A. Linear Algebra and Spectral Theory

A.1. Inner Products

A positive-definite symmetric matrix $D \in \mathbb{R}^{k\times k}$ induces an inner product $\langle \cdot, \cdot \rangle_D$ and norm $\| \cdot \|_D$ on $\mathbb{R}^k$. Specifically, the inner product is written as $\langle v, w \rangle_D = v^\top D w$, and the corresponding norm $\|v\|_D^2 = \langle v, v \rangle_D = v^\top D v$. This corresponds to a Hilbert space $(\mathbb{R}^k, \langle \cdot, \cdot \rangle_D)$. In our work, we equip $\mathbb{R}^n$ (where $n = |\mathcal{S} \times \mathcal{A}|$) with the inner-product induced by the data distribution Ξ. We also equip $\mathbb{R}^d$ (the parameter space) with the usual Euclidean inner product.

Most definitions and constructions with the Euclidean inner product generalize to arbitrary Hilbert spaces, some which we describe on $\mathbb{R}^n$. Two vectors $v, w \in \mathbb{R}^n$ are *orthogonal* if $\langle v, w \rangle_\Xi = v^\top \Xi w = 0$. A matrix $A \in \mathbb{R}^{n\times d}$ is *orthogonal* if the columns have unit norm, and are orthogonal to one another: $A^\top \Xi A = I$. The generalization of transposes and symmetrice matrices comes through the adjoint of a matrix $A \in \mathbb{R}^{n\times n}$, written as $A^* = \Xi^{-1} A^\top \Xi$. A matrix is self-adjoint if $A = A^*$, and for matrices that are not self-adjoint, the symmetric component is given as $\bar{A} = \frac{1}{2}(A + A^*)$. We refer to $\|A\|$ as the matrix norm induced by the equivalent norm on vectors.

Matrix decompositions for a matrix $A \in \mathbb{R}^{n\times n}$ can be re-visited with respect to this inner-product.

- **Spectral Decomposition:** If A is self-adjoint, it admits a decomposition $A = U \Lambda U^\top \Xi$, where $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns are eigenvectors of A and Λ a diagonal matrix with the corresponding eigenvalues.
- **SVD:** A admits a decomposition $A = U \Sigma V^\top \Xi$, where $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns are the left singular vectors of A, $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns are the right singular vectors of A, and Λ a diagonal matrix with the corresponding singular values. Letting $U_d, V_d \in \mathbb{R}^{n\times d}$ correspond to the first d singular vectors and $\Sigma_d \in \mathbb{R}^{d\times d}$ the diagonal matrix with the corresponding singular values, then the low-rank approximation $\hat{A} = U_d \Sigma_d V_d^\top \Xi$ minimizes $\|A - \hat{A}\|_\Xi$ amongst all rank d matrices.

A.2. Eigenvalues

We define the eigenvalues of $A \in \mathbb{C}^{k\times k}$ to be the roots of the characteristic polynomial $p(t) = det(A - tI)$. Some eigenvalues may correspond to a multiple root – we refer to this multiplicity as the algebraic multiplicity. Every eigenvalue λ corresponds to an eigenspace $\mathcal{V}_\lambda$ of eigenvectors with this eigenvalue. If the algebraic multiplicity of any eigenvalue λ does not equal the dimensionality of $\mathcal{V}_\lambda$, then A is said to be *defective*. Otherwise, the matrix A is diagonalizable as PDP^{-1}, where P is a basis of eigenvectors of A, and D the corresponding eigenvalues.

We write $\text{Spec}(A) = \{\lambda_1, \dots \lambda_k\} \subset \mathbb{C}$ to denote the set of eigenvalues of the matrix A. The spectral radius of a matrix is the maximum magnitude of eigenvalues, written as $\rho(A) = \sup_{\lambda \in \text{Spec}(A)} |\lambda|$. For two matrices $A \in \mathbb{C}^{k\times m}, B \in \mathbb{C}^{m\times k}$, we have the following cyclicity: $\text{Spec}(AB)\backslash\{0\} = \text{Spec}(BA)\backslash\{0\}$. As a consequence, we also have that $\rho(AB) = \rho(BA)$. We utilize this cyclicity heavily in the ensuing proofs.

The perturbation of eigenvalues for a diagonalizable matrix can be bounded simply via the Bauer-Fike theorem. Specifically, if $A \in \mathbb{C}^{k\times k}$ is diagonalizable as PDP^{-1}, then eigenvalues of the perturbed matrix $\lambda' \in \text{Spec}(A + E)$ can be bounded in distance from the original eigenvalues as $\inf_{\lambda \in \text{Spec}(A)} |\lambda - \lambda'| \leq \|E\| \kappa(P)$, where $\kappa(P) = \|P\| \|P^{-1}\|$. As a simple corollary of the Bauer-Fike Theorem, we have that $\rho(A + E) \leq \rho(A) + \|E\| \kappa(P)$.

B. Proofs

Proposition 3.1. *TD(0) is stable if and only if the eigenvalues of the implied iteration matrix A_Φ have positive real components, that is*

$$Spec(A_\Phi) \subset \mathbb{C}_+ := \{z : Re(z) > 0\}.$$

*We say that a particular choice of representation Φ is **stable** for (P^π, γ, Ξ) when A_Φ satisfies the above condition.*

Proof of Proposition 3.1. We review the update taken by TD(0) (equation 1), rewritten to express the connection to the implied iteration matrix $A_\Phi = \Phi^\top \Xi (I - \gamma P^\pi)\Phi$. Notice that $A_\Phi \theta^*_{TD} = \Phi^\top \Xi r$.

$$\begin{aligned}
\theta_{k+1} - \theta^*_{TD} &= \theta_k - \eta \left(\Phi^\top \Xi (I - \gamma P^\pi)\Phi\theta_k - \Phi^\top \Xi r\right) - \theta^*_{TD} \\
&= \theta_k - \theta^*_{TD} - \eta \left(A_\Phi \theta_k - A_\Phi \theta^*_{TD}\right) \\
&= (I - \eta A_\Phi)(\theta_k - \theta^*_{TD})
\end{aligned}$$

Unrolling the iteration, the error to the optimal solution takes the form

$$\theta_k - \theta^*_{TD} = (I - \eta A_\Phi)^k (\theta_0 - \theta^*_{TD})$$

This above iteration converges from any initialization θ_0 if and only if the spectral radius is bounded by one: $\rho(I - \eta A_\Phi) < 1$.

From here, we can easily show that TD(0) is stable if and only if $\text{Spec}(A_\Phi) \subset \mathbb{C}_+$. If there is some step-size $\eta > 0$ for which $\rho(I - \eta A_\Phi) < 1$, then $\text{Spec}(A_\Phi) \subset \mathbb{C}_+$. Similarly, if $\text{Spec}(A_\Phi) \subset \mathbb{C}_+$, then letting $\eta = \min_{\lambda \in \text{Spec}(A_\Phi)} \frac{\text{Re}(\lambda)}{|\lambda|^2}$ satisfies that $\rho(I - \eta A_\Phi) < 1$.

□

Proposition 3.2. *An orthogonal representation Φ is stable if and only if the real part of the eigenvalues of the induced transition matrix $\Pi P^\pi \Pi$ is bounded above, according to*

$$Spec(\Pi P^\pi \Pi) \subset \{z \in \mathbb{C} : \text{Re}(z) < \tfrac{1}{\gamma}\}$$

In particular, Φ is stable if $\rho(\Pi P^\pi \Pi) < \frac{1}{\gamma}$.

Proof of Proposition 3.2. For an orthogonal representation, the iteration matrix can be written as $A^\Phi_{TD} = I - \gamma \Phi^\top \Xi P^\pi \Phi$. Then,

$$\begin{aligned}
\text{Spec}(A_\Phi) \subset \mathbb{C}_+ &\iff \text{Spec}(\Phi^\top \Xi P^\pi \Phi) \subset \{z \in \mathbb{C} : \text{Re}(z) < \tfrac{1}{\gamma}\} \\
&\iff \text{Spec}(\Pi P^\pi) \subset \{z \in \mathbb{C} : \text{Re}(z) < \tfrac{1}{\gamma}\} \\
&\iff \text{Spec}(\Pi P^\pi \Pi) \subset \{z \in \mathbb{C} : \text{Re}(z) < \tfrac{1}{\gamma}\}
\end{aligned}$$

The second step falls from the cyclicity of the spectrum and the observation that for an orthogonal representation Φ, the projection can be written as $\Phi\Phi^\top \Xi = \Pi$. The spectral radius condition is immediate. □

Proposition 3.3 (SVD)**.** *The representation Φ_{SVD} is stable if and only if the low-rank approximation $\hat{P^\pi}$ satisfies*

$$\rho(\hat{P^\pi}) < \tfrac{1}{\gamma}.$$

Proof of Proposition 3.3. We can write the SVD factorization of the transition matrix as

$$P^\pi = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} \Xi$$

Then, for $\Phi_{SVD} = U_1$, $\Pi P^\pi = U_1 \Sigma_1 V_1^\top \Xi = \hat{P^\pi}$. The necessary and sufficient conditions follow from Proposition 3.2. □

Proposition 3.4 (Successor Representation)**.** *Recall that Spec*$(\Psi) \subset \mathbb{C}_+$*. The representation* Φ_{SR} *is stable if and only if the low-rank approximation* $\hat{\Psi}$ *satisfies*

$$\textit{Spec}(\hat{\Psi}) \subset \mathbb{C}_+ \cup \{0\}.$$

Proof of Proposition 3.4. We can write the SVD factorization of the successor representation $\Psi = (I - \gamma P^\pi)^{-1}$

$$\Psi = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} \Xi \qquad (I - \gamma P^\pi) = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & \Sigma_2^{-1} \end{bmatrix} \begin{bmatrix} U_1^\top \\ U_2^\top \end{bmatrix} \Xi$$

Then, for $\Phi_{SR} = U_1$, the iteration matrix can be written as $A_\Phi = U_1^\top \Xi V_1 \Sigma_1^{-1}$.

Now, writing $\hat{\Psi}$ as $U_1 \Sigma_1 V_1^\top \Xi$ The cyclicity of the spectrum implies the desired criterion.

$$\text{Spec}(\hat{\Psi}) = \text{Spec}(\hat{\Psi}^+) = \text{Spec}(V_1 \Sigma_1^{-1} U_1^\top \Xi) = \text{Spec}(U_1^\top \Xi V_1 \Sigma_1^{-1}) \bigcup \{0\} = \text{Spec}(A_\Phi) \bigcup \{0\}.$$

□

Theorem 4.1. *An orthogonal invariant representation* Φ *satisfies*

$$\textit{Spec}(\Pi P^\pi \Pi) \subseteq \textit{Spec}(P^\pi) \cup \{0\}$$

and is therefore stable.

Proof of Theorem 4.1. Let λ be an nonzero eigenvalue of $\Pi P^\pi \Pi$ with an eigenvector v. Since $\Pi P^\pi \Pi v = \lambda v$, $v \in \text{Span}(\Phi)$.

Since P^π is invariant on Span(Φ), $P^\pi v = \lambda v$, and therefore λ is an eigenvalue of P^π. Therefore, Spec$(\Pi P^\pi \Pi) \subset$ Spec$(P^\pi) \bigcup \{0\}$.

The spectrum of P^π implies the stability of the representation. P^π is a stochastic matrix satisfying $\rho(P^\pi) = 1$, and thus $\rho(\Pi P^\pi \Pi) \leq 1$, implying stability through Proposition 3.2. □

Proposition 4.1 (Golub & van Loan (2013))**.** *Let* $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ *be the ordered eigenvalues of* P^π*. If* $|\lambda_d| > |\lambda_{d+1}|$ *and* $\Phi_0 \in \mathbb{C}^{n \times d}$*, the sequence* $\Phi_1, \Phi_2, \ldots$ *generated via orthogonal iteration is*

$$\Phi_k = \textsc{Orthog}(\textit{Span}(P^\pi \Phi_{k-1}))$$

where $\textsc{Orthog}(\cdot)$ *finds an orthogonal basis. As* $k \to \infty$*,* $\textit{Span}(\Phi_k)$ *converges to the unique top eigenspace of* P^π*.*

Proof of Proposition 4.1. See Theorem 7.3.1 in Golub & van Loan (2013). □

Theorem 4.2. *Let* Φ *be an orthogonal and* ϵ*-invariant representation for* (P^π, γ, Ξ)*. If* P^π *is diagonalizable with eigenbasis A, then* Φ *is stable if*

$$\epsilon < \frac{1 - \gamma}{\gamma} \frac{1}{\kappa_\Xi(A)}.$$

Proof of Theorem 4.2. We can rewrite the definition of ϵ-invariance in terms of a matrix norm: $\|P^\pi \Pi - \Pi P^\pi \Pi\|_\Xi < \epsilon$. Thus, letting $E = \Pi P^\pi \Pi - P^\pi \Pi$, we have $\|E\|_\Xi < \epsilon$.

Now, suppose that $\Pi P^\pi \Pi$ has an eigenvalue, eigenvector pair (λ, v). This means that $v \in \text{Span}(\Phi)$.

$$\lambda v = \Pi P^\pi \Pi v = P^\pi \Pi v + E v = P^\pi v + E v \implies \lambda \in \text{Spec}(P^\pi + E)$$

Now, the Bauer-Fike Theorem (see Appendix A above) thus implies that $\rho(\Pi P^\pi \Pi) < \rho(P^\pi) + \epsilon \kappa_\Xi(A) < 1 + \epsilon \kappa_\Xi(A)$. Now, if $\epsilon < \frac{1-\gamma}{\gamma} \frac{1}{\kappa_\Xi(A)}$, then $\rho(\Pi P^\pi \Pi) < \gamma^{-1}$, and stability follows from Proposition 3.2. □

Proposition 4.2. *A representation spanning $\mathcal{K}_d(P^\pi, r)$ is ϵ-invariant if*

$$\frac{\|\Pi P^\pi v - P^\pi v\|_\Xi}{\|v\|_\Xi} \leq \epsilon$$

Where $v = (I - \Pi_{d-1})(P^\pi)^{d-1} r$, and Π_{d-1} is a projection onto the $(d-1)$-dimensional Krylov subspace $\mathcal{K}_{d-1}(P^\pi, r)$.

Remark: The vector v can be interpreted as the component of the reward at the d-th timestep that cannot be predicted from the first $d-1$ timesteps.

Proof of Proposition 4.2. Any vector $v \in \mathcal{K}_d(P^\pi, r)$ can be decomposed into two components: $\Pi_{d-1} v + (I - \Pi_{d-1})v$.

$$\begin{aligned}\frac{\|\Pi P^\pi v - P^\pi v\|_\Xi}{\|v\|_\Xi} &= \frac{\|\Pi P^\pi \left(\Pi_{d-1} v + (I - \Pi_{d-1})v\right) - P^\pi \left(\Pi_{d-1} v + (I - \Pi_{d-1})v\right)\|_\Xi}{\|\Pi_{d-1} v + (I - \Pi_{d-1})v\|_\Xi} \\ &= \frac{\|\Pi P^\pi (I - \Pi_{d-1}) - P^\pi (I - \Pi_{d-1})v\|_\Xi}{\|\Pi_{d-1} v\|_\Xi + \|(I - \Pi_{d-1})v\|_\Xi}\end{aligned}$$

This expression is maximized whenever v is nonzero and $\|\Pi_{d-1} v\|_\Xi = 0$, which is true whenever $v = (I - \Pi_{d-1})(P^\pi)^{d-1} r$.

$$\sup_{v \in \text{Span}(\Phi)} \frac{\|\Pi P^\pi v - P^\pi v\|_\Xi}{\|v\|_\Xi} = \frac{\|\Pi P^\pi v - P^\pi v\|_\Xi}{\|v\|_\Xi}$$

□

Theorem 4.3. *A positive-definite representation Φ has a positive-definite iteration matrix A_Φ, and is thus stable.*

Proof of Theorem 4.3. First, we show that the iteration matrix A_Φ is positive-definite, and then show that this implies stability.

For any $x \in \mathbb{R}^d$, let $v = \Phi x$. Because Φ is positive-definite, $v \in \mathcal{S}_{PD}$. Notice that rearranging the definition of positive definiteness implies that $\langle v, (I - \gamma P^\pi)v \rangle_\Xi > 0$.

$$x^\top A_{TD}^\Phi x = v^\top \Xi (I - \gamma P^\pi) v = \langle v, (I - \gamma P^\pi)v \rangle_\Xi > 0.$$

Now, we consider an eigenvalue λ of the iteration matrix A_Φ, and a corresponding unit eigenvector $x \in \mathbb{C}^d$. We know that $\overline{\lambda}$ is also an eigenvalue of A_Φ with unit eigenvector $\overline{x}$. Then,

$$(x + \overline{x})^\top A_\Phi (x + \overline{x}) = \lambda x^\top x + \lambda \overline{x}^\top x + \overline{\lambda} x^\top \overline{x} + \overline{\lambda} \overline{x}^\top \overline{x} = 2(\lambda + \overline{\lambda})$$

Positive-definiteness implies that $2(\lambda + \overline{\lambda}) > 0$, and therefore the real component of λ, $\text{Re}(\lambda) = \frac{1}{2}(\lambda + \overline{\lambda})$, must also be positive. □

Proposition 4.3. *Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of K, in decreasing order, and $u_1, \ldots, u_n$ the corresponding eigenvectors. Define d^* as the smallest integer such that $\lambda_{d^*} < \frac{1}{\gamma}$. For any $i \leq n - d^*$, the safe Laplacian representation Φ, defined as*

$$\Phi = [u_{d^*}, u_{d^*+1}, \ldots, u_{d^*+i}],$$

is positive-definite and stable.

Proof of Proposition 4.3. We shall show that $\text{Span}(\{u_{d^*}, u_{d^*+1}, \ldots, u_n\}) \subseteq \mathcal{S}_{PD}$, which implies the proposition.

$$\langle v, P^\pi v \rangle_\Xi = \langle v, \tfrac{1}{2}(P^\pi + \Xi^{-1}(P^\pi)^\top \Xi) v \rangle_\Xi$$

Consider some $v \in \text{Span}(\{u_{d^*}, u_{d^*+1}, \ldots, u_n\})$ which can be expressed as $\sum_{k=d^*}^{n} \alpha_k u_k$. We have

$$\begin{aligned}
\langle v, P^\pi v\rangle_\Xi &= \langle v, \tfrac{1}{2}(P^\pi + \Xi^{-1}(P^\pi)^\top \Xi)v\rangle_\Xi \\
&= \left\langle \sum_{k=d^*}^{n} \alpha_k u_k, \tfrac{1}{2}(P^\pi + \Xi^{-1}(P^\pi)^\top \Xi) \sum_{k=d^*}^{n} \alpha_k u_k \right\rangle_\Xi \\
&= \left\langle \sum_{k=d^*}^{n} \alpha_k u_k, \sum_{k=d^*}^{n} \lambda_k \alpha_k u_k \right\rangle_\Xi \\
&< \gamma^{-1} \left\langle \sum_{k=d^*}^{n} \alpha_k u_k, \sum_{k=d^*}^{n} \alpha_k u_k \right\rangle_\Xi \\
&= \gamma^{-1} \|v\|_\Xi^2
\end{aligned}$$

Hence, $v \in \mathcal{S}_{PD}$ and $\text{Span}(\{u_{d^*}, u_{d^*+1}, \ldots, u_n\}) \subseteq \mathcal{S}_{PD}$. The second-to-last line is a result of eigenvalues being bounded by γ^{-1}.

Since $\text{Span}(\Phi) \subseteq \text{Span}(\{u_{d^*}, u_{d^*+1}, \ldots, u_n\})$, we also have $\text{Span}(\Phi) \subseteq \mathcal{S}_{PD}$, and stability ensues from Theorem 4.3.

As a sidenote, we can use this same sequence of steps to show that a representation using only the top eigenvectors of K is always *not stable*. Defining the representation $\Phi = [u_1, u_2, \ldots, u_{d^*-1}]$, and following the same set of steps yields that $\langle v, P^\pi v\rangle > \gamma^{-1}\|v\|_\Xi^2$ for any $v \in \text{Span}(\Phi)$. This implies that for this representation, the iteration matrix A_Φ is negative-definite, and has *all* eigenvalues with negative real component, therefore not stable. □

C. Empirical Evaluation

C.1. Experimental Setup

Four-room Domain: The four-room domain (Sutton et al., 1999) has 104 discrete states arranged into four "rooms". At any state, the agent can take one of four actions corresponding to cardinal directions; if a wall blocks movement in the selected direction, the agent remains in place.

Policy Evaluation: We augment this domain with a task where the agent must reach the top right corner of the environment. The corresponding reward function is sparse, with the agent receiving +1 reward when it is in the desired state, and zero otherwise. The policy evaluation problem is to find the value function of a near-optimal policy in the environment Epsilon-Greedy$(\pi^*, \epsilon = 0.1)$, which takes the optimal action with probability 0.9, and a randomly selected action otherwise. Data is collected by rolling out 50-step trajectories from the center of the bottom-left room with a uniform policy, which samples actions uniformly at random. The discount factor is $\gamma = 0.99$.

C.2. Exact Evaluation

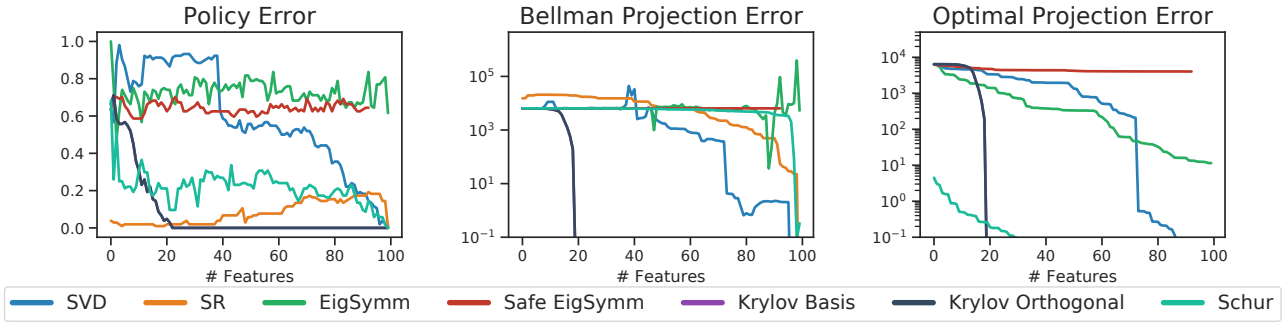
In this setting, the exact transition matrix P^π and data distribution Ξ are used to create the representation. We compute the decompositions according to Table 1 and Appendix A. Stability is measured for a given representation by explicitly creating the induced iteration matrix, computing the eigenvalues, and checking for real positive parts. To measure accuracy, we considered three metrics (Figure C.2).

- **Policy Accuracy: (displayed in paper)** This measures how well the greedy policy for the true value function matches the greedy policy for the estimated value function. This is given as

$$\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \delta(\arg\max_a \hat{Q}(s,a) \neq \arg\max_a Q^\pi(s,a))$$

- **Optimal Projection Error:** This measures how far the true value function is from the subspace of expressible value functions $\|Q^\pi - \Pi Q^\pi\|_\Xi$. As the number of features increases, this error monotonically decreases, but may not be indicative of the quality of the solution.

- **Bellman Projection Error:** This measures how far the solution reached by TD(0) (the TD-fixed point) is from the true value function: $\|Q^\pi - \Phi\theta^*_{TD}\|_\Xi$. This measure of error is nonmonotonic (adding extra features can cause errors to increase) and unbounded. Furthermore, in the regime of a low number of features, this error greatly underestimates the quality of the recovered solution.



C.3. Estimation from Samples

To measure how well the representations can be measured using samples, we consider the difference between the subspace spanned by the estimated and true representations. In particular, we sample t transitions from the data distribution, and reconstruct the empirical transition matrix $\hat{P^\pi}$ given these transitions. If a particular (s, a) pair is never sampled, the prior we use for the transition matrix is that taking this action deterministically leads back to s. We construct the estimated

representation as $\hat{\Phi}$, and measure the distance between the true representation Φ and the estimated representation $\hat{\Phi}$ as $\|\Pi_{\Phi} - \Pi_{\hat{\Phi}}\|_{\Xi,F}$. The Frobenius norm $\|\cdot\|_{\Xi,F}$ is selected in particular as this measures an expected distance, as compared to the maximum distance, measured by the operator norm $\|\cdot\|_{\Xi}$.

C.4. Estimation with Gradient Descent:

When learning the representation using gradient descent, we train a network $f(s, a; \theta)$ with one hidden layer with d units with no activation function, that takes in state-action pairs encoded in one-hot form (as vectors in $\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$) and outputs in $\mathbb{R}^d$. In our experiments, $d = 21$. The value of the units in the hidden layer is the representation $\phi(s, a; \theta)$. The network is trained with a minibatch size of 32 for $100,000$ steps, all implemented in Jax.

- **Schur Decomposition:** To mimic the orthogonal iteration procedure, we use the following training loss function, where θ_t are the parameters for the target network.

$$\mathcal{L}(\theta; \theta_t) = \mathbb{E}_{\substack{(s,a)\sim\xi \\ s'\sim P(\cdot|s,a)}} \left[\|f(s, a; \theta) - \mathbb{E}_{a'\sim\pi}[\phi(s', a'; \theta_t)]\|^2 \right]$$

  This loss is optimized using stochastic gradient descent with a step-size of 4. The target network is updated every $10,000$ steps, and after every target network update, the representation is renormalized to satisfy $\mathbb{E}_{(s,a)\sim\xi}[\phi(s, a; \theta)_i^2] = 1$.

- **Reward Krylov Basis:** We use the following regression training loss function

$$\mathcal{L}(\theta) = \mathbb{E}_{(s_1,a_1)\sim\xi} \left[\sum_{i=1}^{d} \left(f(s, a; \theta)_i - \mathbb{E}_{(s_2,a_2,s_3,a_3,\ldots,s_d,a_d)\sim P^{\pi}}[r(s_i, a_i)] \right)^2 \right]$$

  where the inner expectation comes from trajectories that are generated from the policy π being evaluated starting from (s_1, a_1). Although this loss requires that the evaluated policy be run in the environment, it serves a didactic purpose to show that these Krylov bases can be learned with additional domain knowledge. This loss is optimized using the Adam optimizer with a learning rate of 10^{-3}.