Recent Advances in Target Q-Learning

Ziniu Li

April 18, 2022

The Chinese University of Hong Kong, Shenzhen

Joint work with Tian Xu (NJU).

Table of contents

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 ($0 < \gamma < 1$):

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

Linear Function Approximation

Consider the case where $Q(s, a)$ is linearly parameterized by $\theta \in \mathbb{R}^d$, i.e., $Q(s, a) = \phi(s, a)^\top \theta$, where $\phi(s, a) \in \mathbb{R}^d$ is the given feature.

Suppose we can sample the data pair (s, a, r, s') from a given distribution μ.

- We first sample $(s, a) \sim \mu(s, a)$, then we sample $s' \sim P(\cdot|s, a)$. We assume $\mu(s, a) > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Our goal is to compute the optimal Q-value function $Q^\star(s, a) = \phi(s, a)^\top \theta^\star$.

Q-Learning

Q-Learning is a stochastic approximation method to solve the Bellman optimal equation.

$$
\begin{aligned}
Q_{t+1}(s_t, a_t) &= (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t \left(r(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') \right) \\
&= Q_t(s_t, a_t) + \alpha_t \left[r(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t) \right]
\end{aligned}
$$

In the linear function approximation case, we have

$$
\theta_{t+1} = \theta_t + \alpha_t \left[r(s_t, a_t) + \gamma \max_{a'} \phi(s_{t+1}, a')^\top \theta_t - \phi(s_t, a_t)^\top \theta_t \right] \phi(s_t, a_t).
$$

Divergence of Q-Learning with LFA

Unfortunately, Q-Learning with LFA can diverge.

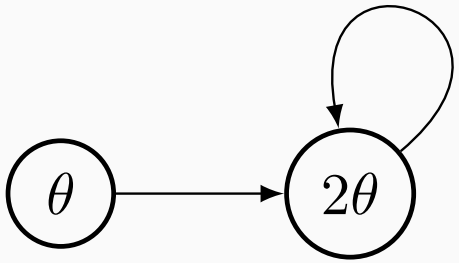

Figure 1: A simple MDP where Q-Learning with LFA can diverge [SB18].

Here $\phi(s_1) = 1$ and $\phi(s_2) = 2$, and $\theta \in \mathbb{R}$ is the optimal parameter to solve. We know that $\theta^\star = 0$. Assume $\mu(s_1) = \mu(s_2) = 0.5$.

$$\begin{aligned}\mathbb{E}\left[\theta_{t+1}|\theta_t\right] &= \theta_t + \alpha_t \mathbb{E}\left\{\left[r(s_t, a_t) + \gamma \max_{a'} \phi(s_{t+1}, a')^\top \theta_t - \phi(s_t, a_t)^\top \theta_t\right]\phi(s_t, a_t)\right\} \\ &= \left[1 - (2.5 - 3\gamma)\alpha_t\right]\theta_t.\end{aligned}$$

When $\gamma > 5/6$, for any $\alpha_t > 0$, we have that $\mathbb{E}[\theta_t]$ will diverge.

Deadly Triad

In the famous book [SB18], Sutton et al. contributed the divergence of Q-learning by three reasons.

- **Off-policy**. "Training on a distribution of transitions other than that produced by the target policy." (In the on-policy case, we can show that policy evaluation is convergent [TVR97].)
- **Function Approximation**. " Stability is guaranteed for function approximation methods that do not extrapolate from the observed targets."
- **Bootstrapping**. " Update targets that include existing estimates (as in dynamic programming or TD methods) rather than relying exclusively on actual rewards and complete returns (as in MC methods)."

Sutton et al. called the combination of the above factors as a deadly triad.

Target Q-Learning

To address the diverge issue, target Q-learning is introduced [MKS+15]:

$$\theta_{t+1} = \theta_t + \alpha_t \left[r(s,a) + \gamma \max_{a'} \phi(s_{t+1}, a')^\top \bar{\theta} - \phi(s_t, a_t)^\top \theta_t \right] \phi(s_t, a_t),$$

where $\bar{\theta}$ is fixed over several iterations.

Specifically, the above update rule can be viewed SGD step of

$$\min_\theta F(\theta; \bar{\theta}) := \sum_{(s,a)} \mu(s,a) \left(\phi(s,a)^\top \theta - \mathcal{T}(\bar{\theta})(s,a) \right)^2,$$

where $\mathcal{T}(\bar{\theta})$ is the Bellman update w.r.t. $\bar{\theta}$:

$$\mathcal{T}(\bar{\theta})(s,a) := r(s,a) + \gamma \mathbb{E}_{s'} [\max_{a'} \phi(s', a')^\top \bar{\theta}].$$

Target Q-learning

Let θ^{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: Sample (s, a, r, s') from μ and update

$$\theta_{t+1} = \theta_t + \alpha_t[r_t + \gamma \max_{a' \in \mathcal{A}} \phi(s_{t+1}, a')^\top \theta^{k-1} - \phi(s_t, a_t)^\top \theta_t]\phi(s_t, a_t).$$

4: **end for**
5: $w^k = w_T$.
6: **end for**

Intuition Behind Target Q-learning

In DP (dynamic programming) based analysis, we care about the criterion:

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{(s,a)}\left|\phi(s,a)^\top\theta_T - \mathcal{T}(\theta^{k-1})(s,a)\right|\right] \\
&\leq \mathbb{E}\left[\frac{1}{\mu_{\min}}\sum_{(s,a)}\mu(s,a)\left|\phi(s,a)^\top\theta_T - \mathcal{T}(\theta^{k-1})(s,a)\right|\right] \\
&\leq \frac{1}{\mu_{\min}}\mathbb{E}\left[\sqrt{\sum_{(s,a)}\mu(s,a)\left(\phi(s,a)^\top\theta_T - \mathcal{T}(\theta^{k-1})(s,a)\right)^2}\right] \\
&= \frac{1}{\mu_{\min}}\mathbb{E}\left[\sqrt{F(\theta_T;\theta^{k-1})}\right].
\end{aligned}
$$

where $\mu_{\min} = \min_{(s,a)}\mu(s,a)$.

Intuition Behind Target Q-learning

Assume $\theta_\star^{k-1}$ belongs to the optimal solution set in the inner loop:

$$\theta_\star^{k-1} \in \operatorname*{argmin}_\theta F(\theta; \theta^{k-1}).$$

After T inner iterations, assume in expectation, we have

$$\mathbb{E}[F(\theta_T; \theta^{k-1})] - F(\theta_\star^{k-1}; \theta^{k-1}) \leq \varepsilon_{\text{opt}}.$$

Then,

$$\underbrace{\text{appro. Bellman error}}_{\sup_{(s,a)} |\phi(s,a)^\top \theta_T - \mathcal{T}(\theta^{k-1})(s,a)|} \leq \underbrace{\text{func. approx. error}}_{F(\theta_\star^{k-1}; \theta^{k-1})} + \underbrace{\text{opt. error}}_{\varepsilon_{\text{opt}}}$$

- For LFA, it is reasonable to assume ε_{opt} is small when T is sufficiently large.
- But, it is not safe to assume the function approximation error is bounded.

Function Approximation Error May not be Bounded

Consider the $\theta - 2\theta$ example again:

$$\begin{aligned}\theta_\star^{k-1} &= \operatorname*{argmin}_{\theta} F(\theta; \theta^{k-1}) \\ &= \operatorname*{argmin}_{\theta \in \mathbb{R}} 0.5(\theta - 2\gamma\theta^{k-1})^2 + 0.5(2\theta - \gamma 2\theta^{k-1})^2 \\ &= \frac{6}{5}\gamma\theta^{k-1}.\end{aligned}$$

Then, the function approximation error is

$$F(\theta_\star^{k-1}; \theta^{k-1}) = \frac{4}{25}\gamma^2(\theta^{k-1})^2.$$

This says that when $\gamma > 5/6$, $\{\theta^{k-1}\}$ diverges and the function approximation error $\{F(\theta_\star^{k-1}; \theta^{k-1})\}$ diverges too.

Solving the Deadly Triad

Over-parameterized LFA

Assume $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$ has a full row rank (over-parameterization).

Over-parameterization allows the function approximation error is 0:

$$F(\theta_\star^{k-1}; \theta^{k-1}) = 0, \quad \forall k \geq 0$$

Consequently, approx. Bellman error becomes:

$$\begin{aligned}\mathbb{E}\left[\sup_{(s,a)} \left|\phi(s,a)^\top \theta_T - \mathcal{T}(\theta^{k-1})(s,a)\right|\right] &\leq \frac{1}{\mu_{\min}} \mathbb{E}\left[\sqrt{F(\theta_T; \theta^{k-1})}\right] \\ &\leq \frac{1}{\mu_{\min}} \sqrt{\varepsilon_{\text{opt}}}.\end{aligned}$$

Approximate Bellman Update

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{(s,a)} \left|\phi(s,a)^\top \theta^K - Q^\star(s,a)\right|\right] \\
&\leq \mathbb{E}\left[\sup_{(s,a)} \left|\phi(s,a)^\top \theta^K - \mathcal{T}(\theta^{K-1})(s,a)\right|\right] + \mathbb{E}\left[\sup_{(s,a)} \left|\mathcal{T}(\theta^{K-1})(s,a) - Q^\star(s,a)\right|\right] \\
&\leq \frac{1}{\mu_{\min}}\sqrt{\varepsilon_{\text{opt}}} + \gamma \mathbb{E}\left[\sup_{(s,a)} \left|\phi(s,a)^\top \theta^{K-1} - Q^\star(s,a)\right|\right] \\
&\leq \cdots\cdots \\
&\leq \underbrace{\frac{\sqrt{\varepsilon_{\text{opt}}}}{\mu_{\min}(1-\gamma)}}_{\text{cumulative opt. error}} + \underbrace{\gamma^{K-1} \sup_{(s,a)} \left|\phi(s,a)^\top \theta^1 - Q^\star(s,a)\right|}_{\text{decaying init. error}}.
\end{aligned}
$$

Deadly Triad in Over-parameterization Case

To justify the target Q-Learning in the over-parameterized case, we can give an example in which the vanilla Q-Learning diverges.

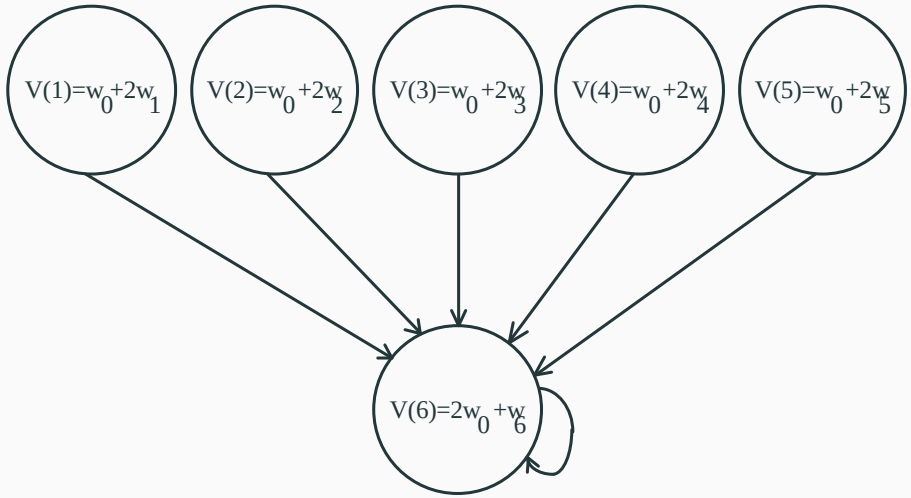

Figure 2: Baird example to show that Q-Learning can diverge in the over-parameterized case [III95].

Optimization Error of Over-Parameterization LS

(Proposition 1) Opt. Error is Bounded for Target Q-Learning

Consider the inner loop in iteration k. Suppose that $\frac{2}{\mu_{\min}}\mathbb{E}\left[F(\theta_T;\theta^i)\right] \leq (1-\gamma)^2, \forall i \leq k-1$.
If we set $\alpha_t = \frac{\eta_0}{\beta+t}$, where $\eta_0 = 2/C_3, \beta = (51\lambda_{\max}\gamma^2)/(8C_3^2) - 1$, then we have that

$$\mathbb{E}\left[F(\theta_t;\theta^{k-1})\right] \leq \frac{102\lambda_{\max}\gamma^2}{C_3^2(1-\gamma)^2}\frac{1}{\beta+t}, \forall t \geq 0,$$

where $C_3 = 1/(\lambda_{\max}C_2^2)$, in which C_2 further depends on the error bound parameter C_1, and $\lambda_{\max}$ is the maximum eigenvalue of the feature matrix $\mathbb{E}_{(s,a)\sim\mu}[\phi(s,a)\phi(s,a)^\top]$.

This result is not published yet.

Proof Sketch of Proposition 1

Proof of Proposition 1 mainly relies on the classical analysis for SGD [BCN18].

There are two things beyond the classical analysis:

- Strong convexity does not hold for the over-parameterized least square problem. Following [SZ17], we use the error bound analysis [LT93] to argue that **PL(Polyak-Łojasiewicz) condition** holds, which is used to show the sublinear convergence.
- Following [LH20], we show that the **variance of stochastic gradient** is upper bounded over iterations k. Instead, the classical SGD assumes the variance is uniformly bounded.

ZIBEL MDP

In the under-parameterized case, we assume $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$ has a full column rank and $\mathbb{E}_{(s,a)\sim\mu}[\phi(s,a)\phi(s,a)^\top]$ is a positive definite matrix.

As discussed, we need a very strong assumption about the function approximation error:

For an $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R)$ with a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, $\mathsf{IBE}(\mathcal{M})$ is defined as:

$$\sup_{\theta \in \mathbb{R}^d} \inf_{\theta' \in \mathbb{R}^d} \sup_{(s,a)} \left| \langle \phi(s,a), \theta' \rangle - r(s,a) - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [\sup_{a' \in \mathcal{A}} \langle \theta, \phi(s',a') \rangle] \right|.$$

(ZIBEL MDP) [ACJ+21]: if $\mathsf{IBE}(\mathcal{M}) = 0$, we call this MDP as a **ZIBEL** (zero inherent Bellman error with linear function approximation).

ZIBEL MDP

Based on $\mathsf{IBE}(\mathcal{M}) = 0$, we can show that for any $\bar{\theta}$, there exists a unique $\bar{\theta}_\star$ such that

$$\phi(s,a)^\top \bar{\theta}_\star = \underbrace{r(s,a) + \gamma \mathbb{E}_{s'}\left[\max_{a'} \phi(s',a')^\top \bar{\theta}\right]}_{\mathcal{T}(\bar{\theta})(s,a)}, \quad \forall (s,a).$$

This implies that the function approximation error is also zero.

Analysis for ZIBEL MDP

In the under-parameterized case, we can have a simple analysis (rather than SGD based analysis) [ACJ+21].

To simplify notation, let $\phi_t = \phi(s_t, a_t)$.

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha_t \left[r(s,a) + \gamma \max_{a'} \phi(s_{t+1}, a')^\top \bar{\theta} - \phi_t^\top \theta_t \right] \phi_t, \\ &= (I - \alpha_t \phi_t \phi_t^\top)\theta_t + \alpha_t \phi_t \left[r(s,a) + \gamma \max_{a'} \phi(s_{t+1}, a')^\top \bar{\theta} \right].\end{aligned}$$

As a result,

$$\begin{aligned}\mathbb{E}\left[\theta_{t+1} - \bar{\theta}_\star \mid \theta_t\right] &= \mathbb{E}\left[(I - \alpha_t \phi_t \phi_t^\top)\theta_t + \alpha_t \phi_t \phi_t^\top \bar{\theta}_\star - \bar{\theta}_\star \mid \theta_t\right] \\ &= (I - \alpha_t \mathbb{E}[\phi_t \phi_t^\top])(\theta_t - \bar{\theta}_\star).\end{aligned}$$

Because $\mathbb{E}[\phi_t \phi_t^\top]$ is PD as assumed, we can show that the above recursion is contractive for some $\alpha_t > 0$. (This analysis cannot be applied in the over-parameterization case)

Beyond ZIBEL MDP

Note (approximate) ZIBEL even does *not* hold for the simple $\theta - 2\theta$ MDP.

Recently, [CCM22] introduced a relaxed approximation error $\mathcal{E}_{\text{approx}}$ is bounded:

$$\sup_{\theta \in \Theta} \inf_{\theta' \in \mathbb{R}^d} \sup_{(s,a)} \left| \langle \phi(s,a), \theta' \rangle - r(s,a) - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [\sup_{a' \in \mathcal{A}} \langle \theta, \phi(s',a') \rangle] \right|$$

where $\Theta = \{\theta : \sup_{(s,a)} |\langle \phi(s,a), \theta \rangle| < 1/(1-\gamma)\}$.

This assumption says that **when θ is bounded, the approximation error is bounded**, which holds for the $\theta - 2\theta$ MDP.

However, such an assumption is *not* sufficient to show the convergence of target Q-learning.

Bounded Approximation Error

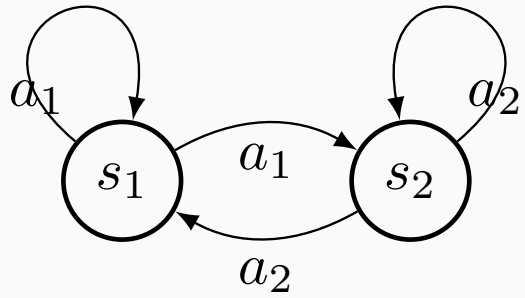

Figure 3: A simple MDP in [CCM22] to show that target Q-Learning can diverge even though $\mathcal{E}_{\text{approx}}$ is bounded.

Reward information: $(r(s_1, a_1), r(s_1, a_2), r(s_2, a_1), r(s_2, a_2)) = (1, 2, 2, 4)$. The feature map is $\Phi = (1, 2, 2, 4)^\top$. Sampling distribution μ is uniform. After calculation, we have that

$$\bar{\theta}_\star = 1 + \frac{9\gamma}{10}\bar{\theta} + \frac{3\gamma\bar{\theta}}{10}\left(\mathbb{I}_{\bar{\theta}\geq 0} - \mathbb{I}_{\bar{\theta}<0}\right).$$

We see that when $\gamma > 5/6$ and the initialization is positive, target Q-Learning would diverge.

Truncation in Target Network

Intuition in Figure 3: even though $\sup_{(s,a)} |\phi(s,a)^\top \bar{\theta}|$ is bounded, $\mathcal{T}(\bar{\theta})$ may lies out the range of bounded approx. error.

To address the "over-estimation" issue, [CCM22] proposed to implement the truncation:

$$\theta_{t+1} = \theta_t + \alpha_t \left[\lceil r(s,a) + \gamma \max_{a'} \phi(s_{t+1}, a')^\top \bar{\theta} \rceil - \phi(s_t, a_t)^\top \theta_t \right] \phi(s_t, a_t),$$

where $\lceil x \rceil = \mathsf{clip}(x, -1/(1-\gamma), 1/(1-\gamma))$.

This ensures that the approximation error encountered by the algorithm is always bounded.

Note that in the over-parameterized case $\mathcal{E}_{\mathsf{approx}} = 0$, and we do not need the truncation.

Real-World Target Q-Learning

Extension to Deep Neural Networks

In practice, we use wide and deep neural networks, which can also perform the “over-parameterization”.

But, the optimization problem is non-convex now. Mathematically, we aim to sequentially solve K non-convex least-square problems.

The theory under LFA tells us that it is fine to run a large inner iterations T, which may not true in the neural network case.

A Deep Target Q-Learning Experiment

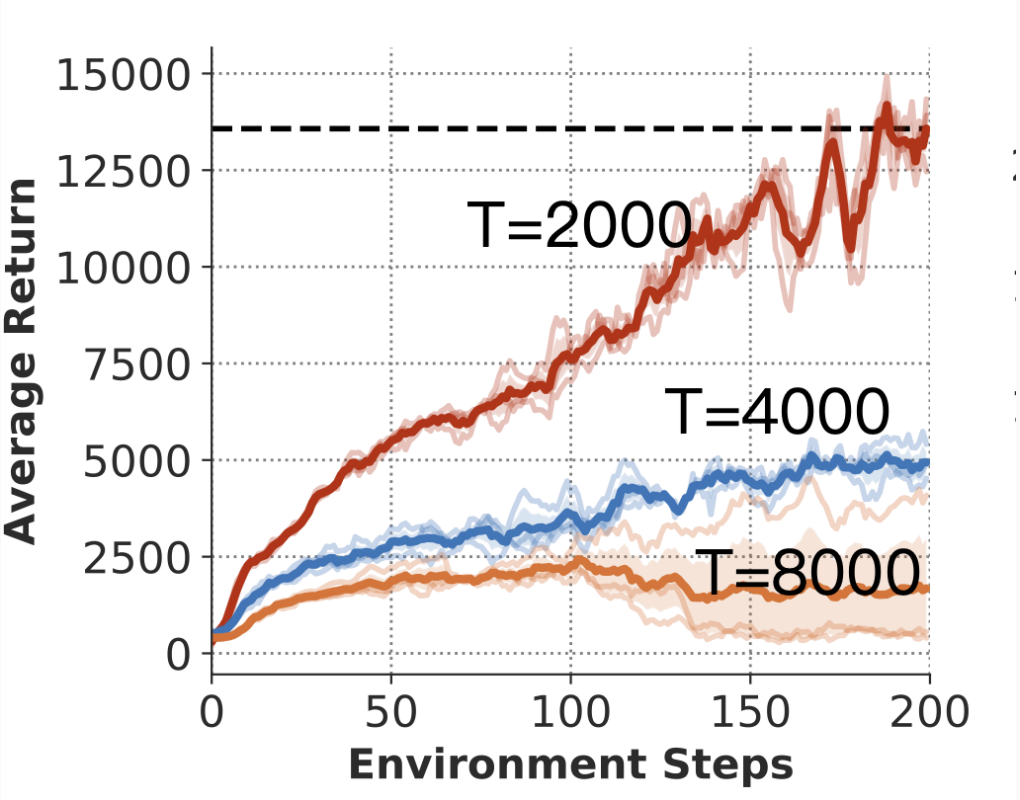

Figure 4: Figure is from [KAGL20].

We find that large inner iterations can hurt the performance.

Understand Deep Target Q-Learning

Deep and wide neural network can have a strong approximation ability. But it does not mean that we can find the optimal parameter from *any* initialization point.

It is a folklore that a random initialization is good in the sense that there exists an optimal parameter close to this initialization.

Unfortunately, deep target Q-Learning inherits the initialization from the last iterate, which may not be good.

We can solve the mentioned issue by re-initialize randomly for each inner loop, but the computation cost is high.

We need more insights and principled methods to solve the "sequential non-convex optimization problem".

References

[ACJ+21] Naman Agarwal, Syomantak Chaudhuri, Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps. *arXiv*, 2110.08440, 2021.

[BCN18] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[CCM22] Zaiwei Chen, John Paul Clarke, and Siva Theja Maguluri. Target network and truncation overcome the deadly triad in q-learning. *arXiv preprint arXiv:2203.02628*, 2022.

References ii

[III95] Leemon C. Baird III. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, pages 30–37, 1995.

[KAGL20] Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. *arXiv preprint arXiv:2010.14498*, 2020.

[LH20] Donghwan Lee and Niao He. Periodic q-learning. In *Learning for Dynamics and Control*, pages 582–598. PMLR, 2020.

References iii

[LT93] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[MKS+15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

References iv

[SZ17] Anthony Man-Cho So and Zirui Zhou. Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. *Optimization Methods and Software*, 32(4):963–992, 2017.

[TVR97] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.