Target Networks and Over-parameterization Stabilize Off-policy Bootstrapping with Function Approximation

Fengdi Che 1  Chenjun Xiao 2  Jincheng Mei 3  Bo Dai 3 4  Ramki Gummadi 3  Oscar A Ramirez 5 6  Christopher K Harris 6 7  A. Rupam Mahmood 1 8  Dale Schuurmans 1 3 8

Abstract

We prove that the combination of a target network and over-parameterized linear function approximation establishes a weaker convergence condition for bootstrapped value estimation in certain cases, even with off-policy data. Our condition is naturally satisfied for expected updates over the entire state-action space or learning with a batch of complete trajectories from episodic Markov decision processes. Notably, using only a target network or an over-parameterized model does not provide such a convergence guarantee. Additionally, we extend our results to learning with truncated trajectories, showing that convergence is achievable for all tasks with minor modifications, akin to value truncation for the final states in trajectories. Our primary result focuses on temporal difference estimation for prediction, providing high-probability value estimation error bounds and empirical analysis on Baird’s counterexample and a Four-room task. Furthermore, we explore the control setting, demonstrating that similar convergence conditions apply to Q-learning.

1. Introduction

Off-policy value evaluation with offline data considers the challenge of evaluating the expected discounted cumulative reward of a given policy using a dataset that was not necessarily collected according to the target policy. Such a scenario is common in real-world applications, such as self-driving vehicles and healthcare, where active data collection with unqualified policy can pose life-threatening risks (Levine et al. 2020). The ability to perform off-policy evaluation can also enhance data efficiency, for example, through techniques like experience replay (Lin 1992). Typically, estimation from offline data does not enforce constraints on the collection procedure, which allows for the inclusion of diverse sources, such as driving data from multiple drivers or online text from alternative platforms. A key challenge, however, is that offline data might only cover part of the state space, which leads to technical difficulties that are not encountered when learning from continual online data.

Temporal difference (TD) estimation (Sutton & Barto 2018), where value estimates are formed by bootstrapping from the Bellman equation, has emerged as one of the most widely deployed value estimation techniques. Despite its popularity, however, the deadly triad can thwart TD algorithms in offline learning. It is well known that the combination of off-policy data, function approximation, and bootstrapping can cause the divergence of such algorithms; in the under-parameterized setting, a TD fixed point might not even exist (Tsitsiklis & Van Roy 1996). To address this issue, L2-regularization is often introduced to ensure the existence of a fixed point, and several algorithms, such as LSTD (Yu & Bertsekas 2009) and TD with a target parameter (Zhang et al. 2021), have been shown to converge to a regularized fixed point. However, this regularized point can result in larger estimation errors than simply using a zero initialization of the parameters (Manek & Kolter 2022).

Over-parameterization is another approach for ensuring the existence of a TD fixed point (Xiao et al. 2021, Thomas 2022). In the offline setting, an over-parameterized model is defined to be one that has more parameters than the number of distinct data points in the dataset. Unfortunately, even over-parameterized TD still suffers from the deadly triad; in Figure 1 below we observe that over-parameterized linear TD still diverges on Baird’s counterexample (Baird 1995).

Our paper establishes the convergence condition for TD with offline data when both a target network and an over-parameterized linear model are incorporated, as shown in Section 3. Our condition is naturally satisfied for expected updates using the entire state-action space, ensuring guaran-
ted convergence. Empirically, we demonstrate on Baird’s counterexample that the over-parameterized target TD converges faster than other existing solutions to the deadly triad, such as residual minimization (RM) or gradient TD methods, while using less memory than convergent methods like LSTD. Our result provides theoretical support for the empirical success of target networks and represents the first demonstration of a practical algorithm that is provably convergent and capable of high-quality empirical performance.

Importantly, over-parameterization ensures that the fixed point remains independent of the state collection distribution. Therefore, state distribution correction is not needed to approximate the value function of a target policy, which grants flexibility in collecting offline data from multiple state distributions. This property also means that the high variance and bias associated with state distribution corrections (Liu et al. 2018) can be avoided, which has been a longstanding concern in the field. We show that the resulting fixed point, given full state coverage and accurate dynamics, approximates the target value function with an error that can be bounded by the distance between the best linearly represented value estimate and the true Q-value, \( \frac{2}{1-\gamma} \inf_{\theta} \| \Phi \theta - q^* \|_\infty \), similar to the under-parameterized on-policy TD fixed point.

Additionally, our convergence condition holds for learning with a batch of trajectory data collected from episodic MDPs. This result can be extended to truncated trajectory data with minor modifications, as explained in Section 4. To compute the temporal difference error under the target policy, we consider two viable approaches: the first involves sampling the next action from the target policy, as detailed in Section 3, while the second, introduced in Section 4, adopts per-step normalized importance sampling (NIS) correction of action choices (not the state distribution) (Hesterberg 1995). The latter method offers the advantage of converging with trajectory data under behaviour policies without making assumptions about the task or dataset. Consequently, we assert that the deadly triad issue can be fully resolved through the introduction of over-parameterized target TD with NIS correction over trajectory data. The value estimation errors of these two approaches are compared empirically in a simple Four Room task in Section 4.

Finally, we extend the results to the offline control case. Q-learning (Watkins & Dayan 1992) is a control algorithm based on temporal difference learning that also suffers from the deadly triad. Here we show that over-parameterized target Q-learning with offline data is also provably convergent.

2. Background

Notation. We let \( \Delta(\mathcal{X}) \) denote the set of probability distributions over a finite set \( \mathcal{X} \). Let \( \mathbb{R} \) denote the set of real numbers, and \( \mathbb{I} \) be the indicator function. For a matrix \( A \in \mathbb{R}^{n \times m} \), we let \( A^\dagger \) denote its Moore-Penrose pseudoinverse and \( \rho(A) \) denote its spectral radius. Finally, we let \( \text{diag}(x) \in \mathbb{R}^{d \times d} \) be a diagonal matrix whose diagonal elements are given by \( x \in \mathbb{R}^d \).

Markov Decision Process. We consider finite Markov Decision Process (MDP) defined by \( \mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\} \), where \( \mathcal{S} \) is a finite state space, \( \mathcal{A} \) is the action space, \( r : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \) is the reward function bounded by one, \( P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) \) is the transition matrix, and \( \gamma < 1 \) is the discount factor. The Q-value represents the expected cumulative rewards starting from a state-action pair \((s, a)\) following a policy \( \pi : \mathcal{S} \to \Delta(\mathcal{A}) \), defined as

\[
q_\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \mid S_0 = s, A_0 = a \right],
\]

where we use \( \mathbb{E}_\pi \) to denote the expectation under the distribution induced by \( \pi \) and the environment. The Bellman operator under the policy \( \pi \) on \( q(s, a) \) is defined as

\[
T_\pi q(s, a) = r(s, a) + \gamma \sum_{s', a' \in \mathcal{S} \times \mathcal{A}} P_\pi(s', a'|s, a) q(s', a'),
\]

with the state-action transition distribution under \( \pi \) defined as \( P_\pi(s', a'|s, a) = P(s'|s, a) \pi(a'|s) \). We represent functions as vectors to enable vector-space operations: the value function and reward function are denoted by \( q_\pi, r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \), and the transition function by \( P_\pi \in \mathbb{R}^{\mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A}} \). Then, we define the Bellman operator on any vector \( q \) as

\[
T_\pi q = r + \gamma P_\pi q.
\]

It is known that the value function satisfies the Bellman equation \( q_\pi = T_\pi q_\pi \).

Linear Function Approximation. In this work, we focus on linear function approximation, \( q_\phi(s, a) \approx \phi(s, a)^\top \theta \), where \( \theta \in \mathbb{R}^d \) is a parameter vector, and \( \phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d \) maps a given state-action pair to a \( d \)-dimensional feature vector \( \phi(s, a) \in \mathbb{R}^d \). We denote \( \Phi \in \mathbb{R}^{\mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A} \times \mathcal{A}} \) as the feature matrix, where each row corresponds to the feature vector of a particular state-action pair \((s, a)\). This matrix form allows us to write the value function approximation as \( q_\phi \approx \Phi \theta \) for some parameter \( \theta \). Finally, we assume that \( \Phi \) is full rank, meaning there are no redundant features.

Offline Value Prediction. We consider offline value prediction, that is, learning to predict the value of a target policy given a prior collected dataset \( D \) that consists of transition data \( \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n \). Let \( \lambda \in \Delta(\mathcal{S} \times \mathcal{A}) \) be an arbitrary data collection distribution; the transition data is collected by first sampling a state-action pair \((s_i, a_i)\) from \( \lambda \), then receiving the reward \( r_i = r(s_i, a_i) \) and next state \( s_i \sim P_\pi(s_i | s_i, a_i) \) from the environment. The problem is known as on-policy if the data collection distribution \( \lambda \) is
the stationary distribution of $\pi$ and off-policy, otherwise (Sutton & Barto 2018). Our results can also be extended for offline policy optimization, aiming to extract a good control policy from the offline data.

For clarity, we introduce the following additional notation. Given offline data $D$, let $n(s, a) = \sum_{i=1}^{n} I[s_i = s, a_i = a]$ be the count of a state-action $(s, a)$ observed in the data, and $\hat{\lambda}(s, a) = n(s, a)/n$ be the empirical distribution of $(s, a)$. We let $\{(s_i, a_i)\}_{i=1}^{k} \subseteq S \times A$ denote the state-actions with $n(s_i, a_i) > 0$, where $k = \sum_{s,a} I[n(s,a) > 0]$ represents the count of state-action pairs with observed outgoing transitions. Following these definitions, we can define a mask matrix $H \in \mathbb{R}^{k \times |S||A|}$, and the empirical distribution of observed data $D_k = \mathbb{R}^{k \times k}$ as:

$$H = \begin{bmatrix} 1^\top_{x_1} \\ \vdots \\ 1^\top_{x_k} \end{bmatrix}, \quad D_k = \begin{bmatrix} \hat{\lambda}(x_1) \\ \vdots \\ \hat{\lambda}(x_k) \end{bmatrix}, \quad (1)$$

where $1_{(s, a)} \in \{0, 1\}^{S \times A}$ is an indicator vector such that $\phi(s, a) = 1_{(s, a)}^\top \Phi$.

To evaluate the value of a target policy $\pi$, we augment each transition $(s_i, a_i, r_i, s'_i)$ to $(s_i, a_i, r_i, s'_i, a'_i)$ by selecting an action $a'_i \sim \pi(\cdot|s'_i)$. The empirical transition matrix between state-action pairs $\hat{P}_{\pi} \in \mathbb{R}^{||A|| \times |S||A|}$ can then be defined for all $s', a'$ as:

$$\hat{P}_\pi(s', a'|s, a) = \frac{\sum_{i=0}^{n} I[s_i = s, a_i = a, s'_i = s', a'_i = a']}{n(s, a)}, \quad (2)$$

if $n(s, a) > 0$; otherwise, $\hat{P}_\pi(s', a'|s, a) = 0$.

Given these notations, we can then define the empirical mean squared Bellman error (EMSBE) as:

$$\text{EMSBE}(\theta) = \frac{1}{2} \| R + \gamma N(\theta) - M\theta \|^2_{D_k}, \quad (3)$$

where $M = H\Phi \in \mathbb{R}^{k \times d}$ denotes the predecessor features observed in the offline data, $N = H\bar{P}_\pi\Phi \in \mathbb{R}^{k \times d}$ gives the next state-action features under the empirical transition, and $R = Hr \in \mathbb{R}^{k}$ gives the rewards of the observed state-action pairs.

**Over-parameterization** This work considers the over-parameterization setting, such that the function approximation applies linear features with dimension $d > k$, the support of the empirical data. This allows all of the Bellman consistency constraints to be satisfied on all transitions in the offline data, driving EMSBE exactly to zero.

3. Over-parameterized Target TD

We first show that leveraging overparameterization and target network can significantly stabilize temporal difference learning with function approximation. We will also use the Baird counterexample to illustrate the effectiveness of over-parameterized target TD (Baird 1995).

3.1. Over-parameterized TD Learning

First, we briefly review the over-parameterized TD (OTD) algorithm for offline value prediction. OTD applies semi-gradients of EMSBE (3) to update the parameter recursively,

$$\theta_{t+1} = \theta_{t} - \eta M^\top D_k [M\theta_{t} - (R + \gamma N\theta_{t})]. \quad (4)$$

where $\eta > 0$ is the learning rate. Xiao et al. (2021) analyzed the convergence properties of OTD. Unfortunately, they neglected a necessary condition that we correct below.

**Proposition 3.1.** For the over-parameterized regime $d > k$, if the following two conditions hold:

- $\rho(I - \eta M^\top D_k (M - \gamma N)) < 1$;
- $NM^\dagger$ has any sub-multiplicative norm smaller than or equal to one,

then there exists a learning rate $\eta$ such that the parameter of OTD updates converges to

$$\theta^*_\text{TD} = M^\dagger (I - \gamma N M^\dagger)^{-1} R,$$

when the initial parameter of OTD equals zero.

It can be verified that the OTD fixed point $\theta^*_\text{TD}$ is the minimum norm solution of EMSBE that lies in the span of $M$. Importantly, Proposition 3.1 characterizes an implicit algorithmic bias of OTD: it implicitly regularizes the solution toward a unique fixed point, even when EMSBE admits infinitely many global minima. However, the convergence of OTD requires certain constraints on the features to guarantee convergence. In particular, it requires the matrix $I - \eta M^\top D_k (M - \gamma N)$ to have a spectral radius of less than one, which cannot be easily met on problems, including the Baird counterexample, and thus causes divergence, as we will show later. The failure of this condition is identified as a core factor behind the deadly triad issue, causing the update to be non-contractive (Sutton et al. 2016, Fellows et al. 2023). Furthermore, a sufficient condition would be to have orthonormal feature vectors and states showing up uniformly to satisfy the spectral radius property, which leaves no generalization space and creates an awkward tradeoff between instability and generalization. We did not find other sufficient conditions that can be easily checked.

3.2. Over-parameterized Target TD

Our first main contribution is to confirm theoretically that leveraging a target network significantly increases TD’s stability with function approximation. Let $\theta_{\text{target}} \in \mathbb{R}^{d}$ be the target parameter at iteration $t$. We refer to the function
Figure 1. On Baird counterexample, states are sampled from a uniform distribution, and there exists only one action at each state. The discount factor is set to be $\gamma = 0.95$. We plot the maximal value prediction error among all states for OTD, OTTD, RM and GTD-algorithms. Other than OTD, the value errors converge to zero for the rest algorithms. OTTD avoids the divergence of TD and slow convergence rate of others.

Theorem 3.2. For the over-parameterized regime $\theta$.

That is, a target parameter $\theta$ is introduced to provide bootstrapping targets for TD updates. The target parameter is initialized with the student parameter $\theta_{\text{targ},0} = \theta_0$, and is kept fixed for a window size $m$. Then, for every $m$ step, we update the target parameter by directly copying from the student parameter,

$\theta_{\text{targ},(n+1)m} = \theta_{nm}$.

Our next result characterizes the convergence of OTTD.

**Theorem 3.2.** For the over-parameterized regime $d > k$, given that the following condition holds:

- $NM^1$ has any sub-multiplicative norm smaller than or equal to one,

there exists a learning rate $\eta$ and an integer $\bar{m}$ such that for all update window sizes of the target parameters $m \geq \bar{m}$, the parameter of OTTD converges to

$\theta_{\text{TD}} = M^1(I - \gamma NM^1)\bar{m}^{-1}R$.

Remark 3.3. The dependence on the initial point is detailed in Theorem A.4 in A.1. The analytical form of $\bar{m}$ depends on the specific norm constraint applied to $NM^1$. For instance, when bounding the infinity norm, $\bar{m} = 1 + \lceil \frac{\log(1 - \gamma) - \log((1 + \gamma)^{-1})}{\log(1 - \eta \lambda_{\min}(MM^1D_k))} \rceil$.

Theorem 3.2 illustrates the efficacy of incorporating a target network in stabilizing bootstrapping with function approximation. This is apparent in the convergence of OTTD to the TD fixed point, eliminating the condition on the spectral radius of $I - \eta M^1D_k(M - \gamma N)$ to be bounded by one. Central to the convergence of OTTD is the role of $NM^1$, which represents the projection coefficient of each row of $N$ onto the row space of $M$. When applying bootstrapping, values are extrapolated using $NM^1M\theta$, based on these projected coefficients. When the infinity norm of $NM^1$ is bounded by one, we have $\|NM^1M\theta\|_\infty \leq \|M\theta\|_\infty$. Thus, the condition on the norm of $NM^1$ prevents overestimation for bootstrapping values outside the current state-action set. Although OTTD still requires this condition for convergence, as we will show later in the paper, it can be resolved when the offline dataset consists of trajectory data.

3.3. Special Case Analysis of Expected Updates

The benefit of incorporating a target network is fully presented when considering expected updates with an off-policy data distribution that covers the entire state-action
Figure 2. In this example, each state has exactly one action, and rewards are labeled next to the transitions. The value functions are parameterized by a scalar parameter $\theta$, and the features are shown in the graph. This counterexample demonstrates a task with fixed transition probabilities and rewards where our convergence condition is satisfied. However, the conditions for under-parameterized target TD fail for certain data distributions.

Off-policy learning means that we do not constrain the data distribution $\lambda$ to be the stationary distribution under a target policy. In this context, OTTD naturally converges with a proven guarantee. In contrast, using only one augmentation—either a target network or an over-parameterized model—can still result in divergence.

The expected update rules for TD and TD with a target network (target TD) are the same for under-parameterized or over-parameterized models and are outlined below:

- **TD expected update:**
  \[
  \theta_{t+1} = \theta_t - \eta \Phi^\top D [\Phi \theta_t - (R + \gamma P \Phi \theta_t)].
  \]

- **Target TD expected update:**
  \[
  \theta_{t+1} = \theta_t - \eta \Phi^\top D \left[ \Phi \theta_t - (R + \gamma P \Phi \theta_{\text{target}, m}[\pi]) \right].
  \]

Several works (Asadi et al. 2023, Fellows et al. 2023) have analyzed the convergence conditions for under-parameterized target TD. However, their condition, which requires $(\Phi^\top D \Phi)^{-1} \Phi^\top D \gamma P \Phi$ to have a spectral radius or norm of less than one as listed in A.8, cannot be met for all off-policy data distributions. This limitation is demonstrated by a Two-state counterexample in Figure 2, where certain data distributions result in the fixed point of under-parameterized TD not existing, causing the convergence condition to fail.

In the counterexample, the feature matrix of these two states equals to $\Phi = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. The learning is off-policy if the state distribution differs from the stationary distribution, which converges on the right state with self-loop. As for a discount factor $\gamma > 0.5$, the off-policy state distribution $(\frac{4\gamma - 4}{2\gamma - 3}, \frac{2\gamma - 3}{2\gamma - 3})$ causes the fixed point to be $(\Phi^\top D(I - \gamma P)\Phi)^{-1} \Phi^\top D R = \frac{0}{n}$, which does not exist. Consequently, the learning of TD, with or without a target network, is stuck at any initialization. This breaks the required condition for the convergence proof in Corollary 2 of Asadi et al. (2023) and the non-asymptotic behavior analysis in Corollary 3.1 of Fellows et al. (2023).

Adding over-parameterization alone does not ensure convergence either. As prove in Proposition 3.1, the convergence of OTD still requires that $\rho(I - \eta \Phi^\top D(\Phi - \gamma P \Phi)) < 1$, which can be easily violated. Baird’s counterexample (Baird 1995) illustrates this expected update with all states observed uniformly. Since each state has only one action, actions must adhere to the target policy, as requested by the data collection procedure in Section 2. Yet, states are sampled from an off-stationary distribution, giving the challenges of off-policy learning. The features are intentionally over-parameterized, having more dimensions than the total number of states. Hyperparameters’ choices are in A.7. As shown in Figure 1, OTD diverges, while OTTD successfully prevents the divergence observed in standard TD.

In summary, over-parameterization fundamentally eliminates the model’s dependency on the data generation distribution and a target network removes a core factor behind the deadly triad issue, causing the update to be non-contractive (Sutton et al. 2016, Fellows et al. 2023). Therefore, the combination of these two simple augmentations resolves the deadly triad, as stated in Proposition 3.4.

**Proposition 3.4.** For the over-parameterized regime $d > k$, there always exists a learning rate $\eta$ and an integer $\bar{m}$ such that for all update window sizes of the target parameters $m \geq \bar{m}$, the parameter of OTTD converges to

\[
\theta^*_{\text{TD}} = \Phi(I - \gamma P)^{-1} R + (I - \Phi^\top \Phi)\theta_0,
\]

where $\theta_0$ is the initial parameter of OTTD.

We also compare OTTD with other standard algorithms empirically on the Baird’s counterexample, including Residual Minimization (RM), Baird RM (Baird 1995), Gradient TD2 (GTD2) (Maei 2011), and TDC (Sutton et al. 2009). Figure 1 indicates that OTTD also mitigates the slow convergence rates of alternative algorithms. The update rules of these algorithms can be expressed through the equation

\[
\theta_{t+1} - \theta^* = C^t(\theta_0 - \theta^*). \]

To gauge the convergence speed of these algorithms, Schoknecht and Merke (2003) introduced a metric defined as $\lim_{t \to \infty} ||C_t||^\top$. A higher value of this metric indicates a slower convergence rate. The convergence metrics, presented in Table 3.3, reveal that the metric exceeds one for TD, leading to divergence. On the other hand, the metrics for RM and GTD2 hover too close to one, resulting in slower convergence rates.

**Table 1.** The table shows a metric for the convergence rate. A higher value of this metric indicates a slower convergence rate, and a larger-than-one value represents the divergence of the algorithm.

| Algorithm | Convergence? | $\lim_{t \to \infty} ||C_t||^\top$ |
|-----------|-------------|-------------------------------|
| TD        | No          | 1.12                          |
| Target TD | Yes         | $1 - 3.8e^{-3}$               |
| RM        | Yes         | $1 - 1.9e^{-5}$               |
| GTD2      | Yes         | $1 - 4.5e^{-6}$               |

**Table 2.** The table shows a metric for the convergence rate. A higher value of this metric indicates a slower convergence rate, and a larger-than-one value represents the divergence of the algorithm.
3.4. Value Prediction Error Bound

We present a worst-case value prediction error bound of OTTD across the entire state-action space.

**Theorem 3.5.** Given \( \|NM^1\|_\infty < 1 \), with probability at least \( 1 - \delta \), the fixed point of OTTD \( \theta^*_TD \) learnt with the dataset \( \mathcal{D} \) gives the following value prediction error bound

\[
\|\Phi\theta^*_TD - q_\pi\|_\infty \leq \epsilon_{\text{stat}} + \epsilon_{\text{projection}} + \epsilon_{\text{approx}},
\]

where error terms are defined as follows:

- **Statistical error** \( \epsilon_{\text{stat}} \) equals

\[
\epsilon_{\text{stat}} = \frac{\|\Phi M^1\|_\infty}{(1 - \gamma)^2} \sqrt{\log\left(\frac{2k|A|}{\delta}\right) \min(s,a) n(s,a)}.
\]

- **Projection error** \( \epsilon_{\text{projection}} \) is defined as

\[
\epsilon_{\text{projection}} = \frac{\|\Phi M^1\|_\infty}{1 - \gamma} \|\Phi(I - M^1M)\theta^*\|_\infty.
\]

- **Approximation error** \( \epsilon_{\text{approx}} \) is given by

\[
\epsilon_{\text{approx}} = \frac{2}{1 - \gamma} \|\Phi M^1\|_\infty \|\Phi\theta^* - q_\pi\|_\infty.
\]

Here, \( \theta^* = \arg\max_\theta \|\Phi\theta - q_\pi\|_\infty \) is the optimal parameter minimizing the difference between \( \Phi\theta \) and \( q_\pi \), \( \epsilon_{\text{stat}} \) counts in the estimation error of the MDP dynamics by the dataset, and \( \epsilon_{\text{projection}} \) accounts for insufficient coverage of the dataset. Notably, in overparameterization, information perpendicular to the row space of \( M \), the space of data features, decides the size of errors instead of distribution shift ratios. The proof is given in A.2.

Next, we narrow our focus on the expected update discussed in Section 3.3. This scenario, characterized by \( M = \Phi \) and \( N = P_\gamma\Phi \), remains off-policy. The fixed point of OTTD \( \theta^*_TD = \Phi(I - \gamma P_\gamma)^{-1}R \), gives the following error bound.

**Corollary 3.6.** For the optimal fixed point \( \theta^*_TD = \Phi(I - \gamma P_\gamma)^{-1}R \), the approximation error is bounded as

\[
\|\Phi\theta^*_TD - q_\pi\|_\infty \leq \frac{2}{1 - \gamma} \inf_\theta \|\Phi\theta - q_\pi\|_\infty.
\]

The value prediction error bound of off-policy OTTD closely aligns with the on-policy results obtained using underparameterized linear models. This significance becomes evident when considering the stringent requirement in on-policy learning, where data must be sampled from the stationary distribution \( d_\pi \) of the target policy \( \pi \). This key observation underscores the remarkable ability of overparameterized models to learn irrespective of the underlying data distribution \( \lambda \). To elaborate, Tsitsiklis and Van Roy (1997) have shown that on-policy TD fixed point \( \theta^*_{\text{under}} \) in the underparameterized setting satisfies that

\[
\|\Phi\theta^*_{\text{under}} - q_\pi\|_{D_{\pi}} \leq \frac{1}{1 - \gamma} \inf_\theta \|\Phi\theta^*_{\text{under}} - q_\pi\|_{D_{\pi}},
\]

where \( D_{\pi} = \text{diag}(d_\pi) \). It aligns with the error bound given here, differing only in a constant and a norm.

4. Learning with Normalized IS Correction

Next, we show that using offline trajectory data can remove the remaining convergence condition of OTTD. Establishing the convergence without relying on specific assumptions about the tasks or features signifies that the deadly triad issue is resolved.

In this section we leverage trajectory data, where the state-action pairs to be bootstrapped are also trained, except for the last states. Thus, the condition of limiting overestimation on out-of-dataset value estimates is no longer required. The dataset \( \mathcal{D} = \{\tau_j\}_{j=1}^{n_I} \) consists of \( n_I \) trajectories. Each trajectory \( \tau_j \) is a sequence of state-action-reward tuples sampled under a behavior policy \( \mu \), defined as

\[
\tau_j = (s_i^j, a_i^j, r_i^j, s_{i+1}^j)_{i=0}^{T_j-1},
\]

with \( T_j \) indicating the length of the trajectory. In this context, \( a_i^j \sim \mu(\cdot|s_i^j) \) and \( s_{i+1}^j \sim P(\cdot|s_i^j, a_i^j) \) is generated by the MDP.

We handle the final states of each collected trajectory \( \{s_j^T \}_{j=1}^{n_I} \), by implementing a looping mechanism that sets these states to transition back to themselves with zero reward. Specifically, an additional transition \( (s_j^T, a_j^T, 0, s_j^T) \) is appended to each trajectory \( \tau_j \), with \( a_j^T \sim \mu(\cdot|s_j^T) \).

This setup aligns with episodic tasks, where each episode ends at a terminal state with zero-reward self-loop transitions (Sutton & Barto 2018). However, it does introduce some challenges in continuing tasks, leading to additional errors in value predictions: the details of these prediction errors are discussed in Section 4.2. Consequently, each trajectory can be decomposed into transitions \( \{(s_i, a_i, r_i, s_{i+1}, a_{i+1})\}_{i=1}^{n} \).

Given this setup, we apply importance sampling (IS) corrections to align off-policy data distributions with the target policy \( \pi \). For each \( (s_i^j, a_i^j) \) the corresponding IS ratio for fixing the next action’s distribution as \( \rho(a_i^j|s_i^j) = \pi(a_i^j|s_i^j)/\mu(a_i^j|s_i^j) \). This per-step action distribution correction suffices without any state distribution correction. The state-action transitions \( \tilde{P}_\pi(s', a'|s, a) \) can be estimated as

\[
\sum_{i=0}^{n} \rho(a_i^j|s_i^j)\mathbb{I}[s_i = s, a_i = a, s_{i+1} = s', a_{i+1} = a'] / n(s, a),
\]

if \( n(s, a) > 0 \); otherwise, zero. However, the high variance introduced by the IS ratio can cause bootstrapping on extremely high values and lead to instability in learning, as illustrated in Figure 3. We thus leverage the Normalized Importance Sampling (NIS) correction (Hesterberg 1995, Kuzborskij et al. 2021) to reduce the variance of the update.

4.1. Normalized Importance Sampling

When using NIS to approximate the transition probability under the target policy \( \pi \) from the state-action pair \( (s, a) \),
the sum of IS ratios for transitions from \((s, a)\) is used as the normalization term instead of the count \(n(s, a)\). More specifically, each element of the estimated transition matrix, denoted as \(P_{\pi, \text{NIS}}(s', a'|s, a)\) for \(n(s, a) > 0\), is defined as
\[
\sum_{i=0}^{n} \rho(a'|s') \mathbb{1}[s_i = s, a_i = a, s'_i = s', a'_i = a']
\]
otherwise, is set to zero. The numerator summarizes the corrected occurrences of transitions into \((s', a')\) from the state-action \((s, a)\), and the denominator represents the total corrected occurrences of the state-action \((s, a)\) in the current data set. We do not need to calculate the normalization term. Instead, the correction is achieved by assigning each transition a weight proportional to its IS ratio \(\rho(a'_i|s'_i)\) for the next action and minimizing the weighted Bellman error.

The following proposition illustrates that our transition estimator is consistent in the absence of artificially added loop transitions. The proof is given in A.3.

**Proposition 4.1.** When behaviour policies cover the support of the target policy, the transition probability estimator \(P_{\pi, \text{NIS}}(s', a'|s, a)\) is consistent for all state-action \((s, a)\) without additional loop transitions, that is the estimator tends to \(P_{\pi}(s', a'|s, a)\) almost surely as \(n(s, a) \to \infty\). Here, \(n(s, a)\) is the counts of the current state-action pair \((s, a)\).

Let \(N_{\text{NIS}} = H P_{\pi, \text{NIS}} \Phi\) be the next state-action feature matrix under the NIS transition estimate \(P_{\pi, \text{NIS}}\). The EMSBE can then be estimated as
\[
\text{EMSBE}_{\text{NIS}}(\theta) = \frac{1}{2} \| R + \gamma N_{\text{NIS}} \theta - M \theta \|^2_{D_k}. \tag{10}
\]

The update rule with NIS correction is given by
\[
\theta_{t+1} = \theta_t - \eta M^T D_k \left[ M \theta_t - (R + \gamma N_{\text{NIS}} \theta_{\text{targ}, m | \pi}) \right]. \tag{11}
\]
The target parameter is still copied fully from the student parameter every \(m\) step.

With modifications to trajectory data and the NIS correction, the condition on the matrix \(N_{\pi, \text{NIS}, M^\dagger}\) is naturally met. This matrix becomes equivalent to the non-zero square matrix in \(P_{\pi}\) and is stochastic, with its infinity norm equal to one. Therefore, the algorithm OTTD with NIS correction effectively addresses the deadly triad for off-policy tasks with trajectory data, as elaborated in the following theorem.

**Theorem 4.2.** For the over-parameterized regime \(d > k\), given a batch of trajectories, there exists a learning rate \(\eta\) and integer \(m\) such that for all update window sizes for target parameters \(m \geq m\), the OTTD update converges to
\[
\theta_{\pi, \text{TD}, \text{NIS}}^* = M^\dagger (I - \gamma N_{\pi, \text{NIS}, M^\dagger})^{-1} R + (I - M^\dagger M) \theta_0,
\]
where \(\theta_0\) is the initial point.

### 4.2. Value Prediction Error Bound

Our analysis first addresses value prediction errors in episodic tasks where trajectories end in terminal states. In these scenarios, loop transitions do not impact the error since terminal states inherently hold zero value. The primary findings of this analysis are detailed in the following corollary. The only distinction from the bound presented in Section 3 is the modification of the statistical error \(\epsilon_{\text{stat}}\). It is adjusted due to the new transition probability estimator with NIS corrections.

**Corollary 4.3.** Given a dataset \(D\) of episodic trajectory data collected under a behaviour policy \(\mu\), when the following condition holds

- \(\mu\) covers the support of the target policy \(\pi\),

with probability at least \(1 - \delta\), the fixed point of OTTD with NIS correction \(\theta_{\pi, \text{TD}}^*\) gives the following value prediction error bound
\[
\| \Phi \theta_{\pi, \text{TD}}^* - q_\pi \|_\infty \leq \epsilon_{\text{stat}}' + \epsilon_{\text{projection}} + \epsilon_{\text{approx}}, \tag{12}
\]
The statistical error $\epsilon_{stat}$ is defined as:

$$\epsilon_{stat} = \frac{\|\Phi M^{\dagger}\|_{\infty} \max\{\rho_M - 1, 1\}}{(1 - \gamma)^2 \sqrt{\min_{(s,a)} n(s,a)} \log\left(\frac{4k|A|}{\delta}\right)},$$

with $\rho_M := \max_{(s',a') \in D} \rho(a'|s') \min_{(s',a') \in D} \rho(a'|s')$.

For truncated trajectories from continuing tasks, however, the integration of loop transitions gives samples that do not follow the MDP and thus introduces extra errors in estimating the reward and transition matrix. These discrepancies prevent establishing a meaningful bound using the infinity norm. Instead, we provide a bound based on mean squared error, elaborated in A.5.

4.3. Empirical Result

In this section, we empirically analyze the value prediction errors in an episodic Four Room task using offline data from trajectories under a random behaviour policy. The target policy is chosen by a human player. The dataset only covers trajectories under a random behaviour policy. The target Q-learning, a widely adopted TD algorithm for learning optimal policies, Q-learning often bootstraps on unobserved actions with the highest value estimates. However, we aim to avoid extrapolation beyond the dataset, thereby circumventing additional assumptions on unobserved data. To address this, we adapt Q-learning by limiting the $\arg\max$ operation to be over actions seen in the dataset, which is a common technique in offline learning (Kostrikov et al. 2021, Xiao et al. 2022). To develop the modified algorithm, first define feature matrices $\Phi_i$ to present features for the state $s_i \in S$ with only seen actions. If this state does not show up in the dataset, set the matrix as a vector of zeros.

Similar to OTTD, the target parameter at iteration $nm$, for $n = 0, 1, \cdots$, copies the student parameter, $\theta_{targ, nm} = \theta_{nm}$, and is kept fixed for $m$ steps. At each iteration $t$, over-parameterized target Q-learning (OTQ) first computes values for bootstrapping using the target parameter

$$q_t(s, a) = \phi(s, a)^	op \theta_{targ, nm}[\frac{1}{m}],$$

$$y_t(s, a) = R(s, a) + \gamma \sum_{s'} \hat{P}(s'|s, a) \max_{a'} q_t(s', a'),$$

where we denote $\hat{P}$ the empirical transition estimated from the data. Then, the student parameter is updated as

$$\theta_{t+1} = \theta_t - \eta \sum_{s,a} \hat{\lambda}(s, a) \phi(s, a)(\phi(s, a)^	op \theta_t - y_t(s, a)).$$

Theorem 5.1. For the over-parameterized regime $d > k$, given that the following condition holds

- $\|\Phi_i M^{\dagger}\|_{\infty} < \frac{1}{\gamma}$, for $i = 1, \cdots, |S|$, there exists an $\hat{m}$ such that for all update window sizes of the target parameters $m \geq \hat{m}$, the OTQ update converges to $\theta^* = M^{\dagger} \hat{q}^* + (I - M^{\dagger} M) \theta_0$, where $\hat{q}^* \in \mathbb{R}^k$ satisfies

$$\hat{q}^* = R + \gamma H \hat{P} \begin{pmatrix} \Phi_1 M^{\dagger} \hat{q}^* \|_{\infty} \\ \Phi_2 M^{\dagger} \hat{q}^* \|_{\infty} \\ \vdots \\ \|\Phi_1|S| M^{\dagger} \hat{q}^* \|_{\infty} \end{pmatrix},$$

and $\theta_0$ is the initial point.

Proof for results in this section are presented in A.6. Here, the $\arg\max$ operator is expressed by the maximum norm. The optimal Q-values for the dataset, $\hat{q}^*$, are defined only for state-action pairs in $\{(s_i, a_i)\}_{i=1}^k$ with outgoing transitions. This value may not be defined for other state-action pairs to be bootstrapped on. Since the dataset may not describe any state-action pairs in $\{(s_i, a_i)\}_{i=1}^k$, their optimal Q-values cannot be evaluated through bootstrapping but only projecting. The term $\Phi_i M^{\dagger}$, for $i = 1, \cdots, |S|$, represents the projection coefficient of features onto the row space of $M$. The values for these state-action pairs depend on extrapolating $\hat{q}^*$ in proportion to the coefficient. The condition on the norm of $\Phi_i M^{\dagger}$ further prevents overestimating those extrapolated values. This condition evidences that avoiding overestimating out-of-distribution actions for learning stability is important.

In the scenario where the dataset consists of episodic trajectory data, seen state-action pairs and their transitions $\hat{P}$ form a truncated empirical MDP. Also, all data used for bootstrapping is included in the training set. As long as we constrain maximization over seen actions, there is no
extrapolation for bootstrapping values, and the convergence is established without further assumptions. Here, $\hat{q}^*$ is the optimal Q-value on this empirical MDP. We present this particular scenario in the following result.

**Corollary 5.2.** For the over-parameterized regime $d > k$, given a batch of episodic trajectories, there exists an integer $m$ such that for all update window sizes of the target parameters $m \geq m_0$, the parameter of OTQ converges to $\theta^* = M^d \hat{q}^* + (I - M^d M)\theta_0$, where

$$\hat{q}^*(s, a) = R(s, a) + \gamma \sum_{s'} \hat{P}(s'|s, a) \max_{a'} \hat{q}^*(s', a').$$

6. Related Work

Most analyses on the convergence of TD are done in online settings. At each step, the parameters are updated by one transition coming in online, either as trajectories or sampled i.i.d. TD with linear function approximation converges when data is sampled as trajectories under the target policy (Tsitsiklis & Van Roy 1996, Dayan 1992) and adding a target network gives the same fixed point (Lee & He 2019). But linear TD with off-stationary state distribution is not guaranteed to converge. This issue is called the deadly triad.

When the fixed point of TD exists, gradient TD methods (Sutton et al. 2009) converge, but much more slowly than regular TD and are sensitive to hyperparameters (Maei 2011, Sutton et al. 2008, Mahadevan et al. 2014). Several algorithms based on gradient TD have been proposed to prevent divergence and gain close-to-TD performance, but do not fully overcome the disadvantages and under-perform empirically (Ghiassian et al. 2020, Mahmood et al. 2017). Regularized least square TD (LSTD) (Boyan 1999, Lagoudakis & Parr 2003, Kolter & Ng 2009, Yu & Bertsekas 2009) with L2 regularization converges to a regularized fixed point, but it stores a feature matrix of dimension $d \times d$, which is extreme in the over-parameterized setting. The convergence and generalization properties of over-parameterized TD has been discussed in (Xiao et al. 2021, Thomas 2022).

Several papers suggest that a target network may help TD overcome the deadly triad (Zhang et al. 2021, Chen et al. 2023, Asadi et al. 2023, Fellows et al. 2023). Some studies employ various modifications to analyze the target network, such as updates with value truncation or parameter projection. Moreover, additional assumptions are frequently required to establish convergence results, which cannot be met for all data distributions, even in expected updates over the entire state-action space.

Residual minimization (RM) (Baird 1995) is known to converge with linear function approximation under any state distribution and the convergence of over-parameterized RM is also confirmed (Xiao et al. 2021). However, RM typically converges slower than TD, as observed empirically (Gordon 1995, Van Hasselt 2011) and proven in the tabular setting (Schoknecht & Merke 2003). Baird’s residual algorithm (Baird 1995) merges the parameter updates for TD and RM, with enhanced stability but still a slower convergence rate than traditional TD learning.

Other methods correct the data distribution by importance sampling (Precup 2000, Precup et al. 2001, Mahmood et al. 2014, Hesterberg 1995). However, these approaches suffer from high variance when correcting the distributions of trajectories with products of IS ratios. Later papers work on estimating state distribution ratios to avoid the ratio product. However, these methods have not yet been adopted for practical algorithms since they are still suffering from more significant variances, biases, and computation requirements. (Sutton et al. 2016, Hallak & Mannor 2017, Gelada & Bellemare 2019, Nachum et al. 2019a-b, Yang et al. 2020, Zhang et al. 2020, Che et al. 2023, He et al. 2023).

In contrast, offline RL, also known as batch RL, considers the setting where no online interactions in environments are allowed and often suffers from insufficient state space coverage and distribution shifts (Levine et al. 2020). Avoiding overestimation for out-of-distribution action values stabilizes the learning (Fujimoto et al. 2018, Kumar et al. 2019) and common techniques include constraining action selection. (Kostrikov et al. 2021, Hu et al. 2023, Xiao et al. 2022), adding pessimism (Jin et al. 2021) and limiting learnt values directly (Kumar et al. 2020). Our modified over-parameterized target Q-learning also confines maximum action selection among seen actions in the dataset.

7. Conclusion

The susceptibility of temporal difference learning to divergence has been a longstanding challenge. Numerous algorithms and techniques have been explored to address this issue, but achieving a balance between stability and performance still needs to be achieved. Our paper first demonstrated that the combination of a target network and an over-parameterized model provided a principled solution to the challenges faced by TD in off-policy learning. While our study is currently confined to linear function approximation, it offers compelling evidence for convergence guarantees and the existence of a qualified fixed point. Extending these results to neural networks would be a crucial next step to understanding TD’s empirical success with target networks.

**Impact Statement**

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.
Acknowledgements

We thankfully acknowledge funding from the Canada CIFAR AI Chairs program, the Reinforcement Learning and Artificial Intelligence (RLAI) laboratory, the Alberta Machine Intelligence Institute (Amii), and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References


Target Networks and Over-parameterization Stabilize Off-policy Bootstrapping


11


A. Appendix

A.1. Over-parameterized Target TD Convergence Proof

In this proposition, we study a sufficient condition for the convergence of over-parameterized TD (OTD).

**Proposition A.1.** A sufficient condition for the matrix $I - \eta M^T D_k(M - \gamma N)$ to have a spectral radius of less than one would be to have orthonormal feature vectors and states showing up uniformly.

**Proof.** Thanks to orthornormality, $MM^T = I$ is the identity matrix and $NM^T = H\hat{P}H^T$.

Thus, the spectral radius equals

$$\rho(I - \eta(M - \gamma N)M^T D_k) = \rho(I - \eta D_k(I - \gamma H\hat{P}H^T))$$

$$= \|I - \eta D_k + \eta D_k\gamma H\hat{P}H^T\|_{\infty}$$

$$\leq \|I - \eta D_k\|_{\infty} + \|\eta D_k\|_{\infty}\|\gamma H\hat{P}H^T\|_{\infty}$$

$$\leq 1 - \eta \min_{s,a} \lambda(s, a) + \gamma \eta \max_{s,a} \hat{\lambda}(s, a) = 1 - \eta \frac{1}{k} + \gamma \eta \frac{1}{k} < 1.$$

Next, we start the proof for the main result, the convergence of OTTD.

**Lemma A.2.** When $M$ is of full rank, $W = NM^T$ has some sub-multiplicative norm smaller than or equal to one and the learning rate satisfies $\rho(\eta MM^T D_k) < 1$, there exists an integer $\bar{m}$ such that for all target parameter update step $m \geq \bar{m}$, the spectral radius of $\gamma W + (I - \gamma W)(I - \eta MM^T D_k)^m$ is strictly smaller than one.

**Proof.** Let $U\Lambda U^{-1}$ be the eigen-decomposition of $MM^T D_k$ where $\Lambda = diag(\lambda_1, \cdots, \lambda_k)$ with $\frac{1}{\eta} > \lambda_1 \geq \cdots \geq \lambda_k > 0$. All eigenvalues are positive due to the symmetry of $D_k^2$ and the full rank of $M$. Denote the matrix $(I - \eta MM^T D_k)^m$ as $A$. Thus,

$$A = (I - \eta MM^T D_k)^m = (I - \eta U\Lambda U^{-1})^m = (UU^{-1} - \eta U\Lambda U^{-1})^m = U(I - \eta\Lambda)^m U^{-1}.$$  \hspace{1cm} (14)

The diagonal matrix $(I - \eta\Lambda)^m$ converges to zero as $m$ tends to infinity. Consequently, any norm of the matrix goes to zero as well. Such that there exists a constant $\bar{m}$, for all step $m \geq \bar{m}$, the norm of $(I - \eta\Lambda)^m$ is bounded by $\frac{1}{1 + \gamma}$. Therefore, the norm of $A$ is bounded as

$$\|A\| \leq \|U\|\|I - \eta\Lambda\|^m\|U^{-1}\| \leq \frac{1 - \gamma}{1 + \gamma}.$$

With the upper bound of the matrix norm, we can proceed and show the desired matrix has spectral radius $\rho(\gamma W + (I - \gamma W)(I - \eta MM^T D_k)^m)$ less than one for all $m \geq \bar{m}$.

$$\rho(\gamma W + (I - \gamma W)(I - \eta MM^T D_k)^m) \leq \|\gamma W + (I - \gamma W)(I - \eta MM^T D_k)^m\|$$

$$\leq \|\gamma W\| + \|(I - \gamma W)\|\|(I - \eta MM^T D_k)^m\|$$

$$\leq \gamma + (1 + \gamma)\|(I - \eta MM^T D_k)^m\| < 1.$$

**Proposition A.3.** Each $m$ steps of over-parameterized target TD updates can be combined to

$$\theta_{(n+1)m} = (I - \eta M^T BD_k(M - \gamma N))\theta_{nm} + \eta M^T BD_k R,$$

where $B = \sum_{i=0}^{m-1}(I - \eta D_k MM^T)^i$. 

13
Proof. The original update rule can be written $m$ times recursively and gives that

$$\theta_{(n+1)m} = (I - \eta M^T D_k M)\theta_{(n+1)m-1} + \eta M^T D_k (R + \gamma N\theta_{\text{targ},nm})$$

$$= (I - \eta M^T D_k M)^m \theta_{nm} + \sum_{i=0}^{m-1} (I - \eta M^T D_k M)^i \eta M^T D_k (R + \gamma N\theta_{\text{targ},nm})$$

$$= (I - \eta M^T D_k M)^m \theta_{nm} + \sum_{i=0}^{m-1} (I - \eta M^T D_k M)^i \eta M^T D_k (R + \gamma N\theta_{nm}).$$

The update rule includes powers of $I - \eta M^T D_k M$ can be reduced to terms dependent on matrix $B = \sum_{i=0}^{m-1} (I - \eta D_k MM^T)^i$.

The $m$-th power of the matrix can be reduced as

$$(I - \eta M^T D_k M)^m = (I - \eta M^T D_k M)^{m-1}(I - \eta M^T D_k M)$$

$$= (I - \eta M^T D_k M)^{m-1} - (I - \eta M^T D_k M)^{m-1} \eta M^T D_k M$$

$$= (I - \eta M^T D_k M)^{m-1} - \eta M^T (I - \eta D_k MM^T)^{m-1} D_k M$$

$$= \ldots$$

$$= I - \eta M^T \sum_{i=0}^{m-1} (I - \eta D_k MM^T)^i D_k M$$

$$= I - \eta M^T BD_k M. \quad (16)$$

Also, notice that sum of matrix power times $M^T$ can be simplified as

$$\sum_{i=0}^{m-1} (I - \eta M^T D_k M)^i \eta M^T = \sum_{i=0}^{m-1} (\eta M^T - \eta M^T D_k M\eta M^T)^i$$

$$= \eta M^T B.$$

Thus, we have

$$\theta_{(n+1)m} = (I - \eta M^T BD_k M)\theta_{nm} + \sum_{i=0}^{m-1} (I - \eta M^T D_k M)^i \eta M^T D_k (R + \gamma N\theta_{nm})$$

$$= (I - \eta M^T BD_k M)\theta_{nm} + \eta M^T BD_k (R + \gamma N\theta_{nm})$$

$$= (I - \eta M^T BD_k (M - \gamma N))\theta_{nm} + \eta M^T BD_k R.$$
**Theorem A.4.** When $M$ has full rank, $W = NM^\dagger$ has some sub-multiplicative norm smaller than or equal to one and the learning rate satisfies $\rho(\gamma NM^\dagger D_k) < 1$, there exists an integer $\bar{m}$ such that for all update steps of the target parameter $m \geq \bar{m}$, the parameter of over-parameterized target TD always converge to

$$\theta^*_\text{target} = M^\dagger(I - \gamma W)^{-1}R + (I - M^\dagger M + M^\dagger(I - \gamma W)^{-1}\gamma N(I - M^\dagger M))\theta_0,$$

where $\theta_0$ is the initial point.

**Proof.** A simple recursive argument shows that for any $\theta \in \mathbb{R}^d$, the update rule given in Proposition A.3.

$$\theta_{(n+1)m} = (I - \eta M^\dagger BD_k(M - \gamma N))\theta_{nm} + \eta M^\dagger BD_k R$$

$$= (I - \eta M^\dagger BD_k(M - \gamma N))^{n+1}\theta_0 + \sum_{i=0}^{n} (I - \eta M^\dagger BD_k(M - \gamma N))^i\eta M^\dagger BD_k R$$

$$= \left( (I - \eta M^\dagger BD_k(M - \gamma N))^{n+1}\theta_0 + \eta M^\dagger BD_k \sum_{i=0}^{n} (I - \gamma W)^i \eta MM^\dagger BD_k R \right). \quad \text{(17)}$$

where $B = \sum_{i=0}^{m-1} (I - \eta D_k MM^\dagger)^i$. Also, the last line uses that $(M - \gamma N)M^\dagger = (I - \gamma NM^\dagger)MM^\dagger$, where we further denote $NM^\dagger$ to $W$.

We work on the second term independent of the initial point first and simplify the summation of matrix powers. Reversing the recursive step in Equation 16, $\eta MM^\dagger BD_k$ can be expressed as $I - (I - \eta MM^\dagger D_k)^m$. Thus, the matrix in the summation can be rewritten as

$$I - (I - \gamma W)\eta MM^\dagger BD_k = \gamma W + (I - \gamma W)(I - \eta MM^\dagger D_k)^m$$

By Lemma A.2, there exists $\bar{m}$ such that for all $m \geq \bar{m}$, $I - (I - \gamma W)\eta MM^\dagger BD_k$ has spectral radius less than one. Thus, by properties of Neumann series, the summation of its matrix powers converges, that is

$$\sum_{i=0}^{n} (I - (I - \gamma W)^i \eta MM^\dagger BD_k) \to [(I - \gamma W)\eta MM^\dagger BD_k]^{-1}, \quad \text{as } n \to \infty. \quad \text{(18)}$$

Notice that $B$ is invertible, since $(I - \eta D_k MM^\dagger)^\dagger$ is strictly diagonal dominant and thus has positive eigenvalues. Then its powers and sum of powers also have positive eigenvalues and thus are invertible. Also, $MM^\dagger$ is invertible, thanks to the full rank of $M$. Therefore,

$$\lim_{n \to \infty} \sum_{i=0}^{n} \eta M^\dagger BD_k(I - (I - \gamma W)\eta MM^\dagger BD_k)^i R$$

$$= \eta M^\dagger BD_k(I - \gamma W)^{-1}R$$

$$= M^\dagger(I - \gamma W)^{-1}R. \quad \text{(19)}$$

Now we start dealing with the first term in Equation 17 and study how the fixed point depends on the initial value.

$$\eta D_k (M - \gamma N)^{n+1}$$

$$= I - \sum_{i=0}^{n} (I - \eta M^\dagger BD_k(M - \gamma N))^i\eta M^\dagger BD_k(M - \gamma N)$$

$$= I - \sum_{i=0}^{n} \eta M^\dagger BD_k(I - (I - \gamma W)\eta MM^\dagger BD_k)^i(M - \gamma N)$$

$$\xrightarrow{n \to \infty} I - M^\dagger BD_k(MM^\dagger BD_k)^{-1}(I - \gamma W)^{-1}(M - \gamma NM^\dagger M - \gamma N(I - M^\dagger M))$$

$$= I - M^\dagger M - M^\dagger(I - \gamma W)^{-1}\gamma N(I - M^\dagger M).$$
Given a dataset $D$, the second line uses the same recursive trick in Equation 16.

In conclusion, the parameter $\theta_{\text{min}}$ converges to $M^\dagger (I - \gamma W)^{-1} R$, as update step $n$ goes to infinity, when the initial point equals zero.

Next, we prove the proposition for the special case with expected updates. The data distribution used for expected updates covers the entire state-action space.

**Proposition A.5.** For the over-parameterized regime $d > k$, there always exists a learning rate $\eta$ and an integer $m$ such that for all update window sizes of the target parameters $m \geq m$, the parameter of OTTD converges to

$$\theta^*_{TD} = \Phi^\dagger (I - \gamma P\pi)^{-1} R + (I - \Phi^\dagger \Phi) \theta_0,$$

where $\theta_0$ is the initial parameter of OTTD.

**Proof.** These updates assume ideal offline data encompassing all state-action pairs such that $k = |S| |A|$, $M = \Phi$, $N = \pi^\dagger \Phi$ and $D = \text{diag}(\lambda)$. It can be easily verified that for the expected update, we have $NM^\dagger = \pi^\dagger \Phi^\dagger = \pi^\dagger$, which naturally satisfies that it has a sub-multiplicative norm smaller than or equal to one with each row summing up to one. Hence, as demonstrated in Theorem 3.2, OTTD converges with a proven guarantee.

**A.2. Bound of Value Estimation Error**

**Theorem A.6.** Given a dataset $D$ and the optimal fixed point $\theta^*_{TD} = M^\dagger (I - \gamma N M^\dagger)^{-1} R$, when $\|NM^\dagger\|_\infty < 1$, with probability at least $1 - \delta$, for $\theta^* = \arg\min_\theta \|\Phi \theta - q\|_\infty$.

$$\|\Phi \theta^*_{TD} - q\|_\infty \leq \frac{\|\Phi M^\dagger\|_\infty}{(1 - \gamma)^2 \sqrt{\min_{s,a} n(s,a)}} \sqrt{\log(\frac{2k|A|}{\delta})} + \|\Phi M^\dagger\|_\infty \|\Phi(I - M^\dagger M)\theta^*\|_\infty + \frac{2\|\Phi M^\dagger\|_\infty}{1 - \gamma} \|\Phi \theta^* - q\|_\infty.$$  

**Proof.**

$$\|\Phi \theta^*_{TD} - q\|_\infty = \|\Phi M^\dagger M \theta^*_{TD} - q\|_\infty 
\leq \|\Phi M^\dagger M \theta^*_{TD} - \Phi M^\dagger H q\|_\infty + \|\Phi M^\dagger H q - q\|_\infty.$$  

Define $\theta^* = \arg\min_\theta \|\Phi \theta - q\|_\infty$. By adding intermediate term, we can bound the second term

$$\|\Phi M^\dagger H q - q\|_\infty = \|\Phi M^\dagger M \theta^* - q\|_\infty 
\leq \|\Phi M^\dagger H(q - \Phi \theta^*)\|_\infty + \|\Phi \theta^* - q\|_\infty + \|\Phi(I - M^\dagger M)\theta^*\|_\infty 
\leq (\|\Phi M^\dagger\|_\infty + 1) \|\Phi \theta^* - q\|_\infty + \|\Phi(I - M^\dagger M)\theta^*\|_\infty.$$

1. $A = \|H \hat{P}(\Phi M^\dagger H q - q)\|_\infty \leq 2\|\Phi \theta^* - q\|_\infty + \|\Phi(I - M^\dagger M)^*\|_\infty.$
2. Define $\epsilon_{est} = \|\gamma H (P - \hat{P}) q\|_\infty + \|R - H r\|_\infty.$
3. $B = \|H \hat{P} \Phi M^\dagger (M \theta^*_{TD} - H q)\|_\infty \leq \frac{1}{1 - \gamma} \epsilon_{est} + \frac{\gamma}{1 - \gamma} A.$

The first equation repeats the steps to bound the term in Equation 23 and uses $\|H \hat{P} \Phi M^\dagger\|_\infty = \|NM^\dagger\|_\infty \leq 1.$

The third equation further uses the extension of $M \theta^*$ and $H q$ following the Bellman update:

$$M \theta^*_{TD} = R + \gamma W M \theta^*_{TD} = R + \gamma H \hat{P} \Phi M^\dagger M \theta^*_{TD},$$

$$H q = H r + \gamma H P q = H r + \gamma H \hat{P} \Phi M^\dagger H q + \gamma H \hat{P}(q - \Phi M^\dagger H q) + \gamma H (P - \hat{P}) q.$$
Next, we can bound the first term in Equation 22 by extending $M_{\theta T,D}$ and $Hq$.

$$\|\Phi M^\dagger M_{\theta T,D} - \Phi M^\dagger Hq\|_{\infty}$$

$$\leq \|\Phi M^\dagger (R - Hr + \gamma H(P - \hat{P})q + \Phi M^\dagger \gamma H \hat{P} \Phi M^\dagger (M_{\theta T,D} - Hq) + \Phi M^\dagger \gamma H \hat{P}(q - \Phi M^\dagger Hq)\|_{\infty}$$

$$\leq \|\Phi M^\dagger\|_{\infty} \epsilon_{est} + \gamma \|\Phi M^\dagger\|_{\infty} A + \gamma \|\Phi M^\dagger\|_{\infty} B$$

$$\leq \|\Phi M^\dagger\|_{\infty} \epsilon_{est} + \gamma \|\Phi M^\dagger\|_{\infty} A$$

$$\leq \|\Phi M^\dagger\|_{\infty} \epsilon_{est} + \frac{2\gamma \|\Phi M^\dagger\|_{\infty}}{1 - \gamma} \|\Phi* - q\|_{\infty} + \gamma \|\Phi M^\dagger\|_{\infty} \|\Phi(I - M^\dagger M)^*\|_{\infty}. \quad (24)$$

Thus, combining the bounds on the first and the second term in Equation 22, we gain

$$\|\Phi_{\theta T,D} - q\|_{\infty}$$

$$\leq \|\Phi M^\dagger\|_{\infty} \epsilon_{est} + \frac{2\gamma \|\Phi M^\dagger\|_{\infty}}{1 - \gamma} \|\Phi(I - M^\dagger M)^*\|_{\infty} + \frac{2\gamma \|\Phi M^\dagger\|_{\infty}}{1 - \gamma} \|\Phi* - q\|_{\infty}$$

$$\leq \|\Phi M^\dagger\|_{\infty} \epsilon_{est} + \frac{2\|\Phi M^\dagger\|_{\infty}}{1 - \gamma} \|\Phi(I - M^\dagger M)^*\|_{\infty} + \frac{2\|\Phi M^\dagger\|_{\infty}}{1 - \gamma} \|\Phi* - q\|_{\infty}, \quad (25)$$

since $\|\Phi M^\dagger\|_{\infty} \geq 1.$

$\epsilon_{est}$ is bounded by Hoeffding’s inequality and a union bound. We have with probability at least $1 - \delta$, for any state $s$ and action $a$ with $d_k(s) > 0$ in the dataset,

$$|(P_{s,a} - \hat{P}_{s,a})^T q| \leq \frac{1}{(1 - \gamma) \sqrt{2n(s,a)}} \sqrt{\log\left(\frac{2k|A|}{\delta}\right)}.$$

$$|R(s,a) - r(s,a)| \leq \frac{1}{\sqrt{2n(s,a)}} \sqrt{\log\left(\frac{2k|A|}{\delta}\right)}.$$

Combining all terms, we gain the bound.

As the dataset gradually covers the entire state space and gain the actual transition matrix and reward, that is, $\hat{P}_\mu \to P_\mu$, $R \to r$ and $M \to \Phi$, $\epsilon_{est} \to 0$ and $\Phi M^\dagger \to I$. Only $\frac{2\gamma \|\Phi* - q\|_{\infty}}{1 - \gamma}$ is left in the bound.

A.3. Normalized Importance Sampling Results

**Proposition A.7.** The transition probability estimator $\hat{P}_\pi(s', a'|s, a)$ is consistent and tends to $P_\pi(s', a'|s, a)$ almost surely as $n(s,a) \to \infty$, which is the counts of the current state-action pair $(s, a)$.

**Proof.** Each time a current state-action pair $(s, a)$ shows up, we can define two random variables:

$$X = \rho(a'|s') \mathbb{1}[S = s, A = a, S' = s', A' = a'], \quad (26)$$

$$Y = \sum_{s,a} \rho(a|s) \mathbb{1}[S = s, A = a, S' = s, A' = a]. \quad (27)$$

These two random variables are sampled $n(s,a)$ times, labeled from $j = 1$ to $n(s,a)$. Since each next action can be sampled from a different behavior policy, $X_1, \cdots, X_{n(s,a)}$ are not sampled from the same distribution but are independent conditioned on the current state-action pair, similar for $Y_j, j = 1, \cdots, n(s,a)$.

Our estimator $\hat{P}_\pi(s', a'|s, a)$ can be seen as a ratio of the average of these random variables, that is,

$$\hat{P}_\pi(s', a'|s, a) = \frac{\sum_{j=1}^{n(s,a)} X_j}{\sum_{j=1}^{n(s,a)} Y_j} \quad (28)$$
Denote given a batch of trajectories under some behaviour policies, if

\[ \text{Theorem A.8.} \]

Then by strong large law of number for martingales, \( \epsilon \)

\[ \text{probability at least} \quad \theta \]

Given a dataset \( \text{Corollary A.10.} \)

\[ \text{Lemma A.9.} \]

First, we state a lemma from Sharoff and colleagues (2020) for the concentration inequality of average ratios.

\[ \text{A.4. Value Estimation Error Bound with NIS for Episodic Tasks} \]

\[ \text{Proof.} \]

The update rule equals \( \theta_{t+1} = \theta_t - \eta M^\top D_k \left[ M \theta_t - (R + \gamma N_{\text{NIS}M^1}) \right] \), which is of the same form as Theorem 3.2.

As long as the assumptions for Theorem 3.2 are satisfied, the convergence is proved. The assumption requires the projected coefficient matrix \( N_{\text{NIS}M^1} = H \hat{\pi}_{\text{NIS}} \Phi M^1 \) to satisfy the condition of having a norm less than one. Here, \( W_{\text{NIS}} \) equals the non-zero square matrix in \( \hat{\pi} \) and is also stochastic. Therefore, the infinity norm is less than one and the condition is satisfied.

\[ \text{A.4. Value Estimation Error Bound with NIS for Episodic Tasks} \]

First, we state a lemma from Sharoff and colleagues (2020) for the concentration inequality of average ratios.

\[ \text{Lemma A.9.} \]

Let \( X, Y \) be possibly dependent random variables with joint distribution \( P \). Consider a sample \( (X_1, Y_1), \ldots, (X_n, Y_n) \) of independent copies of \( (X, Y) \sim P \). Assume that \( X \) takes values in \([0, 1]\) and \( Y \) takes values in \([1, B]\). Define \( \mu_Y := E[Y] \) and let \( \bar{X} \) denote the sample mean of \( X_1, \ldots, X_n \) (likewise for \( Y \) and \( Y_1, \ldots, Y_n \)). For any choice of \( \delta \in [0, 1] \), we have with probability at least \( 1 - \delta \) over the sample,

\[ \left| \bar{X} - \frac{E[X]}{E[Y]} \right| \leq \sqrt{\frac{(B - 1) \log \frac{4}{\delta}}{2n}} + \frac{2(B - 1) \log \frac{4}{\delta}}{3 \mu_Y n} + \sqrt{\frac{4 \log \frac{4}{\delta}}{2n}} \leq \frac{\max\{1, B - 1\} \log \frac{4}{\delta}}{\min\{\mu_Y, 1\} \sqrt{2n}}. \]

\[ \text{Corollary A.10.} \]

Given a dataset \( D \) of episodic trajectory data collected under a behaviour policy \( \mu \) and the optimal fixed point \( \theta_{T,D}^\pi = M^1(I - \gamma H \hat{\pi} \Phi M^1)^{-1} R \), when \( ||H \hat{\pi} \Phi M^1||_\infty < 1 \) and \( \mu \) covers the support of the target policy \( \pi \), with probability at least \( 1 - \delta \),

\[ \| \Phi \theta_{T,D} - q_\pi \|_\infty \]

\[ \leq \epsilon_{\text{stat}} + \epsilon_{\text{projection}} + \epsilon_{\text{approx}}. \]

Denote \( \rho_M := \frac{\max_{(s', a') \in D} \rho(a' \mid s')}{\min_{(s', a') \in D} \rho(a' \mid s')} \),

\[ \epsilon_{\text{stat}} = \frac{\| \Phi M^1 \|_\infty \rho_M}{(1 - \gamma) \sqrt{\min_{(s, a)} n(s, a)}} \log \left( \frac{2k|A|}{\delta} \right) \max\{\rho_M - 1, 1\}. \]

\( \epsilon_{\text{projection}} \) and \( \epsilon_{\text{approx}} \) are the same as in Theorem 3.5.
We have with probability at least \( 1 - \epsilon \), as defined in A.3, this only influences the statistical error, \( \epsilon \). Let Corollary A.11. Given a dataset with trajectory data under a behaviour policy \( \mu \) and the optimal fixed point \( \theta^{*}_{TD} = M^{\dagger}(I - \gamma \hat{P}_{\pi} \Phi^{\dagger} M^{\dagger})^{-1} R \), when \( \| \hat{P}_{\pi} \Phi^{\dagger} M^{\dagger} \|_{\pi} < 1 \) and \( \mu \) covers the support of the target policy \( \pi \), with probability at least \( 1 - \delta \), under the stationary distribution \( d_{\pi} \),

\[
\| \Phi \theta^{*}_{TD} - q \|_{\infty} \leq \frac{\| \Phi M^{\dagger} \|_{\infty}}{1 - \gamma} \epsilon_{\text{stat}} + \frac{\| \Phi M^{\dagger} \|_{\infty}}{1 - \gamma} \| \Phi(I - M^{\dagger} M)\theta^{*} \|_{\infty} + \frac{2\| \Phi M^{\dagger} \|_{\infty}}{1 - \gamma} \| \Phi \theta^{*} - q \|_{\infty}.
\]

Compared to the fixed point of OTTD with sampled target actions, only the transition probability estimation \( \hat{P} \) is changed. This only influences the statistical error, \( \epsilon_{\text{stat}} = \| \gamma H(P - \hat{P})q \|_{\infty} + \| R - Hr \|_{\infty} \), which is bounded by the above lemma and a union bound.

\[
\hat{P}_{\pi, \text{NIS}}(s, a, s', a') = \sum_{i=1}^{n} \rho(a'|s') \mathbb{I}[S_i = s, A_i = a, S'_i = s', A'_i = a'],
\]

As defined in A.3,

\[
X = \rho(a'|s') \mathbb{I}[S = s, A = a, S' = s', A' = a'],
\]

\[
Y = \sum_{s', \hat{a}} \rho(\hat{a}|s) \mathbb{I}[S = s, A = a, S' = \hat{s}, A' = \hat{a}].
\]

We have with probability at least \( 1 - \delta \), for any state \( s \) and action \( a \) with \( d_k(s) > 0 \) in the dataset,

\[
| (P_{s,a} - \hat{P}_{s,a}) \gamma q | \leq \frac{\rho_M \max\{\rho_M - 1, 1\}}{(1 - \gamma) / \sqrt{2n(s,a)}} \log\left( \frac{4k|A|}{\delta} \right),
\]

\[
| R(s,a) - r(s,a) | \leq \frac{1}{\sqrt{2n(s,a)}} \sqrt{2k|A| / \delta}.
\]

A.5. Continuing Tasks’ Error Bound

Corollary A.11. Given a dataset with trajectory data under a behaviour policy \( \mu \) and the optimal fixed point \( \theta^{*}_{TD} = M^{\dagger}(I - \gamma \hat{P}_{\pi} \Phi^{\dagger} M^{\dagger})^{-1} R \), when \( \| \hat{P}_{\pi} \Phi^{\dagger} M^{\dagger} \|_{\pi} < 1 \) and \( \mu \) covers the support of the target policy \( \pi \), with probability at least \( 1 - \delta \), under the stationary distribution \( d_{\pi} \),

\[
\| \Phi \theta^{*}_{TD} - q_{\pi} \|_{\pi} \leq \epsilon_{\text{stat}}'' + \epsilon_{\text{projection}}'' + \epsilon_{\text{approx}}''.
\]

Denote \( \rho_M := \max_{(s', a', s') \in D} \rho(a'|s') \) and \( C := \| \Phi M^{\dagger} H \|_{\pi} \).

\[
\epsilon_{\text{stat}}'' = \frac{C \rho_M}{(1 - \gamma)^2 \min_{n(s,a)} n(s,a)} \log\left( \frac{2k|A|}{\delta} \right) \max\{\rho_M - 1, 1\} + \frac{C}{(1 - \gamma)^2}.
\]

Let \( \theta^{*} \) be defined as \( \theta^{*} = \arg\min_{\theta} \| \Phi \theta - q_{\pi} \|_{\pi} \pi \). Then, \( \epsilon_{\text{projection}}'' = \frac{C(1 + \gamma \| \hat{P}_{\pi, \text{NIS}} \|_{\pi})}{(1 - \gamma)^2} \| \Phi(I - M^{\dagger} M)\theta^{*} \|_{\pi} + \frac{C(1 + \gamma \| \hat{P}_{\pi, \text{NIS}} \|_{\pi})}{(1 - \gamma)^2} \| \Phi(I - M^{\dagger} M)\theta^{*} \|_{\pi} \), and

\[
\epsilon_{\text{approx}}'' = \frac{C(2 + \gamma \| \hat{P}_{\pi, \text{NIS}} \|_{\pi})}{(1 - \gamma)^2} \| \Phi \theta^{*} - q_{\pi} \|_{\pi}.
\]

Proof. The proof is similar as the one in A.2 except for a norm change.

\[
\| \Phi \theta^{*}_{TD} - q \|_{\pi} \leq \| \Phi M^{\dagger} \theta^{*}_{TD} - q \|_{\pi} + \| \Phi M^{\dagger} H q \|_{\pi} + \| \Phi M^{\dagger} H q - q \|_{\pi}.
\]
Define \( \theta^* = \text{argmin}_\theta \| \Phi \theta - q \|_{D_x} \). By adding intermediate term, we can bound the second term

\[
\| \Phi M^\dagger Hq - q \|_{D_x} \\
\leq \| \Phi M^\dagger (q - \Phi \theta^*) \|_{D_x} + \| \Phi \theta^* - q \|_{D_x} + \| \Phi(I - M^\dagger M) \theta^* \|_{D_x} \\
\leq (\| \Phi M^\dagger H \|_{D_x} + 1) \| \Phi \theta^* - q \|_{D_x} + \| \Phi(I - M^\dagger M) \theta^* \|_{D_x}.
\]

1. \( A = \| \hat{P}(\Phi M^\dagger Hq - q) \|_{D_x} \leq (1 + \| \hat{P}_{\pi,\text{NIS}} \|_{D_x}) \| \Phi \theta^* - q \|_{D_x} + \| \hat{P}_{\pi,\text{NIS}} \|_{D_x} \| \Phi(I - M^\dagger M) \theta^* \|_{D_x} \).

2. Define \( \varepsilon_{\text{est}} = \| \gamma(P - \hat{P}_{\pi,\text{NIS}})q \|_{D_x} + \| R - r \|_{D_x} \).

3. \( B = \| \hat{P}_{\pi,\text{NIS}} \Phi M^\dagger (M \theta^*_{TD} - Hq) \|_{D_x} \leq \frac{1}{1 - \gamma} \varepsilon_{\text{est}} + \frac{\gamma}{1 - \gamma} A. \)

These three statements are almost the same as in the proof in A.2 except the norm difference. When bounding \( \varepsilon_{\text{est}} \), we need to count the error from loop transitions. It adds in wrong transitions for the last state-action pair \((s_T, a_T)\) and thus, we can only bound as

\[
| (P_{s_T,a_T} - \hat{P}_{s_T,a_T})^\top q | \leq \frac{2}{(1 - \gamma)}.
\]

In the norm, this term is weighted by the stationary distribution \( d_x(s_T, a_T) \). For state-action pair without additional loop transitions, we use the lemma from Sharoff and colleagues (2020) stated in A.4 and a union bound.

Combining all terms, we gain the bound.

\[\Box\]

A.6. Proof of Q-learning Convergence

**Theorem A.12.** When \( \| \Phi^\top M^\dagger \|_\infty \leq \frac{1}{\gamma} \), there exists an integer \( \bar{m} \) such that for all update steps of the target parameter \( m \geq \bar{m} \), parameter of over-parameterized target Q-learning converges to a fixed point \( \hat{q}^* = M^\dagger \hat{q}^* + (I - M^\dagger M) \theta_0 \), where \( \hat{q}^* \in \mathbb{R}^k \) satisfies

\[
\hat{q}^* = R + \gamma \hat{P} \left( \begin{array}{c}
\| \Phi_1 M^\dagger \hat{q}^* \|_\infty \\
\| \Phi_2 M^\dagger \hat{q}^* \|_\infty \\
\| \Phi_\pi M^\dagger \hat{q}^* \|_\infty
\end{array} \right),
\]

and \( \theta_0 \) is the initial point.

**Proof.** We proceed in the same way to Theorem 1.1. The update rule for student parameter every \( m \) steps equal

\[
\theta_{(n+1)m} = (I - \eta M^\top BD_k M) \theta_{nm} + \eta M^\top BD_k R + \eta M^\top BD_k \gamma \hat{P} \left( \begin{array}{c}
\| \Phi_1 \theta_{nm} \|_\infty \\
\| \Phi_2 \theta_{nm} \|_\infty \\
\| \Phi_\pi \theta_{nm} \|_\infty
\end{array} \right).
\]

When \( \theta_0 = M^\top y \in \text{row} \mathcal{sp}(M) \) for some \( y \in \mathbb{R}^k \), \( \theta_t \) stays in the row space of \( M \) for all \( t \) and the estimated Q-values satisfy

\[
M \theta_{(n+1)m} = M(I - \eta M^\top BD_k M) \theta_{nm} + \eta MM^\top BD_k R + \eta \eta MM^\top BD_k \gamma \hat{P} \left( \begin{array}{c}
\| \Phi_1 M \theta_{nm} \|_\infty \\
\| \Phi_2 M \theta_{nm} \|_\infty \\
\| \Phi_\pi M \theta_{nm} \|_\infty
\end{array} \right).
\]

The last line uses that \( \theta_{nm} = M^\top y = M^\dagger MM^\top y = M^\dagger M \theta_{nm}. \)
Compared to the TD analysis, the only change here is that instead of using next states’ value estimations, we use the maximum Q-values at next states.

According to the above combined $m$ step update rule on $M\theta$ in Equation 42, we define an operator $\mathcal{T} : \mathbb{R}^k \to \mathbb{R}^k$ for $x = M\theta \in \mathbb{R}^k$ as

$$
\mathcal{T}x = (I - \eta \mathcal{M}^T D_k)x + \eta \mathcal{M}^T D_k R + \eta \mathcal{M} M^T D_k \gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T x\|_\infty \\
\|\Phi_2 \mathcal{M}^T x\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T x\|_\infty
\end{pmatrix}
$$

Notice that this operator is contractive. Given any two vectors $\bar{x}$ and $x'$,

$$
\|\mathcal{T}x - \mathcal{T}x'||_\infty \\
= \|(I - \eta \mathcal{M}^T D_k)(\bar{x} - x') + \eta \mathcal{M}^T D_k \gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T \bar{x}\|_\infty \\
\|\Phi_2 \mathcal{M}^T \bar{x}\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T \bar{x}\|_\infty
\end{pmatrix} - \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T x'\|_\infty \\
\|\Phi_2 \mathcal{M}^T x'\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T x'\|_\infty
\end{pmatrix}\|_\infty \\
\leq \|(I - \eta \mathcal{M}^T D_k)^m (\bar{x} - x')\|_\infty + \|(I - \eta \mathcal{M}^T D_k)^m\|_\infty \|\gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T (\bar{x} - x')\|_\infty \\
\|\Phi_2 \mathcal{M}^T (\bar{x} - x')\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T (\bar{x} - x')\|_\infty
\end{pmatrix}\|_\infty \\
\leq \|(I - \eta \mathcal{M}^T D_k)^m\|_\infty + \gamma m \|(I - \eta \mathcal{M}^T D_k)^m\|_\infty \|\gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T \bar{x}\|_\infty \\
\|\Phi_2 \mathcal{M}^T \bar{x}\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T \bar{x}\|_\infty
\end{pmatrix}\|_\infty - \|\gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T x'\|_\infty \\
\|\Phi_2 \mathcal{M}^T x'\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T x'\|_\infty
\end{pmatrix}\|_\infty.
$$

The third line uses $\eta \mathcal{M}^T D_k = I - (I - \eta \mathcal{M}^T D_k)^m$ following the recursive argument in Equation 16.

Set $\bar{m} = 1 + \left[ \log(c - \gamma) - \log(1 + \gamma) \right] / \log(1 - \eta \lambda_k)$ for some constant $c \in (0, 1)$. As shown in Lemma A.2, $\|(I - \eta \mathcal{M}^T D_k)^m\|_\infty < \frac{c - \gamma}{1 + \gamma}$. In this case, the above norm can be bounded by $\frac{c - \gamma}{1 + \gamma} + \gamma [1 + \frac{C - \gamma}{1 + \gamma}] \|\bar{x} - x'||_\infty$ and is smaller than $c \|\bar{x} - x'||_\infty$. Thus, the operator on estimated Q-values is contractive.

The dataset defines a fixed point for estimated Q-values as

$$
\hat{q}^* = R + \gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T \hat{q}^*\|_\infty \\
\|\Phi_2 \mathcal{M}^T \hat{q}^*\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T \hat{q}^*\|_\infty
\end{pmatrix},
$$

and $\hat{q}^*$ exists unique since the right hand side update rule is contractive.

This $\hat{q}^*(s, a)$ is also the unique fixed point of the operator $\mathcal{T}$ in the metric space $(\mathbb{R}^k, \| \cdot \|_\infty)$ by the Banach fixed point theorem, since

$$
\mathcal{T}\hat{q}^* = (I - \eta \mathcal{M}^T D_k)\hat{q}^* + \eta \mathcal{M}^T D_k R + \eta \mathcal{M} M^T D_k \gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T \hat{q}^*\|_\infty \\
\|\Phi_2 \mathcal{M}^T \hat{q}^*\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T \hat{q}^*\|_\infty
\end{pmatrix}
$$

$$
= (I - \eta \mathcal{M}^T D_k)^m \hat{q}^* + [I - (I - \eta \mathcal{M}^T D_k)^m][R + \gamma \hat{P} \begin{pmatrix}
\|\Phi_1 \mathcal{M}^T \hat{q}^*\|_\infty \\
\|\Phi_2 \mathcal{M}^T \hat{q}^*\|_\infty \\
\vdots \\
\|\Phi_S \mathcal{M}^T \hat{q}^*\|_\infty
\end{pmatrix}]
$$

$$
= (I - \eta \mathcal{M}^T D_k)^m \hat{q}^* + [I - (I - \eta \mathcal{M}^T D_k)^m]\hat{q}^*
$$

$$
= \hat{q}^*.
$$

The second equation again uses $\eta \mathcal{M}^T D_k = I - (I - \eta \mathcal{M}^T D_k)^m$ following the recursive argument in Equation 16.
### Table 2.
The table shows hyperparameters for all algorithms tuned on the Baird counterexample. All hyperparameters are found by grid search.

The third line uses the definition of $\hat{q}^*$. 

An initial point can be expressed as $\theta_0 = M^\dagger M \theta_0 + (I - M^\dagger M) \theta_0$.

For the part, $M^\dagger M \theta_0$, lying in the row space of $M$, we have $M M^\dagger M \theta_0$ converges uniquely to the fixed point $\hat{q}^* = M \theta^*$ where $\theta^* = M^\top y$ lies in the row space of $M$ for some $y \in \mathbb{R}^k$. Thus, $M^\dagger M \theta_t \to \theta^* = M^\dagger \hat{q}^*$ converges uniquely as $t \to \infty$.

The other part $(I - M^\dagger M) \theta_0$ is always unmodified by the over-parameterized target Q-learning update rule and thus left unchanged as $(I - M^\dagger M) \theta_t = (I - M^\dagger M) \theta_0$ for all $t$.

Therefore, $\theta_t \to M^\dagger \hat{q}^* + (I - M^\dagger M) \theta_0$ as $t \to \infty$.

#### A.7. Empirical Setting

The features for the four room concatenate one-hot encoding for $x$ and $y$ coordinate separately and the action. Then, we append the matrix $H^\top$, each row of length $k$, representing if state-action pairs show up in the dataset and the showing-up order, to the encoding. Thus, the dimension of features is larger than the number $k$ of state-actions in the dataset and the model is over-parameterized.

The behaviour policy is random, that is, four actions are sampled with the same probability. A human policy is given by a human player, considered as optimal by the player. The target policy is the human policy combined with $\epsilon$-exploration with $\epsilon = 0.08$.

All hyperparameters are tuned with the small dataset of size 300. Since over-parameterized TD converges on the Four Room task, then the target parameter update step is set to one and OTTD is the same as TD. All empirical results are averaged over 10 random seeds.
Target Networks and Over-parameterization Stabilize Off-policy Bootstrapping

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Learning Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Off-Policy without Corrections</td>
<td>0.95</td>
</tr>
<tr>
<td>Target Actions</td>
<td>0.97</td>
</tr>
<tr>
<td>NIS</td>
<td>0.97</td>
</tr>
<tr>
<td>IS</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*Table 3.* The table shows hyperparameters for all algorithms tuned on the Four Room Task. All hyperparameters are found by grid search.

<table>
<thead>
<tr>
<th>Work</th>
<th>Regularity Condition</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours Thm. 3.2</td>
<td>$|NM|^\top &lt; 1$</td>
<td>It is satisfied for expected updates or a batch of complete trajectories naturally.</td>
</tr>
<tr>
<td>Lee and He Thm. 1 (2019)</td>
<td>None</td>
<td>No regularization is needed for the on-policy learning.</td>
</tr>
<tr>
<td>Asadi et al. Prop. 1 (2023)</td>
<td>$\rho((\Phi^\top D\Phi)^{-1}(\gamma\Phi^\top DP_\pi^\top \Phi)) &lt; 1$</td>
<td>The condition fails on a Two-state counterexample even with expected updates.</td>
</tr>
<tr>
<td>Asadi et al. Prop. 5 (2023)</td>
<td>$\frac{\lambda_{\text{max}}(\gamma\Phi^\top DP_\pi^\top \Phi)}{\lambda_{\text{min}}((\Phi^\top D\Phi)^{-1})} &lt; 1$</td>
<td>The condition fails on a Two-state counterexample even with expected updates.</td>
</tr>
<tr>
<td>Fellows et al. Thm. 2 (2023)</td>
<td>$M^\top D_\pi(\gamma N - M)$ has strictly negative eigenvalues</td>
<td>The condition is equivalent to the spectral radius less-than-one condition. Breaking this condition is the main factor behind the divergence with the deadly triad. With this assumption, the paper does not focus on the deadly triad issue.</td>
</tr>
<tr>
<td>Fellows et al. Thm. 4 (2023)</td>
<td>$|((\Phi^\top D\Phi)^{-1}(\gamma\Phi^\top DP_\pi^\top \Phi))| &lt; 1$</td>
<td>The condition fails on a Two-state counterexample even with expected updates.</td>
</tr>
<tr>
<td>Shangtong et al. Thm. 2 (2021)</td>
<td>Projection of the target parameter into a ball and L2 regularization</td>
<td>Projection is hard to realize empirically, and L2 regularization can give a parameter predicting worse than zero values.</td>
</tr>
</tbody>
</table>

*Table 4.* This table compares how strong the regularity conditions are to ensure convergence in the deadly triad under linear function approximation.
Table 5. Comparison of assumptions among analysis of target networks under linear function approximation.

<table>
<thead>
<tr>
<th>Work</th>
<th>MDP</th>
<th>Data Generation Distribution</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours Thm. 3.2</td>
<td>None</td>
<td>None</td>
<td>Full rank</td>
</tr>
<tr>
<td>Lee and He Thm. 1 (2019)</td>
<td>Ergodic under the target policy $\pi$</td>
<td>$s \sim d_\pi$ i.i.d. with $d_\pi(s) &gt; 0$ for all $s$</td>
<td>Full rank</td>
</tr>
<tr>
<td>Asadi et al. Prop. 1 (2023)</td>
<td>None</td>
<td>None</td>
<td>Full rank</td>
</tr>
<tr>
<td>Asadi et al. Prop. 5 (2023)</td>
<td>None</td>
<td>None</td>
<td>Full rank</td>
</tr>
<tr>
<td>Fellows et al. (2023) Thm. 2</td>
<td>None</td>
<td>$s \sim d$ i.i.d. for some off-policy distribution $d$</td>
<td>$|\phi(s, a)\phi(s, a)|<em>\gamma$ and $|\phi(s, a)\phi(s', a')|</em>\gamma$ are bounded, the space of the parameter $\theta$ is convex, and variance of the update is bounded</td>
</tr>
<tr>
<td>Fellows et al. (2023) Thm. 4</td>
<td>None</td>
<td>$s \sim d$ i.i.d. for some off-policy distribution $d$</td>
<td>$|\phi(s, a)\phi(s, a)|<em>\gamma$ and $|\phi(s, a)\phi(s', a')|</em>\gamma$ are bounded, the space of the parameter $\theta$ is convex, and variance of the update is bounded</td>
</tr>
<tr>
<td>Shangtong et al. Thm. 2 (2021)</td>
<td>Ergodic under the behaviour policy</td>
<td>Trajectory data of an infinite length</td>
<td>Full rank and $|\Phi| &lt; C(\eta, |P|<em>{D</em>{\pi}})$</td>
</tr>
</tbody>
</table>

Table 6. Comparison of assumptions among analysis of target networks under linear function approximation.

<table>
<thead>
<tr>
<th>Work</th>
<th>Learning Rate</th>
<th>Target Network Hyperparameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours Thm. 3.2</td>
<td>$\eta &lt; \frac{1}{\mu(MM + D_\pi)}$</td>
<td>$m \geq \bar{m}$</td>
</tr>
<tr>
<td>Lee and He Thm. 1 (2019)</td>
<td>Decaying learning rate $\alpha_t &gt; 0$ such that $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 &lt; \infty$</td>
<td>Share the learning rate with the student or original parameter</td>
</tr>
<tr>
<td>Asadi et al. Prop. 1 (2023)</td>
<td>$\eta = 1$</td>
<td>$m = \infty$</td>
</tr>
<tr>
<td>Asadi et al. Prop. 1 (2023)</td>
<td>$\eta = \frac{1}{\lambda_{\text{max}}(\Phi' D \Phi) + \lambda_{\text{max}}(\Phi' D \Phi)}$</td>
<td>$m \geq 1$</td>
</tr>
<tr>
<td>Fellows et al. (2023) Thm. 2</td>
<td>Decaying learning rate $\alpha_t &gt; 0$ such that $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 &lt; \infty$</td>
<td>None</td>
</tr>
<tr>
<td>Fellows et al. (2023) Thm. 4</td>
<td>$\frac{1}{\eta} &gt; \frac{\lambda_{\text{min}}(\Phi' D \Phi) + \lambda_{\text{max}}(\Phi' D \Phi)}{2}$</td>
<td>$m &gt; \bar{m}$</td>
</tr>
<tr>
<td>Shangtong et al. Thm. 2 (2021)</td>
<td>Decaying learning rate $\alpha_t &gt; 0$ such that $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 &lt; \infty$</td>
<td>Decaying learning rate $\beta_t &gt; 0$ for the target network such that $\sum_{t=0}^{\infty} \beta_t = \infty$, $\sum_{t=0}^{\infty} \beta_t^2 &lt; \infty$ and for some $d &gt; 0$, $\sum_{t=0}^{\infty} (\beta_t / \alpha_t)^d &lt; \infty$</td>
</tr>
</tbody>
</table>
Figure 5. Black blocks are walls which cannot be trespassed, green ones are hallways and the purple block is the terminal state with +1 reward. Each state has $(x, y)$ coordinate and actions include up, down, left and right.

A.8. Comparison Between Convergence Conditions

The size depends on the feature norm, reward norm and the regularization weight. For some dependent constant $C$ on the regularization weight $\eta$ and transition norm.

\[ \bar{m} = 1 + \frac{\log(1 - \gamma) - \log((1 + \gamma)\sqrt{k})}{\log(1 - \eta\lambda_{\min}(MM^T D_k))} \]

When regularizing the infinity norm of $NM^\dagger$.

\[ \bar{m} = 1 + \frac{\log(1 - \|\bar{J}_{FPE}\|) - \log(\|\bar{J}_{FPE}\| + \|\bar{J}_{TD}\|)}{\log(1 - \eta\lambda_{\min}(\Phi^T D\Phi))} \]

where $\|\bar{J}_{FPE}\| = \|\Phi^T D\Phi\|^{-1}(\gamma \Phi^T DP_\pi \Phi)$ and $\|\bar{J}_{TD}\| = \|I - \eta\Phi^T D(I - \gamma P_\pi \Phi)\|$. 

\[ V_{\alpha S \sim d, A \sim \mu, S', A' \sim P_\pi}(\phi(S, A)(r(s, a) + \gamma\phi(S', A')^\top \theta - \phi(S, A)^\top \theta)) \]

is bounded.