Appendix

A General Convergence Considerations for MDPs in Finite State Space

In this section, we will discuss the infinite batch limit and compare the value function obtained with TD to the ground truth value function. We will, for simplicity, consider in this section a Markov reward process with transition matrix $p(s_{t+1} = s'|s_t = s) = \Pi(s, s')$. The general theory described in the main text does not only apply to MDPs, but the convergence analysis for MDPs is much more straightforward so we describe it here. In this case, the ground truth value function satisfies

$$V(s) = R(s) + \gamma \sum_{s'} \Pi(s, s')V(s') \tag{A.1}$$

which gives the vector equation $\boldsymbol{V} = (\boldsymbol{I} - \gamma \boldsymbol{\Pi})^{-1} \boldsymbol{R}$ for $\boldsymbol{V}, \boldsymbol{R} \in \mathbb{R}^{|\mathcal{S}|}$. Suppose the limiting distribution over states is $\boldsymbol{p} \in \mathbb{R}^{|\mathcal{S}|}$ which has entries $p(s) = \frac{1}{T} \sum_{t=1}^{T} p(s_t = s)$. The fixed point of TD dynamics is

$$\boldsymbol{\Psi} \text{diag}(\boldsymbol{p}) \boldsymbol{\Psi}^\top \boldsymbol{w}_{TD} = \boldsymbol{\Psi} \text{diag}(\boldsymbol{p}) \boldsymbol{R} + \gamma \boldsymbol{\Psi} \text{diag}(\boldsymbol{p}) \boldsymbol{\Pi} \boldsymbol{\Psi}^\top \boldsymbol{w}_{TD}. \tag{A.2}$$

We now consider the two possible cases for this fixed point condition.

Case 1: Underparameterized Regime First, if the feature dimension N is smaller than the size of the state space $|\mathcal{S}|$ and the features are maximal rank, then the TD learning fixed point is

$$\boldsymbol{w}_{TD} = \left(\boldsymbol{\Psi} \text{diag}(\boldsymbol{p}) \boldsymbol{\Psi}^\top - \gamma \boldsymbol{\Psi} \text{diag}(\boldsymbol{p}) \boldsymbol{\Pi} \boldsymbol{\Psi}^\top \right)^{-1} \boldsymbol{\Psi} \text{diag}(\boldsymbol{p}) \boldsymbol{R} \tag{A.3}$$

In this case, the value function is not learned perfectly, as can be seen by computing $\hat{\boldsymbol{V}} = \boldsymbol{\Psi}^\top \boldsymbol{w}_{TD}$ and comparing to the ground truth $\boldsymbol{V} = (\boldsymbol{I} - \gamma \boldsymbol{\Pi})^{-1} \boldsymbol{R}$. In this case, we would say that TD learning has an *irreducible value error* due to capturing only a N dimensional projection of the value function.

Case 2: Overparameterized Regime Alternatively, if the feature dimension exceeds the total number of states, then the fixed point equation for TD is underspecified. However, throughout TD learning $\boldsymbol{w}_{TD} \in \text{span}\{\boldsymbol{\psi}(s)\}_{s \in \mathcal{S}}$ so we can instead consider the decompostion $\boldsymbol{w}_V = \sum_s \alpha(s) \boldsymbol{\psi}(s)$, where $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{S}|}$ satisfies

$$\text{diag}(\boldsymbol{p})(\boldsymbol{I} - \gamma \boldsymbol{\Pi}) \boldsymbol{K} \boldsymbol{\alpha} = \text{diag}(\boldsymbol{p}) \boldsymbol{R} \tag{A.4}$$

where $\boldsymbol{K} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the kernel computed with features $K(s, s') = \boldsymbol{\psi}(s) \cdot \boldsymbol{\psi}(s')$. The solution to the above equation is unique and the learned value function $\hat{\boldsymbol{V}} = \boldsymbol{\Psi}^\top \boldsymbol{w}_{TD} = \boldsymbol{K} \boldsymbol{K}^{-1} (\boldsymbol{I} - \gamma \boldsymbol{\Pi})^{-1} \boldsymbol{R} = (\boldsymbol{I} - \gamma \boldsymbol{\Pi})^{-1} \boldsymbol{R} = \boldsymbol{V}$. Therefore, in the over-parameterized limit, the irreducible value error for TD learning is zero. This limit was considered dynamically in the infinite batch (vanishing SGD noise) setting by [41].

B Derivation of Learning Curves

In this section, we now consider the dynamics of TD learning when B random episodes are sampled at a time. In this calculation, the finite batch of episodes leads to non-negligible SGD effects which can cause undesirable plateaus in TD dynamics.

B.1 Field Theory Derivation

In this section we use a Gaussian field theory formalism to compute the learning curve in the high dimensional asymptotic limit $N, B \to \infty$ with $B/N = \alpha$. The episode length T is treated as $\mathcal{O}(1)$. While this paper focuses on the online setting, where fresh trajectories $\{\tau_n^\mu\}$ are sampled at each iteration n, this model can be straightforwardly extended to the case where a fixed number of experience trajectories $\{\tau^\mu\}$ are replayed repeatedly during TD learning. We leave the experience

replay dynamic mean field theory calculation for future work. The starting point of our analysis is tracking the moment generating function for the iterate dynamics

$$Z[\{\boldsymbol{j}_n\}] = \mathbb{E}_{\{\boldsymbol{w}_n\},\{s_n^\mu(t)\}} \exp\left(i \sum_{n=0}^{\infty} \boldsymbol{j}_n \cdot \boldsymbol{w}_n \right). \tag{B.1}$$

To compute this object over random draws of training trajectories, we express the joint average over $\boldsymbol{w}_n, \{s_n^\mu(t)\}$ into conditional averages over $\boldsymbol{w}_n, \{\Delta_n^\mu(t)\}|\{\boldsymbol{\psi}_n^\mu(t)\}$. To simplify the computation, in this section, we will compute the learning curve for mean zero features $\boldsymbol{\mu}(s) = 0$ and

$$\begin{aligned} Z =& \mathbb{E}_{\{\boldsymbol{\psi}_n^\mu(t)\}} \int \prod_n d\boldsymbol{w}_n \delta\left(\boldsymbol{w}_{n+1} - \boldsymbol{w}_n - \frac{\eta}{\sqrt{BT}} \sum_{\mu t} \Delta_n^\mu(t) \boldsymbol{\psi}_n^\mu(t) \right) \exp\left(i \sum_{n=0}^{\infty} \boldsymbol{j}_n \cdot \boldsymbol{w}_n \right) \\ & \times \int \prod_{t\mu n} d\Delta_n^\mu(t)\, \delta\left(\Delta_n^\mu(t) - \frac{1}{\sqrt{N}} (\boldsymbol{w}_R - \boldsymbol{w}_n) \cdot \boldsymbol{\psi}_n^\mu(t) - \frac{\gamma}{\sqrt{N}} \boldsymbol{w}_n \cdot \boldsymbol{\psi}_n^\mu(t+1) \right) \end{aligned} \tag{B.2}$$

Expressing the Dirac-delta function as a Fourier integral $\delta(z) = \int \frac{d\hat{z}}{2\pi} \exp(i\hat{z}z)$ for each of our constraints. Under the *Gaussian equivalence ansatz*, we can easily average over Gaussian $\boldsymbol{\psi}$ to obtain

$$\begin{aligned} Z = & \int \mathcal{D}\Delta\mathcal{D}\hat{\Delta}\mathcal{D}\boldsymbol{w}\mathcal{D}\hat{\boldsymbol{w}} \exp\left(-\frac{\eta^2}{2BT^2} \sum_{n\mu} \sum_{tt'} \Delta_n^\mu(t) \Delta_n^\mu(t') \hat{\boldsymbol{w}}_n^\top \boldsymbol{\Sigma}(t,t') \hat{\boldsymbol{w}}_n \right) \\ & \exp\left(i \sum_n \hat{\boldsymbol{w}}_n \cdot (\boldsymbol{w}_{n+1} - \boldsymbol{w}_n) \right) \\ & \exp\left(-\frac{1}{2N} \sum_{n\mu tt'} \left[(\boldsymbol{w}_R - \boldsymbol{w}_n) \hat{\Delta}_n^\mu(t) \right] \boldsymbol{\Sigma}(t,t') \left[(\boldsymbol{w}_R - \boldsymbol{w}_n) \hat{\Delta}_n^\mu(t') \right] \right) \\ & \exp\left(-\frac{\gamma^2}{2N} \sum_{n\mu tt'} \hat{\Delta}_n^\mu(t-1) \hat{\Delta}_n^\mu(t'-1) \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \right) \\ & \exp\left(-\frac{\gamma}{N} \sum_{n\mu tt'} \hat{\Delta}_n^\mu(t-1) \hat{\Delta}_n^\mu(t') \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t,t') (\boldsymbol{w}_R - \boldsymbol{w}_n) \right) \\ & \exp\left(-\frac{\eta}{\sqrt{NBT}} \sum_{n\mu tt'} \left[\hat{\Delta}_n^\mu(t) (\boldsymbol{w}_R - \boldsymbol{w}_n) + \gamma \hat{\Delta}_n^\mu(t-1) \boldsymbol{w}_n \right]^\top \boldsymbol{\Sigma}(t,t') \hat{\boldsymbol{w}}_n \Delta_n^\mu(t') \right) \\ & \exp\left(i \sum_{n\mu t} \hat{\Delta}_n^\mu(t) \Delta_n^\mu(t) + i \sum_n \boldsymbol{j}_n \cdot \boldsymbol{w}_n \right) \end{aligned} \tag{B.3}$$

where we adopted the shorthand $\mathcal{D}\Delta = \prod_{\mu,n,t} d\Delta_n^\mu(t)$ for the measure for the collection of variables $\{\Delta_n^\mu(t)\}$. Likewise one should interpret $\mathcal{D}\boldsymbol{w} = \prod_n d\boldsymbol{w}_n$. To analyze the high dimensional limit of the above moment generating function, we introduce order parameters for the theory

$$\begin{aligned} & Q_n(t,t') = \frac{1}{B} \sum_{\mu=1}^{B} \Delta_n^\mu(t) \Delta_n^\mu(t') \;,\; C_n(t,t') = \frac{1}{N} \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \\ & C_n^R(t,t') = \frac{1}{N} \boldsymbol{w}_R \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \;,\; D_n(t,t') = -\frac{i}{N} \hat{\boldsymbol{w}}_n^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \;,\; D_n^R(t,t') = -\frac{i}{N} \hat{\boldsymbol{w}}_n^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_R \end{aligned} \tag{B.4}$$

For each of these order parameters, we enforce the definition of the order parameter using the Fourier representation of a Dirac-delta function

$$\begin{aligned}1 &= B\int dQ_n(t,t')\delta\left(BQ_n(t,t') - \sum_\mu \Delta_n^\mu(t)\Delta_n^\mu(t')\right)\\ &= B\int \frac{dQ_n(t,t')d\hat{Q}_n(t,t')}{4\pi i}\exp\left(\frac{B}{2}\hat{Q}_n(t,t')Q_n(t,t') - \frac{1}{2}\sum_\mu \Delta_n^\mu(t)\Delta_n^\mu(t')\hat{Q}_n(t,t')\right).\end{aligned} \tag{B.5}$$

Repeating this procedure for all order parameters $q = \{Q, \hat{Q}, C, \hat{C}, C^R, \hat{C}^R, D, \hat{D}, D^R, \hat{D}^R\}$ and disregarding irrelevant prefactors, we have the following formula for the moment generating function

$$Z \propto \int \mathcal{D}q \exp\left(\frac{N}{2}S[q]\right) \tag{B.6}$$

where the action S has the form

$$\begin{aligned}S =& \sum_n\sum_{tt'}\left[\alpha Q_n(t,t')\hat{Q}_n(t,t') + C_n(t,t')\hat{C}_n(t,t') + C_n^R(t,t')\hat{C}_n^R(t,t')\right]\\ &-2\sum_n\sum_{tt'}\left[D_n(t,t')\hat{D}_n(t,t') + D_n^R(t,t')\hat{D}_n^R(t,t')\right] + \frac{2}{N}\ln\mathcal{Z}_w + 2\alpha\ln\mathcal{Z}_\Delta\\ \mathcal{Z}_w =& \int \mathcal{D}\boldsymbol{w}\mathcal{D}\hat{\boldsymbol{w}}\exp\left(-\frac{\eta^2}{2T^2}\sum_{ntt'}Q_n(t,t')\hat{\boldsymbol{w}}_n^\top\boldsymbol{\Sigma}(t,t')\hat{\boldsymbol{w}}_n + i\sum_n\hat{\boldsymbol{w}}_n\cdot(\boldsymbol{w}_{n+1}-\boldsymbol{w}_n)\right)\\ &\exp\left(-\frac{1}{2}\sum_{ntt'}\hat{C}_n(t,t')\boldsymbol{w}_n^\top\boldsymbol{\Sigma}(t,t')\boldsymbol{w}_n - \frac{1}{2}\hat{C}_n^R(t,t')\boldsymbol{w}_R^\top\boldsymbol{\Sigma}(t,t')\boldsymbol{w}_n\right)\\ &\exp\left(-i\sum_{ntt'}\hat{D}_n(t,t')\hat{\boldsymbol{w}}_n^\top\boldsymbol{\Sigma}(t,t')\boldsymbol{w}_n - i\sum_{ntt'}\hat{D}_n^R(t,t')\hat{\boldsymbol{w}}_n^\top\boldsymbol{\Sigma}(t,t')\boldsymbol{w}_R\right)\\ \mathcal{Z}_\Delta =& \int \mathcal{D}\Delta\mathcal{D}\hat{\Delta}\exp\left(-\frac{1}{2}\sum_{ntt'}\hat{Q}_n(t,t')\Delta_n(t)\Delta_n(t') + i\sum_{nt}\hat{\Delta}_n(t)\Delta_n(t)\right)\\ &\exp\left(-\frac{1}{2}\sum_{ntt'}\hat{\Delta}_n(t)\hat{\Delta}_n(t')\left[\frac{1}{N}\boldsymbol{w}_R^\top\boldsymbol{\Sigma}(t,t')\boldsymbol{w}_R + C(t,t')\right]\right)\\ &\exp\left(\frac{1}{2}\sum_{ntt'}\hat{\Delta}_n(t)\hat{\Delta}_n(t')\left[C^R(t,t') + C^R(t',t)\right]\right)\\ &\exp\left(-\gamma\sum_{t,t'}\hat{\Delta}_n(t)\hat{\Delta}_n(t'-1)C_n^R(t,t')\right)\\ &\exp\left(-\frac{\gamma^2}{2}\sum_{t,t'}\hat{\Delta}_n(t-1)\hat{\Delta}_n(t'-1)C_n(t,t')\right)\\ &\exp\left(-\frac{\eta i}{\sqrt{\alpha}T}\sum_{nt,t'}\hat{\Delta}_n(t)\left[D_n^R(t',t) - D_n(t',t) + \gamma D_n(t',t+1)\right]\Delta_n(t')\right)\end{aligned} \tag{B.7}$$

The function $\mathcal{Z}$ has the interpretation of an effective partition function conditional on order parameters q. To study the $N\to\infty$ limit, we use the steepest descent method and analyze the saddle point

$\frac{\partial S}{\partial q} = 0$. These saddle point equations give

$$\begin{aligned}
\frac{\partial S}{\partial \hat{Q}_n(t,t')} &= \alpha Q_n(t,t') - \alpha \left\langle \Delta_n(t)\Delta_n(t') \right\rangle = 0 \\
\frac{\partial S}{\partial Q_n(t,t')} &= \alpha \hat{Q}_n(t,t') - \frac{\eta^2}{T^2 N} \left\langle \hat{\boldsymbol{w}}_n^\top \boldsymbol{\Sigma}(t,t') \hat{\boldsymbol{w}}_n \right\rangle = 0 \\
\frac{\partial S}{\partial \hat{C}_n(t,t')} &= C_n(t,t') - \frac{1}{N} \left\langle \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \right\rangle = 0 \\
\frac{\partial S}{\partial C_n(t,t')} &= \hat{C}_n(t,t') - \alpha \left\langle \hat{\Delta}_n(t)\hat{\Delta}_n(t') + \gamma^2 \hat{\Delta}_n(t-1)\hat{\Delta}_n(t'-1) \right\rangle = 0 \\
\frac{\partial S}{\partial \hat{C}_n^R(t,t')} &= C_n^R(t,t') - \frac{1}{N} \left\langle \boldsymbol{w}_R^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \right\rangle = 0 \\
\frac{\partial S}{\partial C_n(t,t')} &= \hat{C}_n(t,t') - \alpha \left\langle \hat{\Delta}_n(t)\hat{\Delta}_n(t') + \gamma \hat{\Delta}_n(t)\hat{\Delta}_n(t'-1) \right\rangle = 0 \\
\frac{\partial S}{\partial \hat{D}_n(t,t')} &= -2D_n(t,t') - \frac{2i}{N} \left\langle \hat{\boldsymbol{w}}_n^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \right\rangle = 0 \\
\frac{\partial S}{\partial \hat{D}_n^R(t,t')} &= -2D_n^R(t,t') - \frac{2i}{N} \left\langle \hat{\boldsymbol{w}}_n^\top \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \right\rangle = 0 \\
\frac{\partial S}{\partial D_n(t,t')} &= -2\hat{D}_n(t,t') - \frac{2\alpha\eta i}{\sqrt{\alpha}T} \left\langle \gamma \hat{\Delta}_n(t-1)\Delta_n(t') - \hat{\Delta}_n(t)\Delta_n(t') \right\rangle = 0 \\
\frac{\partial S}{\partial D_n^R(t,t')} &= -2\hat{D}_n^R(t,t') - \frac{2\alpha\eta i}{\sqrt{\alpha}T} \left\langle \hat{\Delta}_n(t)\Delta_n(t') \right\rangle = 0
\end{aligned} \tag{B.8}$$

The brackets $\langle\rangle$ denote averaging over the stochastic processes defined by moment generating functions $\mathcal{Z}_\Delta, \mathcal{Z}_w$. After these saddle point equations are solved the order parameters q are treated as non-random and a Hubbard-Stratonovich transformation is employed. For example,

$$\exp\left(-\frac{1}{2}\hat{\boldsymbol{w}}_n \left[\frac{\eta^2}{T^2} \sum_{tt'} Q_n(t,t') \boldsymbol{\Sigma}(t,t') \right] \hat{\boldsymbol{w}}_n \right) = \mathbb{E}_{\boldsymbol{u}_n^w} \exp\left(i \sum_n \boldsymbol{u}_n^w \cdot \hat{\boldsymbol{w}}_n \right) \tag{B.9}$$

where the average is over $\boldsymbol{u}_n^w \sim \mathcal{N}\left(0, \eta^2 T^{-2} \sum_{tt'} Q_n(t,t') \boldsymbol{\Sigma}(t,t')\right)$. After introducing these Hubbard fields $\boldsymbol{u}_n^w$ and $u_n^\Delta(t)$, we can perform the integrals over $\hat{\boldsymbol{w}}_n$ and $\hat{\Delta}_n(t)$ which collapse to Dirac-Delta functions. The resulting identities of the delta functions define the following stochastic processes on $\boldsymbol{w}_n$ and u_n^Δ

$$\begin{aligned}
\boldsymbol{w}_{n+1} &= \boldsymbol{w}_n + \boldsymbol{u}_n^w + \sum_{tt'} \hat{D}_n^R(t,t') \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_R + \sum_{t,t'} \hat{D}_n(t,t') \boldsymbol{\Sigma}(t,t') \boldsymbol{w}_n \\
\Delta_n(t) &= u_n^\Delta(t) + \frac{\eta}{\sqrt{\alpha}T} \sum_{tt'} [D_n^R(t,t') - D_n(t,t') - \gamma D_n(t',t+1)] \Delta_n(t').
\end{aligned} \tag{B.10}$$

Using a similar trick, we can show that for any observable depending on $\boldsymbol{w}_n$ or $\{\Delta_n(t)\}$ that

$$\begin{aligned}
-i \left\langle \hat{\boldsymbol{w}}_n O(\boldsymbol{w}_n) \right\rangle &= \left\langle \frac{\partial}{\partial \boldsymbol{u}_n} O(\boldsymbol{w}_n) \right\rangle \\
-i \left\langle \hat{\Delta}_n(t) O(\{\Delta_n(t')\}) \right\rangle &= \left\langle \frac{\partial}{\partial u_n^\Delta(t)} O(\{\Delta_n(t')\}) \right\rangle
\end{aligned} \tag{B.11}$$

Since $\boldsymbol{w}_n$ is independent. This can be used to conclude

$$D_n(t,t') = 0 \ , \ D_n^R(t,t') = 0 \tag{B.12}$$

which implies that $\Delta_n(t) = u_n^\Delta(t)$. Consequently the response functions have trivial structure

$$\hat{D}_n(t) = -\frac{\eta\sqrt{\alpha}}{T} \left[\delta(t-t') - \gamma\delta(t-1-t') \right] \ , \ \hat{D}_n^R(t,t') = \frac{\sqrt{\alpha}\eta}{T} \delta(t-t'). \tag{B.13}$$

We therefore obtain a stochastic process of the form

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n + \boldsymbol{u}_n^w + \frac{\eta\sqrt{\alpha}}{T}\sum_t \boldsymbol{\Sigma}(t,t)\boldsymbol{w}_R - \frac{\eta\sqrt{\alpha}}{T}\sum_t \left[\boldsymbol{\Sigma}(t,t) - \gamma\boldsymbol{\Sigma}(t,t+1)\right]\boldsymbol{w}_n$$

$$\boldsymbol{u}_n \sim \mathcal{N}\left(0, \frac{\eta^2}{T^2}\sum_{tt'} Q_n(t,t')\boldsymbol{\Sigma}(t,t')\right) \ , \ \{\Delta_n(t)\} \sim \mathcal{N}(0, \boldsymbol{Q}_n)$$

$$Q_n(t,t') = \langle \Delta_n(t)\Delta_n(t')\rangle = \frac{1}{N}\boldsymbol{w}_R\boldsymbol{\Sigma}(t,t')\boldsymbol{w}_R - C^R(t,t') - C^R(t',t) + C(t,t')$$

$$C_n(t,t') = \frac{1}{N}\left\langle \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t,t')\boldsymbol{w}_n\right\rangle \ , \ C_n^R(t,t') = \frac{1}{N}\left\langle \boldsymbol{w}_R^\top \boldsymbol{\Sigma}(t,t')\boldsymbol{w}_n\right\rangle$$

These are the final equations defining the stochastic evolution of $\boldsymbol{w}_n$ and $\Delta_n(t)$.

B.2 Simplifying the Saddle Point Equations

Using the above saddle point equations, we see that the variables $\{\Delta_n(t)\}$ and $\{\boldsymbol{w}_n\}$ will be Gaussian random variables. It thus suffices to track their mean and covariance. The $\{\Delta_n(t)\}$ variables have zero mean and covariance given by the $Q_n(t,t')$ function. The $\{\boldsymbol{w}_n\}$ variables have the following mean evolution

$$\begin{aligned}\langle \boldsymbol{w}_{n+1}\rangle &= \langle \boldsymbol{w}_n\rangle + \eta\sqrt{\alpha}\left[\bar{\boldsymbol{\Sigma}}\boldsymbol{w}_R - \left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\langle \boldsymbol{w}_n\rangle\right] \\ &= \langle \boldsymbol{w}_n\rangle + \eta\sqrt{\alpha}\left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\left[\boldsymbol{w}_{TD} - \langle \boldsymbol{w}_n\rangle\right] \end{aligned} \tag{B.14}$$

where $\boldsymbol{w}_{TD} = \left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]^{-1}\bar{\boldsymbol{\Sigma}}\boldsymbol{w}_R$ is the fixed point of the TD dynamics. We next compute $\boldsymbol{M}_n = \left\langle (\boldsymbol{w}_n - \boldsymbol{w}_{TD})(\boldsymbol{w}_n - \boldsymbol{w}_{TD})^\top \right\rangle$ which admits the recursion

$$\boldsymbol{M}_{n+1} = \left(\boldsymbol{I} - \eta\sqrt{\alpha}\left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\right)\boldsymbol{M}_n\left(\boldsymbol{I} - \eta\sqrt{\alpha}\left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\right) + \frac{\eta^2}{T^2}\sum_{tt'} Q_n(t,t')\boldsymbol{\Sigma}(t,t') \tag{B.15}$$

To obtain our formulas which hold for finite batch size, we rescale the learning rate by $\eta \to \eta/\sqrt{\alpha}$ giving the following evolution

$$\begin{aligned}\langle \boldsymbol{w}_{n+1}\rangle &= \langle \boldsymbol{w}_n\rangle + \eta\left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\left[\boldsymbol{w}_{TD} - \langle \boldsymbol{w}_n\rangle\right] \\ \boldsymbol{M}_{n+1} &= \left(\boldsymbol{I} - \eta\left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\right)\boldsymbol{M}_n\left(\boldsymbol{I} - \eta\left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\right)^\top + \frac{\eta^2}{T^2\alpha^2}\sum_{tt'} Q_n(t,t')\boldsymbol{\Sigma}(t,t') \end{aligned} \tag{B.16}$$

After this rescaling, we see that the mean evolution for $\boldsymbol{w}_n$ is independent of α but that the variance picks up an additive term on each step on the order of $\mathcal{O}(\eta^2\alpha^{-2})$ which vanishes in the infinite batch limit $B/N \to \infty$. The error for value learning can be obtained from $\boldsymbol{M}_n$ with $\mathcal{L}_n = \frac{1}{N}\mathrm{Tr}\boldsymbol{M}_n\bar{\boldsymbol{\Sigma}}$. Lastly, we note that we can express the formula for $Q_n(t,t')$ entirely in terms of $\boldsymbol{M}_n$ and $\langle \boldsymbol{w}_n\rangle$. This

gives the lengthy expression

$$
\begin{aligned}
Q_n(t,t') &= \frac{1}{N}\left\langle (\boldsymbol{w}_R - \boldsymbol{w}_n)^\top \boldsymbol{\Sigma}(t,t')(\boldsymbol{w}_R - \boldsymbol{w}_n)\right\rangle + \frac{\gamma}{N}\left\langle (\boldsymbol{w}_R - \boldsymbol{w}_n)^\top \boldsymbol{\Sigma}(t,t'+1)\boldsymbol{w}_n\right\rangle \\
&+ \frac{\gamma}{N}\left\langle \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t+1,t')(\boldsymbol{w}_R - \boldsymbol{w}_n)\right\rangle + \frac{\gamma^2}{N}\left\langle \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t+1,t'+1)\boldsymbol{w}_n\right\rangle \\
&= \frac{1}{N}\mathrm{Tr}\boldsymbol{M}_n\boldsymbol{\Sigma}(t,t') + \frac{1}{N}\left(\boldsymbol{w}_{TD} - \langle \boldsymbol{w}_n\rangle\right)\left[\boldsymbol{\Sigma}(t,t') + \boldsymbol{\Sigma}(t',t)\right](\boldsymbol{w}_R - \boldsymbol{w}_{TD}) \\
&+ \frac{1}{N}\left(\boldsymbol{w}_R - \boldsymbol{w}_{TD}\right)^\top \boldsymbol{\Sigma}(t,t')\left(\boldsymbol{w}_R - \boldsymbol{w}_{TD}\right) \\
&- \frac{\gamma}{N}\mathrm{Tr}\boldsymbol{M}_n\left[\boldsymbol{\Sigma}(t,t'+1) + \boldsymbol{\Sigma}(t+1,t')\right] \\
&+ \frac{\gamma}{N}\left(\boldsymbol{w}_{TD} - \langle \boldsymbol{w}_n\rangle\right)\left[\boldsymbol{\Sigma}(t,t'+1) + \boldsymbol{\Sigma}(t+1,t')\right]\boldsymbol{w}_{TD} \\
&+ \frac{\gamma}{N}(\boldsymbol{w}_R - \boldsymbol{w}_{TD})^\top\left[\boldsymbol{\Sigma}(t,t'+1) + \boldsymbol{\Sigma}(t+1,t')\right]\langle \boldsymbol{w}_n\rangle \\
&+ \frac{\gamma^2}{N}\mathrm{Tr}\boldsymbol{M}_n\boldsymbol{\Sigma}(t+1,t'+1) + \frac{2\gamma^2}{N}\left(\langle \boldsymbol{w}_n\rangle - \boldsymbol{w}_{TD}\right)\boldsymbol{\Sigma}(t+1,t'+1)\boldsymbol{w}_{TD} \\
&+ \frac{\gamma^2}{N}\boldsymbol{w}_{TD}^\top\boldsymbol{\Sigma}(t+1,t'+1)\boldsymbol{w}_{TD}
\end{aligned} \tag{B.17}
$$

B.3 Final Result

Below we state in compact form the full final result for our TD learning curves. The below equations give the evolution of the first and second moments of $\boldsymbol{w}_n$ obtained from the mean-field density of the previous section. Concretely, these moments obey dynamics

$$
\begin{aligned}
\langle \boldsymbol{w}_{n+1}\rangle &= \langle \boldsymbol{w}_n\rangle + \eta\left[\bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+\right]\left[\boldsymbol{w}_V - \langle \boldsymbol{w}_n\rangle\right] \\
\boldsymbol{M}_{n+1} &= \left[\boldsymbol{I} - \eta\bar{\boldsymbol{\Sigma}} + \eta\gamma\bar{\boldsymbol{\Sigma}}_+\right]\boldsymbol{M}_n\left[\boldsymbol{I} - \eta\bar{\boldsymbol{\Sigma}} + \eta\gamma\bar{\boldsymbol{\Sigma}}_+\right]^\top + \frac{\eta^2}{\alpha^2 T^2}\sum_{tt'} Q_n(t,t')\boldsymbol{\Sigma}(t,t') \\
Q_n(t,t') &= \frac{1}{N}\left\langle (\boldsymbol{w}_R - \boldsymbol{w}_n)^\top \boldsymbol{\Sigma}(t,t')(\boldsymbol{w}_R - \boldsymbol{w}_n)\right\rangle + \frac{\gamma}{N}\left\langle (\boldsymbol{w}_R - \boldsymbol{w}_n)^\top \boldsymbol{\Sigma}(t,t'+1)\boldsymbol{w}_n\right\rangle \\
&+ \frac{\gamma}{N}\left\langle \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t+1,t')(\boldsymbol{w}_R - \boldsymbol{w}_n)\right\rangle + \frac{\gamma^2}{N}\left\langle \boldsymbol{w}_n^\top \boldsymbol{\Sigma}(t+1,t'+1)\boldsymbol{w}_n\right\rangle .
\end{aligned} \tag{B.18}
$$

These equations can be solved iteratively for $\bar{\boldsymbol{w}}_n, \boldsymbol{M}_n, Q_n$. Finite dimensional versions of this result can be obtained by replacing α with B/N as written in the main text. The value estimation error is

$$
\mathcal{L}_n = \frac{1}{N}\mathrm{Tr}\boldsymbol{M}_n\bar{\boldsymbol{\Sigma}}. \tag{B.19}
$$

B.4 Non-Zero Mean Feature

We can also simply modify the DMFT equations if the mean feature is nonvanishing $\boldsymbol{\mu}(s) \neq 0$. In this case, when averaging over all possible trajectories through state space, there is a mean feature vector at each episodic time $\boldsymbol{\mu}(t)$. The above equations are exact for non-zero mean features if $\boldsymbol{\Sigma}(t,t')$ is regarded as the (non-centered) correlation matrix $\langle \boldsymbol{\psi}(t)\boldsymbol{\psi}(t')\rangle$.

B.5 Tracking Iterate Moments with Direct Recurrence Relation

In this section we give a direct calculation of the first two moments of $\boldsymbol{w}$ over the collection of randomly sampled features $\{\boldsymbol{\psi}_n^\mu(t)\}$ and show which terms can be disregarded.

Letting $\boldsymbol{A} = \bar{\boldsymbol{\Sigma}} - \gamma\bar{\boldsymbol{\Sigma}}_+$, we note that the average evolution of $\boldsymbol{w}$ has the form

$$
\langle \boldsymbol{w}_{n+1}\rangle = \left(\boldsymbol{\Sigma} - \gamma\boldsymbol{\Sigma}_+\right)\left(\boldsymbol{w}_{TD} - \langle \boldsymbol{w}_n\rangle\right) \tag{B.20}
$$

Thus, if we disregarded fluctuations in $\boldsymbol{w}_n$ due to SGD, the model will converge to the correct fixed point. Next, we look at $\boldsymbol{M}_n = \langle (\boldsymbol{w}_n - \boldsymbol{w}_{TD})(\boldsymbol{w}_n - \boldsymbol{w}_{TD})\rangle$. Under the Gaussian equivalence

ansatz, we have

$$\begin{aligned}
\boldsymbol{M}_{n+1} &= \boldsymbol{M}_n - \eta \boldsymbol{A}\boldsymbol{M}_n - \eta \boldsymbol{M}_n \boldsymbol{A}^\top + \frac{\eta^2}{T^2 B^2} \sum_{\mu\nu tt'} \langle \Delta_n^\mu(t) \Delta_n^\nu(t') \psi_n^\mu(t) \psi_n^\nu(t') \rangle \\
&= (\boldsymbol{I} - \eta \boldsymbol{A}) \boldsymbol{M}_n (\boldsymbol{I} - \eta \boldsymbol{A})^\top - \frac{\eta^2}{B} \boldsymbol{A}\boldsymbol{M}_n \boldsymbol{A}^\top + \frac{\eta^2}{T^2 B} \sum_{tt'} \left\langle \Delta_n(t) \Delta_n(t') \boldsymbol{\psi}(t) \boldsymbol{\psi}(t')^\top \right\rangle \\
&= (\boldsymbol{I} - \eta \boldsymbol{A}) \boldsymbol{M}_n (\boldsymbol{I} - \eta \boldsymbol{A})^\top + \frac{\eta^2}{T^2 B} \sum_{tt'} Q_n(t, t') \boldsymbol{\Sigma}(t, t') \\
&+ \frac{\eta^2}{T^2 B} \sum_{tt'} \langle \Delta_n(t') \boldsymbol{\psi}(t) \rangle \left\langle \Delta_n(t) \boldsymbol{\psi}(t')^\top \right\rangle
\end{aligned} \tag{B.21}$$

The mean field theory derived from saddle point integration consists of the first two terms in the final expression. Therefore mean field theory disregards the last term which computes cross time correlations of RPEs with features, effectively making the approximation

$$\frac{\eta^2}{T^2 B} \sum_{tt'} \langle \Delta_n(t') \boldsymbol{\psi}(t) \rangle \left\langle \Delta_n(t) \boldsymbol{\psi}(t')^\top \right\rangle \approx 0. \tag{B.22}$$

After making this approximation, we recover the learning curve obtained in the previous Section B.3. We show in our experiments that dropping this term does not significantly alter the learning curves.

B.6 Scaling of Asymptotic Fixed Points

To identify fixed points in the value error dynamics, we can seek non-vanishing fixed points for the weight error covariance $\boldsymbol{M} = \langle (\boldsymbol{w} - \boldsymbol{w}_{TD})(\boldsymbol{w} - \boldsymbol{w}_{TD}) \rangle$. We note that $\langle \boldsymbol{w} \rangle \sim \boldsymbol{w}_{TD}$ asymptotically. Again, letting $\boldsymbol{A} = \bar{\boldsymbol{\Sigma}} - \gamma \bar{\boldsymbol{\Sigma}}_+$, we obtain the following fixed point condition for $\boldsymbol{M}$ under these assumptions

$$\begin{aligned}
&\boldsymbol{A}\boldsymbol{M} + \boldsymbol{M}\boldsymbol{A}^\top - \eta \boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^\top = \frac{\eta}{BT^2} \sum_{tt'} Q(t, t') \boldsymbol{\Sigma}(t, t') \\
&Q(t, t') = \mathrm{Tr} \boldsymbol{M} \boldsymbol{\Sigma}(t, t') - \gamma \mathrm{Tr} \boldsymbol{M} \left[\boldsymbol{\Sigma}(t, t' + 1) + \boldsymbol{\Sigma}(t + 1, t') \right] + \gamma^2 \mathrm{Tr} \boldsymbol{M} \boldsymbol{\Sigma}(t + 1, t' + 1) \\
&\quad + \gamma^2 \boldsymbol{w}_{TD}^\top \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\Sigma}}_+ \boldsymbol{\Sigma}(t, t') \bar{\boldsymbol{\Sigma}}_+ \bar{\boldsymbol{\Sigma}}^{-1} \boldsymbol{w}_{TD} + \gamma^2 \boldsymbol{w}_{TD}^\top \boldsymbol{\Sigma}(t + 1, t' + 1) \boldsymbol{w}_{TD} \\
&\quad + \gamma^2 \boldsymbol{w}_{TD} \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\Sigma}}_+ \left[\boldsymbol{\Sigma}(t, t' + 1) + \boldsymbol{\Sigma}(t + 1, t') \right] \boldsymbol{w}_{TD}.
\end{aligned} \tag{B.23}$$

Where we used the formula for $Q_n(t, t')$ from Appendix B.5, evaluated at $\langle \boldsymbol{w} \rangle = \boldsymbol{w}_{TD}$ and used the fact that $\boldsymbol{w}_R = \boldsymbol{w}_{TD} - \gamma \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\Sigma}}_+ \boldsymbol{w}_{TD}$. The solution $\boldsymbol{M} = 0$ is a valid fixed point for $\boldsymbol{M}$ in the $\eta \to 0$ and $B \to \infty$ limits because the constant terms on the right-hand side vanish. Similarly, if $\gamma = 0$ (which corresponds to the standard supervised learning case), the right hand side is linear in $\boldsymbol{M}$, allowing $\boldsymbol{M} = 0$ to be a valid fixed point.

However, for finite B and non-zero η and γ, there exists a solution to the above fixed point equation. For small $\frac{\eta \gamma^2}{B}$, we can easily deduce that $\boldsymbol{M}$ must satisfy a self-consistent asymptotic scaling of the form

$$\boldsymbol{M} = \mathcal{O}\left(\frac{\eta \gamma^2}{B} \right) \tag{B.24}$$

impliying an asymptotic value error scaling of $\mathcal{L} \sim \mathrm{Tr} \boldsymbol{M} \bar{\boldsymbol{\Sigma}} \sim \mathcal{O}\left(\frac{\eta \gamma^2}{B} \right)$. These scalings are examined in Figure 3 where experiments obey the expected behavior.

C Reward Shaping

In this section, we consider the role of reward shaping on the dynamics of TD learning. As discussed in the main text, we consider potential based shaping with potential function decomposable in the features $\phi(s) = \boldsymbol{w}_\phi \cdot \boldsymbol{\psi}(s)$. We first describe the change to the average weight evolution $\langle \boldsymbol{w}_n \rangle$ and

then describe the dynamics of the correlations. In potential based shaping, the TD errors take the form

$$\Delta(t) = R(s(t)) + \phi(s(t)) - \gamma\phi(s(t+1)) + \gamma\hat{V}(s(t+1)) - \hat{V}(s(t)) \tag{C.1}$$

Computing from the DMFT equations the evolution of $\langle \boldsymbol{w}_n \rangle$ we have

$$\begin{aligned} \langle \boldsymbol{w}_{n+1} \rangle &= \langle \boldsymbol{w}_n \rangle + \eta\bar{\boldsymbol{\Sigma}}(\boldsymbol{w}_R + \boldsymbol{w}_\phi - \langle \boldsymbol{w}_n \rangle) + \gamma\eta\bar{\boldsymbol{\Sigma}}_+(\langle \boldsymbol{w}_n \rangle - \boldsymbol{w}_\phi) \\ &= \langle \boldsymbol{w}_n \rangle - \eta \boldsymbol{A} \left[\boldsymbol{w}_{TD} + \boldsymbol{w}_\phi - \langle \boldsymbol{w}_n \rangle \right]. \end{aligned} \tag{C.2}$$

We see that including the reward shaping function ϕ offsets the fixed point of the algorithm to be $\boldsymbol{w}_{TD} + \boldsymbol{w}_\phi$. This occurs precisely because the potential-based reward shaping generates an additive correction to the target value function by $\phi(s)$ [61]. When we predict value at evaluation, we use the reshifted value $\hat{V}(s) - \phi(s)$. The natural quantity to track at the level of the mean field equations is the adapted version of $\boldsymbol{M}_n$

$$\boldsymbol{M}_n = \left\langle (\boldsymbol{w}_n - \boldsymbol{w}_{TD} - \boldsymbol{w}_\phi) (\boldsymbol{w}_n - \boldsymbol{w}_{TD} - \boldsymbol{w}_\phi)^\top \right\rangle. \tag{C.3}$$

This correlation matrix has dynamics

$$\boldsymbol{M}_{n+1} = (\boldsymbol{I} - \eta\boldsymbol{A}) \boldsymbol{M}_n (\boldsymbol{I} - \eta\boldsymbol{A})^\top + \frac{\eta^2}{BT^2} \sum_{tt'} Q_n(t,t') \boldsymbol{\Sigma}(t,t') \tag{C.4}$$

and the TD-error correlations $Q_n(t,t')$ have the form

$$\begin{aligned} Q_n(t,t') = &\left\langle (\boldsymbol{w}_R + \boldsymbol{w}_\phi - \boldsymbol{w}_n)^\top \boldsymbol{\Sigma}(t,t') (\boldsymbol{w}_R + \boldsymbol{w}_\phi - \boldsymbol{w}_n) \right\rangle \\ &+ \gamma \left\langle (\boldsymbol{w} - \boldsymbol{w}_\phi)^\top \left[\boldsymbol{\Sigma}(t,t') + \boldsymbol{\Sigma}(t',t) \right] (\boldsymbol{w}_R + \boldsymbol{w}_\phi - \boldsymbol{w}_n) \right\rangle \\ &+ \gamma^2 \left\langle (\boldsymbol{w}_n - \boldsymbol{w}_\phi)^\top \boldsymbol{\Sigma}(t+1, t'+1) (\boldsymbol{w}_n - \boldsymbol{w}_\phi) \right\rangle \end{aligned} \tag{C.5}$$

The value estimation error is again $\mathcal{L}_n = \mathrm{Tr} \boldsymbol{M}_n \bar{\boldsymbol{\Sigma}}$. We see that the two primary ways that reward shaping alters the loss dynamics is

- A change in the initial condition for $\boldsymbol{M}_n$ to be $\boldsymbol{M}_0 = (\boldsymbol{w}_{TD} + \boldsymbol{w}_\phi)(\boldsymbol{w}_{TD} + \boldsymbol{w}_\phi)^\top$
- A change in the TD error covariance term $Q_n(t,t')$

Both effects can generate significant changes in the dynamics and plateaus of the model.

D Numerical methods and additional details

The code to generate the Figures is provided in the Supplementary Material as a Jupyter Notebook. Here, we briefly highlight some of the parameter choices.

For Figures 3 and 4 we use diagonally decoupled, but temporally correlated power law features with $\Sigma_{k\ell}(t,t') = \delta_{k\ell}\, k^{-1.2} \exp\left(-|t-t'|/\tau_k\right)$ with $\tau_k = \frac{10}{k+1}$ and $w_k^R = k^{-1.1}$ for $k \in [N]$ with $N = 300$. This type of feature structure is especially easy to evaluate the theoretical learning curves for. Unless otherwise stated, these figures used $\gamma = 0.9$ and batch size $B = 10$.

For the 2D MDP grid world, we defined a discrete set of states on a 17×17 grid. The agent starts in the middle position and follows a random diffusion policy where each possible movement (up, down, left, right) is taken with equal probability. The features were generated as bell-shaped place cells (shown). We computed $\boldsymbol{\Sigma}(t,t')$ for the theory by sampling 5000 random draws of length $T = 50$. The Gaussian learning curve is obtained with TD learning with $\boldsymbol{\psi}_G \sim \mathcal{N}(0, \boldsymbol{\Sigma})$.

Numerical experiments were performed on a NVIDIA SMX4-A100-80GB GPU. Together numerical experiments (both preliminary experiments and those presented in the paper) took less than 1 hour of compute time.