DIRECTED LAPLACIAN REPRESENTATION LEARNING FOR NON-REVERSIBLE MARKOV CHAINS IN RL

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents a method for Laplacian representation learning in directed Markov chains induced by a fixed reinforcement-learning policy, using the Chung Laplacian as the spectral target. Prior approaches estimate this target from discounted forward transitions, which is exact on reversible chains or at a single step. But a fixed policy generically induces a non-reversible chain, and learning the representation from its stationary transitions poses two challenges, as the energy and orthonormality constraints depend on the unknown stationary distribution, while the forward multi-step sampling can distort the desired representation. Learning in weighted coordinates removes the explicit dependence on the stationary distribution, while a discounting scheme based on the reversible component of the dynamics preserves the target eigenvectors. The method combines forward and time-reversed transitions to construct an unbiased estimator of this energy under stationary sampling. Coupling this estimator with an augmented Lagrangian objective yields local stability guarantees for spectral recovery under a simple-spectrum assumption. Experiments show that the method matches competing approaches in reversible environments and substantially outperforms them in recovering the eigenvectors of the Chung Laplacian in non-reversible settings.

1 INTRODUCTION

Laplacian representations encode the transition structure of a reinforcement-learning (RL) problem in reward-independent features, supporting value approximation (Mahadevan & Maggioni, 2007), option discovery (Machado et al., 2017), and decision-time planning (Shehmar et al., 2026). Sample-based methods learn these features with neural networks without constructing the transition matrix (Wu et al., 2019; Wang et al., 2021; Gomez et al., 2024). When policy-induced dynamics are non-reversible, however, the sampling scheme can change the spectral target: even exact optimization need not recover the intended representation.

The Chung Laplacian (Chung, 2005) provides an established spectral construction for directed Markov chains, and directed spectral bases have previously been studied both in RL and more broadly Johns & Mahadevan (2007); Meila & Pentney (2007); Seabrook & Wiskott (2023). Our contribution is not a new directed Laplacian, but rather an analysis of how sample-based multi-step learning can change this target. Under stationary sampling, the one-step graph drawing objective (GDO) estimates the Chung energy in stationary-weighted coordinates Wu et al. (2019); Giorgi et al. (2026). This correspondence need not survive the forward multi-step sampling used for spectral conditioning Wu et al. (2019); Gomez et al. (2024). We show that forward geometric-lag sampling discounts before reversibilizing, whereas eigenvector-preserving discounting reversibilizes first. For each $\gamma \in (0, 1)$, the resulting operators coincide iff the chain is reversible. Otherwise, the mismatch can alter the minimizing eigenspace even with exact expectations – an operator bias that more data or better optimization cannot remove.

We introduce the *Directed Augmented Lagrangian Laplacian Objective* (DALLO), which corrects the sampler while retaining ALLO's optimization framework (Gomez et al., 2024). At each substep, DALLO independently selects the forward or time-reversed kernel with equal probability, then draws a fresh conditional transition from the current state. Under stationary sampling with exact kernels, this yields an unbiased estimator of the discounted reversible surrogate's energy, preserving the Chung eigenvectors at the population level. Stationary-weighted coordinates avoid ex-

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.

plicit stationary-probability estimation during training. In our finite-state setting, a replay buffer approximates both kernels by resampling outgoing and incoming records; continuous-state extensions would require estimating the reverse conditional.

On reversible chains, DALLO and ALLO have the same population spectral target. Under a simple-spectrum assumption, DALLO inherits local stability guarantees for ordered spectral recovery in stationary-weighted function space. Using ALLO's architecture and optimization hyperparameters, DALLO achieves mean per-vector cosine similarity above 0.97 on directed-cycle benchmarks, versus approximately 0.33–0.49 for ALLO at the same discount. It also improves over one-step ALLO on every reported directed-cycle and coupled-cycle configuration. Correcting the sampler thus retains the benefits of multi-step spectral conditioning without changing the spectral target.

2 Related Work

Laplacian representations and sample-based learning. Proto-value functions use graph-Laplacian eigenvectors as reward-independent bases for value approximation Mahadevan (2005); Mahadevan & Maggioni (2007), with subsequent applications to option discovery, exploration, and planning Machado et al. (2017); Shehmar et al. (2026). Wu et al. (2019) introduced the graph drawing objective (GDO) for learning such representations from samples; Wang et al. (2021) proposed GGDO to identify individual eigenvectors; and Gomez et al. (2024) introduced ALLO, using an augmented-Lagrangian formulation with stability guarantees for ordered eigenvector recovery. We retain ALLO's optimization machinery and modify the transition estimator it is coupled with.

Directed spectral constructions and non-reversible chains. Spectral constructions for directed graphs are well established. Chung (2005) introduced the directed Laplacian used here, based on the stationary distribution and additive reversibilization of the Markov chain. Johns & Mahadevan (2007) used directed-graph spectral bases for value-function approximation in non-reversible MDPs, while (Meila & Pentney, 2007) studied spectral objectives for directed graphs through weighted-cut relaxations. Seabrook & Wiskott (2023) provide a broader treatment of the spectral theory of reversible and non-reversible Markov chains. More recently, Giorgi et al. (2026) clarified the stationary-weighted interpretation of sampled graph-drawing objectives and analyzed how spectral estimation affects value approximation under general policies.

Our contribution is to characterize how multi-step sampling interacts with Chung spectral targets in neural Laplacian learning. We show that sampling and reversibilization generally do not commute, identify when this changes the minimizing spectral subspace, and introduce a forward/time-reversed sampler that preserves the desired Chung operator while remaining compatible with ALLO.

3 Background

Setting. Let $\mathcal{M} = (\rho_0, \mathcal{S}, \mathcal{A}, \mathcal{P})$ be a reward-agnostic Markov decision process, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\rho_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, with $\mathcal{P}(s' \mid s, a)$ the probability of transitioning to s' from s under action a. A fixed policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is provided as input to our method and is not learned. It induces the *policy transition matrix* $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ with entries $P_\pi(s, s') = \sum_{a\in\mathcal{A}} \pi(a \mid s)\, \mathcal{P}(s' \mid s, a)$. In general P_π is asymmetric, and the induced chain is directed. We assume P_π is ergodic, so that it admits a unique stationary distribution $d^\pi \in \Delta(\mathcal{S})$ satisfying $(d^\pi)^T P_\pi = (d^\pi)^T$ and $d^\pi(s) > 0$ for all $s \in \mathcal{S}$. We use the notation $[k] := \{1, \dots, k\}$.

The Laplacian representation. For a weighted graph with adjacency matrix A and degree matrix $D_A := \mathrm{diag}(A\mathbf{1})$, the graph Laplacian is $L = D_A - A$. Its eigenvectors capture directions of variation across the graph, and the associated eigenvalues quantify how much variation each direction carries. In RL, the state space plays the role of the node set, while the policy-induced transition matrix P_π defines the transition structure. When P_π is symmetric, the standard graph Laplacian is $L = I - P_\pi$. For asymmetric P_π, prior Laplacian-representation work commonly uses the symmetrized Laplacian $L = I - \frac{1}{2}(P_\pi + P_\pi^\top)$, which is symmetric by construction (Wu et al., 2019; Gomez et al., 2024). The *Laplacian representation* $f : \mathcal{S} \to \mathbb{R}^k$ collects the entries of the k smallest eigenvectors of L at each state, that is, $f(s) = (v_1(s), \dots, v_k(s))^\top$, where v_i denotes the eigenvector associated with the i-th smallest eigenvalue of L and $v_i(s)$ its entry at state s (Mahadevan

& Maggioni, 2007; Stachenfeld et al., 2014; Machado et al., 2017). Both constructions of L are widely used (Mahadevan, 2005; Wang et al., 2021; Gomez et al., 2024); Section 4.1 examines their mismatch with the directed chain.

Dirichlet energy. For a Laplacian L that is self-adjoint with respect to the inner product under consideration, its *Dirichlet energy* is $\mathcal{E}_L(f) := \langle f, Lf\rangle$. When the inner product is weighted by a positive diagonal matrix D, we write $\mathcal{E}_L^D(f) := \langle f, Lf\rangle_D = f^\top DLf$. For the combinatorial Laplacian of an undirected graph with nonnegative edge weights $w(u,v) = w(v,u)$, the energy in the standard inner product is

$$\mathcal{E}_L(f) = \frac{1}{2}\sum_{u,v} w(u,v)\big(f(u) - f(v)\big)^2.$$

This formula penalizes differences across edges, with larger weights imposing a stronger penalty, so minimizing the energy favors smooth functions on the graph. Similarly, for $L = I - Q$, where Q is a reversible Markov kernel with stationary distribution d and $D = \mathrm{diag}(d)$,

$$\mathcal{E}_{I-Q}^D(f) = \frac{1}{2}\sum_{s,s'} d(s)Q(s,s')\big(f(s) - f(s')\big)^2.$$

In both standard and weighted settings, energy and orthonormality use the same inner product. If $Lu = \lambda u$ and $\langle u, u\rangle = 1$, then $\langle u, Lu\rangle = \lambda$, so each eigenvalue is the energy of its normalized eigenvector. More generally, the k eigenvectors with smallest eigenvalues minimize total energy over all orthonormal k-tuples. Thus, Laplacian representations can be learned by minimizing transition-based energy estimates under orthonormality constraints.

Graph drawing objectives. Since the transition operator is available only through samples, the eigenvectors of L are learned by optimizing objectives whose terms are expectations over transitions. The k smallest eigenvectors globally solve the spectral graph drawing problem Koren (2003)

$$\min_{u_1,\ldots,u_k \in \mathbb{R}^{|\mathcal{S}|}} \sum_{i=1}^{k} \mathcal{E}_L(u_i) \quad \text{s.t.} \quad \langle u_j, u_i\rangle = \delta_{ji}, \;\; 1 \le i \le j \le k. \tag{1}$$

which Wu et al. (2019) relax to an unconstrained graph-drawing objective (GDO) by penalizing violations of the orthonormality constraints, with penalty coefficient $b > 0$. The GDO objective is invariant under orthogonal rotations of the feature matrix and therefore does not identify individual eigenvectors (Gomez et al., 2024). The generalized graph-drawing objective (GGDO) of Wang et al. (2021) breaks this symmetry by assigning distinct, decreasing weights to the individual energy terms. Under the assumptions of Wang et al. (2021), including distinct relevant eigenvalues, the constrained problem identifies the ordered eigenvectors up to sign. In the finite-penalty unconstrained formulation used for neural-network training, however, optimization can still converge to undesirable rotated representations, and the recovered representation can depend on the choice of weights and the penalty coefficient b (Gomez et al., 2024).

Augmented Lagrangian Laplacian objective. Gomez et al. (2024) address both issues with ALLO, which applies a stop-gradient operator $[\![\cdot]\!]$ (identity in the forward pass, zero gradient in the backward pass) to the second factor of each constraint, restricts the inner sum to $i \le j$, and reintroduces the constraints through dual variables $(\beta_{ji})_{1\le i\le j\le k}$,

$$\max_{\beta}\min_{u} \sum_{i=1}^{k} \mathcal{E}_L(u_i) + \sum_{j=1}^{k}\sum_{i=1}^{j} \beta_{ji}\big(\langle u_j, [\![u_i]\!]\rangle - \delta_{ji}\big) + b\sum_{j=1}^{k}\sum_{i=1}^{j}\big(\langle u_j, [\![u_i]\!]\rangle - \delta_{ji}\big)^2. \tag{2}$$

For any barrier coefficient $b > 2$, the asymmetry introduced by the stop-gradient removes the rotational degeneracy, the ordered eigensystem becomes the unique stable equilibrium of the gradient ascent–descent dynamics, and the diagonal duals β_{ii}^* are proportional to the eigenvalues.

Spectral conditioning by discounting. Small spectral gaps can make the recovered eigenspaces sensitive to perturbations and optimization error. Discounted multi-step transitions are used in Laplacian representation learning (Wu et al., 2019; Gomez et al., 2024). For a stochastic operator P and $\gamma \in [0, 1)$, define

$$P^\gamma := (1-\gamma)\sum_{t\ge 1} \gamma^{t-1}P^t = (1-\gamma)P(I - \gamma P)^{-1}. \tag{3}$$

If P is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{D_\pi}$, P^γ is self-adjoint in the same inner product and has the same eigenspaces as P. If λ is an eigenvalue of $I - P$, the corresponding eigenvalue of $I - P^\gamma$ is

$$\lambda^\gamma := h^\gamma(\lambda) = \frac{\lambda}{1 - \gamma + \gamma\lambda}. \tag{4}$$

This strictly increasing map preserves the ordered eigenspaces and the zero eigenspace, and has inverse $\lambda = (1-\gamma)\lambda^\gamma/(1-\gamma\lambda^\gamma)$. For $\gamma > 0$, it expands spectral gaps sufficiently close to zero. These properties and the series identity in equation 3 are derived in Appendix B.

The geometric mixture in equation 3 corresponds to sampling pairs of states K steps apart, with $\mathbb{P}(K = k) = (1-\gamma)\gamma^{k-1}$ for $k \geq 1$ (Wu et al., 2019, Appendix B). For a nonreversible chain, however, the squared-difference objective based on these forward pairs symmetrizes the kernel after discounting. Section 4.3 shows that, for $\gamma > 0$, this estimates a different energy from that obtained by discounting the reversible surrogate.

4 The Directed Augmented Lagrangian Laplacian Objective

For the directed setting we adopt the Chung Laplacian Chung (2005). Let $D_\pi := \mathrm{diag}(d^\pi)$ be the diagonal matrix of the stationary distribution. Since P_π is real-valued, Chung's conjugate transpose reduces to the ordinary transpose:

$$\mathcal{L} := I - \frac{1}{2}\left(D_\pi^{1/2} P_\pi D_\pi^{-1/2} + D_\pi^{-1/2} P_\pi^\top D_\pi^{1/2}\right). \tag{5}$$

The matrix $\mathcal{L}$ is symmetric by construction, and the D_π weighting accounts for the stationary distribution of the chain at every state, so the operator reflects the geometry of the directed random walk. It yields a Cheeger inequality and convergence bounds for non-reversible Markov chains (Chung, 2005). Compared with the symmetrized Laplacian of Section 3, the Chung Laplacian accounts for the stationary distribution of the chain.

Proposition 1. *For an irreducible Markov chain with stationary distribution d^π, the Chung Laplacian coincides with the ordinary symmetrization, $\mathcal{L} = L$, if and only if d^π is uniform.*

The proof appears in Appendix C. In RL the Chung Laplacian has appeared in the tabular setting, where Johns & Mahadevan (2007) used it to construct basis functions on directed state graphs, and Giorgi et al. (2026) recently analyzed the value approximation error of spectral features through an operator related to equation 5 by a similarity transformation, under an oracle assumption on the accuracy of the learned features.

4.1 The Chung Laplacian and the weighted spectral problem

As before, we view functions on $\mathcal{S}$ as vectors and learn features $g_1, \ldots, g_k$ by replacing L with the Chung Laplacian $\mathcal{L}$ in (1), subject to $\langle g_j, g_i \rangle = \delta_{ji}$. Assuming transition pairs $s \sim d^\pi$ and $s' \sim P_\pi(\cdot \mid s)$, the Dirichlet energy of a feature g using Chung Laplacian (Chung, 2005) is:

$$\mathcal{E}_\mathcal{L}(g) = \frac{1}{2}\sum_{s,s'} d^\pi(s) P_\pi(s,s') \left(\frac{g(s)}{\sqrt{d^\pi(s)}} - \frac{g(s')}{\sqrt{d^\pi(s')}}\right)^2. \tag{6}$$

Direct optimization of g requires stationary probabilities at sampled states: the energy uses $1/\sqrt{d^\pi(s)}$, while unbiased orthogonality estimation uses $\mathbb{E}_{s\sim d^\pi}[g_j(s)g_i(s)/d^\pi(s)] = \langle g_j, g_i \rangle$. Thus rare states induce poorly scaled inverse-probability updates, and replacing d^π by $\hat{d}^\pi$ amplifies errors where these weights are largest. We therefore write $g_i = D_\pi^{1/2} f_i$ and optimize in f-coordinates. The energy then becomes

$$\mathcal{E}_\mathcal{L}(D_\pi^{1/2} f) = \frac{1}{2}\sum_{s,s'} d^\pi(s) P_\pi(s,s')\big(f(s) - f(s')\big)^2 = \frac{1}{2}\mathbb{E}_{\substack{s\sim d^\pi \\ s'\sim P_\pi(\cdot|s)}}\big[(f(s) - f(s'))^2\big].$$

This is the expected squared-difference formulation underlying the sampled graph-drawing objective of Wu et al. (2019, Eqs. (3), (5)). Likewise, the orthonormality constraints become

$$\langle g_j, g_i \rangle = \langle f_j, f_i \rangle_{D_\pi} = \mathbb{E}_{s\sim d^\pi}[f_j(s) f_i(s)] = \delta_{ji}.$$

Learning f_i directly removes explicit inverse-probability factors: both the energy and constraint inner products are estimated from stationary samples without evaluating d^π per state.

The quadratic form in these coordinates is represented by

$$\mathcal{W} := D_\pi^{1/2} \mathcal{L} D_\pi^{1/2} = D_\pi - \frac{1}{2}\left(D_\pi P_\pi + P_\pi^\top D_\pi\right),$$

the combinatorial form of the Chung Laplacian (Chung, 2005). The following proposition records the energy equality and the corresponding relation between the spectral problems.

Proposition 2. *For $g = D_\pi^{1/2} f$, $\mathcal{E}_\mathcal{W}(f) = \mathcal{E}_\mathcal{L}(g)$. Moreover, $\mathcal{W} f = \lambda D_\pi f$ if and only if $\mathcal{L} g = \lambda g$.*

The proof is deferred to Appendix C. Every term of the weighted problem is now a buffer average (cf. Wu et al., 2019; Giorgi et al., 2026). For a transition (s, s') drawn from the buffer, the energy $\mathcal{E}_\mathcal{W}(f_j)$ is estimated by $\frac{1}{2}(f_j(s) - f_j(s'))^2$ and the constraint by $f_j(s) f_i(s)$, with no importance weights and no access to d^π (Appendix C). The network learns f_j; the Chung eigenvectors $g_j = D_\pi^{1/2} f_j$ are needed only for evaluation and require an estimate of d^π.

The spectral conditioning of equation 3 is still missing. Its sampler pairs states a geometric number of steps apart, which estimates powers of a transition kernel, and $\mathcal{W}$ is not a kernel. Discounting P_π itself does not escape either, as Section 4.3 shows that the standard sampler then estimates the wrong operator. We therefore need a Markov kernel that is self-adjoint in $\langle \cdot, \cdot \rangle_{D_\pi}$ and shares the Chung spectrum; Section 4.2 extracts one from $\mathcal{W}$.

4.2 The reversible surrogate

The operator $\mathcal{W}$ factors through a reversible Markov kernel. The *time reversal* of P_π is

$$P_\pi^{\mathrm{rev}}(s, s') = \frac{d^\pi(s')}{d^\pi(s)} P_\pi(s', s), \tag{7}$$

which is the adjoint of P_π in $\langle \cdot, \cdot \rangle_{D_\pi}$. The *additive reversibilization* of P_π (Fill, 1991) is

$$P_{\mathrm{alt}} = \frac{1}{2}\left(P_\pi + P_\pi^{\mathrm{rev}}\right). \tag{8}$$

We call P_{alt} the *reversible surrogate* of P_π. The kernel P_{alt} is stochastic, preserves d^π, and is reversible, hence self-adjoint in $\langle \cdot, \cdot \rangle_{D_\pi}$ (Appendix D). Since $D_\pi P_\pi^{\mathrm{rev}} = P_\pi^\top D_\pi$, $\mathcal{W} = D_\pi(I - P_{\mathrm{alt}})$. Under the change of coordinates $g = D_\pi^{1/2} f$, the Chung energy and eigenvalue problem take the form $\mathcal{E}_\mathcal{L}(g) = \mathcal{E}^{D_\pi}_{I - P_{\mathrm{alt}}}(f)$, and $\mathcal{L} g = \lambda g \iff (I - P_{\mathrm{alt}}) f = \lambda f$, by factorization of $\mathcal{W}$ and Proposition 2.

To use discounting while preserving the eigenvectors of $I - P_{\mathrm{alt}}$, we apply equation 3 to P_{alt}: $P_{\mathrm{alt}}^\gamma := (1-\gamma) P_{\mathrm{alt}} (I - \gamma P_{\mathrm{alt}})^{-1}$, $\widetilde{L}_{\mathrm{alt}} := I - P_{\mathrm{alt}}^\gamma$. The corresponding discounted Dirichlet energy is $\mathcal{E}^{D_\pi}_{\widetilde{L}_{\mathrm{alt}}}(f) = \langle f, (I - P_{\mathrm{alt}}^\gamma) f \rangle_{D_\pi}$. The operator $\widetilde{L}_{\mathrm{alt}}$ has the same eigenvectors as $I - P_{\mathrm{alt}}$, with eigenvalues transformed by the strictly increasing map in equation 4 (Appendix B). We next examine whether forward sampling from P_π estimates this discounted energy and how to modify the sampling procedure when it does not. Next we show that the resulting spectrum generally depends on the order of symmetrization and discounting.

4.3 What the sampled objective estimates

We first identify the energy estimated by the forward sampling procedure used by Wu et al. (2019). We then show how combining forward and time-reversed transitions yields an unbiased estimator of $\mathcal{E}^{D_\pi}_{\widetilde{L}_{\mathrm{alt}}}(f)$. Throughout this subsection, the initial state of each sampled pair is drawn from the stationary distribution d^π. Proofs appear in Appendix E.

We state the following lemma for a general transition matrix Q, since we will apply it both to P_π and to the discounted transition matrices considered below.

Lemma 1. *Let Q be a stochastic kernel with stationary distribution d^π, and let Q^{rev} be its adjoint in $\langle \cdot, \cdot \rangle_{D_\pi}$. For $s \sim d^\pi$, $s' \sim Q(\cdot \mid s)$, and any f,*

$$\tfrac{1}{2}\mathbb{E}\left[(f(s) - f(s'))^2\right] = \left\langle f, (I - \tfrac{1}{2}(Q + Q^{\mathrm{rev}})) f \right\rangle_{D_\pi}.$$

For $Q = P_\pi$, Lemma 1 shows that stationary one-step sampling estimates the Chung energy in weighted coordinates (Wu et al., 2019, Eq. 5). For two forward steps, however, the sampled energy corresponds to $I - \frac{1}{2}\big(P_\pi^2 + P_\pi^{\mathrm{rev}\,2}\big)$, which generally differs from $I - P_{\mathrm{alt}}^2$ because P_{alt}^2 also contains the mixed products $P_\pi P_\pi^{\mathrm{rev}}$ and $P_\pi^{\mathrm{rev}} P_\pi$.

For a geometrically distributed number of forward steps, the transition matrix between the sampled states is P_π^γ. Applying Lemma 1 with $Q = P_\pi^\gamma$ gives the following corollary.

Corollary 1. *Let $s \sim d^\pi$, and independently draw K with $\mathbb{P}(K = k) = (1-\gamma)\gamma^{k-1}$ for $k \geq 1$, where $\gamma \in (0,1)$. If $s' \sim P_\pi^K(\cdot \mid s)$, then, for every $f : \mathcal{S} \to \mathbb{R}$,*

$$\tfrac{1}{2}\mathbb{E}[(f(s) - f(s'))^2] = \left\langle f, \left(I - \tfrac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)\right) f \right\rangle_{D_\pi}.$$

Thus, forward geometric-lag sampling discounts the transition kernel before taking its additive reversibilization, whereas the target $\widetilde{L}_{\mathrm{alt}}$ is obtained in the reverse order. The following proposition shows that these two operations commute exactly in the reversible case.

Proposition 3. *For every $\gamma \in [0,1)$, $P_{\mathrm{alt}}^\gamma = \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma) = P_\pi^\gamma$ if and only if P_π is reversible.*

Hence, on a nonreversible chain, discounting and reversibilization do not commute, so Corollary 1 implies that forward multi-step sampling estimates a different energy even with exact expectations. The next result gives a sufficient condition under which this mismatch also changes the leading spectral subspace.

Proposition 4. *Let $f_1, \ldots, f_{|\mathcal{S}|}$ be a D_π-orthonormal eigenbasis of $I - P_{\mathrm{alt}}$, with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_{|\mathcal{S}|}$. Fix $1 \leq k < |\mathcal{S}|$ with $\lambda_k < \lambda_{k+1}$. If $\langle (P_\pi - P_\pi^{\mathrm{rev}}) f_i, (P_\pi - P_\pi^{\mathrm{rev}}) f_j \rangle_{D_\pi} \neq 0$ for some $i \leq k < j$, then, for all sufficiently small $\gamma > 0$, the span of $f_1, \ldots, f_k$ minimizes the spectral objective for $\widetilde{L}_{\mathrm{alt}}$, but not for $I - \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)$.*

We now construct an estimator of the target energy by combining forward and time-reversed transitions. For a stationary chain,

$$\mathbb{P}(s_t = s \mid s_{t+1} = s') = \frac{d^\pi(s) P_\pi(s, s')}{d^\pi(s')} = P_\pi^{\mathrm{rev}}(s', s).$$

Thus, the reverse kernel describes the past state distribution given the current state.

Proposition 5. *Let $s^{(0)} \sim d^\pi$ and independently choose K with $\mathbb{P}(K = k) = (1-\gamma)\gamma^{k-1}$ for $k \geq 1$, where $\gamma \in (0,1)$. Starting from $s^{(0)}$, perform K steps. At each step, independently of K and all previous choices, select P_π or P_π^{rev} with equal probability, and draw a fresh next state from the chosen kernel conditional on the current state. Then, for every $f : \mathcal{S} \to \mathbb{R}$,*

$$\frac{1}{2}\mathbb{E}\Big[\big(f(s^{(0)}) - f(s^{(K)})\big)^2\Big] = \mathcal{E}_{\widetilde{L}_{\mathrm{alt}}}^{D_\pi}(f).$$

Because the forward/reverse kernel is chosen independently at every substep, the k-step transition includes mixed products such as $P_\pi P_\pi^{\mathrm{rev}}$ and $P_\pi^{\mathrm{rev}} P_\pi$, yielding P_{alt}^k. Averaging over the geometric stopping time therefore gives P_{alt}^γ by equation 3. Since P_{alt}^γ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_{D_\pi}$, Lemma 1 gives the stated energy equality.

4.4 The objective

Instantiating the augmented Lagrangian construction equation 2 with the operator $\widetilde{L}_{\mathrm{alt}}$ of Section 4.2 and the inner product $\langle \cdot, \cdot \rangle_{D_\pi}$ yields the *Directed Augmented Lagrangian Laplacian Objective* (DALLO). For $b > 0$ and dual variables $\beta = (\beta_{ji})_{1 \leq i \leq j \leq k}$, define

$$\mathcal{A}_b^\gamma(f, \beta) := \sum_{i=1}^k \mathcal{E}_{\widetilde{L}_{\mathrm{alt}}}^{D_\pi}(f_i) + \sum_{j=1}^k \sum_{i=1}^j \beta_{ji}(\langle f_j, [\![f_i]\!] \rangle_{D_\pi} - \delta_{ji}) + b \sum_{j=1}^k \sum_{i=1}^j (\langle f_j, [\![f_i]\!] \rangle_{D_\pi} - \delta_{ji})^2 \quad (9)$$

with the associated saddle problem $\max_\beta \min_{f_1, \ldots, f_k} \mathcal{A}_b^\gamma(f, \beta)$. As in ALLO, the stop-gradient enters only the constraints; the operator and the inner product are the only changes. The energy terms are estimated from alternating-walk pairs (Proposition 5) and the constraints by plain buffer averages, so equation 9 is a discounted Chung spectral problem estimable from the buffer alone in the finite-state setting.

Proposition 3 immediately yields the reversible-case reduction.

Corollary 2. *If P_π is reversible, then, $P^\gamma_{\mathrm{alt}} = \frac{1}{2}(P^\gamma_\pi + (P^{\mathrm{rev}}_\pi)^\gamma) = P^\gamma_\pi$ for every $\gamma \in [0, 1)$, so DALLO equation 9 and ALLO equation 2 have the same population spectral target $I - P^\gamma_\pi$.*

Thus, DALLO and ALLO share the same population spectral target on reversible chains, while DALLO extends to a different target and sampling procedure under non-reversible dynamics.

4.5 Recovery guarantees

The change of coordinates $g_i = D_\pi^{1/2} f_i$ transforms the weighted inner products into standard inner products and the energy operator into the symmetric matrix $D_\pi^{1/2} \widetilde{L}_{\mathrm{alt}} D_\pi^{-1/2}$. Under this transformation, the continuous-time updates for equation 9, using the D_π-weighted inner product and prescribed stop-gradient, reduce to the corresponding ALLO updates. We therefore apply the equilibrium and stability analysis of Gomez et al. (2024); details appear in Appendices F.1 and F.2.

The equilibrium conditions admit orthonormal eigenfunctions with different selections and orderings. To characterize locally asymptotically stable individual equilibria, we assume the selected eigenvalues are simple and separated from the remaining spectrum; repeated eigenvalues would instead yield continuous families of equilibria through within-eigenspace rotations. Similar assumptions appear in Gomez et al. (2024, Appendix A.4) and Gemp et al. (2021, Theorem 2.1). The following theorem identifies the stable selection and ordering.

Theorem 1. *Let $n = |\mathcal{S}|$, fix $1 \le k < n$, and let $u_1, \dots, u_n$ be a D_π-orthonormal eigenbasis of $\widetilde{L}_{\mathrm{alt}}$, with eigenvalues ordered as $0 = \lambda^\gamma_1 < \cdots < \lambda^\gamma_k < \lambda^\gamma_{k+1} \le \cdots \le \lambda^\gamma_n$. Consider the continuous-time primal–dual updates for equation 9, using the D_π-weighted inner product and the prescribed stop-gradient convention. For any $b > 2$, the locally asymptotically stable equilibria are precisely $f^\star_i = \varepsilon_i u_i$, $\varepsilon_i \in \{-1, 1\}$ for $i \in [k]$ with $\beta^\star_{ii} = -2\lambda^\gamma_i$, $\beta^\star_{ji} = 0$ for $1 \le i < j \le k$. Every equilibrium with a different selection or ordering of the eigenfunctions is linearly unstable.*

The eigenvalues of the Chung Laplacian are recovered from $\lambda^\gamma_i = -\beta^\star_{ii}/2$ by inverting the spectral map in equation 4.

4.6 Limits of bounded representations

The weighted-coordinate formulation avoids evaluating $d^\pi(s)$ during training but does not eliminate rare-state difficulties. Rare states may require large representation values, while direct estimation in Chung coordinates incurs equivalent inverse-density scaling through terms such as $g_j(s)g_i(s)/d^\pi(s)$. Thus, even oracle access to d^π does not resolve the problem, motivating the bounded-output analysis below.

Let d be a positive probability distribution on $\mathcal{S}$, set $D := \mathrm{diag}(d)$, and let $\Phi : \mathcal{S} \to \mathbb{R}^k$ have coordinate functions $\phi_1, \dots, \phi_k$. Define their Gram matrix by $G_\Phi := (\langle \phi_i, \phi_j \rangle_D)^k_{i,j=1}$. Thus, D-orthonormality is equivalent to $G_\Phi = I_k$. We measure its violation by the Frobenius norm $\|G_\Phi - I_k\|_F := (\sum^k_{i,j=1}((G_\Phi)_{ij} - \delta_{ij})^2)^{1/2}$. If most probability mass is concentrated on few states, their contribution to G_Φ is low rank, while bounded outputs control the rest.

Theorem 2. *Suppose $|\phi_i(s)| \le M$ for every $s \in \mathcal{S}$ and $i \in [k]$, where $M > 0$. Let $S_0 \subseteq \mathcal{S}$, write $r := |S_0|$ and $\delta := \sum_{s \notin S_0} d(s)$, and fix $\varepsilon \in (0, 1)$. If $k > r$ and $(r + 1)M^2\delta \le 1 - \varepsilon$, then $\|G_\Phi - I_k\|_F \ge \varepsilon$.*

The proof appears in Appendix G. Under the stated concentration condition, orthonormality error below ε therefore requires $k \le r$, independently of the training algorithm. For example, if two states each have probability 0.495 and the rest total 0.01, then with $M = 2$ the theorem gives $\|G_\Phi - I_k\|_F \ge 0.88$ for all $k \ge 3$, so even perfect optimization cannot make the orthonormality error arbitrarily small.

5 Experiments

Experimental Setup. DALLO retains ALLO's released architecture, optimizer, numerical hyperparameters, and 10^6-transition budget. ALLO uses its released 50-step episodic protocol, while DALLO uses 1,000-step trajectories. We additionally verify that extending ALLO to 1,000-step

episodes does not improve performance (see Appendix K). We denote the learned representations by $f_{\mathrm{DALLO}}, f_{\mathrm{ALLO},\gamma} : \mathcal{S} \to \mathbb{R}^d$, with $d = 11$ in all experiments. We evaluate on eight larger grid mazes from Gomez et al. (2024), together with directed Cycle and CycleNet benchmarks designed to introduce non-reversibility and increasing structural complexity. Detailed descriptions of the environments, policy constructions, and all experimental configurations are provided in Appendix H.

Evaluation. ALLO evaluates gridworlds under a uniform policy, where $D_\pi^{1/2} = cI$, so weighted and Euclidean coordinates differ only by a global scale and cosine similarity is unchanged. Thus, the recovery issue is hidden in that setting. As noted by Giorgi et al. (2026), sample-based Laplacian objectives are expressed in the stationary-distribution weighted geometry. Under non-uniform visitation this distinction matters. By Proposition 2, Chung coordinates are recovered as $g = D_\pi^{1/2} f$; we therefore apply the same readout to both DALLO and ALLO. Results use $\hat{D}_{\mathrm{D}}^{1/2} f$ for DALLO and $\hat{D}_{\mathrm{A}}^{1/2} f$ for ALLO, where $\hat{d}_{\mathrm{D}}^{\pi}$ and $\hat{d}_{\mathrm{A}}^{\pi}$ are fitted to their respective replay-buffer distributions. Oracle $D_\pi^{1/2} f$ and raw f readouts are reported in the appendix. For numerically repeated eigenvalues ($|\lambda_i - \lambda_j| < 10^{-11}$), we rotate the learned vectors within the eigenspace to best align with the true basis before scoring, following ALLO's eigenspace treatment; unlike ALLO's five-decimal grouping, we use a tighter gap to avoid merging distinct small Chung eigenvalues in biased chains.

Grid mazes. The grid chains are reversible, so DALLO and ALLO target the same population operator (Corollary 2). Across the oracle, practical, and raw readouts, ALLO is significantly better on three of eight mazes under uniform visitation. Under non-uniform visitation, DALLO generally dominates: it is significantly better on all eight near-uniform mazes under the oracle and practical readouts and seven of eight using raw outputs; under biased visitation, DALLO wins six of eight, all eight, and four of eight settings, respectively. Full results are reported in Tables 6–8.

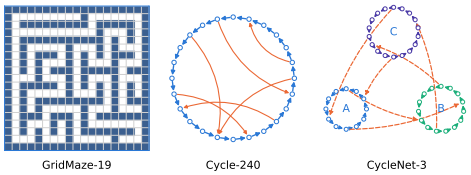

Figure 1: Representative reversible and directed benchmark environments.

Directed cycles and CycleNets. Across the oracle, practical, and raw readouts, DALLO achieves higher recovery than ALLO at $\gamma = 0.9$ in all 15 directed-cycle settings. Setting ALLO to $\gamma = 0$ removes the multi-step discounting mismatch and improves its Chung recovery, yet DALLO still significantly outperforms this one-step control in all 15 settings. On the more complex CycleNets, DALLO significantly outperforms the one-step control in all 12 settings and achieves higher recovery than ALLO at $\gamma = 0.9$ in 11 of 12. Full results are reported in Tables 9–13.

Table 1: Directed-cycle recovery using method-specific practical readouts: $\hat{D}_{\mathrm{D}}^{1/2} f$ for DALLO and $\hat{D}_{\mathrm{A}}^{1/2} f$ for ALLO. "ALLO+γ" uses $\gamma = 0.9$; "ALLO" uses $\gamma = 0$. Mean (s.d.) over 30 seeds; Welch's t-test compares DALLO with the control, with bold indicating $p < 0.01$.

Env	Policy	DALLO + $\hat{d}_{\mathrm{D}}^{\pi}$	ALLO + γ + $\hat{d}_{\mathrm{A}}^{\pi}$	ALLO + $\hat{d}_{\mathrm{A}}^{\pi}$	t-statistic	p-value
Cycle-240	uniform	**0.9979 (0.0007)**	0.4243 (0.0067)	0.9587 (0.0107)	20.0	0.000
	near-uniform	**0.9978 (0.0007)**	0.4327 (0.0053)	0.9602 (0.0102)	20.1	0.000
	biased	**0.9970 (0.0018)**	0.4720 (0.0098)	0.9577 (0.0053)	38.6	0.000
Cycle-500	uniform	**0.9774 (0.0378)**	0.4719 (0.0116)	0.8652 (0.0267)	13.3	0.000
	near-uniform	**0.9763 (0.0237)**	0.4421 (0.0103)	0.8568 (0.0180)	22.0	0.000
	biased	**0.9718 (0.0339)**	0.4783 (0.0128)	0.9285 (0.0164)	6.3	0.000
Cycle-600	uniform	**0.9944 (0.0038)**	0.4362 (0.0131)	0.9125 (0.0096)	43.3	0.000
	near-uniform	**0.9954 (0.0025)**	0.4378 (0.0125)	0.9114 (0.0127)	35.6	0.000
	biased	**0.9960 (0.0018)**	0.4233 (0.0159)	0.8730 (0.0516)	13.1	0.000
Cycle-1100	uniform	**0.9939 (0.0025)**	0.3984 (0.0190)	0.8402 (0.0473)	17.8	0.000
	near-uniform	**0.9932 (0.0026)**	0.3957 (0.0210)	0.8521 (0.0418)	18.5	0.000
	biased	**0.9954 (0.0013)**	0.3240 (0.0084)	0.7730 (0.0368)	33.1	0.000
Cycle-1300	uniform	**0.9929 (0.0029)**	0.3896 (0.0188)	0.9584 (0.0183)	10.2	0.000
	near-uniform	**0.9922 (0.0067)**	0.3851 (0.0179)	0.9604 (0.0096)	14.9	0.000
	biased	**0.9790 (0.0185)**	0.3292 (0.0241)	0.9032 (0.0340)	10.7	0.000

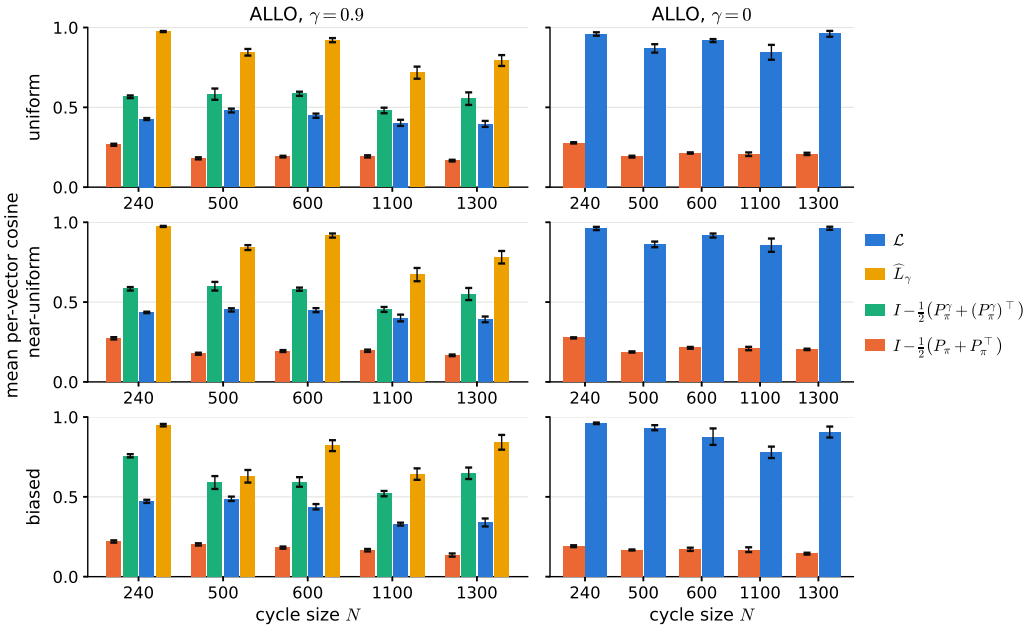

Figure 2: ALLO's spectral target on directed cycles using oracle recovery, reported as mean ± one standard deviation over 30 seeds. At $\gamma = 0.9$, ALLO aligns most closely with $\widehat{L}_\gamma$; at $\gamma = 0$, it aligns with the Chung eigensystem $\mathcal{L} = D_\pi^{1/2}(I - P_{\text{alt}})D_\pi^{-1/2}$. Eigenvectors of $\widehat{L}_\gamma$ are compared in the same Euclidean coordinates via $D_\pi^{1/2}\widehat{L}_\gamma D_\pi^{-1/2}$.

What Does ALLO Learn? For $\gamma > 0$, ALLO's forward sampler has population target $\widehat{L}_\gamma := I - \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\text{rev}})^\gamma)$ (Corollary 1), which differs from the discounted Chung operator $\widetilde{L}_{\text{alt}} = I - P_{\text{alt}}^\gamma$ on non-reversible chains (Proposition 3). Figure 2 supports this prediction: ALLO at $\gamma = 0.9$ aligns with $\widehat{L}_\gamma$, whereas the one-step control $\gamma = 0$ aligns with Chung, as predicted by Lemma 1.

Eigensystem Recovery. DALLO's diagonal dual variables recover discounted eigenvalues (Theorem 1); applying the inverse spectral map in equation 4 recovers the undiscounted eigenvalues. Figure 5 summarizes all directed runs: DALLO is typically closer to Chung eigenvectors than ALLO is to its sampling-induced target $\widehat{L}_\gamma$, and has lower eigenvalue error in every run. Full results are in Tables 14–19.

Effect of Discounting. Larger γ generally improves early-training recovery on CycleNets and often improves final recovery, consistent with the spectral-conditioning analysis. Full learning curves and additional results are reported in Appendix L.

6 Conclusion

We studied Laplacian representation learning for directed, non-reversible Markov chains, where multi-step sampling can alter the spectral target. We showed that the Chung Laplacian is learnable in stationary-weighted coordinates without explicit stationary probabilities, and that forward geometric sampling generally discounts before reversibilization, producing a different eigenspace on non-reversible chains. DALLO corrects this mismatch by combining forward and time-reversed transitions within ALLO's augmented-Lagrangian framework, with local stability guarantees under a simple-spectrum assumption.

Experiments confirm the distinction: DALLO matches ALLO on reversible settings, improves Chung recovery on nearly all directed settings at the same discount, and outperforms one-step ALLO throughout the directed benchmarks. Larger γ generally improves conditioning and convergence at the cost of longer walks. Future work will study when the difference between Chung- and $\widehat{L}_\gamma$-based representations matters for downstream control.

REPRODUCIBILITY STATEMENT

The supplementary material contains the complete code used for all experiments, together with a README listing the commands required to reproduce them. It includes DALLO, all benchmark environments and policies, the evaluation code, and the ALLO baseline. The ALLO baseline is based on the public code of Gomez et al. (2024), released under the MIT license; we added only the directed-graph environments and compatibility updates for current JAX and Gymnasium versions, leaving ALLO's objective, updates, and hyperparameters unchanged.

All hyperparameters are stored in configuration files, and the 30 seeds are listed in a file shared by both methods. Batch scripts reproduce every reported experimental configuration from scratch. Separate scripts recompute the buffer-based density estimates $\hat{d}^{\pi}$ and regenerate all tables and figures from the resulting saved runs. Unless stated otherwise, reported recovery results use the final checkpoint after 200,000 training steps; the discount study additionally reports intermediate checkpoints. Ground-truth eigenvectors are computed from the exact transition matrices and are used only for evaluation, never for training.

The appendices provide proofs for all formal statements in the paper.

Given the same seed and software versions, DALLO reproduces the stored results exactly. ALLO GPU runs are not bitwise deterministic under JAX and may exhibit small run-to-run numerical variation.

AI USE STATEMENT

We used large language models as coding, analysis, and writing assistants. They helped write scripts for generating the directed benchmark environments and checking their structural and spectral properties; scripts to dispatch and monitor the GPU runs; and the analysis pipeline that computes the reported metrics (per-vector cosine to the Chung eigenvectors at the final checkpoint, seed statistics, t-tests, and the stationary-distribution estimators used in the appendix) and assembles the result tables. They also assisted in drafting and editing the text. All LLM-generated code and reported results were reviewed and independently verified by the authors. The research question, the DALLO objective, the theoretical results and proofs, and all scientific decisions are the authors' own. The authors take full responsibility for the content of this submission.

REFERENCES

Sheldon Axler. *Linear Algebra Done Right*. Springer, 4 edition, 2024.

Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9 (1):1–19, 2005.

James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The Annals of Applied Probability*, 1(1):62–87, 1991.

Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. EigenGame: PCA as a Nash equilibrium. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Tommaso Giorgi, Pierriccardo Olivieri, Keyue Jiang, Laura Toni, and Matteo Papini. Impact of connectivity on Laplacian representations in reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.08558.

Diego Gomez, Michael Bowling, and Marlos C. Machado. Proper Laplacian representation learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Jeff Johns and Sridhar Mahadevan. Constructing basis functions from directed graphs for value function approximation. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 385–392, 2007.

Yehuda Koren. On spectral graph drawing. In *International Computing and Combinatorics Conference (COCOON)*, pp. 496–508, 2003.

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 (ICML)*, pp. 2295–2304, 2017.

Sridhar Mahadevan. Proto-value functions: Developmental reinforcement learning. *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 553–560, 2005.

Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8:2169–2231, 2007.

Marina Meila and William Pentney. Clustering by weighted cuts in directed graphs. In *Proceedings of the Seventh SIAM International Conference on Data Mining*, pp. 135–144. SIAM, 2007.

Eddie Seabrook and Laurenz Wiskott. A tutorial on the spectral theory of markov chains. *Neural Comput.*, 35(11):1713–1796, 2023.

Dikshant Shehmar, Matthew Schlegel, Matthew E. Taylor, and Marlos C. Machado. Laplacian representations for decision-time planning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. Design principles of the hippocampal cognitive map. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2528–2536, 2014.

Kaixin Wang, Kuangqi Zhou, Qixin Zhang, Jie Shao, Bryan Hooi, and Jiashi Feng. Towards better Laplacian representation in reinforcement learning with generalized graph drawing. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11003–11012, 2021.

Yifan Wu, George Tucker, and Ofir Nachum. The Laplacian in RL: Learning representations with efficient approximations. In *International Conference on Learning Representations (ICLR)*, 2019.

A Direct optimization in Chung coordinates

The change of coordinates $g = D_\pi^{1/2} f$ is not necessary for the exact spectral problem: with known stationary probabilities, one could optimize g directly. Its importance is computational: parameterizing f removes explicit inverse-density weights and the need to estimate d^π when forming the training objective and constraints.

Under stationary sampling, direct optimization of g uses

$$\mathcal{E}_{\mathcal{L}}(g) = \frac{1}{2}\mathbb{E}_{\substack{s\sim d^\pi \\ s'\sim P_\pi(\cdot|s)}}\left[\left(\frac{g(s)}{\sqrt{d^\pi(s)}} - \frac{g(s')}{\sqrt{d^\pi(s')}}\right)^2\right],$$

$$\langle g_j, g_i\rangle = \mathbb{E}_{s\sim d^\pi}\left[\frac{g_j(s)g_i(s)}{d^\pi(s)}\right].$$

These estimators are unbiased with exact d^π, but contain inverse-probability factors in both the energy and the constraints.

For example, consider an off-diagonal constraint ($i \neq j$):

$$Z_{ij}(s) = \frac{g_j(s)g_i(s)}{d^\pi(s)}, \qquad \frac{\partial Z_{ij}(s)}{\partial g_i(s)} = \frac{f_j(s)}{\sqrt{d^\pi(s)}}.$$

Taking Euclidean gradients with respect to the full tabular output vectors, and assuming $\|f_j\|_{D_\pi} = 1$, gives

$$\mathbb{E}_{s\sim d^\pi}\left[\|\nabla_{g_i} Z_{ij}(s)\|_2^2\right] = \sum_s \frac{g_j(s)^2}{d^\pi(s)} = \sum_s f_j(s)^2,$$

$$\mathbb{E}_{s\sim d^\pi}\left[\|\nabla_{f_i}(f_j(s)f_i(s))\|_2^2\right] = \sum_s d^\pi(s) f_j(s)^2 = 1.$$

Thus, inverse-density weighting can amplify contributions from rare states. This is a comparison of coordinate-dependent output-space moments, not a general guarantee of faster neural-network optimization.

Unknown stationary probabilities introduce an additional issue. For a fixed positive estimate $\hat{d}^\pi$, the plug-in constraint estimator satisfies

$$\mathbb{E}_{s\sim d^\pi}\left[\frac{g_j(s)g_i(s)}{\hat{d}^\pi(s)}\right] = \sum_s \frac{d^\pi(s)}{\hat{d}^\pi(s)} g_j(s)g_i(s),$$

which generally differs from the desired Euclidean inner product. Underestimating stationary probabilities can produce large multiplicative distortions; zero estimates make the estimator undefined without smoothing. The energy likewise depends on inverse square roots of the estimated probabilities.

In f-coordinates, the sampled energy and constraint are simply $\frac{1}{2}(f(s) - f(s'))^2$ and $f_j(s)f_i(s)$. Their expectations equal $\mathcal{E}_{\mathcal{L}}(D_\pi^{1/2} f)$ and $\langle f_j, f_i\rangle_{D_\pi}$, respectively, without evaluating d^π during training. Stationary probabilities, or their estimates, are needed only to construct the post-training Chung readout $g = D_\pi^{1/2} f$. This removes explicit inverse-density weighting, but not the difficulty of sampling and representing rarely visited states.

B Spectral conditioning of self-adjoint Markov operators

Let P be a stochastic matrix that is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{D_\pi}$, and fix $\gamma \in [0, 1)$. By the spectral theorem (Axler, 2024, Theorem 7.29), there exists a D_π-orthonormal eigenbasis $(u_i)_{i=1}^{|\mathcal{S}|}$ with $Pu_i = \alpha_i u_i$. Since P is stochastic and self-adjoint, $\alpha_i \in [-1, 1]$. Hence $|\gamma\alpha_i| \le \gamma < 1$, so the matrix power series converges and gives

$$P^\gamma = (1-\gamma)\sum_{t\ge 1} \gamma^{t-1} P^t = (1-\gamma)P(I-\gamma P)^{-1}.$$

Applying this identity to each eigenvector gives

$$P^\gamma u_i = \frac{(1-\gamma)\alpha_i}{1-\gamma\alpha_i}\, u_i.$$

Thus, P^γ has the same D_π-orthonormal eigenbasis as P, with real eigenvalues, and is therefore self-adjoint in $\langle\cdot,\cdot\rangle_{D_\pi}$.

Define $L := I - P$ and $\widetilde{L} := I - P^\gamma$. Then $Lu_i = \lambda_i u_i$, where $\lambda_i := 1 - \alpha_i \in [0, 2]$, and

$$\widetilde{L}u_i = (I - P^\gamma)u_i = \left(1 - \frac{(1-\gamma)\alpha_i}{1-\gamma\alpha_i}\right) u_i = \frac{\lambda_i}{1-\gamma+\gamma\lambda_i}\, u_i = h^\gamma(\lambda_i)u_i.$$

Here, h^γ is the spectral map defined in equation 4. For $\lambda \in [0, 2]$,

$$\frac{d}{d\lambda}h^\gamma(\lambda) = \frac{1-\gamma}{(1-\gamma+\gamma\lambda)^2} > 0, \qquad h^\gamma(0) = 0.$$

Therefore, L and $\widetilde{L}$ have the same ordered eigenspaces, including multiplicities and the zero eigenspace. Likewise, P and P^γ have the same eigenspaces.

Recall from equation 4 that $\lambda^\gamma = h^\gamma(\lambda) = \lambda/(1-\gamma+\gamma\lambda)$. Solving for λ gives

$$\lambda = \frac{(1-\gamma)\lambda^\gamma}{1-\gamma\lambda^\gamma}, \qquad 0 \le \lambda^\gamma \le \frac{2}{1+\gamma}.$$

For $\gamma > 0$, the derivative at zero is $1/(1-\gamma) > 1$. Thus, discounting expands eigenvalue gaps sufficiently close to zero, although it does not enlarge all spectral gaps. At $\gamma = 0$, $P^\gamma = P$ and $\widetilde{L} = L$.

C THE CHUNG LAPLACIAN AND THE WEIGHTED PROBLEM

C.1 PROOF OF PROPOSITION 1

Proof. If d^π is uniform, then $D_\pi = |\mathcal{S}|^{-1} I$.

Substituting this into the definition of the Chung Laplacian 5 gives

$$\mathcal{L} = I - \frac{1}{2}\left((|\mathcal{S}|^{-1/2} I) P_\pi (|\mathcal{S}|^{1/2} I) + (|\mathcal{S}|^{1/2} I) P_\pi^\top (|\mathcal{S}|^{-1/2} I)\right) = I - \frac{1}{2}(P_\pi + P_\pi^\top).$$

For the reverse implication, suppose that d^π is nonuniform and let

$$\mathcal{M} := \{s \in \mathcal{S} : d^\pi(s) = \max_{x \in \mathcal{S}} d^\pi(x)\}.$$

Suppose, for contradiction, that $\mathcal{L} = I - \frac{1}{2}(P_\pi + P_\pi^\top)$. Equivalently,

$$I - \frac{1}{2}\left(D_\pi^{1/2} P_\pi D_\pi^{-1/2} + D_\pi^{-1/2} P_\pi^\top D_\pi^{1/2}\right) = I - \frac{1}{2}(P_\pi + P_\pi^\top).$$

$s \neq s'$, and therefore the (s, s')-entry of the identity matrix is zero. Hence, comparing the (s, s')-entries of the two operators and cancelling the common factor, we obtain $\left(D_\pi^{1/2} P_\pi D_\pi^{-1/2} + D_\pi^{-1/2} P_\pi^\top D_\pi^{1/2}\right)_{s,s'} = (P_\pi + P_\pi^\top)_{s,s'}$. Equivalently,

$$\sqrt{\frac{d^\pi(s)}{d^\pi(s')}} P_\pi(s, s') + \sqrt{\frac{d^\pi(s')}{d^\pi(s)}} P_\pi(s', s) = P_\pi(s, s') + P_\pi(s', s).$$

Multiplying by $\sqrt{d^\pi(s) d^\pi(s')}$ and rearranging gives

$$\left(\sqrt{d^\pi(s)} - \sqrt{d^\pi(s')}\right)\left(\sqrt{d^\pi(s)} P_\pi(s, s') - \sqrt{d^\pi(s')} P_\pi(s', s)\right) = 0.$$

Since $s \in \mathcal{M}$ and $s' \notin \mathcal{M}$, $d^\pi(s) > d^\pi(s')$, and therefore $\sqrt{d^\pi(s)} P_\pi(s, s') = \sqrt{d^\pi(s')} P_\pi(s', s)$.

By $d^\pi(s') < \sqrt{d^\pi(s) d^\pi(s')}$, $d^\pi(s') P_\pi(s', s) = \sqrt{d^\pi(s) d^\pi(s')} P_\pi(s, s') \leq d^\pi(s) P_\pi(s, s')$, with strict inequality whenever $P_\pi(s, s') > 0$. Since $\mathcal{M}$ is a nonempty proper subset of $\mathcal{S}$ and the chain is irreducible, there exist $s \in \mathcal{M}$ and $s' \notin \mathcal{M}$ such that $P_\pi(s, s') > 0$. Therefore,

$$\sum_{\substack{s \in \mathcal{M} \\ s' \notin \mathcal{M}}} d^\pi(s) P_\pi(s, s') > \sum_{\substack{s \in \mathcal{M} \\ s' \notin \mathcal{M}}} d^\pi(s') P_\pi(s', s).$$

On the other hand, stationarity implies that the total probability flow from $\mathcal{M}$ to its complement equals the flow from the complement into $\mathcal{M}$. Indeed, since $\sum_{s' \in \mathcal{S}} P_\pi(s, s') = 1$, the stationary mass of $\mathcal{M}$ decomposes as

$$d^\pi(\mathcal{M}) = \sum_{s, s' \in \mathcal{M}} d^\pi(s) P_\pi(s, s') + \sum_{\substack{s \in \mathcal{M} \\ s' \notin \mathcal{M}}} d^\pi(s) P_\pi(s, s').$$

Similarly, using the stationarity relation $d^\pi P_\pi = d^\pi$,

$$d^\pi(\mathcal{M}) = \sum_{s, s' \in \mathcal{M}} d^\pi(s') P_\pi(s', s) + \sum_{\substack{s \in \mathcal{M} \\ s' \notin \mathcal{M}}} d^\pi(s') P_\pi(s', s).$$

The first sums on the right-hand sides are equal after relabeling the indices. Hence,

$$\sum_{\substack{s \in \mathcal{M} \\ s' \notin \mathcal{M}}} d^\pi(s) P_\pi(s, s') = \sum_{\substack{s \in \mathcal{M} \\ s' \notin \mathcal{M}}} d^\pi(s') P_\pi(s', s),$$

contradicting the strict inequality above. □

C.2 Proof of Proposition 2

Proof. By the definition of $\mathcal{W}$, $\mathcal{E}_{\mathcal{W}}(f) = f^\top \mathcal{W} f = f^\top D_\pi^{1/2} \mathcal{L} D_\pi^{1/2} f = g^\top \mathcal{L} g = \mathcal{E}_{\mathcal{L}}(g)$. Moreover, $\mathcal{W} f = \lambda D_\pi f$ is equivalent to $D_\pi^{1/2} \mathcal{L} g = \lambda D_\pi^{1/2} g$, which, since $D_\pi^{1/2}$ is invertible, is equivalent to $\mathcal{L} g = \lambda g$. □

Next, we verify the sampled forms quoted in Section 4.1. The constraint identity $\langle f_j, f_i \rangle_{D_\pi} = \mathbb{E}_{s \sim d^\pi}[f_j(s) f_i(s)]$ is the definition of the inner product, written as an expectation. Under stationary sampling, a buffer transition (s, s') has joint law $\Pr[s, s'] = d^\pi(s) P_\pi(s, s')$, so expanding the square,

$$\tfrac{1}{2}\,\mathbb{E}\big[(f(s) - f(s'))^2\big] = \tfrac{1}{2} \sum_{s,s'} d^\pi(s) P_\pi(s, s') \big(f(s)^2 - 2 f(s) f(s') + f(s')^2\big),$$

and we evaluate the three sums separately. Using $P_\pi \mathbf{1} = \mathbf{1}$,

$$\sum_{s,s'} d^\pi(s) P_\pi(s, s') f(s)^2 = \sum_s d^\pi(s) f(s)^2 = f^\top D_\pi f.$$

Using stationarity, $P_\pi^\top d^\pi = d^\pi$,

$$\sum_{s,s'} d^\pi(s) P_\pi(s, s') f(s')^2 = \sum_{s'} d^\pi(s') f(s')^2 = f^\top D_\pi f.$$

Finally,

$$\sum_{s,s'} f(s) d^\pi(s) P_\pi(s, s') f(s') = f^\top D_\pi P_\pi f = \frac{1}{2} f^\top \big(D_\pi P_\pi + P_\pi^\top D_\pi\big) f.$$

where the last equality follows by transposition. Combining the three sums,

$$\tfrac{1}{2}\,\mathbb{E}\big[(f(s) - f(s'))^2\big] = f^\top D_\pi f - \tfrac{1}{2}\, f^\top \big(D_\pi P_\pi + P_\pi^\top D_\pi\big) f = f^\top \mathcal{W} f$$

by definition of $\mathcal{W}$. Substituting $f = D_\pi^{-1/2} g$ gives the Dirichlet form equation 6.

D The reversible surrogate

We record the elementary properties of P_{alt} quoted in Section 4.2. The reversal $P_\pi^{\mathrm{rev}} = D_\pi^{-1} P_\pi^\top D_\pi$ is the adjoint of P_π in $\langle \cdot, \cdot \rangle_{D_\pi}$, since

$$\langle P_\pi f, h \rangle_{D_\pi} = f^\top P_\pi^\top D_\pi h = f^\top D_\pi P_\pi^{\mathrm{rev}} h = \langle f, P_\pi^{\mathrm{rev}} h \rangle_{D_\pi}.$$

It is stochastic ($P_\pi^{\mathrm{rev}} \mathbf{1} = D_\pi^{-1} P_\pi^\top d^\pi = D_\pi^{-1} d^\pi = \mathbf{1}$) and leaves d^π stationary ($(d^\pi)^T P_\pi^{\mathrm{rev}} = \mathbf{1}^\top D_\pi P_\pi^{\mathrm{rev}} = \mathbf{1}^\top P_\pi^\top D_\pi = (d^\pi)^T$), hence so does $P_{\mathrm{alt}} = \frac{1}{2}(P_\pi + P_\pi^{\mathrm{rev}})$. If $P_\pi^{\mathrm{rev}} = P_\pi^\top$, then $d^\pi(s) = d^\pi(s')$ on every transition with $P_\pi(s', s) > 0$, and irreducibility makes d^π uniform, which is the caveat quoted after the definition of the reversal. Reversibility of P_{alt},

$$d^\pi(s) P_{\mathrm{alt}}(s, s') = d^\pi(s') P_{\mathrm{alt}}(s', s),$$

follows entrywise from $d^\pi(s) P_\pi^{\mathrm{rev}}(s, s') = d^\pi(s') P_\pi(s', s)$, and is equivalent to $D_\pi P_{\mathrm{alt}} = P_{\mathrm{alt}}^\top D_\pi$, that is, to self-adjointness of P_{alt} in $\langle \cdot, \cdot \rangle_{D_\pi}$, and to Euclidean symmetry of the conjugated matrix $D_\pi^{1/2} P_{\mathrm{alt}} D_\pi^{-1/2}$:

$$\Big(D_\pi^{1/2} P_{\mathrm{alt}} D_\pi^{-1/2}\Big)^\top = D_\pi^{-1/2} P_{\mathrm{alt}}^\top D_\pi^{1/2} = D_\pi^{1/2} P_{\mathrm{alt}} D_\pi^{-1/2}.$$

Multiplying $\mathcal{W} = D_\pi (I - P_{\mathrm{alt}})$ by $D_\pi^{-1/2}$ on both sides gives the conjugated form used below,

$$\mathcal{L} = D_\pi^{-1/2}\, \mathcal{W}\, D_\pi^{-1/2} = I - D_\pi^{1/2} P_{\mathrm{alt}} D_\pi^{-1/2}.$$

E PROOFS OF THE SAMPLING RESULTS

Throughout this subsection, $\langle f, h\rangle_{D_\pi} = f^\top D_\pi h$, and expectations are over $s \sim d^\pi$ and $s' \sim Q(\cdot \mid s)$ unless stated otherwise. For a general operator X, X^* denotes its adjoint in this inner product; for Markov kernels, we write Q^{rev} for the same adjoint. We use repeatedly that P_π, P_π^{rev}, and P_{alt} are stochastic and preserve d^π (Appendix D), and that time reversal commutes with powers, $(P_\pi^k)^{\mathrm{rev}} = (P_\pi^{\mathrm{rev}})^k$.

E.1 PROOF OF LEMMA 1

Proof. Expanding the square gives

$$\tfrac{1}{2}\,\mathbb{E}[(f(s) - f(s'))^2] = \tfrac{1}{2}\,\mathbb{E}[f(s)^2] + \tfrac{1}{2}\,\mathbb{E}[f(s')^2] - \mathbb{E}[f(s)f(s')].$$

Because d^π is stationary for Q, the marginal distribution of s' is also d^π. Therefore,

$$\mathbb{E}[f(s)^2] = \mathbb{E}[f(s')^2] = \langle f, f\rangle_{D_\pi}, \qquad \mathbb{E}[f(s)f(s')] = \langle f, Qf\rangle_{D_\pi}.$$

Hence,

$$\tfrac{1}{2}\,\mathbb{E}[(f(s) - f(s'))^2] = \langle f, f\rangle_{D_\pi} - \langle f, Qf\rangle_{D_\pi} = \langle f, (I - Q)f\rangle_{D_\pi}.$$

Since Q^{rev} is the adjoint of Q,

$$\langle f, Q^{\mathrm{rev}} f\rangle_{D_\pi} = \langle Qf, f\rangle_{D_\pi} = \langle f, Qf\rangle_{D_\pi},$$

where the last equality follows from symmetry of the real inner product. Thus,

$$\langle f, Qf\rangle_{D_\pi} = \left\langle f, \tfrac{1}{2}(Q + Q^{\mathrm{rev}})f\right\rangle_{D_\pi}.$$

Substituting this identity above yields

$$\tfrac{1}{2}\,\mathbb{E}[(f(s) - f(s'))^2] = \left\langle f, \left(I - \tfrac{1}{2}(Q + Q^{\mathrm{rev}})\right) f\right\rangle_{D_\pi}.$$

□

E.2 PROOF OF COROLLARY 1

Proof. Averaging over K gives

$$\sum_{k\geq 1} \mathbb{P}(K = k) P_\pi^k = (1 - \gamma) \sum_{k\geq 1} \gamma^{k-1} P_\pi^k = P_\pi^\gamma.$$

Moreover, P_π^γ preserves d^π, and its adjoint with respect to $\langle\cdot,\cdot\rangle_{D_\pi}$ is $(P_\pi^{\mathrm{rev}})^\gamma$. Therefore, applying Lemma 1 with $Q = P_\pi^\gamma$ yields

$$\tfrac{1}{2}\mathbb{E}[(f(s) - f(s'))^2] = \left\langle f, \left(I - \tfrac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)\right) f\right\rangle_{D_\pi},$$

which is the claimed identity. □

E.3 PROOF OF PROPOSITION 3

Proof. If P_π is reversible, then $P_\pi = P_\pi^{\mathrm{rev}} = P_{\mathrm{alt}}$, and the claimed equality follows immediately. Conversely, fix $\gamma \in (0, 1)$ and assume that $P_{\mathrm{alt}}^\gamma = \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)$. We will show that P_π is reversible. Set

$$B := I - \gamma P_{\mathrm{alt}}, \qquad C := \tfrac{\gamma}{2}(P_\pi - P_\pi^{\mathrm{rev}}).$$

These matrices separate $I - \gamma P_\pi$ into its self-adjoint and skew-adjoint parts:

$$I - \gamma P_\pi = B - C, \qquad I - \gamma P_\pi^{\mathrm{rev}} = B + C.$$

Here, adjoints are taken with respect to $\langle\cdot,\cdot\rangle_{D_\pi}$. Since P_π^{rev} is the adjoint of P_π, C is skew-adjoint. Moreover, P_{alt} is self-adjoint and stochastic, so its eigenvalues lie in $[-1, 1]$. Thus, the eigenvalues of B are at least $1 - \gamma > 0$, and both B and B^{-1} are positive definite.

Recall from equation 3 that, for any stochastic matrix Q, $Q^\gamma = (1-\gamma)Q(I-\gamma Q)^{-1}$. The inverse exists because $\gamma < 1$ and Q is stochastic. Since

$$(I-\gamma Q)^{-1} - I = \big(I - (I-\gamma Q)\big)(I-\gamma Q)^{-1} = \gamma Q(I-\gamma Q)^{-1},$$

we can equivalently write

$$Q^\gamma = \frac{1-\gamma}{\gamma}\left((I-\gamma Q)^{-1} - I\right).$$

Applying this identity to P_{alt} gives

$$P_{\mathrm{alt}}^\gamma = \frac{1-\gamma}{\gamma}(B^{-1} - I).$$

Similarly, since $I - \gamma P_\pi = B - C$ and $I - \gamma P_\pi^{\mathrm{rev}} = B + C$,

$$\begin{aligned}\tfrac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma) &= \frac{1-\gamma}{2\gamma}\left[\left((B-C)^{-1} - I\right) + \left((B+C)^{-1} - I\right)\right] \\ &= \frac{1-\gamma}{\gamma}\left[\tfrac{1}{2}\left((B-C)^{-1} + (B+C)^{-1}\right) - I\right].\end{aligned}$$

Thus, the assumed equality $P_{\mathrm{alt}}^\gamma = \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)$ implies

$$B^{-1} = \tfrac{1}{2}\left((B-C)^{-1} + (B+C)^{-1}\right),$$

after cancelling the common nonzero factor $(1-\gamma)/\gamma$ and adding I.

To simplify the right-hand side, observe that

$$\begin{aligned}(B-C)^{-1}B(B+C)^{-1} &= \tfrac{1}{2}(B-C)^{-1}\big((B+C) + (B-C)\big)(B+C)^{-1} \\ &= \tfrac{1}{2}\left((B-C)^{-1} + (B+C)^{-1}\right).\end{aligned}$$

This identity does not require B and C to commute. Therefore, $B^{-1} = (B-C)^{-1}B(B+C)^{-1}$.

Multiplying on the left by $B - C$ and on the right by $B + C$ yields $(B-C)B^{-1}(B+C) = B$. Expanding the left-hand side gives

$$(B-C)B^{-1}(B+C) = B + C - C - CB^{-1}C = B - CB^{-1}C.$$

Consequently, $CB^{-1}C = 0$. We now use skew-adjointness of C and positive definiteness of B^{-1} to show that $C = 0$. Since C is skew-adjoint, for any f and g, $\langle Cf, g\rangle_{D_\pi} = -\langle f, Cg\rangle_{D_\pi}$. Taking $g = B^{-1}Cf$, we obtain

$$\langle Cf, B^{-1}Cf\rangle_{D_\pi} = -\langle f, CB^{-1}Cf\rangle_{D_\pi} = 0,$$

where the last equality follows from $CB^{-1}C = 0$. On the other hand, positive definiteness of B^{-1} means that

$$\langle v, B^{-1}v\rangle_{D_\pi} > 0 \qquad \text{for every } v \neq 0.$$

Applying this to $v = Cf$, the preceding equality forces $Cf = 0$. Since f was arbitrary, $C = 0$.

Finally, by the definition of C,

$$0 = C = \tfrac{\gamma}{2}(P_\pi - P_\pi^{\mathrm{rev}}).$$

Because $\gamma > 0$, this implies $P_\pi = P_\pi^{\mathrm{rev}}$, so P_π is reversible. This proves the equivalence. □

Proof of Proposition 4. Let $\mathcal{U}_k := \mathrm{span}\{f_1, \ldots, f_k\}$ and $A := P_\pi - P_\pi^{\mathrm{rev}}$. Since discounting preserves the ordered eigenspaces of $I - P_{\mathrm{alt}}$, the gap $\lambda_k < \lambda_{k+1}$ makes $\mathcal{U}_k$ the unique k-dimensional minimizing eigenspace of the target operator $I - P_{\mathrm{alt}}^\gamma$.

Using the discounted-kernel expansion in equation 3,

$$\frac{1}{2}\left(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma\right) - P_{\mathrm{alt}}^\gamma = (1-\gamma)\sum_{\ell \geq 1}\gamma^{\ell-1}\left[\frac{1}{2}\left(P_\pi^\ell + (P_\pi^{\mathrm{rev}})^\ell\right) - P_{\mathrm{alt}}^\ell\right].$$

The term for $\ell = 1$ vanishes, while

$$\frac{1}{2}\left(P_\pi^2 + (P_\pi^{\mathrm{rev}})^2\right) - P_{\mathrm{alt}}^2 = \frac{1}{4}\left(P_\pi - P_\pi^{\mathrm{rev}}\right)^2 = \frac{1}{4}A^2.$$

Hence

$$\frac{1}{2}\left(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma\right) - P_{\mathrm{alt}}^\gamma = \frac{(1-\gamma)\gamma}{4}A^2 + R_\gamma,$$

where

$$R_\gamma := (1-\gamma)\sum_{\ell\geq 3}\gamma^{\ell-1}\left[\frac{1}{2}\left(P_\pi^\ell + (P_\pi^{\mathrm{rev}})^\ell\right) - P_{\mathrm{alt}}^\ell\right].$$

Thus A^2 is the first non-vanishing term in the lag expansion of the discrepancy between the two orderings, while the remaining terms start at order γ^2. By assumption, there exist $i \leq k < j$ such that $\langle Af_i, Af_j\rangle_{D_\pi} \neq 0$. Since P_π^{rev} is the adjoint of P_π with respect to $\langle\cdot,\cdot\rangle_{D_\pi}$, A is skew-adjoint with respect to $\langle\cdot,\cdot\rangle_{D_\pi}$, and hence $\langle f_j, A^2 f_i\rangle_{D_\pi} = -\langle Af_j, Af_i\rangle_{D_\pi} \neq 0$. Moreover, f_i is an eigenvector of P_{alt}^γ, so

$$\langle f_j, P_{\mathrm{alt}}^\gamma f_i\rangle_{D_\pi} = 0.$$

Consequently,

$$\left\langle f_j, \frac{1}{2}\left(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma\right) f_i\right\rangle_{D_\pi} = -\frac{(1-\gamma)\gamma}{4}\langle Af_j, Af_i\rangle_{D_\pi} + \langle f_j, R_\gamma f_i\rangle_{D_\pi}.$$

It remains to control the remainder. Let $\|f\|_{D_\pi} := \sqrt{\langle f, f\rangle_{D_\pi}}$. For any stochastic kernel Q preserving d^π, Jensen's inequality and stationarity give

$$\|Qf\|_{D_\pi}^2 = \sum_s d^\pi(s)\left(\sum_{s'} Q(s,s')f(s')\right)^2 \leq \sum_{s,s'} d^\pi(s)Q(s,s')f(s')^2 = \|f\|_{D_\pi}^2.$$

Applying this to P_π, P_π^{rev}, and P_{alt}, and using Cauchy–Schwarz and using the D_π-orthonormality of f_i, f_j, we obtain

$$|\langle f_j, R_\gamma f_i\rangle_{D_\pi}| \leq 2(1-\gamma)\sum_{\ell\geq 3}\gamma^{\ell-1} = 2\gamma^2.$$

Therefore the remainder term is $O(\gamma^2)$, whereas

$$-\frac{(1-\gamma)\gamma}{4}\langle Af_j, Af_i\rangle_{D_\pi}$$

has a nonzero coefficient and is of order γ. Hence, for all sufficiently small $\gamma > 0$,

$$\left\langle f_j, \frac{1}{2}\left(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma\right) f_i\right\rangle_{D_\pi} \neq 0.$$

Since $f_i \in \mathcal{U}_k$ and $f_j \in \mathcal{U}_k^\perp$, this shows that $\frac{1}{2}\left(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma\right) f_i \notin \mathcal{U}_k$. Thus $\mathcal{U}_k$ is not invariant under the transition operator underlying the sampled energy, and hence not invariant under $I - \frac{1}{2}\left(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma\right)$. This operator is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{D_\pi}$, so any minimizing spectral subspace is invariant. Therefore $\mathcal{U}_k$ cannot be its k-dimensional minimizing eigenspace. Since $\mathcal{U}_k$ is the unique k-dimensional minimizing eigenspace of the target energy $I - P_{\mathrm{alt}}^\gamma$, the two spectral objectives are minimized by different k-dimensional eigenspaces for all sufficiently small $\gamma > 0$. □

E.4 Proof of Proposition 5

Proof. Conditional on the current state s, averaging over the independent choice of the forward or reverse kernel gives

$$\Pr\left(s^{(t+1)} = s' \mid s^{(t)} = s\right) = \frac{1}{2}P_\pi(s,s') + \frac{1}{2}P_\pi^{\mathrm{rev}}(s,s') = P_{\mathrm{alt}}(s,s').$$

Hence the constructed walk is a Markov chain with transition kernel P_{alt}. Conditioned on $K = k$, the endpoint transition kernel is P_{alt}^k. Averaging over the geometric stopping time and using equation 3 gives

$$(1-\gamma)\sum_{k\geq 1}\gamma^{k-1}P_{\mathrm{alt}}^k = P_{\mathrm{alt}}^\gamma.$$

Since P_{alt} preserves d^π, so does P_{alt}^γ. Moreover, P_{alt}^γ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_{D_\pi}$. Thus, together with $s^{(0)} \sim d^\pi$, Lemma 1 applies with $Q = P_{\mathrm{alt}}^\gamma$ and yields

$$\frac{1}{2}\mathbb{E}\Big[\big(f(s^{(0)}) - f(s^{(K)})\big)^2\Big] = \langle f, (I - P_{\mathrm{alt}}^\gamma) f\rangle_{D_\pi},$$

which is the claimed identity. □

F Recovery analysis

F.1 Variational Characterization and Dual Variables

For brevity, in this appendix we write λ_i for the discounted eigenvalues λ_i^γ.

The following is the weighted form of Lemma 1 of Gomez et al. (2024).

Lemma 2. *Let $\widetilde{L}_{\mathrm{alt}} u_i = \lambda_i u_i$ with $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{|\mathcal{S}|}$, where $(u_i)_{i=1}^{|\mathcal{S}|}$ is a $\langle\cdot,\cdot\rangle_{D_\pi}$-orthonormal eigenbasis. For every $1 \le k \le |\mathcal{S}|$,*

$$\min_{\substack{f_1,\dots,f_k \\ \langle f_i, f_j\rangle_{D_\pi} = \delta_{ij}}} \sum_{i=1}^k \langle f_i, \widetilde{L}_{\mathrm{alt}} f_i\rangle_{D_\pi} = \sum_{i=1}^k \lambda_i,$$

and the tuple $(u_1, \dots, u_k)$ attains the minimum. If $k < |\mathcal{S}|$ and $\lambda_k < \lambda_{k+1}$, a feasible tuple $(f_1, \dots, f_k)$ is a global minimizer if and only if $\mathrm{span}\{f_1, \dots, f_k\} = \mathrm{span}\{u_1, \dots, u_k\}$. Equivalently, every global minimizer is obtained from $(u_1, \dots, u_k)$ by an orthogonal change of basis within this D_π-weighted spectral subspace.

Proof. Let $n := |\mathcal{S}|$. Since $(u_r)_{r=1}^n$ is a $\langle\cdot,\cdot\rangle_{D_\pi}$-orthonormal basis, each f_i admits the expansion

$$f_i = \sum_{r=1}^n c_{ri} u_r, \qquad i = 1, \dots, k.$$

Let $C = (c_{ri}) \in \mathbb{R}^{n\times k}$. The feasibility conditions $\langle f_i, f_j\rangle_{D_\pi} = \delta_{ij}$ imply $C^\top C = I_k$. Set $p_r := \sum_{i=1}^k c_{ri}^2$. Since $CC^\top$ is an orthogonal projection,

$$0 \le p_r \le 1, \qquad \sum_{r=1}^n p_r = k.$$

Using $\widetilde{L}_{\mathrm{alt}} u_r = \lambda_r u_r$ and orthonormality,

$$\sum_{i=1}^k \langle f_i, \widetilde{L}_{\mathrm{alt}} f_i\rangle_{D_\pi} = \sum_{r=1}^n \lambda_r p_r.$$

Since $\sum_{r>k} p_r = \sum_{r=1}^k (1 - p_r)$,

$$\sum_{r=1}^n \lambda_r p_r - \sum_{r=1}^k \lambda_r \ge \lambda_k \left(\sum_{r>k} p_r - \sum_{r=1}^k (1 - p_r)\right) = 0.$$

Thus every feasible tuple satisfies

$$\sum_{i=1}^k \langle f_i, \widetilde{L}_{\mathrm{alt}} f_i\rangle_{D_\pi} \ge \sum_{i=1}^k \lambda_i.$$

The tuple $(u_1, \dots, u_k)$ is feasible and attains equality, proving the first claim.

Now suppose $k < n$ and $\lambda_k < \lambda_{k+1}$. The preceding estimate can be sharpened to

$$\sum_{r=1}^n \lambda_r p_r - \sum_{r=1}^k \lambda_r \ge \lambda_{k+1} \sum_{r>k} p_r - \lambda_k \sum_{r=1}^k (1 - p_r) = (\lambda_{k+1} - \lambda_k) \sum_{r>k} p_r.$$

Hence equality can hold only if $p_r = 0$ for every $r > k$. Since $p_r = \sum_{i=1}^{k} c_{ri}^2$, this implies $c_{ri} = 0$ for every $r > k$ and every i. Thus each f_i belongs to $\mathcal{U}_k := \mathrm{span}\{u_1, \dots, u_k\}$. Because $f_1, \dots, f_k$ are k orthonormal vectors in the k-dimensional space $\mathcal{U}_k$, $\mathrm{span}\{f_1, \dots, f_k\} = \mathcal{U}_k$.

Conversely, any feasible tuple spanning $\mathcal{U}_k$ is an orthonormal basis of this subspace, and therefore

$$\sum_{i=1}^{k} \langle f_i, \widetilde{L}_{\mathrm{alt}} f_i \rangle_{D_\pi} = \mathrm{tr}\left(\widetilde{L}_{\mathrm{alt}}\big|_{\mathcal{U}_k}\right) = \sum_{i=1}^{k} \lambda_i.$$

Hence it is a global minimizer.

Finally, any two $\langle \cdot, \cdot \rangle_{D_\pi}$-orthonormal bases of $\mathcal{U}_k$ differ by an orthogonal change of basis, which proves the final claim. □

The lemma determines the minimizing spectral subspace but not the individual eigenfunctions. To relate this variational characterization to the dual variables of DALLO, we next record the Lagrange multipliers at the eigenbasis minimizer.

Let $\alpha = (\alpha_{jm})_{1 \le m \le j \le k}$ denote the Lagrange multipliers associated with the constraints $\langle f_j, f_m \rangle_{D_\pi} = \delta_{jm}$, $1 \le m \le j \le k$, and define

$$\mathcal{L}(f, \alpha) := \sum_{i=1}^{k} \langle f_i, \widetilde{L}_{\mathrm{alt}} f_i \rangle_{D_\pi} + \sum_{1 \le m \le j \le k} \alpha_{jm} \big(\langle f_j, f_m \rangle_{D_\pi} - \delta_{jm} \big), \tag{10}$$

where $f = (f_1, \dots, f_k)$.

The weighted orthonormality constraints satisfy the linear independence constraint qualification at every feasible tuple, so the KKT conditions apply. With gradients taken with respect to $\langle \cdot, \cdot \rangle_{D_\pi}$, the stationarity conditions are

$$2\widetilde{L}_{\mathrm{alt}} f_i + 2\alpha_{ii} f_i + \sum_{m=1}^{i-1} \alpha_{im} f_m + \sum_{j=i+1}^{k} \alpha_{ji} f_j = 0, \qquad i = 1, \dots, k. \tag{11}$$

We next identify the Lagrange multipliers at the eigenbasis minimizer.

Corollary 3. *For the eigenbasis minimizer $f_i^\star = u_i$, $i = 1, \dots, k$, the Lagrange multipliers associated with equation 10 satisfy*

$$\alpha_{jm}^\star = -\lambda_j \delta_{jm}, \qquad 1 \le m \le j \le k.$$

In particular, $\alpha_{jj}^\star = -\lambda_j$.

Proof. Substituting $f_i^\star = u_i$ and $\widetilde{L}_{\mathrm{alt}} u_i = \lambda_i u_i$ into equation 11 gives

$$2(\lambda_i + \alpha_{ii}^\star) u_i + \sum_{m=1}^{i-1} \alpha_{im}^\star u_m + \sum_{j=i+1}^{k} \alpha_{ji}^\star u_j = 0.$$

Since $u_1, \dots, u_k$ are linearly independent, all coefficients vanish. Therefore $\alpha_{ii}^\star = -\lambda_i$ and all off-diagonal multipliers are zero. □

For the stop-gradient objective equation 9, only one factor of each orthonormality constraint is differentiated. At the eigenbasis minimizer, the stop-gradient differentiates only one factor of each diagonal constraint, so

$$\beta_{ii}^\star = 2\alpha_{ii}^\star = -2\lambda_i,$$

while all off-diagonal dual variables vanish.

F.2 Equilibrium Characterization and Stability Transfer from ALLO

We prove the stability statement of Theorem 1 by reducing the DALLO dynamics to the ALLO dynamics analyzed by Gomez et al. (2024, Theorem 1 and Appendix A.4) and verifying the required assumptions.

The equilibria of equation 9 are characterized as follows, as the weighted counterpart of Lemma 2 of Gomez et al. (2024).

Lemma 3. *For any $b > 0$, a pair $(f^\star, \beta^\star)$ is an equilibrium of the primal–dual gradient dynamics associated with equation 9 if and only if there exist distinct indices $\sigma(1), \dots, \sigma(k) \in \{1, \dots, |\mathcal{S}|\}$ such that*

$$\widetilde{L}_{\mathrm{alt}} f_j^\star = \lambda_{\sigma(j)} f_j^\star, \qquad \langle f_i^\star, f_j^\star \rangle_{D_\pi} = \delta_{ij}, \qquad \beta_{ji}^\star = -2\lambda_{\sigma(j)} \delta_{ji},$$

for $1 \le i \le j \le k$. If the spectrum of $\widetilde{L}_{\mathrm{alt}}$ is simple, then $f_j^\star = \varepsilon_j u_{\sigma(j)}$ with $\varepsilon_j \in \{-1, 1\}$. If an eigenvalue is repeated, the corresponding $f_j^\star$ may be any $\langle \cdot, \cdot \rangle_{D_\pi}$-orthonormal family within its eigenspace.

Proof. At an equilibrium, the dual equations impose

$$\langle f_j^\star, f_i^\star \rangle_{D_\pi} = \delta_{ji}, \qquad 1 \le i \le j \le k.$$

Hence all quadratic-penalty terms vanish in the primal pseudo-gradients, and the stop-gradient stationarity equations reduce to

$$2\widetilde{L}_{\mathrm{alt}} f_j^\star + \sum_{i=1}^{j} \beta_{ji}^\star f_i^\star = 0, \qquad j = 1, \dots, k.$$

For $j = 1$, $2\widetilde{L}_{\mathrm{alt}} f_1^\star + \beta_{11}^\star f_1^\star = 0$, so $f_1^\star$ is an eigenfunction of $\widetilde{L}_{\mathrm{alt}}$.

Now suppose that $f_1^\star, \dots, f_{j-1}^\star$ are eigenfunctions. Taking the D_π-inner product of the j-th stationarity equation with $f_i^\star$, $i < j$, and using self-adjointness of $\widetilde{L}_{\mathrm{alt}}$ and orthonormality gives

$$\beta_{ji}^\star = -2\langle f_i^\star, \widetilde{L}_{\mathrm{alt}} f_j^\star \rangle_{D_\pi} = -2\langle \widetilde{L}_{\mathrm{alt}} f_i^\star, f_j^\star \rangle_{D_\pi} = 0.$$

Therefore

$$2\widetilde{L}_{\mathrm{alt}} f_j^\star + \beta_{jj}^\star f_j^\star = 0,$$

so $f_j^\star$ is also an eigenfunction, with $\beta_{jj}^\star = -2\lambda_{\sigma(j)}$. Induction proves that every equilibrium has the stated form.

Conversely, any D_π-orthonormal family of eigenfunctions with $\beta_{ji}^\star = -2\lambda_{\sigma(j)}\delta_{ji}$ satisfies both the primal and dual equilibrium equations. This proves the characterization. □

Thus, under the stop-gradient, every equilibrium consists of a D_π-orthonormal family of eigenfunctions. In particular, orthogonal mixtures of eigenspaces with distinct eigenvalues are no longer stationary. Lemma 3 characterizes all equilibria, while Theorem 1 determines which selection and ordering is locally stable.

The equilibrium characterization and the stability analysis are the weighted counterparts of the corresponding results for ALLO (Gomez et al., 2024). They transfer under the same change of coordinates used above,

$$g = D_\pi^{1/2} f.$$

Under this transformation, $\langle f_j, f_i \rangle_{D_\pi} = g_j^\top g_i$, and $\widetilde{L}_{\mathrm{alt}}$ becomes the symmetric matrix

$$H := I - D_\pi^{1/2} P_{\mathrm{alt}}^\gamma D_\pi^{-1/2}.$$

Consequently, the objective equation 9 maps term by term onto the ALLO objective equation 2 for H. Let $\mathcal{A}(f, \beta)$ denote the objective in equation 9, and let $\nabla^{(d)}$ denote the gradient with respect to $\langle \cdot, \cdot \rangle_{D_\pi}$. Under the prescribed stop-gradient convention, the associated continuous-time primal–dual dynamics are

$$\dot{f}_i = -\nabla_{f_i}^{(d)} \mathcal{A}(f, \beta), \qquad \dot{\beta}_{ji} = \partial_{\beta_{ji}} \mathcal{A}(f, \beta).$$

To see that the stability analysis transfers as well, define

$$\widehat{\mathcal{A}}(g, \beta) := \mathcal{A}(D_\pi^{-1/2} g, \beta).$$

For each primal variable,

$$\nabla_{g_i} \widehat{\mathcal{A}} = D_\pi^{1/2} \nabla_{f_i}^{(d)} \mathcal{A}.$$

Hence the weighted primal gradient dynamics are mapped exactly to the Euclidean ALLO dynamics, while the dual variables are unchanged. The full primal–dual vector fields are therefore related by an invertible linear change of coordinates. Consequently, the Jacobians at corresponding equilibria are similar and have the same eigenvalues. Thus equilibria, local asymptotic stability, and linear instability are preserved under the transformation.

Applying the stability analysis of Gomez et al. (2024) to the transformed symmetric problem therefore yields Theorem 1. The same sufficient condition $b > 2$ applies. Indeed, P_{alt}^{γ} is stochastic and self-adjoint with respect to $\langle \cdot, \cdot \rangle_{D_\pi}$, so its eigenvalues are real and lie in $[-1, 1]$. Since $D_\pi^{1/2} P_{\text{alt}}^{\gamma} D_\pi^{-1/2}$ is similar to P_{alt}^{γ}, it has the same spectrum, and therefore

$$\sigma(H) \subseteq [0, 2].$$

The stability proof of Gomez et al. (2024) uses this spectral bound to obtain the universal sufficient condition $b > 2$, so the same threshold carries over unchanged.

G Proof of the representation limits

Proof of Theorem 2. Since $k > r$, consider the restrictions of the first $r + 1$ coordinate functions to S_0. These restrictions form $r + 1$ vectors in $\mathbb{R}^r$ and are therefore linearly dependent. Hence there exist coefficients $a_1, \ldots, a_{r+1}$, not all zero, such that

$$\sum_{i=1}^{r+1} a_i \phi_i(s) = 0 \qquad \text{for every } s \in S_0.$$

Normalize these coefficients and extend the resulting vector to $\mathbb{R}^k$ by setting the remaining coordinates to zero. Thus,

$$a := (a_1, \ldots, a_{r+1}, 0, \ldots, 0) \in \mathbb{R}^k, \qquad \sum_{i=1}^{r+1} a_i^2 = \|a\|_2^2 = 1.$$

Define

$$\psi(s) := \sum_{i=1}^{k} a_i \phi_i(s).$$

By construction, $\psi(s) = 0$ for every $s \in S_0$. Moreover, by Cauchy–Schwarz, the normalization of a, and the bound $|\phi_i(s)| \le M$, for every $s \in S$,

$$\psi(s)^2 = \left(\sum_{i=1}^{r+1} a_i \phi_i(s) \right)^2 \le \left(\sum_{i=1}^{r+1} a_i^2 \right) \left(\sum_{i=1}^{r+1} \phi_i(s)^2 \right) \le (r+1) M^2.$$

By the definition of the Gram matrix,

$$a^\top G_\Phi a = \sum_{s \in S} d(s) \psi(s)^2 = \sum_{s \notin S_0} d(s) \psi(s)^2,$$

where the last equality follows from $\psi(s) = 0$ on S_0. Therefore,

$$a^\top G_\Phi a \le (r+1) M^2 \sum_{s \notin S_0} d(s) = (r+1) M^2 \delta \le 1 - \varepsilon.$$

Since $\|a\|_2 = 1$, we also have $a^\top I_k a = 1$, and hence

$$a^\top (I_k - G_\Phi) a \ge \varepsilon.$$

Finally, by Cauchy–Schwarz,

$$\varepsilon \le \left|a^\top (I_k - G_\Phi) a\right| = \left|\sum_{i,j=1}^{k} a_i a_j \big(\delta_{ij} - (G_\Phi)_{ij}\big)\right| \le \sqrt{\sum_{i,j=1}^{k} a_i^2 a_j^2}\, \|G_\Phi - I_k\|_F = \|G_\Phi - I_k\|_F,$$

where the last equality follows from

$$\sum_{i,j=1}^{k} a_i^2 a_j^2 = \left(\sum_{i=1}^{k} a_i^2\right)^2 = 1.$$

Thus,

$$\|G_\Phi - I_k\|_F \ge \varepsilon.$$

□

H Experimental Details

H.1 Training Configuration

DALLO and ALLO use the same network architecture and main numerical optimization hyperparameters. Both use a three-hidden-layer MLP with 256 units per layer, ReLU activations, and an 11-dimensional linear output. Training uses Adam with learning rate 10^{-3}, batch size 1024, no weight decay or gradient clipping, and 200,000 gradient steps. Unless stated otherwise, $\gamma = 0.9$. All main experiments use 30 random seeds; the discount-parameter study uses the first 10 seeds.

Table 2: Shared training configuration used by DALLO and ALLO.

Parameter	Value
Hidden layers	3×256
Activation	ReLU
Representation dimension	11
Optimizer	Adam
Learning rate	10^{-3}
Batch size	1024
Training steps	200,000
Dual step size	10^{-4}
Dual clipping	$[-100, 100]$
Initial barrier coefficient	2
Barrier learning rate	0.01
Maximum barrier coefficient	10^4
Main discount	$\gamma = 0.9$

H.2 Policies

We evaluate three fixed policies, denoted *uniform*, *near-uniform*, and *biased*. The precise construction depends on the environment family.

Grid mazes. The action set is $\mathcal{A} = \{\text{up}, \text{down}, \text{left}, \text{right}\}$, and policy vectors are written in this order. The transition kernel is

$$P(s, s') = \sum_{a \in \mathcal{A}} \pi(a)\, \mathbf{1}[s' = \text{next}(s, a)]\,,$$

where

$$\text{next}(s, a) = \begin{cases} s + \Delta_a, & \text{if } s + \Delta_a \text{ is open,} \\ s, & \text{otherwise.} \end{cases}$$

Thus, the probability of a blocked move is assigned to the self-loop $P(s, s)$ rather than renormalized over the remaining actions. Both methods use the same maze layouts and transition rule.

With $n = |\mathcal{A}| = 4$ and $\epsilon = 0.01$, the three policies are

$$\pi_{\text{unif}} = \left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \frac{1}{n}\right),$$

$$\pi_{\text{near}} = \left(\frac{1}{n} - \epsilon,\ \frac{1}{n} + \epsilon,\ \frac{1}{n} - \epsilon,\ \frac{1}{n} + \epsilon\right),$$

and

$$\pi_{\text{bias}} = (0.2,\ 0.2,\ 0.2,\ 0.4).$$

Directed cycles. At a state i with k outgoing edges, let e_0 denote the ring edge $i \to (i{+}1) \bmod N$, and let $e_1, \ldots, e_{k-1}$ denote the additional outgoing edges, ordered by target index. We use $\epsilon = 0.01$.

Under the uniform policy,

$$\pi_{\text{unif}}(e_j) = \frac{1}{k}, \qquad j = 0, \ldots, k-1.$$

Under the near-uniform policy, the probabilities alternate around the uniform value:

$$\pi_{\text{near}}(e_j) = \frac{1}{k} + (-1)^{j+1}\epsilon, \qquad j = 0, \ldots, 2\lfloor k/2 \rfloor - 1.$$

If k is odd, the final edge e_{k-1} retains probability $1/k$.

Under the biased policy, if $k > 1$, the ring edge receives probability 0.4, while the remaining probability mass is distributed uniformly over the additional outgoing edges:

$$\pi_{\text{bias}}(e_0) = 0.4, \qquad \pi_{\text{bias}}(e_j) = \frac{0.6}{k-1}, \quad j = 1, \ldots, k-1.$$

If $k = 1$, the unique outgoing edge is taken with probability one under all three policies.

CycleNets. Each state lies on a directed ring. Away from bridge sources, the unique outgoing ring edge is taken with probability one. At a bridge source, let e_{ring} denote the ring edge and e_{bridge} the bridge edge. The policy is

$$\pi(e_{\text{bridge}}) = q, \qquad \pi(e_{\text{ring}}) = 1 - q,$$

with

$$q_{\text{unif}} = 0.10, \qquad q_{\text{near}} = q_{\text{unif}} + \epsilon = 0.11, \qquad q_{\text{bias}} = 0.60,$$

where $\epsilon = 0.01$ as above. Thus, the three CycleNet policies differ only in the bridge-transition probability. Under the biased policy, the ring edge therefore receives probability 0.4, as in the directed-cycle construction. The policy denoted *uniform* corresponds to the base CycleNet dynamics and is not uniform over the two outgoing edges at bridge states.

H.3 Environments

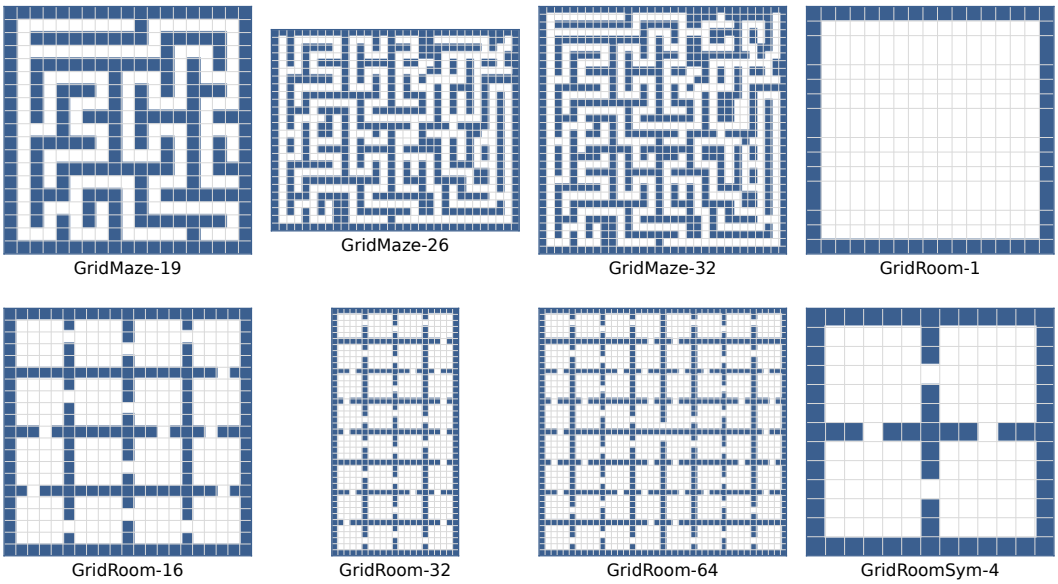

Figure 3: The eight grid environments from Gomez et al. (2024) used in our experiments, with walls shown in blue and accessible cells in white.

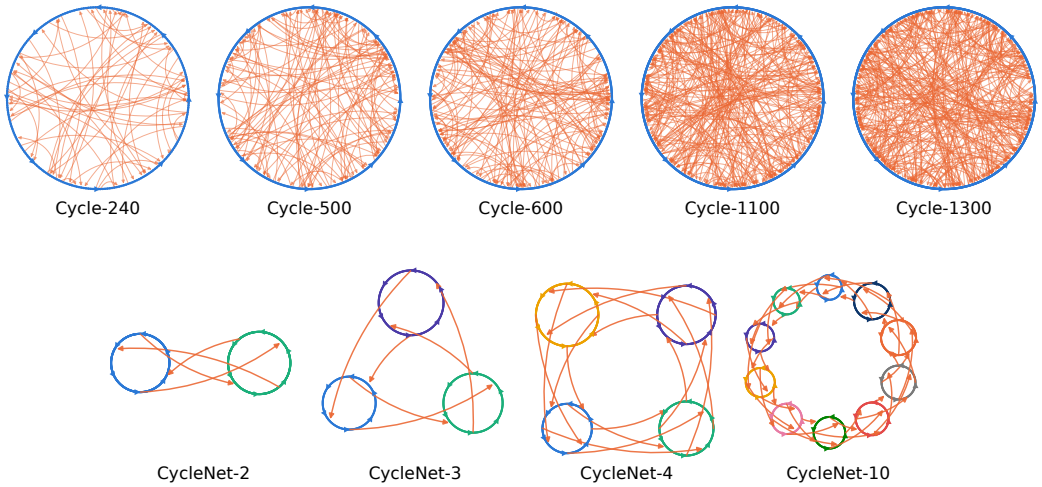

Figure 4: The directed environments used in our experiments. Top: five directed cycles with additional directed edges shown in orange. Bottom: four CycleNets composed of directed rings connected by sparse one-way bridges shown in orange. Arrowheads indicate transition direction.

We evaluate eight grid mazes from the ALLO benchmark (Gomez et al., 2024): GridMaze-19, GridMaze-26, GridMaze-32, GridRoom-1, GridRoom-16, GridRoom-32, GridRoom-64, and GridRoomSym-4 (Figure 3).

A directed Cycle-N, for $N \in \{240, 500, 600, 1100, 1300\}$, consists of the ring $i \to (i+1) \bmod N$ together with $\lfloor N/4 \rfloor$ additional directed edges. These edges are sampled uniformly without replacement, using a fixed random seed, from ordered pairs that are neither self-loops nor existing ring edges. Each graph instance is fixed across all methods, policies, and training seeds.

A CycleNet consists of K directed rings connected by fixed one-way bridges, as summarized in Table 3. All CycleNet instances are deterministic and fixed across methods and seeds.

Table 3: CycleNet benchmark construction.

Environment	States	Ring sizes	Bridges per connection	Topology
CycleNet-2	1500	$700, 800$	2	$R_1 \leftrightarrow R_2$
CycleNet-3	750	$200, 250, 300$	2	$R_1 \to R_2 \to R_3 \to R_1$
CycleNet-4	1300	$250, 300, 350, 400$	4	$R_1 \to \cdots \to R_4 \to R_1$
CycleNet-10	1250	$80, 90, \ldots, 170$	4	$R_1 \to \cdots \to R_{10} \to R_1$

H.4 Replay Buffers and Pair Sampling

DALLO and ALLO use the same replay-buffer budget of 10^6 transitions. DALLO resets every 1000 steps, whereas ALLO retains its released 50-step episodic protocol and resets every 50 steps.

DALLO uses the forward/reverse geometric sampler described in Section 4.3, while ALLO uses its released episodic forward sampler. Exact implementation details are provided in the accompanying code.

H.5 Evaluation

We recover the 11 lowest-frequency Chung eigenvectors. For each method, we consider three readouts:

$$D_\pi^{1/2} f, \qquad \hat{D}_m^{1/2} f, \qquad f,$$

which we refer to as the oracle, practical, and raw readouts, respectively. Here $m \in \{\mathrm{D}, \mathrm{A}\}$ denotes DALLO or ALLO. The corresponding stationary-distribution estimates $\hat{d}_\mathrm{D}$ and $\hat{d}_\mathrm{A}$ are fitted separately to the replay buffer of each method using a density model trained by negative log-likelihood,

with $\hat{D}_m = \mathrm{diag}(\hat{d}_m)$. Thus, the practical readout for DALLO is $\hat{D}_{\mathrm{D}}^{1/2} f$, while the practical readout for ALLO is $\hat{D}_{\mathrm{A}}^{1/2} f$.

Eigenvector recovery is measured by the sign-invariant absolute cosine similarity

$$|\cos(\hat{g}_i, g_i)| = \frac{|\langle \hat{g}_i, g_i \rangle|}{\|\hat{g}_i\|_2 \|g_i\|_2}.$$

The reported per-run score is the arithmetic mean over the 11 recovered vectors, and results are then summarized across seeds by their mean and standard deviation. Final evaluation uses the checkpoint at 200,000 training steps.

For consecutive ground-truth eigenvalues separated by less than 10^{-11}, the corresponding learned vectors are aligned within the repeated-eigenvalue block before computing cosine similarity.

H.6 Statistical Analysis

DALLO and ALLO use independently generated replay buffers and independent training randomness. Comparisons between the methods therefore use Welch's unequal-variance two-sample t-test.

The practical readout is our primary inferential evaluation. Each method is evaluated using the stationary-distribution estimate obtained from its own replay buffer, namely $\hat{d}_{\mathrm{D}}$ for DALLO and $\hat{d}_{\mathrm{A}}$ for ALLO. The oracle and raw readouts are reported as additional diagnostics.

On grid mazes, the statistical comparison is DALLO versus ALLO at $\gamma = 0.9$. On directed cycles and CycleNets, significance is evaluated against the one-step ALLO control at $\gamma = 0$, while ALLO at $\gamma = 0.9$ is reported as an additional comparison reflecting the sampling-induced target.

I EIGENVECTOR RECOVERY

representations under three readouts: the stationary-distribution readout $D_\pi^{1/2} f$, the practical buffer-based readout $\hat{D}_m^{1/2} f$, and the raw network output f, where $m \in \{\mathrm{D}, \mathrm{A}\}$. For the practical readout, each method uses $\hat{d}_m$ estimated from its own replay buffer using the same density-estimation procedure. Results are reported as mean (standard deviation) over 30 seeds per method, with significance assessed by Welch's two-sample t-test.

Eigenvector recovery is measured using the sign-invariant absolute cosine similarity $|\cos(\hat{g}_i, g_i)|$. For numerically repeated eigenvalues ($|\lambda_i - \lambda_j| < 10^{-11}$), we perform an orthogonal Procrustes alignment within the corresponding eigenspace before scoring. The gap between the 11th and 12th Chung eigenvalues exceeds 10^{-6} in every problem, so the $d = 11$ representation never cuts through a repeated eigenspace.

The distinction between D_π and $\hat{D}_\pi$ is relevant because finite replay buffers need not follow the population stationary distribution d^π. The orthogonality constraints encountered during training are instead induced by the empirical visitation distribution in the buffer. We therefore report the oracle, practical, and raw readouts rather than relying on a single evaluation convention.

I.1 REPLAY-BUFFER DISTRIBUTION MISMATCH

Tables 4 and 5 report the distances between $d_{\mathrm{buf},m}$, the true stationary distribution d^π, and the fitted $\hat{d}_m^\pi$, where $m \in \{\mathrm{D}, \mathrm{A}\}$ denotes DALLO and ALLO, respectively.

On the slowly mixing grid environments, the replay-buffer distribution can depart substantially from stationarity under non-uniform visitation. For DALLO, $\mathrm{TV}(d_{\mathrm{buf,D}}, d^\pi)$ reaches 0.696, while for ALLO $\mathrm{TV}(d_{\mathrm{buf,A}}, d^\pi)$ reaches 0.819. Nevertheless, the fitted distributions closely track their respective empirical buffer distributions: $\mathrm{TV}(\hat{d}_{\mathrm{D}}^\pi, d_{\mathrm{buf,D}}) \leq 0.088$ for DALLO and $\mathrm{TV}(\hat{d}_{\mathrm{A}}^\pi, d_{\mathrm{buf,A}}) \leq 0.085$ for ALLO across all settings. These results support the interpretation that the practical readouts based on $\hat{d}_m^\pi$, $m \in \{\mathrm{D}, \mathrm{A}\}$, reflect the finite-buffer geometry under which each representation is trained.

The directed environments generally exhibit smaller deviations from stationarity than the strongly non-uniform grid settings. This effect is particularly pronounced for DALLO, for which $\mathrm{TV}(d_{\mathrm{buf,D}}, d^\pi) \leq 0.023$ across all directed environment-policy settings.

Table 4: Replay-buffer distribution mismatch for DALLO. $d_{\mathrm{buf,D}}$ denotes the empirical state distribution of each replay buffer, $\hat{d}_{\mathrm{D}}^\pi$ the fitted density estimate, and d^π the true stationary distribution. Entries report total-variation distances as mean (s.d.) over 30 seeds.

Env	Policy	$\mathrm{TV}(d_{\mathrm{buf,D}}, d^\pi)$	$\mathrm{TV}(\hat{d}_{\mathrm{D}}^\pi, d_{\mathrm{buf,D}})$	$\mathrm{TV}(\hat{d}_{\mathrm{D}}^\pi, d^\pi)$
GridMaze-19	uniform	0.031 (0.005)	0.011 (0.002)	0.028 (0.006)
	near-uniform	0.151 (0.013)	0.016 (0.004)	0.153 (0.015)
	biased	0.485 (0.014)	0.034 (0.008)	0.487 (0.025)
GridMaze-26	uniform	0.048 (0.005)	0.020 (0.002)	0.042 (0.006)
	near-uniform	0.294 (0.013)	0.033 (0.005)	0.291 (0.019)
	biased	0.525 (0.015)	0.072 (0.012)	0.528 (0.030)
GridMaze-32	uniform	0.055 (0.006)	0.025 (0.003)	0.047 (0.006)
	near-uniform	0.343 (0.015)	0.037 (0.005)	0.341 (0.016)
	biased	0.532 (0.012)	0.077 (0.012)	0.527 (0.026)
GridRoom-1	uniform	0.014 (0.002)	0.007 (0.001)	0.012 (0.002)
	near-uniform	0.021 (0.004)	0.009 (0.002)	0.021 (0.006)
	biased	0.024 (0.001)	0.012 (0.004)	0.030 (0.006)
GridRoom-16	uniform	0.028 (0.006)	0.011 (0.002)	0.026 (0.007)
	near-uniform	0.157 (0.010)	0.016 (0.002)	0.156 (0.010)
	biased	0.383 (0.012)	0.024 (0.007)	0.382 (0.021)
GridRoom-32	uniform	0.041 (0.005)	0.017 (0.002)	0.037 (0.007)

Table 4, continued.

Env	Policy	$\mathrm{TV}(d_{\mathrm{buf}}, d_\pi)$	$\mathrm{TV}(\hat{d}^\pi, d_{\mathrm{buf}})$	$\mathrm{TV}(\hat{d}^\pi, d_\pi)$
	near-uniform	0.314 (0.015)	0.028 (0.004)	0.310 (0.016)
	biased	0.380 (0.013)	0.038 (0.008)	0.383 (0.023)
GridRoom-64	uniform	0.061 (0.007)	0.028 (0.003)	0.055 (0.009)
	near-uniform	0.432 (0.011)	0.042 (0.005)	0.434 (0.019)
	biased	0.696 (0.013)	0.088 (0.023)	0.699 (0.034)
GridRoomSym-4	uniform	0.013 (0.003)	0.005 (0.002)	0.011 (0.005)
	near-uniform	0.025 (0.008)	0.006 (0.002)	0.026 (0.009)
	biased	0.036 (0.003)	0.010 (0.003)	0.041 (0.005)
Cycle-240	uniform	0.006 (0.001)	0.012 (0.002)	0.014 (0.002)
	near-uniform	0.006 (0.001)	0.011 (0.002)	0.013 (0.002)
	biased	0.006 (0.001)	0.012 (0.002)	0.014 (0.002)
Cycle-500	uniform	0.009 (0.002)	0.014 (0.002)	0.017 (0.003)
	near-uniform	0.009 (0.001)	0.015 (0.002)	0.018 (0.002)
	biased	0.010 (0.002)	0.015 (0.002)	0.019 (0.002)
Cycle-600	uniform	0.009 (0.001)	0.014 (0.002)	0.017 (0.002)
	near-uniform	0.010 (0.001)	0.014 (0.002)	0.018 (0.002)
	biased	0.009 (0.001)	0.014 (0.002)	0.018 (0.002)
Cycle-1100	uniform	0.012 (0.001)	0.015 (0.001)	0.019 (0.002)
	near-uniform	0.012 (0.001)	0.015 (0.002)	0.020 (0.002)
	biased	0.012 (0.001)	0.016 (0.002)	0.021 (0.003)
Cycle-1300	uniform	0.013 (0.001)	0.015 (0.002)	0.020 (0.002)
	near-uniform	0.013 (0.001)	0.015 (0.001)	0.020 (0.002)
	biased	0.013 (0.001)	0.016 (0.002)	0.021 (0.002)
CycleNet-2	uniform	0.013 (0.006)	0.006 (0.001)	0.014 (0.006)
	near-uniform	0.013 (0.006)	0.006 (0.001)	0.014 (0.006)
	biased	0.019 (0.004)	0.007 (0.002)	0.020 (0.004)
CycleNet-3	uniform	0.016 (0.009)	0.004 (0.001)	0.016 (0.009)
	near-uniform	0.016 (0.009)	0.004 (0.001)	0.016 (0.009)
	biased	0.009 (0.002)	0.007 (0.004)	0.012 (0.004)
CycleNet-4	uniform	0.017 (0.007)	0.006 (0.001)	0.016 (0.007)
	near-uniform	0.017 (0.008)	0.006 (0.001)	0.016 (0.007)
	biased	0.012 (0.002)	0.007 (0.002)	0.013 (0.002)
CycleNet-10	uniform	0.023 (0.005)	0.006 (0.001)	0.023 (0.005)
	near-uniform	0.023 (0.005)	0.006 (0.001)	0.023 (0.006)
	biased	0.012 (0.002)	0.006 (0.002)	0.011 (0.002)

Table 5: Replay-buffer distribution mismatch for ALLO. $d_{\mathrm{buf,A}}$ denotes the empirical state distribution of each replay buffer, $\hat{d}^\pi_{\mathrm{A}}$ the fitted density estimate, and d^π the true stationary distribution. Entries report total-variation distances as mean (s.d.) over 30 seeds.

Env	Policy	$\mathrm{TV}(d_{\mathrm{buf,A}}, d^\pi)$	$\mathrm{TV}(\hat{d}^\pi_{\mathrm{A}}, d_{\mathrm{buf,A}})$	$\mathrm{TV}(\hat{d}^\pi_{\mathrm{A}}, d^\pi)$
GridMaze-19	uniform	0.016 (0.001)	0.009 (0.001)	0.012 (0.002)
	near-uniform	0.219 (0.004)	0.017 (0.003)	0.218 (0.007)
	biased	0.665 (0.003)	0.035 (0.009)	0.661 (0.017)
GridMaze-26	uniform	0.024 (0.002)	0.018 (0.001)	0.015 (0.002)
	near-uniform	0.346 (0.003)	0.027 (0.002)	0.345 (0.008)
	biased	0.749 (0.002)	0.081 (0.013)	0.743 (0.017)
GridMaze-32	uniform	0.027 (0.001)	0.020 (0.001)	0.017 (0.002)
	near-uniform	0.369 (0.004)	0.031 (0.002)	0.367 (0.008)
	biased	0.727 (0.003)	0.085 (0.009)	0.726 (0.013)
GridRoom-1	uniform	0.010 (0.001)	0.007 (0.001)	0.007 (0.002)
	near-uniform	0.151 (0.003)	0.007 (0.001)	0.151 (0.004)

Table 5, continued.

Env	Policy	$\mathrm{TV}(d_{\mathrm{buf}}, d_\pi)$	$\mathrm{TV}(\hat{d}^\pi, d_{\mathrm{buf}})$	$\mathrm{TV}(\hat{d}^\pi, d_\pi)$
	biased	0.380 (0.002)	0.010 (0.004)	0.382 (0.008)
GridRoom-16	uniform	0.015 (0.001)	0.010 (0.001)	0.011 (0.002)
	near-uniform	0.244 (0.004)	0.018 (0.003)	0.243 (0.009)
	biased	0.669 (0.003)	0.037 (0.008)	0.665 (0.015)
GridRoom-32	uniform	0.021 (0.001)	0.016 (0.001)	0.014 (0.002)
	near-uniform	0.377 (0.003)	0.028 (0.003)	0.375 (0.007)
	biased	0.668 (0.003)	0.047 (0.006)	0.664 (0.018)
GridRoom-64	uniform	0.030 (0.001)	0.024 (0.001)	0.017 (0.002)
	near-uniform	0.464 (0.003)	0.038 (0.002)	0.462 (0.006)
	biased	0.819 (0.003)	0.075 (0.016)	0.820 (0.019)
GridRoomSym-4	uniform	0.009 (0.001)	0.005 (0.001)	0.007 (0.002)
	near-uniform	0.119 (0.003)	0.008 (0.002)	0.119 (0.005)
	biased	0.377 (0.003)	0.016 (0.004)	0.379 (0.011)
Cycle-240	uniform	0.047 (0.001)	0.011 (0.002)	0.049 (0.003)
	near-uniform	0.048 (0.001)	0.012 (0.002)	0.050 (0.003)
	biased	0.051 (0.001)	0.012 (0.002)	0.054 (0.003)
Cycle-500	uniform	0.066 (0.002)	0.014 (0.002)	0.070 (0.003)
	near-uniform	0.066 (0.002)	0.013 (0.002)	0.071 (0.003)
	biased	0.076 (0.002)	0.014 (0.002)	0.081 (0.004)
Cycle-600	uniform	0.078 (0.002)	0.014 (0.003)	0.081 (0.005)
	near-uniform	0.079 (0.001)	0.014 (0.003)	0.082 (0.005)
	biased	0.078 (0.002)	0.014 (0.002)	0.082 (0.004)
Cycle-1100	uniform	0.062 (0.002)	0.015 (0.002)	0.066 (0.005)
	near-uniform	0.062 (0.002)	0.015 (0.002)	0.066 (0.005)
	biased	0.074 (0.002)	0.016 (0.002)	0.079 (0.004)
Cycle-1300	uniform	0.044 (0.001)	0.015 (0.001)	0.049 (0.004)
	near-uniform	0.045 (0.001)	0.015 (0.002)	0.049 (0.003)
	biased	0.053 (0.002)	0.016 (0.003)	0.058 (0.004)
CycleNet-2	uniform	0.026 (0.003)	0.005 (0.001)	0.025 (0.003)
	near-uniform	0.028 (0.003)	0.005 (0.001)	0.027 (0.003)
	biased	0.169 (0.004)	0.007 (0.001)	0.169 (0.003)
CycleNet-3	uniform	0.015 (0.002)	0.005 (0.001)	0.015 (0.002)
	near-uniform	0.016 (0.002)	0.005 (0.001)	0.016 (0.002)
	biased	0.122 (0.003)	0.009 (0.001)	0.123 (0.004)
CycleNet-4	uniform	0.017 (0.001)	0.006 (0.001)	0.017 (0.002)
	near-uniform	0.017 (0.002)	0.006 (0.001)	0.017 (0.002)
	biased	0.066 (0.001)	0.011 (0.002)	0.066 (0.002)
CycleNet-10	uniform	0.014 (0.002)	0.005 (0.001)	0.014 (0.002)
	near-uniform	0.014 (0.002)	0.006 (0.001)	0.014 (0.002)
	biased	0.025 (0.001)	0.007 (0.002)	0.026 (0.002)

I.2 Eigenvector Recovery Results

On the reversible grid mazes, ALLO performs better in several uniform-policy settings, whereas the trend reverses under non-uniform visitation. Using the practical readout, DALLO significantly outperforms ALLO on all eight near-uniform mazes and all eight biased mazes. On the non-reversible directed cycles, DALLO achieves higher recovery than ALLO at $\gamma = 0.9$ in all 15 settings and significantly outperforms the $\gamma = 0$ one-step control in all 15. On the more complex CycleNets, DALLO achieves higher recovery than ALLO at $\gamma = 0.9$ in 11 of 12 settings and significantly outperforms the one-step control in all 12. The following tables report the complete oracle, practical-readout, and raw-output results for each environment family.

Table 6: Grid-maze eigenvector recovery using the exact stationary distribution d^π (oracle). Chung coordinates are recovered as $D_\pi^{1/2} f_{\mathrm{DALLO}}$ and $D_\pi^{1/2} f_{\mathrm{ALLO},0.9}$ and compared with the corresponding Chung eigenvectors using cosine similarity. Results are mean (standard deviation) over 30 seeds per method; significance is assessed with Welch's two-sample t-test, and the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO + d_π	ALLO + d_π	t-statistic	p-value
GridMaze-19	uniform	0.9934 (0.0035)	**0.9966 (0.0018)**	-4.5	0.000
	near-uniform	**0.9750 (0.0049)**	0.9383 (0.0038)	32.2	0.000
	biased	**0.5127 (0.0333)**	0.3622 (0.0049)	24.5	0.000
GridMaze-26	uniform	0.9787 (0.0139)	**0.9935 (0.0051)**	-5.5	0.000
	near-uniform	**0.8056 (0.0425)**	0.6371 (0.0144)	20.6	0.000
	biased	**0.2020 (0.0129)**	0.1379 (0.0146)	18.1	0.000
GridMaze-32	uniform	0.8110 (0.0847)	**0.9180 (0.0500)**	-6.0	0.000
	near-uniform	**0.7532 (0.0498)**	0.5328 (0.0157)	23.1	0.000
	biased	0.2367 (0.0267)	**0.2655 (0.0095)**	-5.6	0.000
GridRoom-1	uniform	**0.9984 (0.0005)**	0.9980 (0.0005)	2.9	0.005
	near-uniform	**0.9980 (0.0005)**	0.9699 (0.0024)	62.0	0.000
	biased	**0.3539 (0.0521)**	0.2008 (0.0056)	16.0	0.000
GridRoom-16	uniform	0.9974 (0.0008)	0.9978 (0.0007)	-2.1	0.037
	near-uniform	**0.9697 (0.0035)**	0.9209 (0.0026)	61.7	0.000
	biased	**0.3441 (0.0217)**	0.3102 (0.0049)	8.4	0.000
GridRoom-32	uniform	0.9850 (0.0229)	0.9937 (0.0059)	-2.0	0.051
	near-uniform	**0.8654 (0.0245)**	0.6905 (0.0413)	20.0	0.000
	biased	**0.3926 (0.0221)**	0.3634 (0.0069)	6.9	0.000
GridRoom-64	uniform	0.9078 (0.0964)	0.9399 (0.0520)	-1.6	0.116
	near-uniform	**0.5684 (0.0516)**	0.5153 (0.0443)	4.3	0.000
	biased	0.2010 (0.0428)	**0.2930 (0.0086)**	-11.5	0.000
GridRoomSym-4	uniform	0.8606 (0.0602)	0.8228 (0.0729)	2.2	0.033
	near-uniform	**0.7932 (0.0966)**	0.6329 (0.0637)	7.6	0.000
	biased	**0.6264 (0.0155)**	0.4714 (0.0119)	43.4	0.000

Table 7: Grid-maze eigenvector recovery using the practical buffer-based readout. Chung coordinates are recovered as $\hat{D}_{\mathrm{D}}^{1/2} f_{\mathrm{DALLO}}$ and $\hat{D}_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0.9}$, where $\hat{d}_{\mathrm{D}}^{\pi}$ and $\hat{d}_{\mathrm{A}}^{\pi}$ are estimated from the respective replay buffers. Results are mean (standard deviation) over 30 seeds per method; significance is assessed with Welch's two-sample t-test, and the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO + $\hat{d}_{\mathrm{D}}^{\pi}$	ALLO + γ + $\hat{d}_{\mathrm{A}}^{\pi}$	t-statistic	p-value
GridMaze-19	uniform	0.9938 (0.0035)	**0.9966 (0.0018)**	-4.0	0.000
	near-uniform	**0.9892 (0.0034)**	0.9709 (0.0032)	21.6	0.000
	biased	**0.6778 (0.0531)**	0.4887 (0.0183)	18.4	0.000
GridMaze-26	uniform	0.9786 (0.0139)	**0.9934 (0.0051)**	-5.5	0.000
	near-uniform	**0.8538 (0.0420)**	0.6868 (0.0133)	20.8	0.000
	biased	**0.6195 (0.0555)**	0.3489 (0.0185)	25.3	0.000
GridMaze-32	uniform	0.8110 (0.0845)	**0.9179 (0.0500)**	-6.0	0.000
	near-uniform	**0.7873 (0.0548)**	0.5589 (0.0179)	21.7	0.000
	biased	**0.3738 (0.0498)**	0.3099 (0.0210)	6.5	0.000
GridRoom-1	uniform	**0.9984 (0.0005)**	0.9980 (0.0005)	3.5	0.001
	near-uniform	**0.9982 (0.0004)**	0.9829 (0.0029)	28.9	0.000
	biased	**0.3720 (0.0563)**	0.1920 (0.0121)	17.1	0.000
GridRoom-16	uniform	0.9977 (0.0006)	0.9978 (0.0007)	-0.7	0.475
	near-uniform	**0.9868 (0.0022)**	0.9588 (0.0019)	52.4	0.000
	biased	**0.4836 (0.0306)**	0.4494 (0.0067)	6.0	0.000
GridRoom-32	uniform	0.9852 (0.0229)	0.9936 (0.0059)	-2.0	0.060
	near-uniform	**0.9197 (0.0264)**	0.7366 (0.0514)	17.4	0.000
	biased	**0.4571 (0.0324)**	0.3989 (0.0125)	9.2	0.000
GridRoom-64	uniform	0.9077 (0.0961)	0.9397 (0.0520)	-1.6	0.115
	near-uniform	**0.6260 (0.0660)**	0.5617 (0.0658)	3.8	0.000
	biased	**0.2636 (0.0443)**	0.2378 (0.0128)	3.1	0.004
GridRoomSym-4	uniform	0.8605 (0.0602)	0.8230 (0.0727)	2.2	0.034
	near-uniform	**0.7935 (0.0966)**	0.6367 (0.0640)	7.4	0.000
	biased	**0.6277 (0.0153)**	0.5420 (0.0101)	25.6	0.000

Table 8: Grid-maze eigenvector recovery using the raw network outputs f_{DALLO} and $f_{\mathrm{ALLO},0.9}$, without stationary-distribution reweighting. Results are mean (standard deviation) over 30 seeds per method; significance is assessed with Welch's two-sample t-test, and the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO – raw	ALLO – raw	t-statistic	p-value
GridMaze-19	uniform	0.9934 (0.0035)	**0.9966 (0.0018)**	-4.5	0.000
	near-uniform	**0.9805 (0.0034)**	0.9694 (0.0031)	13.1	0.000
	biased	**0.5433 (0.0471)**	0.4442 (0.0172)	10.8	0.000
GridMaze-26	uniform	0.9787 (0.0139)	**0.9935 (0.0051)**	-5.5	0.000
	near-uniform	**0.8429 (0.0418)**	0.6857 (0.0132)	19.7	0.000
	biased	**0.5156 (0.0476)**	0.3239 (0.0166)	20.8	0.000
GridMaze-32	uniform	0.8110 (0.0847)	**0.9180 (0.0500)**	-6.0	0.000
	near-uniform	**0.7788 (0.0548)**	0.5578 (0.0181)	21.0	0.000
	biased	0.2920 (0.0405)	0.2842 (0.0195)	1.0	0.347
GridRoom-1	uniform	**0.9984 (0.0005)**	0.9980 (0.0005)	2.9	0.005
	near-uniform	0.9703 (0.0020)	**0.9786 (0.0030)**	-12.7	0.000
	biased	**0.3055 (0.0519)**	0.1492 (0.0130)	16.0	0.000
GridRoom-16	uniform	0.9974 (0.0008)	0.9978 (0.0007)	-2.1	0.037
	near-uniform	**0.9741 (0.0020)**	0.9572 (0.0018)	34.3	0.000
	biased	0.3796 (0.0284)	**0.4183 (0.0064)**	-7.3	0.000
GridRoom-32	uniform	0.9850 (0.0229)	0.9937 (0.0059)	-2.0	0.051
	near-uniform	**0.9076 (0.0261)**	0.7353 (0.0513)	16.4	0.000
	biased	0.2981 (0.0277)	**0.3480 (0.0107)**	-9.2	0.000
GridRoom-64	uniform	0.9078 (0.0964)	0.9399 (0.0520)	-1.6	0.116
	near-uniform	**0.6141 (0.0650)**	0.5601 (0.0657)	3.2	0.002
	biased	0.1875 (0.0351)	**0.2112 (0.0095)**	-3.6	0.001
GridRoomSym-4	uniform	0.8606 (0.0602)	0.8228 (0.0729)	2.2	0.033
	near-uniform	**0.7844 (0.0947)**	0.6350 (0.0639)	7.2	0.000
	biased	**0.5230 (0.0142)**	0.4982 (0.0090)	8.1	0.000

Table 9: Directed-cycle eigenvector recovery using the oracle readout. Chung coordinates are recovered as $D_\pi^{1/2} f_{\mathrm{DALLO}}$, $D_\pi^{1/2} f_{\mathrm{ALLO},0.9}$, and $D_\pi^{1/2} f_{\mathrm{ALLO},0}$. Results are mean (standard deviation) over 30 seeds per method. The reported t-statistic and p-value compare $D_\pi^{1/2} f_{\mathrm{DALLO}}$ with $D_\pi^{1/2} f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO $+ d_\pi$	ALLO $+ \gamma +$ d_π	ALLO $+ d_\pi$	t-statistic	p-value
Cycle-240	uniform	**0.9980 (0.0007)**	0.4267 (0.0063)	0.9594 (0.0107)	19.6	0.000
	near-uniform	**0.9979 (0.0007)**	0.4351 (0.0047)	0.9611 (0.0102)	19.7	0.000
	biased	**0.9972 (0.0018)**	0.4717 (0.0099)	0.9604 (0.0049)	38.9	0.000
Cycle-500	uniform	**0.9777 (0.0378)**	0.4798 (0.0119)	0.8692 (0.0263)	12.9	0.000
	near-uniform	**0.9766 (0.0237)**	0.4511 (0.0102)	0.8613 (0.0177)	21.3	0.000
	biased	**0.9722 (0.0339)**	0.4878 (0.0137)	0.9330 (0.0159)	5.7	0.000
Cycle-600	uniform	**0.9947 (0.0038)**	0.4476 (0.0131)	0.9181 (0.0095)	40.9	0.000
	near-uniform	**0.9957 (0.0025)**	0.4495 (0.0126)	0.9170 (0.0126)	33.6	0.000
	biased	**0.9963 (0.0018)**	0.4380 (0.0167)	0.8769 (0.0516)	12.7	0.000
Cycle-1100	uniform	**0.9942 (0.0024)**	0.4028 (0.0192)	0.8447 (0.0466)	17.6	0.000
	near-uniform	**0.9936 (0.0026)**	0.4003 (0.0212)	0.8565 (0.0416)	18.0	0.000
	biased	**0.9958 (0.0013)**	0.3290 (0.0092)	0.7789 (0.0358)	33.1	0.000
Cycle-1300	uniform	**0.9933 (0.0029)**	0.3961 (0.0188)	0.9604 (0.0184)	9.7	0.000
	near-uniform	**0.9925 (0.0067)**	0.3917 (0.0180)	0.9625 (0.0096)	14.0	0.000
	biased	**0.9793 (0.0184)**	0.3394 (0.0250)	0.9059 (0.0344)	10.3	0.000

Table 10: Raw eigenvector recovery on non-reversible directed cycles. The learned outputs f_{DALLO}, $f_{\mathrm{ALLO},0.9}$, and $f_{\mathrm{ALLO},0}$ are compared directly with the Chung eigenvectors, without stationary-distribution reweighting. Results are mean (standard deviation) over 30 seeds per method. The reported t-statistic and p-value compare f_{DALLO} with $f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO – raw	ALLO + γ – raw	ALLO – raw	t-statistic	p-value
Cycle-240	uniform	**0.9547 (0.0016)**	0.3864 (0.0056)	0.9173 (0.0101)	20.1	0.000
	near-uniform	**0.9528 (0.0017)**	0.3924 (0.0046)	0.9169 (0.0096)	20.1	0.000
	biased	**0.9415 (0.0030)**	0.4199 (0.0079)	0.9004 (0.0055)	35.7	0.000
Cycle-500	uniform	**0.9184 (0.0351)**	0.4098 (0.0089)	0.8071 (0.0266)	13.8	0.000
	near-uniform	**0.9142 (0.0224)**	0.3807 (0.0084)	0.7965 (0.0177)	22.6	0.000
	biased	**0.8875 (0.0301)**	0.3995 (0.0106)	0.8444 (0.0164)	6.9	0.000
Cycle-600	uniform	**0.9165 (0.0040)**	0.3716 (0.0115)	0.8385 (0.0094)	41.7	0.000
	near-uniform	**0.9143 (0.0043)**	0.3715 (0.0108)	0.8353 (0.0123)	33.1	0.000
	biased	**0.9028 (0.0034)**	0.3515 (0.0121)	0.7948 (0.0479)	12.3	0.000
Cycle-1100	uniform	**0.9261 (0.0037)**	0.3492 (0.0178)	0.7717 (0.0455)	18.5	0.000
	near-uniform	**0.9232 (0.0035)**	0.3438 (0.0198)	0.7810 (0.0391)	19.9	0.000
	biased	**0.9014 (0.0032)**	0.2772 (0.0066)	0.6942 (0.0327)	34.5	0.000
Cycle-1300	uniform	**0.9438 (0.0029)**	0.3383 (0.0180)	0.9098 (0.0179)	10.3	0.000
	near-uniform	**0.9412 (0.0073)**	0.3334 (0.0178)	0.9097 (0.0097)	14.2	0.000
	biased	**0.9136 (0.0180)**	0.2742 (0.0187)	0.8432 (0.0306)	10.8	0.000

Table 11: CycleNet eigenvector recovery using the oracle readout. Chung coordinates are recovered as $D_\pi^{1/2} f_{\mathrm{DALLO}}$, $D_\pi^{1/2} f_{\mathrm{ALLO},0.9}$, and $D_\pi^{1/2} f_{\mathrm{ALLO},0}$. Results are mean (standard deviation) over 30 seeds per method. The reported t-statistic and p-value compare $D_\pi^{1/2} f_{\mathrm{DALLO}}$ with $D_\pi^{1/2} f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO + d_π	ALLO + γ + d_π	ALLO + d_π	t-statistic	p-value
CycleNet-2	uniform	**0.7507 (0.1320)**	0.3454 (0.0509)	0.3801 (0.0738)	13.4	0.000
	near-uni.	**0.7614 (0.0809)**	0.4039 (0.0717)	0.3534 (0.0795)	19.7	0.000
	biased	**0.6436 (0.0982)**	0.7858 (0.0400)	0.3846 (0.0628)	12.2	0.000
CycleNet-3	uniform	**0.8978 (0.1255)**	0.3058 (0.0133)	0.6793 (0.1076)	7.2	0.000
	near-uni.	**0.8684 (0.1045)**	0.3044 (0.0124)	0.6493 (0.1115)	7.9	0.000
	biased	**0.9854 (0.0314)**	0.4739 (0.0159)	0.5981 (0.1172)	17.5	0.000
CycleNet-4	uniform	**0.8220 (0.0802)**	0.1353 (0.0117)	0.4892 (0.0834)	15.8	0.000
	near-uni.	**0.7890 (0.0927)**	0.1371 (0.0112)	0.4789 (0.0964)	12.7	0.000
	biased	**0.7427 (0.1168)**	0.5474 (0.0770)	0.4930 (0.1316)	7.8	0.000
CycleNet-10	uniform	**0.7741 (0.1034)**	0.2897 (0.0088)	0.6064 (0.1066)	6.2	0.000
	near-uni.	**0.7005 (0.0994)**	0.2913 (0.0088)	0.6126 (0.0893)	3.6	0.001
	biased	**0.8429 (0.0994)**	0.5298 (0.0460)	0.6501 (0.1261)	6.6	0.000

Table 12: CycleNet eigenvector recovery using the practical buffer-based readout. Chung coordinates are recovered as $\hat{D}_{\mathrm{D}}^{1/2} f_{\mathrm{DALLO}}$, $\hat{D}_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0.9}$, and $\hat{D}_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0}$, where $\hat{d}_{\mathrm{D}}^{\pi}$ and $\hat{d}_{\mathrm{A}}^{\pi}$ are estimated from the respective replay buffers. Results are mean (standard deviation) over 30 seeds per method. The reported t-statistic and p-value compare $\hat{D}_{\mathrm{D}}^{1/2} f_{\mathrm{DALLO}}$ with $\hat{D}_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO + $\hat{d}_{\mathrm{D}}^{\pi}$	ALLO + γ + $\hat{d}_{\mathrm{A}}^{\pi}$	ALLO + $\hat{d}_{\mathrm{A}}^{\pi}$	t-statistic	p-value
CycleNet-2	uniform	**0.7506 (0.1321)**	0.3467 (0.0508)	0.3793 (0.0737)	13.4	0.000
	near-uniform	**0.7613 (0.0809)**	0.4050 (0.0715)	0.3531 (0.0793)	19.7	0.000
	biased	**0.6439 (0.0986)**	0.7866 (0.0395)	0.3803 (0.0623)	12.4	0.000
CycleNet-3	uniform	**0.8979 (0.1254)**	0.3058 (0.0133)	0.6793 (0.1078)	7.2	0.000
	near-uniform	**0.8684 (0.1045)**	0.3044 (0.0123)	0.6493 (0.1115)	7.9	0.000
	biased	**0.9854 (0.0314)**	0.4801 (0.0158)	0.5997 (0.1169)	17.4	0.000
CycleNet-4	uniform	**0.8219 (0.0802)**	0.1356 (0.0116)	0.4889 (0.0833)	15.8	0.000
	near-uniform	**0.7889 (0.0927)**	0.1373 (0.0110)	0.4789 (0.0963)	12.7	0.000
	biased	**0.7426 (0.1167)**	0.5393 (0.0763)	0.4900 (0.1311)	7.9	0.000
CycleNet-10	uniform	**0.7739 (0.1034)**	0.2899 (0.0088)	0.6063 (0.1065)	6.2	0.000
	near-uniform	**0.7004 (0.0994)**	0.2914 (0.0088)	0.6126 (0.0892)	3.6	0.001
	biased	**0.8428 (0.0994)**	0.5294 (0.0459)	0.6499 (0.1260)	6.6	0.000

Table 13: CycleNet eigenvector recovery using the raw network outputs f_{DALLO}, $f_{\mathrm{ALLO},0.9}$, and $f_{\mathrm{ALLO},0}$, without stationary-distribution reweighting. Results are mean (standard deviation) over 30 seeds per method. The reported t-statistic and p-value compare f_{DALLO} with $f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO – raw	ALLO + γ – raw	ALLO – raw	t-statistic	p-value
CycleNet-2	uniform	**0.7506 (0.1320)**	0.3465 (0.0507)	0.3794 (0.0737)	13.4	0.000
	near-uniform	**0.7610 (0.0812)**	0.4048 (0.0714)	0.3531 (0.0793)	19.7	0.000
	biased	**0.6365 (0.0973)**	0.7818 (0.0392)	0.3802 (0.0616)	12.2	0.000
CycleNet-3	uniform	**0.8976 (0.1255)**	0.3055 (0.0131)	0.6791 (0.1078)	7.2	0.000
	near-uniform	**0.8682 (0.1045)**	0.3039 (0.0122)	0.6492 (0.1116)	7.8	0.000
	biased	**0.9714 (0.0310)**	0.4764 (0.0152)	0.5963 (0.1164)	17.1	0.000
CycleNet-4	uniform	**0.8216 (0.0801)**	0.1348 (0.0115)	0.4892 (0.0833)	15.8	0.000
	near-uniform	**0.7886 (0.0927)**	0.1366 (0.0110)	0.4786 (0.0965)	12.7	0.000
	biased	**0.7269 (0.1142)**	0.5202 (0.0741)	0.4817 (0.1285)	7.8	0.000
CycleNet-10	uniform	**0.7739 (0.1033)**	0.2896 (0.0088)	0.6061 (0.1065)	6.2	0.000
	near-uniform	**0.7002 (0.0993)**	0.2911 (0.0087)	0.6124 (0.0892)	3.6	0.001
	biased	**0.8256 (0.0972)**	0.5148 (0.0446)	0.6362 (0.1238)	6.6	0.000

J EIGENVALUE RECOVERY

We report the 11 smallest eigenvalues, corresponding to the $d = 11$ learned representations. DALLO's discounted eigenvalues are recovered from the diagonal dual variables and mapped to the undiscounted Chung scale using the inverse spectral map in equation 4. For ALLO, we apply the same scalar map h_{γ}^{-1} to both the learned eigenvalues and the spectrum of its sampling-induced target $\widehat{L}_{\gamma}$ before comparison. Thus, the learned and target ALLO eigenvalues are evaluated on the same rescaled scale; this rescaling is used only for comparison and does not imply that $\widehat{L}_{\gamma}$ is obtained by discounting an underlying Chung operator. On the reversible grid mazes, $\widehat{L}_{\gamma}$ and the Chung target share the same eigensystem, so no separate $\widehat{L}_{\gamma}$ target is reported. All results are averaged over 30 seeds.

Figure 5 summarizes eigenvalue recovery across all directed environment-policy settings. Each point reports the mean absolute error over the 11 learned eigenvalues, averaged over 30 seeds, comparing DALLO against the Chung spectrum and ALLO against its sampling-induced target $\widehat{L}_{\gamma}$. The dashed diagonal indicates equal error; points below it correspond to lower error for DALLO.

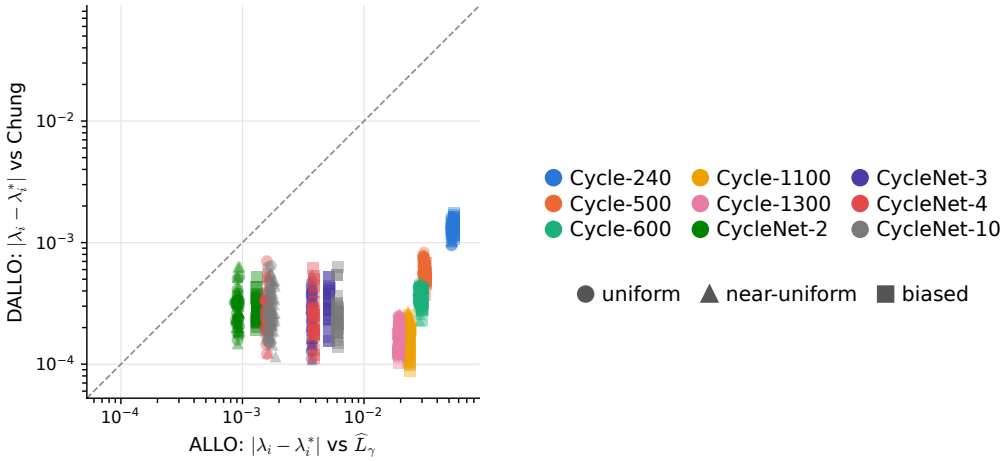

Figure 5: Eigenvalue recovery across the directed benchmarks. Each point corresponds to one environment-policy setting. The horizontal axis reports ALLO's mean absolute eigenvalue error over the 11 learned dimensions relative to its sampling-induced target $\widehat{L}_\gamma$, while the vertical axis reports DALLO's corresponding error relative to the Chung spectrum. Values are averaged over 30 seeds. The dashed diagonal denotes equal error; points below it correspond to lower error for DALLO.

The following tables provide the complete per-eigenvalue results for the grid mazes, directed cycles, and CycleNets under the uniform, near-uniform, and biased policies.

Table 14: Recovery of the 11 smallest eigenvalues on the grid mazes under the *uniform* policy. The "True" column reports the corresponding Chung eigenvalues. Since these chains are reversible, DALLO and ALLO share the same population eigensystem. Learned eigenvalues are mapped to the undiscounted scale using the inverse spectral map h_γ^{-1} before comparison with the target spectrum. DALLO and ALLO entries are mean (standard deviation) over 30 seeds; ground-truth eigenvalues are deterministic.

Env	True	DALLO	ALLO
GridMaze-19	0.00000000	0.00001012 (0.00000173)	0.00000146 (0.00000019)
	0.00017320	0.00025796 (0.00001414)	0.00018012 (0.00000471)
	0.00064661	0.00073503 (0.00001806)	0.00052486 (0.00000736)
	0.00075937	0.00090098 (0.00004069)	0.00061601 (0.00001108)
	0.00157242	0.00167764 (0.00005400)	0.00119787 (0.00001835)
	0.00263513	0.00276548 (0.00007032)	0.00196205 (0.00003305)
	0.00421241	0.00428313 (0.00007421)	0.00308167 (0.00003404)
	0.00466864	0.00480899 (0.00010756)	0.00344316 (0.00005836)
	0.00605108	0.00618093 (0.00012634)	0.00440911 (0.00005951)
	0.00657864	0.00668584 (0.00010240)	0.00481051 (0.00007606)
	0.00993238	0.00998527 (0.00016689)	0.00718541 (0.00009812)
GridMaze-26	0.00000000	0.00001093 (0.00000233)	0.00000194 (0.00000017)
	0.00005617	0.00014498 (0.00000823)	0.00009721 (0.00000324)
	0.00037875	0.00052214 (0.00002507)	0.00034871 (0.00000704)
	0.00043991	0.00061394 (0.00003540)	0.00041344 (0.00001361)
	0.00054111	0.00066678 (0.00002471)	0.00046353 (0.00000790)
	0.00064476	0.00081031 (0.00002397)	0.00055106 (0.00000846)
	0.00105110	0.00123535 (0.00004421)	0.00084640 (0.00001280)
	0.00145331	0.00165811 (0.00005568)	0.00114155 (0.00002118)
	0.00176221	0.00202658 (0.00008446)	0.00137258 (0.00002339)
	0.00220847	0.00249942 (0.00008579)	0.00172602 (0.00003734)
	0.00291317	0.00321166 (0.00009329)	0.00222401 (0.00004123)
GridMaze-32	0.00000000	0.00001367 (0.00000416)	0.00000217 (0.00000021)
	0.00000000	0.00015868 (0.00001566)	0.00009968 (0.00000678)
	0.00004551	0.00025615 (0.00004942)	0.00011078 (0.00001384)

Table 14, continued.

Env	True	DALLO	ALLO
	0.00014427	0.00028703 (0.00003824)	0.00017524 (0.00000700)
	0.00041664	0.00060826 (0.00004583)	0.00038682 (0.00001045)
	0.00048231	0.00067551 (0.00004532)	0.00044110 (0.00001285)
	0.00065051	0.00088618 (0.00004995)	0.00058886 (0.00001180)
	0.00070725	0.00096898 (0.00007569)	0.00061161 (0.00001367)
	0.00084312	0.00124944 (0.00010585)	0.00075520 (0.00002577)
	0.00110651	0.00141258 (0.00006240)	0.00092771 (0.00001586)
	0.00167054	0.00199420 (0.00008860)	0.00134047 (0.00002461)
GridRoom-1	0.00000000	0.00000384 (0.00000090)	0.00000242 (0.00000040)
	0.01092620	0.01060277 (0.00010632)	0.00755658 (0.00005872)
	0.01092620	0.01087550 (0.00011828)	0.00767930 (0.00005742)
	0.02185240	0.02131920 (0.00022840)	0.01518013 (0.00012733)
	0.04322727	0.04147562 (0.00020106)	0.02967566 (0.00017725)
	0.04322727	0.04199356 (0.00015354)	0.03001725 (0.00017613)
	0.05415347	0.05196693 (0.00020903)	0.03716950 (0.00022777)
	0.05415347	0.05256069 (0.00020496)	0.03758620 (0.00025137)
	0.08645454	0.08258631 (0.00035336)	0.05931789 (0.00034162)
	0.09549150	0.09069833 (0.00030425)	0.06528435 (0.00025482)
	0.09549150	0.09127676 (0.00036149)	0.06592334 (0.00031780)
GridRoom-16	0.00000000	0.00000732 (0.00000147)	0.00000130 (0.00000013)
	0.00017551	0.00022732 (0.00001032)	0.00017670 (0.00000418)
	0.00070294	0.00076829 (0.00002048)	0.00055839 (0.00000731)
	0.00156518	0.00161638 (0.00002886)	0.00117949 (0.00001761)
	0.00274984	0.00280014 (0.00004878)	0.00202450 (0.00002237)
	0.00423073	0.00428081 (0.00006978)	0.00308474 (0.00004344)
	0.00598292	0.00598404 (0.00007504)	0.00433643 (0.00004866)
	0.00788379	0.00786425 (0.00010140)	0.00569420 (0.00006938)
	0.01008400	0.01001802 (0.00006691)	0.00727697 (0.00009987)
	0.01244077	0.01238529 (0.00008871)	0.00897536 (0.00008609)
	0.01459569	0.01453427 (0.00012150)	0.01051942 (0.00012442)
GridRoom-32	0.00000000	0.00000888 (0.00000156)	0.00000154 (0.00000014)
	0.00008727	0.00016614 (0.00001097)	0.00012010 (0.00000431)
	0.00020303	0.00031608 (0.00001392)	0.00021703 (0.00000637)
	0.00043769	0.00051921 (0.00001709)	0.00037573 (0.00000448)
	0.00073231	0.00084081 (0.00002810)	0.00059589 (0.00000814)
	0.00152600	0.00162071 (0.00004451)	0.00116204 (0.00001681)
	0.00182221	0.00194741 (0.00005519)	0.00138699 (0.00001689)
	0.00227243	0.00240200 (0.00004847)	0.00171515 (0.00002225)
	0.00308121	0.00320210 (0.00006153)	0.00227906 (0.00002822)
	0.00325308	0.00342098 (0.00006388)	0.00242559 (0.00004002)
	0.00420295	0.00434401 (0.00005288)	0.00310597 (0.00003745)
GridRoom-64	0.00000000	0.00000852 (0.00000150)	0.00000138 (0.00000012)
	0.00004432	0.00011942 (0.00001022)	0.00008522 (0.00000314)
	0.00011027	0.00019393 (0.00001104)	0.00014113 (0.00000397)
	0.00017733	0.00030525 (0.00003270)	0.00021463 (0.00001138)
	0.00022167	0.00033590 (0.00003355)	0.00023130 (0.00001358)
	0.00023066	0.00040688 (0.00002618)	0.00026764 (0.00001180)
	0.00039258	0.00053776 (0.00002701)	0.00037305 (0.00001305)
	0.00045506	0.00062767 (0.00003258)	0.00041640 (0.00001423)
	0.00070126	0.00086309 (0.00002614)	0.00059987 (0.00000657)
	0.00100520	0.00120710 (0.00003900)	0.00083829 (0.00001786)
	0.00109001	0.00129729 (0.00003523)	0.00089204 (0.00001632)
GridRoomSym-4	0.00000000	0.00000855 (0.00000108)	0.00000398 (0.00000070)
	0.00634381	0.00611908 (0.00006258)	0.00429820 (0.00004646)
	0.00634381	0.00628894 (0.00005116)	0.00441407 (0.00005171)
	0.01328642	0.01288534 (0.00009519)	0.00905887 (0.00008058)
	0.09097350	0.08525191 (0.00022838)	0.06026431 (0.00021949)
	0.09117554	0.08571373 (0.00023581)	0.06068214 (0.00022177)
	0.09117554	0.08615907 (0.00029285)	0.06112378 (0.00022627)

Table 14, continued.

Env	True	DALLO	ALLO
	0.09134889	0.08668071 (0.00026868)	0.06154954 (0.00030609)
	0.10741422	0.10076326 (0.00041309)	0.07149405 (0.00028851)
	0.10895382	0.10209985 (0.00032636)	0.07240031 (0.00025414)
	0.10895382	0.10272170 (0.00029058)	0.07293312 (0.00027246)

Table 15: Recovery of the 11 smallest eigenvalues on the grid mazes under the *near-uniform* policy. The "True" column reports the corresponding Chung eigenvalues. Since these chains are reversible, DALLO and ALLO share the same population eigensystem. Learned eigenvalues are mapped to the undiscounted scale using the inverse spectral map h_γ^{-1} before comparison with the target spectrum. DALLO and ALLO entries are mean (standard deviation) over 30 seeds; ground-truth eigenvalues are deterministic.

Env	True	DALLO	ALLO
GridMaze-19	0.00000000	0.00000992 (0.00000206)	0.00000137 (0.00000017)
	0.00008690	0.00020969 (0.00001621)	0.00017693 (0.00000476)
	0.00036402	0.00055628 (0.00002382)	0.00050945 (0.00000857)
	0.00053227	0.00073464 (0.00003878)	0.00060449 (0.00001103)
	0.00167719	0.00160407 (0.00006181)	0.00117335 (0.00001863)
	0.00259459	0.00270514 (0.00008215)	0.00193615 (0.00003331)
	0.00339914	0.00357646 (0.00010560)	0.00288285 (0.00005316)
	0.00418709	0.00421964 (0.00009521)	0.00322326 (0.00005849)
	0.00566722	0.00571625 (0.00010920)	0.00427280 (0.00006072)
	0.00622421	0.00628884 (0.00011081)	0.00461925 (0.00007111)
	0.00986204	0.00984879 (0.00013347)	0.00711318 (0.00008038)
GridMaze-26	0.00000000	0.00001144 (0.00000180)	0.00000189 (0.00000022)
	0.00006194	0.00014740 (0.00001479)	0.00009657 (0.00000371)
	0.00031765	0.00049430 (0.00003093)	0.00034639 (0.00000656)
	0.00048869	0.00054499 (0.00003655)	0.00040705 (0.00001464)
	0.00055495	0.00065891 (0.00003565)	0.00046519 (0.00000681)
	0.00074366	0.00080339 (0.00003315)	0.00055289 (0.00000824)
	0.00109559	0.00124756 (0.00005054)	0.00084745 (0.00001368)
	0.00147620	0.00158840 (0.00007389)	0.00110593 (0.00001905)
	0.00193966	0.00210151 (0.00011050)	0.00136860 (0.00002151)
	0.00233452	0.00250252 (0.00010927)	0.00170674 (0.00003529)
	0.00297666	0.00329016 (0.00016824)	0.00221803 (0.00003812)
GridMaze-32	0.00000000	0.00001260 (0.00000321)	0.00000216 (0.00000023)
	0.00000000	0.00015136 (0.00002099)	0.00009940 (0.00000689)
	0.00001595	0.00026641 (0.00003925)	0.00011075 (0.00001413)
	0.00018217	0.00028913 (0.00005133)	0.00017466 (0.00000675)
	0.00031994	0.00056998 (0.00005952)	0.00038832 (0.00000652)
	0.00038065	0.00073771 (0.00005883)	0.00045365 (0.00000799)
	0.00066118	0.00078989 (0.00005453)	0.00057333 (0.00001375)
	0.00090113	0.00107256 (0.00008053)	0.00060228 (0.00001163)
	0.00124015	0.00135802 (0.00008592)	0.00077675 (0.00002001)
	0.00131261	0.00157120 (0.00018757)	0.00091597 (0.00001687)
	0.00154422	0.00182117 (0.00013994)	0.00130344 (0.00002644)
GridRoom-1	0.00000000	0.00000657 (0.00000141)	0.00000242 (0.00000037)
	0.01131762	0.01084787 (0.00009865)	0.00759802 (0.00005109)
	0.01131762	0.01114856 (0.00013521)	0.00773518 (0.00005461)
	0.02263523	0.02180603 (0.00018913)	0.01525152 (0.00012494)
	0.04359284	0.04174404 (0.00023043)	0.02979998 (0.00019951)
	0.04359284	0.04216739 (0.00019016)	0.03014009 (0.00018312)
	0.05491045	0.05232023 (0.00033613)	0.03728254 (0.00021891)
	0.05491045	0.05304631 (0.00029675)	0.03773426 (0.00024964)
	0.08718567	0.08308383 (0.00052175)	0.05951791 (0.00029334)
	0.09581524	0.09083768 (0.00035407)	0.06545619 (0.00022886)
	0.09581524	0.09149199 (0.00034220)	0.06597650 (0.00025523)
GridRoom-16	0.00000000	0.00000879 (0.00000124)	0.00000128 (0.00000016)
	0.00018045	0.00023969 (0.00001003)	0.00017874 (0.00000396)

Table 15, continued.

Env	True	DALLO	ALLO
	0.00086111	0.00075378 (0.00002671)	0.00055821 (0.00000503)
	0.00185543	0.00175989 (0.00005194)	0.00120374 (0.00001833)
	0.00324332	0.00313912 (0.00005321)	0.00209614 (0.00002274)
	0.00444386	0.00437039 (0.00007503)	0.00310989 (0.00004455)
	0.00644690	0.00634451 (0.00007167)	0.00445409 (0.00005278)
	0.00808327	0.00793271 (0.00008528)	0.00572770 (0.00008061)
	0.01041253	0.01025441 (0.00008843)	0.00739482 (0.00007904)
	0.01237014	0.01224160 (0.00011769)	0.00897474 (0.00008817)
	0.01463830	0.01450375 (0.00012299)	0.01063297 (0.00010117)
GridRoom-32	0.00000000	0.00000925 (0.00000150)	0.00000152 (0.00000019)
	0.00012935	0.00019150 (0.00001181)	0.00012133 (0.00000499)
	0.00018835	0.00028674 (0.00001751)	0.00021853 (0.00000583)
	0.00049153	0.00049135 (0.00002004)	0.00037505 (0.00000441)
	0.00085949	0.00081721 (0.00002947)	0.00059672 (0.00000867)
	0.00181469	0.00179085 (0.00004649)	0.00118773 (0.00001670)
	0.00215474	0.00208976 (0.00006085)	0.00141878 (0.00001971)
	0.00268532	0.00263977 (0.00006036)	0.00174709 (0.00002479)
	0.00350835	0.00349331 (0.00005731)	0.00237097 (0.00003369)
	0.00364349	0.00365854 (0.00007420)	0.00245337 (0.00003432)
	0.00442080	0.00447795 (0.00007649)	0.00313235 (0.00004034)
GridRoom-64	0.00000000	0.00000840 (0.00000113)	0.00000139 (0.00000013)
	0.00004291	0.00011561 (0.00000884)	0.00008551 (0.00000297)
	0.00008475	0.00018608 (0.00001300)	0.00014160 (0.00000510)
	0.00009662	0.00029172 (0.00003148)	0.00021254 (0.00001475)
	0.00016664	0.00032121 (0.00003539)	0.00023278 (0.00001507)
	0.00025135	0.00038579 (0.00003549)	0.00026760 (0.00001283)
	0.00049849	0.00055666 (0.00003648)	0.00037700 (0.00001761)
	0.00059955	0.00061656 (0.00002785)	0.00041675 (0.00001700)
	0.00073677	0.00086443 (0.00003833)	0.00060116 (0.00000884)
	0.00118176	0.00124682 (0.00004715)	0.00083785 (0.00001634)
	0.00124515	0.00133667 (0.00003958)	0.00089908 (0.00001283)
GridRoomSym-4	0.00000000	0.00001036 (0.00000146)	0.00000393 (0.00000061)
	0.00643674	0.00615136 (0.00006624)	0.00431826 (0.00004492)
	0.00643903	0.00633980 (0.00006655)	0.00441852 (0.00004798)
	0.01348702	0.01297749 (0.00011725)	0.00907516 (0.00008680)
	0.09113451	0.08538349 (0.00030293)	0.05996510 (0.00033166)
	0.09134148	0.08590756 (0.00032622)	0.06067969 (0.00035523)
	0.09168016	0.08642477 (0.00026427)	0.06131409 (0.00033482)
	0.09190852	0.08695243 (0.00033718)	0.06193874 (0.00028769)
	0.10739881	0.10065830 (0.00041322)	0.07138595 (0.00031483)
	0.10873161	0.10190398 (0.00045824)	0.07223414 (0.00032534)
	0.10961173	0.10296927 (0.00043747)	0.07296984 (0.00037813)

Table 16: Recovery of the 11 smallest eigenvalues on the grid mazes under the *biased* policy. The "True" column reports the corresponding Chung eigenvalues. Since these chains are reversible, DALLO and ALLO share the same population eigensystem. Learned eigenvalues are mapped to the undiscounted scale using the inverse spectral map h_γ^{-1} before comparison with the target spectrum. DALLO and ALLO entries are mean (standard deviation) over 30 seeds; ground-truth eigenvalues are deterministic.

Env	True	DALLO	ALLO
GridMaze-19	0.00000000	0.00001347 (0.00000416)	0.00000144 (0.00000017)
	0.00000126	0.00009583 (0.00000969)	0.00016756 (0.00000505)
	0.00001920	0.00017525 (0.00001683)	0.00036556 (0.00000751)
	0.00005116	0.00022686 (0.00002536)	0.00048714 (0.00001004)
	0.00010290	0.00035299 (0.00004741)	0.00089212 (0.00002588)
	0.00206479	0.00162946 (0.00006596)	0.00146998 (0.00003857)
	0.00213138	0.00175436 (0.00010112)	0.00201434 (0.00004024)
	0.00306180	0.00319107 (0.00009738)	0.00236091 (0.00004053)

Table 16, continued.

Env	True	DALLO	ALLO
	0.00547344	0.00493188 (0.00013956)	0.00302182 (0.00004326)
	0.00825478	0.00645288 (0.00024994)	0.00333647 (0.00004472)
	0.00873307	0.00760751 (0.00024125)	0.00566793 (0.00007593)
GridMaze-26	0.00000000	0.00001896 (0.00001051)	0.00000197 (0.00000021)
	0.00000092	0.00011016 (0.00000832)	0.00008859 (0.00000264)
	0.00001356	0.00018741 (0.00001430)	0.00027498 (0.00000636)
	0.00001727	0.00021109 (0.00002553)	0.00031630 (0.00001322)
	0.00027153	0.00027807 (0.00002737)	0.00038022 (0.00000710)
	0.00030696	0.00033174 (0.00005242)	0.00043913 (0.00000761)
	0.00044945	0.00064320 (0.00003744)	0.00079242 (0.00001427)
	0.00189635	0.00100474 (0.00006161)	0.00089142 (0.00001847)
	0.00198226	0.00138275 (0.00007001)	0.00119962 (0.00001873)
	0.00206903	0.00156001 (0.00012905)	0.00139022 (0.00003998)
	0.00214006	0.00175204 (0.00015799)	0.00184847 (0.00004563)
GridMaze-32	0.00000000	0.00001985 (0.00000875)	0.00000212 (0.00000021)
	0.00000000	0.00010909 (0.00001451)	0.00009378 (0.00000763)
	0.00000018	0.00012945 (0.00001587)	0.00010331 (0.00001646)
	0.00000130	0.00018499 (0.00002059)	0.00015150 (0.00000621)
	0.00000362	0.00022863 (0.00001712)	0.00031478 (0.00000825)
	0.00002485	0.00026179 (0.00003089)	0.00042837 (0.00001343)
	0.00002547	0.00031726 (0.00005024)	0.00045959 (0.00001166)
	0.00005124	0.00042875 (0.00008968)	0.00051739 (0.00002018)
	0.00101704	0.00076076 (0.00010445)	0.00073438 (0.00001787)
	0.00157678	0.00095451 (0.00010887)	0.00078321 (0.00001626)
	0.00188199	0.00109722 (0.00011418)	0.00105812 (0.00003153)
GridRoom-1	0.00000000	0.00003165 (0.00000736)	0.00000257 (0.00000043)
	0.00874096	0.00783554 (0.00011267)	0.00607729 (0.00005898)
	0.03458182	0.02032974 (0.00085731)	0.01359322 (0.00008427)
	0.04667616	0.02837364 (0.00100650)	0.01907350 (0.00012385)
	0.05541712	0.02971207 (0.00032040)	0.02375888 (0.00018961)
	0.07639320	0.05006661 (0.00163489)	0.03492935 (0.00022478)
	0.08125798	0.05913395 (0.00109355)	0.04413935 (0.00029464)
	0.08322065	0.06301679 (0.00086379)	0.04899433 (0.00035225)
	0.09196161	0.06702543 (0.00180630)	0.05108595 (0.00035545)
	0.11780247	0.08569087 (0.00207029)	0.06061677 (0.00036668)
	0.12306936	0.08948174 (0.00184793)	0.06324936 (0.00048271)
GridRoom-16	0.00000000	0.00002535 (0.00000787)	0.00000151 (0.00000017)
	0.00000059	0.00016623 (0.00001794)	0.00017244 (0.00000416)
	0.00162479	0.00137605 (0.00004415)	0.00053077 (0.00000858)
	0.00527185	0.00204843 (0.00015813)	0.00174661 (0.00003113)
	0.00885759	0.00375267 (0.00036822)	0.00255225 (0.00002645)
	0.00898921	0.00562539 (0.00020268)	0.00362231 (0.00004591)
	0.01259613	0.00831360 (0.00012293)	0.00470183 (0.00006299)
	0.01570675	0.00888397 (0.00011538)	0.00678648 (0.00010309)
	0.01650155	0.01176295 (0.00021810)	0.00879447 (0.00012847)
	0.02021511	0.01335728 (0.00033263)	0.00987586 (0.00017190)
	0.02186993	0.01399684 (0.00029145)	0.01283660 (0.00018841)
GridRoom-32	0.00000000	0.00003275 (0.00001147)	0.00000162 (0.00000016)
	0.00000059	0.00020833 (0.00003929)	0.00011673 (0.00000477)
	0.00000059	0.00023288 (0.00002495)	0.00019514 (0.00000604)
	0.00050062	0.00067326 (0.00003619)	0.00032908 (0.00000721)
	0.00160563	0.00147027 (0.00004066)	0.00055255 (0.00001161)
	0.00433879	0.00221590 (0.00020075)	0.00163562 (0.00002444)
	0.00527193	0.00252468 (0.00031296)	0.00185031 (0.00002668)
	0.00530446	0.00394143 (0.00023913)	0.00199837 (0.00003240)
	0.00885755	0.00459834 (0.00017481)	0.00271093 (0.00003983)
	0.00885782	0.00497534 (0.00029475)	0.00285472 (0.00004133)
	0.00890618	0.00614753 (0.00043570)	0.00364060 (0.00004219)
GridRoom-64	0.00000000	0.00001140 (0.00000275)	0.00000139 (0.00000017)

Table 16, continued.

Env	True	DALLO	ALLO
	0.00000002	0.00007677 (0.00001024)	0.00007887 (0.00000316)
	0.00000003	0.00009416 (0.00001176)	0.00014470 (0.00000514)
	0.00002744	0.00014416 (0.00001389)	0.00019857 (0.00000711)
	0.00002780	0.00016647 (0.00001543)	0.00020600 (0.00001051)
	0.00002835	0.00018560 (0.00002088)	0.00023060 (0.00000918)
	0.00002875	0.00023423 (0.00002452)	0.00032171 (0.00001299)
	0.00053843	0.00032422 (0.00002433)	0.00038927 (0.00001252)
	0.00053843	0.00036625 (0.00003810)	0.00051433 (0.00000971)
	0.00075490	0.00072058 (0.00007649)	0.00071934 (0.00001629)
	0.00111099	0.00095747 (0.00011511)	0.00093656 (0.00002444)
GridRoomSym-4	0.00000000	0.00001840 (0.00000244)	0.00000402 (0.00000066)
	0.00359666	0.00349436 (0.00005908)	0.00277779 (0.00004797)
	0.01621473	0.00968928 (0.00025766)	0.00719279 (0.00007025)
	0.02081805	0.01421779 (0.00027636)	0.01029303 (0.00009761)
	0.07410878	0.06336656 (0.00091801)	0.04391103 (0.00042526)
	0.07414064	0.06725901 (0.00034436)	0.05086080 (0.00049911)
	0.08338511	0.07291294 (0.00112546)	0.05323785 (0.00030794)
	0.08368590	0.07553200 (0.00033808)	0.06044861 (0.00035562)
	0.13513041	0.11473707 (0.00126989)	0.07474053 (0.00043157)
	0.13974242	0.11889674 (0.00136653)	0.07620632 (0.00052720)
	0.14973803	0.12752505 (0.00102769)	0.09009153 (0.00045778)

Table 17: Recovery of the 11 smallest eigenvalues on the directed problems under the *uniform* policy, including directed cycles followed by CycleNets. The "Chung" column reports the spectrum of the Chung Laplacian $I - P_{\mathrm{alt}}$. The $\widehat{L}_\gamma$ column reports h_γ^{-1} applied to the spectrum of ALLO's sampling-induced target $\widehat{L}_\gamma := I - \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)$. DALLO's learned eigenvalues are mapped to the undiscounted Chung scale using h_γ^{-1}. For ALLO, the same map is applied to its learned eigenvalues before comparison with the correspondingly rescaled $\widehat{L}_\gamma$ spectrum. DALLO and ALLO entries are mean (standard deviation) over 30 seeds; target eigenvalues are deterministic.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
Cycle-240	0.00000000	0.00000000	0.00000619 (0.00000110)	0.00000584 (0.00000107)
	0.01331562	0.05879593	0.01293048 (0.00006711)	0.03604102 (0.00017143)
	0.02168071	0.07353152	0.02098056 (0.00010971)	0.04280733 (0.00025034)
	0.02421033	0.08424921	0.02343901 (0.00012608)	0.04859305 (0.00023692)
	0.02984018	0.09320636	0.02889166 (0.00016575)	0.05440720 (0.00019909)
	0.03228776	0.11438018	0.03114604 (0.00015875)	0.06383640 (0.00030643)
	0.03582733	0.12995127	0.03456714 (0.00014005)	0.07244019 (0.00025543)
	0.04235960	0.13709231	0.04071844 (0.00022150)	0.07700226 (0.00055596)
	0.04486419	0.15463486	0.04312686 (0.00031816)	0.08634520 (0.00038449)
	0.04863386	0.16545393	0.04677128 (0.00023341)	0.09156418 (0.00032465)
	0.05869287	0.18780427	0.05607271 (0.00036546)	0.10154408 (0.00039774)
Cycle-500	0.00000000	0.00000000	0.00000759 (0.00000137)	0.00000487 (0.00000083)
	0.01078207	0.03579169	0.01096791 (0.00015325)	0.02240506 (0.00017194)
	0.01231492	0.04442046	0.01204628 (0.00008324)	0.02412760 (0.00015338)
	0.01598959	0.05631393	0.01570473 (0.00009397)	0.03567198 (0.00024820)
	0.02006617	0.06942443	0.01952347 (0.00007965)	0.04287175 (0.00021475)
	0.02043504	0.07542029	0.01986000 (0.00012312)	0.04661420 (0.00035783)
	0.02409353	0.08345002	0.02354119 (0.00016268)	0.04774446 (0.00043959)
	0.02614480	0.08827895	0.02530554 (0.00014072)	0.05240646 (0.00038364)
	0.03052780	0.09260306	0.02935664 (0.00013725)	0.05624470 (0.00026410)
	0.03290864	0.10380571	0.03196695 (0.00025308)	0.05979560 (0.00033869)
	0.03337459	0.11070016	0.03231071 (0.00022491)	0.06324735 (0.00033253)
Cycle-600	0.00000000	0.00000000	0.00000828 (0.00000179)	0.00000494 (0.00000088)
	0.01073920	0.03974356	0.01063414 (0.00006572)	0.02656531 (0.00022025)
	0.01178963	0.04505667	0.01165914 (0.00007209)	0.03009252 (0.00024531)
	0.01475239	0.05288119	0.01454248 (0.00007352)	0.03358672 (0.00019687)
	0.01638611	0.06032696	0.01610360 (0.00005204)	0.03849268 (0.00022428)

Table 17, continued.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
	0.01843088	0.06468217	0.01824931 (0.00013605)	0.03969806 (0.00024304)
	0.01996029	0.07797228	0.02006640 (0.00011872)	0.04594879 (0.00035510)
	0.02116076	0.08171127	0.02061378 (0.00007191)	0.04837864 (0.00029569)
	0.02175763	0.08578664	0.02132925 (0.00009115)	0.05078710 (0.00031019)
	0.02418198	0.09687618	0.02347740 (0.00008908)	0.05275528 (0.00034671)
	0.02783531	0.10103929	0.02685972 (0.00011931)	0.05619319 (0.00044743)
Cycle-1100	0.00000000	0.00000000	0.00000781 (0.00000214)	0.00000463 (0.00000069)
	0.01161490	0.04206436	0.01146059 (0.00007885)	0.02676624 (0.00017260)
	0.01269368	0.04480165	0.01251836 (0.00006573)	0.02743490 (0.00024416)
	0.01367443	0.04682053	0.01350050 (0.00005105)	0.03085577 (0.00018876)
	0.01449753	0.05563184	0.01438868 (0.00007062)	0.03431251 (0.00028538)
	0.01490707	0.05584028	0.01475145 (0.00006381)	0.03601266 (0.00024515)
	0.01613887	0.06320903	0.01593664 (0.00008993)	0.03784458 (0.00021643)
	0.01751373	0.06916103	0.01729283 (0.00007606)	0.04107481 (0.00032896)
	0.01926896	0.07007811	0.01905387 (0.00009917)	0.04326924 (0.00027005)
	0.01996552	0.07285689	0.01985374 (0.00010341)	0.04396829 (0.00024103)
	0.02152110	0.07586270	0.02125595 (0.00009100)	0.04621685 (0.00027347)
Cycle-1300	0.00000000	0.00000000	0.00000890 (0.00000203)	0.00000433 (0.00000063)
	0.00885046	0.03410928	0.00913168 (0.00009999)	0.02262882 (0.00017762)
	0.01063759	0.04037764	0.01054005 (0.00007034)	0.02550520 (0.00021577)
	0.01176763	0.04049247	0.01165760 (0.00006556)	0.02634781 (0.00025531)
	0.01346963	0.04936918	0.01330472 (0.00006692)	0.03196148 (0.00030936)
	0.01437075	0.05376714	0.01418825 (0.00006294)	0.03408907 (0.00034597)
	0.01509634	0.05443173	0.01490490 (0.00007415)	0.03494120 (0.00023502)
	0.01624919	0.05757558	0.01663039 (0.00023512)	0.03638245 (0.00020120)
	0.01709016	0.06112023	0.01682524 (0.00009751)	0.03776285 (0.00030231)
	0.01760256	0.06659748	0.01733865 (0.00006361)	0.03952280 (0.00031039)
	0.01862183	0.06771989	0.01853025 (0.00008220)	0.04200561 (0.00031321)
CycleNet-2	0.00000000	0.00000000	0.00001051 (0.00000219)	0.00000205 (0.00000035)
	0.00003388	0.00046851	0.00017845 (0.00004323)	0.00038056 (0.00002265)
	0.00003586	0.00063514	0.00023671 (0.00006585)	0.00041553 (0.00001555)
	0.00009030	0.00066257	0.00034628 (0.00007906)	0.00043487 (0.00001315)
	0.00009505	0.00099934	0.00038835 (0.00008669)	0.00066145 (0.00002068)
	0.00012348	0.00133736	0.00042380 (0.00009076)	0.00086980 (0.00002990)
	0.00014029	0.00233048	0.00045774 (0.00010389)	0.00129636 (0.00001710)
	0.00017173	0.00255958	0.00050364 (0.00009511)	0.00142590 (0.00001814)
	0.00022327	0.00319032	0.00059464 (0.00010305)	0.00172812 (0.00002913)
	0.00030001	0.00369265	0.00067105 (0.00010458)	0.00214064 (0.00004328)
	0.00036137	0.00543026	0.00075303 (0.00010499)	0.00285197 (0.00003321)
CycleNet-3	0.00000000	0.00000000	0.00000834 (0.00000206)	0.00000189 (0.00000039)
	0.00024838	0.00103304	0.00042041 (0.00006059)	0.00074431 (0.00002561)
	0.00031624	0.00132000	0.00051199 (0.00007713)	0.00096319 (0.00003464)
	0.00031687	0.00431329	0.00055854 (0.00009568)	0.00222823 (0.00003284)
	0.00041186	0.00495101	0.00067444 (0.00009955)	0.00273818 (0.00003456)
	0.00042050	0.00618500	0.00070265 (0.00010841)	0.00318887 (0.00004086)
	0.00071574	0.00718548	0.00099717 (0.00010745)	0.00398024 (0.00005027)
	0.00077845	0.00952832	0.00109916 (0.00011292)	0.00494811 (0.00009151)
	0.00102172	0.01132614	0.00134157 (0.00011606)	0.00624658 (0.00012013)
	0.00115884	0.01692111	0.00149412 (0.00012081)	0.00859592 (0.00006978)
	0.00150600	0.01714596	0.00185585 (0.00012105)	0.00906004 (0.00008177)
CycleNet-4	0.00000000	0.00000000	0.00000821 (0.00000181)	0.00000203 (0.00000035)
	0.00025405	0.00113137	0.00041403 (0.00006714)	0.00082138 (0.00002827)
	0.00025550	0.00131118	0.00047923 (0.00009988)	0.00095284 (0.00002780)
	0.00033498	0.00264261	0.00059277 (0.00011723)	0.00160805 (0.00002092)
	0.00033657	0.00291887	0.00062427 (0.00012778)	0.00165094 (0.00001519)
	0.00047628	0.00292214	0.00076841 (0.00013595)	0.00188383 (0.00005350)
	0.00051281	0.00408310	0.00081888 (0.00014248)	0.00225595 (0.00003100)
	0.00055042	0.00409238	0.00090798 (0.00014531)	0.00232318 (0.00003903)
	0.00070975	0.00549944	0.00106804 (0.00014712)	0.00311329 (0.00003872)

Table 17, continued.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
	0.00073951	0.00550515	0.00110718 (0.00015537)	0.00321093 (0.00004927)
	0.00100702	0.00779318	0.00138248 (0.00015378)	0.00436600 (0.00007324)
CycleNet-10	0.00000000	0.00000000	0.00000791 (0.00000235)	0.00000214 (0.00000032)
	0.00034557	0.00060951	0.00049327 (0.00007096)	0.00049607 (0.00001336)
	0.00037488	0.00069217	0.00059262 (0.00011118)	0.00055861 (0.00001534)
	0.00097589	0.00224234	0.00125117 (0.00012409)	0.00161999 (0.00002683)
	0.00098157	0.00243031	0.00128503 (0.00013579)	0.00176196 (0.00004230)
	0.00135007	0.00416412	0.00161806 (0.00013662)	0.00296379 (0.00005441)
	0.00139750	0.00452386	0.00171162 (0.00014850)	0.00323051 (0.00007163)
	0.00140130	0.00565845	0.00173789 (0.00014958)	0.00403755 (0.00008559)
	0.00143005	0.00665201	0.00176968 (0.00016220)	0.00471876 (0.00011223)
	0.00185026	0.00843855	0.00219995 (0.00015184)	0.00600485 (0.00015864)
	0.00185631	0.01399193	0.00223232 (0.00015972)	0.00738135 (0.00008667)

Table 18: Recovery of the 11 smallest eigenvalues on the directed problems under the *near-uniform* policy, including directed cycles followed by CycleNets. The "Chung" column reports the spectrum of the Chung Laplacian $I - P_{\mathrm{alt}}$. The $\widehat{L}_\gamma$ column reports h_γ^{-1} applied to the spectrum of ALLO's sampling-induced target $\widehat{L}_\gamma := I - \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)$. DALLO's learned eigenvalues are mapped to the undiscounted Chung scale using h_γ^{-1}. For ALLO, the same map is applied to its learned eigenvalues before comparison with the correspondingly rescaled $\widehat{L}_\gamma$ spectrum. DALLO and ALLO entries are mean (standard deviation) over 30 seeds; target eigenvalues are deterministic.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
Cycle-240	0.00000000	0.00000000	0.00000630 (0.00000137)	0.00000587 (0.00000103)
	0.01313723	0.05856360	0.01275194 (0.00006454)	0.03596437 (0.00016495)
	0.02166519	0.07293620	0.02096232 (0.00012056)	0.04242797 (0.00024345)
	0.02416627	0.08390951	0.02340161 (0.00011308)	0.04849825 (0.00023909)
	0.02972382	0.09324328	0.02874607 (0.00018121)	0.05433491 (0.00022371)
	0.03210619	0.11561890	0.03097026 (0.00018836)	0.06425152 (0.00028411)
	0.03592109	0.13069786	0.03459257 (0.00018153)	0.07273751 (0.00032372)
	0.04258047	0.13872671	0.04088952 (0.00026435)	0.07761547 (0.00053349)
	0.04524136	0.15571213	0.04349126 (0.00030037)	0.08659725 (0.00037541)
	0.04873363	0.16619411	0.04683546 (0.00029589)	0.09178890 (0.00034773)
	0.05905459	0.18752910	0.05638844 (0.00039058)	0.10127684 (0.00043198)
Cycle-500	0.00000000	0.00000000	0.00000728 (0.00000150)	0.00000485 (0.00000086)
	0.01077019	0.03522600	0.01092407 (0.00011402)	0.02217326 (0.00017582)
	0.01216191	0.04432716	0.01192296 (0.00007323)	0.02392690 (0.00016246)
	0.01582539	0.05628016	0.01558305 (0.00008716)	0.03568566 (0.00023003)
	0.02006087	0.06904856	0.01953063 (0.00008867)	0.04271536 (0.00022331)
	0.02023336	0.07569364	0.01973127 (0.00008984)	0.04660183 (0.00038699)
	0.02413160	0.08395088	0.02362970 (0.00013957)	0.04789140 (0.00048168)
	0.02624829	0.08911570	0.02543383 (0.00016091)	0.05271085 (0.00045471)
	0.03051082	0.09321869	0.02942157 (0.00016147)	0.05651149 (0.00033817)
	0.03305154	0.10513443	0.03218046 (0.00024139)	0.06007580 (0.00037706)
	0.03359386	0.11183923	0.03252554 (0.00024082)	0.06382004 (0.00039137)
Cycle-600	0.00000000	0.00000000	0.00000816 (0.00000183)	0.00000502 (0.00000084)
	0.01069676	0.03982011	0.01060333 (0.00007844)	0.02660754 (0.00022288)
	0.01172616	0.04500590	0.01162300 (0.00007539)	0.03008622 (0.00024222)
	0.01475120	0.05282895	0.01455602 (0.00007707)	0.03360814 (0.00018543)
	0.01642502	0.06075405	0.01614598 (0.00007760)	0.03860631 (0.00020547)
	0.01845524	0.06527188	0.01829966 (0.00012621)	0.03995645 (0.00025334)
	0.01997355	0.07830512	0.02006829 (0.00008390)	0.04607070 (0.00034101)
	0.02112277	0.08225915	0.02058561 (0.00009740)	0.04848826 (0.00029491)
	0.02173326	0.08662642	0.02135201 (0.00008631)	0.05098204 (0.00033689)
	0.02418024	0.09787511	0.02348061 (0.00011540)	0.05283351 (0.00034639)
	0.02793820	0.10176069	0.02696129 (0.00015786)	0.05622229 (0.00047826)

Table 18, continued.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
Cycle-1100	0.00000000	0.00000000	0.00000795 (0.00000173)	0.00000466 (0.00000066)
	0.01159841	0.04221075	0.01143939 (0.00006742)	0.02685410 (0.00016020)
	0.01252795	0.04495658	0.01233956 (0.00006357)	0.02747230 (0.00021446)
	0.01359740	0.04636687	0.01343011 (0.00004529)	0.03066224 (0.00016948)
	0.01443451	0.05561369	0.01428646 (0.00006413)	0.03451528 (0.00025036)
	0.01473405	0.05631235	0.01456737 (0.00007318)	0.03612480 (0.00025380)
	0.01602948	0.06332851	0.01581237 (0.00006101)	0.03795971 (0.00021022)
	0.01742235	0.06936191	0.01719623 (0.00006179)	0.04113144 (0.00036574)
	0.01928583	0.07058393	0.01902806 (0.00008511)	0.04337318 (0.00026015)
	0.02004812	0.07389346	0.01990568 (0.00009430)	0.04422233 (0.00028121)
	0.02153077	0.07626958	0.02126433 (0.00010444)	0.04643566 (0.00028411)
Cycle-1300	0.00000000	0.00000000	0.00000919 (0.00000236)	0.00000437 (0.00000068)
	0.00886480	0.03399849	0.00917228 (0.00010923)	0.02256220 (0.00018422)
	0.01055449	0.04038974	0.01047015 (0.00006377)	0.02560167 (0.00024114)
	0.01163523	0.04070397	0.01153053 (0.00006539)	0.02631270 (0.00023450)
	0.01340603	0.04933432	0.01326304 (0.00007485)	0.03191112 (0.00027009)
	0.01430787	0.05381474	0.01414275 (0.00006383)	0.03409934 (0.00033805)
	0.01505254	0.05437421	0.01487187 (0.00006192)	0.03487582 (0.00024477)
	0.01620959	0.05787165	0.01664651 (0.00018418)	0.03652045 (0.00021744)
	0.01700038	0.06104862	0.01678008 (0.00009322)	0.03778745 (0.00027814)
	0.01752110	0.06687376	0.01728828 (0.00008173)	0.03968204 (0.00033897)
	0.01858899	0.06794404	0.01849674 (0.00008735)	0.04204919 (0.00025983)
CycleNet-2	0.00000000	0.00000000	0.00001077 (0.00000277)	0.00000209 (0.00000030)
	0.00003384	0.00050848	0.00018513 (0.00006378)	0.00040007 (0.00001543)
	0.00003582	0.00063506	0.00025121 (0.00009589)	0.00042057 (0.00001387)
	0.00009126	0.00066338	0.00035726 (0.00011118)	0.00043730 (0.00001535)
	0.00009584	0.00102339	0.00039640 (0.00011999)	0.00067444 (0.00001720)
	0.00012345	0.00139280	0.00043702 (0.00012395)	0.00090157 (0.00002668)
	0.00014012	0.00232925	0.00046391 (0.00013201)	0.00129694 (0.00001415)
	0.00017358	0.00256458	0.00051972 (0.00013330)	0.00142720 (0.00001742)
	0.00022523	0.00320460	0.00060791 (0.00013513)	0.00173346 (0.00002705)
	0.00029977	0.00376508	0.00068246 (0.00014577)	0.00218574 (0.00004015)
	0.00036138	0.00543785	0.00076870 (0.00013557)	0.00285288 (0.00002865)
CycleNet-3	0.00000000	0.00000000	0.00000844 (0.00000188)	0.00000189 (0.00000037)
	0.00024923	0.00113029	0.00042263 (0.00005784)	0.00079896 (0.00002544)
	0.00031710	0.00144560	0.00052691 (0.00008419)	0.00103698 (0.00003555)
	0.00032413	0.00432694	0.00057138 (0.00009940)	0.00222764 (0.00003306)
	0.00041418	0.00500852	0.00069092 (0.00010962)	0.00277555 (0.00003420)
	0.00043078	0.00620619	0.00072102 (0.00011849)	0.00319555 (0.00004015)
	0.00072561	0.00728591	0.00101802 (0.00011514)	0.00404807 (0.00005024)
	0.00079092	0.00954835	0.00112046 (0.00012491)	0.00496397 (0.00009281)
	0.00102327	0.01156151	0.00135483 (0.00012670)	0.00640917 (0.00012537)
	0.00116751	0.01694813	0.00151449 (0.00013001)	0.00860470 (0.00006785)
	0.00151062	0.01720498	0.00187128 (0.00013246)	0.00911465 (0.00008265)
CycleNet-4	0.00000000	0.00000000	0.00000803 (0.00000159)	0.00000201 (0.00000037)
	0.00025675	0.00124332	0.00041015 (0.00005140)	0.00089101 (0.00002553)
	0.00025826	0.00143987	0.00046999 (0.00007681)	0.00103268 (0.00002442)
	0.00033769	0.00291159	0.00058193 (0.00008807)	0.00162504 (0.00002098)
	0.00033929	0.00294904	0.00061214 (0.00009676)	0.00166670 (0.00001373)
	0.00049365	0.00295270	0.00076772 (0.00010416)	0.00205372 (0.00005523)
	0.00052905	0.00416819	0.00082192 (0.00010422)	0.00229896 (0.00002609)
	0.00054941	0.00417785	0.00088416 (0.00011264)	0.00236666 (0.00003238)
	0.00071307	0.00564430	0.00104991 (0.00011270)	0.00321480 (0.00003891)
	0.00073990	0.00565103	0.00108407 (0.00011106)	0.00331324 (0.00004904)
	0.00102440	0.00799919	0.00138140 (0.00011502)	0.00449717 (0.00007029)
CycleNet-10	0.00000000	0.00000000	0.00000805 (0.00000210)	0.00000217 (0.00000031)
	0.00036527	0.00067438	0.00052007 (0.00005934)	0.00053853 (0.00001407)
	0.00039484	0.00076487	0.00061945 (0.00010002)	0.00060680 (0.00001480)
	0.00098140	0.00247931	0.00126779 (0.00010395)	0.00177797 (0.00003026)

Table 18, continued.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
	0.00098709	0.00268580	0.00129908 (0.00011658)	0.00193173 (0.00004537)
	0.00141361	0.00460151	0.00171482 (0.00012111)	0.00326135 (0.00005954)
	0.00141710	0.00499845	0.00174478 (0.00012240)	0.00355179 (0.00007502)
	0.00143025	0.00625127	0.00177514 (0.00012928)	0.00444452 (0.00009115)
	0.00151168	0.00735060	0.00184092 (0.00013771)	0.00519821 (0.00011513)
	0.00187610	0.00932452	0.00223732 (0.00012814)	0.00661856 (0.00017143)
	0.00188186	0.01403248	0.00226718 (0.00013591)	0.00740075 (0.00007728)

Table 19: Recovery of the 11 smallest eigenvalues on the directed problems under the *biased* policy, including directed cycles followed by CycleNets. The "Chung" column reports the spectrum of the Chung Laplacian $I - P_{\mathrm{alt}}$. The $\widehat{L}_\gamma$ column reports h_γ^{-1} applied to the spectrum of ALLO's sampling-induced target $\widehat{L}_\gamma := I - \frac{1}{2}(P_\pi^\gamma + (P_\pi^{\mathrm{rev}})^\gamma)$. DALLO's learned eigenvalues are mapped to the undiscounted Chung scale using h_γ^{-1}. For ALLO, the same map is applied to its learned eigenvalues before comparison with the correspondingly rescaled $\widehat{L}_\gamma$ spectrum. DALLO and ALLO entries are mean (standard deviation) over 30 seeds; target eigenvalues are deterministic.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
Cycle-240	0.00000000	0.00000000	0.00000668 (0.00000138)	0.00000595 (0.00000106)
	0.01386975	0.06002583	0.01345602 (0.00006800)	0.03680407 (0.00020272)
	0.02096478	0.07180437	0.02026213 (0.00010376)	0.04166669 (0.00021222)
	0.02445856	0.08380816	0.02369451 (0.00015869)	0.04834053 (0.00033630)
	0.02935982	0.09212091	0.02836669 (0.00014737)	0.05258581 (0.00025845)
	0.03226065	0.12162820	0.03120248 (0.00015198)	0.06562298 (0.00034207)
	0.03714953	0.13876062	0.03574687 (0.00017009)	0.07579464 (0.00042413)
	0.04382594	0.14539889	0.04211157 (0.00023209)	0.07801656 (0.00043985)
	0.04720781	0.15908473	0.04528618 (0.00027378)	0.08640844 (0.00037779)
	0.05127352	0.16820773	0.04916584 (0.00031302)	0.09346904 (0.00037545)
	0.06143589	0.18852153	0.05872458 (0.00042269)	0.09836566 (0.00051851)
Cycle-500	0.00000000	0.00000000	0.00000685 (0.00000135)	0.00000489 (0.00000080)
	0.01024271	0.02906502	0.01023279 (0.00011270)	0.01949035 (0.00013517)
	0.01034777	0.04132977	0.01025757 (0.00009107)	0.02130556 (0.00021550)
	0.01443985	0.05266094	0.01430128 (0.00010617)	0.03429578 (0.00029371)
	0.01784020	0.06472886	0.01742251 (0.00011012)	0.04040230 (0.00015057)
	0.02011549	0.07716164	0.01968157 (0.00009489)	0.04677831 (0.00028967)
	0.02448882	0.08362396	0.02405194 (0.00013231)	0.04754410 (0.00036287)
	0.02593752	0.09557431	0.02521221 (0.00012404)	0.05505002 (0.00041244)
	0.02967094	0.09893031	0.02878079 (0.00013140)	0.05860835 (0.00033336)
	0.03355566	0.11159646	0.03236683 (0.00018821)	0.06244153 (0.00042907)
	0.03487382	0.11611232	0.03391580 (0.00020850)	0.06491878 (0.00049253)
Cycle-600	0.00000000	0.00000000	0.00000801 (0.00000172)	0.00000515 (0.00000086)
	0.01054735	0.03965270	0.01051242 (0.00008905)	0.02600568 (0.00024404)
	0.01211043	0.04425324	0.01192919 (0.00005660)	0.02975548 (0.00021886)
	0.01497864	0.05623778	0.01483447 (0.00011351)	0.03480854 (0.00025468)
	0.01599400	0.06042777	0.01568988 (0.00005382)	0.03848481 (0.00026783)
	0.01861803	0.06582780	0.01825089 (0.00008061)	0.03929857 (0.00022238)
	0.01957613	0.07665005	0.01950509 (0.00010234)	0.04652185 (0.00032457)
	0.02083194	0.08330139	0.02046269 (0.00008275)	0.04798428 (0.00024586)
	0.02191202	0.09161469	0.02149895 (0.00010445)	0.05158131 (0.00029647)
	0.02410115	0.10081960	0.02345245 (0.00011131)	0.05376702 (0.00051341)
	0.02743140	0.10716356	0.02654855 (0.00014774)	0.05584347 (0.00048360)
Cycle-1100	0.00000000	0.00000000	0.00000735 (0.00000166)	0.00000473 (0.00000074)
	0.01063711	0.04086509	0.01048077 (0.00005744)	0.02626281 (0.00016176)
	0.01177246	0.04291655	0.01167061 (0.00004026)	0.02808150 (0.00019015)
	0.01268134	0.04754609	0.01260772 (0.00005591)	0.02945454 (0.00021051)
	0.01311128	0.05471695	0.01300384 (0.00008052)	0.03477480 (0.00026889)
	0.01424621	0.06100700	0.01412812 (0.00008156)	0.03755329 (0.00026001)
	0.01510738	0.06310229	0.01497062 (0.00004848)	0.03829436 (0.00025449)

Table 19, continued.

Env	Chung	$\widehat{L}_\gamma$	DALLO	ALLO
	0.01729600	0.06795450	0.01717331 (0.00010438)	0.04105508 (0.00035708)
	0.01871143	0.07222442	0.01839502 (0.00005343)	0.04175192 (0.00026174)
	0.01992062	0.07770637	0.01973761 (0.00008386)	0.04538784 (0.00024912)
	0.02128790	0.07883953	0.02106884 (0.00015712)	0.04698281 (0.00032847)
Cycle-1300	0.00000000	0.00000000	0.00000806 (0.00000153)	0.00000432 (0.00000064)
	0.00880592	0.03272584	0.00901521 (0.00007070)	0.02182270 (0.00017706)
	0.00942135	0.03858038	0.00937685 (0.00005959)	0.02516056 (0.00020419)
	0.01053359	0.04139073	0.01048357 (0.00006248)	0.02586099 (0.00023466)
	0.01252237	0.04840525	0.01243892 (0.00006049)	0.03150739 (0.00023897)
	0.01350898	0.05330043	0.01338126 (0.00005684)	0.03381009 (0.00035234)
	0.01415856	0.05392088	0.01402404 (0.00005492)	0.03431988 (0.00026385)
	0.01520047	0.05941063	0.01542327 (0.00010846)	0.03684344 (0.00036450)
	0.01612817	0.05994961	0.01605311 (0.00008050)	0.03755059 (0.00023419)
	0.01648977	0.06595592	0.01628114 (0.00007034)	0.04032938 (0.00032130)
	0.01798414	0.06809296	0.01775518 (0.00007399)	0.04152146 (0.00036761)
CycleNet-2	0.00000000	0.00000000	0.00001086 (0.00000187)	0.00000222 (0.00000038)
	0.00003097	0.00058713	0.00017548 (0.00004027)	0.00038012 (0.00001335)
	0.00003243	0.00061280	0.00023075 (0.00006002)	0.00041165 (0.00001387)
	0.00010187	0.00153284	0.00034507 (0.00007069)	0.00079895 (0.00001587)
	0.00010650	0.00164242	0.00039264 (0.00007982)	0.00088889 (0.00001662)
	0.00012061	0.00223569	0.00042294 (0.00008194)	0.00127107 (0.00001395)
	0.00012736	0.00240011	0.00045791 (0.00008492)	0.00132750 (0.00001695)
	0.00020743	0.00317312	0.00052451 (0.00009273)	0.00156012 (0.00001910)
	0.00026837	0.00402630	0.00061783 (0.00008538)	0.00189189 (0.00002059)
	0.00027724	0.00519734	0.00066297 (0.00009581)	0.00263280 (0.00002939)
	0.00035763	0.00608100	0.00077562 (0.00009656)	0.00319637 (0.00002804)
CycleNet-3	0.00000000	0.00000000	0.00001137 (0.00000238)	0.00000195 (0.00000039)
	0.00029071	0.00467118	0.00049083 (0.00005472)	0.00198762 (0.00003209)
	0.00035943	0.00490095	0.00059111 (0.00008035)	0.00208216 (0.00002770)
	0.00041103	0.00635273	0.00071813 (0.00009598)	0.00284433 (0.00002900)
	0.00050064	0.00649497	0.00079358 (0.00010217)	0.00303688 (0.00002998)
	0.00058772	0.00767971	0.00092453 (0.00010620)	0.00386392 (0.00003630)
	0.00074779	0.00963816	0.00107434 (0.00010712)	0.00495865 (0.00003404)
	0.00091298	0.01058613	0.00129979 (0.00011348)	0.00633720 (0.00005400)
	0.00097305	0.01482454	0.00133124 (0.00011276)	0.00759043 (0.00005253)
	0.00128004	0.01983837	0.00166390 (0.00011370)	0.01102560 (0.00007995)
	0.00160360	0.02335262	0.00197085 (0.00011387)	0.01288636 (0.00006231)
CycleNet-4	0.00000000	0.00000000	0.00000730 (0.00000200)	0.00000208 (0.00000035)
	0.00018078	0.00236491	0.00030427 (0.00006574)	0.00134742 (0.00001799)
	0.00018305	0.00238124	0.00036576 (0.00010118)	0.00138919 (0.00001337)
	0.00036466	0.00573511	0.00059293 (0.00011739)	0.00309036 (0.00003030)
	0.00045787	0.00664438	0.00071247 (0.00012659)	0.00324879 (0.00002708)
	0.00045976	0.00730249	0.00073474 (0.00013931)	0.00349060 (0.00002888)
	0.00066962	0.00742733	0.00096227 (0.00014099)	0.00355472 (0.00002413)
	0.00067645	0.00744874	0.00099209 (0.00014731)	0.00377904 (0.00003728)
	0.00069919	0.01215763	0.00102868 (0.00014794)	0.00555519 (0.00004137)
	0.00072334	0.01217033	0.00106443 (0.00015976)	0.00601770 (0.00004516)
	0.00094846	0.01252355	0.00131241 (0.00015966)	0.00613627 (0.00004488)
CycleNet-10	0.00000000	0.00000000	0.00000795 (0.00000195)	0.00000215 (0.00000036)
	0.00064579	0.00504421	0.00077719 (0.00006255)	0.00288125 (0.00002651)
	0.00065582	0.00537048	0.00084885 (0.00008823)	0.00315415 (0.00003228)
	0.00084708	0.01035990	0.00109749 (0.00010902)	0.00577725 (0.00006031)
	0.00085538	0.01043824	0.00112830 (0.00012262)	0.00590492 (0.00005029)
	0.00128484	0.01650263	0.00155362 (0.00012226)	0.00899090 (0.00005343)
	0.00130083	0.01654816	0.00158326 (0.00012963)	0.00915140 (0.00005806)
	0.00185228	0.01811212	0.00216661 (0.00013007)	0.01063477 (0.00007860)
	0.00200224	0.01908839	0.00230367 (0.00013017)	0.01131241 (0.00008364)
	0.00201201	0.02010679	0.00232488 (0.00013509)	0.01160484 (0.00005172)
	0.00246168	0.02016825	0.00279671 (0.00014043)	0.01183760 (0.00006576)

K ALLO 1000-STEP CONTROL

To test whether the difference between DALLO's 1000-step trajectories and ALLO's released 50-step episodic protocol explains the observed performance gap, we rerun ALLO with a matched 1000-step collection horizon. We keep ALLO's objective, updates, hyperparameters, and total transition budget unchanged, and evaluate both $\gamma = 0.9$ and the one-step control $\gamma = 0$ on all CycleNets under the three policies. This control uses 10 seeds.

We report three comparisons: DALLO versus matched-horizon ALLO, released ALLO versus matched-horizon ALLO at $\gamma = 0$, and released ALLO versus matched-horizon ALLO at $\gamma = 0.9$. For each comparison, we report the oracle, practical, and raw readouts.

Table 20: CycleNet eigenvector recovery under the matched 1000-step collection horizon using the oracle readout. Chung coordinates are recovered as $D_\pi^{1/2} f_{\mathrm{DALLO}}$, $D_\pi^{1/2} f_{\mathrm{ALLO},0.9}$, and $D_\pi^{1/2} f_{\mathrm{ALLO},0}$. DALLO and ALLO both use 1000-step trajectories. Results are mean (standard deviation) over 10 seeds per method. The reported t-statistic and p-value compare $D_\pi^{1/2} f_{\mathrm{DALLO}}$ with $D_\pi^{1/2} f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO $+ d_\pi$	ALLO$_{1000}$ $+ \gamma +$ d_π	ALLO$_{1000}$ $+ d_\pi$	t-statistic	p-value
CycleNet-2	uniform	**0.7927 (0.1152)**	0.3011 (0.0226)	0.3655 (0.0870)	9.4	0.000
	near-uni	**0.7689 (0.0955)**	0.2983 (0.0254)	0.3602 (0.0372)	12.6	0.000
	biased	**0.6374 (0.0866)**	0.9543 (0.0208)	0.3675 (0.0696)	7.7	0.000
CycleNet-3	uniform	**0.8902 (0.1495)**	0.3218 (0.0105)	0.6235 (0.0978)	4.7	0.000
	near-uni	**0.8500 (0.1196)**	0.3214 (0.0087)	0.6629 (0.1097)	3.6	0.002
	biased	**0.9708 (0.0473)**	0.4457 (0.0039)	0.6352 (0.1269)	7.8	0.000
CycleNet-4	uniform	**0.8295 (0.0818)**	0.1216 (0.0064)	0.5118 (0.1031)	7.6	0.000
	near-uni	**0.7786 (0.0809)**	0.1288 (0.0128)	0.4826 (0.0830)	8.1	0.000
	biased	**0.7572 (0.1105)**	0.3638 (0.0520)	0.4762 (0.1219)	5.4	0.000
CycleNet-10	uniform	**0.8355 (0.0797)**	0.2871 (0.0115)	0.5828 (0.1460)	4.8	0.000
	near-uni	0.7066 (0.1239)	0.2918 (0.0088)	0.6212 (0.0914)	1.8	0.098
	biased	**0.8519 (0.0679)**	0.4930 (0.0594)	0.6484 (0.0894)	5.7	0.000

Table 21: CycleNet eigenvector recovery under the matched 1000-step collection horizon using practical readouts $\hat{D}_{\mathrm{D}}^{1/2} f_{\mathrm{DALLO}}$, $\hat{D}_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0.9}$, and $\hat{D}_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0}$, with each $\hat{d}^\pi$ estimated from the corresponding replay buffer. DALLO and ALLO both use 1000-step trajectories. Results are mean (standard deviation) over 10 seeds per method. The reported t-statistic and p-value compare $\hat{D}_{\mathrm{D}}^{1/2} f_{\mathrm{DALLO}}$ with $\hat{D}_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO $+ \hat{d}_{\mathrm{D}}^\pi$	ALLO$_{1000}$ $+ \gamma + \hat{d}_{\mathrm{A}}^\pi$	ALLO$_{1000}$ $+ \hat{d}_{\mathrm{A}}^\pi$	t-statistic	p-value
CycleNet-2	uniform	**0.7926 (0.1153)**	0.3012 (0.0225)	0.3657 (0.0870)	9.3	0.000
	near-uni	**0.7689 (0.0954)**	0.2984 (0.0253)	0.3601 (0.0370)	12.6	0.000
	biased	**0.6382 (0.0868)**	0.9534 (0.0211)	0.3675 (0.0698)	7.7	0.000
CycleNet-3	uniform	**0.8903 (0.1493)**	0.3217 (0.0103)	0.6233 (0.0977)	4.7	0.000
	near-uni	**0.8501 (0.1196)**	0.3212 (0.0087)	0.6626 (0.1100)	3.6	0.002
	biased	**0.9708 (0.0473)**	0.4459 (0.0039)	0.6348 (0.1270)	7.8	0.000
CycleNet-4	uniform	**0.8294 (0.0818)**	0.1215 (0.0065)	0.5122 (0.1029)	7.6	0.000
	near-uni	**0.7786 (0.0809)**	0.1286 (0.0130)	0.4821 (0.0835)	8.1	0.000
	biased	**0.7571 (0.1105)**	0.3635 (0.0516)	0.4762 (0.1219)	5.4	0.000
CycleNet-10	uniform	**0.8353 (0.0797)**	0.2865 (0.0114)	0.5827 (0.1459)	4.8	0.000
	near-uni	0.7065 (0.1238)	0.2916 (0.0090)	0.6210 (0.0914)	1.8	0.097
	biased	**0.8518 (0.0679)**	0.4930 (0.0592)	0.6483 (0.0894)	5.7	0.000

Table 22: CycleNet eigenvector recovery under the matched 1000-step collection horizon using the raw network outputs f_{DALLO}, $f_{\mathrm{ALLO},0.9}$, and $f_{\mathrm{ALLO},0}$, without stationary-distribution reweighting. DALLO and ALLO both use 1000-step trajectories. Results are mean (standard deviation) over 10 seeds per method. The reported t-statistic and p-value compare f_{DALLO} with $f_{\mathrm{ALLO},0}$ using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	DALLO – raw	$\mathrm{ALLO}_{1000} + \gamma$ – raw	ALLO_{1000} – raw	t-statistic	p-value
CycleNet-2	uniform	**0.7924 (0.1152)**	0.3023 (0.0228)	0.3652 (0.0876)	9.3	0.000
	near-uni	**0.7685 (0.0960)**	0.2995 (0.0253)	0.3606 (0.0367)	12.5	0.000
	biased	**0.6322 (0.0870)**	0.9282 (0.0202)	0.3713 (0.0670)	7.5	0.000
CycleNet-3	uniform	**0.8900 (0.1495)**	0.3215 (0.0107)	0.6231 (0.0978)	4.7	0.000
	near-uni	**0.8497 (0.1196)**	0.3210 (0.0089)	0.6626 (0.1098)	3.6	0.002
	biased	**0.9568 (0.0465)**	0.4462 (0.0039)	0.6247 (0.1238)	7.9	0.000
CycleNet-4	uniform	**0.8292 (0.0817)**	0.1213 (0.0063)	0.5112 (0.1030)	7.6	0.000
	near-uni	**0.7782 (0.0809)**	0.1285 (0.0126)	0.4823 (0.0834)	8.1	0.000
	biased	**0.7412 (0.1081)**	0.3551 (0.0506)	0.4674 (0.1195)	5.4	0.000
CycleNet-10	uniform	**0.8352 (0.0797)**	0.2869 (0.0114)	0.5826 (0.1459)	4.8	0.000
	near-uni	0.7063 (0.1238)	0.2916 (0.0087)	0.6209 (0.0913)	1.8	0.098
	biased	**0.8344 (0.0663)**	0.4789 (0.0573)	0.6337 (0.0872)	5.8	0.000

Table 23: CycleNet eigenvector recovery comparing ALLO's released 50-step protocol with the matched 1000-step collection horizon at $\gamma = 0$, using the practical readout $D_{\mathrm{A}}^{1/2} f_{\mathrm{ALLO},0}$. The density estimate $\hat{d}_{\mathrm{A}}^{\pi}$ is fitted separately to each run's replay buffer. Results are mean (standard deviation) over 10 seeds per setting. The reported t-statistic and p-value compare the two collection horizons using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	$\mathrm{ALLO}_{50} + d_\pi$	$\mathrm{ALLO}_{1000} + d_\pi$	t-statistic	p-value
CycleNet-2	uniform	0.3890 (0.0671)	0.3655 (0.0870)	0.7	0.508
	near-uniform	0.3248 (0.0711)	0.3602 (0.0372)	-1.4	0.185
	biased	0.3831 (0.0714)	0.3675 (0.0696)	0.5	0.626
CycleNet-3	uniform	0.6824 (0.1353)	0.6235 (0.0978)	1.1	0.281
	near-uniform	0.6265 (0.1121)	0.6629 (0.1097)	-0.7	0.473
	biased	0.5983 (0.1355)	0.6352 (0.1269)	-0.6	0.537
CycleNet-4	uniform	0.4877 (0.0519)	0.5118 (0.1031)	-0.7	0.521
	near-uniform	0.4501 (0.0787)	0.4826 (0.0830)	-0.9	0.381
	biased	0.4361 (0.0804)	0.4762 (0.1219)	-0.9	0.398
CycleNet-10	uniform	0.6332 (0.1004)	0.5828 (0.1460)	0.9	0.381
	near-uniform	0.6080 (0.1092)	0.6212 (0.0914)	-0.3	0.772
	biased	0.6386 (0.1826)	0.6484 (0.0894)	-0.2	0.882

Table 24: CycleNet eigenvector recovery comparing ALLO's released 50-step protocol with the 1000-step collection horizon using practical readouts. The two settings use $\hat{D}_{\mathrm{A},50}^{1/2} f_{\mathrm{ALLO},0}$ and $\hat{D}_{\mathrm{A},1000}^{1/2} f_{\mathrm{ALLO},0}$, where $\hat{d}_{\mathrm{A},50}^{\pi}$ and $\hat{d}_{\mathrm{A},1000}^{\pi}$ are fitted separately to their respective replay buffers. Results are mean (standard deviation) over 10 seeds per setting. The reported t-statistic and p-value compare the two collection horizons using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	$\mathrm{ALLO}_{50} + \hat{d}_{\mathrm{A},50}^{\pi}$	$\mathrm{ALLO}_{1000} + \hat{d}_{\mathrm{A},1000}^{\pi}$	t-statistic	p-value
CycleNet-2	uniform	0.3879 (0.0673)	0.3657 (0.0870)	0.6	0.533
	near-uniform	0.3244 (0.0712)	0.3601 (0.0370)	-1.4	0.181
	biased	0.3747 (0.0684)	0.3675 (0.0698)	0.2	0.818
CycleNet-3	uniform	0.6825 (0.1354)	0.6233 (0.0977)	1.1	0.279
	near-uniform	0.6264 (0.1123)	0.6626 (0.1100)	-0.7	0.476
	biased	0.5986 (0.1348)	0.6348 (0.1270)	-0.6	0.544
CycleNet-4	uniform	0.4872 (0.0522)	0.5122 (0.1029)	-0.7	0.505
	near-uniform	0.4500 (0.0788)	0.4821 (0.0835)	-0.9	0.389
	biased	0.4334 (0.0799)	0.4762 (0.1219)	-0.9	0.368
CycleNet-10	uniform	0.6332 (0.1001)	0.5827 (0.1459)	0.9	0.380
	near-uniform	0.6078 (0.1092)	0.6210 (0.0914)	-0.3	0.773
	biased	0.6384 (0.1824)	0.6483 (0.0894)	-0.2	0.880

Table 25: CycleNet eigenvector recovery comparing ALLO's released 50-step protocol with the 1000-step collection horizon at $\gamma = 0$, using the raw outputs $f_{\mathrm{ALLO},0}$ without stationary-distribution reweighting. Results are mean (standard deviation) over 10 seeds per setting. The reported t-statistic and p-value compare the two collection horizons using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	ALLO_{50} – raw	ALLO_{1000} – raw	t-statistic	p-value
CycleNet-2	uniform	0.3880 (0.0670)	0.3652 (0.0876)	0.7	0.522
	near-uniform	0.3243 (0.0711)	0.3606 (0.0367)	-1.4	0.174
	biased	0.3742 (0.0686)	0.3713 (0.0670)	0.1	0.926
CycleNet-3	uniform	0.6822 (0.1355)	0.6231 (0.0978)	1.1	0.280
	near-uniform	0.6265 (0.1121)	0.6626 (0.1098)	-0.7	0.477
	biased	0.5945 (0.1340)	0.6247 (0.1238)	-0.5	0.607
CycleNet-4	uniform	0.4876 (0.0519)	0.5112 (0.1030)	-0.6	0.529
	near-uniform	0.4498 (0.0787)	0.4823 (0.0834)	-0.9	0.382
	biased	0.4249 (0.0785)	0.4674 (0.1195)	-0.9	0.362
CycleNet-10	uniform	0.6329 (0.1004)	0.5826 (0.1459)	0.9	0.382
	near-uniform	0.6077 (0.1091)	0.6209 (0.0913)	-0.3	0.772
	biased	0.6248 (0.1796)	0.6337 (0.0872)	-0.1	0.889

Table 26: CycleNet eigenvector recovery comparing ALLO's released 50-step protocol with the 1000-step collection horizon at $\gamma = 0.9$, using the oracle readout $D_\pi^{1/2} f_{\mathrm{ALLO},0.9}$. Results are mean (standard deviation) over 10 seeds per setting. The reported t-statistic and p-value compare the two collection horizons using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	$\mathrm{ALLO}_{50} + \gamma + d_\pi$	$\mathrm{ALLO}_{1000} + \gamma + d_\pi$	t-statistic	p-value
CycleNet-2	uniform	0.3646 (0.0721)	0.3011 (0.0226)	2.7	0.023
	near-uniform	**0.4352 (0.0852)**	0.2983 (0.0254)	4.9	0.001
	biased	0.7847 (0.0485)	**0.9543 (0.0208)**	-10.2	0.000
CycleNet-3	uniform	0.3085 (0.0112)	0.3218 (0.0105)	-2.8	0.013
	near-uniform	0.3073 (0.0106)	**0.3214 (0.0087)**	-3.2	0.005
	biased	**0.4744 (0.0140)**	0.4457 (0.0039)	6.3	0.000
CycleNet-4	uniform	0.1338 (0.0143)	0.1216 (0.0064)	2.5	0.030
	near-uniform	0.1361 (0.0107)	0.1288 (0.0128)	1.4	0.187
	biased	**0.5509 (0.0863)**	0.3638 (0.0520)	5.9	0.000
CycleNet-10	uniform	0.2919 (0.0088)	0.2871 (0.0115)	1.1	0.307
	near-uniform	0.2926 (0.0096)	0.2918 (0.0088)	0.2	0.844
	biased	0.5286 (0.0344)	0.4930 (0.0594)	1.6	0.123

Table 27: CycleNet eigenvector recovery comparing ALLO's released 50-step protocol with the 1000-step collection horizon at $\gamma = 0.9$, using practical readouts $\hat{D}_{\mathrm{A},50}^{1/2} f_{\mathrm{ALLO},0.9}$ and $\hat{D}_{\mathrm{A},1000}^{1/2} f_{\mathrm{ALLO},0.9}$. The estimates $\hat{d}_{\mathrm{A},50}^{\pi}$ and $\hat{d}_{\mathrm{A},1000}^{\pi}$ are fitted separately to their respective replay buffers. Results are mean (standard deviation) over 10 seeds per setting. The reported t-statistic and p-value compare the two collection horizons using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	$\mathrm{ALLO}_{50} + \gamma + \hat{d}_{\mathrm{A},50}^{\pi}$	$\mathrm{ALLO}_{1000} + \gamma + \hat{d}_{\mathrm{A},1000}^{\pi}$	t-statistic	p-value
CycleNet-2	uniform	0.3658 (0.0718)	0.3012 (0.0225)	2.7	0.021
	near-uniform	**0.4363 (0.0848)**	0.2984 (0.0253)	4.9	0.001
	biased	0.7856 (0.0485)	**0.9534 (0.0211)**	-10.0	0.000
CycleNet-3	uniform	0.3086 (0.0112)	0.3217 (0.0103)	-2.7	0.014
	near-uniform	0.3075 (0.0106)	**0.3212 (0.0087)**	-3.2	0.006
	biased	**0.4810 (0.0137)**	0.4459 (0.0039)	7.8	0.000
CycleNet-4	uniform	0.1337 (0.0143)	0.1215 (0.0065)	2.5	0.029
	near-uniform	0.1358 (0.0107)	0.1286 (0.0130)	1.3	0.197
	biased	**0.5427 (0.0859)**	0.3635 (0.0516)	5.7	0.000
CycleNet-10	uniform	0.2921 (0.0088)	0.2865 (0.0114)	1.2	0.232
	near-uniform	0.2930 (0.0096)	0.2916 (0.0090)	0.3	0.744
	biased	0.5284 (0.0344)	0.4930 (0.0592)	1.6	0.124

Table 28: CycleNet eigenvector recovery comparing ALLO's released 50-step protocol with the 1000-step collection horizon at $\gamma = 0.9$, using the raw outputs $f_{\mathrm{ALLO},0.9}$ without stationary-distribution reweighting. Results are mean (standard deviation) over 10 seeds per setting. The reported t-statistic and p-value compare the two collection horizons using Welch's two-sample t-test; the larger mean is bold when $p < 0.01$.

Env	Policy	$\mathrm{ALLO}_{50} + \gamma$ – raw	$\mathrm{ALLO}_{1000} + \gamma$ – raw	t-statistic	p-value
CycleNet-2	uniform	0.3657 (0.0718)	0.3023 (0.0228)	2.7	0.022
	near-uniform	**0.4361 (0.0848)**	0.2995 (0.0253)	4.9	0.001
	biased	0.7806 (0.0483)	**0.9282 (0.0202)**	-8.9	0.000
CycleNet-3	uniform	0.3082 (0.0109)	0.3215 (0.0107)	-2.8	0.013
	near-uniform	0.3069 (0.0102)	**0.3210 (0.0089)**	-3.3	0.004
	biased	**0.4766 (0.0131)**	0.4462 (0.0039)	7.0	0.000
CycleNet-4	uniform	0.1335 (0.0140)	0.1213 (0.0063)	2.5	0.027
	near-uniform	0.1357 (0.0104)	0.1285 (0.0126)	1.4	0.179
	biased	**0.5233 (0.0840)**	0.3551 (0.0506)	5.4	0.000
CycleNet-10	uniform	0.2917 (0.0088)	0.2869 (0.0114)	1.1	0.308
	near-uniform	0.2924 (0.0095)	0.2916 (0.0087)	0.2	0.845
	biased	0.5138 (0.0332)	0.4789 (0.0573)	1.7	0.117

L Effect of the Discount Parameter

The discount parameter γ controls the geometric lag used by DALLO. As discussed in Section 4.2, the discounted operator in equation 3 preserves the eigenvectors of P_{alt}, while the spectral map in equation 4 expands small spectral gaps increasingly as γ grows. Thus, changing γ does not change the target eigenvectors, but can change the conditioning of their recovery.

We evaluate $\gamma \in \{0.90, 0.95, 0.98\}$ on CycleNet-2, CycleNet-3, CycleNet-4, and CycleNet-10 under the uniform, near-uniform, and biased policies. Each configuration is trained for 200,000 gradient steps using the same ten seeds across all three values of γ, with identical replay buffers across corresponding runs. Recovery is evaluated every 10,000 steps using the metric described in Section I.

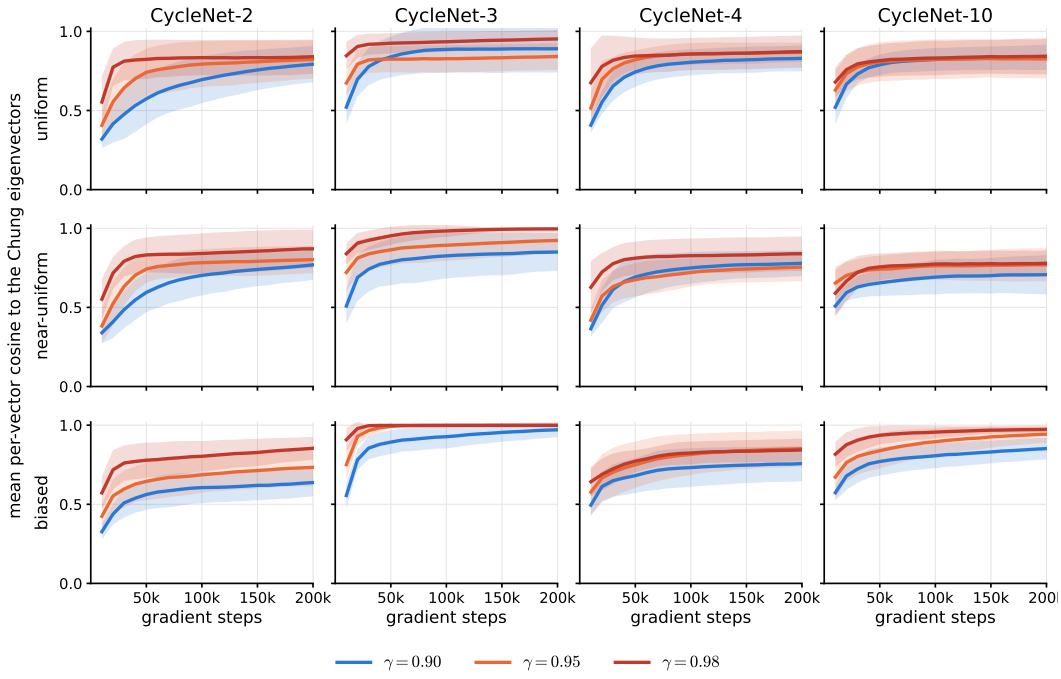

Figure 6: DALLO eigenvector recovery for different discount parameters. Rows correspond to the uniform, near-uniform, and biased policies, and columns correspond to CycleNet-2, CycleNet-3, CycleNet-4, and CycleNet-10. Curves show the mean over ten matched seeds and shaded regions show one standard deviation. Recovery is evaluated every 10,000 gradient steps for 200,000 training steps.

Larger γ improves early-training recovery across these problems. At 20k steps, $\gamma = 0.98$ exceeds $\gamma = 0.90$ in all 12 environment-policy settings, with a mean improvement of 0.20. Similarly, $\gamma = 0.95$ exceeds $\gamma = 0.90$ in all 12 settings at 20k steps, with a mean improvement of 0.10. For $\gamma = 0.98$, the mean advantage over $\gamma = 0.90$ decreases to 0.08 at 200k steps. The advantage of larger γ is therefore strongest early in training, consistent with the spectral-conditioning analysis, while final recovery can also differ across γ.

Larger γ also increases the expected sampled lag, since $\mathbb{E}[K] = 1/(1-\gamma)$. In our implementation, $K \leftarrow \min(K, K_{\max})$, with $K_{\max} = 100$ for $\gamma \leq 0.95$ and $K_{\max} = 300$ for $\gamma = 0.98$. Relative to the untruncated sampler, this clamping leaves the target eigenvectors unchanged and perturbs the discounted eigenvalues by at most $\gamma^{K_{\max}} \leq 6 \times 10^{-3}$ in our experiments.