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/USI/SUPSI, Lugano, Switzerland [2]NNAISENSE, Lugano, Switzerland [3]KAUST, Thuwal, Saudi Arabia

{struplm, francesco, dylan.ashley}@idsia.ch, rupesh@nnaisense.com, juergen@idsia.ch

Summary

- Reward-Weighted Regression (RWR) uses Expectation-Maximization for Reinforcement Learning
- Leads to a widely studied family of simple algorithms that are known to yield monotonic policy improvement
- **Open Question**: do these algorithms learn the **optimal** policy?

We present the first proof that RWR converges to a global optimum when no function approximation is used

Background

$\mathcal{M} = (\mathcal{S}, \mathcal{A}, p_T, R, \gamma, \mu_0)$ an MDP where:

- $\mathcal{S} \subset \mathbb{R}^{n_S}$ ($\mathcal{A} \subset \mathbb{R}^{n_A}$) is a **compact** state (action) space with measurable structure $(\mathcal{S}, \mathcal{B}(\mathcal{S}), \mu_S)$, $((\mathcal{A}, \mathcal{B}(\mathcal{A}), \mu_A))$ where μ_S (μ_A) is a fixed, finite, strict positive reference measure. states and actions with discrete and cont. components
- $p_T(s'|s,a)$ is a density of the transition kernel, which is assumed **continuous in total variation**.
- $R(s,a)$ is a **continuous, bounded, positive** reward function.
- $\gamma \in (0,1)$ discount factor, $\mu_0(s)$ initial state probability density.

• return $R_t := \sum_{k=0}^{\infty} \gamma^k R(s_{t+k+1}, a_{t+k+1})$, • state-value function $V^\pi(s) := \mathbb{E}_\pi[R_t | s_t = s]$, • action-value function $Q^\pi(s,a) := \mathbb{E}_\pi[R_t | s_t = s, a_t = a]$.

Reward-Weighted Regression (RWR)

RWR [1, 3, 2] starts from an initial policy π_0 and generates a sequence of policies (π_n). Each iteration consists of two steps:

1. a batch of episodes is generated using the current policy π_n,
2. a new policy π_{n+1} is fitted to a sample representation of π_n, weighted by the return R_t.

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

which is equivalent to (see Theorem 3.1 for details)

$$\pi_{n+1}(a|s) = \frac{Q^{\pi_n}(s,a)\pi_n(a|s)}{V^{\pi_n}(s)}. \quad (2)$$

Monotonic Improvement Theorem (MIT)

(see Theorem 4.1) Fix arbitrary $s \in \mathcal{S}$. The following holds

$$V^{\pi_{n+1}}(s) \geq V^{\pi_n}(s), \quad (\forall a \in \mathcal{A}) : Q^{\pi_{n+1}}(s,a) \geq Q^{\pi_n}(s,a). \quad (3)$$

Moreover, if $\mathrm{Var}_{a \sim \pi_n(a|s)}[Q^{\pi_n}(s,a)] > 0$ the first inequality above is strict.

When can there be no improvement?

- Deterministic policies
- Stochastic policies which are greedy of their action-value function (optimal policies)

Convergence Results

Problems/Motivation :

- Desirable limit-points (optimal policies) are not always dominated by reference measure μ_A. Note: E.g., consider μ_A being Lebesgue measure, π_n being densities with respect to μ_A, and optimal policy π^* being a kernel concentrating concentrating all mass in single action for some state.
- Optimal policy $\pi^*(\cdot|s)$ can be non-unique, thanks to $\arg\max Q^*(s,\cdot)$ consisting of multiple points (Q^* stands for optimal value function).

Used notion of convergence:(For details see Definition 1 in the paper.) Let $\mathcal{A}$ be a metric space, $F \subset \mathcal{A}$ a compact subset, ν the quotient map $\nu : \mathcal{A} \to \mathcal{A}/F$ ($\mathcal{A}/F$ being topological factor). **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:

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

Trivial Facts: Boundedness of value functions $V_n(s) < B_V, Q_n(s,a) < B_V$, $B_V < +\infty$ and MIT implies existence of point-wise limits V_L and Q_L:

$$V^{\pi_n}(s) \nearrow V_L(s) \leq B_V < +\infty$$
$$Q^{\pi_n}(s,a) \nearrow Q_L(s,a) \leq B_V < +\infty,$$

but to prove something about limiting properties of the sequence (π_n) is a difficult problem (see Convergence Results section in the paper).

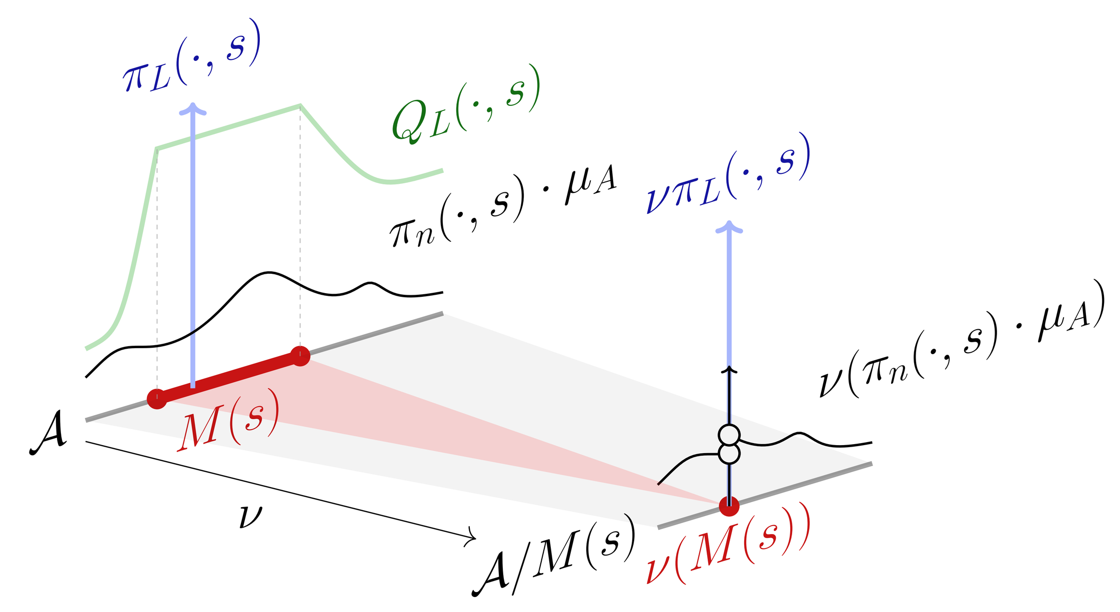
Further notation:

- $M(s) := \arg\max Q_L(s,\cdot)$
- Π_L the set of all probability kernels, greedy with respect to Q_L, i.e. $\pi_L \in \Pi_L \implies (\forall s \in \mathcal{S}) \pi_L(\cdot|s)(M(s)) = 1$
- $\pi_n(\cdot|s) \cdot \mu_A$ stands for the probability kernel formed by reference measure μ_A and the conditional density π_n

Convergence Theorem (see Theorem 5.1):
Let the initial policy π_0 be positive and continuous in actions. Then

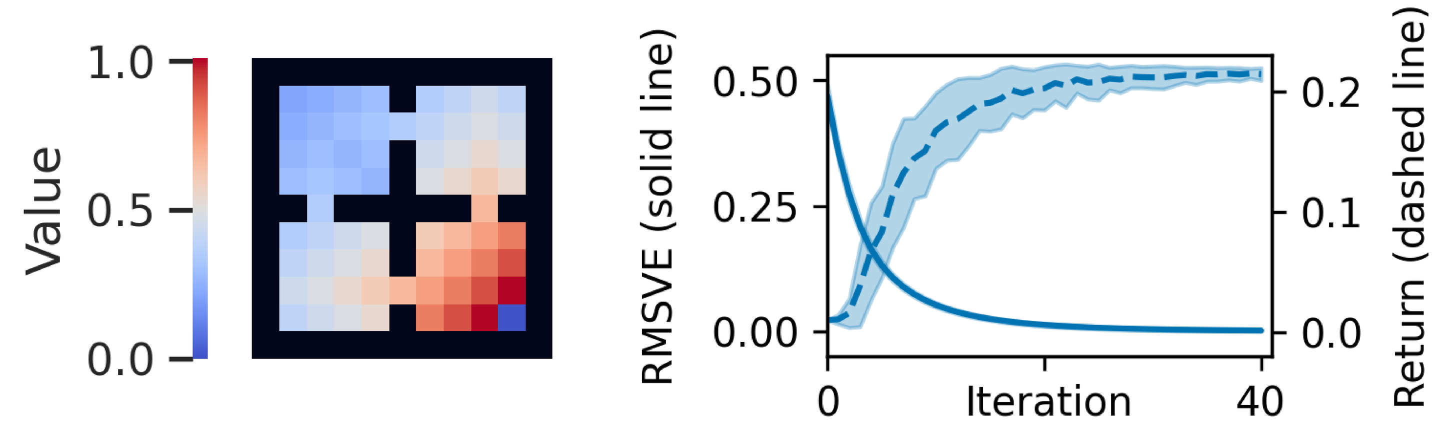
$$(\forall \pi_L \in \Pi_L, \forall s \in \mathcal{S}) : \quad \pi_n(\cdot|s) \cdot \mu_A \to^{w(M(s))} \pi_L(\cdot|s),$$

where Π_L is a set of optimal policies for the MDP. Moreover, V_L, Q_L are the optimal state and action value functions.



Demonstration of RWR Convergence

Convergence of RWR on a modified four-room gridworld domain:



Conclusion

- We provided the **first global convergence proof for RWR** in absence of reward transformation and function approximation.
  - assumes general **compact** state and action spaces $\implies$ **robotic control**.
  - **provides solid theoretical ground** for both previous and future works on RWR [1, 3, 2] and understanding similar algorithms
  - **Techniques developed in the proof are further applicable.** Demonstrated on proof of R-linear convergence order for finite case.
- Established **relationship between improvement of state value function and variance of action-value function** with respect to policy action distribution.
- We also highlighted that **nonlinear reward transformations used in prior work can lead to problems**, potentially resulting in changes to the optimal policy.
- Discussion of undiscounted setting allowing for zero rewards.
- Adaptation of Portmanteau theorem for relative weak convergence.
- Established R-linear convergence for finite case.
- Provided two examples: one for finite case exhibiting Q-linear rate, and one for continuous case exhibiting sublinear order.

Future Work & References

- RWR's convergence under **function approximation.**
- RWR's convergence in **off-policy** settings (Importance Sampling)

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

[2] X. B. Peng, A. Kumar, G. Zhang, and S. Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019.

[3] J. Peters and S. Schaal. Reinforcement learning by reward-weighted regression for operational space control. In Z. Ghahramani, editor, *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*, pages 745--750. ACM, 2007.

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.