Regularized Q-learning

Neurips 2024

Han-Dong Lim, Donghwan Lee

KAIST, Electrical Engineering

Regularized Q-learning

Contents

1. Motivation
2. Background
3. Main Result
4. Experiments
5. Conclusion

Motivation

Convergence of RL algorithms

- RL algorithms shows good performance in practice but its theoretical **convergence** is not well-established even in the **linear function approximation** scheme.

- Can we develop a **convergent** Q-learning algorithm under the **linear function approximation** scheme?

Contributions

1. We propose a Q-learning algorithm that is **convergent** with **linear function approximation**.

2. The convergence of Q-learning with l_2-regularization is established under mild conditions, and its proof is based on the switched system analysis.

3. We analyze the solution of the projected optimal Bellman equation with regularization, where the iterate of the algorithm converges to.

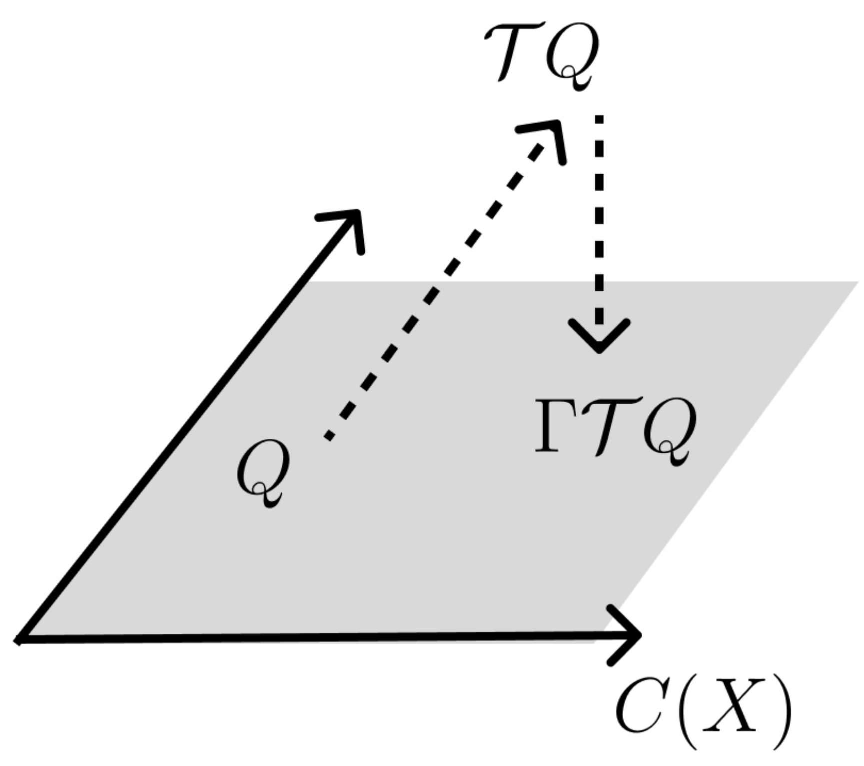
4. Finally, experimental results are provided.

Q-learning with linear function approximation

- We want to approximate the Q-function : $Q^{\pi}(s,a) \approx x(s,a)^{\top}\theta$ where $\theta, x(s,a) \in \mathbb{R}^h$

- The result of Bellman operator may not be in the column space of X. Therefore, we project it back to the column space of X.

- Illustration or projection on to the column of X:



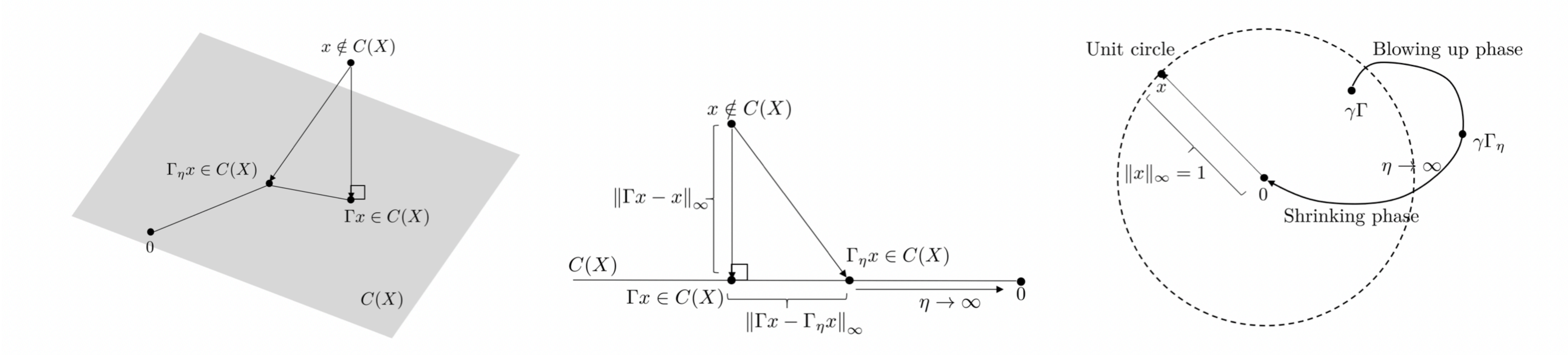
- Projected (Optimal) Bellman Equation :

$$X^{\top}DX - \gamma X^{\top}DP\Pi_{X\theta}X\theta = X^{\top}DR\,.$$

- Does it have a solution?
  - An example of non-existence of the solution was provided by De Farias et al., 2000.

Regularized Projected (optimal) Bellman equation

- Regularized Projected Bellman Equation : $X^\top DX + \eta I - \gamma X^\top DX \Pi_{X\theta} X\theta = X^\top DR\,.$



(a) Regularized projection: $C(X)$ means the range space of X (b) Regularized projection: One dimensional case (c) Boundedness of the projection

- When does it have a solution?
    - A simple condition is $\eta > ||X^\top DX||_\infty + \gamma ||X^\top||_\infty ||DPX||_\infty$
    - Under the assumption that $\max(||X||_\infty, ||X^\top||_\infty) \leq 1, \eta > 2$ is sufficient.
- We provide a simple example where RPBE admits a solution but PBE does not in Appendix A.14 in Lim et al., 2024.

Regularized Projected (optimal) Bellman equation

Error bound on the solution

- Simple algebraic inequalities yield

$$||X\theta_\eta^* - Q^*||_\infty \leq \frac{1}{1-\gamma||\Gamma_\eta||_\infty}||\Gamma_\eta Q^* - Q^*||_\infty .$$

- As $\eta \to 0$, the above inequality reduces to the conventional error bound for Q-learning with linear function approximation in Melo et al., 2008.

- As $\eta \to \infty$, we get $\theta_\eta^* \to 0$.

- With small $\eta \approx 0$, and if the function approximation error is low, the overall error bound is small:

$$||X\theta_\eta^* - Q^*||_\infty \leq \frac{1}{1-\gamma||\Gamma_\eta||_\infty}||\Gamma_\eta Q^* - Q^*||_\infty \leq \frac{1}{1-\gamma||\Gamma_\eta||_\infty}\left(||\Gamma_\eta Q^* - \Gamma Q^*||_\infty + ||\Gamma Q^* - Q^*||_\infty\right).$$

Regularized Q-learning

Algorithm

1. Initialize $\theta_0 \in \mathbb{R}^h$
2. Set the step-size $(\alpha_k)_{k=0}^{\infty}$ and the behavior policy
3. for $k = 0,1,\ldots,$do

    Sample $s_k \sim d^{\mu}$ and $a_k \sim \mu$.

    Sample $s_k' \sim P(s_k, a_k, \cdot\,)$ and $r_{k+1} = r(s_k, a_k, s_k')$.

    Update $\theta_{k+1} = \theta_k + \alpha_k(x(s_k, a_k)\delta_k - \eta\theta_k)$

4. End For

Theorem 5.2 (Informal)

Suppose Assumption 1holds and η satisfies condition 1.

Then, we have $\theta_k \to \theta_\eta^*$ with probability one.

- Assumption 1: Standard assumptions on Markov chain and the feature matrix is non-negative, and the column vectors are orthogonal.
- Condition 1: $\eta > \min \left\{ \gamma ||X^\top D||_\infty ||X||_\infty + ||X^\top DX||_\infty, \lambda_{\max}(C) \left(\max_{\pi,sa} \frac{\gamma d^\top P^\pi(e_a \otimes e_s)}{2d(s,a)} - \frac{2-\gamma}{2} \right) \right\}$

Convergence proof

Switched System Analysis

- Lee et al., 2020 developed an ODE analysis framework for Q-learning based on switched system theory:

- We apply the Borkar-Meyn Theorem which is a tool to prove convergence of stochastic algorithm by its corresponding ODE:

$$\frac{d}{dt}\theta_t = (-X^\top DX - \eta I + \gamma X^\top DP\Pi_{X\theta_t}X)\theta_t + \gamma X^\top DP(\Pi_{X\theta_t} - \Pi_{X\theta_\eta^*})X\theta_\eta^*, \quad \theta_0 \in \mathbb{R}^h.$$

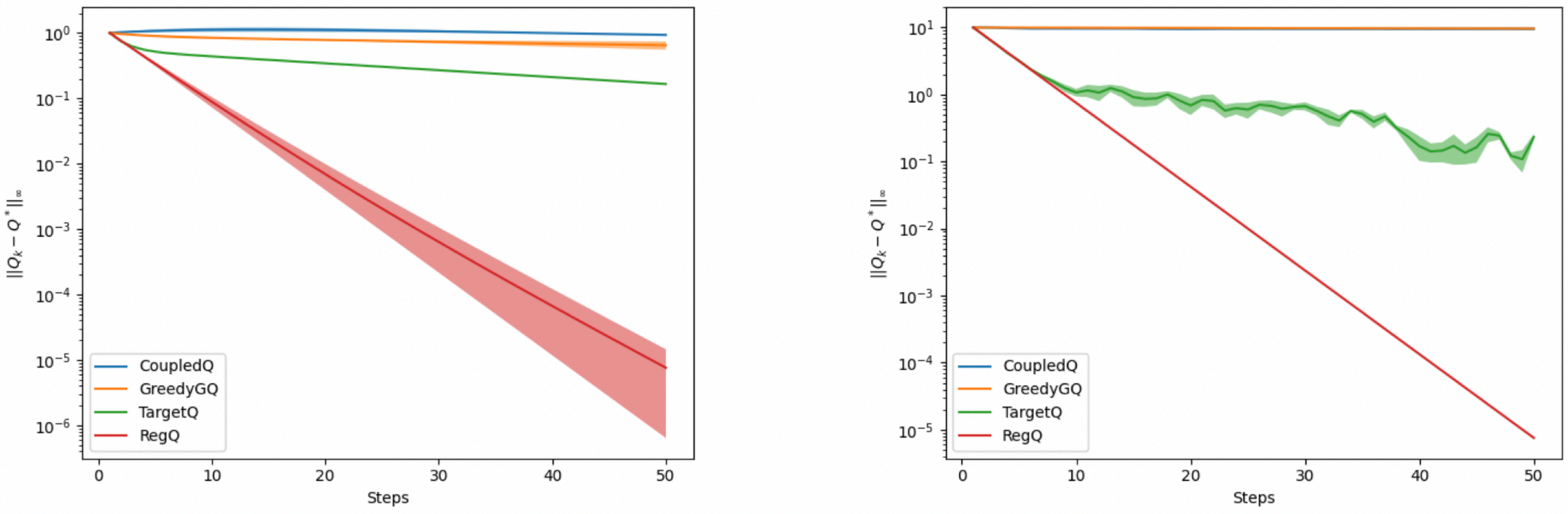
  - The system can be viewed as switched affine linear system, of which stability is difficult to analyze.

- Construct a lower and upper comparison system such that $\theta_t^u \geq \theta_t \geq \theta_t^l$.

$$\frac{d}{dt}\theta_t^u = (-X^\top DX - \eta I + \gamma X^\top DP\Pi_{X\theta_t^u}X)\theta_t^u, \quad \frac{d}{dt}\theta_t^l = (-X^\top DX - \eta I + \gamma X^\top DP\Pi_{X\theta_\eta^*}X)\theta_t^l$$

  - The systems can be viewed as switched linear system and linear time-invariant system.

Experiments



(a) Results in $\theta \to 2\theta$

(b) Results in Baird seven star counter example

- $\theta \to 2\theta$ [Tsitsiklis and Van Roy, 1996] and Baird example [Baird, 2000] is typical example where Q-learning diverges.
- Regularized Q-learning is convergent and shows fast convergence rate.

Conclusion and Future works

- We have proposed regularized Q-learning which is convergence under the linear function approximation scheme and mild assumptions.

- We have analyzed the regularized (projected) optimal Bellman equation.

- As a future work, we can consider neural network approximation case, which is closer to practice.

References

De Farias, Daniela Pucci, and Benjamin Van Roy. "On the existence of fixed points for approximate value iteration and temporal-difference learning." Journal of Optimization theory and Applications 105 (2000): 589-608.

Lee, Donghwan, and Niao He. "A unified switching system perspective and convergence analysis of Q-learning algorithms." Advances in Neural Information Processing Systems 33 (2020): 15556-15567.

Melo, Francisco S., Sean P. Meyn, and M. Isabel Ribeiro. "An analysis of reinforcement learning with function approximation." Proceedings of the 25th international conference on Machine learning. 2008.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic program- ming. Machine Learning, 22(1):59–94, 1996

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In Machine Learning Proceedings 1995, pages 30–37. Elsevier, 1995.