Spectral Alignment in Forward–Backward Representations via Temporal Abstraction

Seyed Mahdi B. Azad, Jasper Hoffmann, Iman Nematollahi, Hao Zhu, Abhinav Valada, Joschka Boedecker

Keywords: Successor Representation, Forward-Backward Learning, Reinforcement Learning

Summary

Forward-backward (FB) representations often suffer from a "spectral mismatch" where low-rank neural bottlenecks struggle to approximate high-rank continuous dynamics, making accurate representation learning difficult. We analyze temporal abstraction as a principled mitigation strategy. Theoretically, we characterize temporal abstraction as a low-pass filter that reduces the effective rank of the successor representation (SR) while maintaining a formal bound on the value function error. Empirically, we demonstrate that spectral alignment is a key factor for effective learning at high discount rates, enabling FB to capture long-horizon structure without representational collapse.

Contribution(s)

1. **Exposing the theory–practice gap in SR learning.** We demonstrate that enforcing a low rank via narrow embeddings or high discount factors fails to yield effective successor representations in continuous control. This reveals a critical mismatch between theoretical low-rank assumptions and practical FB learning with bootstrapping and function approximation.
   Context: Prior works (Blier et al., 2021; Dubail et al., 2025) explore the SR's spectral properties. They also provide theoretical discussions on different mechanisms of arriving at a low-rank approximation of it. We bridge the gap to practice by identifying why these theoretical levers often fail in deep reinforcement learning (RL).
2. **Temporal abstraction as a spectral low-pass filter.** We provide a theoretical study of action-repetition as a spectral filter, showing that it can improve the low-rank approximation of SR via FB in discrete settings.
   Context: Unlike traditional uses of action-repetition for exploration or hierarchical RL (Mnih et al., 2015; Sutton et al., 1999), we recast it as a representational tool that suppresses spectral complexity to facilitate low-rank learning.
3. **Empirical benefit of temporal abstraction for FB.** Empirically, we show that temporal abstraction can be a key factor for effective FB learning in both discrete and continuous settings and across a broad range of bottlenecks and discount factors.
   Context: While deep FB representations can be sensitive to embedding dimension or discount factors, we demonstrate that temporal abstraction acts as a robust regularizer. It enables high-capacity networks to leverage large embedding dimensions without succumbing to the propagation of high-frequency approximation errors common in bootstrapping.

Spectral Alignment in Forward–Backward Representations via Temporal Abstraction

Seyed Mahdi B. Azad[1], **Jasper Hoffmann**[1], **Iman Nematollahi**[1], **Hao Zhu**[1], **Abhinav Valada**[1], **Joschka Boedecker**[1]

basiri,hoffmaja,nematoli,zhuh,valada,jboedeck@cs.uni-freiburg.de

[1]**Department of Computer Science, University of Freiburg, Germany**

Abstract

Forward-backward (FB) representations provide a powerful framework for learning the successor representation (SR) in continuous spaces by enforcing a low-rank factorization. However, a fundamental spectral mismatch often exists between the high-rank transition dynamics of continuous environments and the low-rank bottleneck of the FB architecture, making accurate low-rank representation learning difficult. In this work, we analyze temporal abstraction as a mechanism to mitigate this mismatch. By characterizing the spectral properties of the transition operator, we show that temporal abstraction acts as a low-pass filter that suppresses high-frequency spectral components. This suppression reduces the effective rank of the induced SR while preserving a formal bound on the resulting value function error. Empirically, we show that this alignment is a key factor for stable FB learning, particularly at high discount factors where bootstrapping becomes error-prone. Our results identify temporal abstraction as a principled mechanism for shaping the spectral structure of the underlying MDP and enabling effective long-horizon representations in continuous control.

1 Introduction

Understanding and controlling long-horizon behavior in complex environments requires representations that capture how present actions influence future outcomes. The successor representation (SR) provides such a predictive structure by encoding discounted future state–action occupancies under a policy (Dayan, 1993). By aggregating multi-step dynamics into a single linear operator, the SR reveals the global structure of the environment and provides a principled foundation for value computation across diverse reward functions. Beyond its original tabular formulation, SR-based methods have been extended to deep function approximation and applied to pixel-based control and robotic navigation (Kulkarni et al., 2016; Zhang et al., 2017), demonstrating their relevance in high-dimensional settings. This practical success underscores the importance of learning compact, stable approximations of the SR in continuous domains.

Scaling to continuous state–action spaces requires representations that are both expressive and computationally tractable. Classical formulations scale poorly with state-space size, motivating structured operator approximations. Forward-backward (FB) representations address this challenge by learning a low-rank factorization of the SR directly from interaction (Blier et al., 2021; Touati & Ollivier, 2021). By constraining the spectrum of the learned operator, FB aims to capture dominant long-horizon structure while discarding short-horizon dynamics. Yet a fundamental incompatibility exists: although FB enforces a low-rank constraint for tractability, the true SR in continuous environments often exhibits slow spectral decay. In such settings, high-frequency modes of the transition dynamics contribute significantly to the operator's complexity (Dubail et al., 2025).

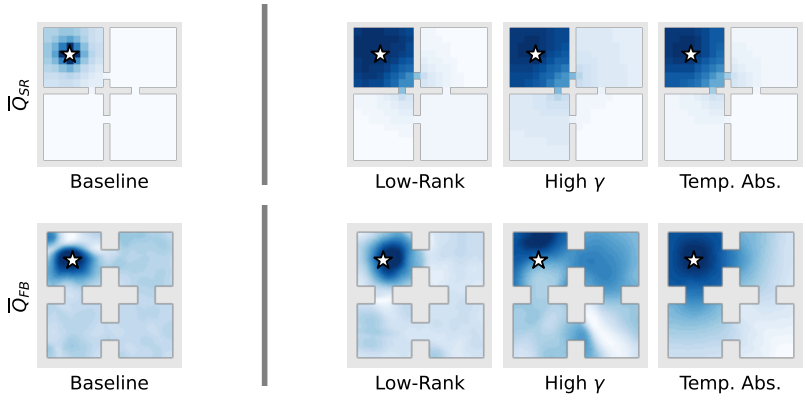


Figure 1: **Q-function via Successor Representation (SR).** The SR enables rapid value inference for arbitrary goals (*star marker*). Low-rank structure is desirable for navigation, as it preserves topological features (e.g., rooms) while suppressing transient dynamics. **Top:** In discrete MDPs, the SR can be computed from the transition matrix. **Bottom:** In continuous domains, forward-backward (FB) learning approximates the SR, where the embedding dimension controls the rank of the approximation. Low-rank structure can arise through (1) explicit constraints (e.g., SVD or small embeddings), (2) long horizons (high γ), or (3) temporal abstraction (e.g., action repetition). We show that, in continuous settings, temporal abstraction provides the spectral alignment needed for effective bootstrapping, whereas high γ or overly restrictive bottlenecks can impair representation learning.

In this work, we identify this spectral mismatch as a key source of difficulty in learning accurate low-rank SR representations with FB. We demonstrate empirically that increasing the capacity of FB networks by expanding the embedding dimension does not reliably improve performance. Instead, we observe a performance degradation beyond a certain capacity threshold. We argue that this is an artifact of high-capacity networks attempting to resolve high-frequency modes of the dynamics that are inherently difficult to predict. When coupled with bootstrapping, errors in these modes propagate through the Bellman updates and hinder accurate low-rank representation learning.

To address this mismatch, we study temporal abstraction via action repetition as a mechanism for regulating the spectral structure of the SR. In finite state–action spaces, we show that multi-step transition operators accelerate spectral decay in the induced SR, improving its low-rank approximation under the FB factorization. As shown in Figure 1, our empirical results support this analysis. Across both discrete and continuous environments, incorporating action repetition, which has previously been used to improve exploration and learning efficiency (Mnih et al., 2015; Biedenkapp et al., 2021), consistently improves the quality of the FB representation and the episodic return. These gains align with the predicted attenuation of high-frequency spectral modes, yielding a more structured and learnable target for the FB objective.

Furthermore, we examine the role of the discount factor, γ. While larger γ increases the effective task horizon, it also degrades the conditioning of the successor representation, amplifying subdominant spectral modes and increasing sensitivity to approximation noise. We demonstrate that temporal abstraction counteracts this effect by improving spectral concentration, thereby facilitating representation learning at high effective discount factors and allowing long-term dependencies to be captured without the interference from fine-grained dynamical variations.

Taken together, our results provide a unified perspective on the spectral requirements of low-rank SR learning with FB representations. Our proposed environment-level intervention shifts the burden of representation learning from the function approximator toward the design of interaction dynamics.

2 Related Works

Successor representations were originally introduced as a task-agnostic predictive representation enabling rapid adaptation to new reward functions (Dayan, 1993). More recent work has leveraged SR for transfer and zero-shot reinforcement learning (Barreto et al., 2017). However, computing the exact SR scales poorly with state-space dimensionality, motivating low-rank and parametric approximations that capture dominant long-horizon dynamics.

Forward-backward representation learning methods (Blier et al., 2021; Touati & Ollivier, 2021; Touati et al., 2022) address this challenge by learning factorizations of the SR that emphasize shared future state occupancies rather than fine-grained state distinctions. These methods implicitly assume that the SR admits a low-rank structure, but provide limited theoretical guidance on when such a structure should emerge. Our work complements FB learning by analyzing how the spectral properties of the transition dynamics, induced by the policy and environment, govern the effective rank of the SR and by proposing practical mechanisms that promote such structure.

The transition operator of a Markov decision process has long been recognized as a fundamental object for understanding long-term behavior, mixing properties, and value estimation. Classical results in Markov chain theory relate the spectral gap of the transition matrix to convergence rates and mixing behavior (Meyn & Tweedie, 2012). In reinforcement learning, spectral perspectives have been used for representation learning and planning, including proto-value functions and Laplacian-based state abstractions (Mahadevan, 2005; Machado et al., 2017a;b; Shehmar et al., 2026). Prior analyses primarily focus on policy evaluation and transfer, but do not examine how properties of the transition operator influence the rank, compressibility, or learnability of low-rank SR under function approximation.

Temporal abstraction has been extensively studied through the framework of semi-Markov decision processes and options (Sutton et al., 1999). A simple and widely used instance of temporal abstraction is action repetition (or frame skipping), in which actions are repeated for multiple environment steps. This technique has been employed in Atari benchmarks (Mnih et al., 2015) and shown to substantially affect learning performance (Machado et al., 2018; Biedenkapp et al., 2021). From an operator perspective, action repetition replaces the one-step transition matrix P with its k-step counterpart $P^{(k)}$, effectively smoothing the transition dynamics. Existing work typically motivates this practice in terms of computational efficiency or exploration, but does not analyze its implications for the spectral structure or low-rank approximations of predictive representations, such as the SR under function approximation.

While the relationship between multi-step transitions and the spectral properties of the SR (Dayan, 1993) is mathematically established (Machado et al., 2017b;a; Dubail et al., 2025), and the effectiveness of FB in reinforcement learning has been demonstrated (Touati & Ollivier, 2021; Touati et al., 2022), the interplay between these two remains unexplored. Specifically, we move beyond viewing temporal abstraction methods, such as action repetition, as an exploration heuristic (Lakshminarayanan et al., 2017). Instead, we characterize temporal abstraction as a spectral alignment tool that bridges the gap between high-rank continuous dynamics and the low-rank inductive bias of FB representations.

3 Background

We represent a finite, reward-free Markov decision process (MDP) as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, respectively, $P(s' \mid s, a)$ is the transition probability from state s to s' given action a, and $\gamma \in (0, 1)$ is the discount factor (Sutton & Barto, 1998). Given a policy π, the policy-induced transition operator is defined as a matrix $P^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$, where $P^\pi(s', a' \mid s, a) = \mathbb{P}(s_{t+1} = s', a_{t+1} = a' \mid s_t = s, a_t = a, \pi)$. The matrix P^π is row-stochastic, i.e., $P^\pi \mathbf{1} = \mathbf{1}$. The (discounted) SR associated with P^π is defined as $M^\pi = (I - \gamma P^\pi)^{-1} = \sum_{t=0}^{\infty} \gamma^t (P^\pi)^t$. In the following, we use the matrix M^π and its functional form interchangeably. We define $M^\pi(s, a, s', a')$ as the expected discounted occupancy of (s', a')

given an initial state-action pair (s, a). In matrix notation, it corresponds to the entry of M^π indexed by row (s, a) and column (s', a').

3.1 Forward-Backward Representation

The FB representation is a parametric framework designed to approximate the SR for all optimal policies in an unsupervised way (Touati & Ollivier, 2021). Let $(\pi_z)_{z\in\mathbb{R}^d}$ be a family of policies parameterized by $z \in \mathbb{R}^d$, and define the embedding functions $F\colon \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}^d$ and $B\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Learning an FB representation entails finding (F, B, π_z) such that:

$$\pi_z(s) \in \operatorname*{argmax}_a F(s, a, z)^\top z \quad \text{and} \quad F(s, a, z)^\top B(s', a') = M^{\pi_z}(s, a, s', a') \tag{1}$$

for all $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$ and $z \in \mathbb{R}^d$. Further, Eq. (1) represents a fixed-point condition for the triplet (F, B, π_z). Given a reward function $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define $z_R = B^\top r$. If the condition holds exactly, the optimal action-value function is recovered by $Q^\star(s, a) = F(s, a, z_R)^\top z_R$.

Given a FB representation (F, B), we define the approximate successor representation as $\hat{M}^z(s, a, s', a') = F(s, a, z)^\top B(s', a')$. The following theorem bounds the approximation error of the optimal action-value function $Q^\star$ by the approximation error in successor representation M^{π_z}:

Theorem 3.1 (Optimality Gap for FB Representations)**.** *Let r be a reward function such that $z_R = B(s, a)r(s, a)$. The approximation error of the optimal Q-function is bounded by:*

$$\left\|F(\cdot, \cdot, z_R)^\top z_R - Q^\star\right\|_\infty \leq \frac{2\|r\|_\infty}{(1-\gamma)} \|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2. \tag{2}$$

Here, $\|\cdot\|_\infty$ denotes the L_∞ or Chebyshev norm, which for a function $f\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $\|f\|_\infty = \sup_{(s,a)} |f(s, a)|$. The norm $\|\cdot\|_2$ denotes the L_2 or spectral norm of a matrix, defined as $\|M\|_2 = \sup_{x\neq 0} \|Mx\|_2/\|x\|_2$, which corresponds to the largest singular value of M. Note that Theorem 3.1 is a simplified version of the result in (Touati & Ollivier, 2021, Theorem 8) tailored to our spectral analysis setting.

3.2 Spectral Bound on Approximation Error

To understand the approximation capacity of the FB framework, we derive a lower bound on the approximation error appearing on the right-hand side of Eq. (2) based on the spectrum of $M^{\pi_{z_R}}$. Related to this is the work in Dubail et al. (2025), which performs a similar study with a focus on finite-sample analysis. In his work, we do not aim to derive the tightest possible bound, but rather to develop a simple theoretical framework that highlights the effect of temporal abstractions on the optimal approximation error. We leave a finite-sample analysis to future work.

Due to limited representational capacity when d is small, the FB criterion cannot generally be fulfilled exactly, even in the finite case. Furthermore, the FB representation must simultaneously reconstruct the successor representation and define a greedy policy, as shown in Eq. (1). By the Eckart–Young–Mirsky theorem (Eckart & Young, 1936), the best rank-d approximation of $M^{\pi_{z_R}}$ is obtained via the truncated singular value decomposition (SVD), denoted by $M^\star$, which satisfies $\|M^\star - M^{\pi_{z_R}}\|_2 = \sigma_{d+1}(M^{\pi_{z_R}})$. Intuitively, $\sigma_{d+1}(M^{\pi_{z_R}})$ corresponds to the first discarded singular value. Motivated by this observation, we define the following:

Definition 3.1 (Forward-backward Realization Error)**.** *Given a reward function $r\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, and a FB representation (F, B), we define the* FB realization error *as the difference to the optimal rank d approximation: $\epsilon_{\text{real}}(r) \coloneqq \|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2 - \sigma_{d+1}(M^{\pi_{z_R}}) \geq 0$.*

Consequently, the optimality gap in Eq. (2) is governed by the decay of the representation's singular values,

$$\left\|F(\cdot, \cdot, z_R)^\top z_R - Q^\star\right\|_\infty \leq \frac{2\|r\|_\infty}{(1-\gamma)} \left(\epsilon_{\text{real}}(r) + \sigma_{d+1}(M^{\pi_{z_R}})\right)$$

assuming that the error $\epsilon_{\text{real}}(r)$ stays bounded. This decomposition separates the FB realization error $\epsilon_{\text{real}}(r)$ from the spectral truncation error $\sigma_{d+1}(M^{\pi_{z_R}})$ determined by the singular values of the successor representation.

4 Temporal Abstraction in Forward-Backward Representations

Our goal is to demonstrate that temporal abstraction is beneficial for learning FB representations. To this end, we introduce a simple temporal abstraction, namely action repetition. Action repetition was introduced in Mnih et al. (2015) and has been shown to be beneficial for exploration and learning performance in model-free RL (Biedenkapp et al., 2021).

4.1 Action Repetition for Temporal Abstraction

In the following, we first formally introduce the concept of action-repeat MDPs, provide the necessary assumptions for this work, and conclude by connecting these concepts to the FB representation.

Definition 4.1 (Action-Repeat MDP)**.** *Given a reward-free MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma)$, an action-repeat MDP $\widetilde{\mathcal{M}}$ with repeat factor $k \in \mathbb{N}$ is defined by the tuple $(\mathcal{S}, \mathcal{A}, \widetilde{P}, \gamma^k)$. The transition probability $\widetilde{P}(s'|s,a)$ represents the probability of reaching state s' after executing action a for k consecutive time steps in $\mathcal{M}$. Mathematically, this is the k-fold composition of the transition operator:*

$$\widetilde{P}(s'|s,a) = \sum_{(s_1,\dots,s_{k-1}) \in S^{k-1}} P(s'|s_{k-1},a) \cdots P(s_1|s,a).$$

Note that for $k = 1$ we define $\widetilde{P}(s'|s,a) = P(s'|s,a)$.

Given this definition, and following Sec. 3, we define the SR $\widetilde{M}^\pi$ and the optimal state-action function $\widetilde{Q}^\star$ accordingly. To measure the error that is introduced by the action repetition, we introduce the following definition:

Definition 4.2 (Action-Repeat Value Error)**.** *For a given repeat factor k, we define the* action-repeat value error *as the worst-case discrepancy between the optimal Q-value function of the original MDP $\mathcal{M}$ and that of the action-repeat MDP $\widetilde{\mathcal{M}}$ as $\epsilon_{\text{repeat}}(k) = \|Q^\star - \widetilde{Q}^\star\|_\infty$.*

For the remainder of this paper, we assume the existence of a repeat factor k such that the resulting action-repeat value error $\epsilon_{\text{repeat}}(k)$ is negligibly small. Furthermore, all representations (F, B) and successor measures $\hat{M}$ are hereafter implicitly assumed to be trained on the action-repeat MDP $\widetilde{\mathcal{M}}$.

4.2 Action Repetition Reduces the Optimality Gap

We first derive the corresponding SR of $\widetilde{\mathcal{M}}$. For a fixed action a_q, let $P_{a_q} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the state-transition dynamics under that action, and let $\pi_{s_p} \in \mathbb{R}^{1\times|\mathcal{A}|}$ represent the row vector of policy probabilities for a given state s_p, i.e.,

$$P_{a_q} = \begin{bmatrix} P(s_1 \mid s_1, a_q) & \dots & P(s_{|\mathcal{S}|} \mid s_1, a_q) \\ \vdots & \ddots & \vdots \\ P(s_1 \mid s_{|\mathcal{S}|}, a_q) & \dots & P(s_{|\mathcal{S}|} \mid s_{|\mathcal{S}|}, a_q) \end{bmatrix}, \quad \pi_{s_p} = \begin{bmatrix} \pi(a_1 \mid s_p) & \dots & \pi(a_{|\mathcal{A}|} \mid s_p) \end{bmatrix}.$$

With these components established, we can express the transition matrix $\widetilde{P}^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ as a product of action-repetition and policy-mapping components:

$$\widetilde{P}^\pi = P_{\text{rep}}^k \, \tilde{\pi}, \tag{3}$$

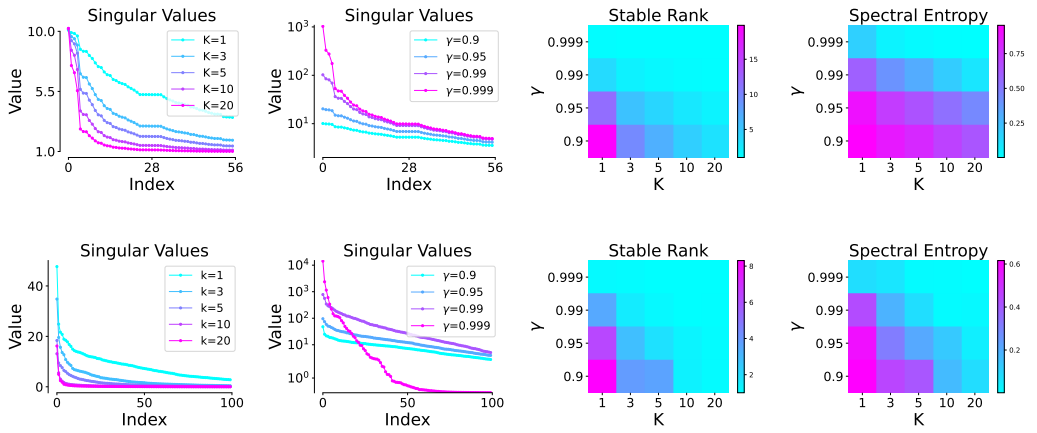


Figure 2: **Effect of temporal abstraction and discount factor on effective rank.** Effective rank decreases by increasing k or γ in both discrete *(top)* and continuous *(bottom)*. Entropy decreases more smoothly as k increases, suggesting a more stable reduction in effective rank as compared to increasing γ.

where

$$P^k_{\text{rep}} = K \begin{bmatrix} P^k_{a_1} \\ \vdots \\ P^k_{a_{|\mathcal{A}|}} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}|} \quad \text{and} \quad \tilde{\pi} = \begin{bmatrix} \pi_{s_1} & & 0 \\ & \ddots & \\ 0 & & \pi_{s_{|\mathcal{S}|}} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}\times\mathcal{A}|}.$$

The matrix $K \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times|\mathcal{S}\times\mathcal{A}|}$ is a commutation matrix that reorders the state-action product space from a state-major to an action-major indexing scheme, enabling a concise block-column representation. When $k = 1$, we write P_{rep} for P^1_{rep}. In the decomposition (3), P^k_{rep} captures the k-step transitions under action repetition, while $\widetilde{\pi}$ maps the policy within the state-action space. Intuitively, the system first evolves for k steps under the same action, after which the next action is selected according to the policy without execution.

With the matrix decomposition established, we can now derive a spectral bound on the optimality gap. By repeating actions, we introduce a bias $\epsilon_{\text{repeat}}(k)$ relative to the original optimal policy, while simultaneously accelerating the spectral decay of the successor representation. This contraction reduces the reconstruction error for a fixed embedding dimension d, potentially leading to a tighter overall bound on the optimality gap.

Lemma 4.1 (Optimality Gap for k-repeat FB Representations)**.** *Given a repeat factor k, a reward function $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and an FB (F, B) representation with dimension d, the error in approximating the original optimal action-value function $Q^\star$ is bounded by:*

$$\left\| F(\cdot,\cdot,z_R)^\top z_R - Q^\star \right\|_\infty \leq \epsilon_{\text{repeat}}(k) + \frac{2\|r\|_\infty}{1-\gamma}\left(\epsilon_{\text{real}}(r) + \frac{1}{1-\gamma(\sigma_{d+1}(P_{\text{rep}}))^k}\right).$$

This Lemma bounds the approximation error of learning (F, B) in the action-repeat MDP $\widetilde{M}^\pi$ specifically in terms of its deviation from the original $Q^\star$, capturing the trade-off between repetition bias and spectral compression. A proof can be found in Sec. B. The utility of this bound is most evident when P_{rep} exhibits a spectral gap such that $\sigma_{d+1} < 1$, in which case the spectral error term decays exponentially with the repetition factor k.

5 Temporal Abstraction in Practice

In this section, we empirically validate the spectral insights developed in the previous sections by analyzing how temporal abstraction shapes the structure and learnability of FB representations. We first introduce spectral metrics that characterize the effective rank of the SR. We then describe the experimental setup and environments used for evaluation. Building on this, we analyze how temporal abstraction reshapes the singular value spectrum of the SR and influences performance. Finally, we study the interaction between temporal abstraction, embedding dimension, and discount factor, highlighting how these components jointly determine the stability and effectiveness of FB learning.

5.1 Spectral Metrics for Representation Complexity

To quantify the structural properties of the SR and the impact of temporal abstraction, we employ two spectral metrics that characterize its effective rank.

Stable Rank. Stable rank measures how many singular directions carry substantial energy relative to the dominant mode, which is defined for some matrix M as

$$\mathrm{SRank}(M) = \frac{\|M\|_F^2}{\|M\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2}.$$

It is particularly sensitive to the presence of strong leading components: when most of the spectral energy concentrates in a few dominant directions, the stable rank becomes small. As such, it provides a direct proxy for how well a representation can be captured by a low-rank bottleneck.

Normalized Spectral Entropy. Normalized spectral entropy quantifies the uniformity of spectral energy across all modes, which is defined for some matrix M as

$$\mathrm{NSE}(M) = \frac{-\sum_i p_i \log(p_i)}{\log(\beta)}, \quad p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2},$$

where σ_i are the singular values of the matrix M and $\beta \in \mathbb{N}$ is a normalization factor representing the number of singular values of M. By normalizing the entropy value with the logarithm of β, it is guaranteed to lie between $[0, 1]$. Values near one indicate a diffused spectrum where many modes contribute comparably, while lower values reflect concentration of energy into a restricted subset of directions. Compared with stable rank, which emphasizes the dominant singular values, spectral entropy is sensitive to the degree to which energy is distributed across the spectrum.

Together, these metrics provide a complementary view of effective rank. They allow us to distinguish between near rank-one collapse (stable rank ≈ 1 and low entropy) and structured spectral concentration, where stable rank is low but entropy remains sufficiently high, indicating that a few dominant modes capture most of the energy while multiple dynamical components are preserved. Furthermore, Sec C and D show that upperbounds for both metrics decrease as temporal abstraction increases.

In discrete environments, where the exact successor representation can be computed, we evaluate these metrics directly on the SR matrix M^π. In continuous environments, where the exact SR is unavailable, we instead compute the metrics on the empirical SR approximation $\hat{M}^\pi = FB^\top$ estimated from batches of sampled state–action transitions collected during training.

5.2 Experimental Setup

To evaluate the benefit of temporal abstraction via action repetition in continuous state-action settings, we consider three maze navigation environments of increasing difficulty shown in Figure 3: *Four-Rooms*, *Maze*, and *Large-Maze*. The environments are implemented using the OGBench benchmark (Park et al., 2025), with random start and goal positions.

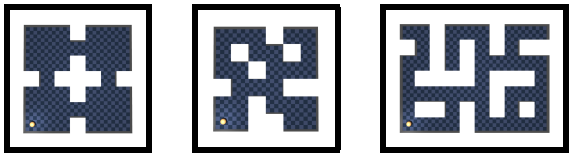
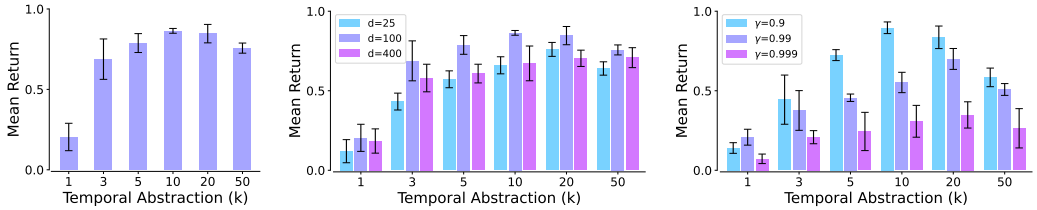


Figure 3: **Continuous Navigation Environments:** *Four-Rooms*, *Maze*, and *Large-Maze*



(a) (b) (c)

Figure 4: **Effect of temporal abstraction on performance.** Ablation over temporal abstraction (k), embedding dimension (d), and discount factor (γ) using a continuous four-rooms environment. Addition of temporal abstraction ($k > 1$) boosts performance, whereas increasing d or γ alone does not.

Unless stated otherwise, ablation experiments are conducted in the Four-Rooms environment with discount factor ($\gamma = 0.95$), forward-backward embedding dimension (d=100), and temporal abstraction via action repetition (k=10). Following Touati & Ollivier (2021), we use state-based inputs where (x, y) coordinates are encoded using an RBF kernel. Similar performance can also be achieved using a learned CNN encoder, as shown in Figure 6a. All results report the mean episodic return $\pm$ standard deviation over 5 random seeds. Table 1 contains the most relevant hyperparameters used in the experiments.

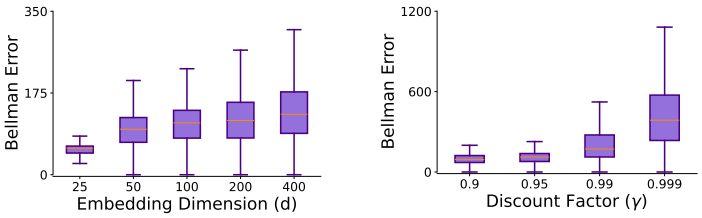
5.3 Temporal Abstraction and Effective Rank

Building on our theoretical results, Figure 2 illustrates how increasing the temporal abstraction step, k, reshapes the singular value spectrum of the SR and consequently, its effective rank. While the SR matrix is approximated using FB in continuous state-action settings, a consistent pattern emerges across both discrete and continuous domains: increasing k accelerates the decay of the tail singular values, encouraging the representation to concentrate on the dominant, principal modes of the dynamics, which are relevant for long-horizon navigation and control tasks. This spectral concentration aligns the structure of the SR with the low-rank inductive bias of FB, facilitating a more effective representation learning.

However, excessive temporal abstraction can negatively impact the representation. As both the stable rank and normalized spectral entropy approach their minima, the representation also begins to lose task-relevant dynamical information. This behavior is also partially predicted by the theory, which highlights a trade-off between spectral compression and the bias introduced by action repetition. Empirically, Figure 4a corroborates this phenomenon, demonstrating performance degradation once k exceeds an optimal threshold.

5.4 Temporal Abstraction and Embedding Dimension of FB

Theoretically, increasing the embedding dimension in FB should enable a more faithful approximation of the SR (Blier et al., 2021; Touati & Ollivier, 2021). One might therefore expect that enlarging the representation capacity would alleviate the mismatch between the high intrinsic rank of the SR and the finite-rank FB approximation. In practice, particularly in continuous environments, this



(a) (b)

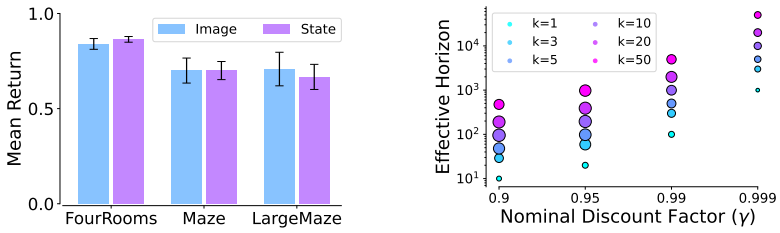
Figure 5: **Ablations: Bellman error.** Increasing the embedding dimension (a) or discount factor (b) without using temporal abstraction ($k = 1$) leads to an increase in the Bellman error.

expectation does not hold. As shown in Figure 5a, increasing the embedding dimension d leads to a systematic increase in Bellman error. Crucially, this increase does not translate into improved performance: Figure 4b shows that without temporal abstraction ($k = 1$), scaling d from 25 to 400 yields no meaningful performance gain. This observation is consistent with the spectral analysis in Section 4, which characterizes how the SR spectrum depends on the singular values of the transition operator. These findings suggest that, in continuous settings, increasing representational capacity alone may encourage the network to fit high-frequency components of the underlying high-rank SR. Such components are difficult to predict accurately, and, in the presence of bootstrapping, their errors can propagate through the Bellman updates, leading to higher Bellman errors without improved control performance.

In contrast, introducing temporal abstraction leads to a substantial performance improvement (Figure 4b). Rather than increasing capacity to match a spectrally complex target, temporal abstraction reduces the effective rank of the SR, thereby simplifying the representation problem. While temporal abstraction provides consistent gains, further improvements can be achieved by tuning the embedding dimension to the specific environment.

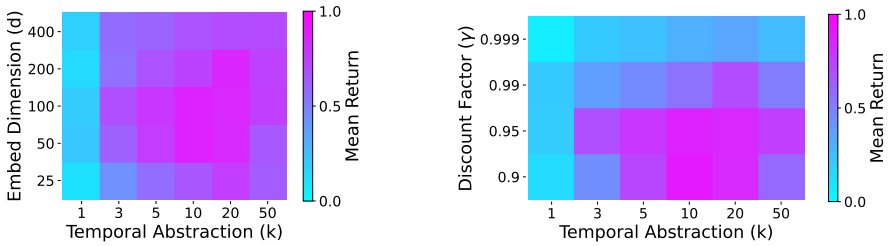
5.5 Spectral Dynamics: Discounting vs. Temporal Abstraction

A low-rank SR concentrates spectral energy into dominant singular values associated with the long-horizon transition dynamics necessary for task completion. In theory, as the discount factor $\gamma \to 1$, more dominant spectral modes undergo amplification, leading to a disproportionate concentration of spectral energy. As shown in Figure 2, this effect is particularly pronounced in continuous settings: FB networks with limited capacity prioritize these dominant long-horizon features, effectively "pruning" the sub-dominant singular components associated with high-frequency, short-term dynamics. However, this spectral compression comes at a cost; Figure 5b demonstrates that Bell-



(a) (b)

Figure 6: **Ablations: Input type and effective discount factor** (a): Image and State inputs use CNN and RBF encodings, respectively. (b): A higher return (larger radius) for a similar task horizon can be achieved by combining a lower nominal γ with a higher k.



(a) (b)

Figure 7: **Ablation: Temporal Abstraction vs. Embedding Dimension vs. Discount Factor** Temporal abstraction offers a considerable boost in performance even with a moderate number of steps k. After the introduction of temporal abstraction, the performance is less sensitive to variations in the embedding dimension *(a)* than to changes in the discount factor *(b)*, especially as $\gamma \to 1$.

man error increases sharply as $\gamma \to 1$, highlighting the inherent training instability of high-discount regimes.

The fundamental difference between discounting and temporal abstraction lies in how they modify the spectral distribution. Increasing γ amplifies existing modes, which preserves the high-frequency "noise" of the dynamics but scales the dominant modes toward infinity, degrading the operator's conditioning. In contrast, increasing the temporal abstraction k fundamentally smooths the dynamics by acting as a non-linear filter that exponentially attenuates sun-dominant, high-frequency modes while preserving the steady-state structure.

Our results highlight a sharp contrast between this discount-driven compression and k-repeat temporal abstraction. While both reduce the effective rank, the reduction under γ is increasingly abrupt and unstable, leading to a sudden collapse in both stable rank and spectral entropy. Conversely, increasing k induces a more controlled spectral decay; the effective rank decreases rapidly for small k before tapering off smoothly, preserving a higher degree of spectral entropy. This suggests that temporal abstraction facilitates a structured simplification of the predictive manifold without the numerical instability associated with near-unity discount factors.

6 A Recipe for Effective Forward-Backward Representations

Our results suggest that optimal performance is not achieved by simply maximizing γ, but rather through a synergy between moderate discounting and temporal abstraction. While lower γ offers stable training at the cost of a shorter task horizon, this loss of horizon can be compensated by increasing the action-repeat factor k.

Critically, we distinguish between the nominal discount factor γ used in the k-repeat MDP and the effective discount γ_{eff} of the original environment, related by $\gamma_{\text{eff}} = \gamma^{1/k}$ (alternatively, for a fixed effective horizon, the nominal γ is scaled as $\gamma = \gamma_{\text{eff}}^{k}$). As shown in Figure 6b, for a fixed 100-step task horizon, the combination of a lower nominal discount ($\gamma = 0.9$) and higher action repetition ($k = 10$) consistently outperforms the standard high-discount approach ($\gamma = 0.99, k = 1$).

Finally, Figure 7 provides a global performance landscape across embedding dimensions d, discount factors γ, and temporal abstraction scales k. The emerging pattern suggests that a moderate level of temporal abstraction ($k \in [5, 10]$) acts as a robust regularizer for FB representations. While the precise optimal k remains an environment-dependent hyperparameter, its inclusion provides a performance boost that is consistent across a wide range of embedding capacities and discounting regimes.

7 Conclusion

In this work, we identify a fundamental mismatch between the theoretical promise of FB and its practical efficacy. Our spectral analysis reveals that while FB imposes a low-rank bias, the underlying SR in continuous settings is inherently high-rank, with a heavy tail of high-frequency modes that resist function approximation and propagate bootstrapping errors.

To bridge this gap, we propose temporal abstraction as a principled spectral regulator. We theoretically demonstrate that action repetition acts as a spectral low-pass filter that reduces FB approximation error by exponentially attenuating high-frequency noise while preserving the steady-state structure of the MDP. This transformation effectively lowers the intrinsic rank of the SR target, providing a cleaner, more learnable signal for FB factorization. Our empirical results across continuous maze navigation tasks confirm that this spectral smoothing enables high-capacity networks to remain stable, even in high-discount regimes where standard FB typically fails.

Ultimately, our work advocates for a shift in perspective for representation learning: rather than attempting to force a complex, high-rank dynamical system through a narrow neural bottleneck, researchers should first shape the spectral structure of the underlying transition process. By leveraging temporal abstraction as a structural regularizer, we shift the burden of representation learning from the function approximator toward a strategic design of interaction dynamics, paving the way for more robust and scalable predictive representations.

8 Limitations and Future Work

In this work, we focus on action repetition as the simplest form of temporal abstraction. While our analysis suggests that its spectral smoothing effect arises from multi-step transition composition, more expressive forms of temporal abstraction, such as options or learned skills, may induce richer forms of spectral conditioning. Extending our theoretical framework to structured temporal abstractions remains an important direction for future research.

Additionally, our experiments rely on online interaction, which is most effective in environments with moderate state dimensionality. Scaling to higher-dimensional observations may benefit from incorporating offline datasets, as explored in recent work on data-augmented and zero-shot reinforcement learning (Sikchi et al., 2025; Tirinzoni et al., 2025). Investigating how temporal abstraction can be combined with offline data to further improve exploration, spectral conditioning, and representation learning is a promising avenue for future work.

A Appendix

In this appendix, we summarize the key hyperparameters used for training the Forward–Backward (FB) representation (Table 1). These settings were kept fixed across experiments unless otherwise stated.

Table 1: Relevant hyperparameters used for training FB representation.

Hidden Layers (Forward, Backward, Actor)	[256, 256]
Ensemble (Forward)	2
Learning Rate (Backward, Actor)	1e-6
Learning Rate (Forward)	1e-5
Batch size (state)	512
Batch size (image)	128
Replay Buffer size	1e6
FB Orthogonal Loss Coef.	1.0
z-latent: buffer data vs. random sampling ratio	0.5
z-latent: hold steps before resampling	10

Acknowledgments

This work was funded by the Carl Zeiss Foundation through the ReScaLe project.

References

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4058–4068, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

André Biedenkapp, Raghu Rajan, Frank Hutter, and Marius Lindauer. Temporl: Learning when to act. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139, pp. 914–924, 2021.

Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *CoRR*, abs/2101.07123, 2021. URL https://arxiv.org/abs/2101.07123.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Comput.*, 5(4):613–624, July 1993. ISSN 0899-7667. DOI: 10.1162/neco.1993.5.4.613. URL https://doi.org/10.1162/neco.1993.5.4.613.

Bastien Dubail, Stefan Stojanovic, and Alexandre Proutière. Shift before you learn: Enabling low-rank representations in reinforcement learning. *arXiv preprint arXiv:2509.05193*, 2025.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. URL https://api.semanticscholar.org/CorpusID:10163399.

Tejas D. Kulkarni, Ardavan Saeedi, Simanta Gautam, and Samuel J. Gershman. Deep successor reinforcement learning. *ArXiv*, abs/1606.02396, 2016. URL https://api.semanticscholar.org/CorpusID:11965834.

Aravind S. Lakshminarayanan, Sahil Sharma, and Balaraman Ravindran. Dynamic action repetition for deep reinforcement learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 2133–2139. AAAI Press, 2017.

Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. A laplacian framework for option discovery in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 2295–2304. JMLR.org, 2017a.

Marlos C. Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. Eigenoption discovery through the deep successor representation. *ArXiv*, abs/1710.11089, 2017b. URL https://api.semanticscholar.org/CorpusID:3300406.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: evaluation protocols and open problems for general agents (extended abstract). In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 5573–5577. AAAI Press, 2018. ISBN 9780999241127.

Sridhar Mahadevan. Proto-value functions: developmental reinforcement learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 553–560, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. DOI: 10.1145/1102351.1102421. URL https://doi.org/10.1145/1102351.1102421.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025.

Dikshant Shehmar, Matthew Schlegel, Matthew E. Taylor, and Marlos C. Machado. Laplacian representations for decision-time planning. *CoRR*, abs/2602.05031, 2026.

Harshit S. Sikchi, Andrea Tirinzoni, Ahmed Touati, Yingchen Xu, Anssi Kanervisto, Scott Niekum, Amy Zhang, Alessandro Lazaric, and Matteo Pirotta. Fast adaptation with behavioral foundation models. *ArXiv*, abs/2504.07896, 2025. URL https://api.semanticscholar.org/CorpusID:277667156.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999. ISSN 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(99)00052-1. URL https://www.sciencedirect.com/science/article/pii/S0004370299000521.

Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-shot whole-body humanoid control via behavioral foundation models. *ArXiv*, abs/2504.11054, 2025. URL https://api.semanticscholar.org/CorpusID:277787058.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? *ArXiv*, abs/2209.14935, 2022. URL https://api.semanticscholar.org/CorpusID:252596234.

Jingwei Zhang, Jost Tobias Springenberg, Joschka Boedecker, and Wolfram Burgard. Deep reinforcement learning with successor features for navigation across similar environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2371–2378. IEEE Press, 2017. DOI: 10.1109/IROS.2017.8206049. URL https://doi.org/10.1109/IROS.2017.8206049.

Supplementary Materials

The following content was not necessarily subject to peer review.

B Proofs

In the following, we provide the proofs of this work. Note, that in Touati & Ollivier (2021), the reward embedding $z_R := B^\top r\nu$ is weighted by a data distribution ν. For this work, we assume a uniform distribution and implicitly absorb its normalization constant into the scaling of B, simplifying the embedding to the matrix-vector product $z_R = B^\top r$.

Theorem 3.1 (Optimality Gap for FB Representations)**.** *Let r be a reward function such that $z_R = B(s,a)r(s,a)$. The approximation error of the optimal Q-function is bounded by:*

$$\left\|F(\cdot,\cdot,z_R)^\top z_R - Q^\star\right\|_\infty \leq \frac{2\|r\|_\infty}{(1-\gamma)}\|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2. \tag{2}$$

Proof. Note, that Theorem 3.1 is a simplified refinement (Touati & Ollivier, 2021, Theorem 8) for our spectral analysis setting. From this theorem we directly have that

$$\left\|F(\cdot,\cdot,z_R)^\top z_R - Q^\star\right\|_\infty \leq \frac{2\|r\|_A}{(1-\gamma)}\sup_{s,a}\|\hat{M}^{z_R}(s,a,\cdot,\cdot) - M^{\pi_{z_R}}(s,a,\cdot,\cdot)\|_B.$$

where we choose $\|\cdot\|_A = \|\cdot\|_\infty$ and $\|\cdot\|_B = \|\cdot\|_2$. Given this we have

$$\begin{aligned}\left\|F(\cdot,\cdot,z_R)^\top z_R - Q^\star\right\|_\infty &\leq \frac{2\|r\|_\infty}{(1-\gamma)}\sup_{s,a}\|\hat{M}^{z_R}(s,a,\cdot,\cdot) - M^{\pi_{z_R}}(s,a,\cdot,\cdot)\|_2 \\ &\leq \frac{2\|r\|_\infty}{(1-\gamma)}\|\hat{M}^{z_R} - M^{\pi_{z_R}}\|_2.\end{aligned}$$

□

Lemma 4.1 (Optimality Gap for k-repeat FB Representations)**.** *Given a repeat factor k, a reward function $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and an FB (F,B) representation with dimension d, the error in approximating the original optimal action-value function $Q^\star$ is bounded by:*

$$\left\|F(\cdot,\cdot,z_R)^\top z_R - Q^\star\right\|_\infty \leq \epsilon_{\text{repeat}}(k) + \frac{2\|r\|_\infty}{1-\gamma}\left(\epsilon_{\text{real}}(r) + \frac{1}{1-\gamma(\sigma_{d+1}(P_{\text{rep}}))^k}\right).$$

Proof. We start by deriving simple bounds for singular values of the successor representation $\widetilde{M}^\pi$ from basic singular value decomposition properties.

First we have $\sigma_i(\widetilde{P}^\pi) \leq \sigma_i(P_{\text{rep}}^k)\sigma_1(\widetilde{\pi}) \leq \sigma_i(P_{\text{rep}}^k)$, where we used that $\sigma_1(\widetilde{\pi}) \leq 1$. The repetition matrix P_{rep}^k consists of the stacked blocks P_{a_q} and the matrix K. As K is a commutation matrix all its singular values are 1, thus the spectrum can be decomposed

$$\sigma(P_{\text{rep}}^k) = \bigcup_{q=1}^{|A|}\sigma(P_{a_q}^k).$$

Combined with the fact that $\sigma_i(P_{a_q}^k) \leq \sigma_i(P_{a_k})^k$ we have that $\sigma_i(P_{\text{rep}}^k) \leq \sigma_i(P_{\text{rep}})^k$.

Given that $\widetilde{M}^\pi = \sum_{t=0}^\infty \gamma^t P^{kt}$, we further have

$$\sigma_i(\widetilde{M}^\pi) \leq \sum_{t=0}^\infty \gamma^t\sigma_i(\widetilde{P}^\pi) \leq \sum_{t=0}^\infty \gamma^t(\sigma_i(\widetilde{P}))^t. \tag{4}$$

Thus, combined we can derive an upper bound of the i-th singular value of $\widetilde{M}^{\pi}$ by

$$\sigma_i(\widetilde{M}^{\pi}) \leq \frac{1}{1-\gamma\sigma_i(\widetilde{P}^{\pi})} = \frac{1}{1-\gamma\sigma_i(P_{\text{rep}}^k)\sigma_1(\widetilde{\pi})} \leq \frac{1}{1-\gamma(\sigma_i(P_{\text{rep}}))^k\sigma_1(\tilde{\pi})}.$$

The final step is using the definitions Def. 3.1, Def. 4.2 and Theorem 3.1:

$$\begin{aligned}
\left\|F(\cdot,\cdot,z_R)^\top z_R - Q^\star\right\|_\infty &\leq \left\|F(\cdot,\cdot,z_R)^\top z_R - \widetilde{Q}^\star\right\|_\infty + \|\widetilde{Q}^\star - Q^\star\|_\infty \\
&= \epsilon_{\text{repeat}}(k) + \left\|F(\cdot,\cdot,z_R)^\top z_R - \widetilde{Q}^\star\right\|_\infty \\
&\leq \epsilon_{\text{repeat}}(k) + \frac{2\|r\|_\infty}{1-\gamma}\|\hat{M}^{z_R} - \widetilde{M}^{\pi_{z_R}}\|_2 \\
&\leq \epsilon_{\text{repeat}}(k) + \frac{2\|r\|_\infty}{1-\gamma}\left(\epsilon_{\text{real}}(r) + \sigma_{d+1}(\widetilde{M}^{\pi_{z_R}})\right) \\
&\leq \epsilon_{\text{repeat}}(k) + \frac{2\|r\|_\infty}{1-\gamma}\left(\epsilon_{\text{real}}(r) + \frac{1}{1-\gamma(\sigma_{d+1}(P_{\text{rep}}))^k}\right).
\end{aligned}$$

This completes the proof. □

C Stable-Rank Bound

We show that the bound for the Stable-Rank value decreases as we increase the temporal abstraction via increasing the number of action-repetitions k.

We start from the definition of Stable-Rank, i.e.,

$$\text{SRank}(M) = \frac{\|M\|_F^2}{\|M\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2}$$

and Eq. (4), we can write the bound for the stable rank as k increases.

First, notice that we have

$$\|\widetilde{M}^{\pi}\|_2 = \sigma_1(\widetilde{M}^{\pi}) = \frac{1}{1-\gamma\sigma_1(\widetilde{P}^{\pi})}.$$

Then, using the singular value bound, we can obtain

$$\|\widetilde{M}^{\pi}\|_F^2 = \sum_i \sigma_i(\widetilde{M}^{\pi})^2 \leq \frac{1}{(1-\gamma\sigma_1(\widetilde{P}^{\pi}))^2} + \sum_{i\geq 2}\frac{1}{(1-\gamma(\sigma_i(\widetilde{P}^{\pi}))^k)^2}.$$

Using $\sigma_i(\widetilde{P}^{\pi}) \leq \rho < 1$ for $i \geq 2$ assumption, we have

$$\|\widetilde{M}^{\pi}\|_F^2 \leq \frac{1}{(1-\gamma\sigma_1(\widetilde{P}^{\pi}))^2} + (|\mathcal{S}|-1)\frac{1}{(1-\gamma\rho^k)^2}.$$

Combining the two expressions, we conclude that

$$\text{SRank}(\widetilde{M}^{\pi}) \leq 1 + (|\mathcal{S}|-1)\left(\frac{1-\gamma\sigma_1(\widetilde{P}^{\pi})}{1-\gamma\rho^k}\right)^2,$$

which shows that as k increases, we have $\rho^k \to 0$, so the stable rank contracts toward 1. This establishes that k-step temporal abstraction induces stronger spectral concentration in the SR, hence decreasing the effective rank of SR.

D Normalized Spectral Entropy Bound

Similar to appendix C, we provide a bound to the normalized spectral entropy of $\widetilde{M}^\pi$ using Eq. (4). Firstly, recall that

$$\mathrm{NSE}(\widetilde{M}^\pi) = \frac{-\sum_i p_i \log p_i}{\log(|\mathcal{S}|)}, \quad p_i = \frac{\sigma_i(\widetilde{M}^\pi)^2}{\sum_j \sigma_j(\widetilde{M}^\pi)^2},$$

which lies in $[0, 1]$ with lower values indicating stronger spectral concentration (i.e., more pronounced low-rank structure). As in the previous discussion, we have

$$\sigma_1(\widetilde{M}^\pi) = \frac{1}{1 - \gamma\sigma_1(\widetilde{P}^\pi)}.$$

Using the singular value bound (4), we have

$$\sigma_i(\widetilde{M}^\pi) \leq \frac{1}{1 - \gamma(\sigma_i(\widetilde{P}^\pi))^k}.$$

For $i \geq 2$, assuming $\sigma_i(\widetilde{P}^\pi) \leq \rho < 1$, and hence,

$$\sigma_i(\widetilde{M}^\pi) \leq \frac{1}{1 - \gamma\rho^k}.$$

Now define

$$E_1 = \frac{1}{(1 - \gamma\sigma_1(\widetilde{P}^\pi))^2}, \qquad E_{\mathrm{rest}} \leq (|\mathcal{S}| - 1)\frac{1}{(1 - \gamma\rho^k)^2},$$

then the total energy satisfies

$$\sum_i \sigma_i(\widetilde{M}^\pi)^2 = E_1 + E_{\mathrm{rest}}.$$

Therefore, the normalized weight of the dominant mode is bounded below by

$$p_1 = \frac{E_1}{E_1 + E_{\mathrm{rest}}} \geq \frac{1}{1 + (|\mathcal{S}| - 1)\left(\frac{1 - \gamma\sigma_1(\widetilde{P}^\pi)}{1 - \gamma\rho^k}\right)^2}.$$

As k increases, $\rho^k \to 0$, implying $p_1 \to 1/(1 + (|\mathcal{S}| - 1)(1 - \gamma\sigma_1(\widetilde{P}^\pi))^2)$. Since the dominant spectral weight increases monotonically with k, the spectral distribution becomes increasingly concentrated, and thus $\mathrm{NSE}(\widetilde{M}^\pi)$ decreases with k.

E Extra Plots and Ablations

Figure 8 provides an overview of the effect of the three main hyperparameters of FB on final episodic return of the Four-Rooms continuous environment. Two values are of particular importance, action-repetition (k=1) and nominal discount factor (γ=0.999). In both cases, the performance suffers significantly regardless of the values of other hyperparameters. In the case of k=1 or no temporal abstraction, FB networks find it challenging to learn a good representation due to the presence of unpredictable high-frequency dynamical modes. In the case high discount factor, γ=0.999, a good representation cannot be achieved as the representation rank approaches singularity. The learning is less sensitive overall to the value of the embedding dimension d.

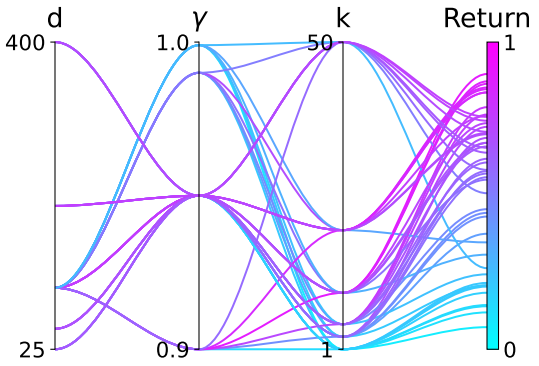
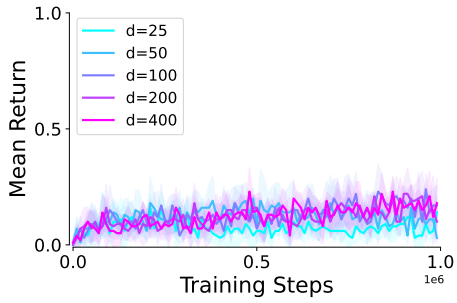
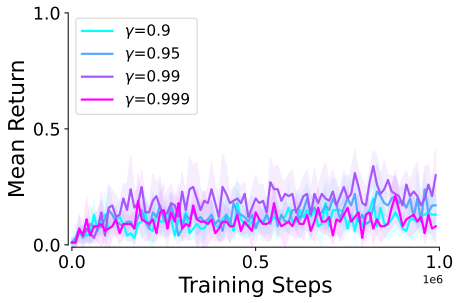


Figure 8: **Relationship between the main hyperparameters and episodic return.** This plot gives an overview of the effect of different combinations of embedding dimension (d), discount factor (γ), and temporal abstraction k over all experiments. Noticeably, $k = 1$ or $\gamma = 0.999$ leads to poor performance in most combinations.

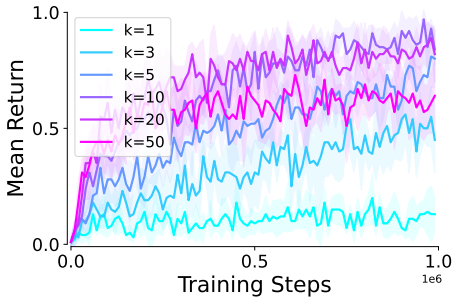
Figure 9, shows the performance of different combination of the main hyperparameters during training with the focus on the effect of introducing temporal abstraction. Figures 9a and 9b highlight that without temporal abstraction (k=1) varying the embedding dimension d or the discount factor γ yields no significant improvement. Figure 9c, on the other hand shows that even a small level of temporal abstraction (k=3) can lead to a significant boost in performance. The figure also shows the limitation of the temporal abstraction where a large temporal abstraction (k=50) can start to have negative impact on the performance, by oversimplification of the SR representation and removing dynamical modes that are useful for the navigation task.



(a) γ=0.95, k=1



(b) d=100, k=1



(c) d=100, γ=0.95

Figure 9: **Ablation: Training plots.** Increasing embedding dimension d or discount factor γ without increasing the temporal abstraction does not yield a meaningful increase in performance.

Figure 10 shows a more complete picture of SR and it's associated Q function (mean over cardinal action directions). The *Baseline* shows the SR and Q using no temporal abstraction (k=1), and moderate discount factor (γ=0.95). For the continuous settings SR is calculated via FB with embedding dimension (d=100).

In discrete settings (top two rows), a low-rank structure can be achieved in three ways: 1) SVD with a small rank *(rank = 4)*, 2) high discount factor *(γ=0.999)*, or 3) using temporal abstraction via action repetition *(k=10)*. In the absence of function approximation and bootstrapping all three paths lead to an overall similar result where a low-rank structure can remove the high frequency dynamical modes and create shared future topology (rooms, corridors,...) where states with similar reachability are grouped together and have similar values.

In continuous settings (bottom two rows), where SR and its associated Q are learned via FB using function approximation and bootstrapping, the results differ. To enforce a low-rank structure via the FB algorithm we reduce the embedding dimension from 100 to 25. The figure shows a small smoothing (grouping of states), but this is not nearly close to the effect of enforcing low-rank structure using SVD in the discrete setting. Increasing the discount factor (γ=0.999) and introducing temporal abstraction via action repetition (k=10) show more promise as they both help spread the SR and Q values to the neighboring rooms. However, a closer look at the Q values show that only temporal abstraction can smoothly distribute the Q values as the states move away from the goal (start marker). The policy based on increased γ will be stuck in local maxima while the policy based on increased k, can follow the Q gradients to the goal.

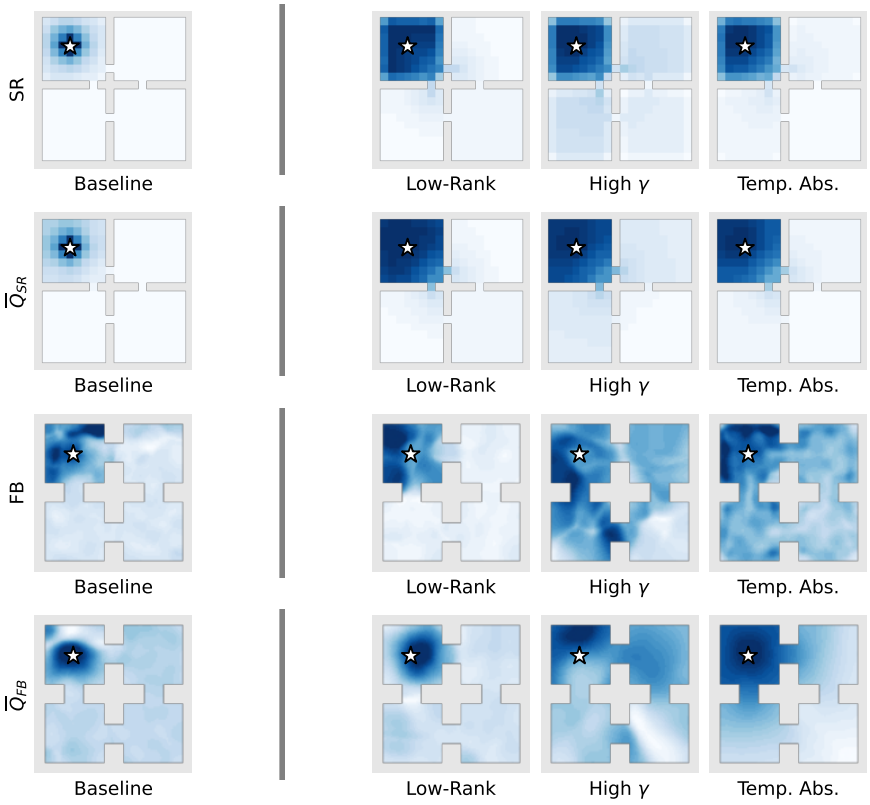


Figure 10: **Successor Representation (SR) and its Q-Function - Discrete and Continuous** This is a more complete picture of Figure 1. The Q-functions are derived from the same SR that is presented here. The *star marker* marks the starting state for SR and the goal state for the Q-function.