Supplementary Material

A. Stability in Off-policy n-step TD Learning

In the context of off-policy n-step TD learning, we discuss the possible occurrence of unstable learning in general as well as characterize a safety region in the policy space that guarantees stable learning.

A.1. Unstable Learning

Following the same notations as in the main paper, let state value functions be approximated by a linear function approximation $\boldsymbol{\theta}_t^T \boldsymbol{\phi}(S_t)$ where $\boldsymbol{\phi}$ are feature maps. For the n-step TD target, the value function update on $\boldsymbol{\theta}_t$ is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i (R_{i+1} + \gamma_{i+1}\boldsymbol{\theta}_t^T \boldsymbol{\phi}(S_{i+1}) - \boldsymbol{\theta}_t^T \boldsymbol{\phi}(S_i))\boldsymbol{\phi}(S_t) \tag{24}$$

$$= \boldsymbol{\theta}_t + \alpha \left(\underbrace{\sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i R_{i+1} \boldsymbol{\phi}(S_t)}_{\boldsymbol{b}_t} - \underbrace{\boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T}_{\boldsymbol{A}_t} \boldsymbol{\theta}_t \right) \tag{25}$$

$$= \boldsymbol{\theta}_t + \alpha(\boldsymbol{b}_t - \boldsymbol{A}_t\boldsymbol{\theta}_t) = (\boldsymbol{I} - \alpha\boldsymbol{A}_t)\boldsymbol{\theta}_t + \alpha\boldsymbol{b}_t. \tag{26}$$

To achieve stability as defined in Sutton et al. (2016), we need $\boldsymbol{b}_t$ and $\boldsymbol{A}_t$ to converge to unique fixed points $\boldsymbol{b}$ and $\boldsymbol{A}$, and we need the steady state updates to be stable regardless of the initial parameters $\boldsymbol{\theta}_0$, equivalently requiring $\boldsymbol{A}$ to be a positive-definite matrix.

$$\boldsymbol{A} = \lim_{t\to\infty} \mathbb{E}[\boldsymbol{A}_t] = \lim_{t\to\infty} \mathbb{E}_\mu \boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \tag{27}$$

$$= \sum_s d_\mu(s)\mathbb{E}_\mu \left[\boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_t = s \right] \tag{28}$$

$$= \sum_{i=t}^{t+n-1} \sum_s d_\mu(s)\mathbb{E}_\mu \left[\boldsymbol{\phi}(S_t) \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_t = s \right] \tag{29}$$

$$= \boldsymbol{\Phi}^T \boldsymbol{D}_\mu \left[(\boldsymbol{I} - \boldsymbol{P}_\pi\boldsymbol{\Gamma}) + (\boldsymbol{P}_\pi\boldsymbol{\Gamma} - \boldsymbol{P}_\pi^2\boldsymbol{\Gamma}^2) + \ldots + (\boldsymbol{P}_\pi^{n-1}\boldsymbol{\Gamma}^{n-1} - \boldsymbol{P}_\pi^n\boldsymbol{\Gamma}^n) \right] \boldsymbol{\Phi} \tag{30}$$

$$= \boldsymbol{\Phi}^T \boldsymbol{D}_\mu \left[\boldsymbol{I} - \boldsymbol{P}_\pi^n\boldsymbol{\Gamma}^n \right] \boldsymbol{\Phi}, \tag{31}$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix with diagnal entries $\boldsymbol{\Gamma}_{t,t} = \gamma_t \doteq \gamma(S_t)$. To see the derivation from Eq. 29 to Eq. 30, take one term in the sum indexed by i. We have

$$\sum_s d_\mu(s)\mathbb{E}_\mu\left[\boldsymbol{\phi}(S_t)\prod_{j=t}^{i-1}(\gamma_{j+1}\rho_j)\rho_i\left[\boldsymbol{\phi}(S_i)-\gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T\middle|S_t=s\right] \tag{32}$$

$$=\sum_s d_\mu(s)\boldsymbol{\phi}(s)\sum_{A_t,S_{t+1}\ldots,A_i,S_{i+1}}\prod_{k=t}^{i}\mu(A_k|S_k)p(S_{k+1}|S_k,A_k)\frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}\prod_{j=t}^{i-1}\gamma_{j+1}\left[\boldsymbol{\phi}(S_i)-\gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \tag{33}$$

$$=\sum_s d_\mu(s)\boldsymbol{\phi}(s)\sum_{A_t,S_{t+1}\ldots,A_i,S_{i+1}}\prod_{k=t}^{i}p(S_{k+1}|S_k,A_k)\pi(A_k|S_k)\prod_{j=t}^{i-1}\gamma_{j+1}\left[\boldsymbol{\phi}(S_i)-\gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \tag{34}$$

$$=\sum_s d_\mu(s)\boldsymbol{\phi}(s)\mathbb{E}_\pi\left[\prod_{j=t}^{i-1}\gamma_{j+1}\left[\boldsymbol{\phi}(S_i)-\gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T\middle|S_t=s\right] \tag{35}$$

$$=\sum_s d_\mu(s)\boldsymbol{\phi}(s)\sum_{S_{t+1}}\gamma_{t+1}[\boldsymbol{P}_\pi]_{S_tS_{t+1}}\cdots\sum_{S_i}\gamma_i[\boldsymbol{P}_\pi]_{S_{i-1}S_i}\left[\boldsymbol{\phi}(S_i)-\sum_{S_{i+1}}\gamma_{i+1}[\boldsymbol{P}_\pi]_{S_iS_{i+1}}\boldsymbol{\phi}(S_{i+1})\right]^T \tag{36}$$

$$=\boldsymbol{\Phi}^T\boldsymbol{D}_\mu(\boldsymbol{P}_\pi^{i-t}\boldsymbol{\Gamma}^{i-t}-\boldsymbol{P}_\pi^{i-t+1}\boldsymbol{\Gamma}^{i-t+1})\boldsymbol{\Phi} \tag{37}$$

Back to Eq. 31, the resulting matrix $\boldsymbol{A}=\boldsymbol{\Phi}^T\boldsymbol{D}_\mu\left[\boldsymbol{I}-\boldsymbol{P}_\pi^n\boldsymbol{\Gamma}^n\right]\boldsymbol{\Phi}$ is not necessarily positive definite since the key matrix $\boldsymbol{D}_\mu\left[\boldsymbol{I}-\boldsymbol{P}_\pi^n\boldsymbol{\Gamma}^n\right]$ can be non-positive definite. For example, in the two-state MDP when $n=2$, let the discount $\gamma=0.99$. We know that the steady state distribution following the behavior policy is equal probability of being in either state, and the target policy always goes right, i.e.

$$\boldsymbol{D}_\mu=\begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix},\boldsymbol{P}_\pi=\begin{bmatrix}0 & 1\\ 0 & 1\end{bmatrix}. \tag{38}$$

Hence the key matrix is

$$\boldsymbol{D}_\mu\left[\boldsymbol{I}-\gamma^2\boldsymbol{P}_\pi^2\right]=\begin{bmatrix}0.5 & -0.99^2/2\\ 0 & (1-0.99^2)/2\end{bmatrix}. \tag{39}$$

It is not a positive definite matrix through checking multiplication on both sides by setting $\boldsymbol{\Phi}=(1,2)^T$.

A.2. Safety Guarantee

Recall that the key matrix of the TD(0) algorithm is given by $\boldsymbol{D}_\mu\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]$. We now briefly summarize a few facts about the key matrix from (Sutton et al., 2016). First, the diagonal entries of the key matrix are positive and the off-diagonal entries are negative, so in order to show its positive definiteness it is enough to show that each row sum plus the corresponding column sum is positive. The row sums are all positive because $\boldsymbol{P}_\pi$ is a stochastic matrix and values in $\boldsymbol{\Gamma}$ are smaller than 1. Thus it only remains to show that the column sums are non-negative. The problem is that this is not true for a general $\boldsymbol{D}_\pi,\boldsymbol{D}_\mu$ as was shown in (Sutton et al., 2016). We further showed that this is not true for general n-step TD learning (App. A.1). Yet, we show next that for a distribution $\boldsymbol{D}_\mu$ that is close enough to $\boldsymbol{D}_\pi$ the key matrix is still positive definite. Intuitively, the implication of this result is that doing off-policy learning with $\boldsymbol{D}_\pi\sim\boldsymbol{D}_\mu$ is stable. To show that, we need to show that the column sums of the key matrix are all positive. We begin by lower bounding them as follows:

$$\mathbf{1}^T\boldsymbol{D}_\mu\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]=\boldsymbol{d}_\mu^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right] \tag{40}$$

$$=(\boldsymbol{d}_\pi+\boldsymbol{d}_\mu-\boldsymbol{d}_\mu)^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right] \tag{41}$$

$$=\boldsymbol{d}_\pi^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]+(\boldsymbol{d}_\mu-\boldsymbol{d}_\pi)^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right] \tag{42}$$

$$\geq\boldsymbol{d}_\pi^T(\boldsymbol{I}-\boldsymbol{\Gamma})+(\boldsymbol{d}_\mu-\boldsymbol{d}_\pi)^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]. \tag{43}$$

It remains to show that Eq. (43) has only positive coordinates. The i-th coordinate is given by $\boldsymbol{d}_\pi^T[\boldsymbol{I}-\boldsymbol{\Gamma}]_i + (\boldsymbol{d}_\mu - \boldsymbol{d}_\mu)^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]_i$, where the subscript i denotes the i-th column of a matrix. By Holder inequality, we have that:

$$\boldsymbol{d}_\pi^T[\boldsymbol{I}-\boldsymbol{\Gamma}]_i + (\boldsymbol{d}_\mu - \boldsymbol{d}_\pi)^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]_i \geq \boldsymbol{d}_\pi^T[\boldsymbol{I}-\boldsymbol{\Gamma}]_i - ||\boldsymbol{d}_\mu - \boldsymbol{d}_\pi||_\infty||\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]_i||_1. \tag{44}$$

The first term in the last equation $\boldsymbol{d}_\pi^T[\boldsymbol{I}-\boldsymbol{\Gamma}]_i$ is positive and does not depend on μ. The second term has two contributions. The first one, $||\boldsymbol{d}_\mu - \boldsymbol{d}_\pi||_\infty$ depends on μ and π but it can become as small as we want in the limit that $\mu \to \pi$. The second quantity, $||\left[\boldsymbol{I}-\boldsymbol{P}_\pi\boldsymbol{\Gamma}\right]_i||_1$ depends only on π. This implies that for a fixed π there exists a μ that is close enough to it such that the key matrix is positive definite.

Note that we can repeat this analysis for n-step TD. In this case we need to show that the key matrix $\boldsymbol{D}_\mu\left[\boldsymbol{I}-\boldsymbol{P}_\pi^n\boldsymbol{\Gamma}^n\right]$ is positive definite (see Appendix A.1 above for its derivation). All the derivation we did above for the TD(0) applies to the n-step TD scenario by replacing $\boldsymbol{P}_\pi$ with $\boldsymbol{P}_\pi^n$. The only step which we need to justify is $\boldsymbol{d}_\pi^T\left[\boldsymbol{I}-\boldsymbol{P}_\pi^n\boldsymbol{\Gamma}^n\right] = \boldsymbol{d}_\pi^T(\boldsymbol{I}-\boldsymbol{\Gamma})$, but to see this recall that $\boldsymbol{d}_\pi^T\boldsymbol{P}_\pi = \boldsymbol{d}_\pi^T$, so $\boldsymbol{d}_\pi^T\boldsymbol{P}_\pi^n = \boldsymbol{d}_\pi^T\boldsymbol{P}_\pi\boldsymbol{P}_\pi^{n-1} = \boldsymbol{d}_\pi^T\boldsymbol{P}_\pi^{n-1} = \ldots = \boldsymbol{d}_\pi^T$.

B. WETD Derivation

B.1. TD(λ_t) as mixed n-step TD target

Defining λ_t as in Eq. 8, we write out the off-policy TD(λ_t) learning target by adding the importance sampling correction to Eq. (12.10) in (Sutton & Barto, 2018) for an arbitrarily large integer q

$$\tilde{G}_t \doteq V(S_t) + \sum_{i=t}^{t+q-1}\left(\prod_{j=t}^{i-1}\rho_j\gamma_{j+1}\lambda_j\right)\lambda_i\rho_i\delta_i. \tag{45}$$

In the sum of weighted TD errors from time t to time $(t+q-1)$, the weight $\prod_{j=t}^{i}\lambda_j$ is zero if any of the values $\lambda_j = 0$, and the TD return bootstraps at the first encounter of $\lambda_j = 0$. For simplicity, first consider when $t=0$, the n-step TD corresponds to the truncated return where all terms in the sum are zero for $i \geq n$, requiring $\lambda_n = 0$. At $t=n$, we have that $\lambda_{2n}=0$ produces the n-step TD learning target. In general, this corresponds to $\lambda_j = 0$ whenever j is a multiple of n. To check that this is the mixed update for any k-th sample in the trajectory,

$$\tilde{G}_{t+k} \doteq V(S_{t+k}) + \sum_{i=t+k}^{t+q-1}\left(\prod_{j=t+k}^{i-1}\rho_j\gamma_{j+1}\lambda_j\right)\lambda_i\rho_i\delta_i = V(S_{t+k}) + \sum_{i=t+k}^{t+n-1}\left(\prod_{j=t+k}^{i-1}\rho_j\gamma_{j+1}\right)\rho_i\delta_i, \tag{46}$$

since $t+n$ is the smallest number bigger than t such that $t+n$ is a multiple of n. Thus we recover the mixed n-step update where each sample $V(S_{t+k})$ in the trajectory is updated with $(n-k)$-step TD error.

B.2. TD(λ_t) as mixed V-trace target

In Eq. 46 above, if we set $\lambda_j = \bar{\rho}_j/\rho_j$ as defined in Eq. 18, we recover the mixed V-trace target.

$$\tilde{G}_{t+k} \doteq V(S_{t+k}) + \sum_{i=t+k}^{t+n-1}\left(\prod_{j=t+k}^{i-1}\rho_j\gamma_{j+1}\lambda_j\right)\lambda_i\rho_i\delta_i = V(S_{t+k}) + \sum_{i=t+k}^{t+n-1}\left(\prod_{j=t+k}^{i-1}\bar{\rho}_j\gamma_{j+1}\right)\bar{\rho}_i\delta_i. \tag{47}$$

C. NETD Derivation

In order to simplify notations in the computation below, we denote the NETD trace as F by omitting the superscript (n). Given the possibly unstable asymptotic updates shown in App. A.1, we can modify the updates with F_t to ensure that the

new limit matrix $\boldsymbol{A}$ is positive definite, i.e. to stabilize learning. The F_t-modified parameter update is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha F_t \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i (R_{i+1} + \gamma_{i+1}\boldsymbol{\theta}_t^T \boldsymbol{\phi}(S_{i+1}) - \boldsymbol{\theta}_t^T \boldsymbol{\phi}(S_i))\boldsymbol{\phi}(S_t) \tag{48}$$

$$= \boldsymbol{\theta}_t + \alpha \left(\underbrace{F_t \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i R_{i+1}\boldsymbol{\phi}(S_t)}_{\boldsymbol{b}_t} - \underbrace{F_t \boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \boldsymbol{\theta}_t}_{\boldsymbol{A}_t} \right). \tag{49}$$

Then the $\boldsymbol{A}$ matrix for the emphatically modified n-step TD update becomes

$$\boldsymbol{A} = \lim_{t\to\infty} \mathbb{E}[\boldsymbol{A}_t] = \lim_{t\to\infty} \mathbb{E}_\mu F_t \boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \tag{50}$$

$$= \sum_s d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t \boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_t = s \right] \tag{51}$$

$$= \sum_s d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t | S_t = s\right] \mathbb{E}_\mu \left[\boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_t = s \right]. \tag{52}$$

Eq. 51 is obtained from Eq. 50 by using the linearity of expectation over the learning updates on different states weighted by their steady state visit frequencies. Eq. 52 is obtained from Eq. 51 since conditioned on the state S_t, emphatic trace F_t computed with variables "from the past" is independent from the TD update using variables "in the future". Since the second expectation term in Eq. 52 does not depend on the time step t but only depends on the state value s under the steady state distribution, we take it out of the limit and for clarity, we re-index its time step with a new variable k. This becomes

$$\sum_s \underbrace{d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t | S_t = s\right]}_{f(s)} \mathbb{E}_\mu \left[\boldsymbol{\phi}(S_k) \sum_{i=k}^{k+n-1} \prod_{j=k}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_k = s \right] \tag{53}$$

$$= \sum_s f(s) \mathbb{E}_\mu \left[\boldsymbol{\phi}(S_k) \sum_{i=k}^{k+n-1} \prod_{j=k}^{i-1} (\gamma_{j+1}\rho_j)\rho_i \left[\boldsymbol{\phi}(S_i) - \gamma\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_k = s \right] \tag{54}$$

$$= \boldsymbol{\Phi}^T \boldsymbol{F} (\boldsymbol{I} - \boldsymbol{P}_\pi^n \boldsymbol{\Gamma}^n) \boldsymbol{\Phi}, \tag{55}$$

where $\boldsymbol{F}$ is a diagonal matrix with diagonal elements $f(s) \doteq d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[F_t | S_t = s]$, which we assume exists. Eq. 55 reorganizes terms in Eq. 54 into their matrices notations and forms the telescoping sum as in Eq. 30.

Recall that the sufficient condition for a matrix to be positive definite from Sutton et al. (2016) is to have all positive diagonal entries and all negative off-diagonal entries in the key matrix, and that its row sum plus column sum is positive. The key matrix is $\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{P}_\pi^n \boldsymbol{\Gamma}^n)$ where $\boldsymbol{F}$ is a diagonal matrix with all positive entries on the diagonal, hence with discounts and transition probabilities smaller than 1, the key matrix has positive diagonal entries and all negative entries on the off-diagonal. The row sum is positive since $\boldsymbol{P}_\pi^n$ is a transition matrix with row sum equal to 1 and the discounts are smaller than 1. Hence to make the key matrix $\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{P}_\pi^n \boldsymbol{\Gamma}^n)$ into a positive definite matrix, we just need the columns sum to be positive. Let $\boldsymbol{f}$ be the diagonal entries of $\boldsymbol{F}$. If we define

$$\boldsymbol{f} \doteq \left[\boldsymbol{I} - (\boldsymbol{P}_\pi^T)^n \boldsymbol{\Gamma}^n\right]^{-1} \boldsymbol{d}_\mu, \tag{56}$$

where $\boldsymbol{d}_\mu$ is the vector of diagonal elements of $\boldsymbol{D}_\mu$, then the column sum of the key matrix is

$$\mathbf{1}^T \boldsymbol{F}(\boldsymbol{I} - \boldsymbol{P}_\pi^n \boldsymbol{\Gamma}^n) = \boldsymbol{f}^T (\boldsymbol{I} - \boldsymbol{P}_\pi^n \boldsymbol{\Gamma}^n) \tag{57}$$

$$= \boldsymbol{d}_\mu^T \left[\boldsymbol{I} - (\boldsymbol{P}_\pi)^n \boldsymbol{\Gamma}^n\right]^{-1} (\boldsymbol{I} - \boldsymbol{P}_\pi^n \boldsymbol{\Gamma}^n) \tag{58}$$

$$= \boldsymbol{d}_\mu^T, \tag{59}$$

which is an all positive vector. Therefore $\boldsymbol{F}$ thus defined, the resulting key matrix is positive definite and the steady state learning updates are stable.

In order to apply the emphatic trace to every update, we need to derive the follow-on trace (NETD) at every time step that corresponds to the thus defined $\boldsymbol{F}$ matrix. We show that this following trace gives the above defined $\boldsymbol{F}$ matrix:

$$F_t = \prod_{i=1}^{n} (\gamma_{t-i+1}\rho_{t-i}) F_{t-n} + 1, \text{ with } F_0, F_1, \ldots, F_{n-1} = 1.$$

Take the $n = 2$ for an example, i.e. $F_t = \gamma_t\gamma_{t-1}\rho_{t-1}\rho_{t-2}F_{t-2} + 1, \text{ with } F_0, F_1 = 1$. Recall for any state s,

$$\begin{aligned}
f(s) &\doteq d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[F_t \mid S_t = s] && (60)\\
&= d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[\gamma_t\gamma_{t-1}\rho_{t-1}\rho_{t-2}F_{t-2} + 1 \mid S_t = s] && (61)\\
&= d_\mu(s) + d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[\rho_{t-1}\rho_{t-2}\gamma_t\gamma_{t-1}F_{t-2} \mid S_t = s] && (62)\\
&= d_\mu(s) + d_\mu(s) \sum_{a',s',a'',s''} P_\mu(S_{t-1} = s', A_{t-1} = a' | S_t = s) P_\mu(S_{t-2} = s'', A_{t-2} = a'' | S_{t-1} = s') \cdot \\
&\qquad \frac{\pi(a'|s')}{\mu(a'|s')} \frac{\pi(a''|s'')}{\mu(a''|s'')} \gamma(s)\gamma(s') \lim_{t\to\infty} \mathbb{E}_\mu[F_{t-2} \mid S_{t-2} = s''] && (63)\\
&= d_\mu(s) + \cancel{d_\mu(s)} \sum_{a',s',a'',s''} \frac{\cancel{d_\mu(s')}\mu(a'|s')p(s|s',a')}{\cancel{d_\mu(s)}} \cdot \frac{\cancel{d_\mu(s'')}\mu(a''|s'')p(s'|s'',a'')}{\cancel{d_\mu(s')}}. && \text{(by Bayes Rule)}\\
&\qquad \frac{\pi(a'|s')}{\cancel{\mu(a'|s')}} \frac{\pi(a''|s'')}{\cancel{\mu(a''|s'')}} \frac{\gamma(s)\gamma(s')f(s'')}{\cancel{d_\mu(s'')}} && (64)\\
&= d_\mu(s) + \gamma(s) \sum_{a',s',a'',s''} p(s|s',a')p(s'|s'',a'')\pi(a'|s')\pi(a''|s'')\gamma(s')f(s'') && (65)\\
&= d_\mu(s) + \gamma(s) \sum_{s'} [\boldsymbol{P}_\pi]_{s',s}\gamma(s') \sum_{s''} [\boldsymbol{P}_\pi]_{s'',s'} f(s''). && (66)
\end{aligned}$$

Thus the vector

$$\boldsymbol{f} = \boldsymbol{d}_\mu + \boldsymbol{P}_\pi^{T^2}\boldsymbol{\Gamma}^2\boldsymbol{f} = (\boldsymbol{I} + \boldsymbol{P}_\pi^{T^2}\boldsymbol{\Gamma}^2 + \boldsymbol{P}_\pi^{T^4}\boldsymbol{\Gamma}^4 + \cdots)\boldsymbol{d}_\mu = (\boldsymbol{I} - \boldsymbol{P}_\pi^{T^2}\boldsymbol{\Gamma}^2)^{-1}\boldsymbol{d}_\mu. \tag{67}$$

Since the case of any other positive integer value of n can be derived exactly in the same way, we conclude

$$\boldsymbol{F} = (\boldsymbol{I} - \boldsymbol{P}_\pi^{T^n}\boldsymbol{\Gamma}^n)^{-1}\boldsymbol{D}_\mu. \tag{68}$$

D. Emphatic traces for the V-trace target

Recall that V-trace was developed in (Espeholt et al., 2018) as a method to reduce the variance in importance sampling based off policy policy evaluation. The motivation for V-trace was to correct small off-policy discrepancies that result from parallelizing, i.e., the parameters of the actors lag behind the parameters of the learner. However, the analysis of V-trace was only performed in the tabular MDP setting and not with function approximation. We now show that with linear function approximation, V-trace suffers from the same stability issues as standard IS method, i.e., that the corresponding key matrix is not positive definite. Recall that V-trace truncates the IS ratios in by some constant $\bar{\rho}$, such that $\bar{\rho}_t = \min\{\bar{\rho}, \rho_t\}$. Before we begin we introduce some notation.

We denote the denominator in Eq. 6 by $\nu(s) = \sum_{a'\in\mathcal{A}} \min(\bar{\rho}\mu(a'|s), \pi(a'|s))$. Using this notation, we have that the importance sampling ratio between the true target V-trace policy $\pi_{\bar{\rho}}$ (Eq. 6) and the behaviour policy μ is given by: $\rho_t^v = \frac{\pi_{\bar{\rho}}(A_t|S_t)}{\mu(A_t|S_t)} = \frac{\bar{\rho}_t}{\nu(S_t)}$.

D.1. WEVtrace

We define the emphatic trace for the V-trace update at $n = 1$ as:

$$F_t^v = F_{t-1}^v\gamma_t\rho_{t-1}^v + 1, \forall t > 0. \tag{69}$$

where $F_0^v = 1$. To see why it stabilizes V-trace learning, we now examine the limit of the $\boldsymbol{A}$ matrix when using the truncated importance sampling ratios $\bar{\rho}_t$ as in V-trace together with the V-trace follow on trace (Eq. 69). We have that:

$$\boldsymbol{A} = \lim_{t\to\infty} \mathbb{E}[\boldsymbol{A}_t] = \lim_{t\to\infty} \mathbb{E}_\mu F_t^v \boldsymbol{\phi}(S_t)\bar{\rho}_t \left[\boldsymbol{\phi}(S_t) - \gamma_{t+1}\boldsymbol{\phi}(S_{t+1})\right]^T \tag{70}$$

$$= \sum_s d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t^v \bar{\rho}_t \boldsymbol{\phi}(S_t) \left[\boldsymbol{\phi}(S_t) - \gamma_{t+1}\boldsymbol{\phi}(S_{t+1})\right]^T \middle| S_t = s\right] \tag{71}$$

$$= \sum_s d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t^v \rho_t^v \nu(S_t)\boldsymbol{\phi}(S_t) \left[\boldsymbol{\phi}(S_t) - \gamma_{t+1}\boldsymbol{\phi}(S_{t+1})\right]^T \middle| S_t = s\right], \tag{72}$$

plugging in the definition of ρ_t^v. Just like the derivation from Eq. 52 to Eq. 53, we use the fact that given S_t, the emphatic trace F_t^v is independent of $\rho_t^v \nu(S_t)\boldsymbol{\phi}_t(\boldsymbol{\phi}_t - \gamma\boldsymbol{\phi}_{t+1})$ and the expected value of the latter term does not depend on the time step t under the steady state distribution, so for clarity we change the time index to a new variable k. Thus we have

$$\sum_s d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu \left[F_t^v | S_t = s\right] \mathbb{E}_\mu \left[\rho_k^v \nu(S_k)\boldsymbol{\phi}(S_k) \left[\boldsymbol{\phi}(S_k) - \gamma_{k+1}\boldsymbol{\phi}(S_{k+1})\right]^T \middle| S_k = s\right] \tag{73}$$

$$= \sum_s f^v(s)\nu(s)\mathbb{E}_\mu \left[\rho_k^v \boldsymbol{\phi}(S_k) \left[\boldsymbol{\phi}(S_k) - \gamma_{k+1}\boldsymbol{\phi}(S_{k+1})\right]^T \middle| S_k = s\right] \tag{74}$$

$$= \sum_s f^v(s)\nu(s)\mathbb{E}_{\pi_{\bar{\rho}}} \left[\boldsymbol{\phi}(S_k) \left[\boldsymbol{\phi}(S_k) - \gamma_{k+1}\boldsymbol{\phi}(S_{k+1})\right]^T \middle| S_k = s\right] \tag{75}$$

$$= \boldsymbol{\Phi}^T \boldsymbol{F}^v \boldsymbol{N} \left[\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}\boldsymbol{\Gamma}\right] \boldsymbol{\Phi}. \tag{76}$$

In Eq. 74 we used the fact that $\nu(s)$ is a function of the state only (and not the action), in Eq. 75 we replace the expectation over μ with an expectation over $\pi_{\bar{\rho}}$, and finally in Eq. 76 we let $\boldsymbol{N}$ be a diagonal matrix with elements $\nu(s)$ on the diagonal.

It is easy to see that without the emphatic trace F_t^v, the V-trace steady state key matrix is $\boldsymbol{N}\boldsymbol{D}_\mu \left[\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}\boldsymbol{\Gamma}\right]$, which is not necessarily positive definite. Thus V-trace may suffer from instability issues with linear function approximation. Following (Sutton et al., 2016), we have that $\boldsymbol{F}^v = \left[\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}^T\boldsymbol{\Gamma}\right]^{-1} \boldsymbol{D}_\mu$ with F_t^v as defined in Eq. 9. Let's check that the key matrix $\boldsymbol{N}\boldsymbol{F}^v \left[\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}\boldsymbol{\Gamma}\right]$ is positive definite. First notice that the $\boldsymbol{F}^v$ and $\boldsymbol{N}$ are diagonal matrices with positive diagonal entries. With transition probabilities smaller than 1, the key matrix must have positive diagonal entries and negative off-diagonal entries. Moreover its row sum is an all positive vector. The column sum of the key matrix is

$$\mathbf{1}^T \boldsymbol{F}^v (\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}\boldsymbol{\Gamma})\boldsymbol{N} = \boldsymbol{f}^T (\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}\boldsymbol{\Gamma})\boldsymbol{N} \tag{77}$$

$$= \boldsymbol{d}_\mu^T \left[\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}\boldsymbol{\Gamma}\right]^{-1} (\boldsymbol{I} - \boldsymbol{P}_{\pi_{\bar{\rho}}}\boldsymbol{\Gamma})\boldsymbol{N} \tag{78}$$

$$= \boldsymbol{d}_\mu^T \boldsymbol{N}, \tag{79}$$

which is an all positive vector. Hence F_t^v stabilized learning. Finally, combined with the definition of λ_t^v in Eq. 18, now we have the WEVtrace.

D.2. NEVtrace

We define the emphatic trace for the n-step V-trace update as:

$$F_t^{(n),v} = \prod_{i=1}^{n} (\gamma_{t-i+1}\rho_{t-i}^v) F_{t-n}^{(n),v} + 1, \tag{80}$$

where $F_0^{(n),v}, F_1^{(n),v}, \ldots, F_{n-1}^{(n),v} = 1$. The $\boldsymbol{A}$ matrix for the emphatically modified V-trace update becomes

$$\boldsymbol{A} = \lim_{t\to\infty} \mathbb{E}[\boldsymbol{A}_t] = \lim_{t\to\infty} \mathbb{E}_\mu F_t \boldsymbol{\phi}(S_t) \sum_{i=t}^{t+n-1} \prod_{j=t}^{i-1} (\gamma_{j+1}\bar{\rho}_j)\bar{\rho}_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \tag{81}$$

$$= \ldots \qquad \text{(similar derivations as Eq. 50 to Eq. 54)}$$

$$= \sum_s f^{(n),v}(s)\mathbb{E}_\mu \left[\boldsymbol{\phi}(S_k) \sum_{i=k}^{k+n-1} \prod_{j=k}^{i-1} (\gamma_{j+1}\bar{\rho}_j)\bar{\rho}_i \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_k = s \right] \tag{82}$$

$$= \sum_{i=k}^{k+n-1} \sum_s f^{(n),v}(s)\mathbb{E}_\mu \left[\boldsymbol{\phi}(S_k) \prod_{j=k}^{i-1} (\gamma_{j+1}\rho_j^v \nu(S_j))\rho_i^v \nu(S_i) \left[\boldsymbol{\phi}(S_i) - \gamma_{i+1}\boldsymbol{\phi}(S_{i+1})\right]^T \middle| S_k = s \right] \tag{83}$$

$$= \sum_{i=k}^{k+n-1} \boldsymbol{\Phi}^T \boldsymbol{N}^{i-k+1} \boldsymbol{F}^{(n),v} \left[\boldsymbol{P}_{\pi_{\bar{\rho}}}^{i-k} \boldsymbol{\Gamma}^{i-k} - \boldsymbol{P}_{\pi_{\bar{\rho}}}^{i-k+1} \boldsymbol{\Gamma}^{i-k+1} \right] \boldsymbol{\Phi} \tag{84}$$

$$\approx \sum_{i=k}^{k+n-1} \boldsymbol{\Phi}^T \boldsymbol{F}^{(n),v} \left[(\boldsymbol{N}^{i-k} \boldsymbol{P}_{\pi_{\bar{\rho}}}^{i-k} \boldsymbol{\Gamma}^{i-k} - \boldsymbol{N}^{i-k+1} \boldsymbol{P}_{\pi_{\bar{\rho}}}^{i-k+1} \boldsymbol{\Gamma}^{i-k+1}) \right] \boldsymbol{\Phi} \tag{85}$$

$$= \boldsymbol{\Phi}^T \boldsymbol{F}^{(n),v} (\boldsymbol{I} - \boldsymbol{N}^n \boldsymbol{P}_{\pi_{\bar{\rho}}}^n \boldsymbol{\Gamma}^n) \boldsymbol{\Phi}, \tag{86}$$

where $f^{(n),v}(s)$ are diagonal entries of $\boldsymbol{F}^{(n),v}$ and $\boldsymbol{F}^{(n),v} \doteq \left[\boldsymbol{I} - \boldsymbol{N}^n {\boldsymbol{P}_{\pi_{\bar{\rho}}}^T}^n \boldsymbol{\Gamma}^n\right]^{-1} \boldsymbol{D}_\mu$. Similar to before, this $\boldsymbol{F}^{(n),v}$ makes the approximate key matrix $\boldsymbol{F}^{(n),v}(\boldsymbol{I} - \boldsymbol{N}^n \boldsymbol{P}_{\pi_{\bar{\rho}}}^n \boldsymbol{\Gamma}^n)$ positive definite. Recall that $\boldsymbol{N}$ is a diagonal matrix with either value 1 or some value in $(0, 1)$ for states where the IS weights are clipped by $\bar{\rho}$, and Eq. 85 is approximate by treating one of the $\boldsymbol{N}$ matrices as an identity matrix.

To see why this follows from the NEVtrace definition in Eq. 80, recall derivations for NETD trace in Eq. 60 for $n = 2$. Here we have

$$f^{(n),v}(s) = d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[F_t^{(n),v} \mid S_t = s] \tag{87}$$

$$= \ldots \tag{88}$$

$$= d_\mu(s) + \gamma(s) \sum_{s'} [\boldsymbol{P}_{\pi_{\bar{\rho}}}]_{s,s'} \nu(s')\gamma(s') \sum_{s''} [\boldsymbol{P}_{\pi_{\bar{\rho}}}]_{s',s''} \nu(s'') f(s''). \tag{89}$$

Thus the vector

$$\boldsymbol{f}^{(n),v} = \boldsymbol{d}_\mu + \boldsymbol{N}^2 {\boldsymbol{P}_{\pi_{\bar{\rho}}}^T}^2 \boldsymbol{\Gamma}^2 \boldsymbol{f} = (\boldsymbol{I} + \boldsymbol{N}^2 {\boldsymbol{P}_{\pi_{\bar{\rho}}}^T}^2 \boldsymbol{\Gamma}^2 + \boldsymbol{N}^4 {\boldsymbol{P}_{\pi_{\bar{\rho}}}^T}^4 \boldsymbol{\Gamma}^4 + \cdots)\boldsymbol{d}_\mu = (\boldsymbol{I} - \boldsymbol{N}^2 {\boldsymbol{P}_{\pi_{\bar{\rho}}}^T}^2 \boldsymbol{\Gamma}^2)^{-1} \boldsymbol{d}_\mu. \tag{90}$$

Extending this to any other positive integer value of n, we conclude

$$\boldsymbol{F}^{(n),v} = (\boldsymbol{I} - \boldsymbol{N}^n {\boldsymbol{P}_{\pi_{\bar{\rho}}}^T}^n \boldsymbol{\Gamma}^n)^{-1} \boldsymbol{D}_\mu. \tag{91}$$

E. Hyperparameters

Architectures.

Table 3. Network architecture

Parameter	
convolutions in block	(2, 2, 2, 2)
channels	(64, 128, 128, 64)
kernel sizes	(3, 3, 3, 3)
kernel strides	(1, 1, 1, 1)
pool sizes	(3, 3, 3, 3)
pool strides	(2, 2, 2, 2)
frame stacking	4
head hiddens	512
activation	Relu

Our DNN architecture is composed of a shared torso, which then splits to different heads. We have a head for the policy and a head for the value function (multiplied by the number of auxiliary tasks). Each head is a two-layered MLP with 512 hidden units, where the output dimension corresponds to 1 for the vale function head. For the policy head, we have $|\mathcal{A}|$ outputs that correspond to softmax logits. We use ReLU activations on the outputs of all the layers besides the last layer. For the policy head, we apply a softmax layer and use the entropy of this softmax distribution as a regularizer.

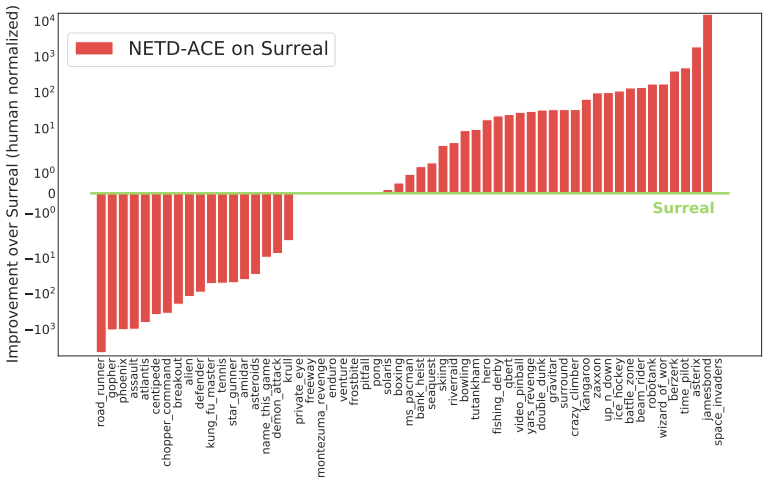
The **torso** of the network is composed from residual blocks. In each block there is a convolution layer, with stride, kernel size, channels specified in Table 3, with an optional pooling layer following it. The convolution layer is followed by n - layers of convolutions (specified by blocks), with a skip contention. The output of these layers is of the same size of the input so they can be summed. The block convolutions have kernel size 3, stride 1.

Hyperparameters. Table 4 lists all the hyperparameters used by our agent. Most of the hyperparameters follow the reported parameters from the IMPALA paper. For completeness, we list all of the exact values that we used below.

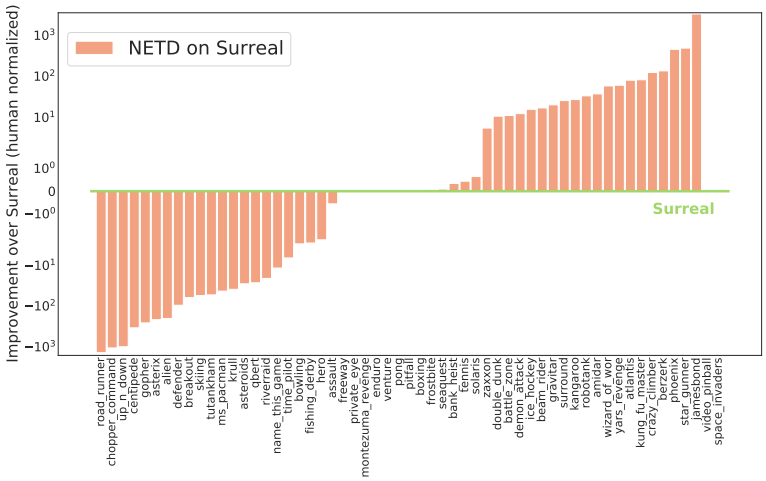
Table 4. Hyperparameters table

Parameter	Value
total environment steps	200e6
optimizer	RMSPROP
start learning rate	$6 \cdot 10^{-4}$ (mixed), $2 \cdot 10^{-4}$ (fixed)
end learning rate	0
decay	0.99
eps	0.1
importance sampling clip	1
gradient norm clip	0.3 (mixed), 1 (fixed)
trajectory n	40 (mixed), 10 (fixed)
batch size (m)	18
discount γ (main)	$\sigma(4.6) \approx .99$
discount γ^1 (1^{st} auxiliary)	$\sigma(4.4) \approx .988$
discount γ^2 (2^{nd} auxiliary)	$\sigma(4.2) \approx .985$

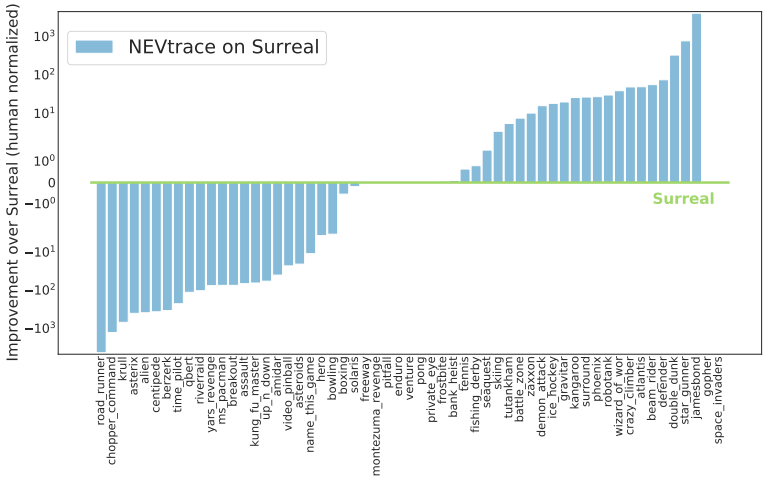
F. Atari experiment results



(a) NETD-ACE on Surreal versus Surreal



(b) NETD on Surreal versus Surreal



(c) NEVtrace on Surreal versus Surreal

Figure 8. Improvement in individual human normalized game scores compared to Surreal: (a) NETD-ACE, (b) NETD and (c) NEVtrace on Surreal versus Surreal in green. Results averaged across 3 seeds and the evaluation phase.

G. Diagnostic experiments

We include the full experiment results and additional implementation details on the diagnostic MDPs in this section.

G.1. Two-state MDP

In Fig. 9, the rightmost column shows three baselines: Off-policy TD(0), Clipped off-policy n-step TD where all IS weights are clipped except for those directly on the TD error, i.e. unbiased V-trace by (Espeholt et al., 2018), and V-trace. All three baselines diverged on this MDP. The emphatic algorithms all converged to the optimal fixed point $\theta = 0$, however, notice that both emphatic TD (NETD/WETD) and emphatic V-trace (NEVtrace/WEVtrace) algorithms exhibited unstable learning, with some runs experiencing large jumps in value error late in training. The clipped emphatic traces (IS clipped at 1) theoretically have a finite variance, and empirically enjoy a faster convergence and stable learning after initial fluctuations.

As we increase the bootstrap length to $n = 5$, the Off-policy 5-step TD baseline converged quickly while the other two baselines Clipped off-policy n-step TD and V-trace exhibited higher variances (see Fig. 10). Similar to before, the clipped traces (row 2) were effective in variance reduction and demonstrated fast convergence. In comparison, WETD exhibited more variance in learning, NEVtrace converged slowly and NETD, WEVtrace were unstable late in training.

G.2. Collision Problem

We present the full results in Fig. 11 and Fig. 12. All algorithms achieved stable learning with their best learning rates from hyperparameter sweeps. Emphatic algorithms consistently achieved the smallest mean RMSE averaged over 200 runs for all values of n tested.

G.3. Baird's counterexample

Baird's counterexample (Fig. 13) is a simple MDP with has seven states and simple linear features that causes TD and other methods to diverge. The features are designed to cause unnecessary generalization, even though the true value function is perfectly representable. This over-parameterization combined with a large mismatch in the target and behavior policies typically causes divergence. See (Sutton & Barto, 2018) for an extensive discussion and analysis of Baird's counterexample.

Using the TD(0) learning update (Fig. 14), both emphatic traces (row 1) converged quickly, with occasional instability in some runs. The clipped emphatic traces (row 2) exhibit slow learning, but exhibit a clear downward trend. The n-step TD baselines and all methods with V-trace targets diverged. This is not surprising as Baird's counterexample is considerably harder than the two-state MDP—Sutton's ETD(λ) diverges on Baird's counterexample, but converges on the two-state MDP (Sutton et al., 2016). Using 5-step TD learning (Fig. 15) improves the performance of several methods. The WETD algorithms performed poorly compared to NETD algorithms and the n-step TD baseline. We see the effect of IS clipping: lowering variance of both WETD and NETD. Vtrace, NEVtrace, and WEVtrace all slowly diverged and WEVtrace exhibited unstable learning late in training. Clipped emphatic methods and n-step TD all benefit from longer n-step targets, significantly improving over their one-step variants in Fig 14. In this challenging MDP, one-step methods are not sufficient for fast and stable learning.

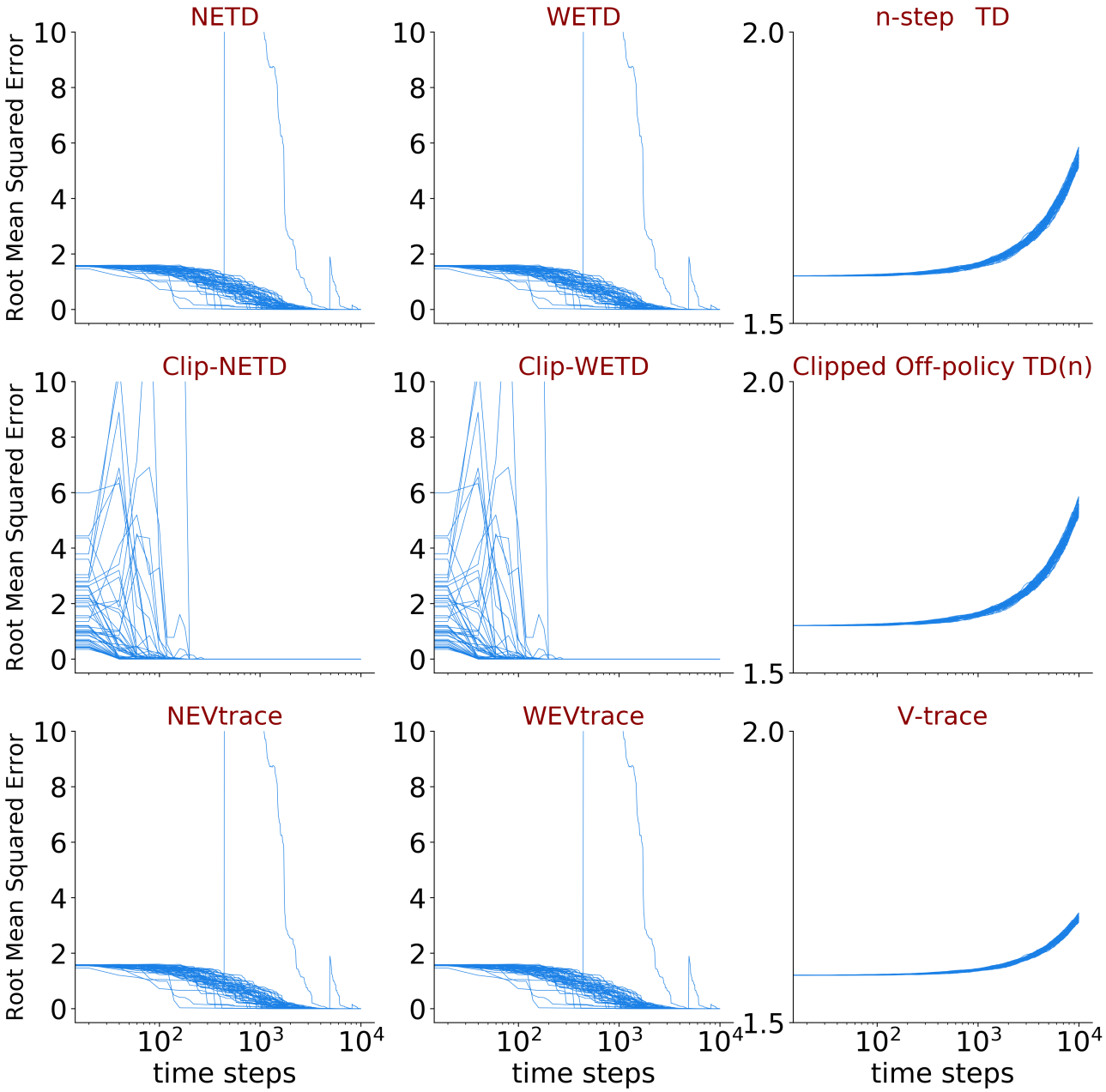


Figure 9. RMSE over time on the two-state MDP with $\gamma = 0.9$ and $n = 1$. Each subplot shows fifty independent runs of each algorithm using the best setting setting of α found in the hyperparameter sweep. Note the log scale on x-axis.

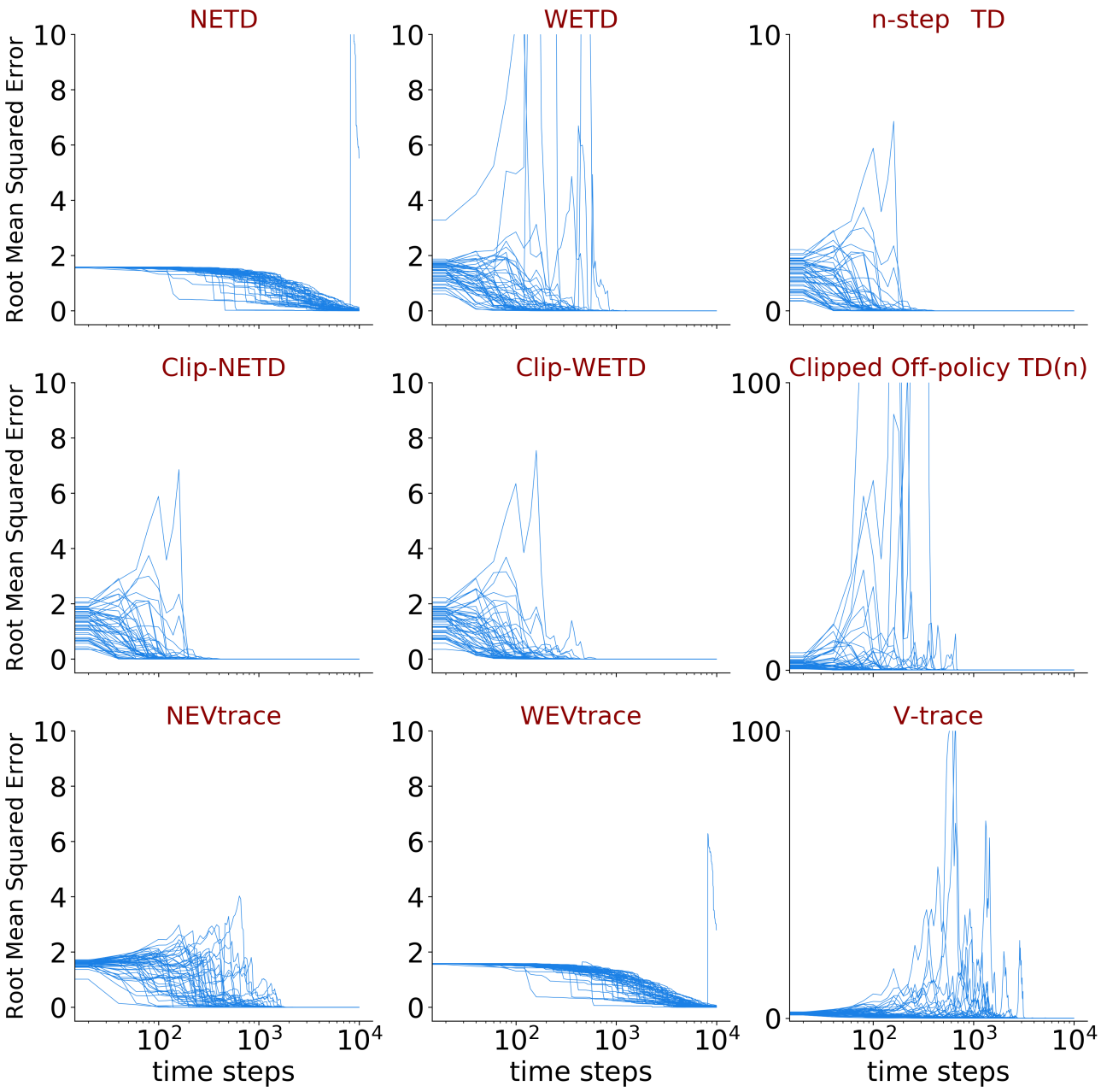


Figure 10. Same experiment setup on the two-state MDP as Fig. 9 except $n = 5$. Note the log scale on x-axis.

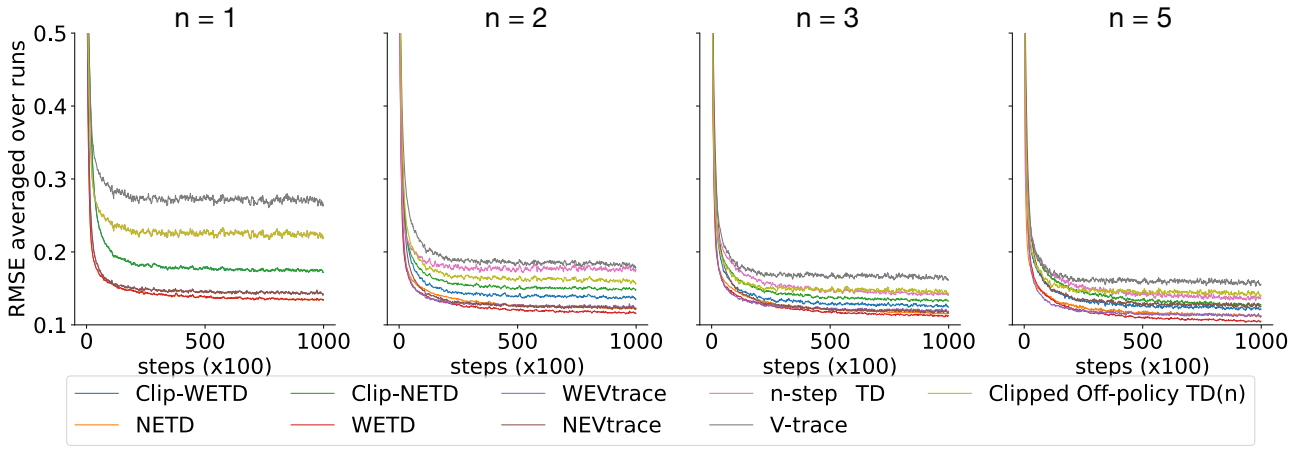


Figure 11. Learning curves comparison on the Collision Episodic MDP. The results are averaged over 200 independent runs with each algorithms best hyper-parameter setting from the sweep. The three baselines produced the highest averaged RMSE errors. Emphatic algorithms ETD, NETD and EVtrace, NEVtrace consistently produced the lowest RMSEs for all bootstrap values n, followed by the clipped emphatic traces.

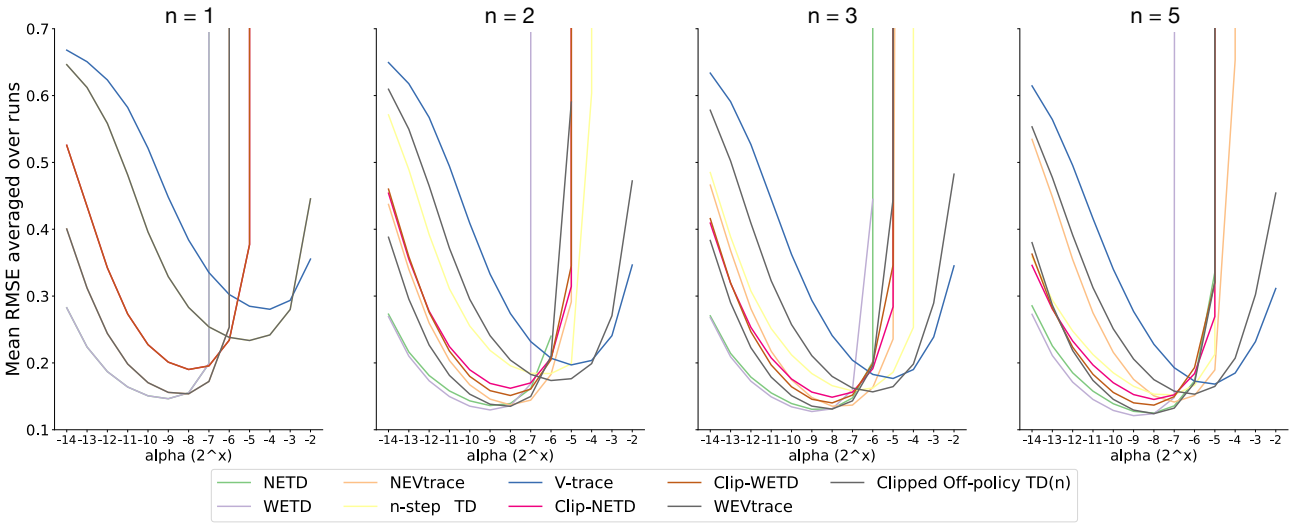


Figure 12. Hyper-parameter sensitivity comparison on the Collision Episodic MDP. Each data point in the plot portrays the mean RMSE averaged over 200 runs for varying learning rate α and the bootstrap length n. The emphatic algorithms achieved the lowest value error with smaller learning rates than the baselines.

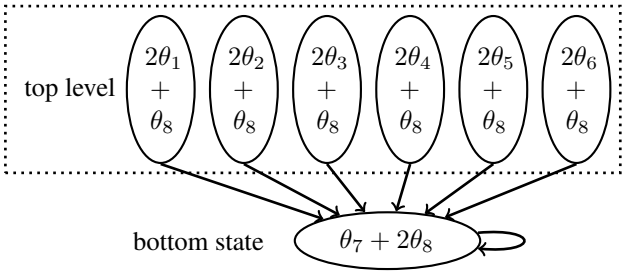


Figure 13. Baird's counterexample MDP. Solid lines indicate the target policy $\pi(\text{down}|\cdot) = 1$, ending up in the bottom state. The behavior policy $\mu(\text{up}|\cdot) = 6/7, \mu(\text{down}|\cdot) = 1/7$. When action is "up", the agent goes to a random state on the top level. When action is "down", the agent goes to the bottom state.

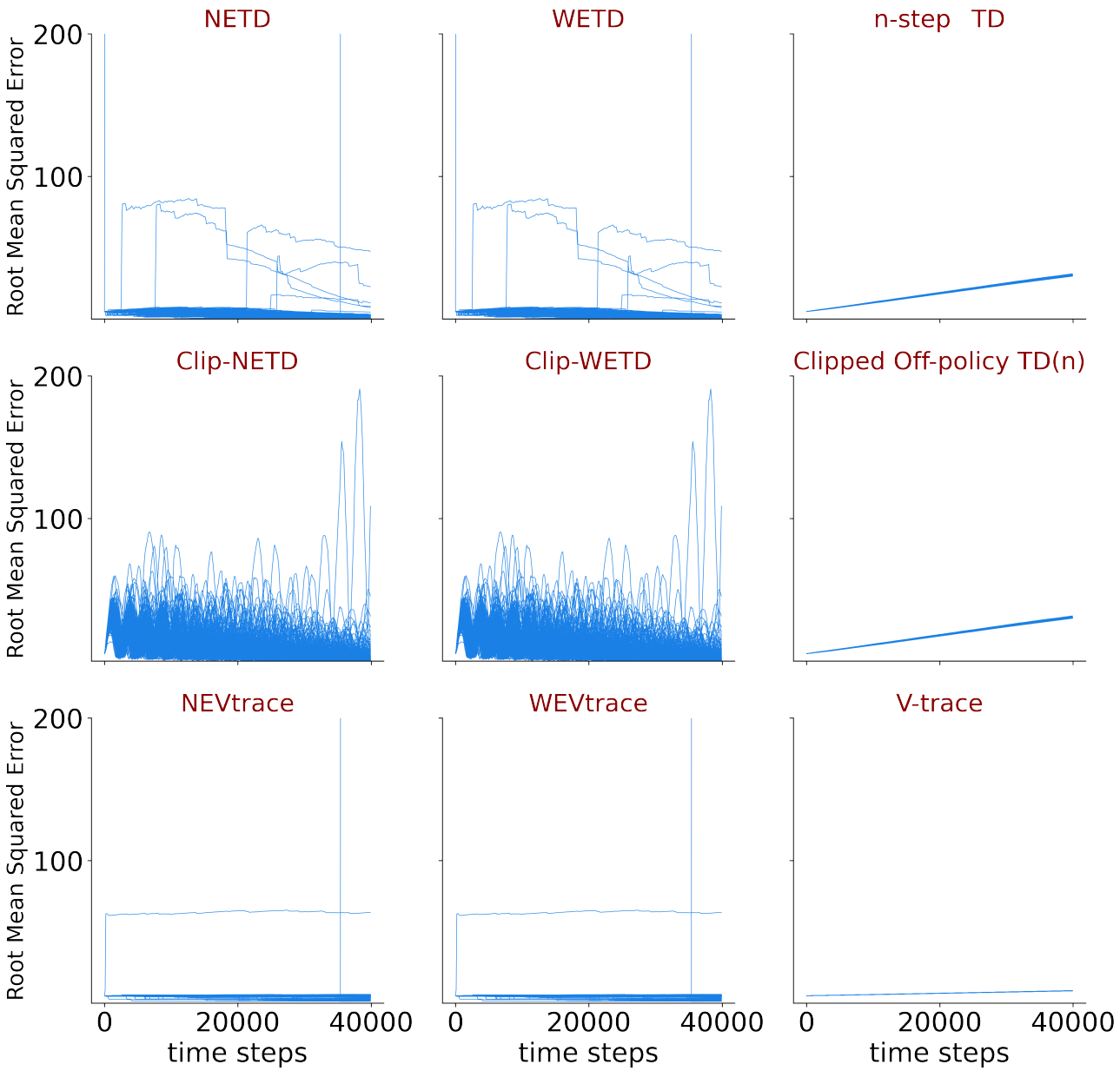


Figure 14. Baird's MDP, $n = 1, \gamma = 0.9$, 200 indepdendent runs. Each algorithm was run with its best hyperparameter setting. Note the log scale on x-axis.

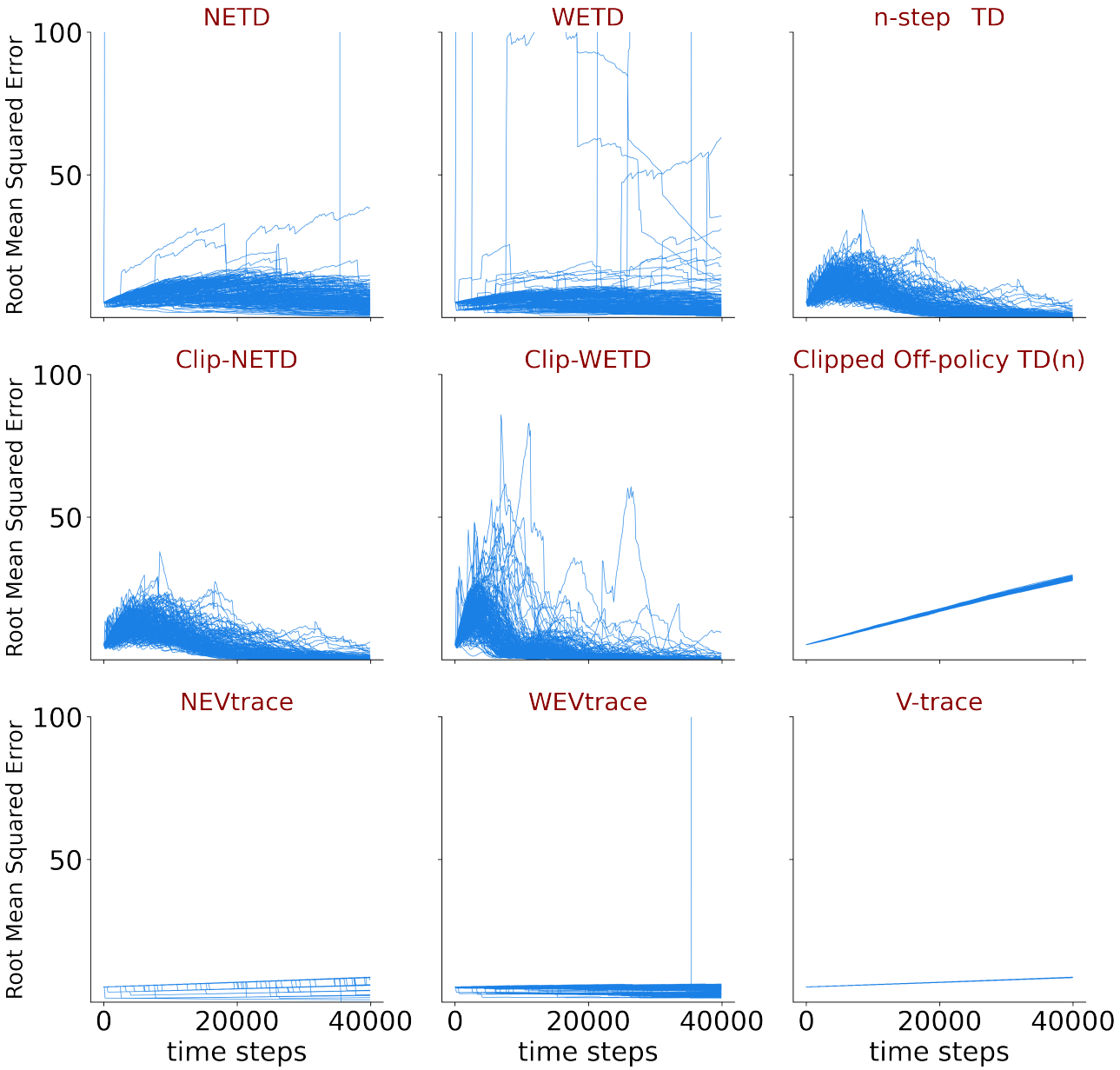


Figure 15. Baird's MDP, $n = 5, \gamma = 0.9$, 200 indepdendent runs. Each algorithm was run with its best hyperparameter setting. NETD has a smaller variance than ETD algorithms. Clipping the IS in emphatic traces help reduce the variance. Note the log scale on x-axis.