Assumption Comparison Table

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[1] and L2 regularization	Projection is hard to realize empirically, and L2 regularization can give a parameter predicting worse than zero values.

Table 1. This table compares how strong the regularity conditions are to ensure convergence in the deadly triad under linear function approximation.

[1]

[2] [3]

[4] [5]

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

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

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

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

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

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[2]
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 $\\|\Phi\\| < C(\eta, \\|P_\pi\\|_{D_u})$[3]

Table 2. Comparison of assumptions among analysis of target networks under linear function approximation.

Work	Learning Rate	Target Network Hyperparameter
Ours Thm. 3.2	$\eta < \frac{1}{\rho(MM^T D_k)}$	$m \geq \bar{m}$[4]
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	$\frac{1}{\eta} > \frac{\lambda_{min}(\Phi^\top D\Phi)+\lambda_{max}(\Phi^\top D\Phi)}{2}$	$m > \tilde{m}$[5]
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 3. Comparison of assumptions among analysis of target networks under linear function approximation.

Reference

Lee, D., He, N. (2019, May). Target-based temporal-difference learning. In International Conference on Machine Learning (pp. 3713-3722). PMLR.

Asadi, K., Sabach, S., Liu, Y., Gottesman, O., Fakoor, R. (2023). TD Convergence: An Optimization Perspective. Advances in Neural Information Processing Systems, 36.

Fellows, M., Smith, M. J., Whiteson, S. (2023, July). Why target networks stabilise temporal difference methods. In International Conference on Machine Learning (pp. 9886-9909). PMLR.

Zhang, S., Yao, H., Whiteson, S. (2021, July). Breaking the deadly triad with a target network. In International Conference on Machine Learning (pp. 12621-12631). PMLR.