Breaking the Deadly Triad with a Target Network

Shangtong Zhang [1]   Hengshuai Yao [2 3]   Shimon Whiteson [1]

Abstract

The deadly triad refers to the instability of a reinforcement learning algorithm when it employs off-policy learning, function approximation, and bootstrapping simultaneously. In this paper, we investigate the target network as a tool for breaking the deadly triad, providing theoretical support for the conventional wisdom that a target network stabilizes training. We first propose and analyze a novel target network update rule which augments the commonly used Polyak-averaging style update with two projections. We then apply the target network and ridge regularization in several divergent algorithms and show their convergence to regularized TD fixed points. Those algorithms are off-policy with linear function approximation and bootstrapping, spanning both policy evaluation and control, as well as both discounted and average-reward settings. In particular, we provide the first convergent linear Q-learning algorithms under nonrestrictive and changing behavior policies without bi-level optimization.

1. Introduction

The deadly triad (see, e.g., Chapter 11.3 of Sutton & Barto (2018)) refers to the instability of a value-based reinforcement learning (RL, Sutton & Barto (2018)) algorithm when it employs off-policy learning, function approximation, and bootstrapping simultaneously. Different from *on-policy* methods, where the policy of interest is executed for data collection, *off-policy* methods execute a different policy for data collection, which is usually safer (Dulac-Arnold et al., 2019) and more data efficient (Lin, 1992; Sutton et al., 2011). *Function approximation* methods use parameterized functions, instead of a look-up table, to represent quantities of interest, which usually cope better with large-scale problems (Mnih et al., 2015; Silver et al., 2016). *Bootstrap-*

[1]University of Oxford [2]Huawei Technologies [3]University of Alberta. Correspondence to: Shangtong Zhang < shangtong.zhang@cs.ox.ac.uk>.

ping methods construct update targets for an estimate by using the estimate itself recursively, which usually has lower variance than *Monte Carlo* methods (Sutton, 1988). However, when an algorithm employs all those three preferred ingredients (off-policy learning, function approximation, and bootstrapping) simultaneously, there is usually no guarantee that the resulting algorithm is well behaved and the value estimates can easily diverge (see, e.g., Baird (1995); Tsitsiklis & Van Roy (1997); Zhang et al. (2021)), yielding the notorious deadly triad.

An example of the deadly triad is Q-learning (Watkins & Dayan, 1992) with linear function approximation, whose divergence is well documented in Baird (1995). However, Deep-Q-Networks (DQN, Mnih et al. (2015)), a combination of Q-learning and deep neural network function approximation, has enjoyed great empirical success. One major improvement of DQN over linear Q-learning is the use of a target network, a copy of the neural network function approximator (the main network) that is periodically synchronized with the main network. Importantly, the bootstrapping target in DQN is computed via the target network instead of the main network. As the target network changes slowly, it provides a stable bootstrapping target which in turn stabilizes the training of DQN. Instead of the periodical synchronization, Lillicrap et al. (2015) propose a Polyak-averaging style target network update, which has also enjoyed great empirical success (Fujimoto et al., 2018; Haarnoja et al., 2018).

Inspired by the empirical success of the target network in RL with deep networks, in this paper, we theoretically investigate the target network as a tool for breaking the deadly triad. We consider a two-timescale framework, where the main network is updated faster than the target network. By using a target network to construct the bootstrapping target, the main network update becomes least squares regression. After adding ridge regularization (Tikhonov et al., 2013) to this least squares problem, we show convergence for both the target and main networks.

Our main contributions are twofold. First, we propose a novel target network update rule augmenting the Polyak-averaging style update with two projections. The balls for the projections are usually large so most times they are just identity mapping. However, those two projections offer sig-

nificant theoretical advantages making it possible to analyze where the target network converges to (Section 3). Second, we apply the target network in various existing divergent algorithms and show their convergence to regularized TD (Sutton, 1988) fixed points. Those algorithms are off-policy algorithms with linear function approximation and bootstrapping, spanning both policy evaluation and control, as well as both discounted and average-reward settings. In particular, we provide the first convergent linear Q-learning algorithms under nonrestrictive and changing behavior policies without bi-level optimization, for both discounted and average-reward settings.

2. Background

Let M be a real positive definite matrix and x be a vector, we use $\|x\|_M \doteq \sqrt{x^\top M x}$ to denote the norm induced by M and $\|\cdot\|_M$ to denote the corresponding induced matrix norm. When M is the identity matrix I, we ignore the subscript I for simplicity. We use vectors and functions interchangeably when it does not cause confusion, e.g., given $f : \mathcal{X} \to \mathbb{R}$, we also use f to denote the corresponding vector in $\mathbb{R}^{|\mathcal{X}|}$. All vectors are column vectors. We use $\mathbf{1}$ to denote an all one vector, whose dimension can be deduced from the context. $\mathbf{0}$ is similarly defined.

We consider an infinite horizon Markov Decision Process (MDP, see, e.g., Puterman (2014)) consisting of a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a transition kernel $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, and a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. At time step t, an agent at a state S_t executes an action $A_t \sim \pi(\cdot|S_t)$, where $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the policy followed by the agent. The agent then receives a reward $R_{t+1} \doteq r(S_t, A_t)$ and proceeds to a new state $S_{t+1} \sim p(\cdot|S_t, A_t)$.

In the *discounted* setting, we consider a discount factor $\gamma \in [0, 1)$ and define the return at time step t as $G_t \doteq \sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i}$, which allows us to define the action-value function $q_\pi(s, a) \doteq \mathbb{E}_{\pi, p}[G_t | S_t = s, A_t = a]$. The action-value function q_π is the unique fixed point of the Bellman operator $\mathcal{T}_\pi$, i.e., $q_\pi = \mathcal{T}_\pi q_\pi \doteq r + \gamma P_\pi q_\pi$, where $P_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ is the transition matrix, i.e., $P_\pi((s, a), (s', a')) \doteq \sum_a p(s'|s, a)\pi(a'|s')$.

In the *average-reward* setting, we assume:

Assumption 2.1. *The chain induced by π is ergodic.*

This allows us to define the *reward rate* $\bar{r}_\pi \doteq \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[R_t | p, \pi]$. The differential action-value function $\bar{q}_\pi(s, a)$ is defined as

$$\lim_{T \to \infty} \sum_{t=0}^{T} \mathbb{E}_{\pi, p}[r(S_t, A_t) - \bar{r}_\pi | S_0 = s, A_0 = a].$$

The differential Bellman equation is

$$\bar{q} = r - \bar{r}\mathbf{1} + P_\pi \bar{q}, \tag{1}$$

where $\bar{q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\bar{r} \in \mathbb{R}$ are free variables. It is well known that all solutions to (1) form a set $\{(\bar{q}, \bar{r}) \mid \bar{r} = \bar{r}_\pi, \bar{q} = q_\pi + c\mathbf{1}, c \in \mathbb{R}\}$ (Puterman, 2014).

The *policy evaluation* problem refers to estimating q_π or $(\bar{q}_\pi, \bar{r}_\pi)$. The *control* problem refers to finding a policy π maximizing $q_\pi(s, a)$ for each (s, a) or maximizing $\bar{r}_\pi$. With *linear* function approximation, we approximate $q_\pi(s, a)$ or $\bar{q}_\pi(s, a)$ with $x(s, a)^\top w$, where $x : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ is a feature mapping and $w \in \mathbb{R}^K$ is the learnable parameter. We use $X \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times K}$ to denote the feature matrix, each row of which is $x(s, a)^\top$, and assume:

Assumption 2.2. *X has linearly independent columns.*

In the average-reward setting, we use an additional parameter $\bar{r} \in \mathbb{R}$ to approximate $\bar{r}_\pi$. In the *off-policy* learning setting, the data for policy evaluation or control is collected by executing a policy μ (behavior policy) in the MDP, which is different from π (target policy). In the rest of the paper, we consider the off-policy linear function approximation setting thus always assume $A_t \sim \mu(\cdot|S_t)$. We use as shorthand $x_t \doteq x(S_t, A_t), \bar{x}_t \doteq \sum_a \pi(a|S_t)x(S_t, a)$.

Policy Evaluation. In the discounted setting, similar to Temporal Difference Learning (TD, Sutton (1988)), one can use Off-Policy Expected SARSA to estimate q_π, which updates w as

$$\delta_t \leftarrow R_{t+1} + \gamma \bar{x}_{t+1}^\top w_t - x_t^\top w_t,$$
$$w_{t+1} \leftarrow w_t + \alpha_t \delta_t x_t, \tag{2}$$

where $\{\alpha_t\}$ are learning rates. In the average-reward setting, (1) implies that $\bar{r}_\pi = d^\top(r + P_\pi \bar{q}_\pi - \bar{q}_\pi)$ holds for any probability distribution d. In particular, it holds for $d = d_\mu$. Consequently, to estimate $\bar{q}_\pi$ and $\bar{r}_\pi$, Wan et al. (2020); Zhang et al. (2021) update w and $\bar{r}$ as

$$w_{t+1} \leftarrow w_t + \alpha_t(R_{t+1} - \bar{r}_t + \bar{x}_{t+1}^\top w_t - x_t^\top w_t)x_t,$$
$$\bar{r}_{t+1} \leftarrow \bar{r}_t + \alpha_t(R_{t+1} + \bar{x}_{t+1}^\top w_t - x_t^\top w_t - \bar{r}_t). \tag{3}$$

Unfortunately, both (2) and (3) can possibly diverge (see, e.g., Tsitsiklis & Van Roy (1997); Zhang et al. (2021)), which exemplifies the deadly triad in discounted and average-reward settings respectively.

Control. In the discounted setting, Q-learning with linear function approximation yields

$$\delta_t \leftarrow R_{t+1} + \gamma \max_{a'} x(S_{t+1}, a')^\top w_t - x_t^\top w_t,$$
$$w_{t+1} \leftarrow w_t + \alpha_t \delta_t x_t. \tag{4}$$

In the average-reward setting, Differential Q-learning (Wan et al., 2020) with linear function approximation yields

$$\delta_t \leftarrow R_{t+1} - \bar{r}_t + \gamma \max_{a'} x(S_{t+1}, a')^\top w_t - x_t^\top w_t,$$
$$w_{t+1} \leftarrow w_t + \alpha_t \delta_t x_t, \quad \bar{r}_{t+1} \leftarrow \bar{r}_t + \alpha_t \delta_t. \tag{5}$$

Unfortunately, both (4) and (5) can possibly diverge as well (see, e.g., Baird (1995); Zhang et al. (2021)), exemplifying the deadly triad again.

Motivated by the empirical success of the target network in deep RL, one can apply the target network in the linear function approximation setting. For example, using a target network in (4) yields

$$\delta_t \leftarrow R_{t+1} + \gamma \max_{a'} x(S_{t+1}, a')^\top \theta_t - x_t^\top w_t,$$

$$w_{t+1} \leftarrow w_t + \alpha_t \delta_t x_t, \tag{6}$$

$$\theta_{t+1} \leftarrow \theta_t + \beta_t(w_t - \theta_t), \tag{7}$$

where θ denotes the target network, $\{\beta_t\}$ are learning rates, and we consider the Polyak-averaging style target network update. The convergence of (6) and (7), however, remains unknown. Besides target networks, regularization has also been widely used in deep RL, e.g., Mnih et al. (2015) consider a Huber loss instead of a mean-squared loss; Lillicrap et al. (2015) consider ℓ_2 weight decay in updating Q-values.

3. Analysis of the Target Network

In Sections 4 & 5, we consider the merits of using a target network in several linear RL algorithms (e.g., (2) (3) (4) (5)). To this end, in this section, we start by proposing and analyzing a novel target network update rule:

$$\theta_{t+1} \doteq \Gamma_{B_1}\big(\theta_t + \beta_t(\Gamma_{B_2}(w_t) - \theta_t)\big). \tag{8}$$

In (8), w denotes the main network and θ denotes the target network. $\Gamma_{B_1} : \mathbb{R}^K \to \mathbb{R}^K$ is a projection to the ball $B_1 \doteq \{x \in \mathbb{R}^K \mid \|x\| \le R_{B_1}\}$, i.e.,

$$\Gamma_{B_1}(x) \doteq x\mathbb{I}_{\|x\| \le R_{B_1}} + (R_{B_1}x/\|x\|)\mathbb{I}_{\|x\| > R_{B_1}},$$

where $\mathbb{I}$ is the indicator function. Γ_{B_2} is a projection onto the ball B_2 with a radius R_{B_2}. We make the following assumption about the learning rates:

Assumption 3.1. *$\{\beta_t\}$ is a deterministic positive nonincreasing sequence satisfying $\sum_t \beta_t = \infty, \sum_t \beta_t^2 < \infty$.*

While (8) specifies only how θ is updated, we assume w is updated such that w can track θ in the sense that

Assumption 3.2. *There exists $w^* : \mathbb{R}^K \to \mathbb{R}^K$ such that $\lim_{t\to\infty} \|w_t - w^*(\theta_t)\| = 0$ almost surely.*

After making some additional assumptions on w^*, we arrive at our general convergent results.

Assumption 3.3. $\sup_\theta \|w^*(\theta)\| < R_{B_2} < R_{B_1} < \infty$.

Assumption 3.4. *w^* is a contraction mapping w.r.t. $\|\cdot\|$.*

Theorem 1. *(Convergence of Target Network) Under Assumptions 3.1-3.4, the iterate $\{\theta_t\}$ generated by (8) satisfies*

$$\lim_{t\to\infty} w_t = \lim_{t\to\infty} \theta_t = \theta^* \quad \text{almost surely,}$$

where θ^* is the unique fixed point of $w^*(\cdot)$.

Assumptions 3.2 - 3.4 are assumed only for now. Once the concrete update rules for w are specified in the algorithms in Sections 4 & 5, we will prove that those assumptions indeed hold. Assumption 3.2 is expected to hold because we will later require that the target network to be updated much slower than the main network. Consequently, the update of the main network will become a standard least-square regression, whose solution w^* usually exists. Assumption 3.4 is expected to hold becuase we will later apply ridge regularization to the least-square regression. Consequently, its solution w^* will not change too fast w.r.t. the change of the regression target.

The target network update (8) is the same as that in (7) except for the two projections, where the first projection Γ_{B_1} is standard in optimization literature. The second projection Γ_{B_2}, however, appears novel and plays a crucial role in our analysis. *First*, if we have only Γ_{B_1}, the iterate $\{\theta_t\}$ would converge to the invariant set of the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = w^*(\theta(t)) - \theta(t) + \zeta(t), \tag{9}$$

where $\zeta(t)$ is a reflection term that moves $\theta(t)$ back to B_1 when $\theta(t)$ becomes too large (see, e.g., Section 5 of Kushner & Yin (2003)). Due to this reflection term, it is possible that $\theta(t)$ visits the boundary of B_1 infinitely often. It thus becomes unclear what the invariant set of (9) is even if w^* is contractive. By introducing the second projection Γ_{B_2} and ensuring $R_{B_1} > R_{B_2}$, we are able to remove the reflection term and show that the iterate $\{\theta_t\}$ tracks the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = w^*(\theta(t)) - \theta(t), \tag{10}$$

whose invariant set is a singleton $\{\theta^*\}$ when Assumption 3.4 holds. See the proof of Theorem 1 in Section A.1 based on the ODE approach (Kushner & Yin, 2003; Borkar, 2009) for more details. *Second*, to ensure the main network tracks the target network in the sense of Assumption 3.2 in our applications in Sections 4 & 5, it is crucial that the target network changes sufficiently slowly in the following sense:

Lemma 1. $\|\theta_{t+1} - \theta_t\| \le \beta_t C_0$ *for some constant $C_0 > 0$.*

Lemma 1 would not be feasible without the second projection Γ_{B_2} and we defer its proof to Section A.2

In Sections 4 & 5, we provide several applications of Theorem 1 in both discounted and average-reward settings, for both policy evaluation and control. We consider a two-timescale framework, where the target network is updated more slowly than the main network. Let $\{\alpha_t\}$ be the learning rates for updating the main network w; we assume

Assumption 3.5. *$\{\alpha_t\}$ is a deterministic positive nonincreasing sequence satisfying $\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty$. Further, for some $d > 0$, $\sum_t (\beta_t/\alpha_t)^d < \infty$.*

4. Application to Off-Policy Policy Evaluation

In this paper, we consider estimating the action-value q_π instead of the state-value v_π for unifying notations of policy evaluation and control. The algorithms for estimating v_π are straightforward up to change of notations and introduction of importance sampling ratios.

Discounted Setting. Using a target network for bootstrapping in (2) yields

$$w_{t+1} \leftarrow w_t + \alpha_t(R_{t+1} + \gamma\bar{x}_{t+1}^\top\theta_t - x_t^\top w_t)x_t.$$

As θ_t is quasi-static for w_t (Lemma 1 and Assumption 3.5), this update becomes least squares regression. Motivated by the success of ridge regularization in least squares and the widespread use of weight decay in deep RL, which is essentially ridge regularization, we add ridge regularization to this least squares, yielding Q-evaluation with a Target Network (Algorithm 1).

Algorithm 1 Q-evaluation with a Target Network
___
 INPUT: $\eta > 0, R_{B_1} > R_{B_2} > 0$
 Initialize $\theta_0 \in B_1$ and S_0
 Sample $A_0 \sim \mu(\cdot|S_0)$
 for $t = 0, 1, \ldots$ **do**
  Execute A_t, get R_{t+1} and S_{t+1}
  Sample $A_{t+1} \sim \mu(\cdot|S_{t+1})$
  $\bar{x}_{t+1} \doteq \sum_{a'} \pi(a'|S_{t+1})x(S_{t+1}, a')$
  $\delta_t \doteq R_{t+1} + \gamma\bar{x}_{t+1}^\top\theta_t - x_t^\top w_t$
  $w_{t+1} \doteq w_t + \alpha_t\delta_t x_t - \alpha_t\eta w_t$
  $\theta_{t+1} \doteq \Gamma_{B_1}\big(\theta_t + \beta_t(\Gamma_{B_2}(w_t) - \theta_t)\big)$
 end for

Let $A = X^\top D_\mu(I - \gamma P_\pi)X, b = X^\top D_\mu r$, where D_μ is a diagonal matrix whose diagonal entry is d_μ, the stationary state-action distribution of the chain induced by μ. Let $\Pi_{D_\mu} \doteq X(X^\top D_\mu X)^{-1}X^\top D_\mu$ be the projection to the column space of X. We have

Assumption 4.1. *The chain in $\mathcal{S} \times \mathcal{A}$ induced by μ is ergodic.*

Theorem 2. *Under Assumptions 2.2, 3.1, 3.5, & 4.1, for any $\xi \in (0, 1)$, let $C_0 \doteq \frac{2(1-\xi)\sqrt{\eta}}{\gamma\|P_\pi\|_{D_\mu}}, C_1 \doteq \frac{\|r\|}{2\xi\sqrt{\eta}} + 1$, then for all $\|X\| < C_0, C_1 < R_{B_1}, R_{B_1} - \xi < R_{B_2} < R_{B_1}$ the iterate $\{w_t\}$ generated by Algorithm 1 satisfies*

$$\lim_{t\to\infty} w_t = w_\eta^* \quad \text{almost surely,}$$

where w_η^* is the unique solution of $(A + \eta I)w - b = \mathbf{0}$, and

$$\|Xw_\eta^* - q_\pi\|$$
$$\leq (\frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^4\sigma_{\min}(D_\mu)^{2.5}}\|q_\pi\|\eta + \|\Pi_{D_\mu}q_\pi - q_\pi\|)/\xi,$$

where $\sigma_{\max}(\cdot), \sigma_{\min}(\cdot)$ denotes the largest and minimum singular values.

We defer the proof to Section A.3. Theorem 2 requires that the balls for projection are sufficiently large, which is completely feasible in practice. Theorem 2 also requires that the feature norm $\|X\|$ is not too large. Similar assumptions on feature norms also appear in Zou et al. (2019); Du et al. (2019); Chen et al. (2019b); Carvalho et al. (2020); Wang & Zou (2020); Wu et al. (2020) and can be easily achieved by scaling.

The solutions to $Aw - b = \mathbf{0}$, if they exist, are TD fixed points for off-policy policy evaluation in the discounted setting (Sutton et al., 2009b;a). Theorem 2 shows that Algorithm 1 finds a regularized TD fixed point w_η^*, which is also the solution of Least-Squares TD methods (LSTD, Boyan (1999); Yu (2010)). LSTD maintains estimates for A and b (referred to as $\hat{A}$ and $\hat{b}$) in an online fashion, which requires $\mathcal{O}(K^2)$ computational and memory complexity per step. As $\hat{A}$ is not guaranteed to be invertible, LSTD usually uses $(\hat{A} + \eta I)^{-1}\hat{b}$ as the solution and η plays a key role in its performance (see, e.g, Chapter 9.8 of Sutton & Barto (2018)). By contrast, Algorithm 1 finds the LSTD solution (i.e., w_η^*) with only $\mathcal{O}(K)$ computational and memory complexity per step. Moreover, Theorem 2 provides a performance bound for w_η^*. Let $w_0^* \doteq A^{-1}b$; Kolter (2011) shows with a counterexample that the approximation error of TD fixed points (i.e., $\|Xw_0^* - q_\pi\|$) can be arbitrarily large if μ is far from π, as long as there is representation error (i.e., $\|\Pi_{D_\mu}q_\pi - q_\pi\| > 0$) (see Section 6 for details). By contrast, Theorem 2 guarantees that $\|Xw_\eta^* - q_\pi\|$ is bounded from above, which is one possible advantage of regularized TD fixed points.

Algorithm 2 Diff. Q-evaluation with a Target Network
___
 INPUT: $\eta > 0, R_{B_1} > R_{B_2} > 0$
 Initialize $[\theta_0^r, \theta_0^{w\top}]^\top \in B_1$ and S_0
 Sample $A_0 \sim \mu(\cdot|S_0)$
 for $t = 0, 1, \ldots$ **do**
  Execute A_t, get R_{t+1} and S_{t+1}
  Sample $A_{t+1} \sim \mu(\cdot|S_{t+1})$
  $\bar{x}_{t+1} \doteq \sum_{a'} \pi(a'|S_{t+1})x(S_{t+1}, a')$
  $\delta_t \doteq R_{t+1} - \theta_t^r + \bar{x}_{t+1}^\top\theta_t^w - x_t^\top w_t$
  $w_{t+1} \doteq w_t + \alpha_t\delta_t x_t - \alpha_t\eta w_t$
  $\bar{r}_{t+1} \doteq \bar{r}_t + \alpha_t(R_{t+1} + \bar{x}_{t+1}^\top\theta_t^w - x_t^\top\theta_t^w - \bar{r}_t)$
  $\begin{bmatrix} \theta_{t+1}^r \\ \theta_{t+1}^w \end{bmatrix} \doteq \Gamma_{B_1}\big(\begin{bmatrix} \theta_t^r \\ \theta_t^w \end{bmatrix} + \beta_t(\Gamma_{B_2}(\begin{bmatrix} \bar{r}_t \\ w_t \end{bmatrix}) - \begin{bmatrix} \theta_t^r \\ \theta_t^w \end{bmatrix}))$
 end for

Average-reward Setting. In the average-reward setting, we need to learn both $\bar{r}$ and w. Hence, we consider target networks θ^r and θ^w for $\bar{r}$ and w respectively. Plugging θ^r and θ^w into (3) for bootstrapping yields Differential Q-evaluation with a Target Network (Algorithm 2), where $\{B_i\}$ are now balls in R^{K+1}. In Algorithm 2, we impose ridge regularization only on w as $\bar{r}$ is a scalar and thus does

not have any representation capacity limit.

Theorem 3. *Under Assumptions 2.1, 2.2, 3.1, 3.5, & 4.1, for any $\xi \in (0, 1)$, there exist constants C_0 and C_1 such that for all $\|X\| < C_0, C_1 < R_{B_1}, R_{B_1} - \xi < R_{B_2} < R_{B_1}$, the iterates $\{\bar{r}_t\}$ and $\{w_t\}$ generated by Algorithm 2 satisfy*

$$\lim_{t \to \infty} \bar{r}_t = d_\mu^\top (r + P_\pi X w_\eta^* - X w_\eta^*),$$
$$\lim_{t \to \infty} w_t = w_\eta^* \quad \textit{almost surely,}$$

where w_η^* is the unique solution of $(\bar{A} + \eta I)w - \bar{b} = \mathbf{0}$ with

$$\bar{A} \doteq X(D_\mu - d_\mu d_\mu^\top)(I - P_\pi)X,$$
$$\bar{b} \doteq X^\top (D_\mu - d_\mu d_\mu^\top)r.$$

If features are zero-centered (i.e., $X^\top d_\mu = \mathbf{0}$), then

$$\|Xw_\eta^* - \bar{q}_\pi^c\| \leq \left(\frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^4 \sigma_{\min}(D_\mu)^{2.5}} \|\bar{q}_\pi^c\| \eta \right.$$
$$\left. + \|\Pi_{D_\mu} \bar{q}_\pi^c - \bar{q}_\pi^c\| \right)/\xi,$$
$$|\bar{r}_\eta^* - \bar{r}_\pi| \leq \|d_\mu^\top (P_\pi - I)\| \inf_c \|(Xw_\eta^* - \bar{q}_\pi^c)\|,$$

where $\bar{q}_\pi^c \doteq \bar{q}_\pi + c\mathbf{1}$.

We defer the proof to Section A.4. As the differential Bellman equation (1) has infinitely many solutions for $\bar{q}$, all of which differ only by some constant offsets, we focus on analyzing the quality of Xw_η^* w.r.t. $\bar{q}_\pi^c$ in Theorem 3. The zero-centered feature assumption is also used in Zhang et al. (2021), which can be easily fulfilled in practice by subtracting all features with the estimated mean. In the on-policy case (i.e., $\mu = \pi$), we have $d_\mu^\top (P_\pi - I) = \mathbf{0}$, indicating $\bar{r}_\eta^* = \bar{r}_\pi$, i.e., the regularization on the value estimate does not pose any bias on the reward rate estimate.

As shown by Zhang et al. (2021), if the update (3) converges, it converges to w_0^*, the TD fixed point for off-policy policy evaluation in the average-reward setting, which satisfies $\bar{A}w_0^* + \bar{b} = 0$. Theorem 3 shows that Algorithm 2 converges to a regularized TD fixed point. Though Zhang et al. (2021) give a bound on $\|Xw_0^* - \bar{q}_\pi^c\|$, their bound holds only if μ is sufficiently close to π. By contrast, our bound on w_η^* in Theorem 3 holds for all μ.

5. Application to Off-Policy Control

Discounted Setting. Introducing a target network and ridge regularization in (4) yields Q-learning with a Target Network (Algorithm 3), where the behavior policy μ_θ depends on θ through the action-value estimate $X\theta$ and can be any policy satisfying the following two assumptions.

Assumption 5.1. *Let $\mathcal{P}$ be the closure of $\{P_{\mu_\theta} \mid \theta \in \mathbb{R}^K\}$. For any $P \in \mathcal{P}$, the Markov chain evolving in $\mathcal{S} \times \mathcal{A}$ induced by P is ergodic.*

Assumption 5.2. *$\mu_\theta(a|s)$ is Lipschitz continuous in θ.*

Assumption 5.1 is standard. When the behavior policy μ is fixed (independent of θ), the induced chain is usually assumed to be ergodic when analyzing the behavior of Q-learning (see, e.g., Melo et al. (2008); Chen et al. (2019b); Cai et al. (2019)). In Algorithm 3, the behavior policy μ_θ changes every step, so it is natural to assume that any of those behavior policies induces an ergodic chain. A similar assumption is also used by Zou et al. (2019) in their analysis of on-policy linear SARSA. Moreover, Zou et al. (2019) assume not only the ergodicity but also the *uniform* ergodicity of their sampling policies. Similarly, in Assumption 5.1, we assume ergodicity for not only all the transition matrices, but also their limits (c.f. the closure $\mathcal{P}$). A similar assumption is also used by Marbach & Tsitsiklis (2001) in their analysis of on-policy actor-critic methods. Assumption 5.2 can be easily fulfilled, e.g., by using a softmax policy w.r.t. $x(s, \cdot)^\top \theta$.

Algorithm 3 Q-learning with a Target Network

INPUT: $\eta > 0, R_{B_1} > R_{B_2} > 0$
Initialize $\theta_0 \in B_1$ and S_0
Sample $A_0 \sim \mu_{\theta_0}(\cdot|S_0)$
for $t = 0, 1, \dots$ **do**
    Execute A_t, get R_{t+1} and S_{t+1}
    Sample $A_{t+1} \sim \mu_{\theta_t}(\cdot|S_t)$
    $\delta_t \doteq R_{t+1} + \gamma \max_{a'} x(S_{t+1}, a')^\top \theta_t - x_t^\top w_t$
    $w_{t+1} \doteq w_t + \alpha_t \delta_t x_t - \alpha_t \eta w_t$
    $\theta_{t+1} \doteq \Gamma_{B_1}(\theta_t + \beta_t(\Gamma_{B_2}(w_t) - \theta_t))$
end for

Theorem 4. *Under Assumptions 2.2, 3.1, 3.5, 5.1, & 5.2, for any $\xi \in (0, 1), R_{B_1} > R_{B_2} > R_{B_1} - \xi > 0$, there exists a constant C_0 such that for all $\|X\| < C_0$, the iterate $\{w_t\}$ generated by Algorithm 3 satisfies*

$$\lim_{t \to \infty} w_t = w_\eta^* \quad \textit{almost surely,}$$

where w_η^* is the unique solution of

$$(A_{\pi_w, \mu_w} + \eta I)w - b_{\mu_w} = \mathbf{0} \tag{11}$$

inside B_1. Here

$$A_{\pi_w, \mu_w} \doteq X^\top D_{\mu_w}(I - \gamma P_{\pi_w})X, \ b_{\mu_w} \doteq X^\top D_{\mu_w}r,$$

and π_w denotes the greedy policy w.r.t. $x(s, \cdot)^\top w$.

We defer the proof to Section A.5. Analogously to the policy evaluation setting, if we call the solutions of $A_{\pi_w, \mu_w}w - b_{\mu_w} = \mathbf{0}$ TD fixed points for control in the discounted setting, then Theorem 4 asserts that Algorithm 3 finds a regularized TD fixed point.

Algorithm 3 and Theorem 4 are significant in two aspects. *First*, in Algorithm 3, the behavior policy is a function of

the target network and thus changes every time step. By contrast, previous work on Q-learning with function approximation (e.g., Melo et al. (2008); Maei et al. (2010); Chen et al. (2019b); Cai et al. (2019); Chen et al. (2019a); Lee & He (2019); Xu & Gu (2020); Carvalho et al. (2020); Wang & Zou (2020)) usually assumes the behavior policy is fixed. Though Fan et al. (2020) also adopt a changing behavior policy, they consider bi-level optimization. At each time step, the nested optimization problem must be solved exactly, which is computationally expensive and sometimes unfeasible. To the best of our knowledge, we are the first to analyze Q-learning with function approximation under a changing behavior policy and without nested optimization problems. Compared with the fixed behavior policy setting or the bi-level optimization setting, our two-timescale setting with a changing behavior policy is more closely related to actual practice (e.g., Mnih et al. (2015); Lillicrap et al. (2015)).

Second, Theorem 4 does not enforce any similarity between μ_θ and π_w; they can be arbitrarily different. By contrast, previous work (e.g., Melo et al. (2008); Chen et al. (2019b); Cai et al. (2019); Xu & Gu (2020); Lee & He (2019)) usually requires the strong assumption that the fixed behavior policy μ is sufficiently close to the target policy π_w. As the target policy (i.e., the greedy policy) can change every time step due to the changing action-value estimates, this strong assumption rarely holds. While some work removes this strong assumption, it introduces other problems instead. In Greedy-GQ, Maei et al. (2010) avoid this strong assumption by computing sub-gradients of an MSPBE objective MSPBE$(w) \doteq \|A_{\pi_w,\mu}w - b_\mu\|^2_{C_\mu^{-1}}$ directly, where $C_\mu \doteq X^\top D_\mu X$. If linear Q-learning (4) under a fixed behavior policy μ converges, it converges to the minimizer of MSPBE(w). Greedy-GQ, however, converges only to a stationary point of MSPBE(w). By contrast, Algorithm 3 converges to a minimizer of our regularized MSPBE (c.f. (11)). In Coupled Q-learning, Carvalho et al. (2020) avoid this strong assumption by using a target network as well, which they update as

$$\theta_{t+1} \leftarrow \theta_t + \alpha_t((x_t x_t^\top)w_t - \theta_t). \quad (12)$$

This target network update deviates much from the commonly used Polyak-averaging style update, while our (8) is identical to the Polyak-averaging style update most times if the balls for projection are sufficiently large. Coupled Q-learning updates the main network w as usual (see (6)). With the Coupled Q-learning updates (6) and (12), Carvalho et al. (2020) prove that the main network and the target network converge to $\bar{w}$ and $\bar{\theta}$ respectively, which satisfy

$$X\bar{w} = XX^\top D_\mu \mathcal{T}_{\pi_{\bar{w}}} X\bar{w}, \quad X\bar{\theta} = \Pi_{D_\mu} \mathcal{T}_{\pi_{\bar{w}}} X\bar{w}.$$

It is, however, not clear how $\bar{w}$ and $\bar{\theta}$ relate to TD fixed points. Yang et al. (2019) also use a target network to avoid

this strong assumption. Their target network update is the same as (8) except that they have only one projection Γ_{B_1}. Consequently, they face the problem of the reflection term $\zeta(t)$ (c.f. (9)). They also assume the main network $\{w_t\}$ is always bounded, a strong assumption that we do not require. Moreover, they consider a fixed sampling distribution for obtaining i.i.d. samples, while our data collection is done by executing the changing behavior policy μ_θ in the MDP.

One limit of Theorem 4 is that the bound on $\|X\|$ (i.e., C_0) depends on $1/R_{B_1}$ (see the proof in Section A.5 for the analytical expression), which means C_0 could potentially be small. Though we can use a small η accordingly to ensure that the regularization effect of η is modest, a small C_0 may not be desirable in some cases. To address this issue, we propose Gradient Q-learning with a Target Network, inspired by Greedy-GQ. We first equip MSPBE(w) with a changing behavior policy μ_w, yielding the following objective $\|A_{\pi_w,\mu_w}w - b_{\mu_w}\|^2_{C_{\mu_w}^{-1}}$. We then use the target network θ in place of w in the non-convex components, yielding

$$L(w,\theta) \doteq \|A_{\pi_\theta,\mu_\theta}w - b_{\mu_\theta}\|^2_{C_{\mu_\theta}^{-1}} + \eta\|w\|^2, \quad (13)$$

where we have also introduced a ridge term. At time step t, we update w_t following the gradient $\nabla_w L(w,\theta_t)$ and update the target network θ_t as usual. Details are provided in Algorithm 4, where the additional weight vector $u \in \mathbb{R}^K$ results from a weight duplication trick (see Sutton et al. (2009b;a) for details) to address a double sampling issue in estimating $\nabla_w L(w,\theta)$.

Algorithm 4 Gradient Q-learning with a Target Network

INPUT: $\eta > 0, R_{B_1} > R_{B_2} > 0$
Initialize $\theta_0 \in B_1$ and S_0
Sample $A_0 \sim \mu_{\theta_0}(\cdot|S_0)$
for $t = 0, 1, \ldots$ **do**
    Execute A_t, get R_{t+1} and S_{t+1}
    Sample $A_{t+1} \sim \mu_{\theta_t}(\cdot|S_t)$
    $\bar{x}_{t+1} \doteq \sum_{a'} \pi_{\theta_t}(a'|S_{t+1})x(S_{t+1}, a')$
    $\delta_t \doteq R_{t+1} + \gamma\bar{x}_{t+1}^\top w_t - x_t^\top w_t$
    $u_{t+1} \doteq u_t + \alpha_t(\delta_t - x_t^\top u_t)x_t$
    $w_{t+1} \doteq w_t + \alpha_t(x_t - \gamma\bar{x}_{t+1})x_t^\top u_t - \alpha_t\eta w_t$
    $\theta_{t+1} \doteq \Gamma_{B_1}\big(\theta_t + \beta_t(\Gamma_{B_2}(w_t) - \theta_t)\big)$
end for

In Algorithm 3, the target policy π_w is a greedy policy, which is not continuous in w. This discontinuity is not a problem there but requires sub-gradients in the analysis of Algorithm 4, which complicates the presentation. We, therefore, impose Assumption 5.2 on π_w as well.

Assumption 5.3. $\pi_\theta(a|s)$ *is Lipschitz continuous in* θ.

Though a greedy policy no longer satisfies Assumption 5.3, we can simply use a softmax policy.

Theorem 5. *Under Assumptions 2.2, 3.1, 3.5, & 5.1-5.3, there exist positive constants C_0 and C_1 such that for all $\|X\| < C_0, R_{B_1} > R_{B_2} > C_1$, the iterate $\{w_t\}$ generated by Algorithm 4 satisfies*

$$\lim_{t\to\infty} w_t = w_\eta^* \quad \text{almost surely,}$$

where w_η^* is the unique solution of

$$(A_{\pi_w,\mu_w}^\top C_{\mu_w}^{-1} A_{\pi_w,\mu_w} + \eta I)w = A_{\pi_w,\mu_w}^\top C_{\mu_w}^{-1} b_{\mu_w}.$$

We defer the proof to Section A.6. Importantly, the C_0 here does not depend on R_{B_1} and R_{B_2}. More importantly, the condition on $\|X\|$ (or equivalently, η) in Theorem 5 is only used to fulfill Assumption 3.4, without which $\{\theta_t\}$ in Algorithm 4 still converges to an invariant set of the ODE (10). This condition is to investigate where the iterate converges to instead of whether it converges or not. If we assume $w_0^* \doteq \lim_{\eta\to 0} w_\eta^*$ exists and $A_{\pi_{w_0^*},\mu_{w_0^*}}$ is invertible, we can see $A_{\pi_{w_0^*},\mu_{w_0^*}} w_0^* - b_{\mu_{w_0^*}} = \mathbf{0}$, indicating w_0^* is a TD fixed point. w_η^* can therefore be regarded as a regularized TD fixed point, though how the regularization is imposed here (c.f. (13)) is different from that in Algorithm 3 (c.f. (11)).

Average-reward Setting. Similar to Algorithm 2, introducing a target network and ridge regularization in (5) yields Differential Q-learning with a Target Network (Algorithm 5). Similar to Algorithm 2, $\{B_i\}$ are now balls in R^{K+1}.

Algorithm 5 Diff. Q-learning with a Target Network

INPUT: $\eta > 0, R_{B_1} > R_{B_2} > 0$
Initialize $[\theta_0^r, \theta_0^{w\top}]^\top \in B_1$ and S_0
Sample $A_0 \sim \mu(\cdot|S_0)$
for $t = 0, 1, \dots$ **do**
    Execute A_t, get R_{t+1} and S_{t+1}
    Sample $A_{t+1} \sim \mu_{\theta_t^w}(\cdot|S_{t+1})$
    $\delta_t \doteq R_{t+1} - \theta_t^r + \max_{a'} x(S_{t+1}, a')^\top \theta_t^w - x_t^\top w_t$
    $w_{t+1} \doteq w_t + \alpha_t \delta_t x_t - \alpha_t \eta w_t$
    $\delta_t' \doteq R_{t+1} + \max_{a'} x(S_{t+1}, a')^\top \theta_t^w - x_t^\top \theta_t^w - \bar{r}_t$
    $\bar{r}_{t+1} \doteq \bar{r}_t + \alpha_t \delta_t'$
    $\begin{bmatrix} \theta_{t+1}^r \\ \theta_{t+1}^w \end{bmatrix} \doteq \Gamma_{B_1}\left(\begin{bmatrix} \theta_t^r \\ \theta_t^w \end{bmatrix} + \beta_t(\Gamma_{B_2}(\begin{bmatrix} \bar{r}_t \\ w_t \end{bmatrix}) - \begin{bmatrix} \theta_t^r \\ \theta_t^w \end{bmatrix}))\right)$
end for

Theorem 6. *Under Assumptions 2.2, 3.1, 3.5, 5.1, & 5.2, let L_μ denote the Lipschitz constant of μ_θ, for any $\xi \in (0, 1), R_{B_1} > R_{B_2} > R_{B_1} - \xi > 0$, there exist constants C_0 and C_1 such that for all $\|X\| < C_0, L_\mu < C_1$, the iterate $\{w_t\}$ generated by Algorithm 5 satisfies*

$$\lim_{t\to\infty} w_t = w_\eta^* \quad \text{almost surely,}$$

where w_η^* is the unique solution of

$(\bar{A}_{\pi_w,\mu_w} + \eta I)w - \bar{b}_{\mu_w} = \mathbf{0}$ *inside B_1, where*

$$\bar{A}_{\pi_w,\mu_w} \doteq X(D_{\mu_w} - d_{\mu_w} d_{\mu_w}^\top)(I - P_{\pi_w})X,$$
$$\bar{b}_{\mu_w} \doteq X^\top(D_{\mu_w} - d_{\mu_w} d_{\mu_w}^\top)r,$$

and π_w is a greedy policy w.r.t. $x(s, \cdot)^\top w$.

We defer the proof to Section A.7. Theorem 6 requires μ_θ to be sufficiently smooth, which is a standard assumption even in the on-policy setting (e.g., Melo et al. (2008); Zou et al. (2019)). It is easy to see that if (5) converges, it converges to a solution of $\bar{A}_{\pi_w,\mu_w} w - \bar{b}_{\mu_w} = \mathbf{0}$, which we call a TD fixed point for control in the average-reward setting. Theorem 6, which shows that Algorithm 5 finds a regularized TD fixed point, is to the best of our knowledge the first theoretical study for linear Q-learning in the average-reward setting.

6. Experiments

All the implementations are publicly available. [1]

We first use Kolter's example (Kolter, 2011) to investigate how η influences the performance of w_η^* in the policy evaluation setting. Details are provided in Section D.1. This example is a two-state MDP with small representation error (i.e., $\|\Pi_{D_\mu} v_\pi - v_\pi\|$ is small). We vary the sampling probability of one state ($d_\mu(s_1)$) and compute corresponding w_η^* analytically. Figure 1a shows that with $\eta = 0$, the performance of w_η^* becomes arbitrarily poor when $d_\mu(s_1)$ approaches around 0.71. With $\eta = 0.01$, the spike exists as well. If we further increase η to 0.02 and 0.03, the performance for w_η^* becomes well bounded. This confirms the potential advantage of the regularized TD fixed points.

We then use Baird's example (Baird, 1995) to empirically investigate the convergence of the algorithms we propose. We use exactly the same setup as Chapter 11.2 of Sutton & Barto (2018). Details are provided in Section D.2. In particular, we consider three settings: policy evaluation (Figure 1b), control with a fixed behavior policy (Figure 1c), and control with an action-value dependent behavior policy (Figure 1d). For the policy evaluation setting, we compare a TD version of Algorithm 1 and standard Off-Policy Linear TD (possibly with ridge regularization). For the two control settings, we compare Algorithm 3 with standard linear Q-learning (possibly with ridge regularization). We use constant learning rates and do not use any projection in all the compared algorithms. The exact update rules are provided in Section D.2. Interestingly, Figures 1b-d show that even with $\eta = 0$, i.e., no ridge regularization, our algorithms with target network still converge in the tested domains. By contrast, without a target network, even when mild regularization is imposed, standard off-policy algorithms still diverge. This confirms the importance of the target network.
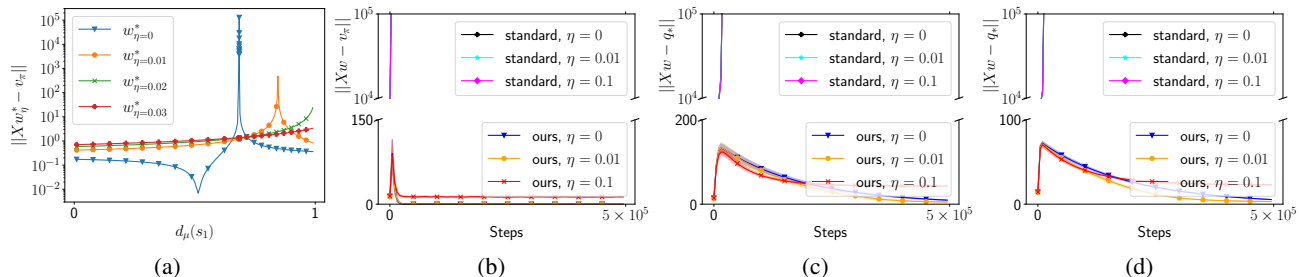
---

[1] https://github.com/ShangtongZhang/DeepRL

Figure 1. (a) Effect of regularization on Kolter's example. v_π is the true state-value function. (b) Policy evaluation on Baird's example. (c) Control on Baird's example with a fixed behavior policy. (d) Control on Baird's example with an action-value-dependent behavior policy. In (b)(c)(d), the curves are averaged over 30 independent runs with shaded region indicating one standard deviation. q_* is the optimal action-value function. η is the weight for the ridge term. Those marked "ours" are curves of algorithms we propose; those marked "standard" are standard semi-gradient off-policy algorithms. Interestingly, the three "standard" curves overlap and get unbounded quickly.

7. Discussion and Related Work

For all the algorithms we propose, both the target network and the ridge regularization are at play. One may wonder if it is possible to ensure convergence with only ridge regularization without the target network. In the policy evaluation setting, the answer is affirmative. Applying ridge regularization in (2) directly yields

$$w_{t+1} \leftarrow w_t + \alpha_t \delta_t x_t - \alpha_t \eta w_t, \qquad (14)$$

where δ_t is defined in (2). The expected update of (14) is

$$\begin{aligned} \Delta_w &\doteq b - (A + \eta I)w \\ &\doteq b - X^\top D_\mu X w + \gamma X^\top D_\mu (P_\pi X w) - \eta w. \end{aligned}$$

If its Jacobian w.r.t. w, denoted as $J_w(\Delta_w)$, is negative definite, the convergence of $\{w_t\}$ is expected (see, e.g., Section 5.5 of Vidyasagar (2002)). This negative definiteness can be easily achieved by ensuring $\eta > \|X\|^2 \|D_\mu(I - \gamma P_\pi)\|$ (see Diddigi et al. (2019) for similar techniques). This direct ridge regularization, however, would not work in the control setting. Consider, for example, linear Q-learning with ridge regularization (i.e., (14) with δ_t defined in (4)). The Jacobian of its expected update is $J_w(b_{\mu_w} - (A_{\pi_w, \mu_w} + \eta I)w)$. It is, however, not clear how to ensure this Jacobian is negative definite by tuning η. By using a target network for bootstrapping, $P_\pi X w$ becomes $P_\pi X \theta$. So $J_w(\Delta_w)$ becomes $-J_w(X^\top D_\mu X w + \eta w)$, which is always negative definite. Similarly, $J_w(b_{\mu_w} - (A_{\pi_w, \mu_w} + \eta)w)$ becomes $-J_w(X^\top D_{\mu_\theta} X w + \eta w)$ in Algorithm 3, which is always negative definite regardless of θ. The convergence of the main network $\{w_t\}$ can, therefore, be expected. The convergence of the target network $\{\theta_t\}$ is then delegated to Theorem 1. Now it is clear that in the deadly triad setting, the target network stabilizes training by ensuring the Jacobian of the expected update is negative definite. One may also wonder if it is possible to ensure convergence with only the target network without ridge regularization. The

answer is unclear. In our analysis, the conditions on $\|X\|$ (or equivalently, η) are only sufficient and not necessarily necessary. We do see in Figure 1 that even with $\eta = 0$, our algorithms still converge in the tested domains. How small η can be in general and under what circumstances η can be 0 are still open problems, which we leave for future work. Further, ridge regularization usually affects the convergence rate of the algorithm, which we also leave for future work.

In this paper, we investigate target network as one possible solution for the deadly triad. Other solutions include Gradient TD methods (Sutton et al. (2009b;a; 2016) for the discounted setting; Zhang et al. (2021) for the average-reward setting) and Emphatic TD methods (Sutton et al. (2016) for the discounted setting). Other convergence results of Q-learning with function approximation include Tsitsiklis & Van Roy (1996); Szepesvári & Smart (2004), which require special approximation architectures, Wen & Van Roy (2013); Du et al. (2020), which consider deterministic MDPs, Li et al. (2011); Du et al. (2019), which require a special oracle to guide exploration, Chen et al. (2019a), which require matrix inversion every time step, and Wang et al. (2019); Yang & Wang (2019; 2020); Jin et al. (2020), which consider linear MDPs (i.e., both p and r are assumed to be linear). Achiam et al. (2019) characterize the divergence of Q-learning with nonlinear function approximation via Taylor expansions and use preconditioning to empirically stabilize training. Van Hasselt et al. (2018) empirically study the role of a target network in the deadly triad setting in deep RL, which is complementary to our theoretical analysis.

Regularization is also widely used in RL. Yu (2017) introduce a general regularization term to improve the robustness of Gradient TD algorithms. Du et al. (2017) use ridge regularization in MSPBE to improve its convexity. Zhang et al. (2020) use ridge regularization to stabilize the training of critic in an off-policy actor-critic algorithm. Kolter & Ng

(2009); Johns et al. (2010); Petrik et al. (2010); Painter-Wakefield et al. (2012); Liu et al. (2012) use Lasso regularization in policy evaluation, mainly for feature selection.

8. Conclusion

In this paper, we proposed and analyzed a novel target network update rule, with which we improved several linear RL algorithms that are known to diverge previously due to the deadly triad. Our analysis provided a theoretical understanding, in the deadly triad setting, of the conventional wisdom that a target network stabilizes training. A possibility for future work is to introduce nonlinear function approximation, possibly over-parameterized neural networks, into our analysis.

Acknowledgments

References

Achiam, J., Knight, E., and Abbeel, P. Towards characterizing divergence in deep q-learning. *arXiv preprint arXiv:1903.08894*, 2019.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. *Machine Learning*, 1995.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009.

Boyan, J. A. Least-squares temporal difference learning. In *Proceedings of the 16th International Conference on Machine Learning*, 1999.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference and q-learning provably converge to global optima. *arXiv preprint arXiv:1905.10027*, 2019.

Carvalho, D., Melo, F. S., and Santos, P. A new convergent variant of q-learning with linear function approximation. *Advances in Neural Information Processing Systems*, 33, 2020.

Chen, S., Devraj, A. M., Bušić, A., and Meyn, S. Zap q-learning with nonlinear function approximation. *arXiv preprint arXiv:1910.05405*, 2019a.

Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *arXiv preprint arXiv:1905.11425*, 2019b.

Diddigi, R. B., Kamanchi, C., and Bhatnagar, S. A convergent off-policy temporal difference algorithm. *arXiv preprint arXiv:1911.05697*, 2019.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Du, S. S., Luo, Y., Wang, R., and Zhang, H. Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, 2019.

Du, S. S., Lee, J. D., Mahajan, G., and Wang, R. Agnostic q-learning with function approximation in deterministic systems: Near-optimal bounds on approximation error and sample complexity. *Advances in Neural Information Processing Systems*, 2020.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*. PMLR, 2020.

Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Johns, J., Painter-Wakefield, C., and Parr, R. Linear complementarity for regularized policy evaluation and improvement. *Advances in neural information processing systems*, 2010.

Kolter, J. Z. The fixed points of off-policy td. In *Advances in Neural Information Processing Systems*, 2011.

Kolter, J. Z. and Ng, A. Y. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th annual international conference on machine learning*, 2009.

Konda, V. R. *Actor-critic algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003.

Lee, D. and He, N. A unified switching system perspective and ode analysis of q-learning algorithms. *arXiv preprint arXiv:1912.02270*, 2019.

Li, L., Littman, M. L., Walsh, T. J., and Strehl, A. L. Knows what it knows: a framework for self-aware learning. *Machine learning*, 2011.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Lin, L.-J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 1992.

Liu, B., Mahadevan, S., and Liu, J. Regularized off-policy td-learning. *Advances in Neural Information Processing Systems*, 2012.

Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Marbach, P. and Tsitsiklis, J. N. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 2001.

Melo, F. S., Meyn, S. P., and Ribeiro, M. I. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

Painter-Wakefield, C., Parr, R., and Durham, N. L1 regularized linear temporal difference learning. *Technical report: Department of Computer Science, Duke University, Durham, NC, TR-2012–01*, 2012.

Petrik, M., Taylor, G., Parr, R., and Zilberstein, S. Feature selection using regularization in approximate linear programs for markov decision processes. *arXiv preprint arXiv:1005.1860*, 2010.

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

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press, 2018.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, 2009a.

Sutton, R. S., Maei, H. R., and Szepesvári, C. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2009b.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., and Precup, D. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, 2011.

Sutton, R. S., Mahmood, A. R., and White, M. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 2016.

Szepesvári, C. and Smart, W. D. Interpolation-based q-learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 100, 2004.

Tikhonov, A. N., Goncharsky, A., Stepanov, V., and Yagola, A. G. *Numerical methods for the solution of ill-posed problems*. Springer Science & Business Media, 2013.

Tsitsiklis, J. N. and Van Roy, B. Feature-based methods for large scale dynamic programming. *Machine Learning*, 1996.

Tsitsiklis, J. N. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, 1997.

Van Hasselt, H., Doron, Y., Strub, F., Hessel, M., Sonnerat, N., and Modayil, J. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

Vidyasagar, M. *Nonlinear systems analysis*. SIAM, 2002.

Wan, Y., Naik, A., and Sutton, R. S. Learning and planning in average-reward markov decision processes. *arXiv preprint arXiv:2006.16318*, 2020.

Wang, Y. and Zou, S. Finite-sample analysis of greedy-gq with linear function approximation under markovian noise. *arXiv preprint arXiv:2005.10175*, 2020.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Watkins, C. J. and Dayan, P. Q-learning. *Machine Learning*, 1992.

Wen, Z. and Van Roy, B. Efficient exploration and value function generalization in deterministic systems. *Advances in Neural Information Processing Systems*, 2013.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

Xu, P. and Gu, Q. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pp. 10555–10565. PMLR, 2020.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Yang, L. F. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. *arXiv preprint arXiv:1902.04779*, 2019.

Yang, Z., Fu, Z., Zhang, K., and Wang, Z. Convergent reinforcement learning with function approximation: A bilevel optimization perspective, 2019. URL https://openreview.net/forum?id=ryfcCo0ctQ.

Yu, H. Convergence of least squares temporal difference methods under general conditions. In *ICML*, 2010.

Yu, H. On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv preprint arXiv:1712.09652*, 2017.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Zhang, S., Wan, Y., Sutton, R. S., and Whiteson, S. Average-reward off-policy policy evaluation with function approximation. *arXiv preprint arXiv:2101.02808*, 2021.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, 2019.

A. Convergence of Target Networks

We first state a result from Borkar (2009) regarding the convergence of a linear system. Consider updating the parameter $y \in \mathbb{R}^K$ recursively as

$$y_{t+1} \doteq y_t + \beta_t(h(y_t) + \epsilon_t), \tag{15}$$

where $h : \mathbb{R}^K \to \mathbb{R}^K$ and $\{\epsilon_t\}$ is a deterministic or random bounded sequence satisfying $\lim_{t\to\infty} \|\epsilon_t\| = 0$. Assuming

Assumption A.1. h is Lipschitz continuous.

Assumption A.2. The learning rates $\{\beta_t\}$ satisfies $\sum_t \beta_t = \infty, \sum_t \beta_t^2 < \infty$.

Assumption A.3. $\sup_t \|y_t\| < \infty$ almost surely.

Theorem 7. *(The third extension of Theorem 2 in Chapter 2 of Borkar (2009)) Under Assumptions A.1- A.3, almost surely, the sequence $\{y_t\}$ generated by (15) converges to a compact connected internally chain transitive invariant set of the ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t} y(t) = h(y(t)).$$

A.1. Proof of Theorem 1

Proof. Similar to Chapter 5.4 of Borkar (2009), we consider $\dot{\Gamma}_{B_1}$, the directional derivative of Γ_{B_1}. At a point $x \in \mathbb{R}^K$, given a direction $y \in R^K$, we have

$$\dot{\Gamma}_{B_1}(x, y) \doteq \lim_{\delta \to 0} \frac{\Gamma_{B_1}(x + \delta y) - \Gamma_{B_1}(x)}{\delta}$$

$$= \begin{cases} y, & x \in int(B_1) \\ y, & x \in \partial B_1, y \in F_x(B_1) \\ -\frac{xx^\top y}{\|x\|^3} + \frac{y}{\|x\|}, & \text{otherwise} \end{cases}$$

where $int(B_1)$ is the interior of B_1, ∂B_1 is the boundary of B_1, $F_x(B_1) \doteq \{y \in \mathbb{R}^K \mid \exists \delta > 0, \quad \text{s.t.} \quad x + \delta y \in B_1\}$ is the feasible directions of B_1 w.r.t. x. The first two cases are trivial and are easy to deal with. The third case is complicated and is the source of the reflection term $\zeta(t)$ in (9). However, thanks to the projection Γ_{B_2}, we succeeded in getting rid of it.

By (8), $\theta_t \in B_1$ always holds. With the directional derivative, we can rewrite the update rule of $\{\theta_t\}$ as

$$\theta_{t+1} = \Gamma_{B_1}\big(\theta_t + \beta_t(\Gamma_{B_2}(w_t) - \theta_t)\big)$$

$$= \theta_t + \beta_t \frac{\Gamma_{B_1}\big(\theta_t + \beta_t(\Gamma_{B_2}(w_t) - \theta_t)\big) - \Gamma_{B_1}(\theta_t)}{\beta_t}$$

$$= \theta_t + \beta_t\big(\dot{\Gamma}_{B_1}(\theta_t, \Gamma_{B_2}(w_t) - \theta_t) + o(\beta_t)\big) \quad \text{(Definition of limit)} \quad .$$

We now compute $\dot{\Gamma}_{B_1}(\theta_t, \Gamma_{B_2}(w_t) - \theta_t)$. We proceed by showing that only the first two cases in $\dot{\Gamma}_{B_1}(x, y)$ can happen and the third case will never occur.

For $\theta_t \in int(B_1)$, we have

$$\dot{\Gamma}_{B_1}(\theta_t, \Gamma_{B_2}(w_t) - \theta_t) = \Gamma_{B_2}(w_t) - \theta_t. \tag{16}$$

For $\theta_t \in \partial B_1$,

$$\langle \theta_t, \Gamma_{B_2}(w_t) - \theta_t \rangle = \langle \theta_t, \Gamma_{B_2}(w_t) \rangle - R_{B_1}^2 \le R_{B_1} R_{B_2} - R_{B_1}^2 < 0. \tag{17}$$

Let $y_0 \doteq \Gamma_{B_2}(w_t) - \theta_t$, (17) implies that we can decompose y_0 as $y_0 = y_1 + y_2$, where $\langle \theta_t, y_1 \rangle = 0$ and $\langle \theta_t, y_2 \rangle = -\|\theta_t\|\|y_2\|$. Here y_2 is the projection of y_0 onto θ_t, which is in the opposite direction of θ_t and y_1 is the remaining orthogonal component. By Pythagoras's theorem, for any $\delta > 0$,

$$\|\theta_t + \delta y_0\|^2 = \|\delta y_1\|^2 + \|\theta_t + \delta y_2\|^2$$

$$= \delta^2 \|y_1\|^2 + \|\theta_t\|^2 - 2\delta\|\theta_t\|\|y_2\| + \delta^2 \|y_2\|^2.$$

For sufficiently small δ, e.g., $\delta^2\|y_1\|^2 - 2\delta\|\theta_t\|\|y_2\| + \delta^2\|y_2\|^2 < 0$, we have

$$\|\theta_t + \delta y_0\|^2 < \|\theta_t\|^2 = R_{B_1}^2,$$

implying $\Gamma_{B_2}(w_t) - \theta_t \in F_{\theta_t}(B_1)$. So we have

$$\dot{\Gamma}_{B_1}(\theta_t, \Gamma_{B_2}(w_t) - \theta_t) = \Gamma_{B_2}(w_t) - \theta_t. \tag{18}$$

Combining (16) and (18) yields

$$\begin{aligned}
\theta_{t+1} &= \theta_t + \beta_t\big(\Gamma_{B_2}(w_t) - \theta_t + o(\beta_t)\big)\\
&= \theta_t + \beta_t\big(\Gamma_{B_2}(w^*(\theta_t)) - \theta_t + o(1) + o(\beta_t)\big) \quad \text{(Assumption 3.2 and the continuity of } \Gamma_{B_2})\\
&= \theta_t + \beta_t\big(w^*(\theta_t) - \theta_t + o(1) + o(\beta_t)\big) \quad \text{(Assumption 3.3)}
\end{aligned}$$

Assumption A.1 is verified by Assumption 3.4; Assumption A.2 is verified by Assumption 3.1; Assumption A.3 is verified directly by the projection in (8). By Theorem 7, almost surely, $\{\theta_t\}$ converges to a compact connected internally chain transitive invariant set of the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = w^*(\theta(t)) - \theta(t).$$

Under Assumption 3.4, the Banach fixed-point theorem asserts that there is a unique θ^* satisfying $w^*(\theta^*) = \theta^*$, i.e., θ^* is the unique equilibrium of the ODE above. We now show θ^* is globally asymptotically stable. Consider the candidate Lyapunov function

$$V(\theta) \doteq \frac{1}{2}\|\theta - \theta^*\|^2.$$

We have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}V(\theta(t)) &= \langle\theta(t) - \theta^*, \frac{\mathrm{d}}{\mathrm{d}t}\theta(t)\rangle\\
&= \langle\theta(t) - \theta^*, w^*(\theta(t)) - \theta(t)\rangle\\
&= \langle\theta(t) - \theta^*, w^*(\theta(t)) - \theta(t) - w^*(\theta^*) + \theta^*\rangle\\
&\leq \varrho\|\theta(t) - \theta^*\|^2 - \|\theta(t) - \theta^*\|^2,
\end{aligned}$$

where $\varrho < 1$ is the Lipschitz constant of w^*. It is easy to see

- $V(\theta) \geq 0$

- $V(\theta) = 0 \iff \theta = \theta^*$

- $\frac{\mathrm{d}}{\mathrm{d}t}V(\theta(t)) \leq 0$

- $\frac{\mathrm{d}}{\mathrm{d}t}V(\theta(t)) = 0 \iff \theta = \theta^*$

Consequently, θ^* is globally asymptotically stable, implying

$$\lim_{t\to\infty}\theta_t = \theta^*,$$
$$\lim_{t\to\infty}w_t = \lim_{t\to\infty}w^*(\theta_t) = w^*(\theta^*) = \theta^*.$$

$\square$

A.2. Proof of Lemma 1

Proof. By (8), $\theta_t \in B_1$ holds for all t. So

$$
\begin{aligned}
&\|\theta_{t+1} - \theta_t\| \\
=&\left\|\Gamma_{B_1}\left(\theta_t + \beta_t(\Gamma_{B_2}(w_t) - \theta_t)\right) - \Gamma_{B_1}(\theta_t)\right\| \\
\leq&\beta_t\|\Gamma_{B_2}(w_t) - \theta_t\| \quad \text{(Nonexpansiveness of projection)} \\
\leq&\beta_t(R_{B_1} + R_{B_2}).
\end{aligned}
$$

$\square$

A.3. Proof of Theorem 2

Proof. Consider the Markov process $Y_t \doteq (S_t, A_t, S_{t+1})$. By Assumption 4.1, Y_t adopts a unique stationary distribution, which we refer to as $d_{\mathcal{Y}}$. We have $d_{\mathcal{Y}}(s, a, s') = d_\mu(s)\mu(a|s)p(s'|s, a)$. We define

$$
h_\theta(s, a, s') \doteq \left(r(s, a) + \gamma \sum_{a'} \pi(a'|s')x(s', a')^\top \Gamma_{B_1}(\theta)\right)x(s, a),
$$

$$
G_\theta(s, a, s') \doteq x(s, a)x(s, a)^\top + \eta I.
$$

As $\theta_t \in B_1$ holds for all t, we can rewrite the update of w_t in Algorithm 1 as

$$
w_{t+1} = w_t + \alpha_t(h_{\theta_t}(Y_t) - G_{\theta_t}(Y_t)w_t).
$$

The asymptotic behavior of $\{w_t\}$ is then governed by

$$
\begin{aligned}
\bar{h}(\theta) &\doteq \mathbb{E}_{(s,a,s')\sim d_{\mathcal{Y}}}[h_\theta(s, a, s')] \\
&= X^\top D_\mu r + \gamma X^\top D_\mu P_\pi X \Gamma_{B_1}(\theta), \\
\bar{G}(\theta) &\doteq \mathbb{E}_{(s,a,s')\sim d_{\mathcal{Y}}}[G_\theta(s, a, s')] \\
&= X^\top D_\mu X + \eta I.
\end{aligned}
$$

Define

$$
w^*(\theta) \doteq \bar{G}(\theta)^{-1}\bar{h}(\theta) = (X^\top D_\mu X + \eta I)^{-1}X^\top D_\mu(r + \gamma P_\pi X \Gamma_{B_1}(\theta)). \tag{19}
$$

We now verify Assumptions 3.2, 3.3, & 3.4 to invoke Theorem 1.

Assumption 3.2 is verified in Lemma 2.

To verify Assumption 3.4, we use SVD and get

$$
D_\mu^{\frac{1}{2}} X = U^\top \Sigma V,
$$

where U, V are two orthogonal matrices and $\Sigma \doteq diag([\ldots, \sigma_i, \ldots])$ is a diagonal matrix. Assumptions 4.1 & 2.2 imply that $\sigma_i > 0$. We have

$$
\left\|(X^\top D_\mu X + \eta I)^{-1}X^\top D_\mu^{\frac{1}{2}}\right\| = \left\|V^\top(\Sigma^2 + \eta I)^{-1}\Sigma U\right\| = \left\|(\Sigma^2 + \eta I)^{-1}\Sigma\right\| \tag{20}
$$

$$
= \left\|diag([\cdots, \frac{\sigma_i}{\sigma_i^2 + \eta}, \cdots])\right\| = \left\|diag([\cdots, \frac{1}{\sigma_i + \eta/\sigma_i}, \cdots])\right\| = \max_i \frac{1}{\sigma_i + \eta/\sigma_i}
$$

$$
\leq \frac{1}{2\sqrt{\eta}}
$$

According to (19), it is then easy to see

$$
\begin{aligned}
\|w^*(\theta_1) - w^*(\theta_2)\| &\leq \frac{\gamma}{2\sqrt{\eta}} \left\| D_\mu^{\frac{1}{2}} P_\pi \right\| \|X\| \|\Gamma_{B_1}(\theta_1) - \Gamma_{B_1}(\theta_2)\| \\
&\leq \frac{\gamma}{2\sqrt{\eta}} \left\| D_\mu^{\frac{1}{2}} P_\pi \right\| \|X\| \|\theta_1 - \theta_2\| \\
&\leq \frac{\gamma}{2\sqrt{\eta}} \left\| D_\mu^{\frac{1}{2}} P_\pi D_\mu^{-\frac{1}{2}} \right\| \left\| D_\mu^{\frac{1}{2}} \right\| \|X\| \|\theta_1 - \theta_2\| \\
&= \frac{\gamma}{2\sqrt{\eta}} \|P_\pi\|_{D_\mu} \left\| D_\mu^{\frac{1}{2}} \right\| \|X\| \|\theta_1 - \theta_2\| \quad \text{(See (20) in Kolter (2011))} \\
&\leq \frac{\gamma}{2\sqrt{\eta}} \|P_\pi\|_{D_\mu} \|X\| \|\theta_1 - \theta_2\|.
\end{aligned}
$$

Take any $\xi \in (0, 1)$, assuming

$$
\|X\| \leq \frac{2(1 - \xi)\sqrt{\eta}}{\gamma \|P_\pi\|_{D_\mu}}, \tag{21}
$$

then

$$
\|w^*(\theta_1) - w^*(\theta_2)\| \leq (1 - \xi)\|\theta_1 - \theta_2\|.
$$

Assumption 3.4, therefore, holds.

We now select proper R_{B_1} and R_{B_2} to fulfill Assumption 3.3. Plugging (20) and (21) into (19) yields

$$
\begin{aligned}
\|w^*(\theta)\| &\leq \frac{1}{2\sqrt{\eta}} \left\| D_\mu^{\frac{1}{2}} r \right\| + (1 - \xi) R_{B_1} \\
&= R_{B_1} - \xi + \left(\frac{1}{2\sqrt{\eta}} \left\| D_\mu^{\frac{1}{2}} r \right\| + \xi - \xi R_{B_1} \right)
\end{aligned}
$$

For sufficiently large R_{B_1}, e.g.,

$$
R_{B_1} \geq \frac{1}{2\xi\sqrt{\eta}} \left\| D_\mu^{\frac{1}{2}} r \right\| + 1, \tag{22}
$$

we have $\sup_\theta \|w^*(\theta)\| \leq R_{B_1} - \xi$. Selecting $R_{B_2} \in (R_{B_1} - \xi, R_{B_1})$ then fulfills Assumption 3.3.

With Assumptions 3.1 - 3.4 satisfied, Theorem 1 then implies that there exists a unique θ_∞ such that

$$
w^*(\theta_\infty) = \theta_\infty \quad \text{and} \quad \lim_{t \to \infty} \theta_t = \lim_{t \to \infty} w_t = \theta_\infty.
$$

Next we show what θ_∞ is. We define

$$
f(\theta) \doteq (X^\top D_\mu X + \eta I)^{-1} X^\top D_\mu (r + \gamma P_\pi X \theta).
$$

Note this is just the right side of equation (19) without the projection. (20) and (21) imply that f is a contraction. The Banach fixed-point theorem then asserts that f adopts a unique fixed point, which we refer to as w_η^*. We have

$$
\begin{aligned}
\|w_\eta^*\| = \|f(w_\eta^*)\| &\leq \frac{1}{2\sqrt{\eta}} \left(\|r\| + \gamma \|P_\pi\|_{D_\mu} \|X\| \|w_\eta^*\| \right) \\
&\leq \frac{\|r\|}{2\sqrt{\eta}} + (1 - \xi) \|w_\eta^*\| \quad \text{(By (21))} \\
\implies \|w_\eta^*\| &\leq \frac{\|r\|}{2\xi\sqrt{\eta}}
\end{aligned}
$$

Then for sufficiently large R_{B_1}, e.g.,

$$R_{B_1} \geq \frac{\|r\|}{2\xi\sqrt{\eta}}, \tag{23}$$

we have $w_\eta^* = \Gamma_{B_1}(w_\eta^*)$, implying w_η^* is a fixed point of $w^*(\cdot)$ (i.e., the right side of (19)) as well. As $w^*(\cdot)$ is a contraction, we have $\theta_\infty = w_\eta^*$. Rewriting $f(w_\eta^*) = w_\eta^*$ yields

$$Aw_\eta^* + \eta w_\eta^* - b = \mathbf{0}.$$

In other words, w_η^* is the unique (due to the contraction of f) solution of $(A + \eta I)w - b = \mathbf{0}$.

Combining (21), (22), and (23), the desired constants are

$$C_0 \doteq \frac{2(1 - \xi)\sqrt{\eta}}{\gamma\|P_\pi\|_{D_\mu}},$$

$$C_1 \doteq \frac{\|r\|}{2\xi\sqrt{\eta}} + 1.$$

We now bound $\left\|Xw_\eta^* - q_\pi\right\|$. For any $y \in \mathbb{R}^{|\mathcal{S}|}$, we define the ridge regularized projection $\Pi_{D_\mu}^\eta$ as

$$\Pi_{D_\mu}^\eta y \doteq X \arg\min_w \left(\|Xw - y\|_{D_\mu}^2 + \eta\|w\|^2\right)$$
$$= X(X^\top D_\mu X + \eta I)^{-1}X^\top D_\mu y.$$

$\Pi_{D_\mu}^\eta$ is connected with f as $\Pi_{D_\mu}^\eta \mathcal{T}_\pi(Xw) = Xf(w)$. We have

$$
\begin{aligned}
\left\|Xw_\eta^* - q_\pi\right\| &\leq \left\|Xw_\eta^* - \Pi_{D_\mu}^\eta q_\pi\right\| + \left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \\
&= \left\|Xf(w_\eta^*) - \Pi_{D_\mu}^\eta q_\pi\right\| + \left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \\
&= \left\|\Pi_{D_\mu}^\eta \mathcal{T}_\pi(Xw_\eta^*) - \Pi_{D_\mu}^\eta \mathcal{T}_\pi q_\pi\right\| + \left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \\
&= \left\|\Pi_{D_\mu}^\eta \gamma P_\pi Xw_\eta^* - \Pi_{D_\mu}^\eta \gamma P_\pi q_\pi\right\| + \left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \\
&= \left\|\gamma \Pi_{D_\mu}^\eta P_\pi(Xw_\eta^* - q_\pi)\right\| + \left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \\
&\leq \left\|\gamma X(X^\top D_\mu X + \eta I)^{-1}X^\top D_\mu P_\pi\right\|\left\|Xw_\eta^* - q_\pi\right\| + \left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \\
&\leq (1 - \xi)\left\|Xw_\eta^* - q_\pi\right\| + \left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \quad \text{(By (20) and (21))}
\end{aligned}
$$

The above equation implies

$$\left\|Xw_\eta^* - q_\pi\right\| \leq \frac{1}{\xi}\left\|\Pi_{D_\mu}^\eta q_\pi - q_\pi\right\| \leq \frac{1}{\xi}\left(\left\|\Pi_{D_\mu}^\eta q_\pi - \Pi_{D_\mu} q_\pi\right\| + \left\|\Pi_{D_\mu} q_\pi - q_\pi\right\|\right),$$

where Π_{D_μ} is shorthand for $\Pi_{D_\mu}^{\eta=0}$. We now bound $\left\| \Pi_{D_\mu}^{\eta} - \Pi_{D_\mu} \right\|$.

$$\left\| \Pi_{D_\mu}^{\eta} - \Pi_{D_\mu} \right\| \leq \|X\| \left\| (X^\top D_\mu X + \eta I)^{-1} - (X^\top D_\mu X)^{-1} \right\| \left\| X^\top D_\mu \right\| \tag{24}$$

$$= \left\| D_\mu^{-\frac{1}{2}} D_\mu^{\frac{1}{2}} X \right\| \left\| V^\top \left((\Sigma^2 + \eta I)^{-1} - \Sigma^{-2} \right) V \right\| \left\| X^\top D_\mu^{\frac{1}{2}} D_\mu^{\frac{1}{2}} \right\|$$

$$\leq \left\| D_\mu^{-\frac{1}{2}} \right\| \|\Sigma\| \left\| (\Sigma^2 + \eta I)^{-1} - \Sigma^{-2} \right\| \|\Sigma\| \left\| D_\mu^{\frac{1}{2}} \right\|$$

$$\leq \left\| D_\mu^{-\frac{1}{2}} \right\| \|\Sigma\| \left\| diag([\ldots, \frac{\eta}{\sigma_i^2(\eta + \sigma_i^2)}, \ldots]) \right\| \|\Sigma\| \left\| D_\mu^{\frac{1}{2}} \right\|$$

$$\leq \left\| D_\mu^{-\frac{1}{2}} \right\| \|\Sigma\| \max_i \frac{\eta}{\sigma_i^2(\eta + \sigma_i^2)} \|\Sigma\| \left\| D_\mu^{\frac{1}{2}} \right\|$$

$$\leq \left\| D_\mu^{-\frac{1}{2}} \right\| \|\Sigma\|^2 \max_i \frac{\eta}{\sigma_i^4} \left\| D_\mu^{\frac{1}{2}} \right\|$$

$$\leq \left\| D_\mu^{-\frac{1}{2}} \right\| \frac{\sigma_{\max}(\Sigma)^2}{\sigma_{\min}(\Sigma)^4} \left\| D_\mu^{\frac{1}{2}} \right\| \eta$$

($\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ indicate the largest and smallest singular values)

$$= \frac{\sigma_{\max}(D_\mu^{\frac{1}{2}} X)^2}{\sigma_{\min}(D_\mu^{\frac{1}{2}} X)^4} \frac{\sigma_{\max}(D_\mu^{\frac{1}{2}})}{\sigma_{\min}(D_\mu^{\frac{1}{2}})} \eta$$

$$\leq \frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^4 \sigma_{\min}(D_\mu)^{2.5}} \eta$$

($\sigma_{\max}(XY) \leq \sigma_{\max}(X)\sigma_{\max}(Y); \sigma_{\min}(XY) \geq \sigma_{\min}(X)\sigma_{\min}(Y)$)

.

Finally, we arrive at

$$\|X w_\eta^* - q_\pi\| \leq \frac{1}{\xi} \left(\frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^4 \sigma_{\min}(D_\mu)^{2.5}} \|q_\pi\| \eta + \left\| \Pi_{D_\mu} q_\pi - q_\pi \right\| \right),$$

which completes the proof. □

A.4. Proof of Theorem 3

Proof. The proof is similar to the proof of Theorem 2 in Section A.3. We, therefore, highlight only the difference to avoid verbatim repetition. Define

$$\theta \doteq \begin{bmatrix} \theta^r \\ \theta^w \end{bmatrix}, u \doteq \begin{bmatrix} \bar{r} \\ w \end{bmatrix},$$

$$h_\theta(s, a, s') \doteq \begin{bmatrix} r(s,a) \\ x(s,a)r(s,a) \end{bmatrix} + \begin{bmatrix} 0 & \sum_{a'} \pi(a'|s')x(s',a')^\top - x(s,a)^\top \\ -x(s,a) & x(s,a)\sum_{a'} \pi(a'|s')x(s',a')^\top \end{bmatrix} \Gamma_{B_1}(\theta),$$

$$G_\theta(s, a, s') \doteq \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & x(s,a)x(s,a)^\top + \eta I \end{bmatrix},$$

We can then rewrite the update of $\bar{r}$ and w in Algorithm 2 as

$$u_{t+1} = u_t + \alpha_t (h_{\theta_t}(Y_t) - G_{\theta_t}(Y_t)u_t).$$

Similarly, we define

$$\bar{h}(\theta) \doteq \mathbb{E}_{(s,a,s')\sim d_\mathcal{Y}}[h_\theta(s,a,s')] = \bar{h}_1 + \bar{H}_2\Gamma_{B_1}(\theta),$$

$$\bar{h}_1 \doteq \begin{bmatrix} d_\mu^\top r \\ X^\top D_\mu r \end{bmatrix}, \bar{H}_2 \doteq \begin{bmatrix} 0 & d_\mu^\top(P_\pi - I)X \\ -X^\top d_\mu & X^\top D_\mu P_\pi X \end{bmatrix}$$

$$\bar{G}(\theta) \doteq \mathbb{E}_{(s,a,s')\sim d_\mathcal{Y}}[G_\theta(s,a,s')] = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & X^\top D_\mu X + \eta I \end{bmatrix},$$

$$u^*(\theta) \doteq \bar{G}(\theta)^{-1}\bar{h}(\theta).$$

We proceed to verifying Assumptions 3.2, 3.3, & 3.4 to invoke Theorem 1.

Assumption 3.2 is verified in Lemma 3.

For Assumption 3.4 to hold, note

$$\|\bar{G}(\theta)^{-1}\| = \max\left\{1, \|(X^\top D_\mu X + \eta I)^{-1}\|\right\} \leq \max\left\{1, \frac{1}{\eta}\right\}$$

$$\|\bar{H}_2\|^2 = \max_{\|u\|=1}\|\bar{H}_2 u\|^2 = \max_{\|u\|=1}\left\| \begin{bmatrix} d_\mu^\top(P_\pi - I)Xw \\ -X^\top d_\mu \bar{r} + X^\top D_\mu P_\pi Xw \end{bmatrix} \right\|^2$$

$$= \max_{\|u\|=1}\left\|d_\mu^\top(P_\pi - I)Xw\right\|^2 + \left\|-X^\top d_\mu\bar{r} + X^\top D_\mu P_\pi Xw\right\|^2$$

$$\leq \max_{\|u\|=1}\left\|d_\mu^\top(P_\pi - I)X\right\|^2\|w\|^2 + 2\|X^\top d_\mu\|^2\|\bar{r}\|^2 + 2\|X^\top D_\mu P_\pi X\|^2\|w\|^2$$

$$\leq \max\left\{\|d_\mu^\top(P_\pi - I)X\|^2 + 2\|X^\top D_\mu P_\pi X\|^2, 2\|X^\top d_\mu\|^2\right\}$$

$$\implies \|\bar{H}_2\| \leq \|X\|\max\left\{\|d_\mu^\top(P_\pi - I)\| + \sqrt{2}\|D_\mu P_\pi\|, \sqrt{2}\|d_\mu\|\right\}.$$

The above equations suggest that

$$\|u^*(\theta_1) - u^*(\theta_2)\| \leq \max\left\{1, \frac{1}{\eta}\right\}\|X\|\max\left\{\|d_\mu^\top(P_\pi - I)\| + \sqrt{2}\|D_\mu P_\pi\|, \sqrt{2}\|d_\mu\|\right\}\|\theta_1 - \theta_2\|.$$

Take any $\xi \in (0,1)$, assuming

$$\|X\| \leq \frac{1-\xi}{\max\left\{1, \frac{1}{\eta}\right\}\max\left\{\|d_\mu^\top(P_\pi - I)\| + \sqrt{2}\|D_\mu P_\pi\|, \sqrt{2}\|d_\mu\|\right\}}, \tag{25}$$

then

$$\|u^*(\theta_1) - u^*(\theta_2)\| \leq (1-\xi)\|\theta_1 - \theta_2\|.$$

Assumption 3.4, therefore, holds.

We now select proper R_{B_1} and R_{B_2} to fulfill Assumption 3.3. Using (25), it is easy to see

$$\|u^*(\theta)\| \leq \max\left\{1, \frac{1}{\eta}\right\}\|\bar{h}_1\| + (1-\xi)R_{B_1}.$$

For sufficiently large R_{B_1}, e.g.,

$$R_{B_1} \geq \max\left\{1, \frac{1}{\eta}\right\}\frac{\|\bar{h}_1\|}{\xi} + 1, \tag{26}$$

we have $\sup_\theta \|w^*(\theta)\| \leq R_{B_1} - \xi$. Selecting $R_{B_2} \in (R_{B_1} - \xi, R_{B_1})$ then fulfills Assumption 3.3.

With Assumptions 3.1 - 3.4 satisfied, Theorem 1 then implies that there exists a unique θ_∞ such that

$$u^*(\theta_\infty) = \theta_\infty \quad \text{and} \quad \lim_{t \to \infty} \theta_t = \lim_{t \to \infty} u_t = \theta_\infty.$$

Next we show what θ_∞ is. We define

$$f(\theta) \doteq \bar{G}(\theta)^{-1}(\bar{h}_1 + \bar{H}_2 \theta).$$

Note this is just $u^*(\theta)$ without the projection. Under (25), it is easy to show f is a contraction. The Banach fixed-point theorem then asserts that f adopts a unique fixed point, which we refer to as u_η^*. Using (25) again, we get

$$\|u_\eta^*\| = \|f(u_\eta^*)\| \leq \max\left\{1, \frac{1}{\eta}\right\}\|\bar{h}_1\| + (1 - \xi)\|w_\eta^*\|$$

$$\implies \|u_\eta^*\| \leq \max\left\{1, \frac{1}{\eta}\right\}\frac{\|\bar{h}_1\|}{\xi}$$

Then for sufficiently large R_{B_1}, e.g.,

$$R_{B_1} \geq \max\left\{1, \frac{1}{\eta}\right\}\frac{\|\bar{h}_1\|}{\xi}, \tag{27}$$

we have $u_\eta^* = \Gamma_{B_1}(u_\eta^*)$, implying u_η^* is a fixed point of $u^*(\cdot)$ as well. As $u^*(\cdot)$ is a contraction, we have $\theta_\infty = u_\eta^*$. Writing u_η^* as $\begin{bmatrix} \bar{r}_\eta^* \\ w_\eta^* \end{bmatrix}$ and expanding $f(u_\eta^*) = u_\eta^*$ yields

$$\bar{r}_\eta^* = d_\mu^\top(r + P_\pi X w_\eta^* - X w_\eta^*),$$
$$(X^\top D_\mu X + \eta I)w_\eta^* = X^\top D_\mu r - X^\top d_\mu \bar{r}_\eta^* + X^\top D_\mu P_\pi X w_\eta^*.$$

Rearranging terms yields $(\bar{A} + \eta I)w_\eta^* - \bar{b} = 0$, i.e., w_η^* is the unique (due to the contraction of f) solution of $(\bar{A} + \eta I)w - \bar{b} = 0$.

We now bound $\|Xw_\eta^* - \bar{q}_\pi^c\|$.

$$\|Xw_\eta^* - \bar{q}_\pi^c\| \leq \left\|Xw_\eta^* - \Pi_{D_\mu}^\eta \bar{q}_\pi^c\right\| + \left\|\Pi_{D_\mu}^\eta \bar{q}_\pi^c - \bar{q}_\pi^c\right\|$$

$$= \left\|X(X^\top D_\mu X + \eta I)^{-1}(X^\top D_\mu r - X^\top d_\mu \bar{r}_\eta^* + X^\top D_\mu P_\pi X w_\eta^*) - \Pi_{D_\mu}^\eta \bar{q}_\pi^c\right\| + \left\|\Pi_{D_\mu}^\eta \bar{q}_\pi^c - \bar{q}_\pi^c\right\|$$

$$= \left\|\Pi_{D_\mu}^\eta(r + P_\pi X w_\eta^*) - \Pi_{D_\mu}^\eta(r + P_\pi \bar{q}_\pi^c)\right\| + \left\|\Pi_{D_\mu}^\eta \bar{q}_\pi^c - \bar{q}_\pi^c\right\| \quad (\text{Using } X^\top d_\mu = \mathbf{0})$$

$$= \left\|\Pi_{D_\mu}^\eta P_\pi(X w_\eta^* - \bar{q}_\pi^c)\right\| + \left\|\Pi_{D_\mu}^\eta \bar{q}_\pi^c - \bar{q}_\pi^c\right\|$$

$$\leq \left\|X(X^\top D_\mu X + \eta I)^{-1}X^\top D_\mu P_\pi\right\|\|X w_\eta^* - \bar{q}_\pi^c\| + \left\|\Pi_{D_\mu}^\eta \bar{q}_\pi^c - \bar{q}_\pi^c\right\|$$

$$\leq \frac{1}{\eta}\|X\|^2\|D_\mu P_\pi\|\|X w_\eta^* - \bar{q}_\pi^c\| + \left\|\Pi_{D_\mu}^\eta \bar{q}_\pi^c - \bar{q}_\pi^c\right\|.$$

Assuming

$$\|X\|^2 \leq \frac{(1 - \xi)\eta}{\|D_\mu P_\pi\|}, \tag{28}$$

we have

$$\|X w_\eta^* - \bar{q}_\pi^c\| \leq \frac{1}{\xi}\left\|\Pi_{D_\mu}^\eta \bar{q}_\pi^c - \bar{q}_\pi^c\right\| \leq \frac{1}{\xi}\left(\frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^4 \sigma_{\min}(D_\mu)^{2.5}}\|\bar{q}_\pi^c\|\eta + \left\|\Pi_{D_\mu}\bar{q}_\pi^c - \bar{q}_\pi^c\right\|\right) \quad (\text{c.f. (24)}) \quad .$$

It is then easy to see

$$|\bar{r}_\eta^* - \bar{r}_\pi| \le \left\| d_\mu^\top (P_\pi - I)(Xw_\eta^* - \bar{q}_\pi^c) \right\| \le \frac{1}{\xi} \left\| d_\mu^\top (P_\pi - I) \right\| \left(\frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^4 \sigma_{\min}(D_\mu)^{2.5}} \|\bar{q}_\pi^c\| \eta + \left\| \Pi_{D_\mu} \bar{q}_\pi^c - \bar{q}_\pi^c \right\| \right).$$

Taking infimum for $c \in \mathbb{R}$ then yields the desired results.

Combining (25), (26), (27), and (28), the desired constants are

$$C_0 \doteq \min \left\{ \frac{1 - \xi}{\max\left\{ 1, \frac{1}{\eta} \right\} \max\left\{ \left\| d_\mu^\top (P_\pi - I) \right\| + \sqrt{2} \|D_\mu P_\pi\|, \sqrt{2} \|d_\mu\| \right\}}, \sqrt{\frac{(1 - \xi)\eta}{\|D_\mu P_\pi\|}} \right\},$$

$$C_1 \doteq \max\left\{ 1, \frac{1}{\eta} \right\} \frac{\|\bar{h}_1\|}{\xi} + 1,$$

which completes the proof. $\qquad\square$

A.5. Proof of Theorem 4

Proof. The proof is similar to the proof of Theorem 2 in Section A.3 but is more involving. We define

$$h_\theta(s, a, s') \doteq \left(r(s, a) + \gamma \max_{a'} x(s', a')^\top \theta \right) x(s, a),$$

$$G_\theta(s, a, s') \doteq x(s, a) x(s, a)^\top + \eta I,$$

where $\bar{\theta} \doteq \Gamma_{B_1}(\theta)$ is shorthand. As $\theta_t \in B_1$ holds for all t, we can rewrite the update of w_t in Algorithm 3 as

$$w_{t+1} = w_t + \alpha_t(h_{\theta_t}(Y_t) - G_{\theta_t}(Y_t)w_t).$$

The expected update given θ is then controlled by

$$\bar{h}(\theta) \doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[h_\theta(s, a, s')]$$
$$= X^\top D_{\mu_\theta} r + \gamma X^\top D_{\mu_\theta} P_{\pi_{\bar{\theta}}} X\bar{\theta},$$
$$\bar{G}(\theta) \doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[G_\theta(s, a, s')]$$
$$= X^\top D_{\mu_\theta} X + \eta I,$$

where Assumption 5.1 ensures the existence of d_{μ_θ} and π_θ is the target policy, i.e. a greedy policy with random tie breaking defined as follows.

Let $\mathcal{A}_{s,\theta}^{\max} \doteq \arg\max_a x(s, a)^\top \theta$ be the set of maximizing actions for state s, we define

$$\pi_\theta(a|s) \doteq \begin{cases} \frac{1}{|\mathcal{A}_{s,\theta}^{\max}|}, & a \in \mathcal{A}_{s,\theta}^{\max} \\ 0, & \text{otherwise} \end{cases}.$$

Similar to the proof in Section A.3, we define

$$w^*(\theta) \doteq \bar{G}(\theta)^{-1} \bar{h}(\theta) = (X^\top D_{\mu_\theta} X + \eta I)^{-1} X^\top D_{\mu_\theta} (r + \gamma P_{\pi_{\bar{\theta}}} X\bar{\theta}) \tag{29}$$

and proceed to verify Assumptions 3.2 - 3.4 to invoke Theorem 1.

Assumption 3.2 is proved in Lemma 4.

For Assumption 3.4, Lemma 10 shows that $w^*(\theta)$ is Lipschitz continuous in θ with

$$C_w \doteq \eta^{-1} \|X\| \|r\|_{L_D} + \eta^{-2} \|X\|^3 \|r\|_{L_D} + \gamma \eta^{-1} L_0 \|X\|^2 + \gamma U_P \|X\| R_{B_1} (\eta^{-1} \|X\|_{L_D} + \eta^{-2} \|X\|^3 L_D)$$

being a Lipschitz constant. Here L_D, L_0, and U_P are positive constants detailed in the proof of Lemma 10. Assuming

$$\|X\| \le 1 \quad \text{and} \quad \gamma U_P \|X\| R_{B_1} \le 1, \tag{30}$$

we have

$$C_w \leq \eta^{-2} \|X\| (\eta \|r\| L_D + \|r\| L_D + \gamma \eta L_0 + \eta L_D + L_D).$$

Take any $\xi \in (0, 1)$, assuming

$$\|X\| \leq (1 - \xi) \eta^2 (\eta \|r\| L_D + \|r\| L_D + \gamma \eta L_0 + \eta L_D + L_D)^{-1}, \tag{31}$$

it then follows that $C_w \leq 1 - \xi$. Assumptions 3.4, therefore, holds.

We now select proper R_{B_1} and R_{B_2} to fulfill Assumption 3.3. Similar to Lemma 10 (see, e.g., the last three rows of Table 1 in the proof of Lemma 10), we can easily get

$$\|w^*(\theta)\| \leq \eta^{-1} \|X\| \|r\| + \gamma \eta^{-1} \|X\| U_P \|X\| R_{B_1}.$$

Using (30) yields

$$\|w^*(\theta)\| \leq \eta^{-1} \|X\| (\|r\| + 1).$$

For sufficiently large R_{B_1}, e.g.,

$$R_{B_1} > \eta^{-1} \|X\| (\|r\| + 1) + \xi, \tag{32}$$

we have $\sup_\theta \|w^*(\theta)\| < R_{B_1} - \xi$. Taking $R_{B_2} \in (R_{B_1} - \xi, R_{B_1})$ then fulfills Assumption 3.3.

With Assumptions 3.1 - 3.4 satisfied, Theorem 1 implies that there exists a unique θ_∞ such that

$$w^*(\theta_\infty) = \theta_\infty \quad \text{and} \quad \lim_{t \to \infty} \theta_t = \lim_{t \to \infty} w_t = \theta_\infty.$$

We now show what θ_∞ is. We define

$$f(\theta) \doteq (X^\top D_{\mu_\theta} X + \eta I)^{-1} X^\top D_{\mu_\theta} (r + \gamma P_{\pi_\theta} X \theta)$$

and consider a ball $B_0 \doteq \{\theta \in \mathbb{R}^K \mid \|\theta\| \leq R_{B_0}\}$ with R_{B_0} to be tuned (for the Brouwer fixed-point theorem). We have

$$\|f(\theta)\| \leq \eta^{-1} \|X\| \|r\| + \gamma \eta^{-1} \|X\|^2 U_P \|\theta\|$$

Assuming

$$\gamma \eta^{-1} \|X\|^2 U_P < 1 - \xi, \tag{33}$$

we have

$$\begin{aligned}
\|f(\theta)\| &\leq \eta^{-1} \|X\| \|r\| + (1 - \xi) \|\theta\| \\
&= R_{B_0} - (R_{B_0} - (1 - \xi) \|\theta\| - \eta^{-1} \|X\| \|r\|).
\end{aligned}$$

Then for sufficiently large R_{B_0}, e.g.,

$$R_{B_0} \geq \frac{\|X\| \|r\|}{\eta \xi},$$

we have

$$\theta \in B_0 \implies f(\theta) \in B_0.$$

The Brouwer fixed-point theorem then asserts that there exists a $w^*_\eta \in B_0$ such that $f(w^*_\eta) = w^*_\eta$. For sufficiently large R_{B_1}, e.g.,

$$R_{B_1} > R_{B_0}, \tag{34}$$

we have $\Gamma_{B_1}(w_\eta^*) = w_\eta^*$, i.e., w_η^* is also a fixed point of $w^*(\cdot)$. The contraction of $w^*(\cdot)$ then implies $\theta_\infty = w_\eta^*$. Rewriting $f(w_\eta^*) = w_\eta^*$ yields

$$A_{\pi_{w_\eta^*}, \mu_{w_\eta^*}} w_\eta^* + \eta w_\eta^* - b_{\mu_{w_\eta^*}} = \mathbf{0}.$$

In other words, w_η^* is the unique solution of $(A_{\pi_w, \mu_w} + \eta I)w - b_{\mu_w} = \mathbf{0}$ inside B_1 (due to the contraction of $w^*(\cdot)$). Combining (30) (31) (33) (32) (34), the desired constant is

$$C_0 \doteq \min\{1, \frac{1}{\gamma U_P R_{B_1}}, \frac{\eta(R_{B_1} - \xi)}{\|r\| + 1}, \sqrt{\frac{\eta(1 - \xi)}{\gamma U_P}}, \frac{R_{B_1}\eta\xi}{\|r\|},$$

$$\frac{(1 - \xi)\eta^2}{\eta\|r\|L_D + \|r\|L_D + \gamma\eta L_0 + \eta L_D + L_D}\},$$

which completes the proof. As R_{B_1} is usually large, in general C_0 is $\mathcal{O}(R_{B_1}^{-1})$. Though C_0 is potentially small, we can use small η as well. So a small C_0 (i.e., $\|X\|$) does not necessarily implies a large regularization bias. □

A.6. Proof of Theorem 5

Proof. Let $\kappa_t \doteq [u_t^\top, w_t^\top]^\top$, Algorithms 4 implies that

$$\kappa_{t+1} = \kappa_t + \alpha_t(h_{\theta_t}(Y_t) - G_{\theta_t}(Y_t)\kappa_t),$$

where

$$G_\theta(s, a, s') \doteq \begin{bmatrix} x(s,a)x(s,a)^\top & -x(s,a)(\gamma\sum_{a'}\pi_\theta(a'|s')x(s,a) - x(s,a))^\top \\ (\gamma\sum_{a'}\pi_\theta(a'|s')x(s,a) - x(s,a))x(s,a)^\top & \eta I \end{bmatrix},$$

$$h_\theta(s, a, s') \doteq \begin{bmatrix} x(s,a)r(s,a) \\ \mathbf{0} \end{bmatrix}.$$

We define

$$C_{\mu_\theta} \doteq X^\top D_{\mu_\theta} X,$$

$$\bar{G}(\theta) \doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[G_\theta(s, a, s')] = \begin{bmatrix} C_{\mu_\theta} & A_{\pi_\theta, \mu_\theta} \\ -A_{\pi_\theta, \mu_\theta}^\top & \eta I \end{bmatrix},$$

$$\bar{h}(\theta) \doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[h_\theta(s, a, s')] = \begin{bmatrix} X^\top D_{\mu_\theta} r \\ \mathbf{0} \end{bmatrix},$$

$$w^*(\theta) \doteq [\bar{G}(\theta)^{-1}\bar{h}(\theta)]_{K+1:2K} = (\eta I + A_{\pi_\theta, \mu_\theta}^\top C_{\mu_\theta}^{-1} A_{\pi_\theta, \mu_\theta})^{-1} A_{\pi_\theta, \mu_\theta}^\top C_{\mu_\theta}^{-1} X^\top D_{\mu_\theta} r, \quad (35)$$

where $[\cdot]_{K+1:2K}$ is the subvector indexed from $K + 1$ to $2K$.

We proceed to verifying Assumptions 3.2, 3.3, & 3.4 thus invoke Theorem 1.

Assumption 3.2 is verified in Lemma 5. Lemma 11 shows that if $\|X\| \leq 1$, there exist a constant $L_w > 0$, independent of $\|X\|$, such that

$$\|w^*(\theta_1) - w^*(\theta_2)\| \leq L_w \|X\| \|\theta_1 - \theta_2\|.$$

As $w^*(\cdot)$ is independent of R_{B_1}, so does L_w. So as long as

$$\|X\| \leq C_0 \doteq \min\left\{1, \frac{1 - \xi}{L_w}\right\},$$

$w^*(\cdot)$ is contractive and Assumption 3.4 is satisfied. In the proof of Lemma 11, we show $\sup_\theta \|C_{\mu_\theta}^{-1}\| < \infty$. It is then easy to see $\sup_\theta \|w^*(\theta)\| < \infty$. Consequently, we can choose sufficiently large R_{B_1} and R_{B_2} such that Assumption 3.3 holds.

With Assumptions 3.1 - 3.4 satisfied, Theorem 1 implies that there exists a unique w_η^* such that

$$w^*(w_\eta^*) = w_\eta^* \quad \text{and} \quad \lim_{t\to\infty} \theta_t = \lim_{t\to\infty} w_t = w_\eta^*.$$

Expanding $w^*(w_\eta^*) = w_\eta^*$ yields

$$w_\eta^* = (A_{\pi_{w_\eta^*}, \mu_{w_\eta^*}}^\top C_{\mu_{w_\eta^*}}^{-1} A_{\pi_{w_\eta^*}, \mu_{w_\eta^*}} + \eta I)^{-1} A_{\pi_{w_\eta^*}, \mu_{w_\eta^*}}^\top C_{\mu_{w_\eta^*}}^{-1} b_{\mu_{w_\eta^*}},$$

which completes the proof. $\qquad\square$

A.7. Proof of Theorem 6

Proof. The proof is combination of the proofs of Theorem 3 and Theorem 4. To avoid verbatim repetition, in this proof, we show only the existence of the constants C_0 and C_1 without showing the exact expressions. We define

$$\theta \doteq \begin{bmatrix} \theta^r \\ \theta^w \end{bmatrix}, u \doteq \begin{bmatrix} \bar{r} \\ w \end{bmatrix}, \begin{bmatrix} \bar{\theta}^r \\ \bar{\theta}^w \end{bmatrix} \doteq \Gamma_{B_1}(\begin{bmatrix} \theta^r \\ \theta^w \end{bmatrix}),$$

$$h_\theta(s, a, s') \doteq \begin{bmatrix} r(s, a) \\ x(s, a) r(s, a) \end{bmatrix} + \begin{bmatrix} 0 & \sum_{a'} \pi_{\bar{\theta}^w}(a'|s') x(s', a')^\top \bar{\theta}^w - x(s, a)^\top \bar{\theta}^w \\ -x(s, a) \bar{\theta}^r & x(s, a) \sum_{a'} \pi_{\bar{\theta}^w}(a'|s') x(s', a')^\top \bar{\theta}^w \end{bmatrix},$$

$$G_\theta(s, a, s') \doteq \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & x(s, a) x(s, a)^\top + \eta I \end{bmatrix},$$

We can then rewrite the update of $\bar{r}$ and w in Algorithm 5 as

$$u_{t+1} = u_t + \alpha_t (h_{\theta_t}(Y_t) - G_{\theta_t}(Y_t) u_t).$$

In the rest of this proof, we write μ_θ and π_θ as shorthand for μ_{θ^w} and π_{θ^w}. We define

$$\bar{h}(\theta) \doteq \mathbb{E}_{(s,a) \sim d_{\mu_\theta}, s' \sim p(\cdot|s,a)}[h_\theta(s, a, s')] = \bar{h}_1(\theta) + \bar{H}_2(\theta),$$

$$\bar{h}_1(\theta) \doteq \begin{bmatrix} d_{\mu_\theta}^\top r \\ X^\top D_{\mu_\theta} r \end{bmatrix}, \bar{H}_2(\theta) \doteq \begin{bmatrix} 0 & d_{\mu_\theta}^\top (P_{\pi_{\bar{\theta}^w}} - I) X \bar{\theta}^w \\ -(X^\top d_{\mu_\theta}) \bar{\theta}^r & X^\top D_{\mu_\theta} P_{\pi_{\bar{\theta}^w}} X \bar{\theta}^w \end{bmatrix},$$

$$\bar{G}(\theta) \doteq \mathbb{E}_{(s,a) \sim d_{\mu_\theta}, s' \sim p(\cdot|s,a)}[G_\theta(s, a, s')] = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & X^\top D_{\mu_\theta} X + \eta I \end{bmatrix},$$

$$u^*(\theta) \doteq \bar{G}(\theta)^{-1} \bar{h}(\theta). \tag{36}$$

We proceed to verifying Assumptions 3.2, 3.3, & 3.4 to invoke Theorem 1.

Assumption 3.2 is verified in Lemma 6.

For Assumption 3.4, Lemma 12 suggests that assuming $\|X\| \le 1$, $L_\mu \le 1$, then

$$C_u = \max\{1, \eta^{-1}\}(\mathcal{O}(\|X\|) + \mathcal{O}(L_\mu)) + \max\{1, \eta^{-2}\}\mathcal{O}(\|X\|)$$

is a Lipschitz constant of $u^*(\theta)$. Take any $\xi \in (0, 1)$, it is easy to see there exists positive constants C_2 and C_3 such that

$$\|X\| \le C_2, L_\mu \le C_3 \implies C_u \le 1 - \xi. \tag{37}$$

Assumption 3.4, therefore, holds.

We now select proper R_{B_1} and R_{B_2} to fulfill Assumption 3.3. Using (37), it is easy to see

$$\|u^*(\theta)\| \le C_4 + (1 - \xi) R_{B_1}$$

for some positive constant C_4. For sufficiently large R_{B_1}, e.g.,

$$R_{B_1} \ge \max \frac{C_4}{\xi} + 1, \tag{38}$$

we have $\sup_\theta \|u^*(\theta)\| \le R_{B_1} - \xi$. Selecting $R_{B_2} \in (R_{B_1} - \xi, R_{B_1})$ then fulfills Assumption 3.3.

With Assumptions 3.1 - 3.4 satisfied, Theorem 1 then implies that there exists a unique θ_∞ such that

$$u^*(\theta_\infty) = \theta_\infty \quad \text{and} \quad \lim_{t \to \infty} \theta_t = \lim_{t \to \infty} u_t = \theta_\infty.$$

We now show what θ_∞ is. We define

$$f(\theta) \doteq \bar{G}(\theta)^{-1}\left(\bar{h}_1(\theta) + \begin{bmatrix} 0 & d_{\mu_\theta}^\top (P_{\pi_\theta} - I)X \\ -X^\top d_{\mu_\theta} & X^\top D_{\mu_\theta} P_{\pi_\theta} X \end{bmatrix}\theta\right).$$

Similar to the proof of Theorem 4 in Section A.5, we can use the Brouwer fixed point theorem to find a $u_\eta^* \in \Gamma_{B_1}$ such that $f(u_\eta^*) = u_\eta^*$ if

$$R_{B_1} \geq C_5 \tag{39}$$

for some constant C_5. Then it is easy to see u_η^* is also the fixed point of $u^*(\cdot)$, implying $\theta_\infty = u_\eta^*$. Rearranging terms of $u_\eta^* = f(u_\eta^*)$ yields

$$(\bar{A}_{\pi_w,\mu_w} + \eta I)w - \bar{b}_{\mu_w} = \mathbf{0},$$

and the desired constants C_0 and C_1 can be deduced from (37), (38), & (39). $\qquad \square$

B. Convergence of Main Networks

This section is a collection of several convergence proofs of main networks. We first state a general convergence result regarding the convergence of time varying linear systems from Konda (2002), which will be repeatedly used.

Consider a stochastic process $\{Y_t\}$ taking values in a finite space $\mathcal{Y}$. Let $\{P_\theta \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|} \mid \theta \in \mathbb{R}^K\}$ be a parameterized family of transition kernels on $\mathcal{Y}$. We update the parameter $w \in \mathbb{R}^K$ recursively as

$$w_{t+1} \doteq w_t + \alpha_t(h_{\theta_t}(Y_t) - G_{\theta_t}(Y_t)w_t),$$

where $h_\theta : \mathcal{Y} \to \mathbb{R}^K$ and $G_\theta : \mathcal{Y} \to \mathbb{R}^{K \times K}$ are two vector- and matrix-valued functions.

Assumption B.1. $\Pr(Y_{t+1} \mid Y_0, w_0, \theta_0, \cdots, Y_t, w_t, \theta_t) = P_{\theta_t}(Y_t, Y_{t+1})$.

Assumption B.2. *The learning rates $\{\alpha_t\}$ is positive deterministic nonincreasing and satisfies*

$$\sum_t \alpha_t = \infty, \sum_t \alpha_t^2 < \infty.$$

Assumption B.3. *The random sequence $\{\theta_t\}$ satisfies*

$$\|\theta_{t+1} - \theta_t\| \leq \beta_t C,$$

where $C > 0$ is a constant, $\{\beta_t\}$ is a deterministic sequence satisfying

$$\sum_t \left(\frac{\beta_t}{\alpha_t}\right)^d < \infty.$$

for some $d > 0$.

Assumption B.4. *For each θ, there exists $\bar{h}(\theta) \in \mathbb{R}^K, \bar{G}(\theta) \in \mathbb{R}^{K \times K}, \hat{h}_\theta : \mathcal{Y} \to \mathbb{R}^K, \hat{G}_\theta : \mathcal{Y} \to \mathbb{R}^{K \times K}$ such that for each $y \in \mathcal{Y}$,*

$$\hat{h}_\theta(y) = h_\theta(y) - \bar{h}(\theta) + \sum_{y'} P_\theta(y, y')\hat{h}_\theta(y'),$$

$$\hat{G}_\theta(y) = G_\theta(y) - \bar{G}(\theta) + \sum_{y'} P_\theta(y, y')\hat{G}_\theta(y').$$

Assumption B.5. $\sup_\theta \|\bar{h}(\theta)\| < \infty, \sup_\theta \|\bar{G}(\theta)\| < \infty$.

Assumption B.6. *For any $f_\theta \in \left\{h_\theta, \hat{h}_\theta, G_\theta, \hat{G}_\theta\right\}, \sup_{\theta,y} \|f_\theta(y)\| < \infty$.*

Assumption B.7. *There exists some constant $C > 0$ such that*

$$\|\bar{h}(\theta) - \bar{h}(\theta')\| \leq C\|\theta - \theta'\|, \|\bar{G}(\theta) - \bar{G}(\theta')\| \leq C\|\theta - \theta'\|.$$

Assumption B.8. *There exists some constant $C > 0$ such that for any $f_\theta \in \left\{h_\theta, \hat{h}_\theta, G_\theta, \hat{G}_\theta\right\}$ and $y \in \mathcal{Y}$,*

$$\|f_\theta(y) - f_{\theta'}(y)\| \le C\|\theta - \theta'\|.$$

Assumption B.9. *There exists a constant $\zeta > 0$ such that for all w, θ,*

$$w^\top \bar{G}(\theta)w \ge \zeta\|w\|^2.$$

Theorem 8. *(Theorem 3.2 in Konda (2002)) Under Assumptions B.1- B.9, almost surely,*

$$\lim_{t\to\infty} \left\|\bar{h}(\theta_t) - \bar{G}(\theta_t)w_t\right\| = 0.$$

We start with the convergence of the main network in Q-evaluation with a Target Network (Algorithm 1).

Lemma 2. *Almost surely, $\lim_{t\to\infty} \|w_t - w^*(\theta_t)\| = 0$.*

Proof. This proof is a special case of the proof of Lemma 4. We, therefore, omit it to avoid verbatim repetition. □

We then show the convergence of the main network in Differential Q-evaluation with a Target Network (Algorithm 2).

Lemma 3. *Almost surely, $\lim_{t\to\infty} \|u_t - u^*(\theta_t)\| = 0$.*

Proof. This proof is the same as the proof of Lemma 2 up to change of notations. We, therefore, omit it to avoid verbatim repetition. □

We now proceed to show the convergence of the main network in Q-learning with a Target Network (Algorithm 3). In Lemma 4 and its proof, we continue using the notations in Theorem 4 and its proof (Section A.5).

Lemma 4. *Almost surely, $\lim_{t\to\infty} \|w_t - w^*(\theta_t)\| = 0$.*

Proof. We proceed by verifying Assumptions B.1- B.9. Let $Y_t \doteq (S_t, A_t, S_{t+1})$ and

$$\mathcal{Y} \doteq \{(s, a, s') \mid s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}, p(s'|s, a) > 0\}.$$

Let $y_1 = (s_1, a_1, s_1'), y_2 = (s_2, a_2, s_2')$, we define $P_\theta \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$ as

$$P_\theta(y_1, y_2) \doteq \mathbb{I}_{s_1' = s_2} \mu_\theta(a_2|s_2)p(s_2'|s_2, a_2). \tag{40}$$

Recall in Section A.5, we define

$$\begin{aligned}
\bar{\theta} &\doteq \Gamma_{B_1}(\theta), \\
h_\theta(s, a, s') &\doteq \left(r(s, a) + \gamma \max_{a'} x(s', a')^\top \bar{\theta}\right)x(s, a), \\
G_\theta(s, a, s') &\doteq x(s, a)x(s, a)^\top + \eta I, \\
\bar{h}(\theta) &\doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[h_\theta(s, a, s')] \\
&= X^\top D_{\mu_\theta} r + \gamma X^\top D_{\mu_\theta} P_{\pi_{\bar{\theta}}} X\bar{\theta}, \\
\bar{G}(\theta) &\doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[G_\theta(s, a, s')] \\
&= X^\top D_{\mu_\theta} X + \eta I.
\end{aligned}$$

According to Algorithm 3, we have

$$\begin{aligned}
\Pr(Y_{t+1}|Y_0, w_0, \theta_0, \dots, Y_t, w_t, \theta_t) &= \Pr(S_{t+1}, A_{t+1}, S_{t+2}|S_t, A_t, S_{t+1}, \theta_t) \\
&= \mu_{\theta_t}(A_{t+1}|S_{t+1})p(S_{t+2}|S_{t+1}, A_{t+1}) \\
&= P_{\theta_t}(Y_t, Y_{t+1}).
\end{aligned}$$

Assumption B.1 therefore holds. Assumptions B.2 & B.3 are satisfied automatically by Assumptions 3.1, 3.5, and Lemma 1.

Consider a Markov Reward Process (MRP) in $\mathcal{Y}$ with transition kernel P_θ and reward function $h_\theta(y)_i$, the i-th element of $h_\theta(y)$ (we need an MRP for each i). The stationary distribution of this MRP is $d_{\mathcal{Y},\theta}(s,a,s') = d_{\mu_\theta}(s,a)p(s'|s,a)$. The reward rate of this MRP is $\sum_y d_{\mathcal{Y},\theta}(y)h_\theta(y)_i = \bar{h}(\theta)_i$, the i-the element of $\bar{h}(\theta)$. We define

$$\hat{h}_\theta(y) \doteq \mathbb{E}[\sum_{k=0}^\infty h_\theta(Y_k) - \bar{h}(\theta)|Y_0 = y, P_\theta].$$

By definition, $\hat{h}_\theta(y)_i$ is the differential value function of this MRP. The differential Bellman equation (1) then implies the first equation in Assumption B.4 holds. Similarly, we define

$$\hat{G}_\theta(y) \doteq \mathbb{E}[\sum_{k=0}^\infty G_\theta(Y_k) - \bar{G}(\theta)|Y_0 = y, P_\theta],$$

the second equation in Assumption B.4 holds as well.

Assumption B.5 follows directly from the definition of $\bar{h}(\theta)$ and $\bar{G}(\theta)$. The boundedness of $h_\theta(y)$ and $G_\theta(y)$ in Assumption B.6 follows directly from their definitions. The differential value function of the MRP can be equivalently expressed as (see, e.g, Eq (8.2.2) in Puterman (2014))

$$\hat{h}_{\theta,i} = (I - P_\theta + \mathbf{1}d_{\mathcal{Y},\theta}^\top)^{-1}(I - \mathbf{1}d_{\mathcal{Y},\theta}^\top)h_{\theta,i},$$
$$\hat{G}_{\theta,ij} = (I - P_\theta + \mathbf{1}d_{\mathcal{Y},\theta}^\top)^{-1}(I - \mathbf{1}d_{\mathcal{Y},\theta}^\top)G_{\theta,ij},$$

where $h_{\theta,i}$ denotes the vector in $\mathbb{R}^{|\mathcal{Y}|}$ whose y-th element is $h_\theta(y)_i$ and $\hat{h}_{\theta,i}, \hat{G}_{\theta,ij}, G_{\theta,ij}$ are similarly defined as $h_{\theta,i}$. Note the existence of the matrix inversion results directly from the ergodicity of the chain (see, e.g., Section A.5 in the appendix of Puterman (2014)). To show the boundedness of $\hat{h}_\theta(y)$ and $\hat{G}_\theta(y)$ required in Assumption B.6, it suffices to show the boundedness of $(I - P_\theta + \mathbf{1}d_{\mathcal{Y},\theta}^\top)^{-1}(I - \mathbf{1}d_{\mathcal{Y},\theta}^\top)$. The intuition is simple: it is a continuous function in a compact set $\mathcal{P}$ (defined in Assumption 5.1), so it obtains its maximum. We now formalize this intuition. We first define a function $f : \mathcal{P} \to \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ that maps a transition matrix in $\mathcal{S}\times\mathcal{A}$ to a transition matrix in $\mathcal{Y}$. For $P \in \mathcal{P}, y_1 \in \mathcal{Y}, y_2 \in \mathcal{Y}$, we have $f(P)(y_1, y_2) \doteq \mathbb{I}_{s_1'=s_2}P((s_1,a_1),(s_2,a_2))p(s_2'|s_2,a_2)/p(s_1'|s_1,a_1)$. Let $\mathcal{P}^\mathcal{Y} \doteq \{f(P) \mid P \in \mathcal{P}\}$. As f is continuous and $\mathcal{P}$ is compact, it is easy to see $\mathcal{P}^\mathcal{Y}$ is also compact. It is easy to verify that $f(P)$ is a valid transition kernel in $\mathcal{Y}$. As Assumption 5.1 asserts P is ergodic, we use d_P to denote the stationary distribution of the chain induced by P. Let $d_{f(P)}(s,a,s') \doteq d_P(s,a)p(s'|s,a)$, it is easy to verify that $d_{f(P)}$ is the stationary distribution of the chain induced by $f(P)$, from which it is easy to see the chain in $\mathcal{Y}$ induced by $f(P)$ is also ergodic, i.e., any $P' \in \mathcal{P}^\mathcal{Y}$ is ergodic. Consider the function $g(P') \doteq (I - P' + \mathbf{1}d_{P'}^\top)^{-1}(I - \mathbf{1}d_{P'}^\top)$. The ergodicity of P' ensures $g(P')$ is well defined. As $\mathcal{P}^\mathcal{Y}$ is compact and g is continuous, g attains its maximum in $\mathcal{P}^\mathcal{Y}$, say, e.g., U_g. For any $\theta \in \mathbb{R}^K$, it is easy to verify that $P_\theta = f(P_{\mu_\theta}) \in \mathcal{P}^\mathcal{Y}$, where P_θ is defined in (40). Consequently, $\|g(P_\theta)\| \le U_g$. The boundedness of $\hat{h}_\theta(y)$ and $\hat{G}_\theta(y)$ then follows directly from the boundedness of $h_\theta(y)$ and $G_\theta(y)$. Assumption B.6 therefore holds.

The Lipschitz continuity of $h_\theta(y), \bar{h}(\theta), G_\theta(y), \bar{G}(\theta)$ in θ follows directly from the Lipschitz continuity of D_{μ_θ} (Lemma 9), the Lipschitz continuity of $P_{\pi_{\bar\theta}}X\bar\theta$ (the proof of Lemma 10), and Lemma 7. We can easily show the boundedness of $(I - P_\theta + \mathbf{1}d_{\mathcal{Y},\theta}^\top)^{-1}$ similar to the boundedness of $(I - P_\theta + \mathbf{1}d_{\mathcal{Y},\theta}^\top)^{-1}(I - \mathbf{1}d_{\mathcal{Y},\theta}^\top)$. The Lipschitz continuity of $\hat{h}_\theta(y)$ and $\hat{G}_\theta(y)$ in θ is then an exercise of Lemmas 7 & 8. Assumptions B.7 & B.8 are then satisfied.

Assumptions B.9 follows directly from the positive definiteness of $\bar{G}(\theta)$.

With Assumptions B.1- B.9 satisfied, invoking Theorem 8 completes the proof. $\qquad\square$

We now show the convergence of the main network in Gradient Q-learning with a Target Network (Algorithm 4). In Lemma 5 and its proof, we continue using the notations in Theorem 5 and its proof (Section A.6).

Lemma 5. *Almost surely,* $\lim_{t\to\infty} \|w_t - w^*(\theta_t)\| = 0$.

Proof. We proceed by verifying Assumptions B.1-B.9. Assumptions B.1-B.8 can be verified in the same way as the proof of Lemma 4. We, therefore, omit it to avoid verbatim repetition. For Assumption B.9, consider $\kappa \doteq [u^\top, w^\top]^\top$, we have

$$\kappa^\top \bar{G}(\theta)\kappa = u^\top C_{\mu_\theta} u + \eta\|w\|^2,$$

where $C_{\mu_\theta} = X^\top D_{\mu_\theta} X$. We first show that there exists a constant $C_1 > 0$ such that $u^\top C_{\mu_\theta} u \geq C_1 \|u\|^2$ holds for any u and θ, which is equivalent to showing for some $C_1 > 0$,

$$u^\top C_{\mu_\theta} u \geq C_1$$

holds for any $u \in \mathcal{U} \doteq \{u \mid \|u\| = 1\}$ and $\theta \in \mathbb{R}^K$. Consider $\mathcal{P}$ defined in 5.1. For any $P \in \mathcal{P}$, we use D_P to denote a diagonal matrix whose diagonal term is the stationary distribution of the chain induced by P. Assumptions 2.2 & 5.1 then ensure that $X^\top D_P X$ is positive definite, i.e.,

$$u^\top X^\top D_P X u > 0$$

holds for any $u \in \mathcal{U}$ and $P \in \mathcal{P}$. As $u^\top X^\top D_P X u$ is a continuous function in both u and P, it obtains its minimum in the compact set $\mathcal{U} \times \mathcal{P}$, say, e.g., $C_1 > 0$, i.e.

$$u^\top X^\top D_P X u \geq C_1$$

holds for any $u \in \mathcal{U}$ and $P \in \mathcal{P}$. It is then easy to see

$$u^\top C_{\mu_\theta} u \geq C_1 \|u\|^2$$

holds for any $u \in \mathbb{R}^K$ and $\theta \in \mathbb{R}^K$, implying

$$\kappa^\top \bar{G}(\theta)\kappa \geq \min\{\eta, C_1\}\|\kappa\|^2.$$

Assumption B.9, therefore, holds. Theorem 8 then implies that

$$\lim_{t \to \infty} \left\| \kappa_t - \bar{G}(\theta_t)^{-1} \bar{h}(\theta_t) \right\| = 0,$$

which completes the proof. $\qquad\square$

We now show the convergence of the main network in Differential Q-learning with a Target Network (Algorithm 5).

Lemma 6. *Almost surely,* $\lim_{t \to \infty} \|u_t - u^*(\theta_t)\| = 0$.

Proof. The proof is the same as the proof of Lemma 4 up to change of notations. We, therefore, omit it to avoid verbatim repetition. $\qquad\square$

C. Technical Lemmas

Lemma 7. *Let $f_1(x)$ and $f_2(x)$ be two Lipschitz continuous functions with Lipschitz constants C_1 and C_2. If they are also bounded by U_1 and U_2, then their product $f_1(x)f_2(x)$ is also Lipschitz continuous with $C_1U_2 + C_2U_1$ being a Lipschitz constant.*

Proof.

$$\|f_1(x)f_2(x) - f_1(y)f_2(y)\| \leq \|f_1(x)(f_2(x) - f_2(y))\| + \|f_2(y)(f_1(x) - f_1(y))\|$$
$$\leq (U_1 C_2 + U_2 C_1)\|x - y\|.$$

$\qquad\square$

Lemma 8. $\left\| Y_1^{-1} - Y_2^{-1} \right\| \leq \left\| Y_1^{-1} \right\| \|Y_1 - Y_2\| \left\| Y_2^{-1} \right\|.$

Proof.

$$\left\| Y_1^{-1} - Y_2^{-1} \right\| = \left\| Y_1^{-1}(Y_1 - Y_2)Y_2^{-1} \right\| \leq \left\| Y_1^{-1} \right\| \|Y_1 - Y_2\| \left\| Y_2^{-1} \right\|.$$

$\qquad\square$

Lemma 9. *Under Assumption 5.1 & 5.2, D_{μ_θ} is Lipschitz continuous in θ.*

Proof. By the definition of stationary distribution, we have $\begin{bmatrix} P_{\mu_\theta}^\top - I \\ \mathbf{1}^\top \end{bmatrix} d_{\mu_\theta} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$. Let $L(P) \doteq \begin{bmatrix} P^\top - I \\ \mathbf{1}^\top \end{bmatrix}$. Assumption 5.1 ensures that for any θ, $L(P_{\mu_\theta})$ has full column rank. So

$$d_{\mu_\theta} = \left(L(P_{\mu_\theta})^\top L(P_{\mu_\theta})\right)^{-1} L(P_{\mu_\theta})^\top \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}.$$

It is easy to see $L(P_{\mu_\theta})$ is Lipschitz continuous in θ and $\sup_\theta L(P_{\mu_\theta}) < \infty$. We now show $\left(L(P_{\mu_\theta})^\top L(P_{\mu_\theta})\right)^{-1}$ is Lipschitz continuous in θ and is uniformly bounded for all θ, invoking Lemma 7 with which will then complete the proof. Note

$$\left(L(P_{\mu_\theta})^\top L(P_{\mu_\theta})\right)^{-1} = \frac{adj\left(L(P_{\mu_\theta})^\top L(P_{\mu_\theta})\right)}{\det\left(L(P_{\mu_\theta})^\top L(P_{\mu_\theta})\right)},$$

where $adj(\cdot)$ denotes the adjugate matrix. By the properties of determinants and adjugate matrices, it is easy to see both the numerator and the denominator are Lipschitz continuous in θ and are uniformly bounded for all θ. It thus remains to show the denominator is bounded away from 0. Consider the compact set $\mathcal{P}$ defined in Assumption 5.1, for any $P \in \mathcal{P}$, Assumption 5.1 ensures $L(P)^\top L(P)$ is invertible, i.e., $|\det\left(L(P)^\top L(P)\right)| > 0$. As $|\det\left(L(P)^\top L(P)\right)|$ is continuous in P, so it obtains its minimum in the compact set $\mathcal{P}$, say $C_0 > 0$. It then follows $\inf_\theta |\det\left(L(P_{\mu_\theta})^\top L(P_{\mu_\theta})\right)| \geq C_0 > 0$, which completes the proof. $\qquad\square$

Lemma 10. *The $w^*(\theta)$ defined in (29) is Lipschitz continuous in θ.*

Proof. Recall

$$w^*(\theta) \doteq \bar{G}(\theta)^{-1}\bar{h}(\theta) = (X^\top D_{\mu_\theta} X + \eta I)^{-1} X^\top D_{\mu_\theta}(r + \gamma P_{\pi_{\bar\theta}} X\bar\theta).$$

We first show $P_{\pi_\theta} X\theta$ is Lipschitz continuous in θ. By definition of π_θ,

$$(P_{\pi_\theta} X\theta)(s,a) = \sum_{s'} p(s'|s,a) \max_{a'} x(s',a')^\top \theta$$
$$= \sum_{s'} p(s'|s,a) \|X_{s'}\theta\|_\infty,$$

where each row of $X_{s'} \in \mathbb{R}^{|\mathcal{A}| \times K}$ is $x(s',a')^\top$. Then

$$|(P_{\pi_{\theta_1}} X\theta_1)(s,a) - (P_{\pi_{\theta_2}} X\theta_2)(s,a)| = |\sum_{s'} p(s'|s,a)(\|X_{s'}\theta_1\|_\infty - \|X_{s'}\theta_2\|_\infty)|$$
$$\leq \sum_{s'} p(s'|s,a)|\|X_{s'}\theta_1\|_\infty - \|X_{s'}\theta_2\|_\infty|$$
$$\leq \sum_{s'} p(s'|s,a)\|X_{s'}\theta_1 - X_{s'}\theta_2\|_\infty \quad \text{(Triangle inequality)}$$
$$\leq \sum_{s'} p(s'|s,a)\|X_{s'}\|_\infty \|\theta_1 - \theta_2\|_\infty$$
$$\leq \|X\|_\infty \|\theta_1 - \theta_2\|_\infty,$$
$$\implies \left\|P_{\pi_{\theta_1}} X\theta_1 - P_{\pi_{\theta_2}} X\theta_2\right\|_\infty \leq \|X\|_\infty \|\theta_1 - \theta_2\|_\infty.$$

The equivalence between norms then asserts that there exists a constant $L_0 > 0$ such that

$$\left\|P_{\pi_{\theta_1}} X\theta_1 - P_{\pi_{\theta_2}} X\theta_2\right\| \leq L_0 \|X\| \|\theta_1 - \theta_2\|.$$

It is easy to see $L_0\|X\|$ is also a Lipschitz constant of $P_{\pi_{\bar\theta}} X\bar\theta$ by the property of projection.

Lemma 9 ensures D_{μ_θ} is Lipschitz continuous in θ and we use L_D to denote the Lipschitz constant.

We can now use Lemmas 7 & 8 to compute the bounds and Lipschitz constants for several terms of interest, which are detailed in Table 1. From Table 1 and Lemma 7, a Lipschitz constant of $w^*(\theta)$ is

$$C_w \doteq \left(\eta^{-1}\|X\|\|r\|L_D + \eta^{-2}\|X\|^3\|r\|L_D\right) + \eta^{-1}\|X\|\gamma L_0\|X\| + \gamma U_P\|X\|R_{B_1}(\eta^{-1}\|X\|L_D + \eta^{-2}\|X\|^3 L_D),$$

which completes the proof.

	Bound	Lipschitz constant
D_{μ_θ}	1	L_D
$(X^\top D_{\mu_\theta} X + \eta I)^{-1}$	η^{-1}	$\eta^{-2}\|X\|^2 L_D$
$X^\top D_{\mu_\theta} r$	$\|X\|\|r\|$	$\|X\|\|r\|L_D$
$(X^\top D_{\mu_\theta} X + \eta I)^{-1} X^\top D_{\mu_\theta} r$	$\eta^{-1}\|X\|\|r\|$	$\eta^{-1}\|X\|\|r\|L_D + \eta^{-2}\|X\|^3\|r\|L_D$
$(X^\top D_{\mu_\theta} X + \eta I)^{-1} X^\top D_{\mu_\theta}$	$\eta^{-1}\|X\|$	$\eta^{-1}\|X\|L_D + \eta^{-2}\|X\|^3 L_D$
$\gamma P_{\pi_{\bar\theta}} X^\top \theta$	$\gamma U_P\|X\|R_{B_1}$	$\gamma L_0\|X\|$

Table 1. $U_P \doteq \sup_\theta \|P_{\pi_\theta}\|$.

$\square$

Lemma 11. *Under Assumptions 2.2, 5.1 - 5.3, if $\|X\| \le 1$, then $w^*(\theta)$ defined in (35) satisfies*

$$\|w^*(\theta_1) - w^*(\theta_2)\| \le \|X\|L_w\|\theta_1 - \theta_2\|,$$

where L_w is a positive constant independent of $\|X\|$.

Proof. Recall

$$w^*(\theta) = (\eta I + A_{\pi_\theta,\mu_\theta}^\top C_{\mu_\theta}^{-1} A_{\pi_\theta,\mu_\theta})^{-1} A_{\pi_\theta,\mu_\theta}^\top C_{\mu_\theta}^{-1} X^\top D_{\mu_\theta} r.$$

Let

$$\tilde{X} \doteq \frac{X}{\|X\|}, \tilde{A}_{\pi_\theta,\mu_\theta} = \tilde{X}^\top D_{\mu_\theta}(I - \gamma P_{\pi_\theta})\tilde{X}, \tilde{C}_{\mu_\theta} = \tilde{X}^\top D_{\mu_\theta} \tilde{X},$$

then

$$w^*(\theta) = \|X\|(\eta I + \|X\|^2 \tilde{A}_{\pi_\theta,\mu_\theta}^\top \tilde{C}_{\mu_\theta}^{-1} \tilde{A}_{\pi_\theta,\mu_\theta})^{-1} \tilde{A}_{\pi_\theta,\mu_\theta}^\top \tilde{C}_{\mu_\theta}^{-1} \tilde{X} D_{\mu_\theta} r.$$

We now show $w^*(\theta)/\|X\|$ is Lipschitz continuous in θ by invoking Lemmas 7 & 8. Let $D_P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ be a diagonal matrix whose diagonal entry is the stationary distribution of the chain induced by P. For any $P \in \mathcal{P}$, Assumption 5.1 ensures that D_P is positive definite. Consequently, $\left\|(\tilde{X}^\top D_P \tilde{X})^{-1}\right\|$ is well defined and is continuous in P, implying it obtains its maximum in the compact set $\mathcal{P}$, say U_g. So $\left\|\tilde{C}_{\mu_\theta}^{-1}\right\| \le U_g$ and importantly, U_g does not depend on $\|X\|$ in the sense that scaling X with any nonzero scalar does not affect U_g. Using Lemma 7, it is easy to see the bound and the Lipschitz constant of $\tilde{A}_{\pi_\theta,\mu_\theta}^\top \tilde{C}_{\mu_\theta}^{-1} \tilde{A}_{\pi_\theta,\mu_\theta}$ do not depend on $\|X\|$. It is easy to see $\left\|(\eta I + \|X\|^2 \tilde{A}_{\pi_\theta,\mu_\theta}^\top \tilde{C}_{\mu_\theta}^{-1} \tilde{A}_{\pi_\theta,\mu_\theta})^{-1}\right\| \le 1/\eta$. If we further assuming $\|X\| \le 1$, Lemma 8 then implies that $(\eta I + \|X\|^2 \tilde{A}_{\pi_\theta,\mu_\theta}^\top \tilde{C}_{\mu_\theta}^{-1} \tilde{A}_{\pi_\theta,\mu_\theta})^{-1}$ has a Lipschitz constant independent of $\|X\|$. It is then easy to see there exists a constant $L_w > 0$, independent of $\|X\|$, such that

$$\|w^*(\theta_1) - w^*(\theta_2)\| \le L_w\|X\|\|\theta_1 - \theta_2\|,$$

which completes the proof. $\square$

Lemma 12. *The $u^*(\theta)$ defined in (36) is Lipschitz continuous in θ.*

Proof. Recall

$$\bar{h}(\theta) \doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[h_\theta(s,a,s')] = \bar{h}_1(\theta) + \bar{H}_2(\theta),$$

$$\bar{h}_1(\theta) \doteq \begin{bmatrix} d_{\mu_\theta}^\top r \\ X^\top D_{\mu_\theta} r \end{bmatrix}, \bar{H}_2(\theta) \doteq \begin{bmatrix} 0 & d_{\mu_\theta}^\top (P_{\pi_{\bar{\theta}^w}} - I) X \bar{\theta}^w \\ -(X^\top d_{\mu_\theta}) \bar{\theta}^r & X^\top D_{\mu_\theta} P_{\pi_{\bar{\theta}^w}} X \bar{\theta}^w \end{bmatrix},$$

$$\bar{G}(\theta) \doteq \mathbb{E}_{(s,a)\sim d_{\mu_\theta}, s'\sim p(\cdot|s,a)}[G_\theta(s,a,s')] = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & X^\top D_{\mu_\theta} X + \eta I \end{bmatrix},$$

$$u^*(\theta) \doteq \bar{G}(\theta)^{-1}\bar{h}(\theta).$$

We now use Lemmas 7 & 8 to compute the bounds and Lipschitz constants for several terms of interest, which are detailed in Table 2. From Table 2 and Lemma 7, a Lipschitz constant of $u^*(\theta)$ is

$$C_u = \max\{1, \eta^{-1}\}(\mathcal{O}(\|X\|) + \mathcal{O}(L_\mu)) + \max\{1, \eta^{-2}\}\mathcal{O}(\|X\|)$$

which completes the proof.

	Bound	Lipschitz constant
D_{μ_θ}	1	L_D
$\bar{G}(\theta)^{-1}$	$\max\{1, \eta^{-1}\}$	$\max\{1, \eta^{-2}\}\mathcal{O}(\|X\|^2)$
$\bar{h}_1(\theta)$	$\mathcal{O}(\|X\|) + \mathcal{O}(1)$	$\mathcal{O}(L_\mu)$
$\bar{H}_2(\theta)$	$\mathcal{O}(\|X\|)$	$\mathcal{O}(\|X\|)$
$\bar{h}(\theta)$	$\mathcal{O}(\|X\|) + \mathcal{O}(1)$	$\mathcal{O}(\|X\|) + \mathcal{O}(L_\mu)$

Table 2. Bounds and Lipschitz constants of several terms, assuming $\|X\| \leq 1, L_\mu \leq 1$.

$\square$

D. Experiment Details

D.1. Kolter's Example

Kolter's example is a simple two-state Markov chain with $P_\pi \doteq \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$. The reward is set in a way such that the state-value function is $v_\pi = \begin{bmatrix} 1 \\ 1.05 \end{bmatrix}$. The feature matrix is $X \doteq \begin{bmatrix} 1 \\ 1.05 + \epsilon \end{bmatrix}$. Kolter (2011) shows that for any $\epsilon > 0, C > 0$, there exists a $D_\mu = diag(\begin{bmatrix} d_\mu(s_1) \\ d_\mu(s_2) \end{bmatrix})$ such that

$$\|\Pi_{D_\mu} v_\pi - v_\pi\| \leq \epsilon \quad \text{and} \quad \|X w_0^* - v_\pi\| \geq C,$$

where w_0^* is the TD fixed point. This suggests that as long as there is representation error (i.e., $\epsilon > 0$), the performance of the TD fixed point can be arbitrarily poor. In our experiments, we set $\epsilon = 0.01$ and find when $d_\mu(s_1)$ approaches around 0.71, $\|X w_0^* - v_\pi\|$ becomes unbounded.

D.2. Baird's Example

Figure 2 shows Baird's example. In the policy evaluation setting (corresponding to Figure 1b), we use the same state features as Sutton & Barto (2018), i.e.,

$$X \doteq \begin{bmatrix} 2I & \mathbf{0} & \mathbf{1} \\ \mathbf{0}^\top & 1 & 2 \end{bmatrix} \in \mathbb{R}^{7\times 8}.$$

The weight w is initialized as $[1,1,1,1,1,1,10,1]^\top$ as suggested by Sutton & Barto (2018). The behavior policy is $\mu_0(\text{dashed}|s_i) = 1/7$ and $\mu_0(\text{solid}|s_i) = 6/7$ for $i = 1, \ldots, 7$. And the target policy always selects the solid action. The standard Off-Policy Linear TD (with ridge regularization) updates w_t as

$$w_{t+1} \leftarrow w_t + \alpha \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}(R_{t+1} + x_{t+1}^\top w_t - x_t^\top w_t)x_t - \alpha\eta w_t,$$
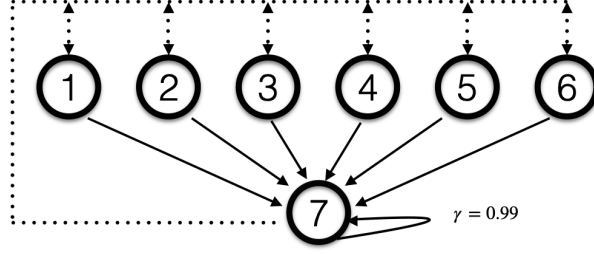
Figure 2. The Baird's example used in Chapter 11.2 of Sutton & Barto (2018). The figure is taken from Zhang et al. (2020). At each state, there are two actions. The solid action always leads to s_7; the dashed action leads to one of $\{s_1, \dots, s_6\}$ with equal probability. The discount factor γ is 0.99.

where we use $\alpha = 0.01$ as used by Sutton & Barto (2018). Note as long as $\alpha > 0$, it diverges. Here we overload x to denote the state feature and $x_t \doteq x(S_t)$. We use a TD version of Algorithm 1 to estimate v_π, which updates w_t as

$$w_{t+1} \leftarrow w_t + \alpha \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}(R_{t+1} + x_{t+1}^\top \theta_t - x_t^\top w_t)x_t - \alpha\eta w_t,$$

$$\theta_{t+1} \leftarrow \theta_t + \beta(w_t - \theta_t),$$

where we set $\alpha = \beta = 0.01, \theta_0 = w_0$.

For the control setting (corresponding to Figures 1c & 1d), we construct the state-action feature in the same way as the Errata of Baird (1995), i.e.,

$$X \doteq \begin{bmatrix} \begin{bmatrix} 2I & \mathbf{0} & \mathbf{1} \\ \mathbf{0}^\top & 1 & 2 \\ & \mathbf{0}^\top \mathbf{0} & \end{bmatrix} & \mathbf{0}^\top \mathbf{0} \\ & I \end{bmatrix} \in \mathbb{R}^{14 \times 15}.$$

The first 7 rows of X are the features of the solid action; the second 7 rows are the dashed action. The weight w is initialized as $[1, 1, 1, 1, 1, 1, 10, 1, 1, 1, 1, 1, 1, 1, 1]^\top$. The standard linear Q-learning (with ridge regularization) updates w_t as

$$w_{t+1} \leftarrow w_t + \alpha(R_{t+1} + \max_{a'} x(S_{t+1}, a)^\top w_t - x_t^\top w_t)x_t - \alpha\eta w_t,$$

where we set $\alpha = 0.01$. The variant of Algorithm 3 we use in this experiment updates w_t as

$$w_{t+1} \leftarrow w_t + \alpha(R_{t+1} + \max_{a'} x(S_{t+1}, a)^\top \theta_t - x_t^\top w_t)x_t - \alpha\eta w_t,$$

$$\theta_{t+1} \leftarrow \theta_t + \beta(w_t - \theta_t),$$

where we set $\alpha = 0.01, \beta = 0.001, \theta_0 = w_0$. The behavior policy for Figure 1c is still μ_0; the behavior policy for Figure 1d is $0.9\mu_0 + 0.1\mu_w$, where μ_w is a softmax policy w.r.t. $x(s, \cdot)^\top w$. For our algorithm with target network, the softmax policy is computed using the target network as shown in Algorithm 3.