Convergence Issues of Q-Learning with Function Approximation

Ziniu Li

The Chinese University of Hong Kong, Shenzhen

Table of contents

1. Introduction

2. Divergence of Q-Learning with Function Approximation

3. Target Q-Learning

Overview

- We use simple examples to illustrate the divergence issues of Q-Learning with function approximation.
- We introduce the practical technique to address this issue: target network.
- We discuss why this technique can work.

Introduction

Markov Decision Processes

- Infinite-horizon MDPs with time-independent dynamics $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R)$.
- Bellman Optimality Equation:

$$Q^\star(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q^\star(s', a') \right], \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

- Bellman operator $\mathcal{T}$:

$$\mathcal{T}(Q)(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s', a') \right].$$

  However, in practice, we do not know P so that $\mathcal{T}$ is not applicable.
- γ-contractility:

$$\max_{(s,a)} \left| \mathcal{T}(Q_1)(s, a) - \mathcal{T}(Q_2)(s, a) \right| \leq \gamma \max_{(s,a)} \left| Q_1(s, a) - Q_2(s, a) \right|.$$

Q-Learning

Assume we have access to the stream data (s_t, a_t, r_t, s_{t+1}).

- Q-learning:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \eta_t \left[r_t + \gamma \max_{a' \in \mathcal{A}} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) \right].$$

- If we try to linearly parameterize $Q(s, a)$ over the feature ϕ. That is, $Q_t(s_t, a_t) = \phi(s_t, a_t)^\top w$. Then, Q-learning becomes:

$$w_{t+1} = w_t + \eta_t \left[r_t + \gamma \max_{a' \in \mathcal{A}} \langle \phi(s_{t+1}, a'), w_t \rangle - \langle \phi(s_t, a_t), w_t \rangle \right] \phi(s_t, a_t).$$

Open Problem

Question: Does Q-Learning converge with any (linear) function approximation?

Answer: No! (See the next page for the counter-example.)

Divergence of Q-Learning with Function Approximation

Baird's Example

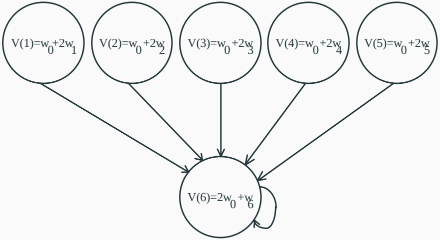
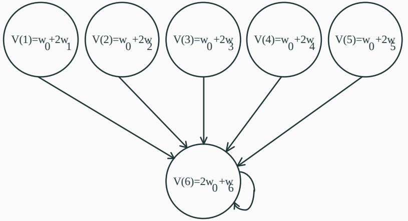


Figure 1: Baird's example [III, 1995].

- Only one action at each state and reward is 0.
- $Q(s, a) = V(s)$ under this case.

Baird's Example



$$\Phi = \begin{bmatrix} 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 2 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \phi_1^\top \\ \phi_2^\top \\ \phi_3^\top \\ \phi_4^\top \\ \phi_5^\top \\ \phi_6^\top \end{bmatrix} \in \mathbb{R}^{6\times 7}$$

$$w = \begin{bmatrix} w^0 & w^1 & w^2 & w^3 & w^4 & w^5 & w^6 \end{bmatrix}^\top \in \mathbb{R}^7.$$

Baird's Example (Important!)

- We have

$$V(1) = w^0 + 2w^1, V(2) = w^0 + 2w^2, \cdots, V(6) = 2w^0 + w^6$$

- Let $\gamma = 0.99$ and $\eta = 0.1$. Suppose $w^0 = 1$ and $w^1 = w^2, \cdots, w^6 = 0$. Q-Learning runs:
  - state s_1: $w_2 = w_1 + \eta(r + \gamma V(6) - V(1))\phi_1 \Longrightarrow w^0 \uparrow, w^1 \uparrow, V(6) \uparrow$

    $$\Delta_1 = r + \gamma V(6) - V(1) = 0.980, w^0 = 1.098, w^1 = 0.196$$

  - state s_2: $w_3 = w_2 + \eta(r + \gamma V(6) - V(2))\phi_2 \Longrightarrow w^0 \uparrow, w^2 \uparrow, V(6) \uparrow$

    $$\Delta_2 = r + \gamma V(6) - V(2) = 1.076, w^0 = 1.206, w^2 = 0.215$$

  - $\ldots\ldots$
  - state s_6: $w_7 = w_6 + \eta(r + \gamma V(6) - V(6))\phi_6 \Longrightarrow w^0 \downarrow, w^6 \downarrow, V(6) \downarrow$

    $$\Delta_6 = r + \gamma V(6) - V(6) = -0.032, w^0 = 1.590, w^6 = -0.003$$

- Repeating the above cycle, w^0 diverges.

Remark on Baird's Example

- Baird's example suggests that Q-Learning + Function Approximation may diverge.

- The divergence is not due to step size or to uncertainties about the environment.
  (we numerically observe that diverges happens even though the step size is very small)

- Divergence is mainly because the extrapolation changes the "target labels".

Question on Baird's Example

Question: Can Q-Learning converge if we use the exact solution rather than taking a gradient step?

Answer: No! (See the next page for Tsitsiklis and Van Roy's counter-example)

Tsitsiklis and Van Roy's Example

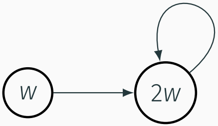


Figure 2: Tsitsiklis and Van Roy's Example [Tsitsiklis and Van Roy, 1997].

(Full-step) Q-learning:

$$w_{k+1} = \operatorname*{argmin}_{w \in \mathbb{R}} (w - 2\gamma w_k)^2 + (2w - \gamma 2 w_k)^2 = \frac{6-4}{5} \gamma w_k.$$

The sequence $\{w_k\}$ diverges when $\gamma > 5/6$ and $w_0 \neq 0$.

Tsitsiklis and Van Roy's Example

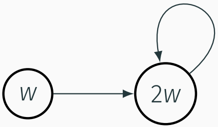


Figure 3: Tsitsiklis and Van Roy's Example [Tsitsiklis and Van Roy, 1997].

(One-step) Q-learning:

$$w_{2k+1} = (1 + 2\gamma\eta - \eta) w_{2k},$$
$$w_{2k+2} = (1 + 4\gamma\eta - 4\eta) w_{2k+1}.$$

Key factor: $(1 + 2\gamma\eta - \eta) \cdot (1 + 4\gamma\eta - 4\eta)$.

When γ is sufficiently large (i.e., $\gamma > 5/6$), and step size η is small (i.e., $0 < \eta < (5 - 6\gamma)/(8\gamma^2 - 12\gamma + 4)$, the sequence $\{w_k\}$ diverges.

Remark

- Tsitsiklis and Van Roy's Example is different from Baird's Example because the former is not over-parameterized.

- Tsitsiklis and Van Roy's Example highlights the off-policy issue: we should update states according to its stationary distribution.
    - In that example, we should update the state "$2w$" more than the state "w".

Target Q-Learning

Literature Review

- Previous examples suggest Q-Learning with function approximation is hard to train.
- However, many deep RL algorithms *work* in practice. Why?
- Both claims are true but people often ignore (or underestimate) an important technique used to train deep RL: target network.

Playing atari with deep reinforcement learning
V Mnih, K Kavukcuoglu, D Silver, A Graves… - arXiv preprint arXiv …, 2013 - arxiv.org
We present the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using reinforcement learning. The model is a convolutional …
被引用次数：7325 相关文章 所有 34 个版本

Figure 4: NIPS 2013 Workshop. It solves 6 tasks.

Human-level control through deep reinforcement learning
V Mnih, K Kavukcuoglu, D Silver, AA Rusu, J Veness… - nature, 2015 - nature.com
The theory of reinforcement learning provides a normative account 1, deeply rooted in psychological 2 and neuroscientific 3 perspectives on animal behaviour, of how agents may …
被引用次数：16938 相关文章 所有 58 个版本

Figure 5: Nature in 2015. It solves 57 tasks.

From NIPS Workshop to Nature: target network is used.

Lessons from DQN

Game	With replay, with target Q	With replay, without target Q	Without replay, with target Q	Without replay, without target Q
Breakout	316.8	240.7	10.2	3.2
Enduro	1006.3	831.4	141.9	29.1
River Raid	7446.6	4102.8	2867.7	1453.0
Seaquest	2894.4	822.6	1003.0	275.8
Space Invaders	1088.9	826.3	373.2	302.0

Figure 6: Ablation study of target Q and experience replay [Mnih et al., 2015].

Experience replay is important; target Q makes it better.

Target Network

- Q-Learning:

$$w_{t+1} = w_t + \eta_t \left[r_t + \gamma \max_{a' \in \mathcal{A}} \langle \phi(s_{t+1}, a'), w_t \rangle - \langle \phi(s_t, a_t), w_t \rangle \right] \phi(s_t, a_t).$$

- Target Q-Learning:

$$w_{t+1} = w_t + \eta_t \left[r_t + \gamma \max_{a' \in \mathcal{A}} \langle \phi(s_{t+1}, a'), \overline{w} \rangle - \langle \phi(s_t, a_t), w_t \rangle \right] \phi(s_t, a_t).$$

  where $\overline{w}$ is the "target parameter", which is fixed over several iterations.

- Target Q-Learning updates $\overline{w}$ periodically with the copy of w_t.

Target Q-Learning

Let w^{k-1} be the target parameter in each epoch k.

Algorithm 1 Target Q-Learning

1: **for** epoch $k = 1, 2, \cdots,$ **do**
2:     **for** iteration $t = 1, 2, \cdots, T-1$ **do**
3:         $w_{t+1} = w_t + \eta_t[r_t + \gamma \max_{a' \in \mathcal{A}} \langle \phi(s_{t+1}, a'), w^{k-1} \rangle - \langle \phi(s_t, a_t), w_t \rangle] \phi(s_t, a_t)$.
4:     **end for**
5:     $w^k = w_T$.
6: **end for**

Baird's Example Revisited

- Let $\gamma = 0.99$ and $\eta = 0.1$. Suppose $w^0 = 1$ and $w^1 = w^2, \cdots, w^6 = 0$. For target Q-Learning:
    - state s_1: $w_2 = w_1 + \eta(r + \gamma\bar{V}(6) - V(1))\phi_1 \Longrightarrow w^0 \uparrow, w^1 \uparrow$.
    $$\Delta_1 = r + \gamma\bar{V}(6) - V(1) = 0.980, w^0 = 1.098, w^1 = 0.196$$
    - state s_2: $w_3 = w_2 + \eta(r + \gamma\bar{V}(6) - V(2))\phi_2 \Longrightarrow w^0 \uparrow, w^2 \uparrow$.
    $$\Delta_2 = r + \gamma\bar{V}(6) - V(2) = 0.882, w^0 = 1.186, w^2 = 0.176$$
    - ......
    - state s_6: $w_7 = w_6 + \eta(r + \gamma\bar{V}(6) - V(6))\phi_6 \Longrightarrow w^0 \downarrow, w^6 \downarrow$
    $$\Delta_6 = r + \gamma\bar{V}(6) - V(6) = -0.823, w^0 = 1.237, w^6 = -0.082$$
- Error does not explode within epoch:
$$\text{Q-Learning}: w_0 \uparrow \Longrightarrow \Delta_1 < \Delta_2 < \cdots < \Delta_5$$
$$\text{Target Q-Learning}: w_0 \uparrow \Longrightarrow \Delta_1 > \Delta_2 > \cdots > \Delta_5$$

Setup of Target Q-Learning

To make notations clean, let $x_{s,a}$ denote $x(s,a)$.

- In iteration k, target Q-Learning amounts to solve the following problem with SGD:

$$F(w; w^{k-1}) = \sum_{(s,a)} \left(\phi_{s,a}^\top w - y_{s,a}\right)^2,$$

$$y_{s,a} = R_{s,a} + \gamma \mathbb{E}_{s'}\left[\max_{a'} \phi_{s',a'}^\top w^{k-1}\right].$$

- Randomness is from the index (s,a) and the label $y_{s,a}$ because the next state s' is also random.

Analysis of Target Q-Learning

$\rightsquigarrow$ Since $\min_w F(w; w^{k-1})$ is an over-determined least square problem, it must exist a minimizer $w^\star_{k-1}$ such that $F(w^\star_{k-1}; w^{k-1}) = 0$.

$\rightsquigarrow$ After T inner iterations, assume in expectation, we have $\mathbb{E}[F(w_T; w^{k-1})] - F(w^\star_{k-1}; w^{k-1}) \leq \varepsilon_{\mathrm{opt}}$ for all outer iteration k. This implies that ($w^k := w_T$)

$$\mathbb{E}\Big[\sup_{(s,a)} |\phi_{s,a}^\top w^k - \mathcal{T}(w^{k-1})_{s,a}|\Big] \leq \sqrt{\mathbb{E}[F(w_T; w^{k-1})]} \leq \sqrt{\varepsilon_{\mathrm{opt}}}.$$

$\rightsquigarrow$ Then we have

$$\begin{aligned}
\mathbb{E}\left[\left\|w^K - w^\star\right\|_\phi\right] &:= \mathbb{E}\Big[\sup_{(s,a)} |\phi_{s,a}^\top w^K - \phi_{s,a}^\top w^\star|\Big] \\
&\leq \mathbb{E}\Big[\sup_{(s,a)} \left|\phi_{s,a}^\top w^K - \mathcal{T}(w_{K-1})_{s,a}\right| + \sup_{(s,a)} \left|\mathcal{T}(w_{K-1})_{s,a} - \phi_{s,a}^\top w^\star\right|\Big] \\
&\leq \sqrt{\varepsilon_{\mathrm{opt}}} + \gamma \mathbb{E}\Big[\left\|w^{K-1} - w^\star\right\|_\phi\Big] \leq \cdots\cdots \\
&\leq \frac{\sqrt{\varepsilon_{\mathrm{opt}}}}{1-\gamma} + \gamma^K \left\|w^0 - w^\star\right\|_\phi .
\end{aligned}$$

Remark on Target Q-Learning

- For target Q-Learning, if we can control $\varepsilon_{\mathrm{opt}}$, the convergence with linear function approximation is guaranteed.
- Target Q-Learning does not contradict with Tsitsiklis and Van Roy's Example because the latter is not in the over-parameterization regime.
- We need additional effort to analyze $\varepsilon_{\mathrm{opt}}$, which depends on w^{k-1} when we try to upper bound the variance of SGD update.
  - Typical SGD analysis assume the variance is upper bound by a constant.
  - For target Q-Learning, we solve multiple least square problems, where the variance changes over different problems.

References

L. C. B. III. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, pages 30–37, 1995.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.