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 an 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-

[1]Department of Computing Science, University of Alberta [2]School of Data Science, The Chinese University of Hong Kong, Shenzhen [3]Google DeepMind [4]School of Computational Science and Engineering, Georgia Tech [5]Figure [6]The work was done while the author was at Google. [7]Uber [8]CIFAR AI Chair, Amii. Fengdi Che <fengdi@ualberta.ca> Correspondence to: Chenjun Xiao <chenjunx@cuhk.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

teed 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, $\frac{2}{1-\gamma}\inf_\theta\|\Phi\theta - q_\pi\|_\infty$, 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 $\mathbb{1}$ 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

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

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

$$\mathcal{T}_\pi q(s,a) = r(s,a) + \gamma \sum_{s',a'\in\mathcal{S}\times\mathcal{A}} P_\pi(s',a'|s,a)q(s',a'),$$

with the state-action transition distribution under π 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 $q_\pi, r \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and the transition function by $P_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$. Then, we define the *Bellman operator* on any vector q as

$$\mathcal{T}_\pi q = r + \gamma P_\pi q\ .$$

It is known that the value function satisfies the *Bellman equation* $q_\pi = \mathcal{T}_\pi q_\pi$.

Linear Function Approximation In this work, we focus on linear function approximation, $q_\pi(s, a) \approx \phi(s, a)^\top\theta$, 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 $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d}$ 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 $q_\pi \approx \Phi\theta$ for some parameter θ. Finally, we assume that Φ 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 $\mathcal{D}$ 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 λ, then receiving the reward $r_i = r(s_i, a_i)$ and next state $s_i \sim P(\cdot|s_i, a_i)$ from the environment. The problem is known as on-policy if the data collection distribution λ is the stationary distribution of π 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 $\mathcal{D}$, let $n(s,a) = \sum_{i=1}^{n} \mathbb{1}[s_i = s, a_i = a]$ 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 $\{(s_i,a_i)\}_{i=1}^{k} \subseteq \mathcal{S} \times \mathcal{A}$ denote the state-actions with $n(s_i,a_i) > 0$, where $k = \sum_{s,a} \mathbb{1}[n(s,a) > 0]$ represents the count of state-action pairs with observed outgoing transitions. Following these definitions, we can define a mask matrix $H \in \mathbb{R}^{k \times |\mathcal{S}||\mathcal{A}|}$, and the empirical distribution of observed data $D_k \in \mathbb{R}^{k \times k}$ as:

$$H = \begin{bmatrix} \mathbf{1}_{x_1}^\top \\ \vdots \\ \mathbf{1}_{x_k}^\top \end{bmatrix}, \qquad D_k = \begin{bmatrix} \hat{\lambda}(x_1) & & \\ & \ddots & \\ & & \hat{\lambda}(x_k) \end{bmatrix}, \tag{1}$$

where $\mathbf{1}_{(s_i,a_i)} \in \{0,1\}^{|\mathcal{S}||\mathcal{A}|}$ is an indicator vector such that $\phi(s_i,a_i) = \mathbf{1}_{(s_i,a_i)}^\top \Phi$.

To evaluate the value of a target policy π, 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 $\hat{P}_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ can then be defined for all s', a' as:

$$\hat{P}_\pi(s',a'|s,a) = \frac{\sum_{i=0}^{n} \mathbb{1}[s_i = s, a_i = a, s_i' = s', a_i' = a']}{n(s,a)}, \tag{2}$$

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:

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

where $M = H\Phi \in \mathbb{R}^{k \times d}$ denotes the predecessor features observed in the offline data, $N = H\hat{P}_\pi\Phi \in \mathbb{R}^{k \times d}$ 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,

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

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 η such that the parameter of OTD updates converges to

$$\theta_{\text{TD}}^* = M^\dagger (I - \gamma NM^\dagger)^{-1} R,$$

when the initial parameter of OTD equals zero.

It can be verified that the OTD fixed point θ_{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 $\theta_{\text{targ},t} \in \mathbb{R}^d$ be the *target parameter* at iteration t. We refer to the function

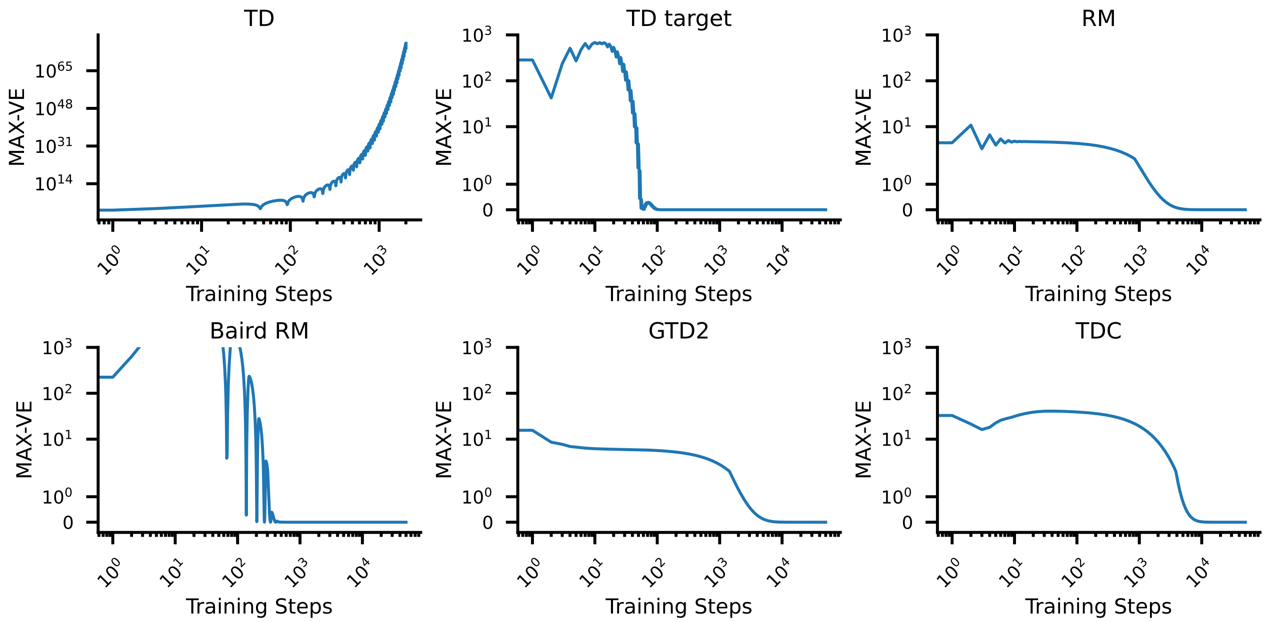

Figure 1. On Baird counterexample, states are sampled from a uniform distribution, and there exists only one action at each state. The discount factor is set to be $\gamma = 0.95$. We plot the maximal value prediction error among all states for OTD, OTTD, RM and GTD-algorithms. Other than OTD, the value errors converge to zero for the rest algorithms. OTTD avoids the divergence of TD and slow convergence rate of others.

approximation's original parameter θ_t as the *student parameter*. Over-parameterized Target TD (OTTD) considers the following update rule using the entire offline dataset:

$$\theta_{t+1} = \theta_t - \eta M^\top D_k \left[M\theta_t - (R + \gamma N \theta_{\mathrm{targ}, m\lfloor \frac{t}{m} \rfloor}) \right]. \tag{5}$$

That is, a target parameter θ_{targ} is introduced to provide bootstrapping targets for TD updates. The target parameter is initialized with the student parameter $\theta_{\mathrm{targ},0} = \theta_0$, 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,

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

Our next result characterizes the convergence of OTTD.

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

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

there exists a learning rate η 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

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

That is, we are considering the hard target network update popularized by (Mnih et al. 2015) instead of Polyak averaging.

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 $NM^\dagger$. For instance, when bounding the infinity norm, $\bar{m} = 1 + \lceil \frac{\log(1-\gamma) - \log((1+\gamma)\sqrt{k})}{\log(1 - \eta \lambda_{\min}(MM^T D_k))} \rceil$.

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 $I - \eta M^\top D_k (M - \gamma N)$ to be bounded by one. Central to the convergence of OTTD is the role of $NM^\dagger$, which represents the projection coefficient of each row of N onto the row space of M. When applying bootstrapping, values are extrapolated using $NM^\dagger M\theta$, based on these projected coefficients. When the infinity norm of $NM^\dagger$ is bounded by one, we have $\|NM^\dagger M\theta\|_\infty \leq \|M\theta\|_\infty$. Thus, the condition on the norm of $NM^\dagger$ 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

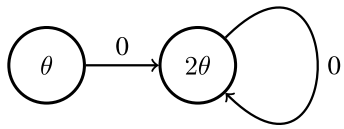

Figure 2. In this example, each state has exactly one action, and rewards are labeled next to the transitions. The value functions are parameterized by a scalar parameter θ, and the features are shown in the graph. This counterexample demonstrates a task with fixed transition probabilities and rewards where our convergence condition is satisfied. However, the conditions for under-parameterized target TD fail for certain data distributions.

space. Off-policy learning means that we do not constrain the data distribution λ 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:
$$\theta_{t+1} = \theta_t - \eta\Phi^\top D\left[\Phi\theta_t - (R + \gamma P_\pi \Phi\theta_t)\right].$$
- Target TD expected update:
$$\theta_{t+1} = \theta_t - \eta\Phi^\top D\left[\Phi\theta_t - (R + \gamma P_\pi \Phi\theta_{\text{targ},m\lfloor \frac{t}{m} \rfloor})\right].$$

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 concentrates on the right state with self-loop. For any discount factor $\gamma > 0.5$, the off-policy state distribution $(\frac{4\gamma-4}{2\gamma-3}, \frac{1-2\gamma}{2\gamma-3})$ causes the fixed point to be $(\Phi^\top D(I-\gamma P)\Phi)^{-1}\Phi^\top DR = \frac{0}{0}$, 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).

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?	$\lim_{t\to\infty}\|C\|^{\frac{1}{t}}$
TD	No	1.12
Target TD	Yes	$1-3.8\text{e}{-3}$
RM	Yes	$1-1.9\text{e}{-5}$
GTD2	Yes	$1-4.5\text{e}{-6}$

Adding over-parameterization alone does not ensure convergence either. As prove in Proposition 3.1, the convergence of OTD still requires that $\rho(I-\eta\Phi^T D(\Phi-\gamma P_\pi\Phi)) < 1$, 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 η 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*
$$\theta^*_{\text{TD}} = \Phi^\dagger(I-\gamma P_\pi)^{-1}R + (I-\Phi^\dagger\Phi)\theta_0,$$
where θ_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 $\theta_t - \theta^* = C^t(\theta_0 - \theta^*)$. To gauge the convergence speed of these algorithms, Schoknecht and Merke (2003) introduced a metric defined as $\lim_{t\to\infty}\|C_{\text{target}}\|^{\frac{1}{t}}$. 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.

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 $\|NM^\dagger\|_\infty < 1$, with probability at least $1-\delta$, the fixed point of OTTD θ^*_{TD} learnt with the dataset $\mathcal{D}$ gives the following value prediction error bound*

$$\|\Phi\theta^*_{\text{TD}} - q_\pi\|_\infty \leq \epsilon_{stat} + \epsilon_{projection} + \epsilon_{approx}, \tag{6}$$

where error terms are defined as follows:

- *Statistical error ϵ_{stat} equals*
$$\epsilon_{stat} = \frac{\|\Phi M^\dagger\|_\infty}{(1-\gamma)^2}\sqrt{\frac{\log(\frac{2k|\mathcal{A}|}{\delta})}{\min_{(s,a)} n(s,a)}}.$$
- *Projection error $\epsilon_{projection}$ is defined as*
$$\epsilon_{projection} = \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi(I - M^\dagger M)\theta^*\|_\infty.$$
- *Approximation error ϵ_{approx} is given by*
$$\epsilon_{approx} = \frac{2\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi\theta^* - q_\pi\|_\infty.$$

Here, $\theta^* = \arg\max_\theta \|\Phi\theta - q_\pi\|_\infty$ is the optimal parameter minimizing the difference between $\Phi\theta$ and q_π, ϵ_{stat} counts in the estimation error of the MDP dynamics by the dataset, and $\epsilon_{projection}$ 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 $N = P_\pi\Phi$, remains off-policy. The fixed point of OTTD $\theta^*_{\text{TD}} = \Phi^\dagger(I - \gamma P_\pi)^{-1} r$, gives the following error bound.

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

$$\|\Phi\theta^*_{\text{TD}} - q_\pi\|_\infty \leq \frac{2}{1-\gamma}\inf_\theta \|\Phi\theta - q_\pi\|_\infty. \tag{7}$$

The value prediction error bound of off-policy OTTD closely aligns with the on-policy results obtained using under-parameterized 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_π of the target policy π. This key observation underscores the remarkable ability of over-parameterized models to learn irrespective of the underlying data distribution λ. To elaborate, Tsitsiklis and Van Roy (1997) have shown that on-policy TD fixed point θ^*_{under} in the under-parameterized 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 $\mathcal{D} = \{\tau_j\}_{j=1}^{n'}$ consists of n' trajectories. Each trajectory τ_j is a sequence of state-action-reward tuples sampled under a behavior policy μ, defined as $\tau_j = \{(s^j_t, a^j_t, r^j_t, s^j_{t+1})\}_{t=0}^{T_j-1}$, with T_j indicating the length of the trajectory. In this context, $a^j_t \sim \mu(\cdot|s^j_t)$ and $s^j_{t+1} \sim P(\cdot|s^j_t, a^j_t)$ is generated by the MDP.

We handle the final states of each collected trajectory $\{s^j_{T_j}\}_{j=1}^{n'}$, by implementing a looping mechanism that sets these states to transition back to themselves with zero reward. Specifically, an additional transition $(s^j_{T_j}, a^j_{T_j}, 0, s^j_{T_j})$ is appended to each trajectory τ_j, with $a^j_{T_j} \sim \mu(\cdot|s^j_{T_j})$. 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 $\{(s_i, a_i, r_i, s'_i, a'_i)\}_{i=1}^n$.

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

$$\frac{\sum_{i=0}^n \rho(a'|s')\mathbb{1}[s_i = s, a_i = a, s'_i = s', a'_i = a']}{n(s,a)}, \tag{8}$$

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 π from the state-action pair (s,a),

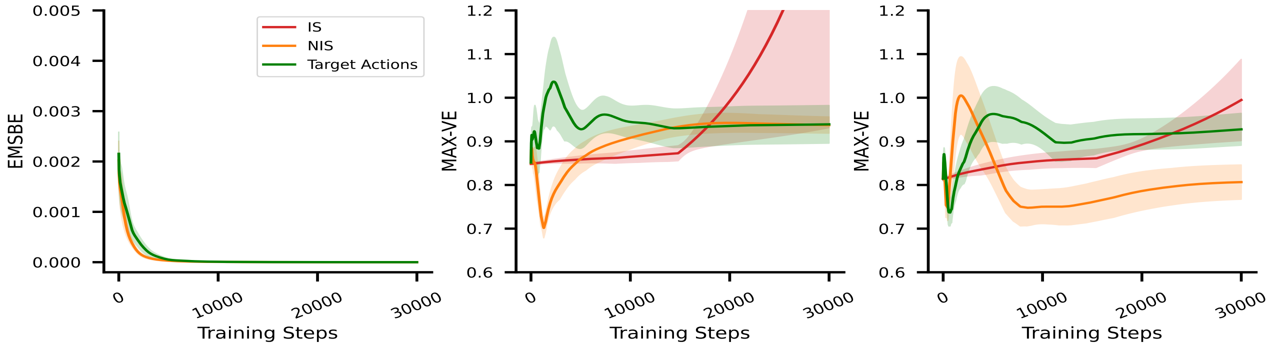

Figure 3. On Four Room, data are sampled as trajectories under the random policy, while the target policy is given by a human player. The left sub-figure shows the training error, EMSBE. The middle and the right sub-figures show the infinity norm of value errors, that is, $\|\Phi\theta^*_{\text{TD}} - q_\pi\|_\infty$. Here, the right sub-figure uses a larger dataset and all results are averaged over 10 random seeds. With off-policy data, per-step normalized IS can correct the action distribution and behave similarly to sampling actions from the target policy. Due to less variance, normalized IS avoids divergence of IS correction.

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 $\hat{P}_{\pi,\text{NIS}}(s', a'|s, a)$ for $n(s, a) > 0$, is defined as

$$\frac{\sum_{i=0}^{n} \rho(a'|s')\mathbb{1}[s_i = s, a_i = a, s'_i = s', a'_i = a']}{\sum_{i=1}^{n}\sum_{\tilde{s},\tilde{a}} \rho(\tilde{a}|\tilde{s})\mathbb{1}[s_i = s, a_i = a, s'_i = \tilde{s}, a'_i = \tilde{a}]}; \quad (9)$$

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 $\hat{P}_\pi(s', a'|s, a)$ 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 $N_{\text{NIS}} = H\hat{P}_{\pi,\text{NIS}}\Phi$ be the next state-action feature matrix under the NIS transition estimate $\hat{P}_{\pi,\text{NIS}}$. The EMSBE can then be estimated as

$$\text{EMSBE}_{\text{NIS}}(\theta) = \frac{1}{2}\|R + \gamma N_{\text{NIS}}\theta - M\theta\|^2_{D_k}. \quad (10)$$

The update rule with NIS correction is given by

$$\theta_{t+1} = \theta_t - \eta M^T D_k \left[M\theta_t - (R + \gamma N_{\text{NIS}}\theta_{\text{targ},m\lfloor\frac{t}{m}\rfloor})\right]. \quad (11)$$

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 $\hat{P}_\pi$ 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 η and integer $\bar{m}$ such that for all update window sizes for target parameters $m \geq \bar{m}$, the OTTD update converges to*

$$\theta^*_{\text{TD,NIS}} = M^\dagger(I - \gamma N_{\text{NIS}}M^\dagger)^{-1}R + (I - M^\dagger M)\theta_0,$$

where θ_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 ϵ_{stat}. It is adjusted due to the new transition probability estimator with NIS corrections.

Corollary 4.3. *Given a dataset $\mathcal{D}$ of episodic trajectory data collected under a behaviour policy μ, when the following condition holds*

- *μ covers the support of the target policy π,*

with probability at least $1 - \delta$, the fixed point of OTTD with NIS correction θ^*_{TD} gives the following value prediction error bound

$$\|\Phi\theta^*_{\text{TD}} - q_\pi\|_\infty \leq \epsilon'_{stat} + \epsilon_{projection} + \epsilon_{approx}, \quad (12)$$

where the projection error $\epsilon_{projection}$ and the approximation error ϵ_{approx} are the same as in Theorem 3.5.

The statistical error ϵ'_{stat} is defined as:

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

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

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 part of the entire state-action space, so generalization is required. Further details are provided in A.7.

As shown in Figure 3, the IS correction, marked in red, can bootstrap on large values due to its high variance, which causes the condition to break, leading to divergence. By contrast, the other two OTTD methods successfully show zero EMSBE, as depicted in the left sub-figure. The first method, in green and detailed in Section 3, selects the next action from the target policy, $a'_i \sim \pi(\cdot|s'_i)$. The second method, in orange and covered in Section 4, leverages per-step normalized importance sampling (NIS) correction. The comparison of the infinity norm of value errors, $\|\Phi\theta^*_{\mathrm{TD}} - q_\pi\|_\infty$, between these two methods is illustrated in the middle and right sub-figures of Figure 3, with the latter sub-figure utilizing a larger dataset. These two methods exhibit equal performance, except as the dataset size increases, the error associated with NIS correction marginally decreases, a trend not observed with the target action sampling approach.

5. Offline Control with Over-parameterized Target Q-learning

The previous convergence result can be extended to 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 Φ_i to present features for the state $s_i \in \mathcal{S}$ 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, $\theta_{\mathrm{targ},nm} = \theta_{nm}$, 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

$$q_t(s,a) = \phi(s,a)^\top \theta_{\mathrm{targ},m\lfloor \frac{t}{m} \rfloor},$$
$$y_t(s,a) = R(s,a) + \gamma \sum_{s'} \hat{P}(s'|s,a) \max_{\text{seen } a'} q_t(s',a'),$$

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

$$\theta_{t+1} = \theta_t - \eta \sum_{s,a} \hat{\lambda}(s,a)\phi(s,a)(\phi(s,a)^\top \theta_t - y_t(s,a)). \tag{13}$$

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

- $\|\Phi_i^\top M^\dagger\|_\infty < \frac{1}{\gamma}$, *for* $i = 1, \cdots, |\mathcal{S}|$,

there exists an $\bar{m}$ such that for all update window sizes of the target parameters $m \geq \bar{m}$, the OTQ update converges to $\theta^* = M^\dagger \hat{q}^* + (I - M^\dagger M)\theta_0$, where $\hat{q}^* \in \mathbb{R}^k$ satisfies

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

and θ_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 exiting transition, their optimal Q-values cannot be evaluated through bootstrapping but only projecting. The term $\Phi_i M^\dagger$, for $i = 1, \cdots, |\mathcal{S}|$, 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 $\bar{m}$ such that for all update window sizes of the target parameters $m \geq \bar{m}$, the parameter of OTQ converges to $\theta^* = M^\dagger \hat{q}^* + (I - M^\dagger M)\theta_0$, where*

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

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 selections (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

Asadi, K., Sabach, S., Liu, Y., Gottesman, O., and Fakoor, R. Td convergence: An optimization perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Boyan, J. A. Least-squares temporal difference learning. In *ICML*, pp. 49–56, 1999.

Che, F., Vasan, G., and Mahmood, A. R. Correcting discount-factor mismatch in on-policy policy gradient methods. In *International Conference on Machine Learning*, volume 202, pp. 4218–4240. PMLR, 2023.

Chen, Z., Clarke, J.-P., and Maguluri, S. T. Target network and truncation overcome the deadly triad in-learning. *SIAM Journal on Mathematics of Data Science*, 5(4): 1078–1101, 2023.

Dayan, P. The convergence of td (λ) for general λ. *Machine learning*, 8:341–362, 1992.

Fellows, M., Smith, M. J., and Whiteson, S. Why target networks stabilise temporal difference methods. In *International Conference on Machine Learning*, pp. 9886–9909. PMLR, 2023.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. PMLR, 2018.

Gelada, C. and Bellemare, M. G. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3647–3655, 2019.

Ghiassian, S., Patterson, A., Garg, S., Gupta, D., White, A., and White, M. Gradient temporal-difference learning with regularized corrections. In *International Conference on Machine Learning*, pp. 3524–3534. PMLR, 2020.

Gordon, G. J. Stable function approximation in dynamic programming. In *Machine learning proceedings 1995*, pp. 261–268. Elsevier, 1995.

Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *International Conference on Machine Learning*, pp. 1372–1383. PMLR, 2017.

He, J., Che, F., Wan, Y., and Mahmood, A. R. Loosely consistent emphatic temporal-difference learning. In *Uncertainty in Artificial Intelligence*, pp. 849–859. PMLR, 2023.

Hesterberg, T. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2): 185–194, 1995.

Hu, X., Ma, Y., Xiao, C., Zheng, Y., and Meng, Z. In-sample policy iteration for offline reinforcement learning. *arXiv preprint arXiv:2306.05726*, 2023.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

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*, pp. 521–528, 2009.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2021.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Kuzborskij, I., Vernade, C., Gyorgy, A., and Szepesvári, C. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pp. 640–648. PMLR, 2021.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4: 1107–1149, 2003.

Lee, D. and He, N. Target-based temporal-difference learning. In *International Conference on Machine Learning*, pp. 3713–3722. PMLR, 2019.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lin, L.-J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8:293–321, 1992.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

Maei, H. R. *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

Mahadevan, S., Liu, B., Thomas, P., Dabney, W., Giguere, S., Jacek, N., Gemp, I., and Liu, J. Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces. *arXiv preprint arXiv:1405.6757*, 2014.

Mahmood, A. R., Van Hasselt, H. P., and Sutton, R. S. Weighted importance sampling for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 27, 2014.

Mahmood, A. R., Yu, H., and Sutton, R. S. Multi-step off-policy learning without importance sampling ratios. *arXiv preprint arXiv:1702.03006*, 2017.

Manek, G. and Kolter, J. Z. The pitfalls of regularization in off-policy td learning. *Advances in Neural Information Processing Systems*, 35:35621–35631, 2022.

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*, 518(7540): 529–533, 2015.

Nachum, O., Chow, Y., Dai, B., and Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019a.

Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*, 2019b.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Precup, D., Sutton, R. S., and Dasgupta, S. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

Schoknecht, R. and Merke, A. Td (0) converges provably faster than the residual gradient algorithm. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 680–687, 2003.

Sharoff, P., Mehta, N., and Ganti, R. A farewell to arms: Sequential reward maximization on a budget with a giving up option. In *International Conference on Artificial Intelligence and Statistics*, pp. 3707–3716. PMLR, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., Maei, H., and Szepesvári, C. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in neural information processing systems*, 21, 2008.

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 annual international conference on machine learning*, pp. 993–1000, 2009.

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*, 17 (1):2603–2631, 2016.

Thomas, V. On the role of overparameterization in off-policy temporal difference learning with linear function approximation. *Advances in Neural Information Processing Systems*, 35:37228–37240, 2022.

Tsitsiklis, J. and Van Roy, B. An analysis of temporal-difference learning with function approximationtechnical. *Rep. LIDS-P-2322). Lab. Inf. Decis. Syst. Massachusetts Inst. Technol. Tech. Rep*, 1996.

Van Hasselt, H. P. *Insights in reinforcement learning*. Hado van Hasselt, 2011.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8:279–292, 1992.

Xiao, C., Dai, B., Mei, J., Ramirez, O. A., Gummadi, R., Harris, C., and Schuurmans, D. Understanding and leveraging overparameterization in recursive value estimation. In *International Conference on Learning Representations*, 2021.

Xiao, C., Wang, H., Pan, Y., White, A., and White, M. The in-sample softmax for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33: 6551–6561, 2020.

Yu, H. and Bertsekas, D. P. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.

Zhang, S., Liu, B., and Whiteson, S. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020.

Zhang, S., Yao, H., and Whiteson, S. Breaking the deadly triad with a target network. In *International Conference on Machine Learning*, pp. 12621–12631. PMLR, 2021.

A. Appendix

A.1. Over-parameterized Target TD Convergence Proof

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

Proposition A.1. *A sufficient condition for the matrix $I - \eta M^\top D_k(M - \gamma N)$ 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, $MM^\top = I$ is the identity matrix and $NM^\dagger = H\hat{P}H^\top$.

Thus, the spectral radius equals

$$
\begin{aligned}
\rho(I - \eta(M - \gamma N)M^\top D_k) &= \rho(I - \eta D_k(I - \gamma H\hat{P}H^\top)) \\
&= \|I - \eta D_k + \eta D_k \gamma H\hat{P}H^\top\|_\infty \\
&\leq \|I - \eta D_k\|_\infty + \|\eta D_k\|_\infty \|\gamma H\hat{P}H^\top\|_\infty \\
&\leq 1 - \eta \min_{s,a} \hat{\lambda}(s,a) + \gamma\eta \max_{s,a} \hat{\lambda}(s,a) \qquad = 1 - \eta\frac{1}{k} + \gamma\eta\frac{1}{k} < 1.
\end{aligned}
$$

□

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

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

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

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

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 $\frac{1-\gamma}{1+\gamma}\frac{1}{\|U\|\|U^{-1}\|}$. Therefore, the norm of A is bounded as

$$
\begin{aligned}
\|A\| &\leq \|U\|\|(I - \eta\Lambda)^m\|\|U^{-1}\| \\
&< \frac{1-\gamma}{1+\gamma}.
\end{aligned}
$$

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

$$
\begin{aligned}
\rho(\gamma W + (I - \gamma W)(I - \eta MM^\top D_k)^m) &\leq \|\gamma W + (I - \gamma W)(I - \eta MM^\top D_k)^m\| \\
&\leq \|\gamma W\| + \|(I - \gamma W)\|\|(I - \eta MM^\top D_k)^m\| \\
&\leq \gamma + (1+\gamma)\|(I - \eta MM^\top D_k)^m\| < 1.
\end{aligned}
$$

□

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

$$
\theta_{(n+1)m} = (I - \eta M^\top BD_k(M - \gamma N))\theta_{nm} + \eta M^\top BD_k R, \tag{15}
$$

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

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

$$\begin{aligned}
\theta_{(n+1)m} &= (I - \eta M^\top D_k M)\theta_{(n+1)m-1} + \eta M^\top D_k (R + \gamma N \theta_{\mathrm{targ},nm}) \\
&= (I - \eta M^\top D_k M)^m \theta_{nm} + \sum_{i=0}^{m-1} (I - \eta M^\top D_k M)^i \eta M^\top D_k (R + \gamma N \theta_{\mathrm{targ},nm}) \\
&= (I - \eta M^\top D_k M)^m \theta_{nm} + \sum_{i=0}^{m-1} (I - \eta M^\top D_k M)^i \eta M^\top D_k (R + \gamma N \theta_{nm}).
\end{aligned}$$

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

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

$$\begin{aligned}
(I - \eta M^\top D_k M)^m &= (I - \eta M^\top D_k M)^{m-1}(I - \eta M^\top D_k M) \\
&= (I - \eta M^\top D_k M)^{m-1} - (I - \eta M^\top D_k M)^{m-1} \eta M^\top D_k M \\
&= (I - \eta M^\top D_k M)^{m-1} - \eta M^\top (I - \eta D_k M M^\top)^{m-1} D_k M \\
&= \cdots \\
&= I - \eta M^\top \sum_{i=0}^{m-1} (I - \eta D_k M M^\top)^i D_k M \\
&= I - \eta M^\top B D_k M. \qquad (16)
\end{aligned}$$

Also, notice that sum of matrix power times $M^\top$ can be simplied as

$$\begin{aligned}
\sum_{i=0}^{m-1} (I - \eta M^\top D_k M)^i \eta M^\top &= \sum_{i=0}^{m-1} (\eta M^\top - \eta M^\top D_k M \eta M^\top)^i \\
&= \eta M^\top B.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\theta_{(n+1)m} &= (I - \eta M^\top B D_k M)\theta_{nm} + \sum_{i=0}^{m-1} (I - \eta M^\top D_k M)^i \eta M^\top D_k (R + \gamma N \theta_{nm}) \\
&= (I - \eta M^\top B D_k M)\theta_{nm} + \eta M^\top B D_k (R + \gamma N \theta_{nm}) \\
&= (I - \eta M^\top B D_k (M - \gamma N))\theta_{nm} + \eta M^\top B D_k R.
\end{aligned}$$

□

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 $\rho(\eta MM^\top D_k) < 1$, 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*

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

where θ_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.

$$\begin{aligned}
&\theta_{(n+1)m} = (I - \eta M^\top BD_k(M - \gamma N))\theta_{nm} + \eta M^\top BD_k R \\
&= (I - \eta M^\top BD_k(M - \gamma N))^{n+1}\theta_0 + \sum_{i=0}^{n}(I - \eta M^\top BD_k(M - \gamma N))^i \eta M^\top BD_k R \\
&= \underbrace{(I - \eta M^\top BD_k(M - \gamma N))^{n+1}\theta_0}_{\text{term 1}} + \underbrace{\eta M^\top BD_k \sum_{i=0}^{n}(I - (I - \gamma W)\eta MM^\top BD_k)^i R}_{\text{term 2}}. \qquad (17)
\end{aligned}$$

where $B = \sum_{i=0}^{m-1}(I - \eta D_k MM^\top)^i$. Also, the last line uses that $(M - \gamma N)M^\top = (I - \gamma NM^\dagger)MM^\top$, 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, $\eta MM^\top BD_k$ can be expressed as $I - (I - \eta MM^\top D_k)^m$. Thus, the matrix in the summation can be rewritten as

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

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

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

Notice that B is invertible, since $(I - \eta D_k MM^\top)^\top$ 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, $MM^\top$ is invertible, thanks to the full rank of M. Therefore,

$$\begin{aligned}
&\lim_{n\to\infty}\sum_{i=0}^{n}\eta M^\top BD_k(I - (I - \gamma W)\eta MM^\top BD_k)^i R \\
&= \eta M^\top BD_k(\eta MM^\top BD_k)^{-1}(I - \gamma W)^{-1}R \\
&= M^\dagger(I - \gamma W)^{-1}R. \qquad (19)
\end{aligned}$$

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

$$\begin{aligned}
&(I - \eta M^\top BD_k(M - \gamma N))^{n+1} \\
&= I - \sum_{i=0}^{n}(I - \eta M^\top BD_k(M - \gamma N))^i \eta M^\top BD_k(M - \gamma N) \\
&= I - \sum_{i=0}^{n}\eta M^\top BD_k(I - (I - \gamma W)\eta MM^\top BD_k)^i(M - \gamma N) \\
&\xrightarrow{n\to\infty} I - M^\top BD_k(MM^\top BD_k)^{-1}(I - \gamma W)^{-1}(M - \gamma NM^\dagger M - \gamma N(I - M^\dagger M)) \\
&= I - M^\dagger M - M^\dagger(I - \gamma W)^{-1}\gamma N(I - M^\dagger M).
\end{aligned}$$

The second line uses the same recursive trick in Equation 16.

In conclusion, the parameter θ_{nm} 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 η 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*

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

where θ_0 is the initial parameter of OTTD.

Proof. These updates assume ideal offline data encompassing all state-action pairs such that $k = |\mathcal{S}||\mathcal{A}|$, $M = \Phi$, $N = P_{\pi}\Phi$ and $D = \text{diag}(\lambda)$. It can be easily verified that for the expected update, we have $NM^{\dagger} = P_{\pi}\Phi\Phi^{\dagger} = P_{\pi}$, 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 $\mathcal{D}$ and the optimal fixed point $\theta^{*}_{TD} = M^{\dagger}(I-\gamma NM^{\dagger})^{-1}R$, when $\|NM^{\dagger}\|_{\infty} < 1$, with probability at least $1-\delta$, for $\theta^{*} = argmin_{\theta}\|\Phi\theta - q\|_{\infty}$,*

$$\|\Phi\theta^{*}_{TD} - q\|_{\infty} \tag{20}$$

$$\leq \frac{\|\Phi M^{\dagger}\|_{\infty}}{(1-\gamma)^2\sqrt{\min_{s,a} n(s,a)}}\sqrt{\log(\frac{2k|\mathcal{A}|}{\delta})} + \frac{\|\Phi M^{\dagger}\|_{\infty}}{1-\gamma}\|\Phi(I-M^{\dagger}M)\theta^{*}\|_{\infty} + \frac{2\|\Phi M^{\dagger}\|_{\infty}}{1-\gamma}\|\Phi\theta^{*}-q\|_{\infty}. \tag{21}$$

Proof.

$$\begin{aligned}\|\Phi\theta^{*}_{TD} - q\|_{\infty} &= \|\Phi M^{\dagger}M\theta^{*}_{TD} - q\|_{\infty} \\ &\leq \underbrace{\|\Phi M^{\dagger}M\theta^{*}_{TD} - \Phi M^{\dagger}Hq\|_{\infty}}_{\text{term 1}} + \underbrace{\|\Phi M^{\dagger}Hq - q\|_{\infty}}_{\text{term 2}}.\end{aligned} \tag{22}$$

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

$$\begin{aligned}&\|\Phi M^{\dagger}Hq - q\|_{\infty} \\ &= \|\Phi M^{\dagger}Hq - \Phi M^{\dagger}M\theta^{*} + \Phi\theta^{*} - q - \Phi(I-M^{\dagger}M)\theta^{*}\|_{\infty} \\ &\leq \|\Phi M^{\dagger}H(q-\Phi\theta^{*})\|_{\infty} + \|\Phi\theta^{*}-q\|_{\infty} + \|\Phi(I-M^{\dagger}M)\theta^{*}\|_{\infty} \\ &\leq (\|\Phi M^{\dagger}\|_{\infty}+1)\|\Phi\theta^{*}-q\|_{\infty} + \|\Phi(I-M^{\dagger}M)\theta^{*}\|_{\infty}.\end{aligned} \tag{23}$$

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

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 Hq following the Bellman update:

$$\begin{aligned}M\theta^{*}_{TD} &= R + \gamma WM\theta^{*}_{TD} = R + \gamma H\hat{P}\Phi M^{\dagger}M\theta^{*}_{TD}, \\ Hq &= Hr + \gamma HPq = Hr + \gamma H\hat{P}\Phi M^{\dagger}Hq + \gamma H\hat{P}(q - \Phi M^{\dagger}Hq) + \gamma H(P-\hat{P})q.\end{aligned}$$

Next, we can bound the first term in Equation 22 by extending $M\theta^*_{TD}$ and Hq.

$$
\begin{aligned}
&\|\Phi M^\dagger M\theta^*_{TD} - \Phi M^\dagger Hq\|_\infty \\
&\leq \|\Phi M^\dagger(R - Hr + \gamma H(P - \hat{P})q) + \Phi M^\dagger \gamma H \hat{P} \Phi M^\dagger (M\theta^*_{TD} - Hq) + \Phi M^\dagger \gamma H \hat{P}(q - \Phi M^\dagger Hq)\|_\infty \\
&\leq \|\Phi M^\dagger\|_\infty \epsilon_{est} + \gamma \|\Phi M^\dagger\|_\infty A + \gamma \|\Phi M^\dagger\|_\infty B \\
&\leq \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\epsilon_{est} + \frac{\gamma\|\Phi M^\dagger\|_\infty}{1-\gamma} A \\
&\leq \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\epsilon_{est} + \frac{2\gamma\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi\theta^* - q\|_\infty + \frac{\gamma\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi(I - M^\dagger M)^*\|_\infty. \qquad (24)
\end{aligned}
$$

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

$$
\begin{aligned}
&\|\Phi\theta^*_{TD} - q\|_\infty \\
&\leq \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\epsilon_{est} + (\frac{\gamma\|\Phi M^\dagger\|_\infty}{1-\gamma} + 1)\|\Phi(I - M^\dagger M)^*\|_\infty + (\|\Phi M^\dagger\|_\infty + 1 + \frac{2\gamma\|\Phi M^\dagger\|_\infty}{1-\gamma})\|\Phi\theta^* - q\|_\infty \\
&\leq \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\epsilon_{est} + \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi(I - M^\dagger M)\theta^*\|_\infty + \frac{2\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi\theta^* - q\|_\infty, \qquad (25)
\end{aligned}
$$

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

ϵ_{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,

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

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

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 $\frac{2}{1-\gamma}\|\Phi\theta^* - q\|_\infty$ is left in the bound. □

A.3. Nomalized 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:

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

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

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 behaviour 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,

$$
\hat{P}_\pi(s', a'|s, a) = \frac{\sum_{j=1}^{n(s,a)} X_j}{\sum_{j=1}^{n(s,a)} Y_j}. \qquad (28)
$$

Notice that for X_j, the mean equals

$$\mathbb{E}X_j = \rho(a'|s')\mathbb{E}\left[\mathbb{1}[S=s, A=a, S'=s', A'=a']\right] \tag{29}$$
$$= \frac{\pi(a'|s')}{\mu_i(a'|s')}\mathbb{P}(S=s, A=a, S'=s', A'=a') \tag{30}$$
$$= P_\pi(s', a'|s, a). \tag{31}$$

Similarly, we gain that $\mathbb{E}Y_j = 1$.

Denote $M_\rho := \max_i \rho(A'_i|S'_i)$. The variance of X_j and Y_j are bounded by M_ρ^2.

Then by strong large law of number for martingales, $\frac{1}{n(s,a)}\sum_{j=1}^{n(s,a)} X_j \to P_\pi(s', a'|s, a)$ and $\frac{1}{n(s,a)}\sum_{j=1}^{n(s,a)} Y_j \to 1$ almost surely. Since $\mathbb{P}(Y_j > 0) = 1$, their ratio $\hat{P}_\pi(s', a'|s, a) \to P_\pi(s', a'|s, a)$ a.s.. $\square$

Theorem A.8. *Given a batch of trajectories under some behaviour policies, if M has full rank and the learning rate satisfies that $\eta < \frac{1}{\rho(MM^\top D_k)}$, there exists an integer $\bar{m}$ such that for all update steps of the target parameter $m \geq \bar{m}$, the parameter of corrected over-parameterized target TD always converge to*

$$\theta^*_{TD,\mathrm{NIS}} = M^\dagger(I - \gamma N_{\mathrm{NIS}M^\dagger})^{-1}R + (I - M^\dagger M)\theta_0,$$

where θ_0 is the initial point.

Proof. The update rule equals $\theta_{t+1} = \theta_t - \eta M^\top D_k \left[M\theta_t - (R + \gamma N_{\mathrm{NIS}}\theta_{\mathrm{targ},m\lfloor \frac{t}{m} \rfloor})\right]$, 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 $N_{\mathrm{NIS}}M^\dagger = H\hat{P}_{\pi,\mathrm{NIS}}\Phi M^\dagger$ to satisfy the condition of having a norm less than one. Here, W_{NIS} equals the non-zero square matrix in $\hat{P}_\pi$ and is also stochastic. Therefore, the its infinity norm is less than one and the condition is satisfied. $\square$

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.

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 $\bar{Y}$ 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,*

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

Corollary A.10. *Given a dataset $\mathcal{D}$ of episodic trajectory data collected under a behaviour policy μ and the optimal fixed point $\theta^*_{TD} = M^\dagger(I - \gamma H\hat{P}_\pi\Phi M^\dagger)^{-1}R$, when $\|H\hat{P}_\pi\Phi M^\dagger\|_\infty < 1$ and μ covers the support of the target policy π, with probability at least $1 - \delta$,*

$$\begin{aligned}&\|\Phi\theta^*_{TD} - q_\pi\|_\infty \\ &\leq \epsilon'_{stat} + \epsilon_{projection} + \epsilon_{approx}.\end{aligned} \tag{32}$$

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

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

$\epsilon_{projection}$ *and* ϵ_{approx} *are the same as in Theorem 3.5.*

Proof. As shown in Equation 25,

$$\begin{aligned}
&\|\Phi\theta^*_{TD} - q\|_\infty \\
&\le \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\epsilon_{est} + \frac{\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi(I - M^\dagger M)\theta^*\|_\infty + \frac{2\|\Phi M^\dagger\|_\infty}{1-\gamma}\|\Phi\theta^* - q\|_\infty.
\end{aligned} \tag{33}$$

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, $\epsilon_{est} = \|\gamma H(P - \hat{P})q\|_\infty + \|R - Hr\|_\infty$, which is bounded by the above lemma and a union bound.

$$\hat{P}_{\pi,\mathrm{NIS}}(s, a, s', a') = \frac{\sum_{i=0}^{n} \rho(a'|s')\mathbb{1}[S_i = s, A_i = a, S'_i = s', A'_i = a']}{\sum_{i=1}^{n}\sum_{\tilde{s},\tilde{a}} \rho(\tilde{a}|\tilde{s})\mathbb{1}[S_i = s, A_i = a, S'_i = \tilde{s}, A'_i = \tilde{a}]}. \tag{34}$$

As defined in A.3,

$$X = \rho(a'|s')\mathbb{1}[S = s, A = a, S' = s', A' = a'], \tag{35}$$

$$Y = \sum_{\tilde{s},\tilde{a}} \rho(\tilde{a}|\tilde{s})\mathbb{1}[S = s, A = a, S' = \tilde{s}, A' = \tilde{a}]. \tag{36}$$

$\frac{X}{\max_{s',a'}\rho(a'|s')} \in [0, 1]$ and $\frac{Y}{\min_{s',a'}\rho(a'|s')} \in [1, \rho_M]$.

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

$$|(P_{s,a} - \hat{P}_{s,a})^\top q| \le \frac{\rho_M \max\{\rho_M - 1, 1\}}{(1-\gamma)\sqrt{2n(s,a)}} log(\frac{4k|\mathcal{A}|}{\delta}).$$

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

□

A.5. Continuing Tasks' Error Bound

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

$$\|\Phi\theta^*_{TD} - q_\pi\|_{D_\pi} \tag{37}$$

$$\le \epsilon''_{stat} + \epsilon''_{projection} + \epsilon''_{approx}. \tag{38}$$

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

$\epsilon''_{stat} = \frac{C\rho_M}{(1-\gamma)^2\sqrt{\min_{(s,a)} n(s,a)}}\log(\frac{4k|\mathcal{A}|}{\delta})\max\{\rho_M - 1, 1\} + \frac{C}{(1-\gamma)^2}$.

Let θ^* be defined as $\theta^* = argmin_\theta\|\Phi\theta - q_\pi\|_{D_\pi}$. Then, $\epsilon''_{projection} = \frac{C(1+\gamma\|\hat{P}_{\pi,\mathrm{NIS}}\|_{D_\pi})}{1-\gamma}\|\Phi(I - M^\dagger M)\theta^*\|_{D_\pi}$, and

$\epsilon''_{approx} = \frac{C(2+\gamma\|\hat{P}_{\pi,\mathrm{NIS}}\|_{D_\pi})}{1-\gamma}\|\Phi\theta^* - q_\pi\|_{D_\pi}$.

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

$$\begin{aligned}
\|\Phi\theta^*_{TD} - q\|_{D_\pi} &= \|\Phi M^\dagger M\theta^*_{TD} - q\|_{D_\pi} \\
&\le \underbrace{\|\Phi M^\dagger M\theta^*_{TD} - \Phi M^\dagger Hq\|_{D_\pi}}_{\text{term 1}} + \underbrace{\|\Phi M^\dagger Hq - q\|_{D_\pi}}_{\text{term 2}}.
\end{aligned} \tag{39}$$

Define $\theta^* = argmin_\theta \|\Phi\theta - q\|_{D_\pi}$. By adding intermediate term, we can bound the second term

$$\begin{aligned}
&\|\Phi M^\dagger H q - q\|_{D_\pi} \\
&\le \|\Phi M^\dagger H(q - \Phi\theta^*)\|_{D_\pi} + \|\Phi\theta^* - q\|_{D_\pi} + \|\Phi(I - M^\dagger M)\theta^*\|_{D_\pi} \\
&\le (\|\Phi M^\dagger H\|_{D_\pi} + 1)\|\Phi\theta^* - q\|_{D_\pi} + \|\Phi(I - M^\dagger M)\theta^*\|_{D_\pi}.
\end{aligned}$$

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

These three statements are almost the same as in the proof in A.2 except the norm difference.

When bounding ϵ_{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

$$|(P_{s_T,a_T} - \hat{P}_{s_T,a_T})^\top q| \le \frac{2}{(1-\gamma)}. \tag{40}$$

In the norm, this term is weighted by the stationary distribution $d_\pi(s_T, a_T)$. 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. □

A.6. Proof of Q-learning Convergence

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

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

and θ_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

$$\theta_{(n+1)m} = (I - \eta M^\top BD_k M)\theta_{nm} + \eta M^\top BD_k R + \eta M^\top BD_k\gamma\hat{P}\begin{pmatrix}\|\Phi_1\theta_{nm}\|_\infty \\ \|\Phi_2\theta_{nm}\|_\infty \\ \cdots \\ \|\Phi_{\mathcal{S}}\theta_{nm}\|_\infty\end{pmatrix}. \tag{41}$$

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

$$\begin{aligned}
M\theta_{(n+1)m} &= M(I - \eta M^\top BD_k M)\theta_{nm} + \eta MM^\top BD_k R + \eta MM^\top BD_k\gamma\hat{P}\begin{pmatrix}\|\Phi_1\theta_{nm}\|_\infty \\ \|\Phi_2\theta_{nm}\|_\infty \\ \cdots \\ \|\Phi_{\mathcal{S}}\theta_{nm}\|_\infty\end{pmatrix} \\
&= (I - \eta MM^\top BD_k)M\theta_{nm} + \eta MM^\top BD_k R + \eta MM^\top BD_k\gamma\hat{P}\begin{pmatrix}\|\Phi_1 M^\dagger M\theta_{nm}\|_\infty \\ \|\Phi_2 M^\dagger M\theta_{nm}\|_\infty \\ \cdots \\ \|\Phi_{\mathcal{S}} M^\dagger M\theta_{nm}\|_\infty\end{pmatrix}.
\end{aligned} \tag{42}$$

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

$$\mathcal{T}x=(I-\eta MM^\top BD_k)x+\eta MM^\top BD_kR+\eta MM^\top BD_k\gamma\hat{P}\begin{pmatrix}\|\Phi_1M^\dagger x\|_\infty\\ \|\Phi_2M^\dagger x\|_\infty\\ \dots\\ \|\Phi_{\mathcal{S}}M^\dagger x\|_\infty\end{pmatrix}$$

.

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

$$\begin{aligned}&\|\mathcal{T}x-\mathcal{T}x'\|_\infty\\ &=\|(I-\eta MM^\top BD_k)(\bar{x}-x')+\eta MM^\top BD_k\gamma\hat{P}[\begin{pmatrix}\|\Phi_1M^\dagger\bar{x}\|_\infty\\ \|\Phi_2M^\dagger\bar{x}\|_\infty\\ \dots\\ \|\Phi_{\mathcal{S}}M^\dagger\bar{x}\|_\infty\end{pmatrix}-\begin{pmatrix}\|\Phi_1M^\dagger x'\|_\infty\\ \|\Phi_2M^\dagger x'\|_\infty\\ \dots\\ \|\Phi_{\mathcal{S}}M^\dagger x'\|_\infty\end{pmatrix}]\|_\infty\\ &\le\|(I-\eta MM^\top D_k)^m(\bar{x}-x')\|_\infty+\|I-(I-\eta MM^\top D_k)^m\|_\infty\|\gamma\hat{P}\begin{pmatrix}\|\Phi_1M^\dagger(\bar{x}-x')\|_\infty\\ \|\Phi_2M^\dagger(\bar{x}-x')\|_\infty\\ \dots\\ \|\Phi_{\mathcal{S}}M^\dagger(\bar{x}-x')\|_\infty\end{pmatrix}\|_\infty\\ &\le[\|(I-\eta MM^\top D_k)^m\|_\infty+\gamma\|I-(I-\eta MM^\top D_k)^m\|_\infty]\|\Phi'M^\dagger\|_\infty\|\bar{x}-x'\|_\infty.\end{aligned}\tag{43}$$

The third line uses $\eta MM^\top BD_k=I-(I-\eta MM^\top D_k)^m$ following the recursive argument in Equation 16.

Set $\bar{m}=1+\lceil\dfrac{\log(c-\gamma)-\log(1+\gamma)}{\log(1-\eta\lambda_k)}\rceil$ for some constant $c\in(0,1)$. As shown in Lemma A.2, $\|(I-\eta MM^\top D_k)^m\|_\infty<\frac{c-\gamma}{1+\gamma}$. 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

$$\hat{q}^*=R+\gamma\hat{P}\begin{pmatrix}\|\Phi_1M^\dagger\hat{q}^*\|_\infty\\ \|\Phi_2M^\dagger\hat{q}^*\|_\infty\\ \dots\\ \|\Phi_{\mathcal{S}}M^\dagger\hat{q}^*\|_\infty\end{pmatrix},$$

and $\hat{q}^*$ exists uniques 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

$$\begin{aligned}\mathcal{T}\hat{q}^*&=(I-\eta MM^\top BD_k)\hat{q}^*+\eta MM^\top BD_kR+\eta MM^\top BD_k\gamma\hat{P}\begin{pmatrix}\|\Phi_1M^\dagger\hat{q}^*\|_\infty\\ \|\Phi_2M^\dagger\hat{q}^*\|_\infty\\ \dots\\ \|\Phi_{\mathcal{S}}M^\dagger\hat{q}^*\|_\infty\end{pmatrix}\\ &=(I-\eta MM^\top D_k)^m\hat{q}^*+[I-(I-\eta MM^\top D_k)^m][R+\gamma\hat{P}\begin{pmatrix}\|\Phi_1M^\dagger\hat{q}^*\|_\infty\\ \|\Phi_2M^\dagger\hat{q}^*\|_\infty\\ \dots\\ \|\Phi_{\mathcal{S}}M^\dagger\hat{q}^*\|_\infty\end{pmatrix}]\\ &=(I-\eta MM^\top D_k)^m\hat{q}^*+[I-(I-\eta MM^\top D_k)^m]\hat{q}^*\\ &=\hat{q}^*.\end{aligned}$$

The second equation again uses $\eta MM^\top BD_k=I-(I-\eta MM^\top D_k)^m$ following the recursive argument in Equation 16.

Algorithm	Learning Rate	Second Learning Rate	Target Update Step
TD	0.5	None	None
Target TD	0.997	None	3
RM	0.8	None	None
Baird RM	0.95	None	None
GTD2	0.6	0.6	None
TDC	0.6	0.4	None

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 $MM^\dagger M\theta_0$ 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 $t \to \infty$. $\square$

A.7. Empirical Setting

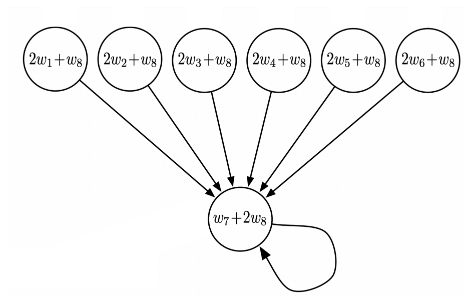

Figure 4. The features are shown in the figure. The transition is labelled with arrows. This Baird Counterexample is a Markov Reward process and only one action is available at each state.

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 ϵ-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.

Algorithm	Learning Rate
Off-Policy without Corrections	0.95
Target Actions	0.97
NIS	0.97
IS	0.02

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

Work	Regularity Condition	Comment
Ours Thm. 3.2	$\|NM^{\dagger}\| < 1$	It is satisfied for expected updates or a batch of complete trajectories naturally.
Lee and He Thm. 1 (2019)	None	No regularization is needed for the on-policy learning.
Asadi et al. Prop. 1 (2023)	$\rho((\Phi^{\top} D\Phi)^{-1}(\gamma\Phi^{\top} DP_{\pi}\Phi)) < 1$	The condition fails on a Two-state counterexample even with expected updates.
Asadi et al. Prop. 5 (2023)	$\frac{\lambda_{max}(\gamma\Phi^{\top} DP_{\pi}\Phi))}{\lambda_{min}((\Phi^{\top} D\Phi)} < 1$	The condition fails on a Two-state counterexample even with expected updates.
Fellows et al. Thm. 2 (2023)	$M^{\top} D_k(\gamma N - M)$ has strictly negative eigenvalues	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.
Fellows et al. Thm. 4 (2023)	$\|(\Phi^{\top} D\Phi)^{-1}(\gamma\Phi^{\top} DP_{\pi}\Phi)\| < 1$	The condition fails on a Two-state counterexample even with expected updates.
Shangtong et al. Thm. 2 (2021)	Projection of the target parameter into a ball and L2 regularization	Projection is hard to realize empirically, and L2 regularization can give a parameter predicting worse than zero values.

Table 4. This table compares how strong the regularity conditions are to ensure convergence in the deadly triad under linear function approximation.

Work	MDP	Data Generation Distribution	Features
Ours Thm. 3.2	None	None	Full rank
Lee and He Thm. 1 (2019)	Ergodic under the target policy π	$s \sim d_\pi$ i.i.d. with $d_\pi(s) > 0$ for all s	Full rank
Asadi et al. Prop. 1 (2023)	None	None	Full rank
Asadi et al. Prop. 5 (2023)	None	None	Full rank
Fellows et al. (2023) Thm. 2	None	$s \sim d$ i.i.d. for some off-policy distribution d	$\\|\phi(s,a)\phi(s,a)^\top\\|$ and $\gamma\\|\phi(s,a)\phi(s',a')^\top\\|$ are bounded, the space of the parameter θ is convex, and variance of the update is bounded
Fellows et al. (2023) Thm. 4	None	$s \sim d$ i.i.d. for some off-policy distribution d	$\\|\phi(s,a)\phi(s,a)^\top\\|$ and $\gamma\\|\phi(s,a)\phi(s',a')^\top\\|$ are bounded, the space of the parameter θ is convex, and variance of the update is bounded
Shangtong et al. Thm. 2 (2021)	Ergodic under the behaviour policy	Trajectory data of an infinite length	Full rank and $\\|\Phi\\| < C(\eta, \\|P_\pi\\|_{D_u})$

Table 5. Comparison of assumptions among analysis of target networks under linear function approximation.

Work	Learning Rate	Target Network Hyperparameter
Ours Thm. 3.2	$\eta < \frac{1}{\rho(MM^T D_k)}$	$m \geq \bar{m}$
Lee and He Thm. 1 (2019)	Decaying learning rate $\alpha_t > 0$ such that $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$	Share the learning rate with the student or original parameter
Asadi et al. Prop. 1 (2023)	$\eta = 1$	$m = \infty$
Asadi et al. Prop. 1 (2023)	$\eta = \frac{1}{\lambda_{max}(\Phi^\top D \Phi)}$	$m \geq 1$
Fellows et al. (2023) Thm. 2	Decaying learning rate $\alpha_t > 0$ such that $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$	None
Fellows et al. (2023) Thm. 4	$\frac{1}{\eta} > \frac{\lambda_{min}(\Phi^\top D\Phi) + \lambda_{max}(\Phi^\top D\Phi)}{2}$	$m > \tilde{m}$
Shangtong et al. Thm. 2 (2021)	Decaying learning rate $\alpha_t > 0$ such that $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$	Decaying learning rate $\beta_t > 0$ for the target network such that $\sum_{t=0}^{\infty} \beta_t = \infty$, $\sum_{t=0}^{\infty} \beta_t^2 < \infty$ and for some $d > 0$, $\sum_{t=0}^{\infty} (\beta_t/\alpha_t)^d < \infty$

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

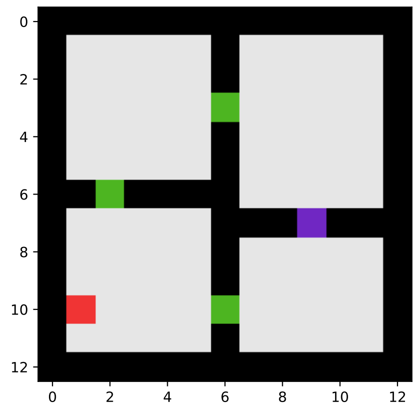

Figure 5. Black blocks are walls which cannot be trespassed, green ones are hallways and the purple block is the terminal state with $+1$ reward. Each state has (x, y) coordinate and actions include up, down, left and right.

A.8. Comparison Between Convergence Conditions

The size depends on the feature norm, reward norm and the regularization weight.

$Var_{S\sim d, A\sim\mu, S'A'\sim P_\pi}(\phi(S,A)(r(s,a)+\gamma\phi(S',A')^\top\theta - \phi(S,A)^\top\theta))$ is bounded.

for some dependent constant C on the regularization weight η and transition norm.

$\bar{m} = 1 + \lceil \frac{\log(1-\gamma) - \log((1+\gamma)\sqrt{k})}{\log(1-\eta\lambda_{\min}(MM^T D_k))} \rceil$ when regularizing the infinity norm of $NM^\dagger$.

$\tilde{m} = 1 + \frac{\log(1 - \|\bar{J}*_{FPE})\|) - \log(\|\bar{J}*_{FPE})\| + \|\bar{J}*_{TD})\|)}{\log(1-\eta\lambda_{\min}(\Phi^\top D\Phi))}$ where $\|\bar{J}*_{FPE})\| = \|(\Phi^\top D\Phi)^{-1}(\gamma\Phi^\top DP_\pi\Phi\|$ and $\|\bar{J}*_{TD})\| = \|I - \eta\Phi^\top D(I-\gamma P_\pi\Phi)\|$.