ensure all rewards are positive.

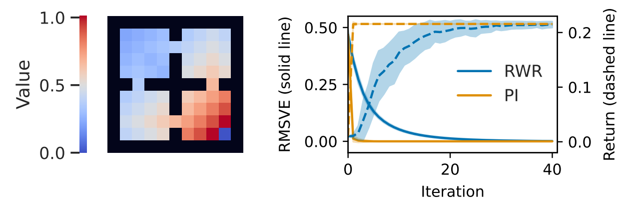

Figure 1. (Left) the value of states under the optimal policy in the four-room gridworld domain. (Right) the root-mean-squared value error (compared to the optimal policy) and return of RWR and policy iteration in the four-room gridworld domain. All lines are averages of 100 runs under different uniform random initial policies. Shading shows standard deviation.

The center 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 standard policy iteration converges more rapidly, smooth convergence can be observed under reward-weighted regression—as would be expected here. The right of Figure 1 shows the return obtained by the learned policy at each iteration. The difference between reward-weighted regression and policy iteration can be explained by the domain naturally favouring the greedy updating as done by policy iteration. The source code for this experiment is available at `https://github.com/dylanashley/reward-weighted-regression`.

7. Related Work

The principle behind expectation-maximization (EM) was first applied to artificial neural networks in Von der Malsburg (1973). The reward-weighted regression (RWR) algorithm, though, originated in the work of Peters & Schaal (2007) which sought to bring earlier work of Dayan & Hinton (1997) to the domain of operational space control and reinforcement learning. However, Peters & 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 & Peters (2011). Wierstra et al. (2008a) went even further and also extended RWR to partially observable Markov decision processes, and Kober & Peters (2011) applied it to motor learning in robotics. Separately, Wierstra et al. (2008b) extended RWR to perform fitness maximization for evolutionary methods. Hachiya et al. (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 more suitable for off-policy RL, using deep neural networks as function approximators.

Other methods based on principles similar to RWR have been proposed. Neumann & Peters (2008), for example, proposed a more efficient version of the well-known fitted Q-iteration algorithm (Riedmiller, 2005; Ernst et al., 2005; Antos et al., 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 has guaranteed monotonic increases in the expected reward. Osa & Sugiyama (2018) subsequently proposed a hierarchical RL method that it is closely related to the episodic version of RWR by (Kober & Peters, 2011). Notably, all of the aforementioned works, as well as a number of other proposed similar RL methods (e.g., Peters et al. (2010), Neumann (2011), Abdolmaleki et al. (2018b), Abdolmaleki et al. (2018a)), are based on the EM framework of Dempster et al. (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. 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. Our RWR is on-policy, using only recent data to update the current policy—future work will study convergence in challenging off-policy settings (using all past data), which require corrections of the mismatch between state-distributions.

Acknowledgements

This work was supported by the ERC Advanced Grant (no: 742870) and by the Swiss National Supercomputing Centre (CSCS, project: s1090). We also thank NVIDIA Corporation for donating a DGX-1 as part of the Pioneers of AI Research Award and to 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. A. Relative entropy regularized policy iteration. *CoRR*, abs/1812.02256, 2018a.

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. A. 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, 2018b.

Antos, A., Munos, R., and Szepesvári, C. 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*, pp. 9–16. Curran Associates, Inc., 2007.

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

Billingsley, P. *Convergence of probability measures*. John Wiley & Sons, 2013.

Dayan, P. and Hinton, G. E. Using expectation-maximization for reinforcement learning. *Neural Comput.*, 9(2):271–278, 1997.

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

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.*, 6: 503–556, 2005.

Hachiya, H., Peters, J., and Sugiyama, M. 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*, pp. 469–484. Springer, 2009.

Hachiya, H., Peters, J., and Sugiyama, M. Reward-weighted regression with sample reuse for direct policy search in reinforcement learning. *Neural Comput.*, 23(11):2798–2832, 2011.

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

Neumann, G. 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*, pp. 817–824. Omnipress, 2011.

Neumann, G. and Peters, J. 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*, pp. 1177–1184. Curran Associates, Inc., 2008.

Osa, T. and Sugiyama, M. 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*, pp. 3860–3867. AAAI Press, 2018.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *CoRR*, abs/1910.00177, 2019.

Peters, J. and Schaal, S. 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*, pp. 745–750. ACM, 2007.

Peters, J. and Schaal, S. Learning to control in operational space. *The International Journal of Robotics Research*, 27(2):197–212, 2008.

Peters, J., Mülling, K., and Altun, Y. 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, 2010.

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

Riedmiller, M. A. 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*, pp. 317–328. Springer, 2005.

Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

Severini, C. 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*, 1910.

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. Mastering the game of go with deep neural networks and tree search. *Nat.*, 529(7587):484–489, 2016.

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

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

Sutton, R. S., Precup, D., and Singh, S. P. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.*, 112(1-2):181–211, 1999.

Ueno, T., Hayashi, K., Washio, T., and Kawahara, Y. 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*, pp. 2366–2374, 2012.

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

Wierstra, D., Schaul, T., Peters, J., and Schmidhuber, J. 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*, pp. 407–416. Springer, 2008a.

Wierstra, D., Schaul, T., Peters, J., and Schmidhuber, J. 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*, pp. 337–346. Springer, 2008b.

Wu, C. J. On the convergence properties of the em algorithm. *The Annals of statistics*, pp. 95–103, 1983.

A. Main results

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:*

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

Proof.

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

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:

$$\begin{aligned}
\int_{\mathcal{S}}\int_{\mathcal{A}} f(s,a)\log\pi(a|s)\, \mathrm{d}a\, \mathrm{d}s &= \int_{\mathcal{S}} f(s) \int_{\mathcal{A}} f(a|s)\log\pi(a|s)\, \mathrm{d}a\, \mathrm{d}s \\
&\leq \int_{\mathcal{S}} f(s) \int_{\mathcal{A}} f(a|s)\log f(a|s)\, \mathrm{d}a\, \mathrm{d}s,
\end{aligned}$$

where the last inequality holds for any policy π, since $\forall s \in S$ we have that $\int_{\mathcal{A}} f(a|s)\log\pi(a|s)\, \mathrm{d}a \leq \int_{\mathcal{A}} f(a|s)\log f(a|s)\, \mathrm{d}a$, 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:

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

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. $\square$

Theorem 4.1. *Fix $n > 0$ and let $\pi_0 \in \Pi$ be a policy*[9]*. Assume $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$, $R(s,a) > 0$. Define the operator $B : \Pi \to \Pi$ such that $\pi_{n+1} = B(\pi_n) = \frac{Q^{\pi_n}(s,a)\pi_n(a|s)}{V^{\pi_n}(s)}$. Then $\forall s \in \mathcal{S}$ we have that $V^{\pi_{n+1}}(s) \geq V^{\pi_n}(s)$ and $Q^{\pi_{n+1}}(s,a) \geq Q^{\pi_n}(s,a)$. Moreover, $\forall s \in \mathcal{S} : \mathrm{Var}_{a \sim \pi_n(a|s)}[Q^{\pi_n}(s,a)] > 0$ the inequalities above are strict.*

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 $\forall s \in \mathcal{S}$, $V^{\pi_{n+1},\pi_n}(s) \geq V^{\pi_n}(s)$, we get that $\forall s \in \mathcal{S}$, $V^{\pi_{n+1}}(s) \geq V^{\pi_n}(s)$. [10]

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

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

Now, let s be fixed:

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

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

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

$\square$

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

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

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 $n > n_0$:

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

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 et al., 1976)) which ensures uniform convergence of Q_n to Q_L. $\square$

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[11]. 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 B.3). In particular, it is enough to show that for all open sets $U \subset A$ such that $U \cap M(s) = \emptyset$ or such that $M(s) \subset U$, we have that $\liminf_n (\pi_n(\cdot|s) \cdot \mu_A) U \geq \pi_L(\cdot|s) U$.

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 $\liminf_n (\pi_n(\cdot|s) \cdot \mu_A) U \geq 1$, 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 $\lambda : A \to \mathbb{R}^+$ and closed sets A_ϵ and B_ϵ:

$$\lambda(a) := \frac{Q_L(a)}{Q_L(a^*)}, \quad A_\epsilon := \{a \in A | \lambda(a) \leq 1 - 2\epsilon\} \quad B_\epsilon := \{a \in A | \lambda(a) \geq 1 - \epsilon\},$$

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 B.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:

$$\frac{Q_L(a)}{Q_L(b)} = \frac{\frac{Q_L(a)}{Q_L(a^*)}}{\frac{Q_L(b)}{Q_L(a^*)}} \leq \frac{Q_L(a)}{Q_L(a^*)(1-\epsilon)} \leq \frac{1-2\epsilon}{1-\epsilon} = 1 - \frac{\epsilon}{1-\epsilon} =: \alpha_1 < 1.$$

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 $n \geq n_0$, 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 $n \geq n_0$, $a \in A_\epsilon$ and $b \in B_\epsilon$:

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

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

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

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 $Q_i(s,a) > 0$). Since $A_\epsilon \times B_\epsilon$ is a compact set, there exists c_m such that $c \leq c_m$. Thus we have that for all $n > n_0$:

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

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:

$$(\pi_n(\cdot|s) \cdot \mu_A)(A_\epsilon) \times (\mu_A B_\epsilon) \leq \alpha^n c_m (\pi_n(\cdot|s) \cdot \mu_A)(B_\epsilon) \times (\mu_A A_\epsilon).$$

[11] 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:

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

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 $C := \{a \in A | \lambda(a) > 1-\epsilon\} \subset B_\epsilon$. Then $\mu_A(B_\epsilon) \ge \mu_A(C) > 0$ (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:*

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

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

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

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

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

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 B.3). The second term can be arbitrarily small since Lemma 5.1 ensures uniform convergence of Q_n to Q_L. □

B. Convergence - lemmas

Lemma B.1. *(on level sets of continuous function on compact metric space) Let (X, d) be a compact metric space and $f : X \to \mathbb{R}$ be a continuous function. Furthermore, let $m := \max_{x\in X} f(x)$ and $F := \{x \in X : f(x) = m\}$. Then for every open $U \subset X$, $F \subset U$ there exists a $\delta > 0$ such that $\{x \in X : f(x) > m - \delta\} \subset U$.*

Proof. First notice that m is defined correctly as f is a continuous function on a compact space and therefore always has a maximum. Also, note that F is compact and $F \neq \emptyset$. Assume that f is not constant (otherwise the conclusion holds trivially). Now consider an open set U and $F \subset U$. If $U = X$, the Lemma holds trivially, thus assume $U \neq X$. From compactness of F we conclude that F is 2ϵ isolated from $U^C := X \setminus U$ for some $\epsilon > 0$. Let us define $V := \{x \in X : d(x, F) < \epsilon\} \subset U$ an open set. Further, define $m' := \max f(X \setminus V)$. Notice that the definition is correct since $X \setminus V$ is closed and therefore compact and also $X \setminus V \neq \emptyset$ as $X \setminus V \supset U^C \neq \emptyset$. Further, $m' < m$ as $X \setminus V$ and F are disjoint ($F \subset V$). Define $\delta := \frac{m-m'}{2}$. It remains to verify that $W := \{x \in X : f(x) > m - \delta\} = \{x \in X : f(x) > m - \frac{m+m'}{2}\} \subset U$. Notice that $f(W) > \frac{m+m'}{2} > m' \geq f(X \setminus V)$. Thus W and $X \setminus V$ must be disjoint and therefore $W \subset V$ $(\subset U)$. $\square$

Lemma B.2. *(quotient of a metric space by a compact subset) Let (X, d) be a metric space and $F \subset X$ compact. Furthermore, let τ denote the topology on X induced by the metric d. Define the equivalence:*

$$(\forall x, y \in X \times X) : x \sim y \iff (x = y \vee (x \in F \wedge y \in F)).$$

Define a (factor) quotient space $\tilde{X} := X/\sim$ and $\nu : X \to \tilde{X}$ the canonical projection $\nu(x) := [x]_\sim$.

1. *Denote by $\tilde{\tau}$ the quotient topology on $\tilde{X}$ (induced by τ and ν). Then it holds:*

$$\tilde{\tau} = \{\nu(U) : U \in \tau, (U \cap F = \emptyset \vee F \subset U)\}.$$

2. *Further, the function $\tilde{d} : \tilde{X} \times \tilde{X} \to \mathbb{R}^+$*

$$\tilde{d}([x]_\sim, [y]_\sim) := d(x, y) \wedge (d(x, F) + d(y, F))$$

defines a metric on $\tilde{X}$ and the topology induced by metric $\tilde{d}$ agrees with $\tilde{\tau}$.

3. *(continuous functions) Let $\tilde{f} : \tilde{X} \to \mathbb{R}$ be a function on $\tilde{X}$. Than it holds:*

$$\tilde{f} \in C(\tilde{X}) \iff \tilde{f} \circ \nu \in C(X),$$

so there is a one to one correspondence between continuous functions on $\tilde{X}$ ($C(\tilde{X})$) and continuous functions on X, which are constant on F (which allow factorisation through ν):

$$\{f \in C(X) : \exists \tilde{f} \in \mathbb{R}^{\tilde{X}} : f = \tilde{f} \circ \nu\} = \{f \in C(X) : \exists c_f \in \mathbb{R} : f|_F = c_f\}.$$

Proof. During the proof, we will assume $F \neq \emptyset$. For the case $F = \emptyset$, the Lemma holds trivially.

1. The quotient topology $\tilde{\tau}$ is the finest topology in which is ν continuous. Suppose $\tilde{U} \in \tilde{\tau}$ (is open in $\tilde{\tau}$) then $U := \nu^{-1}(\tilde{U})$ must be open (otherwise ν would not be continuous). Further, due to the equivalence defined, the pre-images under ν cannot contain F only partially. They either contain the whole F, or are disjoint with F (in the first case we get $F \subset U$ and in the second one we get $F \cap U = \emptyset$). This gives us the inclusion $\tilde{\tau} \subset \{\nu(U) : U \in \tau, (U \cap F = \emptyset \vee F \subset U)\}$. For the reverse inclusion, assume we have $U \in \tau$. Assume $F \subset U$. Then the pre-image $\nu^{-1}(\nu(U)) = U$ (the result would be different from U just when U includes F only partially), which is an open set. Thus, from the fact that $\tilde{\tau}$ is the finest topology in which ν is continuous, it follows that $\nu(U) \in \tilde{\tau}$. Similarly for $U \cap F = \emptyset$.

2. Now we aim to show that $\tilde{d}$ is a metric on $\tilde{X}$. Notice that the definition is correct in the sense that it does not depend on the choice of representants. When we assume that both x, y are not in F, then the choice of representants is unique. So assume that, for example, $x \notin F$,$y \in F$. Then we can choose another representant for $[y]_\sim$, but then $\tilde{d}([x]_\sim, [y]_\sim) = d(x, F)$ is independent of y. Similarly, if x, y are both in F then $\tilde{d}([x]_\sim, [y]_\sim) = 0$ which again does not depend on choice of the representants. Non-negativity and symmetry trivially holds. First, we consider the property:

$$\tilde{d}([x]_\sim, [y]_\sim) = 0 \iff [x]_\sim = [y]_\sim (\iff x \sim y).$$

Assume $x \sim y$, then either $x = y$ or $x, y \in F$. In both cases $\tilde{d}([x]_\sim, [y]_\sim)$ becomes zero. Assume $\tilde{d}([x]_\sim, [y]_\sim) = 0$, then $d(x, y) = 0$ or $d(x, F) + d(y, F) = 0$, where in the first case we get $x = y$ and in the second case (here we use that F is closed) $x, y \in F$. Thus $x \sim y$. The Triangle inequality holds too. The proof follows easily, but is omitted for brevity (it consists of checking multiple cases).

Finally, we have to show that the topology induced by $\tilde{d}$ agrees with $\tilde{\tau}$ (here we will need compactness of F). First we show that every open set in $\tilde{\tau}$ is also open in the topology induced by $\tilde{d}$. Let us consider an open set $\tilde{U} \in \tilde{\tau}$. Now let us fix an arbitrary point $\tilde{x} \in \tilde{U}$. It suffices to show that there exits $r > 0$ such that open ball $U_r(\tilde{x}) := \{\tilde{y} \in \tilde{X} : \tilde{d}(\tilde{x}, \tilde{y}) < r\} \subset \tilde{U}$. From $\tilde{U} \in \tilde{\tau}$ there exists $U \in \tau$ such that $\nu(U) = \tilde{U}$ and moreover $F \subset U$ or $F \cap U = \emptyset$.

Fix $x \in X$ such that $[x]_\sim = \tilde{x}$. We start by considering the case $F \subset U$ and $x \in F$. Notice that the metric reduces to $\tilde{d}([x]_\sim, [y]_\sim) = d(y, F)$. Compactness of F guarantees that there exists $\epsilon > 0$ such that F is ϵ isolated from $U^c := X \setminus U$. So it suffices to choose $r := \epsilon$.

For the second case we consider $F \subset U$ and $x \notin F$. As $U \setminus F$ is open, there exists a $\delta > 0$ such that $U_\delta(x) := \{y \in X : d(x,y) < \delta\} \subset U \setminus F$. Note that $\nu(U_\delta(x))$ is an open set in $\tilde{\tau}$ (has open pre-image and does not contain F) on which the metric simplifies to $\tilde{d}([x]_\sim, [y]_\sim) = d(x,y) \quad (< \delta)$. We conclude that it is an open ball in $\tilde{d}$, whole lying in $\tilde{U}$. So it suffices to put $r := \delta$.

As final case, assume $F \cap U = \emptyset$. This actually reduces to the second case we already considered.

Finally, for the opposite inclusion it suffices to show that every open ball in $\tilde{d}$ is an open set in $\tilde{\tau}$. Thus let us fix an $x \in X$ and positive $r > 0$ and set $\tilde{U} := U_r(\tilde{x})$. In order for $\tilde{U}$ to be open in $\tilde{\tau}$, it must have open pre-image

$$\begin{aligned}\nu^{-1}(\tilde{U}) &= \{y \in X : \nu(y) \in \tilde{U}\} = \{y \in X : d(x,y) \wedge (d(x,F) + d(y,F)) < r\} \\ &= \{y \in X : d(x,y) < r\} \cup \{y \in X : (d(x,F) + d(y,F)) < r\},\end{aligned}$$

where we end up with a union of two sets, both open in τ, which is again open. Thus $\nu^{-1}(\tilde{U})$ is open, so $\tilde{U}$ is open (from $\tilde{\tau}$ is the finest topology in which ν is continuous).

3. (Continuous functions) Assume $\tilde{f} \in C(\tilde{X})$. Since ν is continuous, then $\tilde{f} \circ \nu$ is continuous (composition of continuous maps). For the opposite implication, assume $f := \tilde{f} \circ \nu$ is continuous. We have to show that $\tilde{f}$ is continuous. Thus fix an arbitrary open set $V \subset \mathbb{R}$. We have to show that the pre-image $\tilde{U} := \tilde{f}^{-1}(V)$ is open. We know that $U := f^{-1}(V)$ is open from the continuity of f and that $U = f^{-1}(V) = \nu^{-1}(\tilde{U})$, that means that the pre-image of $\tilde{U}$ under ν is open, but $\tilde{\tau}$ is the finest topology in which ν is continuous, therefore $\tilde{U}$ has to be open. $\square$

Lemma B.3. *(Adaptation of Portmanteau theorem conditions to relative weak convergence) Let (X, d), $(\tilde{X}, \tilde{d})$, F, ν be like above. Let $P, P_n, n \in \mathcal{N}$ be probability measures on $\mathcal{B}(X)$. Then following conditions are equivalent:*

1. *$P_n \to^{w(F)} P$.*
2. *For all continuous $f : X \to \mathbb{R}$ that are constant on F it holds that $P_n f \to Pf$.*
3. *For all $U \subset X$ open satisfying $U \cap F = \emptyset$ or $F \subset U$ it holds that $\liminf P_n U \geq PU$.*

Proof. First we show equivalence of 1. and 2. Point 1. is equivalent to $\nu P_n \to^w \nu P$, (definition 1) which is equivalent to (using Portmanteau theorem):

$$(\forall \tilde{f} \in C(\tilde{X})) : (\nu P_n)\tilde{f} \to (\nu P)\tilde{f},$$

what can be rewritten using definition of image measure:

$$(\forall \tilde{f} \in C(\tilde{X})) : P_n(\tilde{f} \circ \nu) \to P(\tilde{f} \circ \nu).$$

But from Lemma B.2 we already know that there is a one to one correspondence between functions in $C(\tilde{X})$ and functions in $C(X)$, which factors through ν (are constant on F). Thus it is equivalent to:

$$(\forall f \in C(X)) : ((\exists c_f \in \mathbb{R}) : f|_F = c_f) \implies (P_n f \to Pf).$$

Finally, we show equivalence of 1. and 3. Again, point 1. is equivalent (using Portmanteau theorem) to:

$$(\forall \tilde{U} \subset \tilde{X} \text{ open}) : \liminf(\nu P_n)\tilde{U} \geq (\nu P)\tilde{U}.$$

Using the definition of image measure and the one to one correspondence (see Lemma B.2) between all open sets in $\tilde{X}$ and open sets in X we have that at least one of the two conditions $U \cap F = \emptyset$, $F \subset U$ is satisfied. This concludes the result. $\square$

C. Counterexample

Counterexample. Consider the simple two-armed bandit shown in Figure 2 with actions a_0 and a_1, and with $P(r = 1|a_0) = 1$, $P(r = 0|a_1) = 2/3$, and $P(r = 2|a_1) = 1/3$. Note that $q(a_0) = 1 > q(a_1) = 2/3$. Thus the optimal policy always takes action a_0. Now, after applying the transformation $u(r) = e^{\log(3)\, r} = 3^r$, we get $P(u(r) = 3|a_0) = 1$, $P(u(r) = 1|a_1) = 2/3$, and $P(u(r) = 9|a_1) = 1/3$. Hence, under transformation u, we have $q(a_0) = 3 < q(a_1) = 11/3$. So the optimal policy under the transformed rewards always takes action a_1, which is sub-optimal, given the original problem.

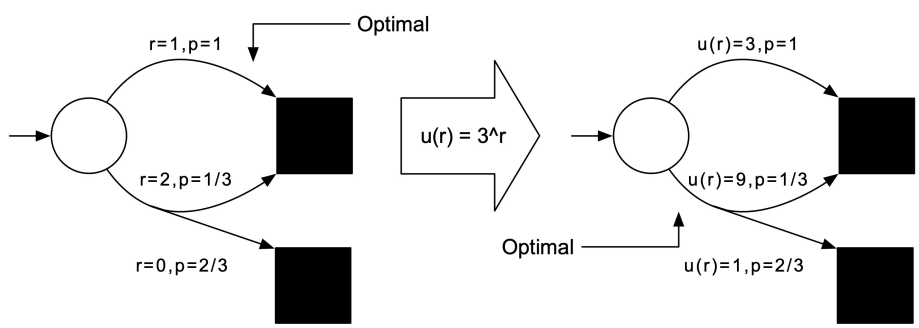

Figure 2. Counterexample demonstrating how applying a naive transformation of the reward function of an MDP may change the optimal policy.