Persistent-Transient Policy Evaluation for Markov Chains via Minimal Peripheral Quotients

Yang Xu
Purdue University, USA
xu1720@purdue.edu

Vaneet Aggarwal
Purdue University, USA
vaneet@purdue.edu

Abstract

We study fixed-policy evaluation for finite Markov chains that may be reducible and periodic. Classical evaluation methods with gain and bias decomposition are not always diagnostic: the gain records only invariant Cesàro averages, while persistent phase-dependent behavior is absorbed into the bias together with genuinely transient effects. We identify the real peripheral invariant subspace $\mathcal{K}(P)$ of the transition matrix P as the source of this ambiguity. Quotienting by $\mathcal{K}(P)$ is the minimal exact quotient that removes all non-decaying modes and makes the remaining dynamics strictly stable. After choosing a gauge projection Π with kernel $\mathcal{K}(P)$, the reward admits a unique decomposition $r = g^\star_\Pi + (I - P)v^\star_\Pi$, where $g^\star_\Pi$ is a persistent regime profile and $v^\star_\Pi$ is a gauge-fixed transient component. An exact comparison with classical normalized gain and bias shows that the new pair reallocates the same information into a persistent profile and a gauge-fixed representative of the transient quotient. This decomposition reconstructs finite-horizon returns, recovers statewise average reward, admits a transient-cost interpretation, and yields a stable estimator under a generative model.

1 Introduction

Average-reward dynamic programming and reinforcement learning evaluate sequential decisions without discounting, and classical theory uses Poisson equations, gain vectors, and bias functions to describe long-run performance [16, 5, 15, 21]. Recent work has extended planning and learning guarantees beyond the unichain setting and into weakly communicating or multichain regimes [28, 13, 29]. We focus on the evaluation problem after a stationary policy has already been fixed. The resulting object is a finite-state Markov reward process with transition matrix P and reward vector r. The central question is: what should summarize the performance of this policy from each initial state when the induced chain may be reducible and periodic?

A useful answer should separate three roles. It should recover the statewise long-term average reward, since different initial states may reach different recurrent classes. It should then describe the persistent regime that the chain eventually exhibits, including cyclic phase structure that repeats forever. It should also describe the transient cost paid before that persistent regime is reached. These roles are distinct in simple examples: a periodic chain can repeat a nonconstant reward pattern forever, a multichain model requires statewise evaluation because recurrent outcomes depend on the start state, and a gridworld random walk can make two policies with the same eventual average behave very differently over finite horizons. Policy evaluation beyond ergodicity should therefore distinguish what persists forever from what is paid only on the way there.

This distinction matters in computing systems with engineered cycles. RotorNet uses a fixed cyclic optical-connectivity schedule; its hardware implementation reports recurring transmission-error bursts at reconfiguration boundaries and rotor defects. In one TCP test, a defect occupying 2% of a usable slot reduced throughput by 11% relative to ideal service [14]. These recurring losses call for schedule,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

guard-interval, or hardware changes, whereas a one-time initialization backlog calls for a different intervention. Opus similarly reconfigures optical rails at communication-phase boundaries within distributed ML training iterations [8], connecting phase-aware accounting to current large-scale ML infrastructure. An average-throughput summary cannot by itself distinguish recurring service limitations, initial phase alignment, and temporary startup cost. A finite model can retain queue or protocol state together with schedule phase: deterministic phase advancement yields a periodic Markov chain even when arrivals and service are stochastic. Distinct closed operating configurations can additionally make outcomes depend on initialization. These examples motivate our evaluation question; they are not deployments of our estimator. Appendix I gives a concrete model and explains the scope of exact versus approximate periodicity.

The classical language for this problem for the Markov chain P with reward function r is the gain-bias Poisson equation in the form of $r = \rho + (I - P)h$, where ρ is the gain vector and h is the bias function [16, 5, 15]. This formalism is already valid for fixed-policy evaluation in general finite chains, and recent multichain fixed-point methods also build on gain-bias structure [29]. However, the gain vector stores only invariant long-run information. The classical gain is the Cesàro average $\rho = P^{\infty}r$, and therefore satisfies $P\rho = \rho$. On each closed irreducible recurrent class, every such invariant vector is constant, so the classical gain collapses cyclic phases to their time average. It can record the average of a periodic reward pattern, but not the phase-resolved pattern itself. Persistent behavior that changes from step to step, such as an alternating reward on a cycle, lasts forever but is not invariant. The only remaining place for this behavior in the classical pair is the bias h.

This creates the entanglement issue. In an ideal diagnostic decomposition, the long-run component would contain all behavior that persists, while the transient component would contain only effects that fade away. Classical gain and bias do not satisfy this principle once periodic persistent structure appears: the invariant gain keeps only the Cesàro part, and the non-invariant persistent remainder is absorbed into the bias. The resulting issue is not that returns cannot be evaluated, but that the classical decomposition does not tell us cleanly what caused them. The practical issue is therefore not loss of evaluability but loss of diagnosability. A practitioner needs to know whether poor finite-horizon behavior is caused by the eventual regime or by the path to that regime, because these imply different interventions. A deterministic cycle makes the point sharp. In such a chain, the process starts inside its recurrent regime and cycles forever. Nevertheless, the invariant gain collapses the phase pattern into an average, so the remaining oscillation must be represented in the bias. Thus a large classical bias can arise even when there is no genuine transient burden to diagnose.

Our answer is to organize evaluation by dynamical persistence. Let $\mathcal{K}(P)$ be the real peripheral invariant subspace of P, containing exactly the non-decaying modes of the chain, which includes the invariant recurrent-class directions as well as periodic phase directions. The quotient by $\mathcal{K}(P)$ is not a design choice. Dynamically, it is the smallest invariant subspace that must be removed before strict contraction is possible. Semantically, it is exactly the collection of modes whose effects remain visible at arbitrarily long horizons and therefore should not be called transient. We prove this minimality statement and then use the quotient to define a unique decomposition $r = g^{\star}_{\Pi} + (I - P)v^{\star}_{\Pi}$ after a gauge projection Π with $\ker(\Pi) = \mathcal{K}(P)$ is fixed. The vector $g^{\star}_{\Pi}$ is the persistent regime profile: its invariant part recovers the classical statewise average reward, while its orbit $\{P^t g^{\star}_{\Pi}\}_{t\geq 0}$ records recurrent-class and phase behavior that the invariant gain discards. The vector $v^{\star}_{\Pi}$ represents the quotient in which all non-decaying modes have been removed. Under the anchor gauge used in this paper, it has an explicit transient-to-regime cost interpretation. A key feature of this viewpoint is that it reorganizes the same Poisson information as classical approaches into two parts with different meanings. We prove that the invariant part of $g^{\star}_{\Pi}$ is exactly the classical gain, and that the classical normalized bias equals the new transient component plus a unique non-invariant peripheral correction. Thus the decomposition makes persistent directions explicit while representing the transient quotient in a chosen gauge. In aperiodic multichain models, where no non-invariant peripheral modes exist, the distinction disappears and the framework reduces to the classical picture up to gauge. The structural viewpoint also has an algorithmic payoff. Because the peripheral quotient restores contraction, it gives a stable target for sample-based estimation under a generative model. Our estimator learns recurrent classes, cyclic phases, anchors, and phase-offset absorption weights; constructs an anchor gauge; solves the projected quotient fixed-point equation for $v^{\star}_{\Pi}$; and reconstructs $g^{\star}_{\Pi}$ from anchor residuals. The output recovers average reward, exposes persistent regime behavior, and quantifies transient-to-regime cost. In this sense, the contribution is not a new way to compute returns. It is

a new way to attribute them. Notably, a detailed comparison of our work with existing literature is given in Appendix A.

Contributions. We introduce a policy-evaluation framework for general finite Markov chains that removes the classical entanglement between persistent non-invariant behavior and transient effects.

- We identify $\mathcal{K}(P)$ as the minimal real P-invariant subspace whose quotient dynamics are strictly contractive.
- We define the gauge-fixed decomposition $r = g^\star_\Pi + (I - P)v^\star_\Pi$ into a persistent regime profile and a transient component.
- We prove an exact comparison with classical normalized gain and bias, showing that the pair reorganizes rather than augments Poisson information.
- We show that the decomposition reconstructs finite-horizon returns, recovers long-run average reward, yields an anchor-gauge transient-cost interpretation, and can be estimated from generative-model samples.

2 Formulation

We fix a stationary policy in a finite Markov decision process and evaluate the induced Markov reward process. The state space is $\mathcal{S} = \{1, 2, \ldots, n\}$, the transition matrix is $P(s'|s) = \sum_{a\in\mathcal{A}} \pi(a|s)\mathcal{P}(s'|s, a)$, and the one step reward is $r(s) = \sum_{a\in\mathcal{A}} \pi(a|s)r_{\mathrm{mdp}}(s, a)$. Thus every object below depends only on the pair (P, r). For an initial state s and a horizon $T \geq 1$, let $J_T(s) = \mathbb{E}_s[\sum_{t=0}^{T-1} r(X_t)]$, or in vector form $J_T = \sum_{t=0}^{T-1} P^t r$. The statewise long run average reward profile is

$$g(s) \coloneqq \lim_{T\to\infty} \frac{1}{T} J_T(s), \tag{1}$$

which exists for every finite Markov chain [16, 5]. The question is how to represent the same fixed policy in a way that distinguishes persistent regime behavior from transient cost. This distinction is about representation rather than about existence of returns. The sequence J_T is always defined, and the average reward profile g is always defined in the finite state setting. What is not automatic is a decomposition whose two pieces answer two different diagnostic questions: what regime behavior remains visible at long horizons, and what finite accumulated effect is caused by the path before that regime is reached.

2.1 Classical normalized gain and bias

The classical representation uses the Cesàro limiting projection $P^\infty = \lim_{T\to\infty} T^{-1} \sum_{t=0}^{T-1} P^t$, whose existence is standard for finite Markov chains [15, 16], and the gain $\rho = P^\infty r$. A normalized bias h is any solution of

$$r = \rho + (I - P)h, \qquad P^\infty h = 0. \tag{2}$$

This pair is well defined in finite state models and is the standard language of average reward evaluation. Its limitation is semantic rather than algebraic. The gain ρ is invariant, so it can only record quantities that remain unchanged after one transition. On a periodic recurrent class, however, the policy may exhibit a reward pattern that persists forever while changing with phase. Such behavior is not transient, but it is also not invariant. The next standard lemma explains why the invariant gain cannot store it [15, 16].

Lemma 1. *Let F be a finite closed irreducible class of P. If x satisfies $Px = x$ on F, then x is constant on F.*

Lemma 1 explains why the classical gain collapses every periodic recurrent class to a single number. A nonconstant reward pattern on a deterministic cycle is not transient, but it cannot appear in ρ. It must therefore be stored in h together with the cost of reaching the recurrent regime. This is the entanglement that the rest of the paper removes. In a unichain aperiodic model this issue is hidden, because the only persistent directions are invariant. When a recurrent class is periodic, however, the set of persistent directions is larger than the invariant subspace. The goal is therefore to keep the classical average reward as one output, but to refine it with the additional persistent coordinates that are visible along the actual trajectory.

3 Minimal peripheral quotient and gauge-fixed decomposition

This section identifies the part of the chain that should be removed before a transient fixed point is solved. The issue is not only how to stabilize an iteration. A direction that does not decay under repeated application of P should not be interpreted as a finite transient correction, because its effect remains visible at arbitrarily long horizons. The persistent subspace below collects exactly these directions. Once it is quotiented out, the remaining dynamics are strictly contractive and can support a unique transient representative.

3.1 The peripheral invariant subspace

Let P act on $\mathbb{C}^n$ and let $\mathcal{K}_{\mathbb{C}}(P)$ be the generalized invariant subspace spanned by eigenvalues with modulus one. The real peripheral invariant subspace is $\mathcal{K}(P) \coloneqq \mathcal{K}_{\mathbb{C}}(P) \cap \mathbb{R}^n$. For a finite Markov chain, this space consists exactly of the modes that do not decay under repeated application of P. It includes the invariant recurrent class directions and, when a class is periodic, the phase directions that oscillate forever. This is the subspace that should be treated as persistent.

3.2 The minimal contractive quotient

For any real P invariant subspace K, write $\bar{P}_K$ for the induced operator on $\mathbb{R}^n/K$. We call K an exact stabilizing quotient if $\rho(\bar{P}_K) < 1$, where $\rho(\cdot)$ denotes spectral radius.

Theorem 1 (Minimal peripheral quotient)**.** *For a real P invariant subspace K, the following are equivalent: $\rho(\bar{P}_K) < 1$; $\mathcal{K}(P) \subseteq K$; and every non decaying mode of P is zero in $\mathbb{R}^n/K$. Hence $\mathcal{K}(P)$ is the unique minimal real invariant subspace whose quotient dynamics are strictly stable.*

The theorem has two consequences. First, it rules out smaller quotients. If a phase direction or recurrent class direction survives in the quotient, then the quotient still contains a mode that does not decay, so no strict contraction can hold. Second, it rules out unnecessary larger quotients. Removing more than $\mathcal{K}(P)$ would also remove directions that are transient in the dynamical sense. Thus the same subspace is forced by both sides of the paper: it is the smallest object needed for stable fixed point analysis, and it is exactly the object that should be treated as persistent in evaluation.

Theorem 2 (Quotient contraction)**.** *On $\mathbb{R}^n/\mathcal{K}(P)$ there exists a norm $\|\cdot\|_{\mathrm{q}}$ and a constant $\gamma < 1$ such that $\|\bar{P}x\|_{\mathrm{q}} \le \gamma\|x\|_{\mathrm{q}}$ for every quotient class x. Equivalently, the pullback semi-norm on $\mathbb{R}^n$ satisfies $\|Pv\|_{\mathrm{q}} \le \gamma\|v\|_{\mathrm{q}}$ for every $v \in \mathbb{R}^n$.*

This is the computational counterpart of Theorem 1. After quotienting by $\mathcal{K}(P)$, the Bellman type map no longer has unit modulus directions left to obstruct contraction. The norm in Theorem 2 is not meant to be a new evaluation metric for users. It is a proof device showing that the transient part is a stable fixed point problem once all persistent directions have been removed. The proofs of Theorems 1 and 2 are given in Appendices F.1 and F.2.

3.3 Gauge fixed decomposition

The quotient fixed point identifies only an equivalence class modulo $\mathcal{K}(P)$. This is the analogue of fixing an additive constant in classical average reward theory, except that the nullspace is now the whole persistent subspace rather than only constants. To obtain a concrete vector with evaluation meaning, choose a projection Π with $\ker(\Pi) = \mathcal{K}(P)$ and let $\mathrm{range}(\Pi)$ be the gauge slice. Define the projected Bellman map

$$\mathcal{T}_\Pi(v) \coloneqq \Pi(r + Pv). \tag{3}$$

Theorem 3 (Gauge fixed persistent transient decomposition)**.** *For every reward r and every projection Π with kernel $\mathcal{K}(P)$, there is a unique $v_\Pi^\star \in \mathrm{range}(\Pi)$ satisfying $v_\Pi^\star = \mathcal{T}_\Pi(v_\Pi^\star)$. With $g_\Pi^\star \coloneqq r - (I - P)v_\Pi^\star$, we have*

$$r = g_\Pi^\star + (I - P)v_\Pi^\star, \quad g_\Pi^\star \in \mathcal{K}(P), \quad v_\Pi^\star \in \mathrm{range}(\Pi), \tag{4}$$

$$v_\Pi^\star = \Pi(r + Pv_\Pi^\star), \tag{5}$$

$$g_\Pi^\star = r - (I - P)v_\Pi^\star \in \mathcal{K}(P), \tag{6}$$

$$[v_\Pi^\star] = [r] + \bar{P}[v_\Pi^\star]. \tag{7}$$

The theorem turns the quotient contraction into an evaluation decomposition. The vector $g^\star_\Pi$ is the part of the reward that remains in the persistent subspace, and $v^\star_\Pi$ is the transient representative selected by the gauge. The gauge is a normalization, not an additional modeling assumption. Different projections with kernel $\mathcal{K}(P)$ choose different representatives of the same quotient class, just as different normalizations choose different classical bias functions. The anchor gauge below is used because it gives the representative a direct hitting time interpretation. Here transience refers to the quotient: $\Pi P^t v^\star_\Pi \to 0$. An anchor representative need not itself decay under P^t, so its persistent profile and boundary correction must be interpreted jointly. The commuting spectral gauge gives literal separation into an asymptotic reward profile and a decaying remainder. Appendix I explains the distinction.

Corollary 1 (Recovery of average reward)**.** *Let g be the statewise average reward profile from* (1)*. Then $g = P^\infty g^\star_\Pi = P^\infty r$.*

Thus $g^\star_\Pi$ refines the classical average reward. Its invariant part is the usual gain, while its non invariant peripheral part records phase resolved persistent behavior.

Proposition 1 (Reduction to the aperiodic case)**.** *If every closed irreducible class of P is aperiodic, then $\mathcal{K}(P) = \mathrm{range}(P^\infty)$. In this case the persistent profile carries no non invariant phase information, and the decomposition reduces to the classical multichain gain bias picture up to the chosen normalization.*

3.4 Comparison with classical gain and bias

A natural concern is that the new pair may add information beyond the classical gain and bias. The following theorem shows that it does not. The decomposition contains the same Poisson information, but makes phase-dependent accounting explicit through the persistent profile and the selected gauge.

Theorem 4 (Comparison with classical gain and bias)**.** *Let $(g^\star_\Pi, v^\star_\Pi)$ be the decomposition from Theorem 3, and let (ρ, h) be the normalized classical pair from* (2)*. Then $P^\infty g^\star_\Pi = \rho$. Moreover, there is a unique $\psi_\Pi \in \mathcal{K}(P) \cap \ker(P^\infty)$ such that $g^\star_\Pi = \rho + (I-P)\psi_\Pi$, and the classical normalized bias satisfies $h = v^\star_\Pi + \psi_\Pi - P^\infty v^\star_\Pi$.*

Corollary 2 (Projected classical bias)**.** *The gauge projection recovers the transient component, $\Pi h = v^\star_\Pi$. Equivalently, $g^\star_\Pi = \rho + (I-P)(I-\Pi)h$.*

Theorem 4 is the formal link to the classical picture. The invariant part of $g^\star_\Pi$ is exactly the classical gain. The correction ψ_Π is the persistent phase component that the classical gain cannot store because it is restricted to invariant vectors. The classical bias therefore contains this correction together with the transient component. The term $P^\infty v^\star_\Pi$ is only a normalization effect and disappears after applying the gauge projection. Thus periodic phase information is explicit in the new accounting coordinates, while $v^\star_\Pi$ represents the transient quotient in the selected gauge.

3.5 Anchor gauge basis

The paper uses an anchor gauge because it turns the abstract quotient into concrete coordinates. Decompose the chain into transient states T and closed irreducible classes F_i, and let $\mathsf{F} = \cup_i F_i$ be the recurrent set. If F_i has period d_i, write its cyclic classes as $C_{i,0}, \dots, C_{i,d_i-1}$, indexed so that one step from $C_{i,\ell}$ enters $C_{i,\ell+1 \bmod d_i}$. For a recurrent state $y \in F_i$, define $\mathsf{cls}(y) = i$ and $\mathsf{phase}(y) = \ell$ when $y \in C_{i,\ell}$. Choose one anchor $a_{i,k} \in C_{i,k}$. Let $\mathcal{I} = \{(i,k) : 1 \le i \le m, 0 \le k < d_i\}$ and $N = |\mathcal{I}| = \sum_i d_i \le |\mathsf{F}| \le n$. The basis below is a phase corrected absorption basis. Starting from a transient state, it records which closed class is eventually reached and which cyclic phase is reached after subtracting the elapsed time. This correction makes the span of the basis invariant under one step, with $Pb_{i,k} = b_{i,k-1 \bmod d_i}$ on class i.

Definition 1 (Phase offset absorption basis)**.** *For $(i,k) \in \mathcal{I}$, define $b_{i,k}$ by $b_{i,k}(s) = \mathbf{1}\{s \in C_{i,k}\}$ on recurrent states. For $s \in \mathsf{T}$, let τ_F be the first hitting time of the recurrent set and set*

$$b_{i,k}(s) = \mathbb{P}_s\{\mathsf{cls}(X_{\tau_\mathsf{F}}) = i,\ \mathsf{phase}(X_{\tau_\mathsf{F}}) - \tau_\mathsf{F} \equiv k \pmod{d_i}\}. \tag{8}$$

Lemma 2 (Anchor gauge identifies the peripheral subspace)**.** *The functions $\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ form a basis of $\mathcal{K}(P)$. The projection*

$$(\Pi v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k}) b_{i,k}(s) \tag{9}$$

has kernel $\mathcal{K}(P)$ and satisfies

$$b_{i,k}(a_{j,\ell}) = \mathbf{1}\{(i,k) = (j,\ell)\}. \tag{10}$$

Lemma 2 is where the abstract peripheral subspace becomes an implementable gauge. The projection Πv subtracts the unique persistent profile that agrees with v on all anchors, so the remaining representative vanishes at those anchors. This is the normalization used in the transient cost interpretation below and the quantity learned by the sample based estimator in Section 5.

4 Evaluation consequences of the decomposition

The previous section constructs the decomposition. We now show what it says about returns. Finite-horizon return can be written as a persistent orbit plus a boundary correction, and this is stronger than an asymptotic statement since at every horizon, it separates the contribution of behavior that continues from the contribution of the transient representative. For any potential v, define $g_v := r - (I - P)v$.

Lemma 3 (Return identity)**.** *For every v and every $T \geq 1$,*

$$J_T = \sum_{t=0}^{T-1} P^t g_v + v - P^T v. \tag{11}$$

Applying Lemma 3 to $v^\star = v^\star_\Pi$ and $g^\star = g^\star_\Pi$ gives

$$J_T = \sum_{t=0}^{T-1} P^t g^\star + v^\star - P^T v^\star. \tag{12}$$

This identity is the main reason for separating $g^\star$ and $v^\star$. The sum $\sum_{t<T} P^t g^\star$ keeps the recurrent class and phase resolved persistent profile visible at each time. The term $v^\star - P^T v^\star$ is a boundary term. It can affect finite horizon returns, but it does not create a new long run rate. Thus the decomposition matches the diagnostic distinction in the introduction: persistent behavior contributes throughout the horizon, while the transient component enters through the path to the selected representative.

Under the anchor gauge, Theorem 3 and Lemma 2 give $v^\star(a_{i,k}) = 0$ for all $(i,k) \in \mathcal{I}$. Let $\tau_{\mathcal{A}}$ be the first hitting time of the anchor set.

Proposition 2 (Transient cost under the anchor gauge)**.** *For every state s,*

$$v^\star(s) = \mathbb{E}_s\left[\sum_{t=0}^{\tau_{\mathcal{A}}-1} \big(r(X_t) - g^\star(X_t)\big)\right]. \tag{13}$$

Thus $v^\star(s)$ is the cumulative excess reward, relative to the persistent profile, paid before the chain first reaches the anchor set. The anchors are only a normalization device, but they make the sign and magnitude of $v^\star$ interpretable. If the path to the eventual regime is costly relative to the persistent profile, the transient component records that cost. The next proposition is a sanity check for the interpretation. A deterministic cycle has no transient path once the initial state is fixed inside the cycle. Any nonzero classical bias in this example must therefore come from persistent phase structure rather than from transient cost.

Proposition 3 (Deterministic cycles have no transient burden)**.** *Let P be the deterministic cycle on d states, $P(k, k+1 \bmod d) = 1$. Then $\mathcal{K}(P) = \mathbb{R}^d$, so the anchor gauge has $\Pi = 0$, $v^\star_\Pi = 0$, and $g^\star_\Pi = r$. In contrast, the classical normalized pair is $\rho = \bar{r}\mathbf{1}$, with $\bar{r} = d^{-1}\sum_k r(k)$, and h solves $h(k) - h(k+1 \bmod d) = r(k) - \bar{r}$ with $\sum_k h(k) = 0$. Thus h is generally nonzero even though there is no transient path to diagnose. If d is even and $r(k) = 1$ for $k < d/2$ and $r(k) = 0$ otherwise, then $\|h\|_\infty = d/8$ while $v^\star_\Pi = 0$.*

Proposition 3 gives the sharpest separation from the classical bias: a large bias can be caused entirely by persistent phase structure.

Corollary 3 (Return error from decomposition error)**.** *Let $(g^\star, v^\star)$ be the exact decomposition and define $\widehat{J}_T = \sum_{t=0}^{T-1} P^t \widehat{g} + \widehat{v} - P^T \widehat{v}$ for any candidate pair $(\widehat{g}, \widehat{v})$. Then, for every $T \geq 1$,*

$$\|\widehat{J}_T - J_T\|_\infty \leq T\|\widehat{g} - g^\star\|_\infty + 2\|\widehat{v} - v^\star\|_\infty. \tag{14}$$

The corollary is the bridge to sample-based evaluation: estimating the persistent profile and the transient component gives a direct finite horizon return guarantee. The factor T in front of the persistent profile error is unavoidable because the persistent profile contributes at every time step, while the transient component appears only through two boundary terms. This matches the intended meaning of the two quantities. Proofs of the results in this section are given in Appendix D.

5 Sample-based estimation of the decomposition

We now define an evaluation report with two quantities, $g^\star_\Pi$ and $v^\star_\Pi$, and estimate them under a tabular generative access model. Since our decomposition depends on closed classes, periods, cyclic phases, and absorption probabilities from transient states. From a single trajectory, or from logged data that does not cover some branch, these objects are not identifiable for general chains. A generative model is one clean way to impose the needed coverage and other data models with comparable coverage could be used. The estimator then follows the same order as the decomposition: learn the persistent coordinates, learn the anchor gauge, solve the stable quotient equation, and reconstruct the persistent profile. The detailed algorithms and proofs formally stated in Appendices B and C while this section provides the statistical statement without technical constants. Numerical diagnostics and sample-based experiments are also reported in Appendix H.

5.1 Learning the persistent coordinates

Assume access to a simulator that returns an independent sample from $P(\cdot|s)$ for any queried state s. The first task is structural. The estimator samples every state, forms the empirical support graph, computes closed strongly connected components, and then computes the period and cyclic partition of each closed class. This gives estimates of F_i, d_i, $C_{i,k}$, and the anchors $a_{i,k}$. If $p_{\min}$ is the smallest positive transition probability, then $K \geq p_{\min}^{-1} \log(n^2/\delta)$ samples per state recover the support graph with probability at least $1-\delta$. On this event, the recurrent classes, periods, cyclic partitions, and anchors are correct up to cyclic relabeling. This recovery step is important because the persistent coordinates are not numerical values that can be chosen after the fact. They are determined by the communication structure and periods of the chain. If a closed class or a cyclic phase is missed, the estimator would project out the wrong subspace and the later contraction statement would apply to the wrong quotient.

The second task is to estimate the basis $b_{i,k}$ from Definition 1. For each transient state, the simulator runs independent episodes until the recurrent set is reached and records the terminal class together with the phase offset. The empirical frequencies define $\widehat{b}_{i,k}(s)$. With $N = |\mathcal{I}|$ and M episodes per transient state, the concentration bound in Appendix B.1 gives $\max_{s\in\mathsf{T}} \|\widehat{b}(s) - b(s)\|_1 \leq \varepsilon_b$ with high probability once M is of order $\varepsilon_b^{-2}(N + \log(|\mathsf{T}|/\delta))$. The learned projection is $\widehat{\Pi}v = v - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\widehat{b}_{i,k}$. On the same event, $\|(\widehat{\Pi} - \Pi)v\|_\infty \leq \varepsilon_b \|v\|_\infty$. This bound is the reason the estimator separates structural learning from numerical fixed point iteration. Once the projection is accurate, every subsequent error term can be measured as a perturbation of the stable quotient equation from Theorem 2.

The probabilistic meaning of $b_{i,k}$ is important for interpretation. Starting from a transient state, $b_{i,k}(s)$ is the probability that the chain eventually enters class F_i with the phase offset corresponding to anchor $a_{i,k}$. Thus the learned basis does not approximate an arbitrary spectral vector. It estimates the weights with which the initial state is assigned to persistent regimes. This is the point at which the abstract peripheral subspace becomes a concrete tabular object. The projection $\widehat{\Pi}$ removes the estimated persistent coordinates by subtracting the anchor values extended through these absorption probabilities.

Algorithm 1 is deliberately aligned with the decomposition. The first two steps estimate the gauge and therefore determine which coordinates are persistent. The fourth step estimates only the quotient fixed point, where contraction is available. The last step reconstructs $g^\star_\Pi$ from the residual $r + (P - I)v^\star_\Pi$ on anchors and then extends those coordinates through the estimated basis.

This order is useful because the three tasks have different statistical roles. Support recovery is a discrete structural problem. Basis estimation is a probability estimation problem over absorption events. Stochastic approximation is a numerical fixed point problem on a stable quotient. If these

Algorithm 1 Main estimator for persistent transient policy evaluation

Require: Generative model for P, reward vector r, budgets K, M, T, J

1: Learn the support graph, closed classes, periods, cyclic phases, and anchors.
2: Estimate the phase offset absorption basis $\widehat{b}_{i,k}$ by absorption episodes.
3: Form the learned anchor projection $\widehat{\Pi}v = v - \sum_{(i,k)} v(a_{i,k})\widehat{b}_{i,k}$.
4: Run projected stochastic approximation on the anchor gauge for T iterations.
5: Estimate anchor residuals with J samples per anchor and reconstruct $\widehat{g}$.
6: **return** Persistent profile $\widehat{g}$ and transient component $\widehat{v}$.

tasks are mixed together, the meaning of the resulting value vector becomes unclear. The estimator keeps them separate, so each error term can be traced back to one part of the evaluation report. This is also why the algorithm is suited to tabular policy evaluation. Once a stationary policy is fixed, the induced Markov reward process is the object being evaluated, and the recurrent classes, periods, and phases are properties of this policy induced process.

5.2 Projected stochastic approximation on the quotient

Let $W = \{v \in \mathbb{R}^n : v(a_{i,k}) = 0 \text{ for all } (i,k) \in \mathcal{I}\}$ be the anchored subspace. By Lemma 2, this subspace intersects $\mathcal{K}(P)$ only at zero, so the quotient semi-norm becomes a genuine norm on W. The exact projected map $v \mapsto \Pi(r + Pv)$ is contractive on this space. The learned map uses $\widehat{\Pi}$ and one sample of the next state for each coordinate. If $\widetilde{s} \sim P(\cdot|s)$, one stochastic update has the form

$$v_{t+1} = \widehat{\Pi}\big((1-\alpha_t)v_t + \alpha_t \widehat{T}(v_t)\big), \qquad \widehat{T}(v_t)(s) = r(s) + v_t(\widetilde{s}). \tag{15}$$

This update shows where the quotient matters. Without removing $\mathcal{K}(P)$, the fixed point map may have eigenvalues on the unit circle. After projection, the deterministic part is a contraction. With a small basis error, the learned projected operator has contraction factor $\gamma + O(\varepsilon_b)$ and remains stable whenever ε_b is small enough. This stability statement is the algorithmic counterpart of the minimality theorem. The estimator does not attempt to average away periodicity. It removes the full peripheral subspace, estimates the transient representative in the anchored complement, and later puts the persistent coordinates back into $\widehat{g}$.

The resulting transient component guarantee has the usual stochastic approximation form, plus the price of learning the gauge. Under the structural recovery event, the basis concentration event, a bounded second moment condition for the oracle noise, and a standard stepsize choice, Appendix B.2 proves

$$\mathbb{E}\|[v_T] - [v_\Pi^\star]\|_{\mathrm{q}} \le O(T^{-1/2}) + O(\varepsilon_b). \tag{16}$$

Equivalently, setting T of order ε^{-2} and choosing the basis accuracy at the same scale gives quotient error of order ε. The constants depend on the quotient contraction, the norm equivalence on W, and the noise moment bound. These dependencies are unavoidable because the algorithm estimates a fixed point in a norm induced by the quotient. They are also useful diagnostically. A small quotient error means that the estimated transient component is accurate after ignoring persistent coordinates, exactly as the theory requires. It does not require the stochastic approximation iterates to learn phase oscillations as part of the transient value.

Periodicity alone does not imply failure of temporal-difference (TD) learning: differential TD admits convergence guarantees under suitable unichain, coverage, and stepsize assumptions [23]. A converged classical bias can correctly encode persistent phase effects; our projected update instead targets $v_\Pi^\star = \Pi h$. For a deterministic two-state cycle with rewards $(0, 2)$, the normalized classical bias is $(-1/2, 1/2)$, whereas our transient component is zero: both evaluations are correct, but they attribute phase effects differently. The minimality theorem concerns the original transition dynamics; relaxation can stabilize non-invariant phase modes without removing them. General multichain evaluation poses a separate issue: a scalar gain cannot represent closed classes with unequal averages, and one continuing trajectory does not generally cover all classes or repeatedly visit transient states. Generative access supplies that coverage at the structural-learning costs stated here. Thus classical estimation followed by the same gauge postprocessing is a valid alternative; our results do not establish universal TD failure or superior sample efficiency.

5.3 Reconstructing the persistent profile and returns

Once v_T has been estimated, the persistent profile is recovered from anchor residuals. For each anchor, the estimator samples $Y \sim P(a_{i,k}, \cdot)$ and estimates $Pv_T(a_{i,k})$ by an empirical average. The anchor coordinate is then $\widehat{\theta}_{i,k} = r(a_{i,k}) + \widehat{Pv_T}(a_{i,k}) - v_T(a_{i,k})$, and the full profile is reconstructed as $\widehat{g}(s) = \sum_{(i,k)\in\mathcal{I}} \widehat{\theta}_{i,k}\widehat{b}_{i,k}(s)$. This formula mirrors Theorem 3: the residual lies in the peripheral subspace, and the anchor basis extends its coordinates from anchors to all states.

The persistent profile bound in Appendix B.2 states that, for J next state samples per anchor,

$$\mathbb{E}\|\widehat{g} - g^\star_\Pi\|_\infty \le C_1\mathbb{E}\|v_T - v^\star_\Pi\|_{\mathrm{q}} + C_2\sqrt{\frac{\log N}{J}} + C_3\varepsilon_b. \tag{17}$$

Combining this with (16) yields the end to end form $\mathbb{E}\|\widehat{g} - g^\star_\Pi\|_\infty \le O(T^{-1/2}) + O(\sqrt{\log N/J}) + O(\varepsilon_b)$. Together with Corollary 3, the estimator therefore controls finite horizon prediction by estimating the two components that have evaluation meaning. The query complexity has three parts: nK samples for the support graph, expected $|\mathsf{T}|MH_{\mathrm{abs}}$ samples for absorption episodes, where $H_{\mathrm{abs}} = \max_{s\in\mathsf{T}} \mathbb{E}_s[\tau_{\mathsf{F}}]$, and $nT + NJ$ samples for stochastic approximation and anchor residuals. Thus the theory is operational in tabular MDP evaluation: it estimates the average reward, the persistent phase resolved regime, and the transient to regime cost from the same generative access used by standard model based methods. The output can be used in two ways. For asymptotic evaluation, $P^\infty\widehat{g}$ estimates the statewise average reward. For diagnostic evaluation, the orbit of $\widehat{g}$, together with the boundary correction, provides phase-aware return accounting, and $\widehat{v}$ reports the transient cost of reaching the anchored regime. This is the sense in which the estimator implements the persistent transient decomposition rather than only solving an average reward equation.

This structural recovery step is part of the statistical target instead of merely a preprocessing convenience. The quotient and the anchor gauge need to be determined by the support graph through the closed classes, their periods, cyclic partitions, and anchors. If this map is incorrect, the learned projection may have a different kernel and the algorithm may solve a different quotient problem. Thus a finite-sample guarantee for the stated decomposition must work on a high-probability event where this structure is recovered correctly. The rare-edge parameter $p_{\min}$ is used only to give a sufficient sample size for this event. It is not an input to the estimator. Its appearance reflects an identifiability limit for exact support recovery. Specifically, if a rare transition is never sampled, then a chain with that transition and a chain without it are statistically indistinguishable, although they may induce different recurrent structure and hence a different decomposition. Appendix B.1 gives the formal support-recovery argument.

Theorem 5 (sample-based decomposition guarantee)**.** *Let $\mathcal{E}_{\mathrm{str}}$ be the event that the support graph, recurrent classes, periods, cyclic partitions, and anchors are recovered correctly. On $\mathcal{E}_{\mathrm{str}}$, suppose the basis estimation error is at most ε_b, the oracle noise has a bounded conditional second moment, and the learned quotient contraction factor is smaller than one. Then the estimator in Algorithm 1 satisfies*

$$\mathbb{E}\|[\widehat{v}] - [v^\star_\Pi]\|_{\mathrm{q}} \le O(T^{-1/2}) + O(\varepsilon_b) \tag{18}$$

and, with J samples per anchor for residual estimation,

$$\mathbb{E}\|\widehat{g} - g^\star_\Pi\|_\infty \le O(T^{-1/2}) + O\left(\sqrt{\frac{\log N}{J}}\right) + O(\varepsilon_b). \tag{19}$$

Consequently, the finite horizon prediction error is controlled by Corollary 3.

Together with the support recovery guarantee in Appendix B.1, Theorem 5 gives a high probability structural guarantee followed by estimation bounds on the recovered quotient. The term $T^{-1/2}$ is the stochastic approximation error, ε_b is the price of learning the anchor projection, and $\sqrt{\log N/J}$ appears only when reconstructing the persistent profile from anchor residuals. These errors mirror the decomposition itself: errors in $\widehat{v}$ affect boundary terms, while errors in $\widehat{g}$ accumulate over the horizon.

The parameters in the guarantee have structural meanings. The number of persistent coordinates is $N = |\mathcal{I}| = \sum_i d_i \le |\mathsf{F}| \le n$, and H_{abs} only converts absorption episodes into expected simulator queries. A dense plug in estimator could estimate P and compute the same decomposition from

the learned model, so our claim is not uniform dominance over plug in methods. The point is that the estimator follows the persistent transient structure directly. For tabular MDP evaluation, the procedure is applied after fixing a stationary policy, so the learned recurrent classes and phases are properties of the policy induced chain. The output fits the usual evaluation tasks: $P^{\infty}\widehat{g}$ estimates the statewise average reward, the orbit of $\widehat{g}$ supplies the persistent accounting profile, and $\widehat{v}$ gives the anchor gauge transient diagnostic.

6 Conclusion

This paper studies fixed policy evaluation in finite-state Markov chains beyond the aperiodic setting. Classical normalized gain and bias remain valid, but the bias can mix persistent phase behavior with transient cost. The real peripheral invariant subspace resolves this ambiguity. Quotienting by it is minimal for strict contraction, and an anchor gauge gives the decomposition $r = g^{\star}_{\Pi} + (I - P)v^{\star}_{\Pi}$. The pair contains the same Poisson information as classical gain and bias, but attributes it differently: $g^{\star}_{\Pi}$ records the persistent regime, while $v^{\star}_{\Pi}$ records the cost of reaching the anchored representative of that regime. The contribution is therefore not a new way to compute returns, but a new way to attribute them. The sample-based estimator follows the same structure by learning the recurrent phases, estimating the anchor projection, solving the stable quotient problem, and reconstructing the persistent profile. Its guarantees translate directly into finite horizon return error. The result is a tabular policy evaluation object that supports average reward evaluation, phase resolved diagnostics, and transient cost attribution from the same generative access.

References

[1] Jinane Abounadi, Dimitrib Bertsekas, and Vivek S Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

[2] Shubhada Agrawal, L. A. Prashanth, and Siva Theja Maguluri. Markov chain variance estimation: A stochastic approximation approach. *arXiv preprint arXiv:2409.05733*, 2024.

[3] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[4] Peter L Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42, 2009.

[5] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

[6] Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[7] Zaiwei Chen, Sheng Zhang, Zhe Zhang, Shaan Ul Haque, and Siva Theja Maguluri. A non-asymptotic theory of seminorm lyapunov stability: From deterministic to stochastic iterative algorithms. *arXiv preprint arXiv:2502.14208*, 2025.

[8] Eric Ding, Barry Lyu, Bhaskar Kataria, and Rachee Singh. Opus: Photonic rail-optimized fabric in ml datacenters. In *Proceedings of the ACM SIGCOMM 2026 Conference*, pages 922–939, 2026.

[9] Abhishek Gupta, Rahul Jain, and Peter W Glynn. An empirical algorithm for relative value iteration for average-cost mdps. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5079–5084. IEEE, 2015.

[10] Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

[11] Shaan Ul Haque and Siva Theja Maguluri. Stochastic approximation with unbounded markovian noise: A general-purpose theorem. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3718–3726. PMLR, 2025.

[12] Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for mdps via anchoring. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 32907–32929. PMLR, 2025.

[13] Jongmin Lee and Ernest K Ryu. Optimal non-asymptotic rates of value iteration for average-reward markov decision processes. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] William M Mellette, Alex Forencich, Rukshani Athapathu, Alex C Snoeren, George Papen, and George Porter. Realizing rotornet: Toward practical microsecond scale optical networking. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pages 392–414, 2024.

[15] Carl D Meyer, Jr. The role of the group generalized inverse in the theory of finite markov chains. *Siam Review*, 17(3):443–464, 1975.

[16] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[17] Zachary Roch, Chi Zhang, George K. Atia, and Yue Wang. Provably sample-efficient robust reinforcement learning with average reward. *arXiv preprint arXiv:2505.12462*, 2025.

[18] Paul J Schweitzer. Perturbation theory and finite markov chains. *Journal of Applied Probability*, 5(2):401–413, 1968.

[19] Paul J Schweitzer and Awi Federgruen. The asymptotic behavior of undiscounted value iteration in markov decision problems. *Mathematics of Operations Research*, 2(4):360–381, 1977.

[20] Paul J Schweitzer and Awi Federgruen. Geometric convergence of value-iteration in multichain markov decision problems. *Advances in Applied Probability*, 11(1):188–217, 1979.

[21] PJ Schweitzer and Awi Federgruen. Foolproof convergence in multichain policy iteration. *Journal of Mathematical Analysis and Applications*, 64(2):360–368, 1978.

[22] John N Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

[23] Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR, 2021.

[24] Yang Xu, Washim Uddin Mondal, and Vaneet Aggarwal. Finite-sample analysis of policy evaluation for robust average reward reinforcement learning. *arXiv preprint arXiv:2502.16816*, 2025.

[25] Huizhen Yu and Dimitri P Bertsekas. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.

[26] Shangtong Zhang, Yi Wan, Richard S Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. In *international conference on machine learning*, pages 12578–12588. PMLR, 2021.

[27] Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242, 2021.

[28] Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward mdps. *Advances in Neural Information Processing Systems*, 37:33455–33504, 2024.

[29] Matthew Zurek and Yudong Chen. Faster fixed-point methods for multichain mdps. *arXiv preprint arXiv:2506.20910*, 2025.

A Related work

We review prior work through the lens of the evaluation question posed in the Introduction: what object is used to summarize a fixed policy-induced Markov reward process, and which part of the analysis concerns control, prediction, fixed-point stabilization, or Markov-chain structure? This distinction is important because much of the average-reward literature gives valid gain-bias formalisms or efficient control algorithms, while our paper studies a different object: a phase-aware persistent profile together with an anchor-gauge transient component for an arbitrary finite reducible or periodic policy-induced chain.

A.1 Average-reward control and planning beyond ergodicity.

Classical dynamic programming treats average-reward MDPs through gain and bias equations, including multichain policy and value iteration methods [16, 5, 19–21]. Modern learning work has refined this picture for communicating, weakly communicating, robust, and general average-reward MDPs. Representative examples include regret and sample-complexity guarantees for communicating or weakly communicating models [3, 4, 28, 12], robust average-reward control via Halpern-type iteration [17], and multichain fixed-point and relative-value methods [9, 13, 29]. These works primarily address how to learn, plan, or optimize a policy, usually through an invariant gain and a bias or relative value function. Our setting is fixed-policy evaluation. We do not claim that the classical gain-bias object is invalid; rather, we show that when recurrent classes are periodic it can be diagnostically entangled because non-invariant persistent modes are stored in the bias. The peripheral quotient is introduced to separate that persistent information from transient-to-regime cost.

A.2 Average-reward policy evaluation and temporal-difference methods.

Average-reward prediction is also well studied in ergodic or unichain settings. Classical and modern analyses estimate a scalar reward rate and a differential value function, possibly with linear function approximation or off-policy sampling [22, 25, 1, 23, 27, 26, 11]. Robust average-reward policy evaluation has recently been analyzed under contraction in an appropriate semi-norm [24]. These methods are complementary to ours. They address stochastic approximation, off-policy learning, function approximation, robustness, or unbounded Markovian noise, whereas our main difficulty is structural: for an arbitrary finite policy-induced chain, the non-decaying directions of the original transition dynamics are not merely constants or invariant recurrent-class shifts. Periodic classes introduce additional unit-modulus phase directions, and the evaluation object must decide whether those directions are persistent behavior or transient cost. Our decomposition answers this question before the stochastic approximation step is run.

A.3 Nonexpansive and semi-norm-contractive fixed-point methods.

Several recent average-reward algorithms use anchoring or Halpern iteration to stabilize nonexpansive Bellman operators [10, 12, 17]. More generally, [7] gives a non-asymptotic theory of semi-norm Lyapunov stability and characterizes when a linear iteration contracts modulo a subspace containing all modes with modulus at least one. This perspective is closely related to the contraction part of our analysis. The difference is what is being constructed. Those results give general fixed-point or control tools once the relevant semi-norm or quotient kernel is specified. We identify the Markov-chain-specific kernel for fixed-policy evaluation, prove that it is exactly the real peripheral invariant subspace, construct a learnable anchor gauge from recurrent classes, periods, cyclic phases, and absorption weights, and interpret the resulting coordinates as persistent regime profile and transient-to-regime cost.

A.4 Poisson equations and generalized inverses for Markov chains.

Poisson equations, fundamental matrices, and generalized inverses are standard tools in finite Markov-chain theory [18, 15, 16, 5]. They characterize solutions of $r = \rho + (I - P)h$, sensitivity, and transient corrections once a normalization is chosen. Recent work also uses Poisson equations as an analytic device for other statistical targets, such as Markov-chain asymptotic variance estimation [2]. Our contribution is not the existence of a Poisson equation or a generalized inverse. It is the decomposition of the same Poisson information along the dynamical boundary between persistence

and transience. The exact comparison theorem shows that the classical normalized gain and bias can be recovered from our pair, and conversely that the classical bias consists of our transient component plus a non-invariant peripheral correction. This is why the construction is a refinement of classical Markov-chain Poisson theory rather than a replacement for it.

B Detailed sample-based estimation theory

This appendix gives the detailed guarantees behind the sample-based estimator summarized in Section 5. The estimator first learns the support graph, recurrent classes, cyclic phases, and anchors; then estimates the phase-offset basis from Definition 1; finally it runs projected stochastic approximation for $v^\star_\Pi$ and reconstructs $g^\star_\Pi$ from anchor residuals. Algorithms are listed in Appendix C, and proofs are in Appendices G.1–G.3.

B.1 Learning the anchor gauge

Assume access to a generative model that returns an independent sample from $P(\cdot|s)$ for any queried state s. Let

$$p_{\min} := \min\{P(s'|s) : P(s'|s) > 0\}.$$

The first stage samples each state K times, forms the empirical support graph, computes closed strongly connected components, their periods, cyclic partitions, and one anchor per phase.

Theorem 6 (Exact structural recovery)**.** *If $p_{\min} > 0$ and $K \geq p_{\min}^{-1} \log(n^2/\delta)$, then with probability at least $1 - \delta$ the learned support graph equals the true support graph. Consequently, the recurrent classes, their periods, and their cyclic partitions are correctly recovered up to cyclic relabeling.*

Remark 1 (Rare edges and support recovery)**.** *The parameter $p_{\min}$ is not required by the algorithm. It appears only in the sufficient sample size for exact support recovery. This dependence is unavoidable in the worst case. Consider two chains that differ only at one state s. In the first chain, the edge $s \to v$ is absent. In the second chain, the edge $s \to v$ has probability $p_{\min}$ and the remaining mass is assigned to an otherwise identical transition. If K samples are drawn from s, then under the second chain the edge is unseen with probability $(1 - p_{\min})^K$. On this event, the sample transcript is compatible with the first chain. Thus any procedure that distinguishes the two supports with error probability at most δ must have $(1 - p_{\min})^K \leq \delta$, which gives $K = \Omega(p_{\min}^{-1} \log(1/\delta))$.*

Remark 2 (Good event interpretation)**.** *The end to end guarantee is stated on the event that the support graph and the induced recurrent and cyclic structure are recovered correctly. If the support graph is misspecified, the learned projection may have a different kernel and may target a different quotient problem. The theorem therefore gives a high probability exact structure guarantee followed by a conditional estimation guarantee on that event. An unconditional bound can be obtained by adding the failure probability times a uniform bad event bound whenever such a bound is imposed.*

The second stage estimates the basis $b_{i,k}$ by independent absorption episodes. For each transient state s, let $\widehat{b}(s) = (\widehat{b}_{i,k}(s))_{(i,k)\in\mathcal{I}}$ be the empirical distribution of the terminal class-phase offset in Definition 1, and let $N := |\mathcal{I}|$.

Lemma 4 (Basis estimation)**.** *If*

$$M \geq \frac{8}{\varepsilon_b^2}\left(N + \log\frac{|\mathsf{T}|}{\delta}\right),$$

then with probability at least $1 - \delta$,

$$\max_{s\in\mathsf{T}} \|\widehat{b}(s) - b(s)\|_1 \leq \varepsilon_b.$$

Define the learned anchor projection by

$$(\widehat{\Pi}v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\widehat{b}_{i,k}(s). \tag{20}$$

Lemma 5 (Learned projection deviation)**.** *On the event of Lemma 4, assuming the recurrent-state basis values are exact,*

$$\|(\widehat{\Pi} - \Pi)v\|_\infty \leq \varepsilon_b \|v\|_\infty, \qquad \forall v \in \mathbb{R}^n.$$

B.2 Projected stochastic approximation and reconstruction

Let

$$W := \{v \in \mathbb{R}^n : v(a_{i,k}) = 0 \text{ for all } (i,k) \in \mathcal{I}\}$$

be the anchored subspace. The quotient semi-norm becomes a genuine norm on W.

Lemma 6 (Norm equivalence on the anchored subspace)**.** *We have $W \cap \mathcal{K}(P) = \{0\}$. Hence $\|\cdot\|_{\mathrm{q}}$ is a norm on W, and there exists $C_W < \infty$ such that*

$$\|w\|_\infty \le C_W \|w\|_{\mathrm{q}}, \qquad \forall w \in W.$$

Given $v_t \in W$, the SA update draws one next-state sample $\widetilde{s} \sim P(\cdot|s)$ for each s, forms $\widehat{T}(v_t)(s) = r(s) + v_t(\widetilde{s})$, and sets

$$v_{t+1} = \widehat{\Pi}\big((1-\alpha_t)v_t + \alpha_t \widehat{T}(v_t)\big).$$

The next lemma gives the stability reason for using the peripheral quotient.

Lemma 7 (Perturbed quotient contraction)**.** *On the good structural and basis-estimation events, the learned projected operator is contractive on W with contraction factor*

$$\widehat{\gamma} = \gamma + O(\varepsilon_b).$$

In particular, for sufficiently small ε_b, $\widehat{\gamma} < 1$.

Theorem 7 (Transient-component estimation)**.** *Assume the structural learning and weight estimation steps succeed, the oracle noise has a uniform conditional second moment bound, and $\widehat{\gamma} < 1$. Then the projected SA iterates satisfy*

$$\mathbb{E}\|[v_T] - [v^\star]\|_{\mathrm{q}} \le O(T^{-1/2}) + O(\varepsilon_b),$$

where the constants depend on the contraction and second-moment parameters.

Corollary 4 (End-to-end query complexity)**.** *Let $H_{\mathrm{abs}} := \max_{s\in\mathsf{T}} \mathbb{E}_s[\tau_{\mathsf{F}}]$. To achieve $\mathbb{E}\|[v_T] - [v^\star]\|_{\mathrm{q}} \le \varepsilon$, it suffices to choose*

$$K = \widetilde{O}(1/p_{\min}), \qquad T = \widetilde{O}(\varepsilon^{-2}), \qquad M = \widetilde{O}(N\varepsilon^{-2}),$$

with expected simulator queries at most

$$nK + |\mathsf{T}|MH_{\mathrm{abs}} + nT.$$

Remark 3 (Structural parameters)**.** *The quantity $N = |\mathcal{I}| = \sum_i d_i$ is the number of recurrent class and phase coordinates. Since $d_i \le |F_i|$ for each recurrent class, we have $N \le |\mathsf{F}| \le n$. The quantity $H_{\mathrm{abs}} = \max_{s\in\mathsf{T}} \mathbb{E}_s[\tau_{\mathsf{F}}]$ only converts absorption episodes into expected simulator queries. The norm equivalence constant C_W and the constants in Theorem 8 depend on the conditioning of the quotient and the anchor gauge geometry, but not on the sample budgets K, M, T, J.*

Finally, estimate the anchor residual coordinates

$$\widehat{\theta}_{i,k} = r(a_{i,k}) + \widehat{Pv_T}(a_{i,k}) - v_T(a_{i,k})$$

using J samples per anchor, and reconstruct $\widehat{g}(s) = \sum_{(i,k)} \widehat{\theta}_{i,k}\widehat{b}_{i,k}(s)$.

Theorem 8 (Persistent-profile reconstruction)**.** *Under the conditions of Theorem 7, if the iterates lie in W, then*

$$\mathbb{E}\|\widehat{g} - g^\star_\Pi\|_\infty \le C_1 \mathbb{E}\|v_T - v^\star_\Pi\|_{\mathrm{q}} + C_2\sqrt{\frac{\log N}{J}} + C_3\varepsilon_b.$$

Consequently,

$$\mathbb{E}\|\widehat{g} - g^\star_\Pi\|_\infty \le O(T^{-1/2}) + O\left(\sqrt{\frac{\log N}{J}}\right) + O(\varepsilon_b).$$

Together with Corollary 3, these guarantees convert decomposition estimation into finite-horizon evaluation error. The learned output estimates the objects introduced in Section 3: the persistent regime profile, the anchor-gauge transient component, and the classical average reward obtained from the invariant part of the profile.

Algorithm 2 Learn support graph, recurrent classes, cyclic structure, and anchors

Require: Generative model for P, sample budget K
Ensure: Closed classes $\{F_i\}_{i=1}^m$, recurrent set $\mathcal{F}$, transient set $\mathcal{T}$
Ensure: Periods $\{d_i\}_{i=1}^m$, cyclic sets $\{C_{i,k}\}$, maps cls and phase
Ensure: Anchors $\{a_{i,k}\}$ and index set $\mathcal{I}$

Initialize directed edge set $\widehat{E} \leftarrow \emptyset$
for $s \in \mathcal{S}$ **do**
  for $k = 1, 2, \ldots, K$ **do**
    Draw $\tilde{s} \sim P(s, \cdot)$
    Add edge $(s, \tilde{s})$ to $\widehat{E}$
  end for
end for
Form directed graph $\widehat{G} = (\mathcal{S}, \widehat{E})$
Compute strongly connected components of $\widehat{G}$ and identify closed components
Let $\{F_i\}_{i=1}^m$ be the closed components
Set $\mathcal{F} \leftarrow \cup_{i=1}^m F_i$ and $\mathcal{T} \leftarrow \mathcal{S} \setminus \mathcal{F}$
for each F_i **do**
  Compute period d_i
  Compute cyclic partition $F_i = \cup_{k=0}^{d_i-1} C_{i,k}$ with transitions from $C_{i,k}$ to $C_{i,k+1 \bmod d_i}$
  for $k = 0, 1, \ldots, d_i - 1$ **do**
    Choose one anchor $a_{i,k} \in C_{i,k}$
  end for
end for
Define $\mathrm{cls}(s) = i$ and $\mathrm{phase}(s) = k$ for $s \in C_{i,k} \subseteq \mathcal{F}$
Define $\mathcal{I} = \{(i,k) : i \in \{1, \ldots, m\},\ k \in \{0, \ldots, d_i - 1\}\}$
return $\{F_i\}$, $\mathcal{F}$, $\mathcal{T}$, $\{d_i\}$, $\{C_{i,k}\}$, cls, phase, $\{a_{i,k}\}$, $\mathcal{I}$

C Algorithms for the sample-based estimator

The main text summarizes the estimator; the detailed procedures are listed here.

D Additional proofs for Section 4

Proof of Lemma 3. By the definition of g_v, we have $r = g_v + v - Pv$. Multiplying by P^t gives $P^t r = P^t g_v + P^t v - P^{t+1} v$. Summing over $t = 0, \ldots, T-1$ telescopes: $\sum_{t=0}^{T-1} P^t r = \sum_{t=0}^{T-1} P^t g_v + v - P^T v$. Evaluating at coordinate s gives (11). □

Proof of Proposition 2. From (6), rearrange to $r - g^\star = v^\star - Pv^\star$. Define $M_t := v^\star(X_t) + \sum_{u=0}^{t-1} (r(X_u) - g^\star(X_u))$. A one-step conditional expectation gives

$$\mathbb{E}[M_{t+1} - M_t | X_t] = (Pv^\star)(X_t) - v^\star(X_t) + (r(X_t) - g^\star(X_t)) = 0,$$

so $(M_t)_{t \geq 0}$ is a martingale. By optional stopping at $\tau_{\mathcal{A}}$ (integrability holds since $\mathbb{E}_s[\tau_{\mathcal{A}}] < \infty$ and everything is bounded), $\mathbb{E}_s[M_{\tau_{\mathcal{A}}}] = \mathbb{E}_s[M_0] = v^\star(s)$. Using the anchor normalization gives $v^\star(X_{\tau_{\mathcal{A}}}) = 0$ almost surely, yielding (13). □

Proof of Proposition 3. For a deterministic d-cycle, $P^d = I$, so all eigenvalues are dth roots of unity and the real peripheral subspace is all of $\mathbb{R}^d$. Hence any projection with kernel $\mathcal{K}(P)$ is the zero projection, and Theorem 3 gives $v_\Pi^\star = 0$ and $g_\Pi^\star = r$. The classical gain is the Cesàro average $\rho = \bar{r}\mathbf{1}$. The normalized bias equation is exactly $h - Ph = r - \rho$ with $P^\infty h = 0$, which gives the displayed recurrence and zero-mean condition. For the even-cycle example $d = 2m$, the recurrence decreases by $1/2$ for the first m states and increases by $1/2$ for the remaining m states. The zero-mean solution has maximum $m/4$ and minimum $-m/4$, so $\|h\|_\infty = m/4 = d/8$. □

Algorithm 3 Estimate phase offset absorption weights

Require: Generative model for P
Require: Recurrent set $\mathcal{F}$, transient set $\mathcal{T}$
Require: Periods $\{d_i\}$, cyclic sets $\{C_{i,k}\}$, maps cls and phase, index set $\mathcal{I}$
Require: Episode budget M
Ensure: Estimated weights $\{\widehat{b}_{i,k}(s)\}_{s\in\mathcal{S},(i,k)\in\mathcal{I}}$

Initialize $\widehat{b}_{i,k}(s) \leftarrow 0$ for all $s \in \mathcal{S}$ and $(i,k) \in \mathcal{I}$
for $s \in \mathcal{F}$ **do**
  for $(i,k) \in \mathcal{I}$ **do**
    Set $\widehat{b}_{i,k}(s) \leftarrow \mathbf{1}\{s \in C_{i,k}\}$
  end for
end for
for $s \in \mathcal{T}$ **do**
  Initialize counts $c_{i,k} \leftarrow 0$ for all $(i,k) \in \mathcal{I}$
  for $j = 1, 2, \ldots, M$ **do**
    Set $X \leftarrow s$ and $\tau \leftarrow 0$
    while $X \notin \mathcal{F}$ **do**
      Draw $X' \sim P(X, \cdot)$, set $X \leftarrow X'$, set $\tau \leftarrow \tau + 1$
    end while
    Set $i \leftarrow \mathrm{cls}(X)$ and $\ell \leftarrow \mathrm{phase}(X)$
    Set $k \leftarrow (\ell - \tau) \bmod d_i$
    Increment $c_{i,k} \leftarrow c_{i,k} + 1$
  end for
  for $(i,k) \in \mathcal{I}$ **do**
    Set $\widehat{b}_{i,k}(s) \leftarrow c_{i,k}/M$
  end for
end for
return $\widehat{b}$

Algorithm 4 Build learned gauge map

Require: Anchors $\{a_{i,k}\}$, index set $\mathcal{I}$, estimated weights $\widehat{b}$
Ensure: A callable linear map $\widehat{\Pi} : \mathbb{R}^n \to \mathbb{R}^n$

Define $\widehat{\Pi}$ by the rule below for any $v \in \mathbb{R}^n$ and any $s \in \mathcal{S}$

$$(\widehat{\Pi}v)(s) \leftarrow v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\widehat{b}_{i,k}(s)$$

return $\widehat{\Pi}$

Proof of Corollary 3. Subtract the exact identity (12) from the definition of $\widehat{J}_T$:

$$\widehat{J}_T - J_T = \sum_{t=0}^{T-1} P^t(\widehat{g} - g^\star) + (\widehat{v} - v^\star) - P^T(\widehat{v} - v^\star).$$

Since P is row-stochastic, $\|P^t x\|_\infty \le \|x\|_\infty$ for all $t \ge 0$. Taking sup norms gives the result. □

E Relation to plug in estimation

A natural alternative is to estimate the full transition matrix $\widehat{P}$ and then compute recurrent classes, periods, the peripheral subspace, the gauge projection, and the Poisson solution from the learned model. This approach is valid when $\widehat{P}$ is accurate enough to recover the relevant structure, and on small dense tabular problems it may have smaller final numerical error. Our claim is not that the proposed estimator uniformly dominates plug in estimation. The distinction is that the proposed procedure estimates only the structural and quotient objects needed for the decomposition, avoids storing a dense transition matrix during stochastic approximation, and separates the sources of error

Algorithm 5 Projected quotient stochastic approximation

Require: Generative model for P, reward vector $r \in \mathbb{R}^n$

Require: Learned map $\widehat{\Pi}$, stepsizes $\{\alpha_t\}_{t=0}^{T-1}$, iteration budget T

Ensure: Value iterate $v_T \in \mathbb{R}^n$ in the range of $\widehat{\Pi}$

Initialize $v_0 \in \mathbb{R}^n$ and set $v_0 \leftarrow \widehat{\Pi} v_0$
for $t = 0, 1, \ldots, T-1$ **do**
  Initialize vector $\widehat{T}(v_t) \in \mathbb{R}^n$
  for $s \in \mathcal{S}$ **do**
    Draw $\tilde{s} \sim P(s, \cdot)$
    Set $\widehat{T}(v_t)(s) \leftarrow r(s) + v_t(\tilde{s})$
  end for
  Set $v_{t+\frac{1}{2}} \leftarrow (1 - \alpha_t) v_t + \alpha_t \widehat{T}(v_t)$
  Set $v_{t+1} \leftarrow \widehat{\Pi} v_{t+\frac{1}{2}}$
end for
return v_T

Algorithm 6 Estimate and reconstruct persistent regime profile

Require: Generative model for P, reward vector r, iterate v_T

Require: Anchors $\{a_{i,k}\}$, index set $\mathcal{I}$, weights $\widehat{b}$, sample budget J

Ensure: Coordinate estimates $\widehat{\theta}$ and reconstructed profile $\widehat{g}$

for each $(i,k) \in \mathcal{I}$ **do**
  Draw J independent samples $Y^{(1)}, \ldots, Y^{(J)}$ from $P(a_{i,k}, \cdot)$
  Set $\widehat{(Pv_T)}(a_{i,k}) \leftarrow \frac{1}{J} \sum_{j=1}^{J} v_T(Y^{(j)})$
  Set $\widehat{\theta}_{i,k} \leftarrow r(a_{i,k}) + \widehat{(Pv_T)}(a_{i,k}) - v_T(a_{i,k})$
end for
for $s \in \mathcal{S}$ **do**
  Set $\widehat{g}(s) \leftarrow \sum_{(i,k) \in \mathcal{I}} \widehat{\theta}_{i,k} \, \widehat{b}_{i,k}(s)$
end for
return $\widehat{\theta}$ and $\widehat{g}$

according to the decomposition. This is why the guarantees are stated in terms of support recovery, anchor basis estimation, quotient stochastic approximation, and residual reconstruction.

F Proofs for Section 3

F.1 Proof of Theorem 1

Proof. Write the three statements in the order used in Theorem 1. We first compare statements two and three. By definition, $\mathcal{K}(P)$ is the real span of all generalized eigenmodes whose eigenvalues have modulus one. Hence quotienting by K kills every non decaying mode exactly when $\mathcal{K}(P) \subseteq K$.

Assume next that $\mathcal{K}(P) \subseteq K$. Since P is a finite state stochastic matrix, all eigenvalues lie in the closed unit disk. The finite dimensional spectral decomposition gives a real P invariant subspace L such that $\mathbb{R}^n = \mathcal{K}(P) \oplus L$ and $\rho(P|_L) < 1$. Since K is invariant and contains $\mathcal{K}(P)$, the quotient $\mathbb{R}^n / K$ is naturally a quotient of L. The induced operator on this quotient has spectral radius at most $\rho(P|_L)$, and is therefore strictly smaller than one.

It remains to prove the converse. Suppose $\mathcal{K}(P)$ is not contained in K. Then some nonzero peripheral mode survives in the quotient. After complexification, the quotient operator has an eigenvalue λ with $|\lambda| = 1$ on the image of $\mathcal{K}_{\mathbb{C}}(P)$. Hence $\rho(\bar{P}_K) \geq 1$. This proves that strict stability of the quotient forces $\mathcal{K}(P) \subseteq K$. Combining the two directions proves the equivalence and the minimality of $\mathcal{K}(P)$. □

F.2 Proof of Theorem 2

Proof. We start by showing that $\bar{P}([v]) = [Pv]$ is well-defined on $\mathbb{R}^n/\mathcal{K}(P)$. By construction, $\mathcal{K}_{\mathbb{C}}(P)$ is a P-invariant subspace of $\mathbb{C}^n$ since it is a direct sum of generalized eigenspaces. Hence its realification

$$\mathcal{K}(P) = \{\Re z : z \in \mathcal{K}_{\mathbb{C}}(P)\} + \{\Im z : z \in \mathcal{K}_{\mathbb{C}}(P)\} \subseteq \mathbb{R}^n \tag{21}$$

is also P-invariant. Therefore, if $v - w \in \mathcal{K}(P)$ then $Pv - Pw = P(v - w) \in \mathcal{K}(P)$, which implies $[Pv] = [Pw]$. Thus the induced operator $\bar{P}([v]) = [Pv]$ on $\mathbb{R}^n/\mathcal{K}(P)$ is well-defined and linear.

Since P is row-stochastic, $\|P\|_\infty = 1$, hence every eigenvalue λ of P satisfies $|\lambda| \leq \rho(P) \leq \|P\|_\infty = 1$.

Work over $\mathbb{C}^n$ and let J be the Jordan normal form of P: there exists an invertible matrix S such that

$$S^{-1}PS = J = \operatorname{diag}(J_1, \ldots, J_L),$$

where each J_ℓ is a Jordan block associated with an eigenvalue λ_ℓ. Reorder the blocks so that the first L_0 blocks correspond exactly to eigenvalues with $|\lambda_\ell| = 1$. Let $U \subseteq \mathbb{C}^n$ be the span of the columns of S associated with these first L_0 blocks. By construction, U is precisely the generalized invariant subspace spanned by generalized eigenvectors with $|\lambda| = 1$, so $U = \mathcal{K}_{\mathbb{C}}(P)$.

Consider the induced operator on the quotient $\mathbb{C}^n/U$. In the basis induced by S, the action of P is block diagonal, and quotienting by U removes the first L_0 blocks. Hence the induced map $\overline{P}_{\mathbb{C}}$ on $\mathbb{C}^n/U$ is similar to the block diagonal matrix $\operatorname{diag}(J_{L_0+1}, \ldots, J_L)$, so its spectrum is exactly $\{\lambda_\ell : \ell > L_0\}$ (counting algebraic multiplicities), and therefore we have

$$\rho(\overline{P}_{\mathbb{C}}) = \max_{\ell > L_0} |\lambda_\ell|. \tag{22}$$

All eigenvalues remaining after removing the $|\lambda| = 1$ generalized eigenspace satisfy $|\lambda_\ell| < 1$, hence $\rho(\overline{P}_{\mathbb{C}}) < 1$.

Finally, the real quotient $\mathbb{R}^n/\mathcal{K}(P)$ is the realification of $\mathbb{C}^n/\mathcal{K}_{\mathbb{C}}(P)$, and $\bar{P}$ is the corresponding real linear operator. Therefore $\rho(\bar{P}) = \rho(\overline{P}_{\mathbb{C}}) < 1$, which yields $\rho(\bar{P}) < 1$.

We now construct the contractive quotient norm. Fix any $\gamma \in (\rho(\bar{P}), 1)$ and define the scaled operator

$$\bar{B} := \gamma^{-1}\bar{P} \quad \text{on } \mathbb{R}^n/\mathcal{K}(P).$$

Then $\rho(\bar{B}) = \rho(\bar{P})/\gamma < 1$.

Let $d := \dim(\mathbb{R}^n/\mathcal{K}(P))$ and fix any linear isomorphism $\Phi : \mathbb{R}^n/\mathcal{K}(P) \to \mathbb{R}^d$. Let $B \in \mathbb{R}^{d\times d}$ denote the matrix representation of $\bar{B}$ under Φ: $B := \Phi \circ \bar{B} \circ \Phi^{-1}$. Then $\rho(B) = \rho(\bar{B}) < 1$.

Define the following discrete-time Lyapunov series [6]

$$H := \sum_{t=0}^{\infty} (B^\top)^t B^t \in \mathbb{R}^{d\times d}. \tag{23}$$

Since $\rho(B) < 1$, the series converges in any matrix norm and H is symmetric positive definite. This Lyapunov series construction is standard in discrete-time stability theory. Moreover, H satisfies the discrete Lyapunov equation

$$H - B^\top H B = I_d, \tag{24}$$

which follows by multiplying (23) by $B^\top(\cdot)B$ and telescoping.

Now define a norm on the quotient by

$$\|z\|_{\mathrm{q}} := \sqrt{(\Phi z)^\top H (\Phi z)} \qquad \text{for } z \in \mathbb{R}^n/\mathcal{K}(P). \tag{25}$$

This is a valid norm because $H \succ 0$ and Φ is an isomorphism.

We claim that $\bar{P}$ is γ-contractive under $\|\cdot\|_{\mathrm{q}}$. Indeed, for any $z \in \mathbb{R}^n/\mathcal{K}(P)$, let $x := \Phi z \in \mathbb{R}^d$. Then

$$\begin{aligned}
\|\bar{P}z\|_{\mathrm{q}}^2 &= \|\gamma \bar{B} z\|_{\mathrm{q}}^2 = \gamma^2 \|\bar{B}z\|_{\mathrm{q}}^2 \\
&= \gamma^2 (\Phi(\bar{B}z))^\top H (\Phi(\bar{B}z)) = \gamma^2 (Bx)^\top H (Bx) \\
&= \gamma^2 x^\top B^\top H B x.
\end{aligned}$$

Using (24), $B^\top H B = H - I_d \preceq H$, hence

$$\|\bar{P}z\|_{\mathrm{q}}^2 \le \gamma^2\, x^\top H x = \gamma^2 \|z\|_{\mathrm{q}}^2,$$

which gives the quotient contraction.

We now lift the quotient norm back to $\mathbb{R}^n$. Define $\|v\|_{\mathrm{q}} := \|[v]\|_{\mathrm{q}}$ for $v \in \mathbb{R}^n$. This is a semi-norm because it is the pullback of a norm by the quotient map. Its kernel is

$$\ker(\|\cdot\|_{\mathrm{q}}) = \{v : \|[v]\|_{\mathrm{q}} = 0\} = \{v : [v] = 0\} = \mathcal{K}(P),$$

and for any $v \in \mathbb{R}^n$ we have

$$\|Pv\|_{\mathrm{q}} = \|[Pv]\|_{\mathrm{q}} = \|\bar{P}([v])\|_{\mathrm{q}} \le \gamma \|[v]\|_{\mathrm{q}} = \gamma \|v\|_{\mathrm{q}}.$$

This proves the pullback contraction and completes the proof of Theorem 2. □

F.3 Proof of Theorem 3

We start by the following proposition on the semi-norm,

Proposition 4. *Let Π be a linear projection with $\ker(\Pi) = \mathcal{K}(P)$. Then for all $v \in \mathbb{R}^n$, we have $[\Pi v] = [v]$. In particular, for the quotient semi-norm $\|v\|_{\mathrm{q}} := \|[v]\|_{\mathrm{q}}$, $\|\Pi v\|_{\mathrm{q}} = \|v\|_{\mathrm{q}}$.*

Proof. Fix $v \in \mathbb{R}^n$. Since Π is a projection, we can write

$$v - \Pi v = (I - \Pi)v.$$

Moreover, $(I-\Pi)v \in \ker(\Pi)$ because $\Pi(I-\Pi) = \Pi - \Pi^2 = 0$. By the assumption $\ker(\Pi) = \mathcal{K}(P)$, we obtain $v - \Pi v \in \mathcal{K}(P)$, i.e., v and Πv belong to the same equivalence class in the quotient. Hence $[\Pi v] = [v]$.

For the semi-norm claim, by definition $\|\Pi v\|_{\mathrm{q}} = \|[\Pi v]\|_{\mathrm{q}} = \|[v]\|_{\mathrm{q}} = \|v\|_{\mathrm{q}}$. □

Let $W \coloneqq \mathrm{range}(\Pi)$ and let $Q \coloneqq \mathbb{R}^n/\mathcal{K}(P)$. We first Identify W with the Q. Define the linear map $S : W \to Q$ by $S(w) = [w]$. We claim S is a linear isomorphism.

Injectivity: if $S(w) = 0$, then $w \in \mathcal{K}(P) = \ker(\Pi)$. Since also $w \in W = \mathrm{range}(\Pi)$, we have $w = \Pi u$ for some u and $0 = \Pi w = \Pi^2 u = \Pi u = w$, hence $w = 0$.

Surjectivity: given any class $[v] \in Q$, set $w := \Pi v \in W$. By Proposition 4, $[\Pi v] = [v]$, hence $S(w) = [v]$. Thus S is bijective.

On Q we have the affine map $\bar{T}(z) = [r] + \bar{P}z$. Because $\rho(\bar{P}) < 1$, the linear operator $I - \bar{P}$ is invertible on Q. Hence the equation $z = \bar{T}(z)$ has the unique solution $z^\star = (I - \bar{P})^{-1}[r] \in Q$. Since S is bijective, we then define $v^\star \coloneqq S^{-1}(z^\star) \in W$. We show $v^\star$ satisfies (5) and is the unique such element in W.

First, we verify that $[v^\star] = z^\star$ solves the quotient fixed point equation:

$$[v^\star] = z^\star = [r] + \bar{P}z^\star = [r] + \bar{P}([v^\star]) = [r + Pv^\star].$$

Using Proposition 4 with $x := r + Pv^\star$ gives

$$[\Pi(r + Pv^\star)] = [r + Pv^\star] = [v^\star].$$

Both $\Pi(r + Pv^\star)$ and $v^\star$ lie in W. Since S is injective on W, equality of their quotient classes implies

$$\Pi(r + Pv^\star) = v^\star,$$

which is exactly (5). This proves existence.

For uniqueness, let $\tilde{v} \in W$ satisfy $\tilde{v} = \Pi(r + P\tilde{v})$. Applying Proposition 4 yields

$$[\tilde{v}] = [\Pi(r + P\tilde{v})] = [r + P\tilde{v}] = [r] + \bar{P}([\tilde{v}]),$$

so $[\tilde{v}]$ is a fixed point of $\bar{T}$ on Q. By uniqueness of the fixed point of $\bar{T}$ on Q, we have $[\tilde{v}] = z^\star = [v^\star]$. Since both $\tilde{v}$ and $v^\star$ lie in W and S is injective, we conclude $\tilde{v} = v^\star$.

Let $g^\star := r + Pv^\star - v^\star$. Applying Π and using $\Pi v^\star = v^\star$ gives

$$\Pi g^\star = \Pi(r + Pv^\star) - \Pi v^\star = v^\star - v^\star = 0,$$

so $g^\star \in \ker(\Pi) = \mathcal{K}(P)$.

Finally, (7) holds since $[v^\star] = [r + Pv^\star] = \bar{T}([v^\star])$.

F.4 Proof of Corollary 1

Proof. Fix $s \in \mathcal{S}$. By Lemma 3 applied to $v^\star$ (so that $g_{v^\star} = g^\star$ by (6); see Section 4), for any horizon $T \geq 1$,

$$\mathbb{E}_s\left[\sum_{t=0}^{T-1} r(X_t)\right] = \sum_{t=0}^{T-1}(P^t g^\star)(s) + v^\star(s) - (P^T v^\star)(s).$$

Divide both sides by T and let $T \to \infty$. Since the state space is finite, $v^\star$ is bounded, hence $\big(v^\star(s) - (P^T v^\star)(s)\big)/T \to 0$. By the definition of the long-run average reward in (1), the left-hand side converges to $g(s)$. Therefore,

$$g(s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1}(P^t g^\star)(s),$$

which shows that g^{avg} exists and coincides with $g(\cdot)$ componentwise.

It remains to show $g^{\mathrm{avg}} \in \ker(I - P)$. Let $A_T := \frac{1}{T}\sum_{t=0}^{T-1} P^t$. Then $PA_T = A_T + \frac{P^T - I}{T}$, hence

$$\|PA_T g^\star - A_T g^\star\|_\infty = \left\|\frac{(P^T - I)g^\star}{T}\right\|_\infty \leq \frac{2\|g^\star\|_\infty}{T} \to 0.$$

Passing to the limit yields $Pg^{\mathrm{avg}} = g^{\mathrm{avg}}$, i.e., $g^{\mathrm{avg}} \in \ker(I - P)$. □

F.5 Proof of Proposition 1

Proof. Write the finite state chain in its standard transient and recurrent block form. The recurrent blocks are the closed communicating classes. For each recurrent class, the peripheral spectrum of the corresponding stochastic block consists of the roots of unity determined by its period. If the class is aperiodic, then its period is one, so the only peripheral eigenvalue is 1.

Therefore, if every recurrent class is aperiodic, the only eigenvalue of P on the unit circle is 1. It follows that the real peripheral invariant subspace coincides with the invariant subspace:

$$\mathcal{K}(P) = \ker(I - P).$$

Hence the decomposition in Theorem 3 reduces to the usual gain and bias decomposition up to the choice of gauge. □

F.6 Proof of Theorem 4

Proof. Let $(g^\star_\Pi, v^\star_\Pi)$ be the decomposition from Theorem 3. By definition, we have $r = g^\star_\Pi + (I - P)v^\star_\Pi$. Let (ρ, h) be the normalized classical gain and bias pair, so

$$r = \rho + (I - P)h, \qquad \rho = P^\infty r, \qquad P^\infty h = 0.$$

Apply P^∞ to the decomposition $r = g^\star_\Pi + (I - P)v^\star_\Pi$. Since $P^\infty(I - P) = 0$, we obtain $P^\infty r = P^\infty g^\star_\Pi$. Because $\rho = P^\infty r$, this gives $P^\infty g^\star_\Pi = \rho$.

We already know from Theorem 3 that $g^\star_\Pi \in \mathcal{K}(P)$. Thus,

$$P^\infty(g^\star_\Pi - \rho) = 0.$$

Also, $\rho \in \ker(I - P) \subseteq \mathcal{K}(P)$, so

$$g^\star_\Pi - \rho \in \mathcal{K}(P) \cap \ker(P^\infty) := \mathcal{K}_\circ(P).$$

Since $\mathcal{K}_\circ(P)$ is P invariant and contains no invariant directions, the operator $I - P$ is injective on $\mathcal{K}_\circ(P)$. Because $\mathcal{K}_\circ(P)$ is finite dimensional, injectivity implies bijectivity. Therefore there exists a unique vector $\psi_\Pi \in \mathcal{K}_\circ(P)$ such that

$$(I - P)\psi_\Pi = g^\star_\Pi - \rho,$$

which also means $g^\star_\Pi = \rho + (I - P)\psi_\Pi$. Substitute the identity above into the decomposition:

$$r = g^\star_\Pi + (I - P)v^\star_\Pi = \rho + (I - P)\psi_\Pi + (I - P)v^\star_\Pi = \rho + (I - P)(v^\star_\Pi + \psi_\Pi).$$

Compare this with the classical equation $r = \rho + (I - P)h$ and subtracting the two equations gives

$$(I - P)\big(h - v_\Pi^\star - \psi_\Pi\big) = 0.$$

Hence, we have $h - v_\Pi^\star - \psi_\Pi \in \ker(I - P)$. Apply P^∞ to this identity. Since $P^\infty h = 0$ and $\psi_\Pi \in \mathcal{K}_\circ(P) \subseteq \ker(P^\infty)$, we obtain

$$P^\infty(h - v_\Pi^\star - \psi_\Pi) = -P^\infty v_\Pi^\star.$$

Because the left-hand side equals $h - v_\Pi^\star - \psi_\Pi$, we conclude

$$h - v_\Pi^\star - \psi_\Pi = -P^\infty v_\Pi^\star.$$

Thus we have $h = v_\Pi^\star + \psi_\Pi - P^\infty v_\Pi^\star$. □

F.7 Proof of Corollary 2

Proof. Start from the identity in Theorem 4:

$$h = v_\Pi^\star + \psi_\Pi - P^\infty v_\Pi^\star.$$

Apply Π to both sides. Since $v_\Pi^\star \in \mathrm{range}(\Pi)$, we have $\Pi v_\Pi^\star = v_\Pi^\star$. Since $\psi_\Pi \in \mathcal{K}_\circ(P) \subseteq \mathcal{K}(P) = \ker(\Pi)$, we have $\Pi\psi_\Pi = 0$. Also, $P^\infty v_\Pi^\star \in \ker(I - P) \subseteq \mathcal{K}(P) = \ker(\Pi)$, so $\Pi P^\infty v_\Pi^\star = 0$. Therefore

$$\Pi h = v_\Pi^\star.$$

This proves the first identity.

For the second identity, use the first identity together with

$$g_\Pi^\star = r + P v_\Pi^\star - v_\Pi^\star$$

and the classical equation

$$r = \rho + (I - P)h.$$

Substituting $v_\Pi^\star = \Pi h$ gives

$$g_\Pi^\star = \rho + (I - P)h + P\Pi h - \Pi h = \rho + (I - P)(I - \Pi)h.$$

This proves the second identity. □

G Proofs for Appendix B

G.1 Proofs in Section B.1

We first formally define several notations. Let $E = \{(i, j) : P_{ij} > 0\}$ be the directed support. Then we can redefine the intrinsic edge rarity parameter as

$$p_{\min} := \min_{(i,j)\in E} P_{ij}. \tag{26}$$

If $p_{\min} = 0$, exact support recovery is impossible from finite samples; all guarantees below hold conditional on recovering the support graph, or stated in terms of $p_{\min}$ when $p_{\min} > 0$.

Lemma 8. *Let $G = (S, E)$ be the directed support graph of P (i.e., $(i, j) \in E \iff P_{ij} > 0$). Then the closed strongly connected components of G are exactly the recurrent communicating classes $\{F_i\}_{i=1}^m$. Their union $\mathsf{F} = \cup_i F_i$ is the set of recurrent states and $\mathsf{T} = S \setminus \mathsf{F}$ is the set of transient states.*

Proof. Let $G = (S, E)$ be the directed support graph, so $(i, j) \in E$ if and only if $P_{ij} > 0$.

First, let $C \subseteq S$ be a closed strongly connected component of G. Closed means there is no edge from C to $S \setminus C$, hence $P_{ij} = 0$ for all $i \in C$ and $j \notin C$. Therefore the Markov chain started in C stays in C almost surely. Since C is strongly connected, the restriction of P to C is irreducible on a finite state space, hence every state in C is recurrent. Thus each closed SCC is a recurrent communicating class.

Conversely, let F be a recurrent communicating class. If there were an edge $(i, j) \in E$ with $i \in F$ and $j \notin F$, then $P_{ij} > 0$ and, by definition of communicating classes, there is no directed path

from j back to i in G. Hence with probability at least P_{ij} the chain leaves F and never returns to i, contradicting recurrence of i. Thus no such edge exists and F is closed in G. Since F is a communicating class, it is strongly connected, hence it is a closed SCC.

Therefore the closed SCCs of G are exactly the recurrent communicating classes. Their union is the recurrent set F and the complement is the transient set T. □

Theorem 9 (Formal version of Theorem 6)**.** *Suppose we have a generative model that can sample K i.i.d. next-states from $P(i,\cdot)$ for every $i \in S$. Let $\widehat{E}$ include edge (i,j) iff j appears at least once among the K samples from state i. If $p_{\min} > 0$ and*

$$K \;\geq\; \frac{1}{p_{\min}} \log \frac{n^2}{\delta}, \tag{27}$$

then with probability at least $1-\delta$, $\widehat{E} = E$ and hence $\widehat{G}$ equals the true support graph. Consequently, the closed communicating classes $\{F_i\}$ are correctly recovered, along with their periods d_i and cyclic partitions $\{C_{i,k}\}$ (up to cyclic relabeling within each class).

Proof. If $P_{ij} = 0$, then j is sampled from $P(i,\cdot)$ with probability 0, hence $(i,j) \notin \widehat{E}$ almost surely. Therefore $\widehat{E} \subseteq E$ always, meaning there are no false positives. We now consider a fixed edge $(i,j) \in E$, so $P_{ij} > 0$. The probability that j never appears among the K i.i.d. samples from $P(i,\cdot)$ equals

$$\Pr\big((i,j) \notin \widehat{E}\big) = (1-P_{ij})^K \leq e^{-KP_{ij}} \leq e^{-Kp_{\min}}, \tag{28}$$

with $1 - x \leq e^{-x}$ and $P_{ij} \geq p_{\min}$.

By a union bound over all possible pairs $(i,j) \in S \times S$ (at most n^2 pairs),

$$\Pr(\widehat{E} \neq E) = \Pr(\exists (i,j) \in E :\ (i,j) \notin \widehat{E}) \leq \sum_{(i,j)\in E} \Pr\big((i,j) \notin \widehat{E}\big) \leq n^2 e^{-Kp_{\min}}. \tag{29}$$

If $K \geq \frac{1}{p_{\min}} \log \frac{n^2}{\delta}$, then $n^2 e^{-Kp_{\min}} \leq \delta$, so $\Pr(\widehat{E} = E) \geq 1-\delta$.

On the event $\widehat{E} = E$, the estimated support graph $\widehat{G} = (S, \widehat{E})$ equals the true support graph $G = (S, E)$. By Lemma 8, the closed SCCs of G are exactly the recurrent communicating classes $\{F_i\}_{i=1}^m$. Therefore F and T are recovered exactly.

Fix a recovered recurrent class F_i. Its period d_i depends only on the directed edges inside the induced subgraph $G|_{F_i}$. Equivalently, d_i is the greatest common divisor of lengths of directed cycles in $G|_{F_i}$, so it is determined by the support graph. Since $\widehat{G}|_{F_i} = G|_{F_i}$ on the event $\widehat{E} = E$, the period d_i is recovered correctly.

Given d_i, the standard cyclic decomposition yields a partition $F_i = C_{i,0} \cup \cdots \cup C_{i,d_i-1}$ such that all transitions from $C_{i,k}$ go into $C_{i,k+1}$, where indices are modulo d_i. This cyclic partition is unique up to a cyclic relabeling of the indices. Hence the cyclic partition is also recovered correctly. □

We use the peripheral basis functions $\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ introduced in Definition 1 of the main paper.

Lemma 9. *Fix $(i,k) \in \mathcal{I}$ and define $b_{i,k}$ as in Definition 1. Then for all $s \in S$,*

$$(Pb_{i,k})(s) \;=\; b_{i,k-1 \bmod d_i}(s),$$

where the index addition is taken modulo d_i within class i (and $b_{i,k} \equiv 0$ outside class i on F). In particular, $\mathrm{span}\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ is P-invariant and carries a block-cyclic action.

Proof. Fix a recurrent class index $i \in [m]$ and $k \in \{0,\ldots,d_i-1\}$. Let $b_{i,k}$ be defined as in Definition 1. For any state $s \in S$,

$$(Pb_{i,k})(s) = \sum_{s'\in S} P(s,s')\, b_{i,k}(s') = \mathbb{E}_s[b_{i,k}(X_1)]\,. \tag{30}$$

We prove that $(Pb_{i,k})(s) = b_{i,k-1 \bmod d_i}(s)$ by considering cases.

Case 1: $s \in \mathsf{F}$ **and** $s \in F_i$**.** Let $\ell := \mathsf{phase}(s)$, so $s \in C_{i,\ell}$. By the cyclic decomposition of the periodic class F_i, we have $P(s,\cdot)$ supported on $C_{i,\ell+1 \bmod d_i}$. Hence,

$$(Pb_{i,k})(s) = \sum_{s' \in C_{i,\ell+1}} P(s,s')\mathbf{1}\{s' \in C_{i,k}\} = \mathbf{1}\{\ell+1 \equiv k \pmod{d_i}\} = \mathbf{1}\{\ell \equiv k-1 \pmod{d_i}\}.$$

Since $b_{i,k-1}(s) = \mathbf{1}\{s \in C_{i,k-1}\} = \mathbf{1}\{\ell \equiv k-1 \pmod{d_i}\}$, this shows $(Pb_{i,k})(s) = b_{i,k-1}(s)$.

Case 2: $s \in \mathsf{F}$ **and** $s \notin F_i$**.** Then $s \in C_{j,\ell}$ for some $j \neq i$ and some ℓ. By Definition 1, $b_{i,k}(s) = \mathbf{1}\{s \in C_{i,k}\} = 0$. Also, since F_j is closed, $P(s,\cdot)$ is supported on F_j and therefore on $\mathsf{F} \setminus F_i$. But $b_{i,k}(\cdot)$ equals 0 on $\mathsf{F} \setminus F_i$, so

$$(Pb_{i,k})(s) = \sum_{s' \in S} P(s,s')b_{i,k}(s') = 0 = b_{i,k-1}(s). \tag{31}$$

Case 3: $s \in \mathsf{T}$**.** Let $\tau_{\mathsf{F}} := \min\{t \geq 0 : X_t \in \mathsf{F}\}$ denote the hitting time of F when starting from $X_0 = s$. Since s is transient in a finite Markov chain, $\tau_{\mathsf{F}} \geq 1$ almost surely.

Define the shifted hitting time from X_1 as $\tau_{\mathsf{F}}^+ := \min\{t \geq 0 : X_{t+1} \in \mathsf{F}\}$. On the event $\{\tau_{\mathsf{F}} \geq 1\}$, we have the identities $\tau_{\mathsf{F}}^+ = \tau_{\mathsf{F}} - 1$, and $X_{1+\tau_{\mathsf{F}}^+} = X_{\tau_{\mathsf{F}}}$. We further compute $(Pb_{i,k})(s) = \mathbb{E}_s[b_{i,k}(X_1)]$.

By Definition 1, for the possibly transient state X_1,

$$b_{i,k}(X_1) = \Pr\Big(\mathsf{cls}(X_{1+\tau_{\mathsf{F}}^+}) = i,\ (\mathsf{phase}(X_{1+\tau_{\mathsf{F}}^+}) - \tau_{\mathsf{F}}^+) \bmod d_i = k | X_1\Big). \tag{32}$$

Using the Markov property and the identities above,

$$(\mathsf{phase}(X_{1+\tau_{\mathsf{F}}^+}) - \tau_{\mathsf{F}}^+) \bmod d_i = (\mathsf{phase}(X_{\tau_{\mathsf{F}}}) - (\tau_{\mathsf{F}}-1)) \bmod d_i = (\mathsf{phase}(X_{\tau_{\mathsf{F}}}) - \tau_{\mathsf{F}} + 1) \bmod d_i. \tag{33}$$

Therefore the condition $(\mathsf{phase}(X_{1+\tau_{\mathsf{F}}^+}) - \tau_{\mathsf{F}}^+) \bmod d_i = k$ is equivalent to $(\mathsf{phase}(X_{\tau_{\mathsf{F}}}) - \tau_{\mathsf{F}}) \bmod d_i = k-1$. Hence,

$$b_{i,k}(X_1) = \Pr(\mathsf{cls}(X_{\tau_{\mathsf{F}}}) = i,\ (\mathsf{phase}(X_{\tau_{\mathsf{F}}}) - \tau_{\mathsf{F}}) \bmod d_i = k-1 | X_1). \tag{34}$$

Taking expectation over X_1 yields

$$(Pb_{i,k})(s) = \Pr_s(\mathsf{cls}(X_{\tau_{\mathsf{F}}}) = i,\ (\mathsf{phase}(X_{\tau_{\mathsf{F}}}) - \tau_{\mathsf{F}}) \bmod d_i = k-1) = b_{i,k-1}(s), \tag{35}$$

where the last equality is exactly Definition 1 for transient s.

Based on the above, all cases give $(Pb_{i,k})(s) = b_{i,k-1 \bmod d_i}(s)$ for every $s \in S$. Therefore, $\mathrm{span}\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ is P-invariant and the action of P on this span is block-cyclic within each recurrent class i. □

Lemma 10. *Let $\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ be as in Definition 1, and fix anchors $\{a_{i,k}\}$ with $a_{i,k} \in C_{i,k}$. Define $\Pi : \mathbb{R}^n \to \mathbb{R}^n$ by*

$$(\Pi v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\, b_{i,k}(s). \tag{36}$$

Then:

1. *Π is a linear projection ($\Pi^2 = \Pi$), and $(\Pi v)(a_{i,k}) = 0$ for all (i,k).*
2. *$\ker(\Pi) = \mathrm{span}\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ and $\dim(\ker(\Pi)) = |\mathcal{I}| = \sum_{i=1}^m d_i$.*

Proof of Lemma 10. Since $a_{i,k} \in C_{i,k} \subset \mathsf{F}$ and $b_{j,\ell}(s) = \mathbf{1}\{s \in C_{j,\ell}\}$ on F, we have for all $(i,k),(j,\ell) \in \mathcal{I}$:

$$b_{j,\ell}(a_{i,k}) = \mathbf{1}\{a_{i,k} \in C_{j,\ell}\} = \mathbf{1}\{(j,\ell) = (i,k)\}. \tag{37}$$

We first note that the linearity of Π is immediate from its definition. For any $(i,k) \in \mathcal{I}$,

$$(\Pi v)(a_{i,k}) = v(a_{i,k}) - \sum_{(j,\ell)\in\mathcal{I}} v(a_{j,\ell})\, b_{j,\ell}(a_{i,k}) = v(a_{i,k}) - \sum_{(j,\ell)\in\mathcal{I}} v(a_{j,\ell})\mathbf{1}\{(j,\ell) = (i,k)\} = 0, \tag{38}$$

using (37). Hence Πv satisfies $(\Pi v)(a_{i,k}) = 0$ for all anchors.

To show $\Pi^2 = \Pi$, fix any v and write

$$(\Pi^2 v)(s) = (\Pi(\Pi v))(s) = (\Pi v)(s) - \sum_{(i,k)\in\mathcal{I}} (\Pi v)(a_{i,k})\, b_{i,k}(s) = (\Pi v)(s), \tag{39}$$

because we already proved $(\Pi v)(a_{i,k}) = 0$ for every (i,k). Thus Π is a projection.

For any coefficients $\{\theta_{i,k}\}$, let $v = \sum_{(i,k)\in\mathcal{I}} \theta_{i,k} b_{i,k}$. Then by (37), $v(a_{i,k}) = \theta_{i,k}$, and therefore

$$(\Pi v)(s) = \sum_{(i,k)} \theta_{i,k} b_{i,k}(s) - \sum_{(i,k)} v(a_{i,k}) b_{i,k}(s) = \sum_{(i,k)} \theta_{i,k} b_{i,k}(s) - \sum_{(i,k)} \theta_{i,k} b_{i,k}(s) = 0. \tag{40}$$

So $\operatorname{span}\{b_{i,k}\} \subseteq \ker(\Pi)$.

Conversely, if $\Pi v = 0$, then for every $s \in S$,

$$v(s) = \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\, b_{i,k}(s), \tag{41}$$

which shows $v \in \operatorname{span}\{b_{i,k}\}$. Hence $\ker(\Pi) = \operatorname{span}\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$.

Finally, $\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ is linearly independent: if $\sum_{(i,k)} \theta_{i,k} b_{i,k} = 0$, evaluate at $s = a_{j,\ell}$ and use (37) to get $\theta_{j,\ell} = 0$ for all (j,ℓ). Thus $\dim(\ker(\Pi)) = |\mathcal{I}| = \sum_{i=1}^{m} d_i$. □

Lemma 11 (Formal version of Lemma 2)**.** *The P-invariant subspace $\operatorname{span}\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ coincides with the real peripheral invariant subspace $\mathcal{K}(P)$. Consequently, $\ker(\Pi) = \mathcal{K}(P)$.*

Proof. Let

$$\mathcal{B} \;\coloneqq\; \operatorname{span}\{b_{i,k}\}_{(i,k)\in\mathcal{I}} \subseteq \mathbb{R}^n.$$

We first show $\mathcal{B} \subseteq \mathcal{K}(P)$. By Lemma 9,

$$P b_{i,k} \;=\; b_{i,k-1 \bmod d_i} \qquad \text{for each } (i,k) \in \mathcal{I}, \tag{42}$$

and $b_{i,k} \equiv 0$ on $\mathsf{F} \setminus F_i$.

Fix a class i. Let $\omega_i \coloneqq e^{2\pi \mathrm{i}/d_i}$ and define, for $\ell = 0, 1, \ldots, d_i - 1$, the complex "Fourier mode"

$$\psi_{i,\ell} \;\coloneqq\; \sum_{k=0}^{d_i-1} \omega_i^{\ell k}\, b_{i,k} \;\in\; \mathbb{C}^n.$$

Using (42) and a change of index,

$$P\psi_{i,\ell} = \sum_{k=0}^{d_i-1} \omega_i^{\ell k}\, P b_{i,k} = \sum_{k=0}^{d_i-1} \omega_i^{\ell k}\, b_{i,k-1} = \sum_{j=0}^{d_i-1} \omega_i^{\ell(j+1)}\, b_{i,j} = \omega_i^{\ell}\, \psi_{i,\ell}. \tag{43}$$

Thus each $\psi_{i,\ell}$ is an eigenvector of P with eigenvalue ω_i^{ℓ}, and $|\omega_i^{\ell}| = 1$. Therefore $\psi_{i,\ell} \in \mathcal{K}_{\mathbb{C}}(P)$ for all i, ℓ. Since $b_{i,k}$ is a real linear combination of $\{\psi_{i,\ell}\}_{\ell=0}^{d_i-1}$ (invert the discrete Fourier transform), we conclude that each $b_{i,k}$ lies in the realification $\mathcal{K}(P)$, hence $\mathcal{B} \subseteq \mathcal{K}(P)$.

We now show that $\dim(\mathcal{K}(P)) = |\mathcal{I}|$. Permute the states so that transients come first and recurrent classes are grouped:

$$P \;=\; \begin{pmatrix} Q & R \\ 0 & J \end{pmatrix}, \qquad J = \operatorname{diag}(P_1, \ldots, P_m),$$

where Q is the transient-to-transient submatrix and each P_i is the transition matrix restricted to recurrent class F_i.

Because Q corresponds to transient states in a finite chain, we have $\rho(Q) < 1$; in particular, Q has no eigenvalues on the unit circle. Since P is block upper-triangular, the multiset of eigenvalues of P equals the union of the eigenvalues of Q and those of the blocks P_i. Hence all eigenvalues of P with $|\lambda| = 1$ come from the recurrent blocks $\{P_i\}$.

Now fix i. The block P_i is row-stochastic, irreducible, and has period d_i. A standard Perron–Frobenius periodicity result implies: the eigenvalues of P_i on the unit circle are exactly $\{e^{2\pi \mathrm{i}\ell/d_i} : \ell = 0, \dots, d_i - 1\}$, each with algebraic (and geometric) multiplicity 1. Consequently, we have $\dim\big(\mathcal{K}_{\mathbb{C}}(P)\big) = \sum_{i=1}^{m} d_i = |\mathcal{I}|$.

Finally, since the peripheral eigenvalues occur in complex-conjugate pairs and the peripheral spectrum is semisimple for stochastic P (no Jordan blocks on $|\lambda| = 1$), the realification does not change the total dimension of the corresponding real invariant subspace; hence

$$\dim\big(\mathcal{K}(P)\big) \;=\; |\mathcal{I}|.$$

Since we have $\mathcal{B} \subseteq \mathcal{K}(P)$, and by Lemma 10, $\dim(\mathcal{B}) = |\mathcal{I}|$. Step 2 gives $\dim(\mathcal{K}(P)) = |\mathcal{I}|$. Therefore $\mathcal{B} = \mathcal{K}(P)$, and together with Lemma 10,

$$\ker(\Pi) = \mathcal{B} = \mathcal{K}(P). \qquad \square$$

Remark 4 (Semisimplicity of the peripheral spectrum)**.** *Since P is row-stochastic, $\|P^t\|_\infty \le 1$ for all $t \ge 0$. Hence all eigenvalues with $|\lambda| = 1$ are semisimple (no Jordan blocks), and $\mathcal{K}_{\mathbb{C}}(P)$ equals the direct sum of eigenspaces for $|\lambda| = 1$.*

Lemma 12 (Formal version of Lemma 4)**.** *Assume Theorem 9 holds (so $\{F_i\}, \{C_{i,k}\}$ are known). For each transient state $s \in \mathsf{T}$, estimate $\widehat{b}(s) = (\widehat{b}_{i,k}(s))_{(i,k)\in\mathcal{I}}$ by M i.i.d. absorption episodes as in Algorithm 3. Let $N := |\mathcal{I}| = \sum_{i=1}^{m} d_i$. Then for any $\varepsilon_{\mathrm{b}} \in (0,1)$, if*

$$M \;\ge\; \frac{8}{\varepsilon_{\mathrm{b}}^2}\Big(N + \log\frac{|\mathsf{T}|}{\delta}\Big),$$

we have with probability at least $1 - \delta$:

$$\max_{s\in\mathsf{T}} \|\widehat{b}(s) - b(s)\|_1 \le \varepsilon_{\mathrm{b}}. \tag{44}$$

Proof of Lemma 12. Fix a transient state $s \in \mathsf{T}$. One absorption episode from s produces a trajectory $(X_t)_{t\ge 0}$ until $\tau_{\mathsf{F}} = \min\{t \ge 0 : X_t \in \mathsf{F}\}$ and yields a terminal index $Y \in \mathcal{I}$ (the recovered recurrent class and phase). By definition, Y is a categorical random variable with distribution $b(s) = (b_{i,k}(s))_{(i,k)\in\mathcal{I}}$, and $\widehat{b}(s)$ is the empirical distribution of M i.i.d. copies of Y.

Define $f_s := \|\widehat{b}(s) - b(s)\|_1$ as a function of the M samples. Changing one sample changes $\widehat{b}(s)$ by at most $2/M$ in ℓ_1, hence f_s has bounded differences with constants $2/M$. By McDiarmid's inequality, for any $t > 0$,

$$\Pr(f_s - \mathbb{E}f_s \ge t) \le \exp\!\left(-\frac{Mt^2}{2}\right).$$

Also, by Cauchy–Schwarz and Jensen,

$$\mathbb{E}f_s \le \sqrt{N}\,\mathbb{E}\|\widehat{b}(s) - b(s)\|_2 \le \sqrt{N}\,\sqrt{\mathbb{E}\|\widehat{b}(s) - b(s)\|_2^2}.$$

A direct computation for the multinomial estimator gives

$$\mathbb{E}\|\widehat{b}(s) - b(s)\|_2^2 = \sum_{(i,k)\in\mathcal{I}} \mathrm{Var}\big(\widehat{b}_{i,k}(s)\big) = \frac{1 - \|b(s)\|_2^2}{M} \le \frac{1}{M},$$

so $\mathbb{E}f_s \le \sqrt{N/M}$.

Now set $t = \sqrt{\frac{2\log(|\mathsf{T}|/\delta)}{M}}$ and take a union bound over all $s \in \mathsf{T}$ to obtain

$$\max_{s\in\mathsf{T}} f_s \le \sqrt{\frac{N}{M}} + \sqrt{\frac{2\log(|\mathsf{T}|/\delta)}{M}}$$

with probability at least $1 - \delta$. The stated sufficient condition on M makes the right-hand side at most ε_{b}. $\square$

Corollary 5 (Choosing M to ensure $\widehat{\gamma} < 1$)**.** *Fix $\gamma \in (0,1)$ and constants $C_{\mathrm{q},\infty}$ and $C_{\infty,\mathrm{q}}$. Set*

$$\varepsilon_b^{\mathrm{tar}} := \frac{1-\gamma}{2\, C_{\mathrm{q},\infty}\, C_{\infty,\mathrm{q}}}.$$

If Lemma 12 holds with $\varepsilon_b \le \varepsilon_b^{\mathrm{tar}}$, then we have $\widehat{\gamma} \le \frac{1+\gamma}{2} < 1$. In particular, by Lemma 12, it suffices to choose

$$M \;\ge\; \frac{8}{(\varepsilon_b^{\mathrm{tar}})^2}\Big(N + \log\frac{|\mathsf{T}|}{\delta}\Big)$$

to guarantee $\widehat{\gamma} < 1$ on the event of Lemma 12.

G.2 Proofs in Section B.2

Lemma 13. *Work on the event of Theorem 9 and Lemma 12. Under Algorithm 5 with the recursion $v_{t+1} = \widehat{\Pi}\Big((1-\alpha_t)v_t + \alpha_t\, \widehat{T}(v_t)\Big)$ and $\alpha_t = \alpha/(t+1)$. Assume the oracle noise along the iterates satisfies the uniform conditional second-moment bound*

$$\mathbb{E}\Big[\|\widehat{T}(v_t) - (r + Pv_t)\|_\infty^2 | \mathcal{F}_t\Big] \le \sigma^2 \qquad \text{for all } t, \tag{45}$$

where $\mathcal{F}_t = \sigma(v_0, \dots, v_t)$. Then there exists a finite constant C_{stab} (depending on $\alpha, \gamma, N, \|r\|_\infty$, and norm-equivalence constants) such that

$$\sup_{t \ge 0} \mathbb{E}\|v_t - v^\star\|_{\mathrm{q}}^2 \le C_{\mathrm{stab}}. \tag{46}$$

Consequently, by norm equivalence on $\mathrm{range}(\Pi)$, there is a finite constant $C_{\infty,q}$ such that

$$\sup_{t \ge 0} \mathbb{E}\|v_t\|_\infty^2 \le 2C_{\infty,q}^2\, C_{\mathrm{stab}} + 2\|v^\star\|_\infty^2 =: B_{\sup}^2 < \infty. \tag{47}$$

Proof of Lemma 13. Work on the event

$$\mathsf{E} \;:=\; \Big(\text{Theorem 9 holds}\Big) \cap \Big(\text{Lemma 12 holds}\Big), \tag{48}$$

so that in particular the anchor exactness and transient accuracy hold:

$$\widehat{b}_{j,\ell}(a_{i,k}) = b_{j,\ell}(a_{i,k}) = \mathbf{1}\{(j,\ell) = (i,k)\}, \qquad \max_{s \in \mathsf{T}} \|\widehat{b}(s) - b(s)\|_1 \le \varepsilon_{\mathrm{b}}. \tag{49}$$

Recall the projections

$$(\Pi v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\, b_{i,k}(s), \qquad (\widehat{\Pi} v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\, \widehat{b}_{i,k}(s), \tag{50}$$

and define the anchored subspace

$$W \;:=\; \Big\{v \in \mathbb{R}^n :\; v(a_{i,k}) = 0 \;\; \forall (i,k) \in \mathcal{I}\Big\}. \tag{51}$$

Using anchor exactness, for any v and any anchor $a_{i,k}$,

$$(\widehat{\Pi} v)(a_{i,k}) = v(a_{i,k}) - \sum_{(j,\ell)\in\mathcal{I}} v(a_{j,\ell})\, \widehat{b}_{j,\ell}(a_{i,k}) = v(a_{i,k}) - v(a_{i,k}) = 0. \tag{52}$$

Hence $\mathrm{range}(\widehat{\Pi}) \subseteq W$. Conversely, if $w \in W$, then all anchor values vanish and thus $\widehat{\Pi} w = w$, implying $W \subseteq \mathrm{range}(\widehat{\Pi})$. Therefore,

$$\mathrm{range}(\widehat{\Pi}) = W \qquad \text{and} \qquad \widehat{\Pi}|_W = I_W. \tag{53}$$

The same holds for Π. In particular, for all $t \ge 1$,

$$v_t \in W, \qquad v^\star \in W, \qquad e_t := v_t - v^\star \in W. \tag{54}$$

We now formulate the error recursion. Let

$$T(v) := r + Pv, \qquad \xi_{t+1} := \widehat{T}(v_t) - T(v_t). \tag{55}$$

By the oracle condition, $\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = 0$ and $\mathbb{E}[\|\xi_{t+1}\|_\infty^2|\mathcal{F}_t] \le \sigma^2$. Define $\eta_{t+1} := \widehat{\Pi}\xi_{t+1} \in W$. Since $\widehat{\Pi}$ is $\mathcal{F}_t$-measurable and linear,

$$\mathbb{E}[\eta_{t+1}|\mathcal{F}_t] = 0. \tag{56}$$

The update is

$$v_{t+1} = \widehat{\Pi}\big((1-\alpha_t)v_t + \alpha_t\widehat{T}(v_t)\big) = (1-\alpha_t)v_t + \alpha_t\widehat{\Pi}T(v_t) + \alpha_t\eta_{t+1}. \tag{57}$$

Subtract $v^\star = \Pi T(v^\star)$ and add/subtract $\Pi T(v_t)$ to obtain

$$e_{t+1} = (1-\alpha_t)e_t + \alpha_t\Pi(Pe_t) + \alpha_t\delta_t + \alpha_t\eta_{t+1}, \tag{58}$$

where the projection-mismatch term is

$$\delta_t := (\widehat{\Pi} - \Pi)T(v_t) = (\widehat{\Pi} - \Pi)(r + Pv_t) \in W. \tag{59}$$

Let $D := 1 - \gamma \in (0,1)$. By Theorem 2 and the definition of $\|\cdot\|_{\mathrm{q}}$,

$$\|\Pi(Pe_t)\|_{\mathrm{q}} \le \gamma\|e_t\|_{\mathrm{q}}, \tag{60}$$

and hence by the triangle inequality,

$$\big\|(1-\alpha_t)e_t + \alpha_t\Pi(Pe_t)\big\|_{\mathrm{q}} \le (1-\alpha_t)\|e_t\|_{\mathrm{q}} + \alpha_t\gamma\|e_t\|_{\mathrm{q}} = \big(1-\alpha_t D\big)\|e_t\|_{\mathrm{q}}. \tag{61}$$

Let

$$u_t := (1-\alpha_t)e_t + \alpha_t\Pi(Pe_t) \in W. \tag{62}$$

Then (58) becomes $e_{t+1} = u_t + \alpha_t\delta_t + \alpha_t\eta_{t+1}$. Since $\|\cdot\|_{\mathrm{q}}$ is induced by an inner product (by construction in Theorem 2), the martingale cross term cancels:

$$\mathbb{E}\big[\langle u_t + \alpha_t\delta_t, \eta_{t+1}\rangle_{\mathrm{q}}|\mathcal{F}_t\big] = \Big\langle u_t + \alpha_t\delta_t, \mathbb{E}[\eta_{t+1}|\mathcal{F}_t]\Big\rangle_{\mathrm{q}} = 0. \tag{63}$$

Therefore,

$$\mathbb{E}[\|e_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t] = \|u_t + \alpha_t\delta_t\|_{\mathrm{q}}^2 + \alpha_t^2\mathbb{E}[\|\eta_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t]. \tag{64}$$

Use Young's inequality $2\langle u, \delta\rangle \le D\|u\|^2 + D^{-1}\|\delta\|^2$ to bound

$$\|u_t + \alpha_t\delta_t\|_{\mathrm{q}}^2 \le \|u_t\|_{\mathrm{q}}^2 + \alpha_t D\|u_t\|_{\mathrm{q}}^2 + \frac{\alpha_t}{D}\|\delta_t\|_{\mathrm{q}}^2 + \alpha_t^2\|\delta_t\|_{\mathrm{q}}^2. \tag{65}$$

Together with (61), this yields

$$\mathbb{E}[\|e_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t] \le (1-\alpha_t D)\|e_t\|_{\mathrm{q}}^2 + \frac{\alpha_t}{D}\|\delta_t\|_{\mathrm{q}}^2 + \alpha_t^2\|\delta_t\|_{\mathrm{q}}^2 + \alpha_t^2\mathbb{E}[\|\eta_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t]. \tag{66}$$

To bound the termsδ_t and η_{t+1}, let $N := |\mathcal{I}|$. By the definition of δ_t and the accuracy event, for all $s \in S$,

$$|(\delta_t)(s)| \le \|T(v_t)\|_\infty\|b(s) - \widehat{b}(s)\|_1 \le \varepsilon_{\mathrm{b}}\,\|T(v_t)\|_\infty. \tag{67}$$

Hence, for a norm-equivalence constant $C_{q,\infty} > 0$ on W,

$$\|\delta_t\|_{\mathrm{q}} \le C_{q,\infty}\|\delta_t\|_\infty \le C_{q,\infty}\varepsilon_{\mathrm{b}}\,\|r + Pv_t\|_\infty \le C_{q,\infty}\varepsilon_{\mathrm{b}}\,(R + \|v_t\|_\infty). \tag{68}$$

Similarly, $\|\widehat{\Pi}x\|_\infty \le 2\|x\|_\infty$, so for $\eta_{t+1} = \widehat{\Pi}\xi_{t+1}$,

$$\mathbb{E}[\|\eta_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t] \le C_{q,\infty}^2\mathbb{E}[\|\eta_{t+1}\|_\infty^2|\mathcal{F}_t] \le 4C_{q,\infty}^2\mathbb{E}[\|\xi_{t+1}\|_\infty^2|\mathcal{F}_t] \le 4C_{q,\infty}^2\sigma^2. \tag{69}$$

We now provide uniform boundedness of $\mathbb{E}\|e_t\|_{\mathrm{q}}^2$. Taking expectations in (66) and using (68)-(69), we get a scalar recursion of the form

$$m_{t+1} \le (1-\alpha_t D)m_t + \frac{\alpha_t}{D}\mathbb{E}\|\delta_t\|_{\mathrm{q}}^2 + \alpha_t^2\Big(\mathbb{E}\|\delta_t\|_{\mathrm{q}}^2 + \sigma_\eta^2\Big), \tag{70}$$

where $m_t := \mathbb{E}\|e_t\|_{\mathrm{q}}^2$ and $\sigma_\eta := 2C_{q,\infty}\sigma$. Using (68) and $\|Pv_t\|_\infty \le \|v_t\|_\infty$,

$$\mathbb{E}\|\delta_t\|_{\mathrm{q}}^2 \le 2(C_{q,\infty}\varepsilon_{\mathrm{b}})^2\Big(R^2 + \mathbb{E}\|v_t\|_\infty^2\Big). \tag{71}$$

On W, norms are equivalent, so there is $C_{\infty,q} > 0$ such that

$$\|v_t\|_\infty \le \|v^\star\|_\infty + \|e_t\|_\infty \le \|v^\star\|_\infty + C_{\infty,q}\|e_t\|_{\mathrm{q}}. \tag{72}$$

Thus $\mathbb{E}\|v_t\|_\infty^2 \le C_0 + C_1 m_t$ for finite constants C_0, C_1, and hence

$$\mathbb{E}\|\delta_t\|_{\mathrm{q}}^2 \le C_\delta^2\varepsilon_{\mathrm{b}}^2\big(1 + m_t\big) \tag{73}$$

for a finite constant C_δ. Plugging into (70) yields, for all large enough t,

$$m_{t+1} \le \Big(1 - \alpha_t D + \tfrac{\alpha_t}{D}C_\delta^2\varepsilon_{\mathrm{b}}^2\Big)m_t + \tfrac{\alpha_t}{D}C_\delta^2\varepsilon_{\mathrm{b}}^2 + \alpha_t^2 \cdot C_{\mathrm{noise}}, \tag{74}$$

with $C_{\mathrm{noise}} < \infty$. If ε_{b} is chosen so that $D - \frac{1}{D}C_\delta^2\varepsilon_{\mathrm{b}}^2 \ge D/2$, then the above is a standard stable Robbins–Monro recursion with stepsize $\alpha_t = \alpha/(t+1)$, which implies

$$\sup_{t\ge 0} m_t \le C_{\mathrm{stab}} < \infty. \tag{75}$$

Finally, using norm equivalence on W,

$$\sup_{t\ge 0}\mathbb{E}\|v_t\|_\infty^2 \le 2C_{\infty,q}^2 \sup_{t\ge 0}\mathbb{E}\|e_t\|_{\mathrm{q}}^2 + 2\|v^\star\|_\infty^2 < \infty. \tag{76}$$

This completes the proof. □

Lemma 14 (Formal version of Lemma 5)**.** *Assume $\widehat{\Pi}$ is built from anchors $\{a_{i,k}\}$ and weights $\{\widehat{b}_{i,k}\}$ via*

$$(\widehat{\Pi}v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\,\widehat{b}_{i,k}(s), \qquad (\Pi v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\,b_{i,k}(s), \tag{77}$$

and suppose

$$\max_{s\in\mathcal{T}}\|\widehat{b}(s) - b(s)\|_1 \le \varepsilon_b, \qquad \widehat{b}_{i,k}(s) = b_{i,k}(s) \text{ for all } s \in \mathcal{F}, \tag{78}$$

where $\widehat{b}(s) = (\widehat{b}_{i,k}(s))_{(i,k)\in\mathcal{I}}$ and $b(s) = (b_{i,k}(s))_{(i,k)\in\mathcal{I}}$.

Then for every $v \in \mathbb{R}^n$,

$$\|(\widehat{\Pi} - \Pi)v\|_\infty \le \varepsilon_b\,\|v\|_\infty. \tag{79}$$

Moreover, for any semi-norm $\|\cdot\|_{\mathrm{q}}$ satisfying a norm-equivalence

$$\|x\|_{\mathrm{q}} \le C_{\mathrm{q},\infty}\|x\|_\infty, \tag{80}$$

we also have

$$\|(\widehat{\Pi} - \Pi)v\|_{\mathrm{q}} \le C_{\mathrm{q},\infty}\,\varepsilon_b\,\|v\|_\infty. \tag{81}$$

Proof. Fix $s \in S$. From (77),

$$\big((\widehat{\Pi} - \Pi)v\big)(s) = -\sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\big(\widehat{b}_{i,k}(s) - b_{i,k}(s)\big). \tag{82}$$

On $s \in \mathcal{F}$ the difference is zero by assumption. On $s \in \mathcal{T}$, taking absolute values and using $|v(a_{i,k})| \le \|v\|_\infty$ gives

$$\big|\big((\widehat{\Pi} - \Pi)v\big)(s)\big| \le \|v\|_\infty \sum_{(i,k)\in\mathcal{I}}\big|\widehat{b}_{i,k}(s) - b_{i,k}(s)\big| = \|v\|_\infty\,\|\widehat{b}(s) - b(s)\|_1 \le \varepsilon_b\,\|v\|_\infty. \tag{83}$$

Taking $\max_s$ yields (79). Then (81) follows from (80) applied to $x = (\widehat{\Pi} - \Pi)v$. □

Lemma 15 (Well-posedness of the $\widehat{\Pi}$-projected fixed point)**.** *Work on the "good events". Recall the projections*

$$(\Pi v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\, b_{i,k}(s), \qquad (\widehat{\Pi} v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\, \widehat{b}_{i,k}(s), \tag{84}$$

and define the anchored subspace

$$W \;:=\; \Big\{ v \in \mathbb{R}^n :\; v(a_{i,k}) = 0 \;\; \forall (i,k) \in \mathcal{I} \Big\}. \tag{85}$$

Using anchor exactness, for any v and any anchor $a_{i,k}$,

$$(\widehat{\Pi} v)(a_{i,k}) = v(a_{i,k}) - \sum_{(j,\ell)\in\mathcal{I}} v(a_{j,\ell})\, \widehat{b}_{j,\ell}(a_{i,k}) = v(a_{i,k}) - v(a_{i,k}) = 0. \tag{86}$$

Hence $\mathrm{range}(\widehat{\Pi}) \subseteq W$*. Conversely, if $w \in W$, then all anchor values vanish and thus $\widehat{\Pi} w = w$, implying $W \subseteq \mathrm{range}(\widehat{\Pi})$. Therefore,*

$$\mathrm{range}(\widehat{\Pi}) = W \qquad \textit{and} \qquad \widehat{\Pi}|_W = I_W. \tag{87}$$

Define the affine map

$$\widehat{F}(v) \;:=\; \widehat{\Pi}(r + Pv). \tag{88}$$

Let $\|\cdot\|_{\mathrm{q}}$ be the quotient norm used in Theorem 2, and assume norm equivalence on W:

$$\|x\|_{\mathrm{q}} \le C_{\mathrm{q},\infty} \|x\|_\infty, \qquad \|x\|_\infty \le C_{\infty,\mathrm{q}} \|x\|_{\mathrm{q}}, \qquad \forall x \in W. \tag{89}$$

Assume also the uniform projection estimation error bound (Lemma 14):

$$\|(\widehat{\Pi} - \Pi)u\|_\infty \le \varepsilon_b \|u\|_\infty, \qquad \forall u \in \mathbb{R}^n. \tag{90}$$

Let $B : W \to W$ be the linear operator

$$B \;:=\; I_W - \Pi P|_W, \tag{91}$$

and denote

$$\kappa_B \coloneqq \|B^{-1}\|_{\mathrm{q}\to\mathrm{q}} < \infty. \tag{92}$$

If ε_b is small enough so that

$$\kappa_B\, C_{\mathrm{q},\infty}\, C_{\infty,\mathrm{q}}\, \varepsilon_b < 1, \tag{93}$$

then the fixed point equation

$$\widehat{v}^\star = \widehat{\Pi}(r + P\widehat{v}^\star) \tag{94}$$

admits a unique solution $\widehat{v}^\star \in W$.

Proof. If v satisfies (94), then

$$v = \widehat{\Pi}(r + Pv) \in \mathrm{range}(\widehat{\Pi}) = W. \tag{95}$$

Conversely, for $v \in W$ we have $\widehat{\Pi} v = v$, hence (94) is equivalent to

$$v = \widehat{\Pi} r + \widehat{\Pi} P v \qquad \Longleftrightarrow \qquad (I_W - \widehat{\Pi} P|_W)\, v = \widehat{\Pi} r. \tag{96}$$

Therefore existence/uniqueness of $\widehat{v}^\star$ is equivalent to invertibility of the operator

$$\widehat{B} \;:=\; I_W - \widehat{\Pi} P|_W : W \to W. \tag{97}$$

We show $\ker(B) = \{0\}$. Let $x \in W$ satisfy $Bx = 0$. Then

$$x = \Pi P x. \tag{98}$$

Since $\ker(\Pi) = \mathcal{K}(P)$ and Π is the canonical representative map of the quotient (Theorem 3), we have

$$[x] = [\Pi x] \quad \text{and} \quad [\Pi y] = [y] \quad \text{for all } y, \tag{99}$$

where $[\cdot]$ denotes the equivalence class in $Q := \mathbb{R}^n/\mathcal{K}(P)$. Applying the quotient map $[\cdot]$ to (98) gives

$$[x] = [Px] = \overline{P}[x]. \tag{100}$$

Hence

$$(I - \overline{P})[x] = 0. \tag{101}$$

By Theorem 2, $\rho(\overline{P}) < 1$, so 1 is not an eigenvalue of $\overline{P}$ and therefore $I - \overline{P}$ is invertible on Q. Thus (101) implies

$$[x] = 0 \quad \Longrightarrow \quad x \in \mathcal{K}(P). \tag{102}$$

But $x \in W = \mathrm{range}(\Pi)$ and $\ker(\Pi) = \mathcal{K}(P)$ implies

$$W \cap \mathcal{K}(P) = \{0\}. \tag{103}$$

Therefore $x = 0$, proving $\ker(B) = \{0\}$. Since W is finite-dimensional, B is invertible and $\kappa_B < \infty$ in (92) is well-defined.

We now bound the perturbation $\widehat{B} - B$ in operator norm. From (91)–(97),

$$\widehat{B} - B = (\Pi - \widehat{\Pi})P|_W. \tag{104}$$

Fix $x \in W$. Using (89), (90), and $\|Px\|_\infty \le \|x\|_\infty$,

$$\begin{aligned}
\|(\widehat{B} - B)x\|_{\mathrm{q}} &= \|(\Pi - \widehat{\Pi})Px\|_{\mathrm{q}} && (105)\\
&\le C_{\mathrm{q},\infty}\|(\Pi - \widehat{\Pi})Px\|_\infty && (106)\\
&\le C_{\mathrm{q},\infty}\,\varepsilon_b\,\|Px\|_\infty && (107)\\
&\le C_{\mathrm{q},\infty}\,\varepsilon_b\,\|x\|_\infty && (108)\\
&\le C_{\mathrm{q},\infty}\,C_{\infty,\mathrm{q}}\,\varepsilon_b\,\|x\|_{\mathrm{q}}. && (109)
\end{aligned}$$

Therefore

$$\|\widehat{B} - B\|_{\mathrm{q}\to\mathrm{q}} \le C_{\mathrm{q},\infty}\,C_{\infty,\mathrm{q}}\,\varepsilon_b. \tag{110}$$

We now write

$$\widehat{B} = B - (\widehat{B} - B) = B\Big(I_W - B^{-1}(\widehat{B} - B)\Big). \tag{111}$$

Let

$$E := B^{-1}(\widehat{B} - B). \tag{112}$$

Then by (92) and (110),

$$\|E\|_{\mathrm{q}\to\mathrm{q}} \le \|B^{-1}\|_{\mathrm{q}\to\mathrm{q}} \cdot \|\widehat{B} - B\|_{\mathrm{q}\to\mathrm{q}} \le \kappa_B C_{\mathrm{q},\infty}\,C_{\infty,\mathrm{q}}\,\varepsilon_b. \tag{113}$$

Under the small-gain condition (93), we have $\|E\|_{\mathrm{q}\to\mathrm{q}} < 1$, so $I_W - E$ is invertible with Neumann series

$$(I_W - E)^{-1} = \sum_{m=0}^{\infty} E^m, \tag{114}$$

and hence $\widehat{B}$ is invertible by (111).

To show the existence and uniqueness of $\widehat{v}^\star$), since $\widehat{B}$ is invertible on W, the linear system (96) has the unique solution

$$\widehat{v}^\star = \widehat{B}^{-1}\widehat{\Pi}r \in W. \tag{115}$$

By Step 1 this $\widehat{v}^\star$ is exactly the unique fixed point of (94). □

Lemma 16 (Simplex property of the extended basis)**.** *For every state $s \in S$ and every index $(i,k) \in \mathcal{I}$, we have $b_{i,k}(s) \ge 0$ and*

$$\sum_{(i,k)\in\mathcal{I}} b_{i,k}(s) = 1.$$

Moreover, Algorithm 3 outputs $\widehat{b}_{i,k}$ satisfying $\widehat{b}_{i,k}(s) \ge 0$ and

$$\sum_{(i,k)\in\mathcal{I}} \widehat{b}_{i,k}(s) = 1 \qquad \textit{for all } s \in S.$$

Proof. If $s \in \mathcal{F}$, then exactly one pair (i,k) satisfies $s \in C_{i,k}$, so the sum equals 1 by definition.

If $s \in \mathcal{T}$, then the chain hits $\mathcal{F}$ almost surely and the random pair

$$\Big(\mathsf{cls}(X_{\tau_{\mathcal{F}}}),\ (\mathsf{phase}(X_{\tau_{\mathcal{F}}}) - \tau_{\mathcal{F}}) \bmod d_{\mathsf{cls}(X_{\tau_{\mathcal{F}}})}\Big)$$

takes values in $\mathcal{I}$ and is uniquely defined on each trajectory. The events indexed by $(i,k) \in \mathcal{I}$ form a partition of the sample space, hence their probabilities sum to 1.

For $\widehat{b}$, the recurrent case is set deterministically to indicators. For $s \in \mathcal{T}$, each simulated absorption episode increments exactly one coordinate (i,k) and the subsequent normalization by M forces the sum over $(i,k) \in \mathcal{I}$ to equal 1. □

Lemma 17 (Infinity-norm operator bound for the gauge map)**.** *Assume $\widehat{b}_{i,k}(s) \geq 0$ for all $s \in S$ and $(i,k) \in \mathcal{I}$, and*

$$\sum_{(i,k)\in\mathcal{I}} \widehat{b}_{i,k}(s) = 1 \qquad \text{for all } s \in S.$$

Let $\widehat{\Pi}$ be defined by (77)*. Then for all $v \in \mathbb{R}^n$,*

$$\|\widehat{\Pi} v\|_\infty \leq 2\|v\|_\infty.$$

The same bound holds for Π.

Proof. Fix $s \in S$. By definition,

$$(\widehat{\Pi} v)(s) = v(s) - \sum_{(i,k)\in\mathcal{I}} v(a_{i,k})\, \widehat{b}_{i,k}(s).$$

Taking absolute values and using $\widehat{b}_{i,k}(s) \geq 0$ gives

$$|(\widehat{\Pi} v)(s)| \leq |v(s)| + \sum_{(i,k)\in\mathcal{I}} |v(a_{i,k})|\, \widehat{b}_{i,k}(s) \leq \|v\|_\infty + \|v\|_\infty \sum_{(i,k)\in\mathcal{I}} \widehat{b}_{i,k}(s) = 2\|v\|_\infty.$$

Taking the maximum over s proves the claim. The proof for Π is identical. □

Lemma 18 (Contraction and bounded noise imply bounded iterates in expectation)**.** *Assume the conditions of Lemma 19. Assume the stepsizes satisfy $\alpha_t = \frac{\alpha}{t+t_0}$ and $t_0 > 1$. Let v_t be generated by* (125) *and let $\widehat{v}^\star$ be the fixed point in* (127)*. Then*

$$\sup_{t\geq 0} \mathbb{E}\|v_t - \widehat{v}^\star\|_{\mathrm{q}}^2 \leq \|v_0 - \widehat{v}^\star\|_{\mathrm{q}}^2 + 4\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2 \sum_{t=0}^{\infty} \frac{1}{(t+t_0)^2} \leq \|v_0 - \widehat{v}^\star\|_{\mathrm{q}}^2 + \frac{4\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2}{t_0 - 1}. \tag{116}$$

Moreover, using norm equivalence on $\mathrm{range}(\widehat{\Pi})$ *as in* (124)*,*

$$\sup_{t\geq 0} \mathbb{E}\|v_t\|_\infty^2 \leq 2C_{\infty,\mathrm{q}}^2 \left(\|v_0 - \widehat{v}^\star\|_{\mathrm{q}}^2 + \frac{4\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2}{t_0 - 1} \right) + 2\|\widehat{v}^\star\|_\infty^2. \tag{117}$$

Proof. Let $e_t := v_t - \widehat{v}^\star$. From (133) we have

$$e_{t+1} = (1-\alpha_t) e_t + \alpha_t\, \widehat{\Pi} P e_t + \alpha_t\, \widehat{\Pi} \xi_{t+1}, \tag{118}$$

where ξ_{t+1} is the oracle noise with $\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = 0$ and $\mathbb{E}\big[\|\xi_{t+1}\|_\infty^2 |\mathcal{F}_t\big] \leq \sigma^2$. Furthermore, by the contraction bound in the proof of Lemma 19,

$$\|(1-\alpha_t) e_t + \alpha_t\, \widehat{\Pi} P e_t\|_{\mathrm{q}} \leq \|e_t\|_{\mathrm{q}}. \tag{119}$$

Since $\|\widehat{\Pi} z\|_\infty \leq 2\|z\|_\infty$ by Lemma 17, we obtain

$$\mathbb{E}\big[\|\widehat{\Pi}\xi_{t+1}\|_{\mathrm{q}}^2 |\mathcal{F}_t\big] \leq C_{\mathrm{q},\infty}^2 \mathbb{E}\big[\|\widehat{\Pi}\xi_{t+1}\|_\infty^2 |\mathcal{F}_t\big] \leq 4C_{\mathrm{q},\infty}^2 \sigma^2. \tag{120}$$

Expanding the square in the $\|\cdot\|_{\mathrm{q}}$ inner product and using $\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = 0$ to drop the cross term gives

$$\mathbb{E}\big[\|e_{t+1}\|_{\mathrm{q}}^2 |\mathcal{F}_t\big] \leq \|e_t\|_{\mathrm{q}}^2 + 4\alpha_t^2 C_{\mathrm{q},\infty}^2 \sigma^2. \tag{121}$$

Taking expectation and summing over t yields

$$\mathbb{E}\|e_T\|_{\mathrm{q}}^2 \le \|e_0\|_{\mathrm{q}}^2 + 4\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2 \sum_{t=0}^{T-1} \frac{1}{(t+t_0)^2} \le \|e_0\|_{\mathrm{q}}^2 + \frac{4\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2}{t_0 - 1}. \tag{122}$$

The $\|\cdot\|_\infty$ bound follows from $\|v_t\|_\infty \le \|e_t\|_\infty + \|\widehat{v}^\star\|_\infty$ and $\|e_t\|_\infty \le C_{\infty,\mathrm{q}} \|e_t\|_{\mathrm{q}}$ on $\mathrm{range}(\widehat{\Pi})$. □

Lemma 19 (Formal version of Lemma 7)**.** *Let $T(v) = r + Pv$ with $\|r\|_\infty \le R$ and P row-stochastic. Assume the quotient semi-norm $\|\cdot\|_{\mathrm{q}}$ satisfies*

$$\|Px\|_{\mathrm{q}} \le \gamma \|x\|_{\mathrm{q}} \qquad \textit{for all } x \in \mathbb{R}^n, \tag{123}$$

for some $\gamma \in (0,1)$. Assume also there is a constant $C_{\infty,\mathrm{q}}$ such that on the subspace $\mathrm{range}(\widehat{\Pi})$,

$$\|x\|_\infty \le C_{\infty,\mathrm{q}} \|x\|_{\mathrm{q}}. \tag{124}$$

Consider the recursion (initialized with $v_0 \in \mathrm{range}(\widehat{\Pi})$)

$$v_{t+1} = \widehat{\Pi}\Big((1-\alpha_t)v_t + \alpha_t \widehat{T}(v_t)\Big), \qquad \alpha_t = \frac{\alpha}{t+t_0}, \qquad t_0 \ge \alpha, \tag{125}$$

where the oracle satisfies

$$\mathbb{E}[\widehat{T}(v_t)|\mathcal{F}_t] = r + Pv_t, \qquad \mathbb{E}\Big[\|\widehat{T}(v_t) - (r+Pv_t)\|_\infty^2 |\mathcal{F}_t\Big] \le \sigma^2. \tag{126}$$

with $\mathcal{F}_t = \sigma(v_0, \dots, v_t)$.

Let $\widehat{v}^\star$ be the unique fixed point of $\widehat{F}(v) := \widehat{\Pi}T(v)$ on $\mathrm{range}(\widehat{\Pi})$, i.e.

$$\widehat{v}^\star = \widehat{\Pi}(r + P\widehat{v}^\star). \tag{127}$$

Define

$$\widehat{\gamma} := \gamma + C_{\mathrm{q},\infty} \varepsilon_b \, C_{\infty,\mathrm{q}}. \tag{128}$$

Assume ε_b is chosen so that $\widehat{\gamma} < 1$. Here $C_{\mathrm{q},\infty}$ is from (80) *and ε_b is from Lemma 14.*

Then there exists a finite constant C_{stab} (depending on $\alpha, \widehat{\gamma}, \sigma, R$, and the norm-equivalence constants) such that

$$\sup_{t \ge 0} \mathbb{E}\|v_t - \widehat{v}^\star\|_{\mathrm{q}}^2 \le C_{\mathrm{stab}}. \tag{129}$$

Consequently, there is a finite constant $B_{\sup}$ such that

$$\sup_{t \ge 0} \mathbb{E}\|v_t\|_\infty^2 \le B_{\sup}^2 < \infty. \tag{130}$$

Proof. Since $v_0 \in \mathrm{range}(\widehat{\Pi})$ and $\widehat{\Pi}$ is a projection, we have $v_t = \widehat{\Pi} v_t$ for all t. Thus (125) becomes

$$v_{t+1} = (1-\alpha_t)v_t + \alpha_t \widehat{\Pi}\widehat{T}(v_t) = v_t + \alpha_t\Big(\widehat{\Pi}T(v_t) - v_t\Big) + \alpha_t \widehat{\Pi}\xi_{t+1}, \tag{131}$$

where $\xi_{t+1} := \widehat{T}(v_t) - T(v_t)$ satisfies

$$\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi_{t+1}\|_\infty^2 |\mathcal{F}_t] \le \sigma^2. \tag{132}$$

Define the error $e_t := v_t - \widehat{v}^\star$. Using (127), subtract $\widehat{v}^\star$ from (131):

$$e_{t+1} = (1-\alpha_t)e_t + \alpha_t \widehat{\Pi} P e_t + \alpha_t \widehat{\Pi}\xi_{t+1}. \tag{133}$$

For any $x \in \mathrm{range}(\widehat{\Pi})$, write

$$\widehat{\Pi}Px = \Pi Px + (\widehat{\Pi} - \Pi)Px. \tag{134}$$

Hence, using (123), Lemma 14 (in $\|\cdot\|_{\mathrm{q}}$), and $\|Px\|_\infty \le \|x\|_\infty$,

$$\begin{aligned}
\|\widehat{\Pi}Px\|_{\mathrm{q}} &\le \|\Pi Px\|_{\mathrm{q}} + \|(\widehat{\Pi} - \Pi)Px\|_{\mathrm{q}} && (135)\\
&\le \gamma\|x\|_{\mathrm{q}} + C_{\mathrm{q},\infty}\varepsilon_b \|Px\|_\infty && (136)\\
&\le \gamma\|x\|_{\mathrm{q}} + C_{\mathrm{q},\infty}\varepsilon_b \|x\|_\infty && (137)\\
&\le \Big(\gamma + C_{\mathrm{q},\infty}\varepsilon_b \, C_{\infty,\mathrm{q}}\Big)\|x\|_{\mathrm{q}} = \widehat{\gamma}\|x\|_{\mathrm{q}}. && (138)
\end{aligned}$$

Apply this to $x = e_t \in \mathrm{range}(\widehat{\Pi})$ and combine with (133):

$$\|(1-\alpha_t)e_t + \alpha_t \widehat{\Pi} P e_t\|_{\mathrm{q}} \le \big(1 - \alpha_t(1-\widehat{\gamma})\big)\|e_t\|_{\mathrm{q}}. \tag{139}$$

Let $d_t := (1-\alpha_t)e_t + \alpha_t \widehat{\Pi}(Pe_t)$, then (133) becomes

$$e_{t+1} = d_t + \alpha_t \widehat{\Pi}\xi_{t+1}. \tag{140}$$

Since $\|\cdot\|_{\mathrm{q}}$ is induced by an inner product, we can expand the square and use $\mathcal{F}_t$-measurability of d_t and $\widehat{\Pi}$ to get

$$\begin{aligned}\mathbb{E}\big[\|e_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t\big] &= \|d_t\|_{\mathrm{q}}^2 + 2\alpha_t\,\mathbb{E}\Big[\langle d_t, \widehat{\Pi}\xi_{t+1}\rangle_{\mathrm{q}}|\mathcal{F}_t\Big] + \alpha_t^2\,\mathbb{E}\Big[\|\widehat{\Pi}\xi_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t\Big] \\ &= \|d_t\|_{\mathrm{q}}^2 + \alpha_t^2\,\mathbb{E}\Big[\|\widehat{\Pi}\xi_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t\Big], \end{aligned} \tag{141}$$

where the cross term vanishes because $\mathbb{E}[\xi_{t+1}|\mathcal{F}_t] = 0$.

Next, using (139) and $\alpha_t \in (0,1]$,

$$\|d_t\|_{\mathrm{q}} \le (1-\alpha_t)\|e_t\|_{\mathrm{q}} + \alpha_t\widehat{\gamma}\,\|e_t\|_{\mathrm{q}} = \big(1-\alpha_t(1-\widehat{\gamma})\big)\|e_t\|_{\mathrm{q}}. \tag{142}$$

Therefore

$$\|d_t\|_{\mathrm{q}}^2 \le \big(1-\alpha_t(1-\widehat{\gamma})\big)^2\|e_t\|_{\mathrm{q}}^2. \tag{143}$$

Finally, Lemma 17 gives $\|\widehat{\Pi}z\|_\infty \le 2\|z\|_\infty$, so with $\|x\|_{\mathrm{q}} \le C_{\mathrm{q},\infty}\|x\|_\infty$ we obtain

$$\mathbb{E}\Big[\|\widehat{\Pi}\xi_{t+1}\|_{\mathrm{q}}^2|\mathcal{F}_t\Big] \le 4C_{\mathrm{q},\infty}^2\,\mathbb{E}\big[\|\xi_{t+1}\|_\infty^2|\mathcal{F}_t\big] \le 4C_{\mathrm{q},\infty}^2\sigma^2. \tag{144}$$

Combining (141), (143), and (144) and taking total expectation yields

$$\mathbb{E}\|e_{t+1}\|_{\mathrm{q}}^2 \le \big(1-\alpha_t(1-\widehat{\gamma})\big)^2\,\mathbb{E}\|e_t\|_{\mathrm{q}}^2 + 4\alpha_t^2 C_{\mathrm{q},\infty}^2\sigma^2. \tag{145}$$

Since $\sum_t \alpha_t^2 < \infty$ and $\sum_t \alpha_t = \infty$, the recursion (145) implies $\sup_t \mathbb{E}\|e_t\|_{\mathrm{q}}^2 < \infty$ by a standard Robbins–Siegmund argument. This proves (129).

Finally, (130) follows from (124) and $\|v_t\|_{\mathrm{q}} \le \|v_t - \widehat{v}^\star\|_{\mathrm{q}} + \|\widehat{v}^\star\|_{\mathrm{q}}$, plus norm equivalence between $\|\cdot\|_{\mathrm{q}}$ and $\|\cdot\|_\infty$ on the finite-dimensional space. □

Theorem 10 (Formal version of Theorem 7)**.** *Assume the conditions of Lemma 19 hold and that $\alpha_t = \alpha/(t+t_0)$ with $t_0 \ge \max\{\alpha, 1\}$. Let v_t be generated by* (125)*, and let $\widehat{v}^\star$ be the unique fixed point of $\widehat{F}(v) = \widehat{\Pi}(r + Pv)$ in $W = \mathrm{range}(\widehat{\Pi})$.*

Define $a := \alpha(1-\widehat{\gamma})$. Assume $a > 1/2$. Then for every $T \ge 0$,

$$\mathbb{E}\|v_T - \widehat{v}^\star\|_{\mathrm{q}} \le \Big(\frac{t_0}{T+t_0}\Big)^a\,\|v_0 - \widehat{v}^\star\|_{\mathrm{q}} + \frac{2^{a+1}\alpha\, C_{\mathrm{q},\infty}\sigma}{\sqrt{(2a-1)(T+t_0)}}. \tag{146}$$

Let $v^\star$ be the fixed point of $F(v) = \Pi(r+Pv)$ in $\mathrm{range}(\Pi)$*. On the event* (78)*, the fixed points satisfy*

$$\|\widehat{v}^\star - v^\star\|_{\mathrm{q}} \le \frac{C_{\mathrm{q},\infty}\varepsilon_b}{1-\gamma}\Big(R + \|\widehat{v}^\star\|_\infty\Big). \tag{147}$$

Moreover,

$$\|\widehat{v}^\star\|_\infty \le \frac{2C_{\infty,\mathrm{q}}C_{\mathrm{q},\infty}R}{1-\widehat{\gamma}}. \tag{148}$$

Combining (146)*,* (147)*, and* (148) *yields an explicit bound on $\mathbb{E}\|v_T - v^\star\|_{\mathrm{q}}$.*

Proof. Let $e_t := v_t - \widehat{v}^\star$. By Lemma 19, we have the recursion

$$\mathbb{E}\|e_{t+1}\|_{\mathrm{q}}^2 \le \big(1-\alpha_t(1-\widehat{\gamma})\big)^2\,\mathbb{E}\|e_t\|_{\mathrm{q}}^2 + 4\alpha_t^2\,C_{\mathrm{q},\infty}^2\sigma^2. \tag{149}$$

With $\alpha_t = \alpha/(t+t_0)$ and $a = \alpha(1-\widehat{\gamma})$, this becomes

$$m_{t+1} \le \Big(1 - \frac{a}{t+t_0}\Big)^2 m_t + \frac{4\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2}{(t+t_0)^2}, \tag{150}$$

where $m_t := \mathbb{E}\|e_t\|_{\mathrm{q}}^2$.

Define $w_t := (t+t_0)^{2a} m_t$. Using $(1+1/(t+t_0))^{2a} \le \exp(2a/(t+t_0))$ and $(1-a/(t+t_0))^2 \le \exp(-2a/(t+t_0))$, we obtain

$$(t+t_0+1)^{2a}\Big(1 - \frac{a}{t+t_0}\Big)^2 \le (t+t_0)^{2a}. \tag{151}$$

Multiplying (150) by $(t+t_0+1)^{2a}$ yields

$$w_{t+1} \le w_t + 4\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2 \cdot \frac{(t+t_0+1)^{2a}}{(t+t_0)^2}. \tag{152}$$

Since $(t+t_0+1)^{2a} \le 2^{2a}(t+t_0)^{2a}$, we obtain

$$w_{t+1} \le w_t + 2^{2a+2}\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2 \,(t+t_0)^{2a-2}. \tag{153}$$

Summing from $t=0$ to $T-1$ gives

$$w_T \le w_0 + 2^{2a+2}\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2 \sum_{t=0}^{T-1} (t+t_0)^{2a-2}. \tag{154}$$

When $a > 1/2$, we have $2a-2 > -1$, so

$$\sum_{t=0}^{T-1} (t+t_0)^{2a-2} \le \int_{t_0-1}^{T+t_0} x^{2a-2}\,dx = \frac{(T+t_0)^{2a-1} - (t_0-1)^{2a-1}}{2a-1} \le \frac{(T+t_0)^{2a-1}}{2a-1}. \tag{155}$$

Therefore,

$$m_T = \frac{w_T}{(T+t_0)^{2a}} \le \Big(\frac{t_0}{T+t_0}\Big)^{2a} m_0 + \frac{2^{2a+2}\alpha^2 C_{\mathrm{q},\infty}^2 \sigma^2}{(2a-1)(T+t_0)}. \tag{156}$$

Taking square roots and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ yields (146).

For the fixed point gap, start from

$$\widehat{v}^\star - v^\star = (\widehat{\Pi} - \Pi)(r + P\widehat{v}^\star) + \Pi\big(P(\widehat{v}^\star - v^\star)\big). \tag{157}$$

Taking $\|\cdot\|_{\mathrm{q}}$ and using $\|\Pi(Px)\|_{\mathrm{q}} \le \gamma\|x\|_{\mathrm{q}}$ gives

$$\|\widehat{v}^\star - v^\star\|_{\mathrm{q}} \le \|(\widehat{\Pi} - \Pi)(r + P\widehat{v}^\star)\|_{\mathrm{q}} + \gamma\|\widehat{v}^\star - v^\star\|_{\mathrm{q}}. \tag{158}$$

Rearranging yields

$$\|\widehat{v}^\star - v^\star\|_{\mathrm{q}} \le \frac{1}{1-\gamma}\|(\widehat{\Pi} - \Pi)(r + P\widehat{v}^\star)\|_{\mathrm{q}}. \tag{159}$$

On (78), Lemma 14 implies $\|(\widehat{\Pi} - \Pi)u\|_\infty \le \varepsilon_b\|u\|_\infty$, so

$$\|(\widehat{\Pi} - \Pi)(r + P\widehat{v}^\star)\|_{\mathrm{q}} \le C_{\mathrm{q},\infty}\varepsilon_b\|r + P\widehat{v}^\star\|_\infty \le C_{\mathrm{q},\infty}\varepsilon_b\big(R + \|\widehat{v}^\star\|_\infty\big), \tag{160}$$

which gives (147).

Finally, from the fixed point equation $\widehat{v}^\star = \widehat{\Pi}r + \widehat{\Pi}(P\widehat{v}^\star)$, we get

$$\|\widehat{v}^\star\|_{\mathrm{q}} \le \|\widehat{\Pi}r\|_{\mathrm{q}} + \widehat{\gamma}\|\widehat{v}^\star\|_{\mathrm{q}}, \qquad \|\widehat{v}^\star\|_{\mathrm{q}} \le \frac{\|\widehat{\Pi}r\|_{\mathrm{q}}}{1-\widehat{\gamma}}. \tag{161}$$

Using $\|\widehat{\Pi}r\|_{\mathrm{q}} \le C_{\mathrm{q},\infty}\|\widehat{\Pi}r\|_\infty \le 2C_{\mathrm{q},\infty}R$ and then $\|\cdot\|_\infty \le C_{\infty,\mathrm{q}}\|\cdot\|_{\mathrm{q}}$ yields (148). □

Corollary 6. *Work on the event of Theorem 9 and Lemma 12. Assume* (126) *holds and assume* $\widehat{\gamma} < 1$ *in* (128)*.*

Run Algorithm 5 with stepsizes

$$\alpha_t = \frac{\alpha}{t + t_0}, \qquad t_0 \geq \alpha, \tag{162}$$

and initialization $v_0 \in \mathrm{range}(\widehat{\Pi})$*.*

Let $\widehat{v}^\star$ *be the unique fixed point of* $\widehat{F}(v) = \widehat{\Pi}(r + Pv)$ *on* $W = \mathrm{range}(\widehat{\Pi})$ *and let* $v^\star$ *be the fixed point of* $F(v) = \Pi(r + Pv)$ *on* $\mathrm{range}(\Pi)$*.*

Define

$$a := \alpha(1 - \widehat{\gamma}), \qquad \text{and assume } a > \frac{1}{2}.$$

Then for every $T \geq 0$*,*

$$\mathbb{E}\|v_T - v^\star\|_{\mathrm{q}} \leq \Big(\frac{t_0}{T + t_0}\Big)^a \|v_0 - \widehat{v}^\star\|_{\mathrm{q}} + \frac{2^{a+1}\alpha\, C_{\mathrm{q},\infty}\sigma}{\sqrt{(2a-1)(T+t_0)}} + \frac{C_{\mathrm{q},\infty}\varepsilon_b}{1-\gamma}\Big(R + \|\widehat{v}^\star\|_\infty\Big). \tag{163}$$

Moreover, $\|\widehat{v}^\star\|_\infty$ *is bounded by* (148)*.*

Proof. Apply Theorem 10 with $K = t_0$. □

Corollary 7 (Formal version of Corollary 4)**.** *Fix a target accuracy* $\varepsilon \in (0, 1)$*. Assume* $p_{\min} > 0$ *in* (26)*. Assume the SA stepsizes are* $\alpha_t = \alpha/(t + t_0)$ *with* $t_0 \geq \alpha$ *and* $a = \alpha(1 - \widehat{\gamma}) > 1/2$*.*

Choose

$$K_{\mathrm{graph}} \geq \frac{1}{p_{\min}} \log \frac{n^2}{\delta_1}.$$

Choose ε_b *so that* $\widehat{\gamma} < 1$ *and*

$$\frac{C_{\mathrm{q},\infty}\varepsilon_b}{1-\gamma}\Big(R + \|\widehat{v}^\star\|_\infty\Big) \leq \frac{\varepsilon}{3}.$$

Choose

$$M_{\mathrm{abs}} \;\geq\; \frac{8}{\varepsilon_b^2}\Big(N + \log\frac{|\mathcal{T}|}{\delta_2}\Big).$$

Choose the SA horizon T *so that*

$$\Big(\frac{t_0}{T + t_0}\Big)^a \|v_0 - \widehat{v}^\star\|_{\mathrm{q}} \leq \frac{\varepsilon}{3}, \qquad \frac{2^{a+1}\alpha\, C_{\mathrm{q},\infty}\sigma}{\sqrt{(2a-1)(T+t_0)}} \leq \frac{\varepsilon}{3}.$$

Then on the event of Theorem 9 and Lemma 12,

$$\mathbb{E}\|[v_T] - [v^\star]\|_{\mathrm{q}} = \mathbb{E}\|v_T - v^\star\|_{\mathrm{q}} \leq \varepsilon.$$

Moreover, the expected number of simulator queries is at most

$$nK_{\mathrm{graph}} + |\mathcal{T}|\, M_{\mathrm{abs}}\, H_{\mathrm{abs}} + nT.$$

Remark 5. *If one insists on analyzing the recursion* (125) *directly relative to the true fixed point* $v^\star$ *(and treats* $(\widehat{\Pi} - \Pi)T(v_t)$ *as a bounded drift bias), then because* $\sum_{t \leq T} \alpha_t = \Theta(\log T)$ *one can indeed derive a term of order* $\varepsilon_b \log T/(1-\gamma)$*, exactly like the biased-noise term in standard contractive SA.*

The sharper analysis above avoids this artifact by (i) introducing the perturbed contraction $\widehat{F}(v) = \widehat{\Pi}T(v)$*, (ii) proving SA converges to its fixed point* $\widehat{v}^\star$ *at the usual* $O(1/\sqrt{T})$ *rate, and (iii) bounding the time-independent fixed-point gap* $\|\widehat{v}^\star - v^\star\|$*. This converts a would-be* $\varepsilon \log T$ *accumulation into an* $\varepsilon/(1-\gamma)$ *error floor.*

G.3 Proofs in Section B.2

Recall from Theorem 3 that the unique gauge-fixed solution $v^\star \in \mathrm{range}(\Pi)$ induces a peripheral residual

$$g^\star := r + Pv^\star - v^\star \in \mathcal{K}(P). \tag{164}$$

When the peripheral space is represented via the basis $\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ from Definition 1 with anchors $\{a_{i,k}\}$, we can recover $g^\star$ coordinate-wise.

Lemma 20. *Let $\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ be as in Definition 1, and choose anchors $a_{i,k} \in C_{i,k}$. Then for all $(i,k),(j,\ell) \in \mathcal{I}$,*

$$b_{j,\ell}(a_{i,k}) = \mathbf{1}\{(j,\ell) = (i,k)\}. \tag{165}$$

Consequently, every $g \in \mathrm{span}\{b_{i,k}\}_{(i,k)\in\mathcal{I}}$ admits the unique expansion

$$g(\cdot) = \sum_{(i,k)\in\mathcal{I}} g(a_{i,k})\, b_{i,k}(\cdot). \tag{166}$$

Proof. For $a_{i,k} \in C_{i,k} \subseteq \mathcal{F}$, Definition 1 gives $b_{j,\ell}(a_{i,k}) = \mathbf{1}\{a_{i,k} \in C_{j,\ell}\}$, which equals $\mathbf{1}\{(j,\ell) = (i,k)\}$, proving (10). For (166), write $g = \sum_{(j,\ell)} \theta_{j,\ell} b_{j,\ell}$. Evaluating at $a_{i,k}$ and using (10) yields $g(a_{i,k}) = \theta_{i,k}$. Uniqueness follows immediately. □

Definition 2 (Peripheral coordinates of the residual)**.** *Let $v^\star$ and $g^\star$ be as in Theorem 3. Define the peripheral coordinates*

$$\theta^\star_{i,k} := g^\star(a_{i,k}), \qquad (i,k) \in \mathcal{I}. \tag{167}$$

Then $g^\star(\cdot) = \sum_{(i,k)\in\mathcal{I}} \theta^\star_{i,k} b_{i,k}(\cdot)$ by Lemma 20.

Theorem 11 (Formal version of Theorem 8)**.** *Fix anchors $\{a_{i,k}\}$ and basis functions $\{b_{i,k}\}$ from Definition 1. Let $v^\star$ and $\theta^\star$ be as in Definition 2. Let $v \in \mathbb{R}^n$ be any possibly random gauge-fixed vector with $v(a_{i,k}) = 0$ for all $(i,k) \in \mathcal{I}$.*

Run Algorithm 6 with J samples per anchor. Define $B_v^2 := \mathbb{E}\|v\|_\infty^2$ and $N := |\mathcal{I}|$. Then

$$\mathbb{E}\Big[\max_{(i,k)\in\mathcal{I}} |\hat\theta_{i,k} - \theta^\star_{i,k}|\Big] \le \mathbb{E}\|v - v^\star\|_\infty + \sqrt{\frac{2B_v^2 \log(2N)}{J}}. \tag{168}$$

Moreover, if (78) *holds with parameter ε_b, then*

$$\mathbb{E}\|\hat g - g^\star\|_\infty \le \mathbb{E}\Big[\max_{(i,k)\in\mathcal{I}} |\hat\theta_{i,k} - \theta^\star_{i,k}|\Big] + \|g^\star\|_\infty\, \varepsilon_b. \tag{169}$$

Proof. Fix (i,k) and condition on v. The estimator $\widehat{(Pv)}(a_{i,k})$ is the sample mean of J i.i.d. draws of $v(Y)$ with $Y \sim P(a_{i,k},\cdot)$. The centered mean $\widehat{(Pv)}(a_{i,k}) - (Pv)(a_{i,k})$ is subgaussian with proxy variance at most $\|v\|_\infty^2/J$. A maximal inequality for N subgaussian variables gives

$$\mathbb{E}\Big[\max_{(i,k)} |\widehat{(Pv)}(a_{i,k}) - (Pv)(a_{i,k})| \;\Big|\; v\Big] \le \|v\|_\infty \sqrt{\frac{2\log(2N)}{J}}. \tag{170}$$

Unconditioning and applying Cauchy–Schwarz yields

$$\mathbb{E}\Big[\max_{(i,k)} |\widehat{(Pv)}(a_{i,k}) - (Pv)(a_{i,k})|\Big] \le \sqrt{\frac{2B_v^2 \log(2N)}{J}}. \tag{171}$$

For the bias term, $(Pv)(a) - (Pv^\star)(a) \le \|v - v^\star\|_\infty$ since P is row-stochastic. Combining gives the first inequality.

For the residual, write $\hat g - g^\star = \sum_{(i,k)} (\hat\theta_{i,k} - \theta^\star_{i,k})\widehat{b}_{i,k} + \sum_{(i,k)} \theta^\star_{i,k}(\widehat{b}_{i,k} - b_{i,k})$. For the residual, fix $s \in S$ and write

$$(\hat g - g^\star)(s) = \sum_{(i,k)} (\hat\theta_{i,k} - \theta^\star_{i,k})\, \widehat{b}_{i,k}(s) + \sum_{(i,k)} \theta^\star_{i,k} \big(\widehat{b}_{i,k}(s) - b_{i,k}(s)\big).$$

Table 1: Experiment design map. Each row targets a specific claim or possible objection.

Question tested	Experimental object	Evidence reported
Can classical bias be large without transient cost?	Deterministic periodic cycle	$\|h\|_\infty = 3$ while $\|v_\Pi^\star\|_\infty = 0$
Can two starts have the same eventual regime but different finite-horizon behavior?	Two deterministic paths into the same recurrent anchor	Same $g_\Pi^\star$, different $v_\Pi^\star$ and J_H curves
Can average reward hide persistent phase structure?	Two cycles with equal Cesàro average	Same ρ, different $g_\Pi^\star$
Does the refinement disappear when no periodic persistence exists?	Aperiodic reduction check	$\|g_\Pi^\star - \rho\|_\infty$ at numerical round-off
Can the estimator learn the decomposition from samples?	Periodic two-class MRP with transient branching	Errors in $g_\Pi^\star$, $v_\Pi^\star$, and J_{40}
Why not only estimate the classical gain?	Oracle avg-only comparator	Avg-only has no statistical error in ρ but large diagnostic error
Why not estimate the full transition matrix?	Dense plug-in baseline	Plug-in is competitive; ours is compared honestly and not claimed to dominate uniformly
Which pipeline component causes residual error?	Structure, weight, projection ablation	Exact absorption weights make profile error numerical
How fragile is structural recovery?	Single-edge deletion check	$0/582$ deletions change structural invariants in this benchmark

By Lemma 16, $\widehat{b}_{i,k}(s) \geq 0$ and $\sum_{(i,k)} \widehat{b}_{i,k}(s) = 1$, hence

$$\left| \sum_{(i,k)} (\hat{\theta}_{i,k} - \theta^\star_{i,k})\, \widehat{b}_{i,k}(s) \right| \leq \max_{(i,k)} |\hat{\theta}_{i,k} - \theta^\star_{i,k}|.$$

For the second term, use $|\theta^\star_{i,k}| \leq \|g^\star\|_\infty$ and (78) to obtain

$$\sum_{(i,k)} |\theta^\star_{i,k}| \cdot |\widehat{b}_{i,k}(s) - b_{i,k}(s)| \leq \|g^\star\|_\infty \sum_{(i,k)} |\widehat{b}_{i,k}(s) - b_{i,k}(s)| = \|g^\star\|_\infty \, \|\widehat{b}(s) - b(s)\|_1 \leq \|g^\star\|_\infty \, \varepsilon_b.$$

Taking $\max_s$ and expectations yields (169). □

H Numerical Experiments

The experiments are designed to validate the central attribution claim of the paper: the decomposition separates persistent regime behavior from transient-to-regime cost. We therefore do not use average-reward error as the only metric. Instead, we report errors for the three quantities that appear in the theory: the persistent profile $g_\Pi^\star$, the anchor-gauge transient component $v_\Pi^\star$, and the finite-horizon return J_H. The experiments are synthetic by design. They isolate exactly the reducible and periodic pathologies studied in the paper, allow exact ground truth for all decomposition components, and permit component-level ablations of the estimator. Throughout, the estimator has access only to a tabular generative model for transitions. The exact transition matrix is used only to compute ground truth and plug-in baselines.

Table 1 summarizes which question each experiment is intended to answer. The table is included to make clear that the numerical section is not a generic benchmark suite: each experiment targets one claim made in the theory.

H.1 Experiment design and metrics

We evaluate finite Markov reward processes (P, r), which are the policy-induced objects obtained after fixing a stationary policy in an MDP. Rewards are deterministic and bounded, and a simulator query at state s returns an independent sample from $P(s, \cdot)$. The main metrics are

$$E_g = \|\widehat{g} - g_\Pi^\star\|_\infty, \qquad E_v = \|\widehat{v} - v_\Pi^\star\|_\infty, \qquad E_J(H) = \|\widehat{J}_H - J_H\|_\infty,$$

where

$$\widehat{J}_H = \sum_{t=0}^{H-1} P^t \widehat{g} + \widehat{v} - P^H \widehat{v}.$$

These metrics match Corollary 3: persistent-profile error accumulates over the horizon, while transient-component error enters through boundary terms. We use $H = 40$ in the sample-based benchmark.

We compare three estimators. **Ours** learns the recurrent classes, cyclic phases, anchors, and phase-offset absorption weights, then runs projected stochastic approximation in the learned anchor gauge and reconstructs $\widehat{g}$ from anchor residuals. **Avg-only** is an oracle classical-gain comparator: it is given the exact phase-averaged profile $\rho = P^\infty r$ and predicts returns using only $H\rho$. This comparator is intentionally strong statistically, but it cannot represent non-invariant persistent phase behavior or transient cost. It is not a TD baseline: a full gain–bias estimator retains the bias and can be postprocessed using the same gauge. **Plug-in** estimates the full transition matrix $\widehat{P}$ from samples, recovers the structure of $\widehat{P}$, and computes the same decomposition from the empirical model. Plug-in is a natural and strong tabular baseline. Our claim is not uniform numerical dominance over plug-in solvers on small tabular models; the claim is that the proposed estimator targets the persistent-transient decomposition directly and makes the sources of evaluation error interpretable.

H.2 Exact diagnostic checks

We first run four exact diagnostic checks, shown in Figure 1 and summarized in Table 2. These experiments use exact transition matrices and are meant to test whether the decomposition has the intended semantics.

For reproducibility, the four exact checks use the following constructions. In panel (a), $S = \{0, \dots, 23\}$, $P(k, k+1 \bmod 24) = 1$, and $r(k) = 1\{k < 12\}$. Since the whole state space is recurrent and periodic, $\mathcal{K}(P) = \mathbb{R}^{24}$ and the anchor-gauge projection is the zero map. In panel (b), a single absorbing recurrent anchor a has reward $r(a) = 1$, while two deterministic transient chains of lengths 3 and 14 have reward zero and then enter a. A state at distance ℓ from a has the same persistent profile $g^\star_\Pi = 1$ but transient component $v^\star_\Pi = -\ell$. In panel (c), two deterministic 4-cycles have reward profiles $r_A = (1, 1, 0, 0)$ and $r_B = (1, 0, 1, 0)$. Both have classical gain $1/2$, but their persistent profiles differ in phase. In panel (d), we compare an aperiodic chain, where Proposition 1 implies $\mathcal{K}(P) = \ker(I - P)$ and therefore no non-invariant persistent residual remains, with a periodic chain whose phase profile differs from its invariant average.

Pure periodic chain. The first example is a deterministic 24-cycle with reward one on the first half of the cycle and zero on the second half. Since the process starts already inside its recurrent regime, there is no transient-to-regime path. The proposed decomposition gives $v^\star_\Pi = 0$ and $g^\star_\Pi = r$. In contrast, the classical gain is the constant average $\rho = 1/2$, and the classical normalized bias is nonzero because it must absorb the phase pattern discarded by ρ. This confirms the motivation behind Proposition 3: a large classical bias need not mean a large transient burden.

Same eventual regime, different transient burden. The second example compares two starts that eventually enter the same unit-reward regime but after different deterministic path lengths. The persistent profile is identical, $g^\star_\Pi = 1$, but the anchor-gauge transient costs differ: $v^\star_\Pi = -3$ versus $v^\star_\Pi = -14$. The finite-horizon return curves differ accordingly. This is the attribution behavior the paper is designed to recover: the eventual regime is good in both cases, but one path pays a much larger transient cost before reaching it.

Same average reward, different persistent profile. The third example compares two periodic profiles with the same Cesàro average reward. The classical gain cannot distinguish them, so $\|\rho_A - \rho_B\|_\infty = 0$. The persistent profiles differ by $\|g^\star_A - g^\star_B\|_\infty = 1$, however, because the phase-resolved regimes are different. This verifies that average reward alone is too coarse when phase behavior persists forever.

Aperiodic reduction. The fourth example checks that the proposed refinement disappears when it should. In an aperiodic chain, the non-invariant persistent part is numerically zero: $\|g^\star_\Pi - \rho\|_\infty =$

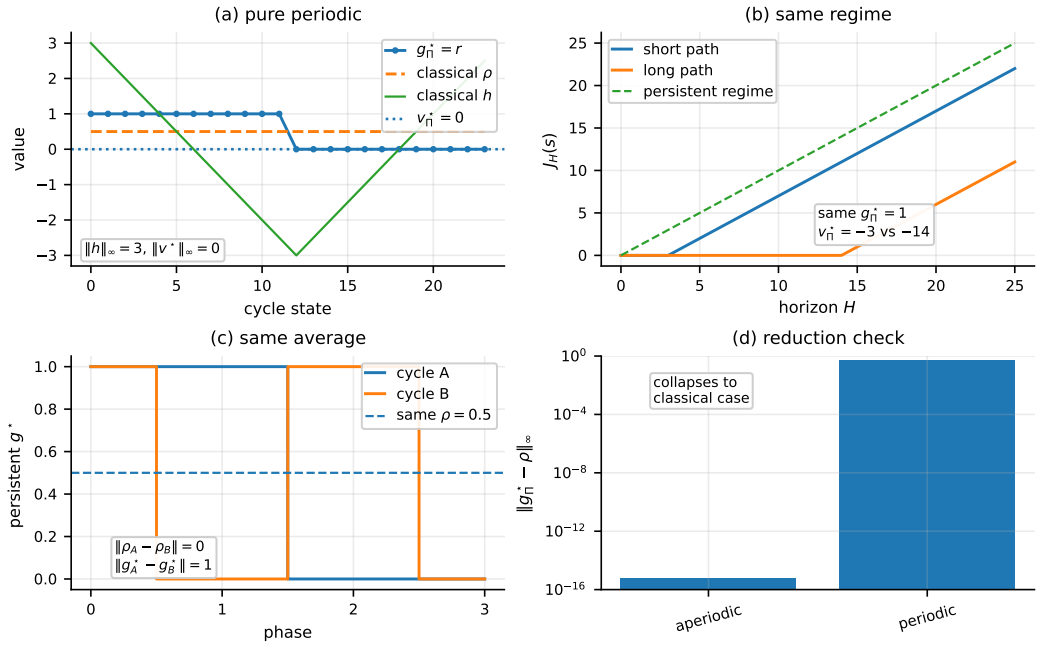

Figure 1: Exact diagnostic checks for the persistent-transient decomposition. Panel (a) shows that a deterministic periodic chain has nonzero classical bias even though the anchor-gauge transient component is zero. Panel (b) separates identical eventual regimes from different transient path costs. Panel (c) shows that equal average reward can hide different persistent phase profiles. Panel (d) verifies collapse to the classical picture in the aperiodic case.

Table 2: Exact diagnostic checks. The numbers isolate the attribution issue rather than benchmarking statistical efficiency.

Diagnostic claim	Observed values
Pure periodic: classical bias need not be transient	$\|h\|_\infty = 3,\ \|v^\star_\Pi\|_\infty = 0$
Same regime, different path cost	$g^\star_\Pi = 1$ for both starts, $v^\star_\Pi = -3$ versus -14
Same average, different persistent profile	$\|\rho_A - \rho_B\|_\infty = 0,\ \|g^\star_A - g^\star_B\|_\infty = 1$
Aperiodic reduction	$\|g^\star_\Pi - \rho\|_\infty = 5.6 \times 10^{-16}$ aperiodic versus 0.5 periodic

5.6×10^{-16}. In a periodic comparison chain the corresponding gap is 0.5. Thus the decomposition reduces to the classical picture in aperiodic models and refines it only when non-invariant peripheral behavior exists.

H.3 Sample-based decomposition estimation

We next test the learned estimator on a periodic two-class MRP with transient branching. The benchmark has two recurrent classes with periods $d_1 = 2$ and $d_2 = 3$, phase sizes $m_1 = 10$ and $m_2 = 9$, and a transient line of length $L = 35$, giving $n = 82$ states and $N = d_1 + d_2 = 5$ persistent coordinates. The recurrent phase rewards are

$$(0.05, 0.95) \quad \text{and} \quad (0.10, 0.55, 0.95),$$

while transient rewards are zero. From transient state t_j, the chain either self-loops, moves forward along the transient line, or exits to one of the two recurrent classes. We use self-loop parameter $\epsilon = 0.20$, exit mass $\eta = 0.08$, and a linearly varying class-one exit probability $q_j \in [0.15, 0.85]$. This construction creates both periodic phase structure and state-dependent absorption into recurrent classes.

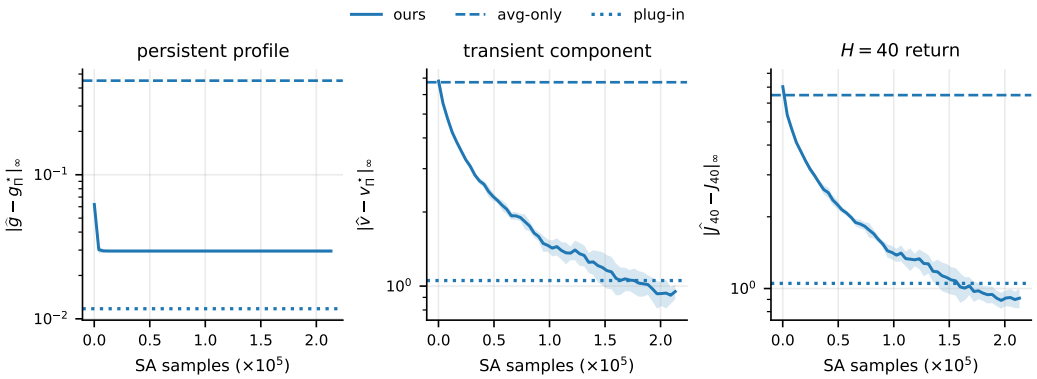

Figure 2: Sample-based estimation of the persistent-transient decomposition on a periodic multichain MRP. The average-only comparator is an oracle phase-averaged gain profile and therefore has no statistical error in estimating ρ, but it discards phase-resolved persistence and transient cost. The plug-in comparator estimates the full transition matrix and computes the same decomposition from the empirical model. Ours directly estimates the decomposition and substantially improves over the average-only comparator on persistent profile, transient component, and $H = 40$ return prediction.

Table 3: Final sample-based errors on the periodic multichain benchmark. Entries are mean $\pm$ one standard deviation over five seeds. The average-only comparator is an oracle classical-gain profile; the plug-in comparator estimates the full transition matrix before computing the same decomposition from the empirical model.

Metric	Ours	Avg-only	Plug-in
Persistent profile $\|\widehat{g} - g^\star_\Pi\|_\infty$	0.0295 ± 0.0009	0.45	0.0118 ± 0.0005
Transient component $\|\widehat{v} - v^\star_\Pi\|_\infty$	0.951 ± 0.08	6.73	1.05 ± 0.32
Return prediction $\|\widehat{J}_{40} - J_{40}\|_\infty$	0.911 ± 0.10	6.48	1.05 ± 0.32

For the learned estimator, support recovery uses $K = 180$ transition samples per state. Phase-offset absorption weights use $M = 900$ absorption episodes per transient state. Projected stochastic approximation runs for 2600 synchronous iterations, drawing one transition sample per state at each iteration, with stepsize $\alpha_t = 1.5(t+80)^{-0.72}$. Anchor residuals are estimated with $J = 100$ samples per anchor. Results are averaged over five random seeds. Figure 2 plots the three errors as a function of stochastic-approximation transition samples; the support and absorption-weight budgets are fixed for the curve. The shaded bands are approximate 95% confidence intervals over seeds.

Table 3 gives the final errors. Relative to the oracle average-only comparator, the learned quotient estimator reduces persistent-profile error by about 93%, transient-component error by about 86%, and $H = 40$ return-prediction error by about 86%. This gap is not caused by a weak statistical baseline: the average-only comparator is given the exact invariant gain. Its failure is semantic. It targets the wrong object for diagnostic evaluation because it removes phase-resolved persistent behavior and has no transient component. The plug-in baseline is strongest on persistent-profile reconstruction in this small tabular instance, while ours is comparable or slightly better on transient and return errors. This is consistent with the paper's positioning: plug-in is a strong model-based comparator, whereas the proposed method estimates the decomposition through its structural components and reports errors that correspond directly to the theory.

H.4 Ablation and support-structure robustness

The estimator has three conceptually separate stages: structural recovery, phase-offset absorption-weight estimation, and projected stochastic approximation. Table 4 ablates these stages on the same benchmark. The ablation uses $M = 500$ absorption episodes, 1200 stochastic-approximation iterations, $J = 80$ anchor-residual samples, and three random seeds. The *Full learned* row uses

Table 4: Component ablation on the sample-based benchmark. Exact absorption weights make persistent-profile error essentially zero, indicating that the residual profile error is due to Monte Carlo phase-offset weight estimation rather than the quotient construction. Entries are mean $\pm$ one standard deviation over three seeds.

Variant	$\|\widehat{g} - g^\star_\Pi\|_\infty$	$\|\widehat{v} - v^\star_\Pi\|_\infty$	$\|\widehat{J}_{40} - J_{40}\|_\infty$
Full learned	0.0446 ± 0.002	1.45 ± 0.13	1.44 ± 0.11
True structure + MC weights	0.040 ± 0.004	1.48 ± 0.15	1.44 ± 0.12
Learned structure + exact weights	$3.15 \times 10^{-11} \pm 9.6 \times 10^{-12}$	1.47 ± 0.11	1.43 ± 0.11
Oracle projection	$3.15 \times 10^{-11} \pm 9.6 \times 10^{-12}$	1.47 ± 0.11	1.43 ± 0.11

Table 5: Sparse storage proxy for direct quotient estimation versus dense plug-in model storage. Here $N = 5$ is the number of persistent coordinates. The dense plug-in proxy is n^2 entries, while the quotient proxy counts observed support plus nN basis-weight entries.

n	N	dense P entries	dense / quotient proxy
290	5	84,100	28.9×
1040	5	1,081,600	112.0×
1540	5	2,371,600	167.5×

learned structure and learned absorption weights. The *True structure + MC weights* row gives the estimator the true recurrent classes and periods but still estimates absorption weights by Monte Carlo. The *Learned structure + exact weights* row uses learned graph structure but exact absorption weights. The *Oracle projection* row uses the exact anchor projection.

The ablation identifies the dominant error source. Replacing learned structure by true structure changes persistent-profile error only slightly, from 0.0446 to 0.040. In contrast, replacing Monte Carlo absorption weights by exact weights drives persistent-profile error to numerical precision. Thus the quotient construction and structural recovery are not the bottleneck on this benchmark; the remaining profile error is the expected finite-sample error in phase-offset absorption probabilities. The transient and finite-horizon errors are similar across the exact-weight rows because, at this budget, they are dominated by stochastic approximation error in $\widehat{v}$.

We also tested a concrete support-structure perturbation. In an exhaustive single-edge deletion check over the benchmark support graph, removing one true support edge changed the recovered structural invariants in $0/582$ cases. Equivalently, the recurrent-class count, recurrent-set size, and periods were preserved for every single-edge deletion in this benchmark. This should not be read as a universal robustness theorem: the theoretical guarantee still requires recovering the correct support graph, and rare edges can be information-theoretically hard to detect. The result is instead a benchmark-level sanity check showing that the structural invariants used here are not fragile to every single missed edge.

H.5 Storage proxy for the plug-in comparison

A dense plug-in estimator stores an $n \times n$ transition matrix and then solves the decomposition from the learned model. Our estimator instead stores the observed support and the learned anchor-basis weights needed by the quotient evaluator. Table 5 reports a simple storage proxy for larger instances with the persistent dimension $N = 5$. The ratio is not a runtime theorem or a statistical optimality claim; it illustrates why the algorithm is structure-aware rather than full-model based.

The storage proxy complements the statistical comparison. On small tabular problems, plug-in can be very accurate and should be treated as a strong baseline. The proposed method is valuable for a different reason: it estimates the persistent coordinates, the transient representative, and the finite-horizon return diagnostic through the quotient structure, without making dense full-model estimation the central object.

H.6 Implementation details and reproducibility

All experiments use deterministic rewards and independent generative-model transition samples. Exact diagnostic checks are deterministic and use exact transition matrices. In the sample-based benchmark, each reported seed relearns the support structure, estimates the phase-offset absorption basis, runs projected stochastic approximation, and estimates anchor residuals. Ground truth is computed only for evaluation by first constructing the exact support graph, computing the exact phase-offset basis, solving the gauge-fixed linear system for $v^\star_\Pi$, and then forming $g^\star_\Pi = r + Pv^\star_\Pi - v^\star_\Pi$. The plug-in baseline estimates a dense empirical transition matrix from simulator samples and then applies the same exact decomposition routine to the empirical model. Thus plug-in is given the natural full-model route, while the proposed method follows the quotient estimator analyzed in the paper.

The main numerical comparison should be read in this light. The average-only comparator is deliberately favorable to classical average-reward evaluation because it is given the exact invariant gain; it still fails on $g^\star_\Pi$, $v^\star_\Pi$, and J_{40} because those are not invariant-gain objects. The dense plug-in comparator is deliberately strong on a small tabular problem and can have smaller persistent-profile error. This is not a contradiction of the paper's claim. The claim is that the peripheral quotient gives a diagnostic decomposition and a stable estimator for that decomposition, not that every structure-aware estimator must numerically dominate dense model estimation at every tabular sample budget.

H.7 Takeaway

The exact diagnostics validate the semantic claim: classical gain and bias can entangle persistent phase behavior with transient cost, while the proposed decomposition separates them. The sample-based benchmark validates the statistical claim: the estimator can learn the two decomposition components and thereby control finite-horizon return prediction. The plug-in comparison addresses the natural model-based alternative, and the ablations show that the quotient construction itself is not the source of the remaining profile error. Overall, the experiments support the paper's main message: the contribution is not merely a way to compute returns, but a way to attribute them to persistent regime behavior and transient-to-regime cost.

I Interpretation in cyclic systems

A finite model of a cyclic service system can use $X_t = (Z_t, \phi_t)$, where Z_t contains bounded queues and protocol variables and $\phi_t \in \{0, \dots, d-1\}$ is schedule phase, with $d \geq 2$. For phase-conditioned kernels Q_ϕ, the homogeneous transition matrix is

$$P\big((z,\phi),(z',\phi')\big) = Q_\phi(z,z')\mathbf{1}\{\phi' = \phi + 1 \pmod d\}. \tag{172}$$

Every return time is divisible by d, so stochastic operational transitions do not eliminate periodicity. Initial phase determines alignment with recurring service; the operational state can also produce relaxing initialization effects. Multiple closed configurations arise only when the fixed controller does not subsequently switch between them. A calibrated finite Markov abstraction and suitable sampling access are needed before applying our estimator to a real system.

For a simple example, let Z_t be cold or warm. A cold system becomes warm with probability $p \in (0,1]$ per step; warm states remain warm. Let $r(b,\phi) = a_\phi - c\mathbf{1}\{b = \text{cold}\}$, with $c > 0$, and use warm states as anchors. Direct substitution gives $g^\star_\Pi(b,\phi) = a_\phi$ and $v^\star_\Pi(b,\phi) = -(c/p)\mathbf{1}\{b = \text{cold}\}$. Every start has the same average reward, but the phase-aligned return difference is

$$J_T(\text{cold},\phi) - J_T(\text{warm},\phi) = -\frac{c}{p}\big(1 - (1-p)^T\big). \tag{173}$$

Thus a_ϕ records recurring service, while c/p records startup burden. Changing the initial phase changes timing rather than startup cost. This is an illustrative model, not a measured result for the cited systems.

Gauge choice matters outside this simple example. Let D be a common multiple of recurrent periods and $E = \lim_{m\to\infty} P^{mD}$, the commuting peripheral projection. The gauge $I - E$ gives $g_{\text{c}} = Er$ and $v_{\text{c}} = \sum_{t\geq 0} P^t(I-E)r$, with $P^t v_{\text{c}} \to 0$. A general anchor pair instead satisfies $v^\star_\Pi = \Pi v_{\text{c}}$ and

$g^\star_\Pi = g_c + (I - P)(I - \Pi)v_c$. Its representative need not decay under the original dynamics. For example, with

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{pmatrix}, \qquad r = (0, 2, 0)^\top,$$

anchors $1, 3$ give $v^\star_\Pi = (0, 2, 0)^\top$ and $g^\star_\Pi = (0, 0, 1)^\top$, whereas $Er = (1, 1, 0)^\top$; $P^t v^\star_\Pi$ oscillates. The anchor value nevertheless has a phase-aligned comparison interpretation:

$$v^\star_\Pi(s) = \lim_{T\to\infty} \left[J_T(s) - \sum_{(i,k)\in\mathcal{I}} b_{i,k}(s) J_T(a_{i,k}) \right]. \tag{174}$$

Indeed, applying Π to the return identity gives $\Pi J_T = v^\star_\Pi - \Pi P^T v^\star_\Pi \to v^\star_\Pi$ by quotient contraction. This separates anchor-relative initialization accounting from literal spectral decay without changing the estimator's target.

Finally, approximate periodicity does not imply exact persistence. For a deterministic cycle C, the perturbed chain $(1 - \eta)C + \eta I$, $0 < \eta < 1$, has all nonconstant eigenvalues strictly inside the unit disk. Its oscillations decay, potentially slowly. Our guarantees concern exact finite-chain structure, not uniform recovery of approximately persistent modes or deployment-scale models. Similarly, damping a classical expected TD update changes its error operator to $(1 - \alpha)I + \alpha P$, $0 < \alpha < 1$; every mode with $|\lambda| \le 1$ and $\lambda \neq 1$ then has $|1 - \alpha + \alpha\lambda| < 1$. Periodic phase modes therefore do not by themselves imply TD divergence.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper's approach have only been validated under tabular finite-state Markov chains. The finite-state setting has been constantly emphasized throughout this paper.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper's approach have only been validated under tabular finite-state Markov chains. The finite-state setting has been constantly emphasized throughout this paper. For the sample-based learning approach, the generative model is also explicitly assumed.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Although the primary contribution of this paper is theoretical, the paper's experiment section in the appendix covers sufficient details for reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This paper is primarily theoretical. The code and experiments only serve for numerical validations for the theoretical results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Although the primary contribution of this paper is theoretical, the paper's experiment section in the appendix covers sufficient details for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Although the primary contribution of this paper is theoretical, the paper's experiment section in the appendix covers sufficient details for reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Although the primary contribution of this paper is theoretical, the paper's experiment section in the appendix covers sufficient details for reproducibility. The experiments are simple numerical validations which can be ran on simple laptop devices.

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: This paper is primarily theoretical.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is primarily theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper is primarily theoretical.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: This paper is primarily theoretical.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper is primarily theoretical.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper is primarily theoretical.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper is primarily theoretical.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is mainly used for polishing the draft.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.