BEYOND THE SPECTRAL BULK: SPIKE ACCELERATION FOR CRITIC LEARNING IN DEEP REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Value-function approximation can destabilize deep reinforcement learning because critic errors affect both bootstrapped targets and subsequent policy updates. Existing methods mainly stabilize learning targets, reduce estimation error, or constrain critic parameters, but do not directly promote weak task-related directions that remain unresolved within the spectral bulk. We propose Spike Spectral Acceleration Regularization (SSAR), which selectively amplifies critic singular modes near the estimated upper bulk edge and removes the additional pressure after sufficient spectral separation.

SSAR is motivated by the Baik–Ben Arous–Péché (BBP) transition, under which an emerging low-rank signal becomes increasingly aligned with an empirical singular mode as it approaches and crosses the separation threshold. For a fixed-target critic update, we decompose the SSAR-induced Bellman-loss change into a task-aligned descent term and a remainder, and characterize their scaling across the BBP transition. Under a rank-one spiked model, the two terms have the same asymptotic order in the fixed-subcritical regime, providing no sign guarantee. At non-degenerate criticality and in the fixed-supercritical regime, the aligned term dominates, making the additional SSAR step loss-decreasing with probability tending to one.

This analysis shows that spectral proximity determines where SSAR acts, while the emergence of task alignment determines when the additional update becomes beneficial. Controlled experiments verify the predicted finite-width scaling and loss-sign transition. We further evaluate SSAR with Soft Actor-Critic on six continuous-control tasks from the DeepMind Control Suite against vanilla SAC and matched L_2 critic regularization.

1 INTRODUCTION

Value functions are central to deep reinforcement learning because they propagate reward information and guide policy improvement. However, function approximation, bootstrapping, and off-policy data can destabilize learning, especially in actor–critic methods where critic errors influence subsequent policy updates Zhang et al. (2021); Fellows et al. (2023); Tang & Berseth (2024).

Existing approaches mainly stabilize critic learning by modifying targets or controlling estimation error. Target networks, Double Q-learning, averaging, and ensemble methods improve robustness against target drift, overestimation, and uncertainty Mnih et al. (2013); Hasselt (2010); Van Hasselt et al. (2016); Anschel et al. (2017); Fujimoto et al. (2018); Lee et al. (2021); Chen et al. (2021). Other works study representation degradation, including capacity loss, primacy bias, dormant neurons, and loss of plasticity Kumar et al. (2021; 2022); Lyle et al. (2022); Nikishin et al. (2022); Sokar et al. (2023); Lyle et al. (2023); D'Oro et al. (2023); Dohare et al. (2024). However, these methods do not directly promote weak task-related directions that remain unresolved within the critic spectrum.

Random matrix theory (RMT) provides a tool for studying such unresolved structure through high-dimensional spectra Pennington & Worah (2017); Louart et al. (2018); Mahoney & Martin (2019); Martin & Mahoney (2021). The Marchenko–Pastur law describes the random spectral bulk, while deviations from this baseline reveal structured components. The BBP transition Baik et al. (2005)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

further characterizes when a weak low-rank signal becomes spectrally identifiable: near the transition, the empirical singular vector gradually aligns with the underlying signal, and above it the spike separates from the bulk Benaych-Georges & Nadakuditi (2012).

Motivated by this observation, we propose Spike Spectral Acceleration Regularization (SSAR), a critic-side intervention that selectively amplifies singular modes near the estimated upper bulk edge. The key distinction is that spectral proximity only determines where SSAR acts; the effectiveness depends on the task alignment already contained in the selected mode. SSAR therefore does not assume that every near-edge mode is useful.

Unlike conventional spectral normalization or spectrum-preserving methods Sanyal et al. (2020); Gogianu et al. (2021); Bjorck et al. (2021); Lewandowski et al. (2025), SSAR temporarily amplifies modes inside a BBP-scale edge window and removes the additional pressure after spectral separation. For a fixed-target critic update, we decompose the induced Bellman-loss change into a task-aligned descent term and a remainder. Under a rank-one spiked model, both terms have order $O_p(m^{-1})$ in the fixed-subcritical regime, while the aligned term dominates at criticality ($\Theta_p(m^{-2/3})$) and in the fixed-supercritical regime ($\Theta_p(m^{-1/2})$). Hence the additional update becomes loss-decreasing with probability tending to one once sufficient alignment emerges.

We verify these predictions using controlled spiked-model experiments and evaluate SSAR with Soft Actor-Critic on six DeepMind Control Suite tasks. The results confirm the predicted scaling behavior and show improved sample efficiency compared with vanilla SAC and matched L_2 critic regularization.

2 Related Work

Stabilizing critic learning. A large body of work addresses instability in value-based and actor–critic reinforcement learning through target control, estimation-error reduction, and uncertainty modeling Mnih et al. (2013); Hasselt (2010); Van Hasselt et al. (2016); Anschel et al. (2017); Fujimoto et al. (2018); Lee et al. (2021); Chen et al. (2021). SSAR is complementary to these approaches: it leaves the Bellman target unchanged and instead modifies critic spectral dynamics.

Representation degradation and spectral methods. Recent studies have analyzed capacity loss, primacy bias, dormant neurons, and loss of plasticity in deep RL Kumar et al. (2021; 2022); Lyle et al. (2022); Nikishin et al. (2022); Sokar et al. (2023); Lyle et al. (2023); D'Oro et al. (2023); Dohare et al. (2024). Meanwhile, RMT has been used to characterize neural-network spectra and bulk-plus-spike structures Pennington & Worah (2017); Louart et al. (2018); Mahoney & Martin (2019); Martin & Mahoney (2021). Existing spectral regularization methods mainly constrain or normalize spectra Sanyal et al. (2020); Gogianu et al. (2021); Bjorck et al. (2021); Lewandowski et al. (2025). In contrast, SSAR selectively amplifies near-edge modes and provides a regime-dependent analysis of how its effect changes across the BBP transition.

3 Preliminaries and Notation

3.1 Reinforcement Learning

We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, R, \mu, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(s' \mid s, a)$ is the transition kernel, $R(s, a) \in \mathbb{R}$ is the reward function, μ is the initial-state distribution, and $\gamma \in (0, 1)$ is the discount factor. The goal is to learn a stochastic policy $\pi^\phi(a \mid s)$ that maximizes the expected discounted return $J(\pi^\phi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$, where actions are sampled $a_t \sim \pi^\phi(\cdot \mid s_t)$ and transitions follow $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.

For a fixed policy π^ϕ, the state–action value function (Q-function) $Q^{\pi^\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $Q^{\pi^\phi}(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid (s_0, a_0) = (s, a)]$.

We approximate Q^{π^ϕ} with a critic $Q^\theta(s, a)$ parameterized by θ. Let y denote the target used in the current critic update. As in fitted-Q and target-network methods, y is treated as fixed when differentiating the critic objective. We define the Bellman residual $r(s, a) = Q^\theta(s, a) - y$, and the

corresponding loss

$$L_0(Q^\theta) = \frac{1}{2}\mathbb{E}\left[r(s,a)^2\right]. \tag{1}$$

A standard gradient-based update is

$$\theta' = \theta - \eta\,\nabla_\theta L_0(Q^\theta) = \theta - \eta\,\mathbb{E}_{(s,a,s')\sim\mathcal{D}}\big[r(s,a)\nabla_\theta Q^\theta(s,a)\big], \tag{2}$$

where $\eta > 0$ denotes the step size.

To connect our spectral analysis to network parameters, consider a typical parametrization of the critic

$$Q^\theta(h) = o^\top \varphi(Wh), \tag{3}$$

where $W \in \mathbb{R}^{n\times m}$ denotes the penultimate-layer weight matrix, $h \in \mathbb{R}^m$ is its input feature, $o \in \mathbb{R}^n$ is the output weight, and $\varphi(\cdot)$ the penultimate-layer activation.

Let $\{s_k, u_k, v_k\}_{k\geq 1}$ denote the singular triplets of W, ordered as $s_1 \geq s_2 \geq \cdots$. Then $Wv_k = s_k u_k$ and $W^\top u_k = s_k v_k$. We define $\lambda_k = s_k^2/m$ as the normalized squared singular values, which are the eigenvalues of $WW^\top/m$.

3.2 Random Matrix Theory

We recall two standard results from random matrix theory that we use in the paper.

Throughout the paper, we use standard asymptotic notation. For deterministic sequences, $a_m = O(b_m)$ means that $|a_m|/|b_m|$ remains bounded as $m \to \infty$, while $a_m = \Theta(b_m)$ means that the two sequences have the same asymptotic order. For random quantities, $X_m = O_p(a_m)$ means that X_m/a_m is bounded in probability, and $X_m = \Theta_p(a_m)$ means that $|X_m|/a_m$ is bounded both above and away from zero in probability. We write $X_m \xrightarrow{p} X$ for convergence in probability. Throughout the spectral analysis, we use $\beta = n/m$ for the layer aspect ratio. We let σ^2 denote the entrywise variance prescribed by the initialization of the selected critic layer. This fixed initialization variance serves as the background variance parameter in the Marchenko–Pastur reference spectrum.

Theorem 3.1 (Marchenko–Pastur law Marčenko & Pastur (1967))**.** *Let X be an $n \times m$ random matrix with i.i.d. entries having zero mean and variance $\sigma^2 < \infty$. Define the sample covariance $Y = \frac{1}{m}XX^\top$ and let $\rho_Y(\lambda)$ be the empirical spectral distribution (ESD) of Y. If $n, m \to \infty$ with $n/m \to \beta \in (0,\infty)$, then almost surely the ESD ρ_Y converges weakly to the Marchenko–Pastur distribution*

$$\rho_Y^\beta(\lambda) = \frac{1}{2\pi\beta\sigma^2}\,\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}\,\mathbf{1}_{[\lambda_-,\lambda_+]},$$

where $\lambda_\pm = \sigma^2(1 \pm \sqrt{\beta})^2$ are the edges of the support.

Theorem 3.2 (BBP phase transition Baik et al. (2005))**.** *Let $x_1, \ldots, x_m \in \mathbb{C}^n$ be independent complex Gaussian sample vectors with population covariance matrix $\Sigma = \mathrm{diag}(\ell_1, \ldots, \ell_n)$, and let $X = [x_1, \ldots, x_m]$. Define the sample covariance matrix $Y = \frac{1}{m}XX^*$, and let λ_1 denote its largest eigenvalue. Fix an integer $r \geq 0$ and suppose that $\ell_{r+1} = \cdots = \ell_n = 1$. Assume $n, m \to \infty$ with $\beta = n/m$ lying in a compact subset of $(0, 1]$. If, for some $0 \leq k \leq r$, the first k population spikes satisfy $\ell_1 = \cdots = \ell_k = 1 + \sqrt{\beta}$, while $\ell_{k+1}, \ldots, \ell_r$ lie in a compact subset of $(0, 1 + \sqrt{\beta})$, then $\mathbb{P}\Big(\frac{\beta^{1/6}}{(1+\sqrt{\beta})^{4/3}} m^{2/3}\left[\lambda_1 - (1+\sqrt{\beta})^2\right] \leq x\Big) \to F_k(x)$ as $n, m \to \infty$.*

If, for some $1 \leq k \leq r$, the top population spikes satisfy $\ell_1 = \cdots = \ell_k$ and lie in a compact subset of $(1 + \sqrt{\beta}, \infty)$, while $\ell_{k+1}, \ldots, \ell_r$ lie in a compact subset of $(0, \ell_1)$, then $\mathbb{P}\left(\frac{\left[\lambda_1 - \left(\ell_1 + \frac{\beta\ell_1}{\ell_1 - 1}\right)\right]\sqrt{m}}{\sqrt{\ell_1^2 - \frac{\beta\ell_1^2}{(\ell_1-1)^2}}} \leq x\right) \to G_k(x)$. Here, F_k denotes the corresponding Tracy–Widom-type limiting distribution and G_k is the largest-eigenvalue distribution of a $k \times k$ GUE matrix, as defined in Baik et al. (2005).

The Marchenko–Pastur law characterizes the limiting spectral distribution of a high-dimensional random matrix and identifies the deterministic bulk edges λ_- and λ_+. The BBP result describes how this spectral picture changes in the presence of a low-rank population signal. Before the signal

exceeds the critical strength, the leading sample eigenvalue remains at the upper bulk edge and fluctuates on the $m^{-2/3}$ scale. Once the signal becomes supercritical, the leading eigenvalue separates from the bulk and fluctuates around an outlier location on the $m^{-1/2}$ scale.

4 METHOD

We propose a lightweight spectral-acceleration module that can be added to standard value-network training to encourage earlier separation of task-relevant singular modes from the random spectral bulk. We first summarize the spectral regimes around the BBP transition, and then introduce the regularizer, stopping rule, and training procedure.

Consider the normalized weight matrix $\widetilde{W} = \frac{1}{\sqrt{m}}W = Z + \theta u^\star v^{\star\top}$, where Z is the random background, θ is the normalized rank-one signal strength, and $u^\star, v^\star$ are unit population signal directions. Since W and $\widetilde{W}$ have identical singular vectors, $\lambda_1 = s_1(W)^2/m = s_1(\widetilde{W})^2$.

Let θ_c denote the corresponding BBP threshold and define $\tau_m = |\langle v_1, v^\star\rangle|$ as the alignment between the empirical leading right singular vector and the population signal direction.

Corollary 4.1 (Spectral regimes around the BBP transition)**.** *For the spiked model $\widetilde{W}$, the leading spectral mode has the following asymptotic behavior.*

Subcritical regime. *If $\theta \le \theta_c - \varepsilon$ for some fixed $\varepsilon > 0$, then $\lambda_1 = \lambda_+ + O_p(m^{-2/3})$ and $\tau_m = O_p(m^{-1/2})$.*

BBP transition regime. *If $|\theta - \theta_c| = O(m^{-1/3})$, then $\lambda_1 = \lambda_+ + O_p(m^{-2/3})$, while the characteristic singular-vector overlap scale is $\tau_m = O_p(m^{-1/6})$.*

Supercritical regime. *If $\theta \ge \theta_c + \varepsilon$, then $\lambda_1 = f(\theta) + O_p(m^{-1/2})$ with $f(\theta) > \lambda_+$, and $\tau_m \xrightarrow{p} \tau_\infty(\theta) > 0$.*

A detailed derivation of Corollary 4.1 is given in Appendix A.

Motivation We view the selected critic-layer weight matrix during training as $W = W_{\text{init}} + \Delta W$, where W_{init} behaves like a random matrix and ΔW encodes the learned, low-rank task signal $\sum_{i=1}^{r} \theta_i u_i v_i^\top$.

At initialization, the empirical spectrum of $\frac{1}{m}W_{\text{init}}W_{\text{init}}^\top$ is described by the Marchenko–Pastur law (Theorem 3.1). According to Corollary 4.1, before the BBP transition the largest sample eigenvalue remains near the upper bulk edge λ_+, with $O(m^{-2/3})$ fluctuations. Once a signal direction becomes supercritical, the corresponding sample eigenvalue separates from the bulk and concentrates around an outlier location $f(\theta_i) > \lambda_+$, with $O(m^{-1/2})$ fluctuations. This spectral separation marks the formation of a spike. Our key idea is to accelerate this process by selectively increasing near-edge modes, so that task-related spectral structure can emerge earlier during critic training.

Spike regularizer Guided by the above, we design a simple per-update regularizer that targets near-edge singular modes. Let λ_+ denote the right edge of the MP bulk for the background ensemble of W. Fix a constant $c > 0$ and define the edge-window width $T = cm^{-2/3}$ and the spectral edge window $I = [\lambda_+, \lambda_+ + T]$. We call an empirical singular value s_k a *near-edge* mode if $s_k^2/m \in I$. For each such mode we add the per-mode penalty $(\lambda_+ + T - s_k^2/m)_+$, where $(x)_+ = \max\{x, 0\}$. The multi-mode spike regularizer is $L_{\text{SSAR}}(W) = \sum_{k:\, s_k^2/m \in I}(\lambda_+ + T - s_k^2/m)_+$.

Given the standard Bellman residual loss L_0, the total training loss is $L_{\text{tot}} = L_0 + \gamma L_{\text{SSAR}}$, where $\gamma > 0$ is a small regularization weight. For a simple singular value, $\nabla_W s_k = u_k v_k^\top$ and hence $\nabla_W(s_k^2/m) = (2s_k/m)u_k v_k^\top$. Therefore, when the k-th penalty is active, minimizing L_{SSAR} adds a positive update along $u_k v_k^\top$ and increases the corresponding singular value.

Stopping rule The spiked-matrix approximation is most appropriate during the early and intermediate stages of training. At later stages, weight spectra may develop stronger correlations and heavy-tailed structure Martin & Mahoney (2021). We therefore monitor the number of modes that

have moved beyond the acceleration window, $N_{\text{sep}}(W) = \sum_k \mathbf{1}\{s_k^2/m > \lambda_+ + T\}$. SSAR remains active while $N_{\text{sep}}(W) < N_c$ and is disabled once $N_{\text{sep}}(W) \geq N_c$, where N_c is the stopping threshold.

Algorithm Algorithms 1 and 2 summarize the integration of SSAR into a standard actor–critic update and the corresponding spectral loss computation. The SSAR term is evaluated every e_{spike} critic updates.

Algorithm 1 SSAR (high-level)

Input: policy parameters ϕ, critic parameters θ, coefficient γ
Initialize the base RL agent and set $t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  $L_{\text{SSAR}} \leftarrow$ COMPUTE-SSAR-LOSS(θ, t)
  $\theta \leftarrow \theta - \eta \nabla_\theta \big(L_0(Q^\theta) + \gamma L_{\text{SSAR}}\big)$
  Update ϕ using the base policy objective
until convergence

Algorithm 2 Compute-SSAR-Loss

Input: $\theta, \sigma, c, N_c, t, e_{\text{spike}}$
if $t \bmod e_{\text{spike}} \neq 0$ **then**
  return 0
end if
Extract W and compute top singular values $s_1, \ldots, s_r$
$\lambda_+ \leftarrow \sigma^2(1 + \sqrt{n/m})^2$, $T \leftarrow cm^{-2/3}$, $I \leftarrow [\lambda_+, \lambda_+ + T]$
$N_{\text{sep}} \leftarrow \sum_i \mathbf{1}\{s_i^2/m > \lambda_+ + T\}$
if $N_{\text{sep}} \geq N_c$ **then**
  return 0
end if
$L_{\text{SSAR}} \leftarrow \sum_{i:\, s_i^2/m \in I} (\lambda_+ + T - s_i^2/m)_+$
return L_{SSAR}

5 THEORETICAL ANALYSIS

We now study whether the spectral acceleration introduced by SSAR reduces the critic Bellman loss. The analysis focuses on one active leading singular mode and isolates the additional parameter update generated by the SSAR regularizer. Formal assumptions and complete proofs are given in Appendix B.

Let (s, u, v) denote the leading singular triplet of the selected critic weight matrix W. When this mode lies in the active edge window, its contribution to loss is $L_{\text{spike}}(W) = (\lambda_+ + T - s^2/m)_+$. Inside the active region, using $\nabla_W s = uv^\top$, we have $\nabla_W L_{\text{spike}}(W) = -(2s/m)uv^\top$.

A gradient step on the total critic objective $L_{\text{tot}} = L_0 + \gamma L_{\text{spike}}$ is therefore

$$\begin{aligned} W_{\text{new}} &= W - \eta \nabla_W L_0 - \eta\gamma \nabla_W L_{\text{spike}} \\ &= W - \eta \nabla_W L_0 + \frac{2\eta\gamma s}{m} uv^\top. \end{aligned} \tag{4}$$

Thus the additional update produced by SSAR is

$$\Delta W_{\text{SSAR}} = \alpha_m uv^\top, \qquad \alpha_m = \frac{2\eta\gamma s}{m} > 0. \tag{5}$$

For an active near-edge mode, $s^2/m = \Theta(1)$. Thus, when η and γ are independent of the matrix width, $\alpha_m = \Theta(m^{-1/2})$.

We analyze the change in the fixed-target Bellman loss caused by this additional spectral component,

$$\Delta L_{\text{SSAR}} = L_0(W + \Delta W_{\text{SSAR}}) - L_0(W). \tag{6}$$

Corollary 4.1 shows that modes close to the upper spectral edge can correspond to different stages of the BBP transition. In the far-subcritical regime, the task alignment is only $O_p(m^{-1/2})$; in the non-degenerate transition regime it increases to $\Theta_p(m^{-1/6})$; and for a fixed supercritical spike it becomes order one. The following theorem shows that these different alignment scales lead to qualitatively different effects on the Bellman loss.

Theorem 5.1. *Under mild regularity conditions on the critic, the input features, and the task-relevant signal, formally stated in Appendix B, we write the Bellman-loss change as* $\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m$*, where* $\mathcal{D}_m \geq 0$ *denotes the task-aligned descent contribution and* $\mathcal{E}_m$ *collects the task-orthogonal and higher-order terms.*

Subcritical regime. *If the leading signal remains a fixed distance below the BBP threshold, then* $\tau_m = O_p(m^{-1/2})$ *and* $\mathcal{D}_m = O_p(m^{-1})$*,* $\mathcal{E}_m = O_p(m^{-1})$*. Since the latter is not sign constrained, the Bellman-loss change has no deterministic sign under the stated assumptions.*

BBP transition regime. *Under a non-degenerate critical scaling for which* $\tau_m = \Theta_p(m^{-1/6})$*, we have* $\mathcal{D}_m = \Theta_p(m^{-2/3})$ *and* $\mathcal{E}_m = O_p(m^{-1}) = o_p(\mathcal{D}_m)$*, so* $\Delta L_{\mathrm{SSAR}} < 0$ *with probability tending to one.*

Supercritical regime. *For a fixed supercritical signal,* $\tau_m \xrightarrow{p} \tau_\infty > 0$*,* $\mathcal{D}_m = \Theta_p(m^{-1/2})$*, and* $\mathcal{E}_m = O_p(m^{-1}) = o_p(\mathcal{D}_m)$*, so* $\Delta L_{\mathrm{SSAR}} < 0$ *with probability tending to one.*

A complete proof of Theorem 5.1 is provided in Appendix B.

Theorem 5.1 explains why SSAR operates on a spectral window rather than requiring an explicit classifier for the BBP stage. A far-subcritical mode and a mode inside the BBP transition can both satisfy $\lambda_1 = \lambda_+ + O_p(m^{-2/3})$ and therefore lie inside $I = [\lambda_+, \lambda_+ + T]$, but their effects on critic optimization differ. In the subcritical regime, task alignment remains on the same $m^{-1/2}$ scale as the orthogonal fluctuation, so the SSAR perturbation has only an $O_p(m^{-1})$ Bellman effect with no fixed sign. Near the non-degenerate transition, the $m^{-1/6}$ alignment produces an $m^{-2/3}$ task-aligned descent term that dominates the $O_p(m^{-1})$ remainder.

Thus, SSAR does not need to infer the latent BBP regime from the observed eigenvalue alone. It can act on unresolved near-edge modes: the update has no systematic leading-order advantage while a mode remains genuinely subcritical, but becomes reliably loss-decreasing once sufficient task alignment emerges. As separated modes move beyond $\lambda_+ + T$, the stopping rule counts them through N_{sep} and removes the additional pressure once $N_{\mathrm{sep}} \geq N_c$.

6 Experiments

In this section, we evaluate SSAR in two stages. First, we use a controlled spiked-model setting to directly test the finite-width predictions of Theorem 5.1. We then evaluate whether the same critic-side spectral intervention improves learning in full reinforcement-learning environments using SAC on the DeepMind Control Suite.

6.1 Controlled Verification of the Bellman-Loss Theory

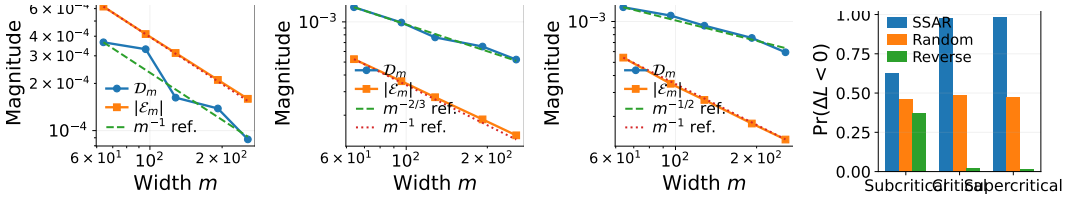
Before evaluating SSAR in full reinforcement-learning environments, we first construct a controlled setting to examine the predictions of Theorem 5.1. This controlled setting isolates the local critic update analyzed by the theorem and allows its constituent terms to be measured directly.

We consider the linear critic $Q_W(h) = o^\top W h$ with the rank-one spiked weight model $\widetilde{W} = W/\sqrt{m} = Z + \theta u^\star v^{\star\top}$, where $Z_{ij} \sim \mathcal{N}(0, \sigma^2/m)$. To satisfy the input conditions used in Appendix C, we generate $h = zv^\star + \xi_\perp/\sqrt{m}$, where $z \in \{-1, +1\}$ is a Rademacher variable and $\xi_\perp$ is a Rademacher noise vector projected onto the orthogonal complement of $v^\star$. The full construction and parameter settings are provided in Appendix C.

Using the same setting, we perform two tests. We first examine the finite-width scaling of the task-aligned contribution $\mathcal{D}_m$ and the task-orthogonal fluctuation $\mathcal{E}_m$ in $\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m$. We then examine whether the resulting scale separation leads to a reliable decrease of the fixed-target critic loss.

6.1.1 Finite-Width Scaling

We vary $m \in \{64, 96, 128, 192, 256\}$ and report medians over 100 independent trials for each spectral regime. Figure 1(a)–(c) shows the results.



(a) Subcritical. (b) Critical. (c) Supercritical. (d) Sign transition.

Figure 1: **Controlled verification of the Bellman-loss theory.** (a)–(c) Finite-width scaling of the task-aligned contribution $\mathcal{D}_m$ and task-orthogonal fluctuation $|\mathcal{E}_m|$. The two remain comparable below the transition, while $|\mathcal{E}_m|$ decreases faster at and beyond it. (d) Probability that the additional rank-one update decreases the fixed-target critic loss.

In the subcritical regime, the fitted slopes of $\mathcal{D}_m$ and $|\mathcal{E}_m|$ are -1.058 and -0.976, respectively, showing that the two terms remain at comparable scales. In the critical regime, the slopes are -0.638 and -0.948, close to the predicted $m^{-2/3}$ and $O(m^{-1})$ rates. Thus, the task-orthogonal fluctuation decreases faster than the task-aligned contribution, with $|\mathcal{E}_m|/\mathcal{D}_m$ decreasing approximately as $m^{-1/3}$.

In the fixed-supercritical regime, the corresponding slopes are -0.547 and -0.997, consistent with the predicted $m^{-1/2}$ and $O(m^{-1})$ rates. In this case, $|\mathcal{E}_m|/\mathcal{D}_m$ decreases approximately as $m^{-1/2}$. These results show that the two contributions remain comparable below the transition, whereas the task-aligned contribution increasingly dominates at and beyond it.

6.1.2 Bellman-Loss Sign Transition

We next fix $m = 256$ and perform 200 independent trials in each regime. Figure 1(d) reports the probability that the additional update satisfies $\Delta L < 0$.

For SSAR, the loss-decrease probability increases from 62.5% in the subcritical regime to 98.0% and 98.5% in the critical and supercritical regimes. The subcritical result does not imply a 50% baseline; rather, the theory states that the task-aligned contribution and task-orthogonal fluctuation remain comparable, so no consistently negative loss change is guaranteed.

As controls, a matched Random-right update remains near chance level across all three regimes, whereas reversing the SSAR direction largely reverses the effect. This indicates that the transition is associated with both the spectral direction selected by SSAR and the direction in which that mode is accelerated.

Additional diagnostics are provided in Appendix C, including the full distribution of ΔL_{SSAR}, Random-right, Reverse, and Oracle direction controls, direction controls across all three regimes, and the loss-decrease probability as a function of width.

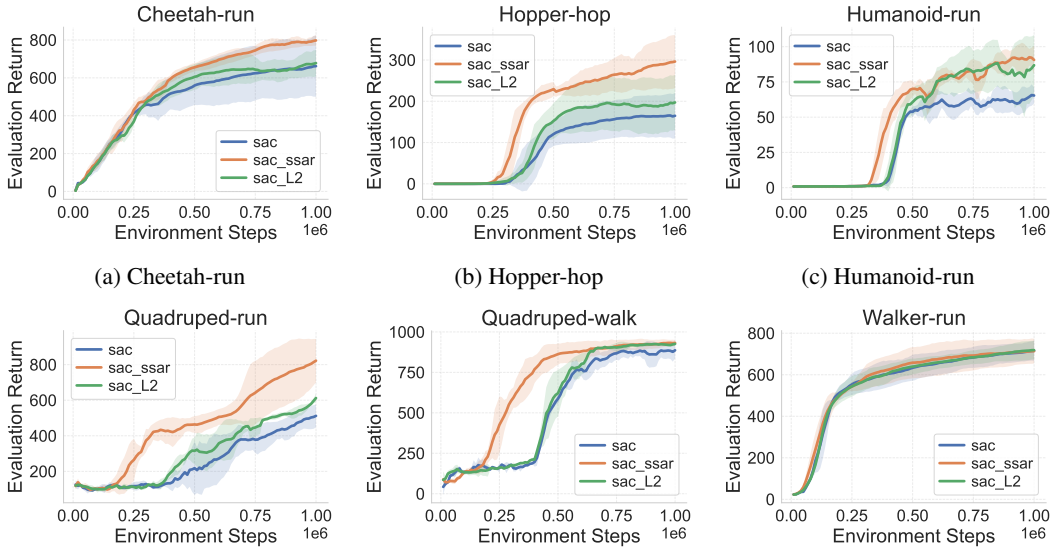
Together, the two experiments verify both the finite-width scaling predicted by Theorem 5.1 and its consequence for the Bellman-loss sign. We next examine whether this mechanism translates into improved performance in full reinforcement-learning environments.

6.2 Reinforcement Learning Experiments

6.2.1 Experimental Setup

We evaluate on six continuous-control tasks from the DeepMind Control Suite (DMC) Tassa et al. (2018); Tunyasuvunakool et al. (2020): Cheetah-run, Hopper-hop, Humanoid-run, Quadruped-run, Quadruped-walk, and Walker-run. We use state observations and Soft Actor-Critic (SAC) Haarnoja et al. (2018) as the base algorithm. Both the actor and twin critics use two 1024-unit hidden layers with ReLU activations.

SSAR is applied only to the penultimate linear layer of each critic. We compare vanilla SAC, SAC with matched L_2 regularization applied to the same critic layers, and SAC with SSAR. All methods use the same network architecture, initialization, optimization schedule, and training budget. Each agent is trained for 10^6 environment steps and evaluated every 10,000 steps over 30 deterministic



(a) Cheetah-run (b) Hopper-hop (c) Humanoid-run

(d) Quadruped-run (e) Quadruped-walk (f) Walker-run

Figure 2: **Performance on DeepMind Control Suite tasks.** Evaluation returns of SAC, SAC with matched L_2 critic regularization, and SAC with SSAR. Curves show the mean over five random seeds and shaded regions denote one standard deviation.

episodes. We run each method with five random seeds. In the learning curves, the solid line denotes the mean across seeds and the shaded region denotes one standard deviation. Full implementation details are provided in Appendix D.

6.2.2 Performance on Continuous-Control Tasks

Figure 2 shows the learning curves on the six DMC tasks. The results show three main patterns.

The clearest improvements appear on Cheetah-run, Hopper-hop, and Quadruped-run. On these tasks, SSAR learns faster than both vanilla SAC and SAC+L_2, and the advantage remains visible near the end of training. The difference is particularly clear on Quadruped-run, where SSAR reaches high returns substantially earlier and maintains a large gap over the two baselines. Hopper-hop shows a similar trend, while Cheetah-run exhibits a smaller but persistent improvement.

A second pattern appears on Humanoid-run and Quadruped-walk. SSAR improves learning mainly during the early and intermediate stages. On Quadruped-walk, the baselines gradually close much of the gap later in training, while on Humanoid-run part of the advantage remains near the end. Walker-run shows relatively small differences among the three methods, with all methods reaching similar performance by the end of training.

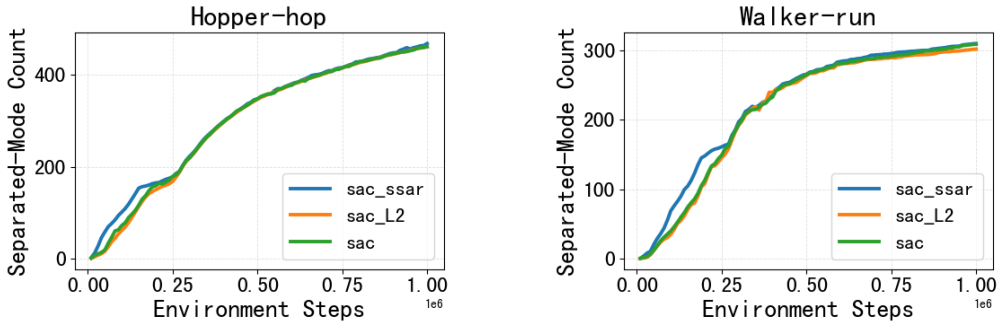
Overall, the learning curves indicate that SSAR mainly improves sample efficiency. On Cheetah-run, Hopper-hop, and Quadruped-run, this earlier improvement is maintained throughout training. On Quadruped-walk and Humanoid-run, the gain is more pronounced during the earlier stages, while Walker-run shows little difference among the methods.

To compare performance near the end of training more directly, Table 1 reports the average evaluation return over the final 100k environment steps. For each seed, we first average the final ten evaluation checkpoints and then compute the mean and standard deviation over the five seeds.

The final-window results follow the same overall trend. SSAR achieves the largest gains on Cheetah-run, Hopper-hop, and Quadruped-run, and also improves the final return on Humanoid-run. On Quadruped-walk the final difference is small, while on Walker-run the three methods remain nearly identical.

Table 1: Average evaluation return over the final 100k environment steps (mean $\pm$ standard deviation over five seeds).

Task	SAC	SAC+L_2	SAC+SSAR
Cheetah-run	654.8 $\pm$ 154.6	672.8 $\pm$ 66.8	793.0 $\pm$ 24.2
Hopper-hop	164.3 $\pm$ 53.3	193.4 $\pm$ 65.6	275.9 $\pm$ 53.3
Humanoid-run	63.8 $\pm$ 7.5	83.7 $\pm$ 22.7	91.3 $\pm$ 10.3
Quadruped-run	501.9 $\pm$ 57.5	581.5 $\pm$ 4.9	803.5 $\pm$ 140.3
Quadruped-walk	884.4 $\pm$ 49.6	924.4 $\pm$ 8.8	931.4 $\pm$ 21.2
Walker-run	711.4 $\pm$ 45.8	716.4 $\pm$ 45.5	712.0 $\pm$ 59.6



(a) Hopper-hop

(b) Walker-run

Figure 3: **Separated spectral modes during training.** Number of normalized squared singular values above the upper boundary $\lambda_+ + T$ of the SSAR acceleration window, averaged across the twin critics. SSAR advances spectral separation during early training in both tasks, while the trajectories become similar later.

6.2.3 Spectral Modes Beyond the Acceleration Window

To examine whether SSAR changes the spectral dynamics as intended, we track the number of singular modes that move beyond the upper boundary of the acceleration window. For critic j, define $N_{\text{sep}}^{(j)}(t) = \#\{k : s_{k,j}(t)^2/m > \lambda_+ + T\}$; for the twin critics we report the average $N_{\text{sep}}(t) = \big(N_{\text{sep}}^{(1)}(t) + N_{\text{sep}}^{(2)}(t)\big)/2$.

Figure 3 reports this statistic on Hopper-hop and Walker-run. In both tasks, SAC+SSAR produces a faster increase in separated modes during the early stage of training, while the curves gradually approach one another later. This behavior is consistent with the intended role of SSAR: it accelerates near-edge modes early rather than maintaining a permanently larger separated-mode count. Hopper-hop also exhibits a clear return improvement under SSAR, whereas Walker-run shows little performance difference among the three methods. Thus, earlier spectral separation is a direct diagnostic of the intervention, but it does not by itself guarantee a return improvement when critic spectral development is not the dominant learning bottleneck.

A sensitivity analysis of the SSAR coefficient is provided in Appendix E.

7 Conclusion

We introduced Spike Spectral Acceleration Regularization (SSAR), an RMT-guided critic regularizer that accelerates near-edge singular modes. Our analysis shows that the additional update is not guaranteed to improve the loss in the subcritical regime, but becomes loss-decreasing with probability tending to one after sufficient task alignment emerges. Controlled experiments verify the predicted transition, and SAC experiments on six continuous-control tasks demonstrate the effectiveness of SSAR.

AI USE STATEMENT

Large language models were used to assist with writing and presentation, literature retrieval and discovery, and research ideation and execution. All technical claims, experimental results, citations, and final content were reviewed and verified by the authors. The authors take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

Complete assumptions and proofs for the theoretical results are provided in the appendix. Appendix C describes the controlled spiked-model construction, parameters, and additional experiments. Appendix D provides the network architecture, training hyperparameters, SSAR and baseline configurations, random seeds, evaluation protocol, and spectral diagnostic used in the reinforcement-learning experiments. Sensitivity to the SSAR coefficient is reported in Appendix E.

REFERENCES

Oron Anschel, Nir Baram, and Nahum Shimkin. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 176–185. PMLR, 2017.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. 2005.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021. doi: 10.1287/opre.2020.2024.

Johan Bjorck, Carla P. Gomes, and Kilian Q. Weinberger. Towards deeper deep reinforcement learning with spectral normalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8242–8255, 2021.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith W. Ross. Randomized ensembled double q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td(0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 6144–6160, 2018. doi: 10.1609/aaai.v32i1.12079.

Xiucai Ding. High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli*, 26(1):387–417, 2020.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G. Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *International Conference on Learning Representations*, 2023.

Mattie Fellows, Matthew J. A. Smith, and Shimon Whiteson. Why target networks stabilise temporal difference methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9886–9909. PMLR, 2023.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596. PMLR, 2018.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021. doi: 10.1214/20-AOS1990.

Florin Gogianu, Tudor Berariu, Mihaela C. Rosca, Claudia Clopath, Lucian Busoniu, and Razvan Pascanu. Spectral normalisation for deep reinforcement learning: An optimisation perspective. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3734–3744. PMLR, 2021.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Hado Hasselt. Double q-learning. *Advances in neural information processing systems*, 23, 2010.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2021.

Aviral Kumar, Rishabh Agarwal, Tengyu Ma, Aaron Courville, George Tucker, and Sergey Levine. DR3: Value-based deep reinforcement learning requires explicit regularization. In *International Conference on Learning Representations*, 2022.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International conference on machine learning*, pp. 6131–6141. PMLR, 2021.

Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, András György, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C. Machado. Learning continually by spectral regularization. In *International Conference on Learning Representations*, 2025.

Zhenyu Liao and Romain Couillet. On the spectrum of random features maps of high dimensional data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3063–3071. PMLR, 2018.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 504–513. AUAI Press, 2015.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018. doi: 10.1214/17-AAP1328.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations*, 2022.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211. PMLR, 2023.

Michael W. Mahoney and Charles H. Martin. Traditional and heavy tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4284–4293. PMLR, 2019.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022. doi: 10.1214/22-AOS2211.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16828–16847. PMLR, 2022.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2638–2647, 2017.

Amartya Sanyal, Philip H. S. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2020.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32145–32168. PMLR, 2023.

Hongyao Tang and Glen Berseth. Improving deep reinforcement learning by reducing the chain effect of value and policy churn. *Advances in Neural Information Processing Systems*, 37:15320–15355, 2024.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. doi: 10.1016/j.simpa.2020.100022.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12621–12631. PMLR, 2021.

A Proof of the Spectral-Regime Corollary

We provide the derivation of Corollary 4.1. The eigenvalue fluctuation scales follow from the BBP transition Baik et al. (2005) and its rectangular finite-rank extensions Benaych-Georges & Nadakuditi (2012); Ding (2020), while the singular-vector behavior follows from the corresponding results for low-rank rectangular perturbations Benaych-Georges & Nadakuditi (2012); Ding (2020).

A.1 From Distributional Limits to O_p Rates

We first recall a standard consequence of convergence in distribution.

Lemma A.1 (Distributional convergence implies O_p)**.** *Let X_m be random variables, and let $a_m > 0$ and $b_m \in \mathbb{R}$ be deterministic sequences. If*

$$a_m(X_m - b_m) \xrightarrow{d} Y, \tag{7}$$

where Y is a non-degenerate random variable, then

$$X_m - b_m = O_p(a_m^{-1}). \tag{8}$$

Proof. Fix $\varepsilon > 0$. Since Y is a proper random variable, there exists $K > 0$ such that

$$\mathbb{P}(|Y| > K) < \frac{\varepsilon}{2}. \tag{9}$$

By convergence in distribution, for all sufficiently large m,

$$\mathbb{P}\left(|a_m(X_m - b_m)| > K\right) < \varepsilon. \tag{10}$$

Equivalently,

$$\mathbb{P}\left(|X_m - b_m| > \frac{K}{a_m}\right) < \varepsilon. \tag{11}$$

Hence

$$X_m - b_m = O_p(a_m^{-1}), \tag{12}$$

which completes the proof. □

A.2 Largest-Eigenvalue Fluctuation

At the BBP transition, Theorem 3.2 gives a non-degenerate limiting distribution under the $m^{2/3}$ edge scaling Baik et al. (2005). In particular,

$$A_1\left(\lambda_1 - \lambda_+\right) \xrightarrow{d} F_k, \tag{13}$$

where

$$A_1 = \Theta(m^{2/3}). \tag{14}$$

Applying Lemma A.1 gives

$$\lambda_1 - \lambda_+ = O_p(m^{-2/3}), \tag{15}$$

and therefore

$$\lambda_1 = \lambda_+ + O_p(m^{-2/3}). \tag{16}$$

For the rectangular additive deformation considered in our method, the same edge scale is obtained for non-outlier modes. In particular, Eq. (1.14) and Theorem 3.4 of Ding (2020) show that a mode below the outlier transition remains at the upper spectral edge on the $m^{-2/3}$ scale, up to the arbitrarily small polynomial slack in the general high-probability statement.

For a fixed supercritical signal, Theorem 3.2 gives the $m^{-1/2}$ fluctuation scale in the classical spiked covariance model Baik et al. (2005). For rectangular finite-rank perturbations, Theorem 2.8

of Benaych-Georges & Nadakuditi (2012) establishes the emergence of a separated singular-value outlier. The corresponding local rate is given by Eq. (1.13) and Theorem 3.4 of Ding (2020):

$$\lambda_1 = f(\theta) + O_p(m^{-1/2}), \qquad f(\theta) > \lambda_+, \tag{17}$$

when the signal remains a fixed positive distance above the critical threshold. Thus, the $m^{-2/3}$ and $m^{-1/2}$ scales used in Corollary 4.1 are respectively the edge and fixed-supercritical fluctuation scales of the rectangular spiked model Benaych-Georges & Nadakuditi (2012); Ding (2020).

A.3 Singular-Vector Alignment

Recall

$$\tau_m = |\langle v_1, v^\star \rangle|, \tag{18}$$

where v_1 is the leading empirical right singular vector and $v^\star$ is the population signal direction.

Subcritical regime. Suppose

$$\theta \le \theta_c - \varepsilon \tag{19}$$

for some fixed $\varepsilon > 0$. Below the BBP threshold, the empirical singular vector does not acquire a non-vanishing projection onto the population signal direction; this singular-vector phase transition is characterized in Theorem 2.10 of Benaych-Georges & Nadakuditi (2012).

More quantitatively, Eq. (1.18) and the non-outlier part of Theorem 3.5 in Ding (2020) give the squared-overlap scale

$$\tau_m^2 = O_p(m^{-1}) \tag{20}$$

for a signal below the transition scale. Taking square roots yields

$$\tau_m = O_p(m^{-1/2}). \tag{21}$$

BBP transition regime. The characteristic BBP transition window for the rectangular model is

$$|\theta - \theta_c| = O(m^{-1/3}), \tag{22}$$

as described in Eq. (1.12) of Ding (2020). At this scale, the corresponding eigenvalue displacement is of order $m^{-2/3}$, comparable to the natural edge fluctuation; see Eqs. (1.13)–(1.14) and Theorem 3.4 of Ding (2020).

The singular-vector scale can be obtained from the overlap formula in Ding (2020). For the right singular vector, Eq. (2.12) defines

$$a_2(d) = \frac{d^4 - c^{-1}}{d^2(d^2+1)}, \tag{23}$$

where the critical signal strength in the notation of Ding (2020) is $d_c = c^{-1/4}$. Theorem 3.5 and Eq. (1.17) show that, on the supercritical side, the empirical squared overlap is controlled by this deterministic quantity.

Since

$$a_2(d_c) = 0 \tag{24}$$

and the numerator satisfies

$$d^4 - d_c^4 = 4d_c^3(d - d_c) + O((d - d_c)^2), \tag{25}$$

while the denominator remains non-zero at d_c, we have

$$a_2(d) = O(d - d_c) \tag{26}$$

near the transition. Therefore, when

$$d - d_c = O(m^{-1/3}), \tag{27}$$

the characteristic squared-overlap scale is

$$\tau_m^2 = O_p(m^{-1/3}), \tag{28}$$

consistent with the local singular-vector estimates in Theorem 3.5 of Ding (2020). On the subcritical side of the same window, the non-outlier estimate is no larger than this scale Ding (2020). Hence,

$$\tau_m = O_p(m^{-1/6}). \tag{29}$$

Here we suppress the arbitrarily small polynomial slack appearing in the general high-probability bounds of Ding (2020).

For the Bellman analysis in Theorem 5.1, we consider a non-degenerate critical regime in which the overlap actually attains this characteristic scale,

$$\tau_m = \Theta_p(m^{-1/6}). \tag{30}$$

This last relation is an additional non-degeneracy condition used in our Bellman-loss analysis; it is stronger than the upper-scale result required for Corollary 4.1.

Supercritical regime. For

$$\theta \geq \theta_c + \varepsilon, \tag{31}$$

the empirical leading singular vector develops a non-zero asymptotic projection onto the population signal direction. This singular-vector phase transition is established in Theorems 2.9–2.10 of Benaych-Georges & Nadakuditi (2012).

For the rectangular signal-plus-noise model, Eq. (1.17), Eq. (2.12), and Theorem 3.5 of Ding (2020) give the corresponding deterministic overlap. For a signal separated from the threshold by a fixed positive amount, this overlap is strictly positive. Consequently,

$$\tau_m \xrightarrow{p} \tau_\infty(\theta) > 0. \tag{32}$$

Combining equation 16, equation 17, equation 21, equation 29, and equation 32 gives the three regimes stated in Corollary 4.1.

B Bellman-Loss Analysis of the SSAR Update

We now prove Theorem 5.1. The spectral scales used below are those established in Corollary 4.1 and justified in Appendix A. The remaining arguments are derived directly from the critic model and the assumptions stated below.

B.1 Setup and Assumptions

Consider the critic representation

$$Q_W(h) = o^\top \varphi(Wh), \tag{33}$$

where

$$W \in \mathbb{R}^{n\times m}, \qquad h \in \mathbb{R}^m, \qquad o \in \mathbb{R}^n. \tag{34}$$

The target y is treated as fixed during the critic update, and we write

$$r(h) = Q_W(h) - y. \tag{35}$$

The corresponding Bellman loss is

$$L_0(W) = \frac{1}{2}\mathbb{E}[r(h)^2]. \tag{36}$$

Let (s, u, v) denote the singular triplet of the active mode,

$$Wv = su, \qquad W^\top u = sv, \qquad \|u\|_2 = \|v\|_2 = 1. \tag{37}$$

The analysis uses the following local assumptions.

Assumption 1 (Spiked spectral model)**.** *The normalized weight matrix admits the local rank-one representation*

$$\frac{W}{\sqrt{m}} = Z + \theta u^\star v^{\star\top}, \tag{38}$$

where $u^\star$ and $v^\star$ are unit population signal directions. The aspect ratio satisfies

$$\frac{n}{m} \to c \in (0, \infty). \tag{39}$$

Assumption 1 is the same spiked model used in Corollary 4.1. The corresponding spectral behavior is supported by the rectangular finite-rank perturbation results reviewed in Appendix A.

Assumption 2 (Input geometry)**.** *There exist constants $C_h, \sigma_h < \infty$, independent of m, such that*

$$\|h\|_2 \le C_h \tag{40}$$

almost surely, and for every unit vector a orthogonal to $v^\star$,

$$\mathbb{E}\left[(a^\top h)^2 \mid W\right] \le \frac{\sigma_h^2}{m}. \tag{41}$$

Discussion of Assumption 2. Assumption 2 combines two standard regularity conditions from related reinforcement-learning and high-dimensional learning analyses. Uniform boundedness, or bounded second moments, of feature vectors is commonly used in finite-sample analyses of temporal-difference methods Dalal et al. (2018); Liu et al. (2015); Bhandari et al. (2021). Such conditions prevent individual samples from producing arbitrarily large updates and provide uniform control of the stochastic terms appearing in the analysis.

The second condition is a high-dimensional isotropy-type assumption. Let $P_\perp = I - v^\star v^{\star\top}$. The directional condition in Assumption 2 is equivalently expressed as

$$\left\|P_\perp \,\mathbb{E}[hh^\top \mid W]\, P_\perp\right\|_{\mathrm{op}} \le \frac{\sigma_h^2}{m}. \tag{42}$$

Thus, the feature energy orthogonal to the task direction is distributed across the high-dimensional feature space rather than concentrated in a single direction. The same m^{-1} directional-energy scaling arises in standard spherical or isotropic input models used in high-dimensional analyses of random features and neural networks Ghorbani et al. (2021); Mei & Montanari (2022); Montanari & Zhong (2022). Closely related random-matrix models also use high-dimensional covariance scalings of order C/m Liao & Couillet (2018).

We do not require the critic features to be exactly spherical or Gaussian. Assumption 2 retains only the boundedness and task-orthogonal directional-energy scaling needed by the local Bellman-loss analysis. The controlled construction in Appendix C satisfies these conditions explicitly.

Assumption 3 (Local critic regularity)**.** *There exist constants $C_o, C_r, L_1, L_2 < \infty$, independent of m, such that*

$$\|o\|_2 \le C_o, \qquad |r(h)| \le C_r, \tag{43}$$

and

$$|\varphi'(x)| \le L_1, \qquad |\varphi''(x)| \le L_2. \tag{44}$$

The active singular value s is simple.

Assumption 4 (Task-relevant signal direction)**.** *There exists a constant $c_0 > 0$, independent of m, such that*

$$-\mathbb{E}\left[r(h)\left(o^\top \operatorname{Diag}(\varphi'(Wh))u\right)(v^{\star\top}h)\right] \ge c_0. \tag{45}$$

Assumption 4 states that moving the right singular direction toward the population signal direction is locally useful for the Bellman objective. It is the only assumption used to determine the sign of the task-aligned contribution.

B.2 The SSAR Update Direction

We first verify the singular-value derivative used by SSAR.

Lemma B.1 (Derivative of a simple singular value)**.** *Let s be a simple singular value of W with left and right singular vectors u and v. For an arbitrary perturbation H,*

$$\frac{d}{d\epsilon} s(W + \epsilon H)\Big|_{\epsilon=0} = u^\top H v. \tag{46}$$

Consequently,

$$\nabla_W s = uv^\top. \tag{47}$$

Proof. Differentiate

$$Wv = su \tag{48}$$

in the direction H. At $\epsilon = 0$,

$$Hv + W\dot{v} = \dot{s}\,u + s\dot{u}. \tag{49}$$

Multiplying by $u^\top$ gives

$$u^\top H v + u^\top W \dot{v} = \dot{s} + s\,u^\top \dot{u}. \tag{50}$$

Since $W^\top u = sv$,

$$u^\top W \dot{v} = s\,v^\top \dot{v}. \tag{51}$$

The normalizations $\|u\|_2 = \|v\|_2 = 1$ imply

$$u^\top \dot{u} = 0, \qquad v^\top \dot{v} = 0. \tag{52}$$

Hence

$$\dot{s} = u^\top H v. \tag{53}$$

Using the Frobenius inner product,

$$u^\top H v = \langle uv^\top, H \rangle_{\mathrm{F}}, \tag{54}$$

which gives

$$\nabla_W s = uv^\top. \tag{55}$$

□

For one active mode,

$$L_{\mathrm{spike}} = \left(\lambda_+ + T - \frac{s^2}{m}\right)_+. \tag{56}$$

Inside the active region, Lemma B.1 gives

$$\nabla_W L_{\mathrm{spike}} = -\frac{2s}{m} uv^\top. \tag{57}$$

Therefore, one gradient step on

$$L_{\text{tot}} = L_0 + \gamma L_{\text{spike}} \tag{58}$$

contains the additional SSAR displacement

$$\Delta W_{\text{SSAR}} = \alpha_m u v^\top, \tag{59}$$

where

$$\alpha_m = \frac{2\eta\gamma s}{m}. \tag{60}$$

For an active near-edge mode,

$$\frac{s^2}{m} \in [\lambda_+, \lambda_+ + T], \qquad T = cm^{-2/3}. \tag{61}$$

Since the MP edge λ_+ is a positive constant in the asymptotic regime considered here,

$$\frac{s^2}{m} = \Theta(1), \tag{62}$$

and therefore

$$s = \Theta(\sqrt{m}). \tag{63}$$

When η and γ do not scale with m,

$$\alpha_m = \Theta(m^{-1/2}). \tag{64}$$

B.3 Directional Bellman Gain

Choose the sign of v such that

$$v^\top v^\star = \tau_m \geq 0. \tag{65}$$

Then

$$v = \tau_m v^\star + \sqrt{1 - \tau_m^2}\, v_\perp, \tag{66}$$

where

$$v_\perp^\top v^\star = 0, \qquad \|v_\perp\|_2 = 1. \tag{67}$$

Define

$$S(h) = o^\top \operatorname{Diag}(\varphi'(Wh))u. \tag{68}$$

Lemma B.2 (Directional derivative of the critic)**.** *For the direction $uv^\top$,*

$$\frac{d}{d\alpha} Q_{W+\alpha uv^\top}(h)\Big|_{\alpha=0} = S(h)v^\top h. \tag{69}$$

Proof. Since

$$(W + \alpha uv^\top)h = Wh + \alpha u(v^\top h), \tag{70}$$

the chain rule gives

$$\frac{d}{d\alpha} Q_{W+\alpha uv^\top}(h)\Big|_{\alpha=0} = o^\top \operatorname{Diag}(\varphi'(Wh))u(v^\top h) \tag{71}$$

$$= S(h)v^\top h. \tag{72}$$

□

We define the Bellman gain along the spectral direction as

$$G_m = -\left\langle \nabla_W L_0(W), uv^\top \right\rangle_{\text{F}}. \tag{73}$$

Using Lemma B.2,

$$G_m = -\mathbb{E}\left[r(h)S(h)v^\top h\right]. \tag{74}$$

Substituting equation 66 yields

$$G_m = \tau_m A_m + R_m, \tag{75}$$

where

$$A_m = -\mathbb{E}\left[r(h)S(h)(v^{\star\top}h)\right] \tag{76}$$

and

$$R_m = -\sqrt{1-\tau_m^2}\,\mathbb{E}\left[r(h)S(h)(v_\perp^\top h)\right]. \tag{77}$$

Assumption 4 gives

$$A_m \geq c_0. \tag{78}$$

On the other hand,

$$|S(h)| \leq \|o\|_2 \left\|\mathrm{Diag}(\varphi'(Wh))\right\|_{\mathrm{op}} \|u\|_2 \leq C_o L_1. \tag{79}$$

Together with $\|h\|_2 \leq C_h$,

$$|A_m| \leq \mathbb{E}\left[|r(h)S(h)(v^{\star\top}h)|\right] \tag{80}$$

$$\leq C_r C_o L_1 C_h. \tag{81}$$

Hence

$$A_m = \Theta(1). \tag{82}$$

B.4 Task-Orthogonal Contribution

Lemma B.3 (Task-orthogonal contribution)**.** *Under Assumptions 2 and 3,*

$$R_m = O_p(m^{-1/2}). \tag{83}$$

Proof. From equation 77,

$$|R_m| \leq \mathbb{E}\left[|r(h)S(h)(v_\perp^\top h)| \mid W\right]. \tag{84}$$

Applying Cauchy–Schwarz,

$$|R_m| \leq \left(\mathbb{E}[r(h)^2 S(h)^2 \mid W]\right)^{1/2} \left(\mathbb{E}[(v_\perp^\top h)^2 \mid W]\right)^{1/2}. \tag{85}$$

By Assumption 3,

$$|r(h)S(h)| \leq C_r C_o L_1. \tag{86}$$

Since $v_\perp$ is a unit vector orthogonal to $v^\star$, Assumption 2 gives

$$\mathbb{E}[(v_\perp^\top h)^2 \mid W] \leq \frac{\sigma_h^2}{m}. \tag{87}$$

Therefore

$$|R_m| \leq \frac{C_r C_o L_1 \sigma_h}{\sqrt{m}}, \tag{88}$$

which proves

$$R_m = O_p(m^{-1/2}). \tag{89}$$

□

B.5 Local Change of the Bellman Loss

We next control the finite SSAR step rather than only its directional derivative.

Lemma B.4 (Local Bellman expansion)**.** *There exist constants $\alpha_0 > 0$ and $L_B < \infty$, independent of m, such that for $|\alpha| \le \alpha_0$,*

$$L_0(W + \alpha uv^\top) - L_0(W) = -\alpha G_m + \xi_m(\alpha), \tag{90}$$

where

$$|\xi_m(\alpha)| \le \frac{L_B}{2}\alpha^2. \tag{91}$$

Proof. Let

$$b(\alpha) = L_0(W + \alpha uv^\top), \tag{92}$$

and define

$$q_\alpha(h) = Q_{W+\alpha uv^\top}(h). \tag{93}$$

From Lemma B.2,

$$b'(0) = -G_m. \tag{94}$$

For general α,

$$q'_\alpha(h) = (v^\top h)\, o^\top \operatorname{Diag}\left(\varphi'((W + \alpha uv^\top)h)\right) u. \tag{95}$$

Using Assumptions 2 and 3,

$$|q'_\alpha(h)| \le C_o L_1 C_h. \tag{96}$$

Differentiating once more,

$$q''_\alpha(h) = (v^\top h)^2 \sum_{j=1}^{n} o_j u_j^2 \varphi''\left(((W + \alpha uv^\top)h)_j\right). \tag{97}$$

Since

$$\sum_{j=1}^{n} |o_j| u_j^2 \le \|o\|_2 \left(\sum_{j=1}^{n} u_j^4\right)^{1/2} \le C_o, \tag{98}$$

we obtain

$$|q''_\alpha(h)| \le C_o L_2 C_h^2. \tag{99}$$

Furthermore, by equation 96,

$$|q_\alpha(h) - q_0(h)| \le |\alpha| C_o L_1 C_h. \tag{100}$$

Thus, for $|\alpha| \le \alpha_0$,

$$|q_\alpha(h) - y| \le C_r + \alpha_0 C_o L_1 C_h. \tag{101}$$

Differentiating the Bellman loss twice gives

$$b''(\alpha) = \mathbb{E}\left[(q'_\alpha(h))^2 + (q_\alpha(h) - y) q''_\alpha(h)\right]. \tag{102}$$

Equations equation 96 and equation 99 imply that there exists a constant L_B, independent of m, such that

$$|b''(\alpha)| \le L_B \tag{103}$$

throughout this neighborhood.

Taylor's theorem around $\alpha = 0$ therefore gives

$$b(\alpha) = b(0) + \alpha b'(0) + \xi_m(\alpha), \tag{104}$$

with

$$|\xi_m(\alpha)| \le \frac{L_B}{2}\alpha^2. \tag{105}$$

Using $b'(0) = -G_m$ proves the result. □

B.6 Task-Aligned and Remainder Terms

Set $\alpha = \alpha_m$ in Lemma B.4. From equation 75,

$$\Delta L_{\mathrm{SSAR}} = -\alpha_m \left(\tau_m A_m + R_m\right) + \xi_m(\alpha_m) \tag{106}$$

$$= -\alpha_m \tau_m A_m + \left[-\alpha_m R_m + \xi_m(\alpha_m)\right]. \tag{107}$$

We therefore define

$$\mathcal{D}_m = \alpha_m \tau_m A_m \tag{108}$$

and

$$\mathcal{E}_m = -\alpha_m R_m + \xi_m(\alpha_m). \tag{109}$$

Then

$$\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m. \tag{110}$$

Since

$$\alpha_m > 0, \qquad \tau_m \geq 0, \qquad A_m \geq c_0, \tag{111}$$

we have

$$\mathcal{D}_m \geq 0. \tag{112}$$

The remainder $\mathcal{E}_m$ has no prescribed sign.

Using

$$\alpha_m = \Theta(m^{-1/2}), \qquad R_m = O_p(m^{-1/2}), \tag{113}$$

we have

$$\alpha_m R_m = O_p(m^{-1}). \tag{114}$$

The Taylor remainder satisfies

$$|\xi_m(\alpha_m)| \leq \frac{L_B}{2}\alpha_m^2 = O(m^{-1}). \tag{115}$$

Therefore

$$\mathcal{E}_m = O_p(m^{-1}). \tag{116}$$

B.7 A Useful Order Comparison

The following elementary fact will be used in the critical and supercritical cases.

Lemma B.5. *Let $X_m = O_p(a_m)$ and let $Y_m = \Theta_p(b_m)$ with $Y_m > 0$. If*

$$\frac{a_m}{b_m} \to 0, \tag{117}$$

then

$$\frac{X_m}{Y_m} \xrightarrow{p} 0. \tag{118}$$

Equivalently,

$$X_m = o_p(Y_m). \tag{119}$$

Proof. Since $X_m = O_p(a_m)$,

$$\frac{X_m}{a_m} = O_p(1). \tag{120}$$

Since $Y_m = \Theta_p(b_m)$ and $Y_m > 0$,

$$\frac{b_m}{Y_m} = O_p(1). \tag{121}$$

Therefore

$$\frac{X_m}{Y_m} = \frac{X_m}{a_m}\frac{a_m}{b_m}\frac{b_m}{Y_m}. \tag{122}$$

The first and third factors are bounded in probability, while the middle factor converges deterministically to zero. Hence

$$\frac{X_m}{Y_m} \xrightarrow{p} 0. \tag{123}$$

□

B.8 Proof of Theorem 5.1

Proof of Theorem 5.1. We apply the decomposition

$$\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m \tag{124}$$

to the three regimes in Corollary 4.1.

Subcritical regime. Appendix A gives

$$\tau_m = O_p(m^{-1/2}) \tag{125}$$

for a signal that remains a fixed distance below the BBP threshold.

Together with

$$\alpha_m = \Theta(m^{-1/2}) \tag{126}$$

and

$$A_m = \Theta(1), \tag{127}$$

this gives

$$\mathcal{D}_m = \alpha_m \tau_m A_m = O_p(m^{-1}). \tag{128}$$

Equation equation 116 gives

$$\mathcal{E}_m = O_p(m^{-1}). \tag{129}$$

Thus neither contribution is asymptotically dominant. Since $\mathcal{E}_m$ has no prescribed sign, the stated assumptions do not determine the sign of

$$\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m. \tag{130}$$

This proves the subcritical statement.

BBP transition regime. For the non-degenerate critical regime used in the theorem,

$$\tau_m = \Theta_p(m^{-1/6}). \tag{131}$$

As noted in Appendix A, the $m^{-1/6}$ scale is the characteristic critical overlap scale obtained from the rectangular singular-vector results of Ding (2020); the lower-order non-degeneracy required for Θ_p rather than O_p is an explicit condition of Theorem 5.1.

Using

$$\alpha_m = \Theta(m^{-1/2}) \tag{132}$$

and

$$A_m = \Theta(1), \tag{133}$$

we obtain

$$\mathcal{D}_m = \Theta_p(m^{-2/3}). \tag{134}$$

The remainder satisfies

$$\mathcal{E}_m = O_p(m^{-1}). \tag{135}$$

Since

$$\frac{m^{-1}}{m^{-2/3}} = m^{-1/3} \to 0, \tag{136}$$

Lemma B.5 gives

$$\mathcal{E}_m = o_p(\mathcal{D}_m). \tag{137}$$

Hence

$$\frac{\mathcal{E}_m}{\mathcal{D}_m} \xrightarrow{p} 0. \tag{138}$$

Because $\mathcal{D}_m > 0$ with probability tending to one,

$$\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m \tag{139}$$

$$= -\mathcal{D}_m \left(1 - \frac{\mathcal{E}_m}{\mathcal{D}_m}\right) < 0 \tag{140}$$

with probability tending to one.

This proves the critical-regime statement.

Supercritical regime. For a fixed supercritical signal, Corollary 4.1 gives

$$\tau_m \xrightarrow{p} \tau_\infty(\theta) > 0. \tag{141}$$

The origin of this non-vanishing overlap is the rectangular singular-vector phase transition established in Benaych-Georges & Nadakuditi (2012); Ding (2020), as discussed in Appendix A. Hence

$$\tau_m = \Theta_p(1). \tag{142}$$

It follows that

$$\mathcal{D}_m = \alpha_m \tau_m A_m = \Theta_p(m^{-1/2}). \tag{143}$$

Again,

$$\mathcal{E}_m = O_p(m^{-1}). \tag{144}$$

Since

$$\frac{m^{-1}}{m^{-1/2}} = m^{-1/2} \to 0, \tag{145}$$

Lemma B.5 gives

$$\mathcal{E}_m = o_p(\mathcal{D}_m). \tag{146}$$

Therefore

$$\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m < 0 \tag{147}$$

with probability tending to one.

This completes the proof. □

B.9 Relation to the Edge Window and Stopping Rule

The result also clarifies the role of the spectral window used by SSAR. Both a far-subcritical mode and a mode near the BBP transition may satisfy

$$\lambda_1 = \lambda_+ + O_p(m^{-2/3}), \tag{148}$$

as established in Appendix A. Consequently, the leading eigenvalue alone does not distinguish the two regimes within the window

$$I = [\lambda_+, \lambda_+ + cm^{-2/3}]. \tag{149}$$

Their Bellman effects are nevertheless different. In the subcritical regime,

$$\mathcal{D}_m = O_p(m^{-1}), \qquad \mathcal{E}_m = O_p(m^{-1}), \tag{150}$$

so no sign is guaranteed. In the non-degenerate transition regime,

$$\mathcal{D}_m = \Theta_p(m^{-2/3}), \qquad \mathcal{E}_m = O_p(m^{-1}) = o_p(\mathcal{D}_m), \tag{151}$$

and the task-aligned descent term dominates.

Once a mode becomes clearly separated from the spectral bulk, it no longer belongs to the near-edge window and the stopping rule removes the additional spectral acceleration. The supercritical part of Theorem 5.1 describes the Bellman geometry of the same spectral direction after separation; it does not require SSAR to continue accelerating a mode that has already satisfied the stopping criterion.

C Controlled Synthetic Experiment

C.1 Experimental Construction and Design Rationale

The controlled experiment is designed to isolate the local Bellman-loss mechanism analyzed in Theorem 5.1. In particular, we require a setting in which the task-aligned and task-orthogonal contributions can be tracked separately, while the input geometry and boundedness conditions used in the theoretical analysis are explicitly satisfied. We therefore consider a single fixed-target critic update, excluding additional effects caused by policy evolution, exploration, and non-stationary data.

For each independent trial, we generate a rectangular rank-one spiked matrix $\widetilde{W} = Z + \theta u^\star v^{\star\top} \in \mathbb{R}^{n\times m}$, where $Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2/m)$ and $u^\star$ and $v^\star$ are independently sampled unit vectors. We use $W = \sqrt{m}\,\widetilde{W}$ as the critic weight matrix, so that $s_k(W)^2/m = s_k(\widetilde{W})^2$, consistent with the normalization in our analysis. Here, $v^\star$ represents the population task direction.

Let (s, u, v) denote the leading singular triplet of W. We orient the SVD pair such that $\langle v, v^\star\rangle \geq 0$ and define $\tau = \langle v, v^\star\rangle$. The empirical right singular direction can then be decomposed as $v = \tau v^\star + \sqrt{1-\tau^2}\, v_\perp$, where $v_\perp \perp v^\star$ and $\|v_\perp\|_2 = 1$.

We use the linear critic $Q_W(h) = o^\top W h$ and set $o = u$. This critic is a linear special case of the network considered in the main text and is chosen to isolate the spectral mechanism without introducing additional effects from nonlinear activations. Since $o^\top u = 1$, a rank-one perturbation along $uv^\top$ changes the critic output according to $Q_{W+\alpha uv^\top}(h) - Q_W(h) = \alpha v^\top h$. The effect of the additional update is therefore directly determined by the empirical right singular direction studied in our analysis.

To satisfy the input assumptions used in the Bellman-loss analysis, each feature is generated as $h = zv^\star + \xi_\perp/\sqrt{m}$, where $z \in \{-1, +1\}$ is a Rademacher variable, $\xi \in \{-1, +1\}^m$ has independent Rademacher coordinates, and $\xi_\perp = \xi - v^\star(v^{\star\top}\xi)$ removes the component along the task direction. Hence $v^{\star\top} h = z$. For any fixed unit vector $a \perp v^\star$, we have $a^\top h = (a^\top \xi)/\sqrt{m}$ and therefore $\mathbb{E}[(a^\top h)^2] = 1/m$, which gives the input geometry assumed in the theorem. The construction also satisfies the bounded-input condition, since $\|h\|_2 \leq \sqrt{2}$.

A further consideration is required for the task-orthogonal component of the Bellman loss. If the residual contained only the task term $r(h) = -\rho z$, the population task-orthogonal contribution would vanish by independence. At finite batch size, only sampling fluctuations would remain, making the controlled problem substantially cleaner than the non-degenerate case allowed by the theorem. In particular, this can make the subcritical SSAR direction appear artificially reliable.

We therefore introduce a bounded task-orthogonal nuisance. Using $v_\perp$ defined above, let $g = \sqrt{m}\, v_\perp^\top h$ and $\nu(h) = \tanh(g)$. For each independent matrix trial, we draw a single Rademacher sign $\epsilon \in \{-1, +1\}$ and define $r(h) = -\rho z + \epsilon\kappa\nu(h)$. The corresponding frozen target is $y(h) = Q_W(h) - r(h)$.

This construction is chosen so that the nuisance does not systematically alter the task-aligned signal. Since z is independent of the task-orthogonal variable g, the population task coefficient satisfies $A_m = -\mathbb{E}[r(h)z] = \rho$.

At the same time, the construction produces a non-degenerate task-orthogonal term. Since $v_\perp^\top h = g/\sqrt{m}$, we obtain

$$R_m = -\epsilon \frac{\kappa\sqrt{1-\tau^2}}{\sqrt{m}} \mathbb{E}[g\tanh(g)].$$

For a non-degenerate symmetric distribution of g, $\mathbb{E}[g\tanh(g)] > 0$. Hence $|R_m| = \Theta(m^{-1/2})$ whenever $\sqrt{1-\tau^2}$ remains bounded away from zero. Sampling ϵ once per trial makes this contribution sign-indefinite across independent trials without averaging it away within a batch.

Finally, $\tanh(\cdot)$ keeps the nuisance bounded: $|\nu(h)| \leq 1$ and therefore $|r(h)| \leq \rho + \kappa$. The resulting construction therefore satisfies the theoretical input and boundedness conditions, preserves the task-aligned signal, and realizes a non-degenerate task-orthogonal fluctuation. This construction is intended as a controlled test of the assumptions and mechanism in Theorem 5.1, rather than as a model of a particular natural reinforcement-learning environment.

Table 2: Parameters used in the controlled synthetic experiments.

Parameter	Value
Aspect ratio $\beta = n/m$	0.5
Noise scale σ	1.0
Task strength ρ	1.0
Nuisance strength κ	4.0
Batch size B	256
Learning rate η	0.05
SSAR coefficient γ	0.10
Subcritical ratio	0.70
Critical offset	1.0
Supercritical ratio	1.30

C.2 SSAR Update and Experimental Details

For the leading singular triplet (s, u, v), the additional SSAR perturbation is $\Delta W_{\mathrm{SSAR}} = \alpha_m uv^\top$, where $\alpha_m = 2\eta\gamma s/m$. The empirical fixed-target critic loss is $L_0(W) = \frac{1}{2B}\sum_{i=1}^{B}(Q_W(h_i) - y_i)^2$.

For each trial, we compute $A_m = -B^{-1}\sum_i r_i(v^{\star\top}h_i)$ and $R_m = -\sqrt{1-\tau^2}\,B^{-1}\sum_i r_i(v_\perp^\top h_i)$. The directional term is $G_m = \tau A_m + R_m$, and the task-aligned contribution is $\mathcal{D}_m = \alpha_m \tau A_m$. The remaining contribution is defined as $\mathcal{E}_m = -\alpha_m R_m + \xi_m$, giving the decomposition $\Delta L_{\mathrm{SSAR}} = -\mathcal{D}_m + \mathcal{E}_m$.

For the linear critic used here, the finite-step term can be written as

$$\xi_m = \frac{\alpha_m^2}{2B}\sum_{i=1}^{B}(v^\top h_i)^2.$$

The implementation evaluates the actual loss before and after the SSAR perturbation and checks the decomposition numerically in every trial.

We use the additive-spike threshold $\theta_c = \sigma\beta^{1/4}$, where $\beta = n/m$. The three spectral regimes are instantiated as $\theta_{\mathrm{sub}} = 0.70\theta_c$, $\theta_{\mathrm{crit}} = \theta_c + \sigma m^{-1/3}$, and $\theta_{\mathrm{super}} = 1.30\theta_c$. Thus, the critical setting remains within the finite-width transition window, while the other two settings stay below and above the threshold, respectively.

The common experimental parameters are summarized in Table 2.

For the finite-width experiment, we use $m \in \{64, 96, 128, 192, 256\}$ with 100 independent trials for each width and spectral regime. For the fixed-width sign experiment, we set $m = 256$ and use 200 independent trials for each regime.

C.3 Additional Experimental Results

C.3.1 Distribution of the SSAR Loss Change

The main text reports the probability that the SSAR perturbation decreases the fixed-target critic loss. Figure 4 additionally shows the complete distribution of ΔL_{SSAR} over independent trials.

In the subcritical regime, the distribution extends across zero, showing that both loss-decreasing and loss-increasing SSAR steps occur. In contrast, the critical and supercritical distributions are concentrated predominantly below zero. The full distributions therefore provide a complementary view of the sign transition reported in the main text.

C.3.2 Direction Controls

To determine whether the observed loss reduction is specific to the spectral direction selected by SSAR, we compare four rank-one perturbations using the same step magnitude α_m and the same empirical left singular vector u.

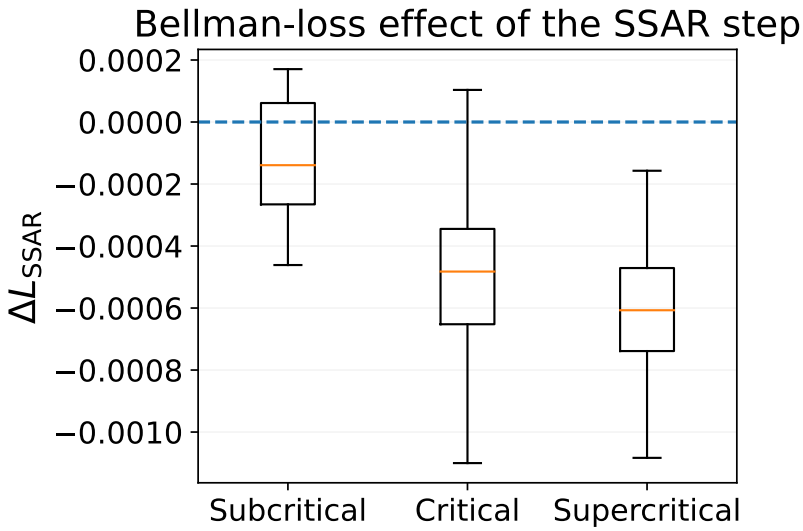


Figure 4: **Distribution of the SSAR-induced loss change.** The subcritical distribution crosses zero, whereas the critical and supercritical distributions are concentrated predominantly on negative loss changes.

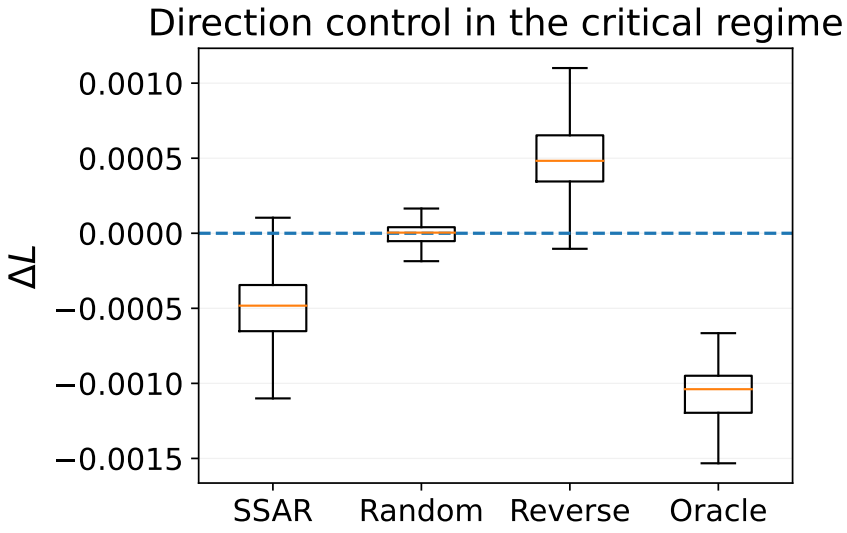


Figure 5: **Direction controls in the critical regime.** SSAR is compared with a matched random direction, the reversed SSAR direction, and an Oracle direction using the population task vector.

SSAR uses $\Delta W_{\mathrm{SSAR}} = \alpha_m u v^\top$. Random-right replaces the empirical right singular vector by an independently sampled unit vector v_{rand} and uses $\Delta W_{\mathrm{random}} = \alpha_m u v_{\mathrm{rand}}^\top$. Reverse SSAR changes only the sign of the SSAR perturbation, $\Delta W_{\mathrm{reverse}} = -\alpha_m u v^\top$. Finally, the Oracle control uses $\Delta W_{\mathrm{oracle}} = \alpha_m u v^{\star\top}$. Since the population task direction $v^\star$ is unavailable to the learner, Oracle is included only as an idealized reference for perfect task alignment.

Figure 5 compares the four directions in the critical regime. The Random-right distribution remains centered close to zero, showing that a rank-one perturbation of the same magnitude alone does not reproduce the SSAR effect. Reverse SSAR moves the distribution to the opposite side of zero, demonstrating that the sign of the spectral perturbation is essential. The Oracle direction produces the strongest negative shift because it uses the exact task direction.

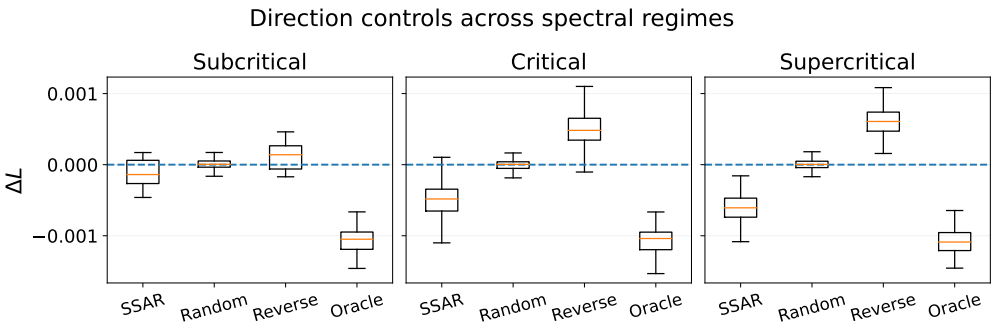


Figure 6: **Direction controls across the three spectral regimes.** Random-right provides a matched unstructured direction, Reverse changes the sign of the SSAR perturbation, and Oracle provides an idealized reference using the population task direction.

Table 3: Probability of a negative loss change for the four direction controls at $m = 256$.

Regime	SSAR	Random	Reverse	Oracle
Subcritical	0.625	0.460	0.375	1.000
Critical	0.980	0.485	0.020	1.000
Supercritical	0.985	0.475	0.015	1.000

Figure 6 extends the same comparison to all three spectral regimes. Random-right remains close to zero throughout the three settings, while Reverse SSAR produces the opposite sign behavior. Oracle remains strongly loss-decreasing because it directly uses $v^\star$. In contrast, SSAR becomes progressively more favorable from the subcritical to the critical and supercritical regimes.

For completeness, Table 3 reports the loss-decrease probability of all four directions at $m = 256$. The Oracle direction is consistently loss-decreasing in this controlled setting, whereas Random-right remains close to an unstructured baseline. Reverse SSAR increasingly fails to decrease the loss in the critical and supercritical regimes.

C.3.3 Loss-Decrease Probability Across Widths

The sign experiment in the main text fixes the width at $m = 256$. As an additional finite-width experiment, we measure $\Pr(\Delta L_{\mathrm{SSAR}} < 0)$ at every width used in the scaling experiment.

Figure 7 shows that the critical and supercritical regimes maintain high loss-decrease probabilities throughout the tested width range. In contrast, the subcritical probability remains substantially lower and does not exhibit the same concentration toward one. This complements the fixed-width sign experiment by showing that the qualitative difference between the regimes persists across the tested network widths.

D Reinforcement-Learning Implementation Details

This section provides the implementation details for the DeepMind Control Suite experiments. SAC, SAC+L_2, and SAC+SSAR use the same base SAC implementation, network architecture, optimizer settings, training budget, and random seeds unless otherwise stated.

D.1 Environments and Evaluation

We evaluate on six state-based continuous-control tasks from the DeepMind Control Suite: Cheetah-run, Hopper-hop, Humanoid-run, Quadruped-run, Quadruped-walk, and Walker-run. Pixel observations are not used.

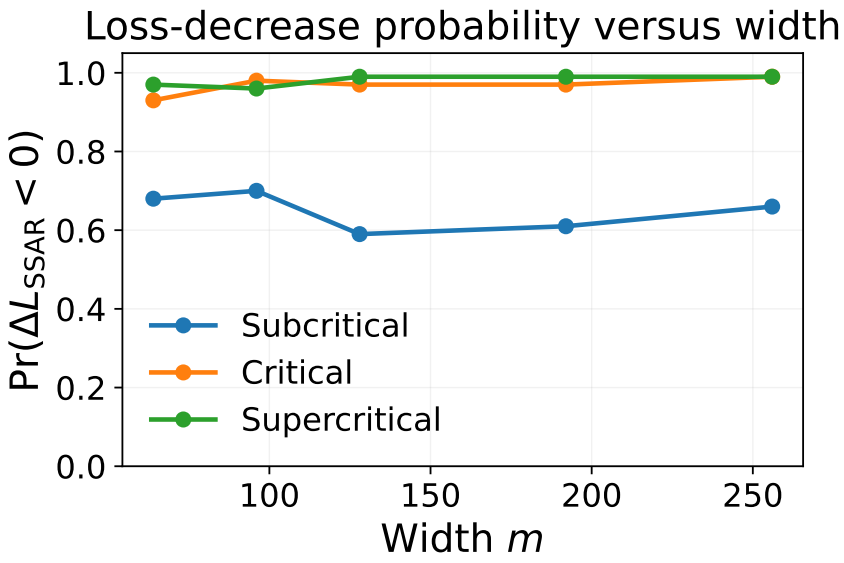


Figure 7: **Loss-decrease probability as a function of network width.** The critical and supercritical regimes maintain high probabilities of a negative SSAR-induced loss change, whereas the subcritical regime remains substantially less reliable.

Each agent is trained for 10^6 environment steps. The first 5,000 steps use uniformly sampled actions to initialize the replay buffer, after which actions are sampled from the SAC policy. The replay-buffer capacity is 10^6 transitions.

Evaluation is performed every 10,000 environment steps using 30 episodes. During evaluation, deterministic actions given by the mean of the policy distribution are used. We use five random seeds, $\{100, 200, 300, 400, 500\}$, for every main experiment.

The learning curves report the mean across the five seeds, with shaded regions denoting one standard deviation. For the final-performance results, we first average the final ten evaluation checkpoints of each seed, corresponding to the final 100k environment steps, and then report the mean and standard deviation over the five seed-level averages.

D.2 SAC Configuration

The actor and both critics are multilayer perceptrons with two hidden layers of 1024 units and ReLU activations. SAC uses two independent Q-functions. The actor parameterizes a squashed Gaussian policy with the log-standard-deviation constrained to $[-5, 2]$.

The actor and critic networks are optimized using Adam with learning rate 10^{-4} and coefficients $(0.9, 0.999)$. The entropy temperature is also learned using Adam with learning rate 10^{-4} and is initialized to 0.1. The target entropy is set to the negative action dimension.

The discount factor is 0.99, the minibatch size is 1024, and the replay buffer contains at most 10^6 transitions. The actor is updated after every critic update. The target critic networks are updated every two critic updates using Polyak averaging with coefficient 0.005.

D.3 SSAR Implementation

SSAR is applied only to the penultimate linear layer of each of the two critics. For SAC, SAC+L_2, and SAC+SSAR, these selected critic matrices use the same initialization so that differences are not caused by different initial spectral structures.

For a selected matrix $W \in \mathbb{R}^{n\times m}$, we compute the singular values of $W/\sqrt{m}$ and use the normalized squared singular values $\lambda_k = s_k^2/m$. The MP upper edge is $\lambda_+ = \sigma^2(1 + \sqrt{n/m})^2$, where σ^2

Table 4: Common hyperparameters used in the DeepMind Control Suite experiments.

Parameter	Value
Training steps	10^6
Initial random steps	5,000
Replay-buffer capacity	10^6
Batch size	1024
Hidden layers	2
Hidden units per layer	1024
Activation	ReLU
Discount factor	0.99
Actor learning rate	10^{-4}
Critic learning rate	10^{-4}
Temperature learning rate	10^{-4}
Adam coefficients	$(0.9, 0.999)$
Initial temperature	0.1
Actor-update frequency	1
Target-update frequency	2
Target-network coefficient	0.005
Evaluation interval	10,000 steps
Evaluation episodes	30
Random seeds	$100, 200, 300, 400, 500$

is the recorded initialization variance of the selected layer. The acceleration window is $[\lambda_+, \lambda_+ + T]$ with $T = cm^{-2/3}$, as defined in the main text.

The spectral operation is evaluated every 20 critic updates. For modes inside the acceleration window, the SSAR penalty is added to the original critic objective. The total critic loss is $L_{\text{tot}} = L_0 + \gamma L_{\text{SSAR}}$, and the main reinforcement-learning experiments use $\gamma = 2$.

For critic $j \in \{1, 2\}$, define $N_{\text{sep}}^{(j)} = \#\{k : \lambda_k^{(j)} > \lambda_+ + T\}$. The stopping statistic is the average across the twin critics, $N_{\text{sep}} = (N_{\text{sep}}^{(1)} + N_{\text{sep}}^{(2)})/2$. We use $N_c = 150$. Once $N_{\text{sep}} \geq 150$, the additional SSAR term is disabled for subsequent critic updates.

D.4 Matched L_2 Baseline

SAC+L_2 applies L_2 regularization to exactly the same penultimate critic weight matrices used by SSAR. No additional L_2 penalty is applied to the actor or the remaining critic layers. If $W^{(1)}$ and $W^{(2)}$ denote the selected matrices of the twin critics, the added regularizer has the form $L_{L_2} = \frac{\lambda_{L_2}}{2}(\|W^{(1)}\|_F^2 + \|W^{(2)}\|_F^2)$. All other architecture, optimization, training, and evaluation settings are identical to vanilla SAC and SAC+SSAR.

D.5 Spectral Diagnostic

The separated-mode statistic in Figure 3 is recorded for SAC, SAC+L_2, and SAC+SSAR using the same MP edge and acceleration window. At each spectral measurement, we compute $N_{\text{sep}}^{(1)}$ and $N_{\text{sep}}^{(2)}$ and report their average. Thus, the diagnostic uses the same definition as the stopping statistic of SSAR and directly compares the spectral evolution of the three methods.

E Sensitivity to the SSAR Coefficient

We examine the sensitivity of SSAR to the spectral regularization coefficient γ on Hopper-hop while keeping all other SAC and SSAR settings unchanged.

The coefficient γ controls the strength of the SSAR term in $L_{\text{tot}} = L_0 + \gamma L_{\text{SSAR}}$. The main experiments use $\gamma = 2$. We additionally evaluate $\gamma = 0.2$ and $\gamma = 20$, spanning two orders of magnitude around the default value.

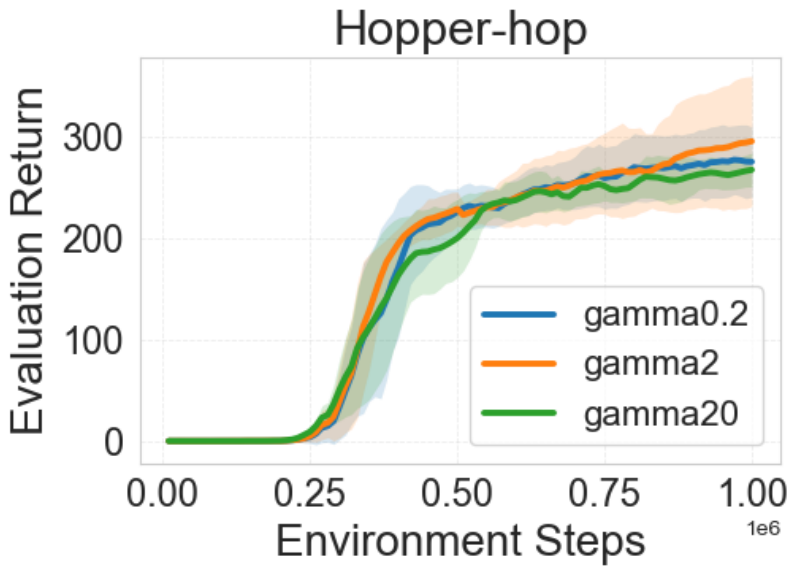


Figure 8: **Sensitivity to the SSAR coefficient γ on Hopper-hop.** We compare $\gamma \in \{0.2, 2, 20\}$ while keeping all other settings unchanged. Curves show the mean evaluation return over three random seeds, and shaded regions denote one standard deviation. Performance remains similar across a wide range of regularization strengths.

Figure 8 shows that the three settings produce broadly similar learning trajectories. Although some differences appear during training and $\gamma = 2$ attains a somewhat higher return near the end, the curves remain close and their variability substantially overlaps. This suggests that SSAR is relatively insensitive to the precise value of γ over the tested range.

A possible explanation is that γ primarily changes how quickly an active near-edge mode is pushed across the upper boundary $\lambda_+ + T$. Once the mode leaves the acceleration window, it is no longer directly accelerated by the corresponding SSAR penalty. Therefore, changing γ can alter the crossing time without substantially changing the later spectral dynamics, which may reduce sensitivity of overall performance to the exact regularization strength.