On The Effect of Auxiliary Tasks on Representation Dynamics: Appendices

A Additional results

In this section, we state and prove some additional lemmas that are useful in proving the results stated in the main paper.

Lemma A.1. Let $x \in \mathbb{R}^d$, and let $(v_t)_{t\geq 0}$ be a sequence of vectors in $\mathbb{R}^d$ satisfying $v_t = f(t)x + o(f(t))$, for some function $f : [0, \infty) \to (0, \infty)$. Then $d(\langle v_t\rangle, \langle x\rangle) \to 0$ as $t \to \infty$.

Proof. The Grassmann distance $d(\langle v_t\rangle, \langle x\rangle)$ between two one-dimensional subspaces has a particular simple form, given by

$$d(\langle v_t\rangle, \langle x\rangle) = \min\left(\arccos\left(\frac{\langle v_t, x\rangle}{\|v_t\|\|x\|}\right), \arccos\left(\frac{\langle -v_t, x\rangle}{\|v_t\|\|x\|}\right)\right).$$

In our case, for sufficiently large t this yields

$$\begin{aligned} d(\langle v_t\rangle, \langle x\rangle) &= \arccos\left(\frac{\langle f(t)x + o(f(t)), x\rangle}{\|f(t)x + o(|f(t)|)\|\|x\|}\right) \\ &= \arccos\left(\frac{\langle x + o(1), x\rangle}{\|x + o(1)\|\|x\|}\right) \\ &\to \arccos\left(\frac{\langle x, x\rangle}{\|x\|\|x\|}\right) \\ &= 0\,. \end{aligned}$$

□

Lemma A.2. Let $U_1, \ldots, U_{|\mathcal{X}|}$ be a basis for $\mathbb{R}^{\mathcal{X}}$, let $K < |\mathcal{X}|$, and let $(a_{ij}|i \in [K], j \in [|\mathcal{X}|])$ be real coefficients. Let $0 < \beta_1 < \cdots < \beta_{|\mathcal{X}|}$, and consider time-dependent vectors $W_1(t), \ldots, W_d(t)$ defined by

$$W_i(t) = \sum_{j=1}^{|\mathcal{X}|} a_{ij} e^{-\beta_j t} U_j\,, \quad t \geq 0\,.$$

Then for almost all sets of coefficients $(a_{ij}|i \in [K], j \in [|\mathcal{X}|])$, we have

$$d(W_{1:K}(t), U_{1:K}) \to 0\,.$$

Proof. Without loss of generality, we may take the vectors $U_1, \ldots, U_{|\mathcal{X}|}$ to be the canonical basis vectors. Under the assumptions of the theorem, we exclude initial conditions for which the matrix A with $(i, j)^{\text{th}}$ element a_{ij} is not full rank. Note that under this condition, the matrix A_t with $(k, i)^{\text{th}}$ element $a_{ki}e^{\beta_i t}$ is also full rank for all but finitely many t. By performing row reduction operations and scaling rows, for all such t we may pass from $(W_k(t) \mid k \in [K])$ to an alternative spanning set $(\widetilde{W}_k(t) \mid k \in [K])$ of the same subspace such that $\widetilde{W}_k(t) - U_k \in \langle U_{K+1:|\mathcal{X}|}\rangle$, and $\|\widetilde{W}_k(t) - U_k\| = O(e^{-t(\beta_{K+1}-\beta_k)}) = o(1)$. We therefore obtain an orthonormal basis for this subspace of the form $U_1 + o(1), \ldots, U_K + o(1)$.

We now use the singular value decomposition characterisation of Grassmann distance in Definition 3.2. Since we have obtained an orthonormal basis for the subspace $\langle W_k(t) \mid k \in [K]\rangle$, the top-$K$ singular values of the matrix $(\sum_{k=1}^K U_k U_k^\top)(\sum_{k=1}^K (U_k + o(1))(U_k + o(1))^\top)$ determine the Grassmann distance. However, this matrix is equal to $\text{diag}(1, \ldots, 1, 0, \ldots, 0) + o(1)$, with K entries of 1 in the diagonal matrix. But the top-K singular values this matrix are $1 + o(1)$, and so the principal angles between the subspaces are $o(1)$, and hence the Grassmann distance between the subspaces is $o(1)$, as required. □

Lemma A.3. For $M \in \mathbb{N}$, let $(r^m)_{m=1}^M$ be independent random variables drawn from some fixed mean-zero distribution in $\mathscr{P}(\mathbb{R}^{\mathcal{X}\times\mathcal{A}})$ such that the covariance between coordinates $(x,a),(y,a)$ is Σ_{xy}, independent of $a \in \mathcal{A}$. Let $(\mathbf{w}^m)_{m=1}^M$ be independent random variables taking values in $\mathbb{R}^{K\times\mathcal{A}}$, with columns drawn independently from $\mathcal{N}(0,(1/M)I)$. Then $\sum_{m=1}^M r^m(\mathbf{w}^m)^\top$ converges (in distribution) to a mean-zero Gaussian distribution over $\mathbb{R}^{\mathcal{X}\times K}$, with independent columns, and individual columns having covariance matrix Σ.

Proof. The proof simply follows by noting that $\sum_{m=1}^M r^m\mathbf{w}^m$ may be written $1/\sqrt{M}\sum_{m=1}^M r^m\varepsilon^m$, with $(\varepsilon^m)_{m=1}^\infty$ i.i.d. $N(0,I)$ random variables. The individual terms have the desired mean and variance, and the resulting converge in distribution now follows from the central limit theorem. □

Lemma A.4. For fixed M, let $(\mathbf{w}^m)_{m=1}^M$, $\mathbf{w}^m \in \mathbb{R}^d$, be sampled i.i.d. according to $\mathcal{N}(0,\frac{1}{M}I)$. Then the following hold.

$$\lim_{M\to\infty}\sum_{m=1}^M \mathbf{w}^m(\mathbf{w}^m)^\top = I \text{ and } \lim_{M\to\infty}\sum_{m=1}^M \mathbf{w}^m \overset{D}{=} \epsilon \sim \mathcal{N}(0,I) \tag{17}$$

Proof. We prove two results on the limit of $W = \sum_{m=1}^M \mathbf{w}^m(\mathbf{w}^m)^\top$ as $k\to\infty$. First

$$\lim_{M\to\infty}\sum_{m=1}^M \mathbf{w}^m(\mathbf{w}^m)^\top \overset{P}{=} I\,,$$

which we observe by evaluating an arbitrary diagonal and off-diagonal element of $\sum_{m=1}^M \mathbf{w}^m(\mathbf{w}^m)^\top$. For the diagonal terms, note that

$$\left(\sum_{m=1}^M \mathbf{w}^m(\mathbf{w}^m)^\top\right)[j,j] = \sum_{m=1}^M (\mathbf{w}_j^m)^2$$

Now observe that

$$\mathbb{E}\left[\sum_{m=1}^M (\mathbf{w}_j^m)^2\right] = M\frac{1}{M} = 1\,,\text{ and }\quad \mathrm{Var}\left(\sum_{m=1}^M (\mathbf{w}_j^m)^2\right) = M\frac{1}{M^2}\to 0$$

Similarly, for the off-diagonal terms, let $j\neq\ell$. Then we have

$$\left(\sum_{m=1}^M \mathbf{w}^m(\mathbf{w}^m)^\top\right)[j,\ell] = \sum_{m=1}^M \mathbf{w}_j^m\mathbf{w}_\ell^m\,,$$

and further

$$\mathbb{E}\left[\sum_{m=1}^M \mathbf{w}_j^m\mathbf{w}_\ell^m\right] = 0\,,\text{ and }\quad \mathrm{Var}\left(\sum_{m=1}^M \mathbf{w}_\ell^m\mathbf{w}_j^m\right) = M\frac{1}{M^2}\to 0;$$

The limit in probability is immediately implied by Chebyshev's inequality. The result on $\sum_{m=1}^M \mathbf{w}^m$ follows immediately from part 1 and the fact that a sum of Gaussian random variables is another Gaussian random variable whose mean and variance in this case will be a standard normal. □

B Proofs

Lemma 3.1. If $(V_t)_{t\geq 0}$ satisfies Equation (3) with initial condition V_0 at time $t=0$, then we have

$$V_t = \exp(-t(I-\gamma P^\pi))(V_0 - V^\pi) + V^\pi\,. \tag{4}$$

Proof. Equation (4) can be verified as a solution to Equation (3) by direct differentiation. Uniqueness of the solution follows since this is an autonomous initial value problem that satisfies the Lipschitz condition, and so the Picard-Lindelhöf theorem applies. □

Proposition 3.4. Under Assumption 3.3, and $(V_t)_{t\geq 0}$ the solution to Equation (3), for almost every[3] initial condition V_0, we have

$$d(\langle V_t - V^\pi \rangle, \langle U_1 \rangle) \to 0\,.$$

Proof. By Assumption 3.3, P^π is diagonalisable, with eigenbasis $U_1, \ldots, U_{|\mathcal{X}|}$, with corresponding eigenvalues $\lambda_{1:|\mathcal{X}|}$ with strictly decreasing magnitudes $|\lambda_1| > \cdots > |\lambda_{|\mathcal{X}|}|$. We note then that $\exp -(t(I - \gamma P^\pi))$ is also diagonaisable under the same basis, with eigenvalues $\exp(t(\gamma\lambda_i - 1))$, for $i = 1, \ldots, |\mathcal{X}|$. We may therefore expand V_0 with respect to this eigenbasis, and write

$$V_0 - V^\pi = \sum_{i=1}^{|\mathcal{X}|} \alpha_i U_i\,,$$

for some $\alpha_{1:|\mathcal{X}|} \in \mathbb{R}^{|\mathcal{X}|}$. Now note from the differential equation (4), we have

$$V_t - V^\pi = \exp(-t(I - \gamma P^\pi))(V_0 - V^\pi) = \sum_{i=1}^{|\mathcal{X}|} \alpha_i \exp(t(\gamma\lambda_i - 1))U_i\,.$$

Note that as P^π is a stochastic matrix, we have $|\lambda_i| \leq 1$ for all $i = 1, \ldots, |\mathcal{X}|$, and hence $\exp(t(\gamma\lambda_i - 1)) \to 0$ for all $i = 1, \ldots, |\mathcal{X}|$. Further, $\exp(t(\gamma\lambda_i - 1)) = o(\exp(t(\gamma\lambda_1 - 1)))$ for all $i = 2, \ldots, |\mathcal{X}|$. We make the additional assumption that $\alpha_1 \neq 0$, which makes the 'almost every initial condition' assumption in the statement precise. Under this assumption, we therefore have

$$V_t - V^\pi = \alpha_1 \exp(t(\gamma\lambda_1 - 1))U_1 + \sum_{i=2}^{|\mathcal{X}|} \alpha_i \exp(t(\gamma\lambda_i - 1))U_i = \alpha_1 \exp(t(\gamma\lambda_1 - 1))U_1 + o(\exp(t(\gamma\lambda_1 - 1)))\,.$$

Then Lemma A.1 applies to give $d(\langle V_t - V^\pi \rangle, \langle U_1 \rangle) \to 0$, as required. □

Proposition 3.5. Under Assumption 3.3, and $(V_t^{(k)})_{t\geq 0}$ the solution to Equation (3) for each $k = 1, \ldots, K$, for almost every initial condition $(V_0^{(k)})_{k=1}^K$, we have

$$d(\langle V_t^{(k)} - V^\pi \mid k \in [K] \rangle, \langle U_{1:K} \rangle) \to 0\,.$$

Proof. Expanding $V_0^{(k)} - V^\pi$ with respect to $U_1, \ldots, U_{|\mathcal{X}|}$ for each $k = 1, \ldots, |\mathcal{X}|$, we obtain expressions of the form

$$V_0^{(k)} - V^\pi = \sum_{i=1}^{|\mathcal{X}|} a_{ki} U_i\,.$$

By the ODE solution in Lemma 3.1, we then have

$$V_t^{(k)} - V^\pi = \sum_{i=1}^{|\mathcal{X}|} a_{ki} e^{-t(1-\gamma\lambda_i)} U_i\,.$$

We may now apply Lemma A.2 to obtain the desired result. □

Lemma 3.6. Let Φ_t and $\mathbf{w}_t$ parameterize a value function approximator as defined above. Then

$$\partial_t \Phi_t = \alpha(R^\pi + \gamma P^\pi \Phi_t \mathbf{w}_t - \Phi_t \mathbf{w}_t)\mathbf{w}_t^\top\,, \tag{7}$$

$$\partial_t \mathbf{w}_t = \beta \Phi_t^\top (R^\pi + \gamma P^\pi \Phi_t \mathbf{w}_t - \Phi_t \mathbf{w}_t)\,. \tag{8}$$

Proof. This follows immediately by computing the derivatives in Equations (5) & (6), and so we omit the direct calculations. □

[3] In the measure-theoretic sense that the set of excluded initial conditions V_0 has Lebesgue measure 0.

Theorem 4.1. For $M \in \mathbb{N}$, let $(\Phi_t^M)_{t\geq 0}$ be the solution to Equation (9), with each $\mathbf{w}_t^m$ for $m = 1, \ldots, M$ initialised independently from $N(0, \sigma_M^2)$, and fixed throughout training ($\beta = 0$). We consider two settings: first, where the learning rate α is scaled as $\frac{1}{M}$ and $\sigma_M^2 = 1$ for all M, and second where $\sigma_M^2 = \frac{1}{M}$ and the learning rate α is equal to 1. These two settings yield the following dynamics, respectively:

$$\lim_{M\to\infty} \partial_t \Phi_t^M \overset{P}{=} -(I - \gamma P^\pi)\Phi_t^M \quad \text{, and} \tag{11}$$

$$\lim_{M\to\infty} \partial_t \Phi_t^M \overset{D}{=} -(I - \gamma P^\pi)\Phi_t^M + R^\pi \epsilon^\top \text{ , } \epsilon \sim \mathcal{N}(0, I)\,. \tag{12}$$

The corresponding limiting trajectories for a fixed initialisation $\Phi_0 \in \mathbb{R}^{\mathcal{X}\times K}$, are therefore given respectively by

$$\lim_{M\to\infty} \Phi_t^M \overset{P}{=} \exp(-t(I - \gamma P^\pi))\Phi_0 \quad \text{, and} \tag{13}$$

$$\lim_{M\to\infty} \Phi_t^M \overset{D}{=} \exp(-t(I - \gamma P^\pi))(\Phi_0 - (I - \gamma P^\pi)^{-1} R^\pi \varepsilon^\top) + (I - \gamma P^\pi)^{-1} R^\pi \varepsilon^\top \text{ , } \epsilon \sim \mathcal{N}(0, I)\,. \tag{14}$$

Proof. We write the dynamics on Φ_t^M as follows and apply the results of Lemma A.4. We first consider the scaled initialization setting (implicitly setting the learning rate $\alpha = 1$), where we find

$$\partial_t \Phi_t^M = (I - \gamma P^\pi)\Phi_t^M \sum_{m=1}^M \mathbf{w}^m (\mathbf{w}^m)^\top + \sum_{m=1}^M R^\pi (\mathbf{w}^m)^\top \tag{18}$$

$$\lim_{M\to\infty} \partial_t \Phi_t^M = (I - \gamma P^\pi)\Phi_t^M \lim_{M\to\infty} \sum_{m=1}^M \mathbf{w}^m (\mathbf{w}^m)^\top + \lim_{M\to\infty} R^\pi (\sum_{m=1}^M \mathbf{w}^m)^\top \tag{19}$$

$$\overset{D}{=} (I - \gamma P^\pi)\Phi_t^M I + R^\pi \epsilon^\top,\ \epsilon \sim \mathcal{N}(0, I). \tag{20}$$

We further observe that, for any finite interval, in the setting of zero reward we obtain *uniform* convergence of the induced trajectory Φ_t^M to the trajectory of the limiting dynamics. We first observe that for a fixed initialization, we have that the induced dynamics are linear (in the zero-reward setting, affine otherwise) function of Φ_t^M, and so

$$\partial_t \Phi_t^M = (I - \gamma P^\pi)\Phi_t^M \sum_{m=1}^M w^m (w^m)^\top = \mathcal{L}^M \Phi_t^M$$

$$\text{where } \mathcal{L}^M(A) = (I - \gamma P^\pi) A \sum_{m=1}^M w^m (w^m)^\top$$

$$\implies \Phi_t^M = \exp(t\mathcal{L}^M)\Phi_0^M\,.$$

Because the function $t \mapsto \exp(tA)$ is Lipschitz on a bounded interval for any A, this implies that for any finite interval $[0, T]$, the functions $t \mapsto \Phi_t^M$, as well as limiting solution, are L-Lipschitz for some L. Further, since the exponential is continuous,

$$\lim_{M\to\infty} \Phi_t^M = \lim_{M\to\infty} \exp(t\mathcal{L}^M)\Phi_0^M = \exp(t \lim_{M\to\infty} \mathcal{L}^M)\Phi_0$$
$$= \exp(-t(I - \gamma P^\pi))\Phi_0 = \Phi_t^\infty\,.$$

Therefore, the functions $t \mapsto \Phi_t^M$ are L-Lipschitz and converge to the limit Φ_t^∞ on the interval $[0, T]$, which implies that they converge uniformly.

To evaluate the scaled learning rate setting, we observe that we now have

$$\partial_t \Phi_t^M = \frac{1}{M}(I - \gamma P^\pi)\Phi_t^M \sum_{m=1}^M \mathbf{w}^m (\mathbf{w}^m)^\top + \sum_{m=1}^M R^\pi (\mathbf{w}^m)^\top \tag{21}$$

$$\lim_{M\to\infty} \partial_t \Phi_t^M = (I-\gamma P^\pi)\Phi_t^M \lim_{M\to\infty} \frac{1}{M}\sum_{m=1}^{M} \mathbf{w}^m(\mathbf{w}^m)^\top + \lim_{M\to\infty}\frac{1}{M} R^\pi (\sum_{m=1}^{M} \mathbf{w}^m)^\top \tag{22}$$

$$= (I-\gamma P^\pi)\Phi_t^M I. \tag{23}$$

$$\implies \lim_{M\to\infty} \Phi_t^M = \exp(-t(I-\gamma P^\pi))\Phi_0\,, \tag{24}$$

almost surely. The principal difference between this and the scaled initialization setting is that here we divide the $R^\pi \mathbf{w}^\top$ term by $\frac{1}{M}$, whereas the scaled initialization is equivalent to scaling by $\frac{1}{\sqrt{M}}$. Therefore the scaled learning rate limit can be computed by the law of large numbers and converges in probability to its mean (zero), whereas under the scaled initialization it converges via the central limit theorem to a Gaussian distribution. □

Corollary 4.2. Under the feature flow (9) with $\mathbf{w}_t^m$ fixed at initialization for each $i = 1, \dots, M$ and Assumption 3.3, for almost all initialisations Φ_0, we have when $R^\pi = 0$

$$d(\langle \Phi_t \rangle, \langle U_{1:K} \rangle) \to 0\,, \quad \text{as } t \to \infty.$$

Proof. As described in the proof of Theorem 4.1, we have $\Phi_t = \exp(-t(I-\gamma P^\pi))(\Phi_0 - \Phi_\infty) + \Phi_\infty$. Under Assumption 3.3, we may now apply an analogous argument as in Proposition 3.5 to the columns of $\Phi_t - \Phi_\infty$, and apply Lemma A.2 to obtain the desired result. □

Theorem 4.3. For fixed $M \in \mathbb{N}$, let the random rewards $(r^m)_{m=1}^M$ and weights $(\mathbf{w}^m)_{m=1}^M$ be as defined above, let $\alpha = 1$, and consider the representation dynamics in Equation (15), with weights fixed throughout training ($\beta = 0$). Let Σ denote the covariance matrix of the random cumulant distribution. Then

$$\lim_{M\to\infty} \sum_{m=1}^{M} r^m(\mathbf{w}^m)^\top \overset{D}{=} Z_\Sigma \sim \mathcal{N}(0, \Sigma), \text{ and}$$

$$\lim_{M\to\infty} \Phi_t^M \overset{D}{=} \exp(-t(I-\gamma P^\pi))(\Phi_0 - (I-\gamma P^\pi)^{-1} Z_\Sigma) + (I-\gamma P^\pi)^{-1} Z_\Sigma\,.$$

As the columns of Z_Σ are mean-zero, uncorrelated, with covariance matrices Σ, the limiting distribution of each column of $\Phi_\infty = \lim_{t\to\infty}\lim_{M\to\infty}\Phi_t^M$ has covariance $\Psi\Sigma\Psi^\top$, where Ψ is the resolvent $(I-\gamma P^\pi)^{-1}$.

Proof. We recall from Theorem 4.1 that the limiting dynamics follow the distribution

$$\lim_{t\to\infty}\lim_{M\to\infty} \Phi_t \overset{D}{=} \lim_{t\to\infty} \exp(-t(I-\gamma P^\pi))(\Phi_0 - (I-\gamma P^\pi)^{-1} Z_\Sigma) + (I-\gamma P^\pi)^{-1} Z_\Sigma \tag{25}$$

$$\overset{D}{=} (I-\gamma P^\pi)^{-1} Z_\Sigma \tag{26}$$

for which we can straightforwardly apply known properties of Gaussian distributions: namely, that the distribution of a linear transformation A of a Gaussian random variable with parameters μ, Σ is also Gaussian with mean $A\mu$ and covariance $A\Sigma A^\top$. Letting $A = (I-\gamma P^\pi)$ therefore gives the desired result. □

Corollary 4.4. Under the feature flow (15) with $\mathbf{w}_t^m$ fixed at initialization for each $i = 1, \dots, M$ and Assumption 3.3, for almost all initialisations Φ_0, we have when $R^\pi = 0$

$$d(\lim_{M\to\infty} \langle \Phi_t^M - \Phi_\infty \rangle, \langle U_{1:K} \rangle) \to 0\,, \quad \text{as } t \to \infty.$$

Proof. As described in the proof of Theorem 4.3, we have $\Phi_t = \exp(-t(I-\gamma P^\pi))(\Phi_0 - (I-\gamma P^\pi)^{-1} Z_\Sigma) + (I-\gamma P^\pi)^{-1} Z_\Sigma$. Under Assumption 3.3, we may now apply an analogous argument as in Proposition 3.5 to the columns of $\Phi_t - (I-\gamma P^\pi)^{-1} Z_\Sigma$, and apply Lemma A.2 to obtain the desired result. □

C Additional results from Table 1

We begin this section by noting the following property of systems following linear dynamics.

Lemma C.1. Let $\Phi_t \in \mathbb{R}^{\mathcal{X}\times M}$ follow the dynamics $\partial_t \Phi_t \overset{D}{=} A\Phi_t + B$, where A is a linear operator for which all eigenvalues have negative real part, and B is a vector. Then

$$\lim_{t\to\infty} \Phi_t = -A^{-1}B\,. \tag{27}$$

Further, if A is diagonalisable, with all eigenvalues of different magnitudes,

$$\lim_{t\to\infty} d(\langle \Phi_t - \Phi_\infty \rangle, \langle U_{1:K}(A) \rangle) = 0\,, \tag{28}$$

where $U_i(A)$ is the eigenvector of A corresponding to the eigenvalue with i^{th} largest magnitude.

Proof. We observe that the dynamics $\partial_t \Phi_t = A\Phi_t$ induce the trajectory

$$\Phi_t = \exp(tA)\Phi_0 + (I - \exp(tA))(-A^{-1}B)\,, \tag{29}$$

with limit $\Phi_\infty = -A^{-1}B$. When A is diagonalizable, we can therefore straightforwardly apply the results of Lemma A.2 to get that the limiting subspace will be characterized by the top k eigenvectors of A. In the settings we are interested in, $A = -(I - \gamma P^\pi)$ for some π and some γ, and so the principal eigenvectors of A will be the principal eigenvectors of P^π. □

The following two theorems characterize the learning dynamics under the past policies and multiple timescale auxiliary tasks listed in Table 1. With these characterizations, it becomes straightforward to deduce Φ_∞ and the limiting subspace error as a direct consequence of the previous lemma.

Theorem C.2. Let $\pi_1, \ldots, \pi_L$ be a fixed set of policies. Given fixed M and L, we define the indexing function $i_m = \lceil \frac{L}{m} \rceil$ for $m \in [1, M]$. Let Φ_t^M follow the dynamics

$$\partial_t \Phi_t^M = \sum_{m=1}^{M} -((I - \gamma P^{\pi_{i_m}})\Phi_t^M \mathbf{w}_t^m + R^{\pi_{i_m}})(\mathbf{w}_t^m)^\top \tag{30}$$

Then Φ_t^M satisfies the following dynamics and trajectory in the limit as $M \to \infty$, where $\bar{\pi} = \sum_{i=1}^{L} \pi_i$ and $\epsilon_i \in \mathbb{R}^d$ is an isotropic Gaussian with variance $\frac{1}{L}$. Note that we cannot naively average the rewards without changing the variance of the induced distribution unless $R^{\pi_i} = R^{\pi_j}$ for all i, j.

$$\lim_{M\to\infty} \partial_t \Phi_t^M \overset{D}{=} -(I - \gamma P^{\bar{\pi}})\Phi_t + \sum_{i=1}^{L} R^{\pi_i}\epsilon_i \tag{31}$$

$$\lim_{M\to\infty} \Phi_t^M \overset{D}{=} \exp(-t(I - \gamma P^{\bar{\pi}}))(\Phi_0 - \Phi_\infty) + (I - \gamma P^{\bar{\pi}})^{-1}\left(\sum_{i=1}^{L} R^{\pi_i}\epsilon_i^\top\right) \tag{32}$$

Proof. The result on the trajectories follows immediately from the result on the dynamics, so it suffices to prove convergence of the dynamics. We approach this problem by decomposing the dynamics of Φ_t^M as follows.

$$\partial_t \Phi_t^M = \sum_{m=1}^{M} -(I - \gamma P^{\pi_{i_m}})\Phi_t^M \mathbf{w}_t^m (\mathbf{w}_t^m)^\top - \sum_{m=1}^{M} R^{\pi_{i_m}} (\mathbf{w}_t^m)^\top\,. \tag{33}$$

We first consider the random variables in the term which includes the rewards R^π. For this, we can directly apply the results from the previous theorems to the random variables $\epsilon_j = \sum_{m:i_m=j} w^m$, whose limiting variance is easily computed to be

$$\lim_{M\to\infty} \mathrm{Var}\left(\sum_{m:i_m=j} \mathbf{w}^m\right) = \lim_{M\to\infty} \sum_{\lfloor \frac{j}{n} M \rfloor}^{\lfloor \frac{j+1}{n} M \rfloor} \frac{1}{M} I = \frac{1}{L} I\,. \tag{34}$$

For the term which depends on Φ_t, we see

$$\sum_{m=1}^{M} -(I - \gamma P^{\pi_{i_m}})\Phi_t^M \mathbf{w}_t^m (\mathbf{w}_t^m)^\top = \sum_{i=1}^{L} \sum_{m:i_m=i}^{M} -(I - \gamma P^{\pi_{i_m}})\Phi_t^M \mathbf{w}_t^m (\mathbf{w}_t^m)^\top \tag{35}$$

$$= \sum_{i=1}^{L} -(I - \gamma P^{\pi_i})\Phi_t^M \sum_{m:i_m=i}^{M} \mathbf{w}_t^m (\mathbf{w}_t^m)^\top . \tag{36}$$

Since L is finite and fixed, $\sum_{m:i_m=i}^{M} \mathbf{w}_t^m (\mathbf{w}_t^m)^\top$ converges to $\frac{1}{L} I$

$$\xrightarrow[M\to\infty]{} \sum_{i=1}^{L} -(I - \gamma P^{\pi_i})\Phi_t^M \frac{1}{L} I \tag{37}$$

$$= -(I - \gamma \frac{1}{L} \sum_{i=1}^{L} P^{\pi_i})\Phi_t^M \tag{38}$$

$$= -(I - \gamma P^{\bar{\pi}})\Phi_t^M . \tag{39}$$

And so the limiting distribution becomes

$$\lim_{M\to\infty} \partial_t \Phi_t^M = -(I - \gamma P^{\bar{\pi}})\Phi_t^M - \left(\sum_{i=1}^{L} R^{\pi_i} \epsilon_i \right). \tag{40}$$

□

Corollary C.3. The above result can be readily adapted to the setting in which each head predicts a randomly selected (deterministic) policy in MDPs with finite state and action spaces. Let $L = |\mathcal{A}|^{|\mathcal{X}|}$, $\{\pi_1, \ldots, \pi_L\}$ be an enumeration of $\mathcal{A}^{\mathcal{X}}$, and i_m denote the index of the policy randomly assigned to head m; then the above result still holds, and $\bar{\pi}$ is the uniform policy.

Theorem C.4. We consider the task of predicting the value functions of a fixed policy under multiple discount rates $\gamma_1, \ldots, \gamma_L$. For fixed M, L, let i_m denote the indexing function defined in Theorem C.2 Let Φ_t follow the dynamics

$$\partial_t \Phi_t^M = \sum_{m=1}^{M} -((I - \gamma_{i_m} P^\pi)\Phi_t^M \mathbf{w}_t^m + R^\pi)(\mathbf{w}_t^m)^\top . \tag{41}$$

Then the limiting dynamics as $M \to \infty$ of Φ_t^M are as follows, where $\bar{\gamma} = \sum \frac{1}{L}\gamma_i$

$$\lim_{M\to\infty} \partial_t \Phi_t^M \overset{D}{=} -(I - \bar{\gamma} P^\pi)\Phi_t + R^\pi \epsilon^\top \tag{42}$$

and

$$\lim_{M\to\infty} \Phi_t^M \overset{D}{=} \exp(-t(I - \gamma P^{\bar{\pi}}))(\Phi_0 - \Phi_\infty) + (I - \gamma P^\pi)^{-1} R^\pi \epsilon^\top) . \tag{43}$$

Proof. We follow a similar derivation as for Theorem C.2 in deriving the component of the dynamics which depends on Φ_t^M. The result of Theorem 4.1 immediately applies to the $\sum R^\pi (\mathbf{w}_t^m)^\top$ term:

$$\sum_{m=1}^{M} -(I - \gamma_{i_m} P^\pi)\Phi_t^M \mathbf{w}_t^m (\mathbf{w}_t^m)^\top = \sum_{i=1}^{L} \sum_{m:i_m=i}^{M} -(I - \gamma_i P^\pi)\Phi_t^M \mathbf{w}_t^m (\mathbf{w}_t^m)^\top \tag{44}$$

$$= \sum_{i=1}^{L} -(I - \gamma_i P^\pi)\Phi_t^M \sum_{m:i_m=i}^{M} \mathbf{w}_t^m (\mathbf{w}_t^m)^\top . \tag{45}$$

Since L is finite and fixed, $\sum_{m:i_m=i}^{M} \mathbf{w}_t^m (\mathbf{w}_t^m)^\top$ converges to $\frac{1}{L} I$ as before:

$$\xrightarrow[M\to\infty]{} \sum_{i=1}^{L} -(I - \gamma_i P^\pi)\Phi_t^M \frac{1}{L} I \tag{46}$$

$$= -(I - \sum_{i=1}^{L} \frac{\gamma_i}{L} P^\pi)\Phi_t^M \tag{47}$$

$$= -(I - \bar{\gamma} P^\pi)\Phi_t^M \,. \tag{48}$$

□

D Experimental details

D.1 Experimental details for Figure 2

In our evaluations of the evolution of single feature vectors, we compute the continuous-time feature evolution defined in Equation (5), using P^π defined by a random walk on a simple Four-Rooms Gridworld with no reward. We use a randomly initialized representation $\Phi \in \mathbb{R}^{|\mathcal{X}| \times 10}$, and use a single column of this matrix in our feature visualization (we observed similar behaviour in each feature). To compute trajectories, we use the SciPy ODE solver `solve_ivp` (Virtanen et al., 2020).

D.2 Experimental details for Section 5.1

Here, we provide details of the environment used in producing Figure 5.1. The environment is a 30-state chain, with two actions, `left` and `right`, which move the agent one state to the left or right, respectively. When the agent cannot move further left or right (due to being at an end state of the chain), the result of the corresponding action keeps the agent in the same state. There is additionally environment stochasticity of 0.01, meaning that with this probability, a uniformly random action is executed instead. This stochasticity ensures that P^π satisfies the conditions of Assumption 3.3. Taking the action left in the left-most state incurs a reward of $+2$, and taking the action `right` in the right-most state incurs a reward of $+1$; all other rewards are zero.

D.3 Experimental details for Section 5.2

We modify a base Double DQN agent (Van Hasselt et al., 2016) and evaluate on the ALE without sticky actions (Bellemare et al., 2013). Our agents are implemented in Jax (Bradbury et al., 2018), and are based on the DQN Zoo (Quan and Ostrovski, 2020). Unless otherwise mentioned, all hyperparameters are as for the default Double DQN agent, with the exception of the epsilon parameter in the evaluation policy, which is set to 0.001 in all agents, and the optimizer, which for agents using auxiliary tasks CV, REM and Ensemble is Adam with epsilon $0.1/32^2$, and a lightly tuned learning rate; see below for further details.

Experimental results shown in bar plots, such as Figures 5 and 7, report a "relative score" which is the per-game score normalized by the maximum average score achieved by any agent or configuration. The same, per-game, normalization values are used for all such figures.

Auxiliary task details. In this section, we describe the implementations of all auxiliary tasks considered in the main paper.

- *QR-DQN.* The implementation and hyperparameters match QR-DQN-1 in Dabney et al. (2018).
- *DDQN+RC.* We use a many-head DQN network which is identical to the standard neural network used for DQN, except that the output dimension is $(M+1) \times |\mathcal{A}|$ instead of $|\mathcal{A}|$, where M is the number of auxiliary heads. Random cumulants are generated using a separate neural network with the same architecture as a standard DQN, but with output dimension equal to the number of auxiliary heads. The width of the Huber loss for each auxiliary head is equal to the number of auxiliary tasks. Let $\phi(x) \in \mathbb{R}^M$ be the output of the cumulant network given input observation x, with M the number of auxiliary heads. Then the cumulant for auxiliary head m, at time step t, is given by $c_t = s \times (\phi(x_{t+1}) - \phi(x_t))$, where $s \in \mathbb{R}$ is a scaling factor. We

performed a small hyperparameter sweep over scaling factors in $\{1, 10, 100, 500\}$, finding $s = 100$ to provide the best performance and use this value for all reported experiments. Note that this auxiliary task and the details are nearly identical to the *CumulantValues* auxiliary task of Dabney et al. (2020), except that we do not pass the values through a tanh non-linearity as this did not appear to have any impact in practice. We performed a hyperparameter sweep over learning rates and gradient norm clipping for this agent, considering learning rates $\{0.00025, 0.0001, 0.00005\}$ and gradient clipping in $\{10, 40\}$. We found that a learning rate of 0.00005 and gradient norm clipping of 40 to work best and use these values for all experiments.

- *DDQN+REM.* We use a many-head variant of Double DQN, with heads trained according to the REM loss of Agarwal et al. (2020). For the agent's policy, an argmax over a uniform average of the heads is used. We swept over learning rates of 0.0001 and 0.00005, generally finding 0.00005 to perform best.
- *DDQN+Ensemble.* As for the REM auxiliary task, we use a many-head variant of Double DQN. Each head is trained using its own double DQN loss, and the resulting losses are averaged. For the agent's policy, an argmax over a uniform average of the heads is used. We swept over learning rates of 0.0001 and 0.00005, generally finding 0.00005 to perform best.

Modified dense-reward games. We modified four Atari games (Pong, MsPacman, Seaquest, and Q*bert) to obtain sparse, harder versions of these games to test the performance of random cumulants and other auxiliary tasks. The details of these games are given below. In each case a low-valued, commonly encountered reward is 'censored', which means that during training the agent observes a reward of 0 instead of the targeted reward. When evaluated, and thus for all empirical results reported, the standard uncensored rewards are reported.

- *Sparse Pong.* All negative rewards are censored (i.e. set to 0 before being fed to the agent), so the agent receives a reward of $+1$ for scoring against the opponent, but no reward when it concedes a point to the opponent. As 0, 1, and -1 are the only rewards in Pong, this modification makes Pong significantly harder. The agent can no longer learn to 'avoid losing points', but can only improve by learning to score points directly.
- *Sparse MsPacman.* All rewards less than or equal to 10 are censored. This corresponds to rewards for the numerous small pellets that MsPacman eats, but not the larger pellets or ghosts. Each level ends when all of the small pellets are consumed, thus, by hiding these from the agent we may have significantly changed the primary incentive for the agent to advance the game.
- *Sparse Seaquest.* All rewards less than or equal to 20 are censored. This corresponds to the rewards for shooting the sharks underwater, but not the rewards for picking up divers or surfacing. Additionally, even the rewards for sharks increase beyond this level, and thus become visible, once the agent has surfaced and collected enough divers.
- *Sparse Q*bert.* All rewards less than or equal to 25 are censored. These are the rewards for flipping the colour of a tile, which is the primary source of reward and the mechanism for advancing to the next level of the game. Once all tiles are flipped, the agent will go to the next level. However, the agent can still observe rewards for going to the next level and for dispatching the enemies.

As described in the main paper,we found that the sparse versions of MsPacman, Seaquest, and Q*bert were too difficult for any agent we tested to achieve a reasonable level of performance. In Figure 6, we display the performance of several auxiliary tasks on these games, noting that the performance achieved is extremely low in comparison to the agents trained on the standard versions of these games (see Figure 4).

Hyperparameter sweeps. In Figure 7 we vary the weight of the auxiliary loss for the random cumulants agents, with the aim of understanding how this hyperparameters affect each method's performance. Next, in Figures 8 and 9 we present the results of a hyperparameter sweep for Ensemble and REM respectively. For these two, since there is no separate auxiliary loss as in RC, we vary number of heads and the learning rate. Results presented in the main text use the best settings for each algorithm found from these sweeps.

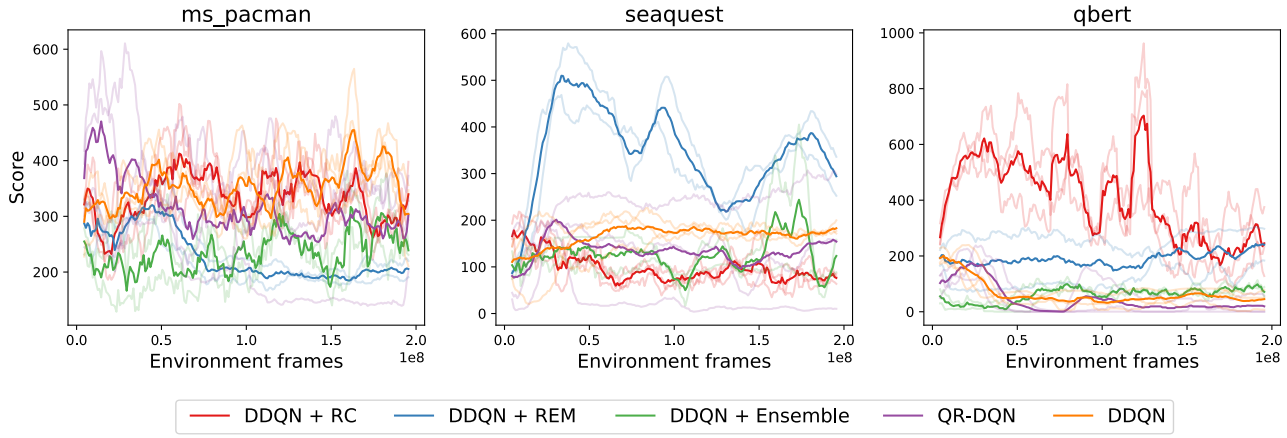

Figure 6: Learning curves on sparsified MsPacman (left), sparsified Seaquest (centre), and sparisifed Q*bert (right).

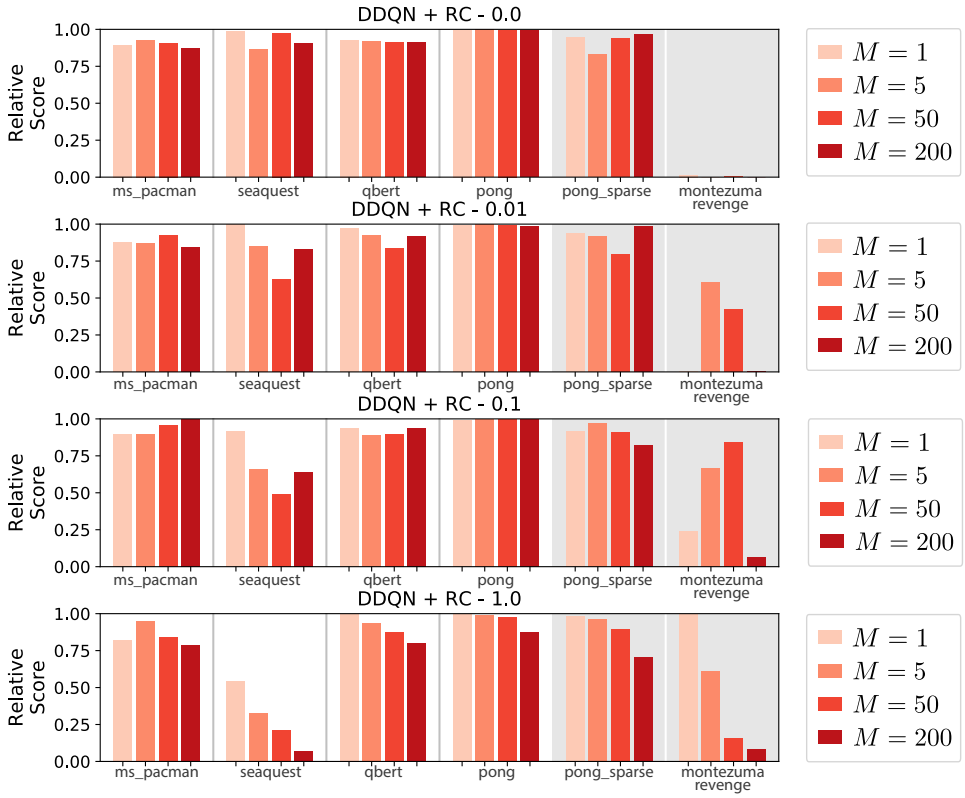

Figure 7: Results of hyper-parameter sweep for Random Cumulant (RC) method, where each row is for a different value of multiplicative scale applied to the auxiliary losses and each bar corresponds to the number of auxiliary heads (M). Note that the first row of results corresponds to initializing a network with the auxiliary heads, but setting the weight to zero, effectively disabling the auxiliary task.

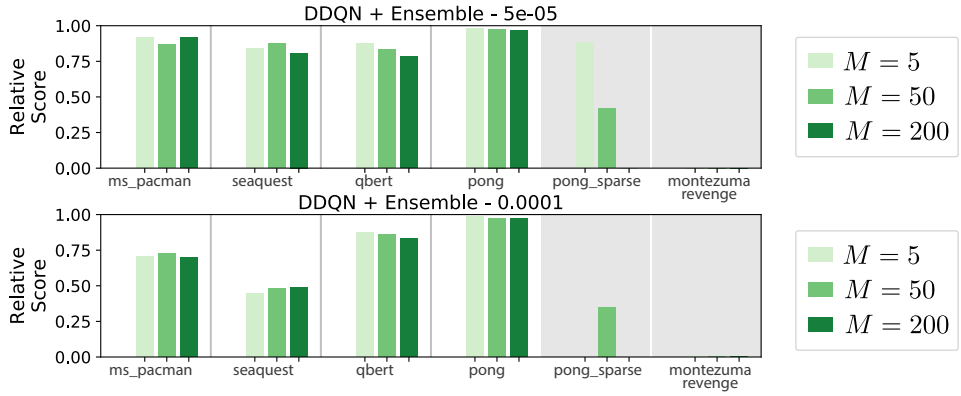

Figure 8: Results of hyper-parameter sweep for the Ensemble method, where each row is for a different learning rate and each bar corresponds to the number of auxiliary heads (M).

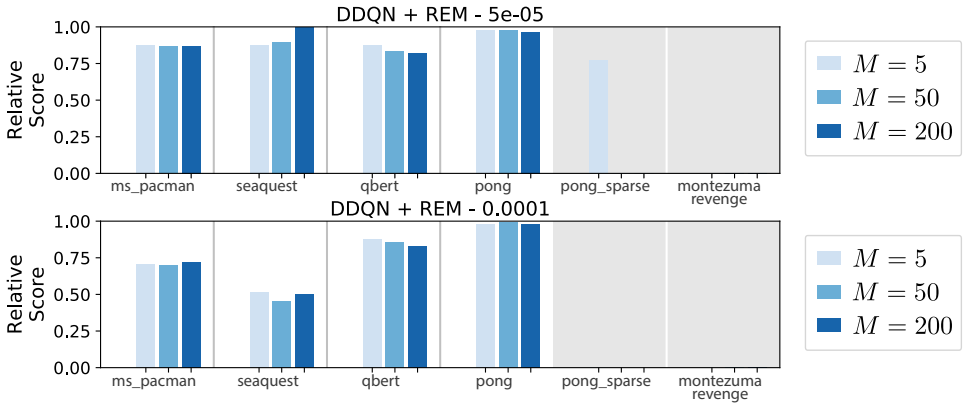

Figure 9: Results of hyper-parameter sweep for the REM method, where each row is for a different learning rate and each bar corresponds to the number of auxiliary heads (M).

E Extensions beyond one-step temporal difference learning

Our analysis in the main paper has focused on the case of learning dynamics under one-step temporal difference learning. This choice is largely because one-step temporal difference learning is such a popular algorithm, not because the results do not hold more generally. In this section, we describe the elements of analogous results for n-step learning and TD(λ) for interested readers. We focus on the case of value function dynamics, and believe extensions of the representation dynamics analysis in the main paper along these lines will be interesting directions for future work.

E.1 Temporal difference learning with n-step returns

In the case of n-step returns, the dynamics on the value function $(V_t)_{t\geq 0}$ are given by

$$\partial_t V_t(x) = \mathbb{E}_\pi \left[\sum_{k=0}^{n-1} \gamma^k R_k + \gamma^n V_t(X_n) \middle| X_0 = x \right] - V_t(x) \,.$$

In full vector notation, we have

$$\partial_t V_t = -(I - \gamma^n (P^\pi)^n) V_t + \left[\sum_{k=0}^{n-1} (\gamma P^\pi)^k \right] R^\pi \,.$$

The solution to this differential equation is

$$V_t = \exp(-t(I - (\gamma P^\pi)^n))(V_0 - V^\pi) + V^\pi \,.$$

This bears a close relationship with the result obtained for 1-step temporal difference learning in the main paper. As expected, we obtain the same limit point. Further, under Assumption 3.3, $(P^\pi)^n$ has the same eigenvectors as P^π, and so results analogous to Propositions 3.4 & 3.5 hold for n-step temporal difference learning too under these conditions.

E.2 Temporal difference learning with λ-returns

In the case of temporal difference learning with λ-returns (for $\lambda \in [0, 1)$), the dynamics on the value function $(V_t)_{t\geq 0}$ are given by

$$\partial_t V_t(x) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} (\lambda\gamma)^k (P^\pi)^k (R^\pi + \gamma P^\pi V_t(X_{k+1}) - V_t(X_k)) \middle| X_0 = x \right] - V_t(x) \,.$$

In full vector notation, we have

$$\partial_t V_t = \sum_{k=0}^{\infty} (\lambda\gamma)^k (P^\pi)^k (R^\pi + \gamma P^\pi V_t - V_t)$$

The solution to this differential equation is

$$V_t = \exp\left(t \left((1-\lambda) \sum_{k=1}^{\infty} \lambda^{k-1} \gamma^k (P^\pi)^k - I \right) \right) (V_0 - V^\pi) + V^\pi \,.$$

As with n-step temporal difference learning, this bears a close relationship with the result obtained for 1-step temporal difference learning in the main paper. As expected, we obtain the same limit point. Further, under Assumption 3.3, each $(P^\pi)^k$ has the same eigenvectors as P^π, and so results analogous to Propositions 3.4 & 3.5 hold for n-step temporal difference learning too under these conditions.

F Beyond diagonalisability assumptions

In this section, we briefly describe extensions of the results of the main paper in scenarios where Assumption 3.3 does not hold. There are two main cases we consider: (i) those in which P^π is still diagonalisable, but does not have all eigenvalues with distinct magnitudes; and (ii) those in which P^π is not diagonalisable.

In the former case, we do not have the different convergence rates of coefficients of different eigenvectors as in the proof of Proposition 3.5. By similar arguments we can still deduce convergence of V_t to the span of the eigenspaces with highest magnitude eigenvalues, but we can no longer deduce convergence to individual eigenspaces if there are several other eigenvalues with the same magnitude as the eigenvalue concerned. Note also that this includes the case where the matrix P^π is complex- but not real-diagonalisable, since in such case non-real eigenvalues must come in conjugate pairs (which are necessarily of the same absolute value).

In the latter case, we no longer have an eigenbasis for $\mathbb{R}^{\mathcal{X}}$ based on P^π. However, we can consider the Jordan normal decomposition, and may still recover analogous results to those in main paper, where convergence is now to the subspaces generated by *Jordan blocks* with high absolute value eigenvalues. See Parr et al. (2008) for further commentary on Jordan normal decompositions in feature analysis.

G Further discussion of features and operator decompositions

Proto-value functions (PVFs), were first defined by Mahadevan and Maggioni (2007) as the eigenvectors of the *incidence matrix* induced by the environment transition matrix P. In the ensuing years, the term PVF has been used to refer to a number of related but not necessarily equivalent concepts. To clarify our use of the term and the relationship of our decompositions of the resolvent and transition matrices of an MDP, we provide a brief discussion here; a summary is provided in Table 2.

We will use A to refer to the adjacency matrix of the unweighted, undirected graph induced by the matrix P (i.e. $A[i,j]$ is 1 if there exists some action with nonzero probability of taking the agent from state i to state j or from state j to state i, and 0 otherwise). L_G will refer to the graph Laplacian based on this matrix A.

We can additionally consider the Laplacian of the weighted, directed graph defined by P^π; we will refer to this matrix as L_{P^π}, in reference to its dependence on the probability of transitioning. T denotes the matrix defined by a collection of sampled transitions indexed by t, with entries $T_{it} = -1$ if the transition t leaves i and $+1$ if it enters state i.

Our first observation is that eigendecomposition and SVD are equivalent for symmetric matrices because any real symmetric matrix has an orthogonal eigenbasis; this means that performing either decomposition yields the same eigenvectors and easily related eigenvalues. Our second observation is that when P^π is *not* symmetric, its singular value decomposition and eigendecomposition may diverge; further, the relationship between the SVD of the resolvent matrix $\Psi = (I - \gamma P^\pi)^{-1}$ and of P^π is no longer straightforward, despite the eigenspaces of the two matrices being analogous. This means that analysis of the singular value decomposition of P^π does not immediately imply any results about the resolvent matrix.

Matrix	SVD	Eigendecomposition (ED)
L_G	PVFs (Mahadevan and Maggioni, 2007)	Equivalent to SVD
T	sometimes $\equiv$ ED(L_G) (Machado et al., 2017)	not discussed
L_{P^π}	$\neq$ ED(L_{P^π})	Stachenfeld et al. (2014)
$(I - \gamma P^\pi)^{-1}$	RSBFs	$\equiv L_{P^\pi}$
P^π	Behzadian and Petrik (2018)	$\equiv L_{P^\pi}$

Table 2: Summary of decompositions of various matrices associated with MDP transition operators, and associated features.

Finally, we note that applying a uniform random walk policy may not be sufficient to guarantee that P^π will be symmetric, and that in general it will not be possible to obtain a policy which will symmetrize the transition matrix. For example: when G is a connected, non-regular graph (as is the case in many environments such as chains), there must be a node v of degree d adjacent to a node v' of degree $d' \neq d$. A random walk policy will assign $p(v, v') = \frac{1}{d}$, while $p(v', v)$ will receive probability $\frac{1}{d'}$; thus, P^π will not be symmetric. Fortunately, this

is not a barrier to spectral analysis; the eigenvectors and eigenvalues of P^π will still be real, as their transition matrix will be *similar* to a symmetric matrix. We defer to Machado et al. (2017) for a more detailed discussion of this relationship.

H Bayes-optimality of RSBFs

We can develop the discussion of RSBFs beyond their properties as a matrix decomposition described in Section G to observe that the RSBFs characterize the Bayes-optimal features for predicting an unknown value function given an isotropic Gaussian prior distribution on the reward, and further characterize a Bayesian posterior over value functions given by conditioning on the known dynamics of the MDP. We will denote by $V_K(\Psi)$ the top K eigenvectors of the matrix $\Psi\Psi^\top$, i.e. the top K left singular vectors of Ψ.

Corollary H.1. Under an isotropic Gaussian prior on reward function $r \in \mathbb{R}^{\mathcal{X}}$, the subspace $V_K(\Psi)$ corresponds to the optimal subspace with respect to the following regression problem.

$$\min_{\Phi \in \mathbb{R}^{\mathcal{X} \times K}} \mathbb{E}_{r \sim \mathcal{N}(0,I)} \left[\| \Pi_{\Phi^\perp} (I - \gamma P^\pi)^{-1} r \|^2 \right] , \tag{49}$$

where $\Pi_{\Phi^\perp}$ denotes orthogonal projection onto the orthogonal complement of Φ.

Proof. Let S denote some subspace $S \subset V$.

$$\mathbb{E}[\|\Pi_s \Psi r\|^2] = \mathbb{E}[r^\top \Psi^\top \Pi_s^\top \Pi_s \Psi r] \tag{50}$$

We note that for any real symmetric matrix A we can rewrite $A = \sum \alpha_i v_i v_i^\top$.

$$\mathbb{E}[r^\top \Psi^\top \Pi_s^\top \Pi_S \Psi r] = \mathbb{E}[r^\top (\sum \alpha_i v_i v_i^\top) r] = \mathbb{E}[\sum \alpha_i (r^\top v_i)(v_i^\top r)] \tag{51}$$

$$= \mathbb{E}[\sum \alpha_i v_i^\top r r^\top v_i] = \sum \alpha_i v_i^\top \mathbb{E}[r r^\top] v_i \tag{52}$$

$$= \sum \alpha_i v_i^\top v_i = \mathrm{Tr}(\Psi^\top \Pi_S^\top \Pi_S \Psi) = \mathrm{Tr}(\Psi^\top \Pi_S \Psi) \tag{53}$$

Finally, we can re-express the minimization problem as follows

$$\mathrm{argmin}_{S:\mathrm{Dim}(S)=k} \mathrm{Tr}(\Psi^\top (\Pi_{S^\perp}) \Psi) = \mathrm{argmax}_{S:\mathrm{Dim}(S)=k} \mathrm{Tr}(\Psi^\top \Pi_S \Psi) \tag{54}$$

Now, because the subspace spanned by the top k left-singular vectors $\{u_1, \ldots, u_k\}$ of Ψ is known to be the maximizer of the above equation, we finally obtain

$$= \langle u_1, \ldots, u_k \rangle = V_K(\Psi) . \tag{55}$$

□

Corollary H.2. The limiting distribution of Φ_t^M under the random cumulant auxiliary task described in Theorem 4.3 is equivalent to the Bayesian posterior over value functions obtained by conditioning on the dynamics P^π, and given a prior distribution on the reward function equal to $\mathcal{N}(0, \Sigma)$.

Proof. Each column of Z_Σ is sampled from an isotropic Gaussian distribution, and therefore each feature $\phi_i \overset{D}{=} (I - \gamma P^\pi)\epsilon_i$. It therefore suffices to show that under a suitable prior distribution, the distribution of ϕ_i is equal to a Bayesian posterior. For this, it suffices to show that such a posterior can be obtained by conditioning on the transition dynamics P^π, and looking at the induced pushforward measure on the reward distribution. Noting that $(I - \gamma P^\pi)$ is invertible, we then obtain the following prior over V^π, assuming an isotropic Gaussian prior on $p_r(r)$ and any arbitrary distribution over potential transition dynamics $p_\pi(P^\pi)$ which covers $\mathbb{R}^{|S| \times d}$.

$$P(V^\pi) = \int_{(r, P^\pi)} \mathbb{1}[(I - \gamma P^\pi)^{-1} r = V^\pi] dp_r(r) dp_\pi(P^\pi) \tag{56}$$

We observe that the random variable V^π has conditional distribution $P(V^\pi|P^\pi) = P((I-\gamma P^\pi)^{-1}r)$, whose density is proportional to $p_r((I-\gamma P^\pi)V)$ by the change of variables formula.

$$P(V^\pi|P^\pi) = cp_r(r = (I-\gamma P^\pi)V^\pi) \tag{57}$$

Because our prior over r is equal to the initialization distribution of ϵ_i, we obtain

$$= cp_{\text{init}}(\epsilon_i = (I-\gamma P^\pi)V^\pi) \tag{58}$$

which is precisely the limiting distribution p_∞ of ϕ_i (again applying the change of variables formula).

$$= p_\infty(\phi_i = (I-\gamma P^\pi)^{-1}\epsilon_i = V^\pi) \tag{59}$$

So we see that the limiting distribution of ϕ_i is equal to the prior over value functions conditioned on the transition dynamics. □

I Learning Dynamics for Ensemble Prediction

We provide some visualizations of the induced behaviour on features as a result of training an ensemble with multiple heads and zero reward, replicating the analysis of Section 2.2, to highlight how the eigendecomposition of P^π affects the learned representations. We run our evaluations on the Four-Rooms Gridworld by initializing $\Phi \in \mathbb{R}^{105\times 10}$ (i.e. $|\mathcal{X}| = 105$ and the number of features $d = 10$) and simulating the ODE defined in Equation 9 for time $t = 100$ with transition matrix P^π defined by the uniform random policy on this Gridworld. In some cases, the features converged to zero quickly and so we show a final $t < 100$ to highlight the behaviour of the representation before it reaches zero.

We consider three variables which we permit to vary: the initialization scheme of features, in one case sampled from an isotropic Gaussian `rand` or from a randomly initialized 2-layer MLP `nn`); whether the weight matrix is fixed at initialization `fix` or permitted to follow the flow defined by Equation 6 `train`; and finally the number of 'heads', `M`=1, 20, and 200.

In Figure 10, we plot the output of an arbitrary head $\mathbf{w}^m$ of the ensemble. In Figure 11 we visualize the value of a single feature (i.e. a single column of Φ). In Figure 12, we track the dot product of the columns of Φ with the eigenfunctions of P^π.

We observe, as predicted, that for fixed heads in the overparameterized regime, the features (and the value functions they induce) converge to smooth eigenfunctions. We do not see meaningful convergence of the features trained in conjunction with a single weight vector. In contrast, the value functions and features trained in conjunction with ensembles with more heads than the feature dimension consistently resemble the eigenfunctions of P^π. When $(\mathbf{w}^m)$ are held fixed, we see convergence to smooth eigenfunctions as predicted by our theory; when $(\mathbf{w}^m)$ are permitted to vary according to the flow in Equation 10, we see convergence to the most eigenfunction corresponding to the most negative eigenfunction of P^π. We can observe the evolution of the dot product between the features and the EBFs of P^π more clearly in Figure 12. Here, each red line corresponds to the dot product between a feature and an EBF. The colour of the line indicates the order i of the eigenvalue λ_i to which the EBF corresponds, interpolating between red λ_1 and blue λ_{105}. Lower values of i correspond to smoother eigenfunctions. We observe that for sufficiently large M, the representations exhibit higher dot product with the smoother eigenfunctions, while for $M = 1$ the features stay largely fixed during training.

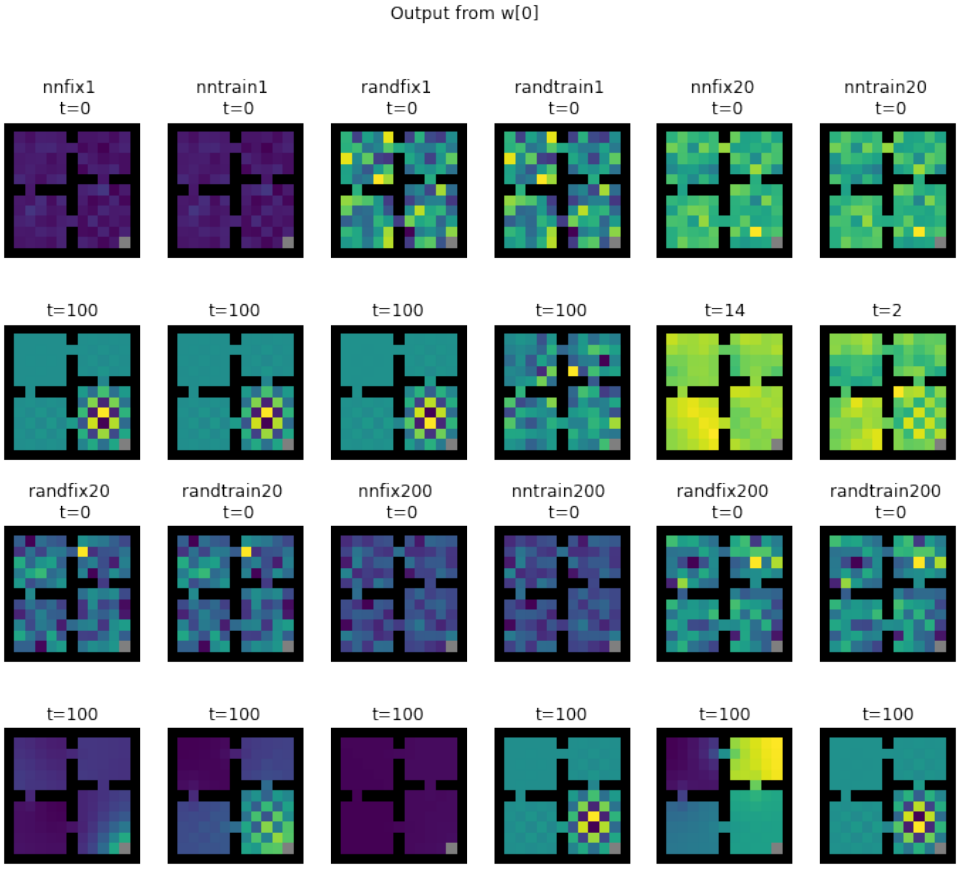

Figure 10: Value functions learned by the ensemble head at index 0 for different training regimes. Plot titles of form (feature initialization scheme, train/fix weight matrix, number of heads in ensemble). Observe that the representation learned with fixed weights tends to converge to smoother eigenfunctions than those learned with weights that are also allowed to train.

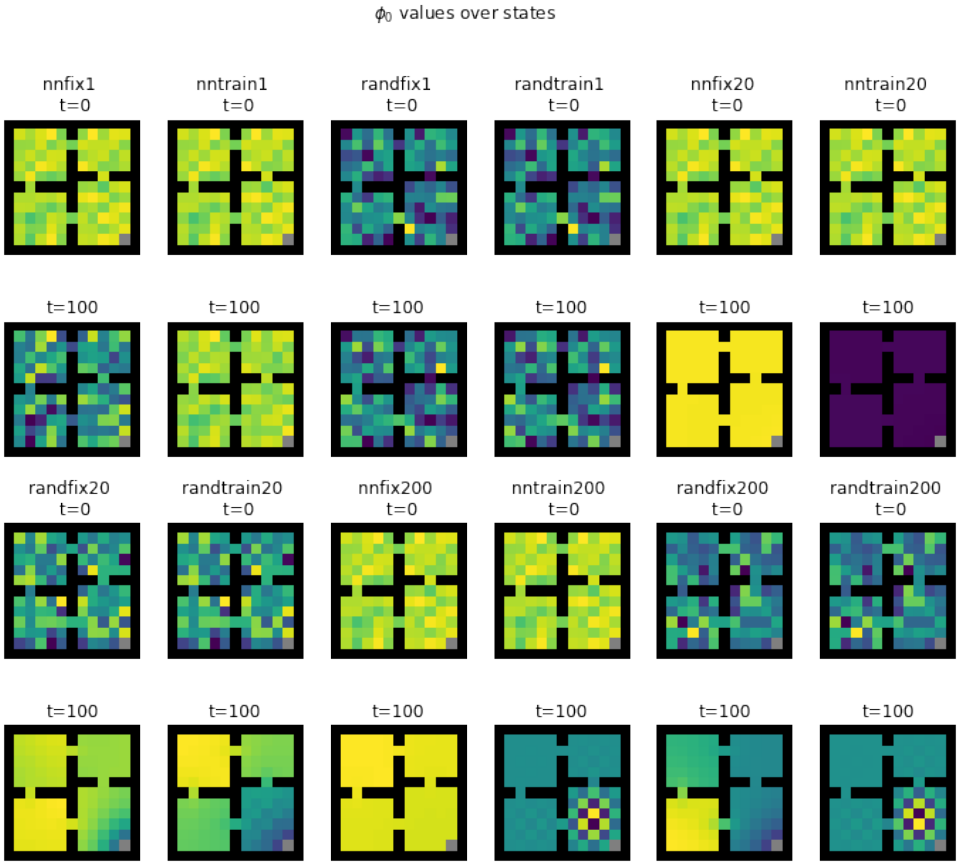

Figure 11: Values of ensemble feature at index 0 for different training regimes. Plot titles of form (feature initialization scheme, train/fix weight matrix, number of heads in ensemble). Observe that the representation learned with fixed weights tends to converge to smoother eigenfunctions than those learned with weights that are also allowed to train.

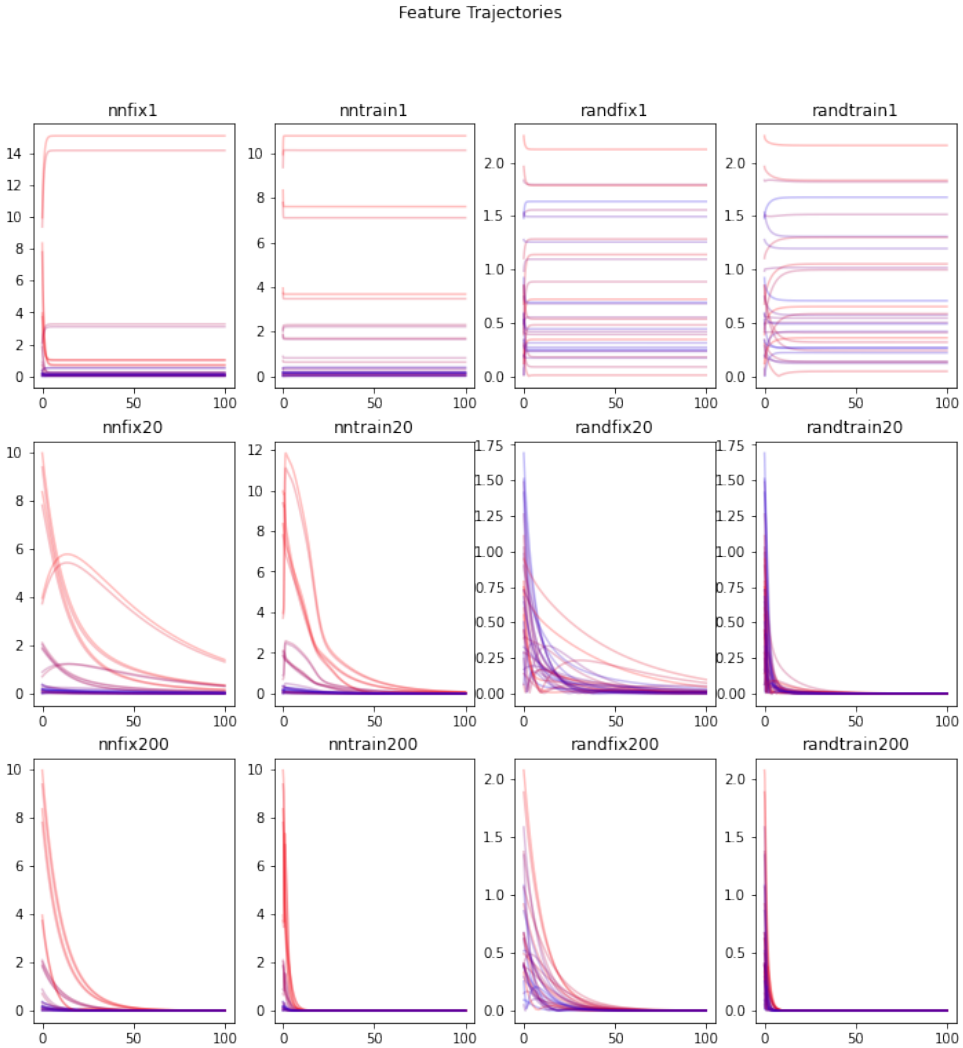

Figure 12: Projection of features onto eigenvectors of P^π. Red lines correspond to projection onto eigenvectors of higher eigenvalues, blue lines to lower eigenvalues.