Recursive Interaction Systems: Non-Identifiability, Emergence, and Metastability

Claudio Irrgang

May 4, 2026

Abstract

We define Recursive Interaction Systems (RIS), a class of dynamical systems in which a high-dimensional latent state is projected through a non-invertible emission map and the emitted observation is recursively re-injected into the next-step dynamics. Under explicit assumption packages, we derive three structural consequences.

First, non-identifiability: if the emission map has nontrivial kernel and the initial distribution assigns positive mass to two distinct states in the same emission fiber, observable trajectories do not uniquely determine latent trajectories.

Second, typical-set separation: if the component predictors satisfy positive divergence rate and cross-entropy separation, the mixture concentrates on commitment chains to which each component assigns exponentially small probability.

Third, spectral metastability: under exact or approximate lumpability with slow/fast spectral separation, the induced kernel exhibits near-unit eigenvalues and slow mixing.

Independence: the three textbook sufficient conditions for typical-set separation ($\Delta_i > 0$, hard token partition, geometric basin separation) are each individually non-necessary; loop-driven divergence accumulation holds in concrete systems where each fails (Theorem 6.16, formally verified in Lean).

For the separation and metastability results, the δ-separated regime provides an explicit solvable instance (spectral gap $\Theta(\delta)$, positive divergence rate). We construct an explicit finite-state RIS exhibiting all three phenomena in one inspectable model. The contribution is identifying one operator class (recursive lossy feedback) in which all three phenomena arise from a single recursive mechanism. Empirical realization in trained systems is outside the scope of the present work.

Contents

1 Introduction **3**
- 1.1 The object class . . . 3
- 1.2 Scope . . . 4
- 1.3 Main results . . . 4
- 1.4 Scope and non-claims . . . 6
- 1.5 Paper organization . . . 6

2 Core Framework **7**
- 2.1 Notation . . . 7
- 2.2 Definitions . . . 7
- 2.3 Architecture-independent pipeline . . . 8
- 2.4 Training quotient . . . 8
- 2.5 Relation to irreversible compression . . . 8

3 Supporting Propositions **8**
3.1 I. Irreversibility and information loss . . . 9
3.2 II. Discretization and dynamics . . . 9
3.3 III. Mixture and emergence . . . 9

4 Operator Theory **10**
4.1 Markov operator formalization . . . 10
4.2 Spectral classification . . . 10
4.3 Typed separation parameters . . . 10
4.4 Metastability in the exact-lumpable regime . . . 11
4.5 Basin decomposition . . . 17
4.6 Slow-mode recovery and gap tracking . . . 17
4.7 Operator stability . . . 18
4.8 Non-normal transient growth . . . 18

5 Mixture and Emergence **19**
5.1 Emergence gap . . . 19
5.2 Exponential amplification under sequential generation . . . 20
5.3 Token-level containment and sequence-level separation . . . 20

6 Emergence Theorems **21**
6.1 Model class . . . 21
6.2 Positive divergence lemma . . . 22
6.3 Constructive emergence . . . 22
6.4 Typical-set separation . . . 23
6.5 Loop Closure independence . . . 27

7 Non-Identifiability **28**

8 A Concrete Finite-State RIS **31**
8.1 Model definition . . . 31
8.2 Non-identifiability . . . 32
8.3 Metastability and spectral structure . . . 32
8.4 Typical-set separation . . . 33
8.5 Summary . . . 33
8.6 Typed separation existence . . . 33

9 Discussion **33**
9.1 Nature of the contribution . . . 33
9.2 What the framework constrains . . . 34
9.3 Framework predictions . . . 34
9.4 Scope and limits . . . 34
9.5 Future directions . . . 35

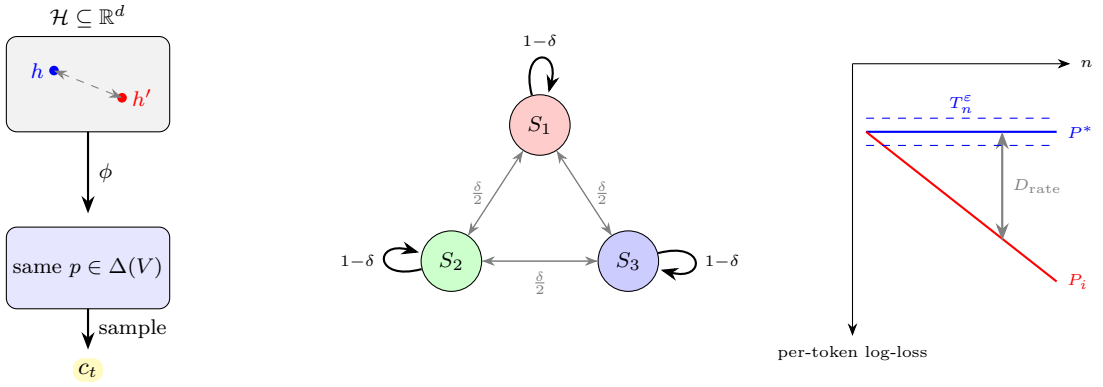
10 Conclusion **36**

A Symbolic Verification **37**

B Lean Formalization Artifacts **38**

1 Introduction

A recurring structure in sequential systems is the composition of lossy projection with recursive re-injection: a high-dimensional internal state is reduced to a finite committed observation, and that observation becomes part of the input to the next step. This composition is not a one-time compression but an iterated commitment process. Under additional structural assumptions, its recursive application can produce trajectory-level effects not visible at a single step, yielding non-identifiability of latent trajectories, typical-set separation, and spectral metastability. This paper defines an abstract operator class, establishes three structural results within it under explicit assumptions, and constructs a concrete finite-state instance exhibiting all three phenomena.



(a) Emission **(b) Lumped chain** **(c) Separation**

Figure 1: Three views of one system. A 3-domain δ-separated mixture exhibits three analytic levels of the RIS framework. **(a)** Non-identifiability (Theorem 7.1): distinct latent states h, h' in the same emission fiber induce the same emission distribution (one-step non-injectivity; multi-step persistence is handled separately in Remark 7.2). **(b)** Metastability (Theorem 4.3): the lumped B=3 chain has self-loop probability $1-\delta$ and cross-domain leakage $\delta/2$ per neighbor ($B-1=2$ other domains). **(c)** Typical-set separation (Theorem 6.11): per-token log-loss under P_i exceeds that under P^* by $D_{\text{rate}}(P^* \| P_i)$ asymptotically.

1.1 The object class

We define *Recursive Interaction Systems* (RIS), a class of dynamical systems with the following structure.

Definition 1.1 (Recursive Interaction System)**.** A *Recursive Interaction System* consists of:

(i) a latent state space $\mathcal{H} \subseteq \mathbb{R}^d$;

(ii) a finite observation alphabet V;

(iii) a non-invertible projection $\phi\colon \mathcal{H} \to \Delta(V)$;

(iv) a re-injection map $f\colon \mathcal{H} \times V \to \mathcal{H}$.

The recursive operator is the composition

$$H_{t+1} = f(H_t,\, C_t), \qquad C_t \sim \phi(H_t), \tag{1}$$

where $C_t \in V$ is the committed observation (a sampled token, not the distribution $\phi(H_t)$ from which it is drawn) at step t. The projection ϕ is information-losing by assumption. In the softmax-linear case $\phi(h) = \mathrm{softmax}(Wh)$, $\dim\ker(W) > 0$ whenever $d > |V| - 1$, provided W has full rank on the quotient space (i.e., rank $|V| - 1$).

Remark 1.2 (Architecture-agnostic illustration). The definition is architecture-agnostic: it requires only a latent state, a lossy emission, and recursive re-injection of the emitted observation. As an illustration, a residual-stream state is projected through a linear map and softmax to produce a token that is re-embedded into the next forward pass.

Remark 1.3 (Transformer state vs. sliding-window state). In a full autoregressive transformer the latent state is recomputed from the entire token history via attention at each step, and the full key-value cache is not the same object as a sliding window V^k; the finite-state Markov structure used in Section 4 is an approximation in state representation, not only in context length. None of the theorems depends on transformer-specific mechanisms (see Section 1.4 for the precise boundary).

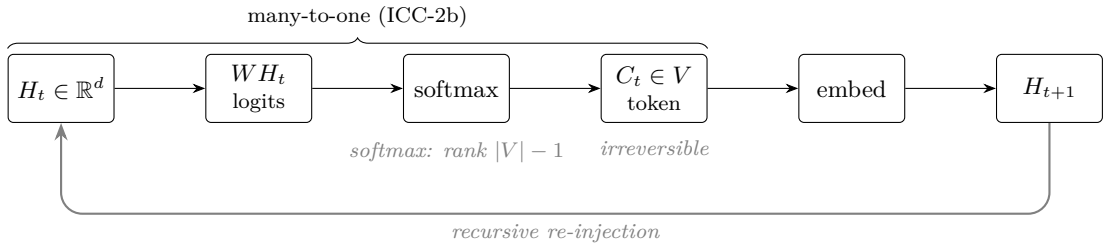


Figure 2: The RIS feedback architecture. H_t is projected through softmax (rank $|V|-1$, many-to-one) to a committed token C_t, which is re-embedded into the next latent state. The feedback loop defines the recursive operator.

1.2 Scope

This paper treats a single RIS with dynamics $H_{t+1} = f(H_t, C_t)$. On a finite context state space V^k, the recursion induces a Markov operator whose spectrum classifies the dynamics as ergodic, periodic, or metastable (Section 4). Coupling of multiple RIS through exchanged commitments is outside the scope of the present paper.

The δ-separation condition used in the typical-set separation and metastability results is a sufficient structural regime chosen because it yields clean exponents and explicit spectral structure. It is not a claim that real recursive generative systems generally satisfy hard domain separation; weaker overlap-based conditions may suffice for analogous but less sharp conclusions.

RIS is distinct from adjacent frameworks. Hidden Markov models treat the observation as a noisy readout of a hidden state without committing to it; partially observable Markov decision processes add control but retain the observation-as-readout structure. RIS adds emission-induced commitment: the observation is not a noisy view but an irreversible projection that the next state is constructed atop. This structural addition is what couples the three phenomena (Theorem 7.1, Theorem 6.11, Theorem 4.3) to a single recursive mechanism, rather than producing them in distinct mathematical classes.

1.3 Main results

Theorem 1.4 (RIS theorem packages and coexistence structure)**.** *Let $(\mathcal{H}, V, \phi, f)$ be a Recursive Interaction System (Definition 1.1). The following hold under the assumption packages below:*

(I) ***Non-identifiability.*** *If $\phi(h) = \mathrm{softmax}(Wh)$ with $\mathrm{rank}(W) = |V|-1$, $\dim(\mathcal{H}) > |V|-1$, and the initial distribution assigns positive mass to two distinct states in the same emission fiber, then the latent-to-observation relation is non-identifiable: there exist distinct latent trajectories compatible with the same observed token history with positive probability. The theorem establishes non-identifiability at the initial step; persistence across multiple steps*

additionally requires transition regularity as in Remark 7.2. (Proved as Theorem 7.1; no finite-state or separation hypotheses required.)

Takeaway: you cannot recover the hidden trajectory from the token output.

(II) ***Typical-set separation.*** *If additionally the component predictors satisfy positive divergence rate* $D_{\text{rate}}(P^*\|P_i) > 0$ *and cross-entropy separation* $\Delta_i := H_{\text{rate}}(P_i, P^*) - h(P^*) > 0$*, then the mixture typical set carries exponentially vanishing probability under each component, with good-subset exponent* D_{rate}*. The* δ*-separated regime (Definition 6.1) is one sufficient regime establishing these conditions (Section 6). Typical-set separation requires both positive divergence rate and cross-entropy separation; neither alone suffices.* (Proved as Theorem 6.11.)

Takeaway: the mixture can concentrate on sequences to which each single component assigns exponentially small probability.

(III) ***Spectral metastability.*** *Under exact lumpability in the* δ*-separated regime, the induced kernel has* B *near-unit eigenvalues with spectral gap* $\Theta(\delta)$*. Under approximate lumpability with spectral separation between slow and fast sectors, the* B*-dimensional slow subspace persists with projector error* $O(B\varepsilon)$ *relative to the macro spectrum.* (Proved as Theorems 4.3 and 4.10.)

Takeaway: the system gets stuck in coarse regimes for long stretches before switching.

(IV) ***Independence of sufficient conditions.*** *The three textbook sufficient conditions for typical-set separation in (II) (*$\Delta_i > 0$*, hard token partition, geometric basin separation) are each individually non-necessary: there exist concrete finite RIS in which any one of them fails while loop-driven divergence accumulation holds.* (Proved as Theorem 6.16; three countermodels formally verified in Lean.)

Takeaway: recursive feedback is sufficient for divergence accumulation on its own; no single textbook condition is required.

The concrete RIS of Section 8 satisfies all three conditions simultaneously: six latent states in three basins of two, with observation alphabet $\{a, b, c\}$*, a token-dependent transition kernel, and emission distributions identical within each basin (*$\phi(1) = \phi(2)$*, etc.). The* 3×3 *coarse kernel has eigenvalues* 1 *and* $1 - \frac{3}{2}\delta$ *(spectral gap* $\frac{3}{2}\delta$*), the six-state kernel is exactly lumpable, and all three phenomena are verified analytically and numerically.*

The coexistence example is existential, not generic: it shows that non-identifiability, typical-set separation, and spectral metastability can arise jointly within one explicit RIS under one coherent assumption package. The contribution is an operator class in which non-identifiability, typical-set separation, and metastability arise from one recursive mechanism and are provable under explicit, individually checkable hypotheses. The δ-separated regime is one sufficient regime that yields clean exponents and explicit spectral structure; it is not claimed to be necessary.

A further refinement distinguishes predictive, dynamical, and geometric notions of separation (Definition 4.2), which coincide in aligned special cases but can diverge under overlapping emissions.

Figure 3 summarizes sufficient assumptions, not a frequency claim about naturally occurring systems.

Supporting propositions collect standard facts and auxiliary consequences used in the RIS setting (Sections 3 and 4).

Table 1 records assumption dependencies for the theorem statements only; it should not be read as an empirical checklist for real architectures.

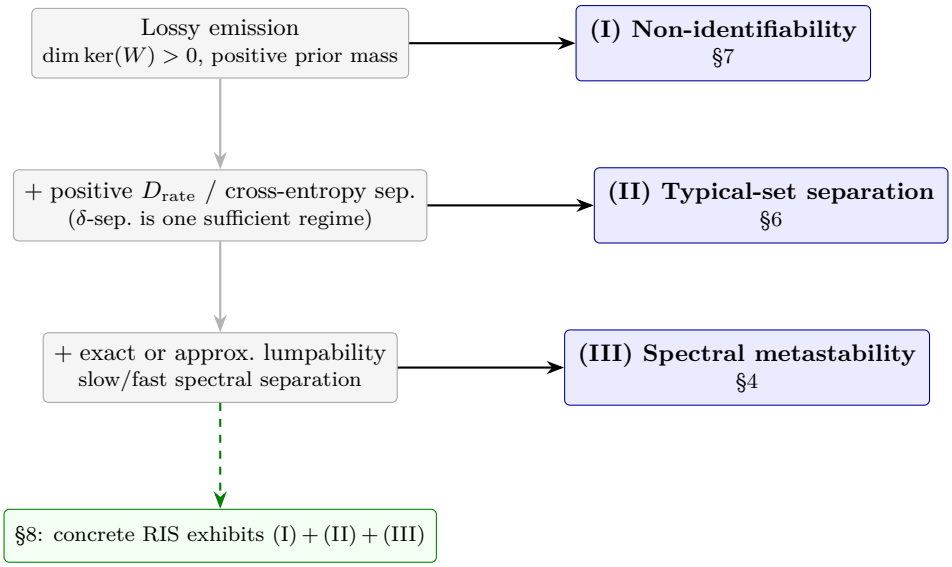


Figure 3: The RIS theorem packages (Theorem 1.4). Each level adds an assumption and yields a structural consequence (right). The concrete RIS of Section 8 satisfies all three levels simultaneously.

1.4 Scope and non-claims

This paper makes structural, theorem-level claims only. Its statements are conditional: once the listed assumptions hold, the conclusions follow independently of architecture-specific implementation. The paper does not assert that any particular trained architecture satisfies the full RIS assumption package. For transformers, RIS is an analogy at the level of lossy committed emission plus recursive token feedback (Remark 1.2), not a proof that the effective latent dynamics are Markov on a fixed finite-dimensional state (Remark 1.3). Outside scope are empirical behavior of deployed systems, mechanistic identity between biological and artificial instantiations, and claims about consciousness or subjective experience. Toy-model and symbolic verifications are illustrative only and do not strengthen the proofs.

1.5 Paper organization

Note: expository order differs from logical strength. Non-identifiability (the weakest-assumption result) appears in Section 7 rather than earlier for expository flow: the partition and mixture machinery of Sections 5–6 motivates the question of what remains unobservable.

- Sections 2–4 establish the operator-theoretic backbone (framework, propositions, spectral theory).
- Sections 5–6 develop mixture separation and its typical-set consequences.
- Section 7 proves non-identifiability.
- Section 8 constructs an explicit finite-state RIS exhibiting all three phenomena.
- Section 9 discusses scope and future directions.

The paper can be read along two tracks: *Theorem-first*, Sections 1, 4, 6, 7, and 8 (main results and coexistence model), referring to Sections 2–3 as needed for definitions and supporting facts; *Sequential*, Sections 1–8 in order, where Sections 2–3 build the operator-theoretic and mixture-theoretic vocabulary used throughout.

Table 1: Assumption dependencies of the main results. A checkmark (✓) indicates the result uses this hypothesis; a tilde (∼) indicates an approximate version suffices; a dash (—) indicates no dependence.

Result	Finite state	δ-sep.	$D_{\text{rate}} > 0$	$\Delta_i > 0$	Exact lump.	Conclusion
Theorem 7.1	—	—	—	—	—	Latent non-identifiability
Theorem 6.11	✓	∼*	✓	✓	—	Exponential typical-set separation
Proposition 6.5	✓	—	✓	—	—	$D_{\text{rate}} > 0$ without δ-sep.
Theorem 4.3	✓	✓	—	—	✓	Exact slow-sector spectrum
Theorem 4.10	✓	∼	—	—	∼	Slow-sector persistence
Proposition 4.13†	✓	—	—	—	—	Pred. coherence → approx. lump.
Remark 4.16‡	✓	—	—	—	—	Metastability from δ_{ker} alone

*δ-separation is one sufficient regime for establishing positive D_{rate} and cross-entropy separation; it is not a direct hypothesis of Theorem 6.11.

†Requires predictive coherence and token-block alignment (hypotheses 1–2 of Proposition 4.13); these are joint conditions not decomposable into the column headers above.

‡Requires small stationary cross-block flow δ_{ker} and reversibility (Remark 4.16); no lumpability needed. Bound is gap $\leq O(\delta_{\text{ker}})$ (upper bound only; lower bound is $\Phi^2/2 \geq \delta_{\text{ker}}^2/2$, quadratically loose).

2 Core Framework

2.1 Notation

Symbol	Meaning
$\mathcal{H} \subseteq \mathbb{R}^d$	Latent state space
V	Finite observation alphabet
$X = V^k$	Context state space (sliding window; Section 4)
$\phi\colon \mathcal{H} \to \Delta(V)$	Emission map (non-invertible projection)
$f\colon \mathcal{H} \times V \to \mathcal{H}$	Re-injection map
P_i	Component predictor ($i = 1, \ldots, B$)
$P^* = \sum_i \pi_i P_i$	Mixture predictor
π_i	Mixture weight of component i
K	Markov transition kernel on X
K_L	Lumped kernel on domain-level chain
$h(\cdot)$	Entropy rate
$D_{\text{rate}}(\cdot\|\cdot)$	Relative entropy rate
$H_{\text{rate}}(P, Q)$	Cross-entropy rate: $h(P) + D_{\text{rate}}(P\|Q)$
$T_n^\varepsilon(P^*)$	P^*-typical set at length n, tolerance ε
δ	Separation parameter
$\delta_{\text{pred}}, \delta_{\text{ker}}, \delta_{\text{geom}}$	Typed separation (Definition 4.2)

2.2 Definitions

Definition 2.1 (Commitment)**.** A *commitment* is an element $c \in V$ drawn from the distribution $\phi(H_t)$ and appended to the observation history. (The term is used in the information-theoretic sense of an irreversible output, not in the behavioral-economics sense of a binding constraint on future choices.) Once emitted, c is fixed in the observation history. The updated history is $C_{0:t+1} = C_{0:t} \cdot c_{t+1}$.

Definition 2.2 (Commitment chain)**.** A *commitment chain* of length n is a sequence $C_n = (c_1, c_2, \ldots, c_n)$ where each c_k is drawn from $\phi(H_{k-1})$ and $H_k = f(H_{k-1}, c_k)$. The history grows

monotonically by appending one token at each step. C_n determines C_m for $m < n$ by truncation, but C_m does not determine C_n.

2.3 Architecture-independent pipeline

The generation of a commitment chain factors into four stages:

$$\mathcal{S} \xrightarrow{T} \theta^* \xrightarrow{G} K_{\theta^*} \xrightarrow{\text{iterate}} \mu_t \xrightarrow{\text{commit}} c_t. \tag{2}$$

Here $\mathcal{S}$ is a source distribution on training data, $\theta^* \in \Theta$ is a finite parameter object, K_{θ^*} is a Markov kernel on the observation state space X, $\mu_t \in \Delta(X)$, and $c_t \in V$.

T: training map. T sends a source distribution $\mathcal{S}$ to a finite parameter object θ^* by loss minimization: $\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathcal{S}}[\ell_\theta]$.

G: kernel induction. $\theta^* \mapsto K_{\theta^*}$, the next-commitment transition kernel induced by the trained parameters.

Iterate. $\mu_{t+1} = K_{\theta^*}\, \mu_t$ (Markov evolution on the observation state space).

Commit. $c_t \sim \mu_t$ (irreversible sampling from the predictive distribution).

Each arrow is non-invertible or non-expansive, and typically compressive in the information-theoretic sense. Training selects θ^*; recursion applies K_{θ^*} with θ^* fixed. These are distinct operations at different levels.

2.4 Training quotient

This subsection provides conceptual framing only and is not used in the proofs of Theorem 4.3, Theorem 6.11, or Theorem 7.1.

Definition 2.3 (Training equivalence)**.** Two source distributions $\mathcal{D}_1$, $\mathcal{D}_2$ are *training-equivalent* under loss ℓ and hypothesis class Θ if

$$\arg\min_{\theta \in \Theta} \mathbb{E}_{\mathcal{D}_1}[\ell_\theta] \;=\; \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathcal{D}_2}[\ell_\theta].$$

The training map T factors through equivalence classes: $[\mathcal{D}]_{\ell,\Theta} \mapsto \theta^*$. Distinctions in the source that do not change expected loss are not preserved in θ^*. This is the formal content of the observation that training compresses a broad source distribution into a finite parameter object.

2.5 Relation to irreversible compression

RIS belongs to the broader family of many-to-one emission architectures formalized in the ICC framework [9]; no ICC-specific machinery is required for the spectral or mixture-separation results proved here.

3 Supporting Propositions

This section collects standard facts used repeatedly later; readers may skim it on a first pass and return as needed. The results are organized in three groups.

3.1 I. Irreversibility and information loss

Proposition 3.1 (Softmax quotient)**.** *The softmax map* $\mathrm{softmax}\colon \mathbb{R}^V \to \Delta(V)$ *satisfies* $\mathrm{softmax}(z) = \mathrm{softmax}(z+\alpha\mathbf{1})$ *for all* $\alpha \in \mathbb{R}$*. At interior points* $p \in \mathrm{int}(\Delta(V))$*, its Jacobian is* $J = \mathrm{diag}(p) - pp^\top$ *with* $\mathrm{rank}(J) = |V| - 1$ *and null direction* $\mathbf{1}$*.*

Proposition 3.2 (Fiber characterization)**.** *For* p *in the interior of* $\Delta(V)$*, the full preimage under softmax is the affine line*

$$\mathrm{softmax}^{-1}(p) = \{\log p + \alpha\mathbf{1} : \alpha \in \mathbb{R}\}.$$

Proposition 3.3 (TV contraction)**.** *For any stochastic kernel* K *and distributions* μ*,* ν*:*

$$\|\mu K - \nu K\|_{\mathrm{TV}} \le \|\mu - \nu\|_{\mathrm{TV}}.$$

Every stochastic stage is non-expansive in total variation. Equivalently, the Dobrushin coefficient [12] satisfies $\eta(K) \le 1$*.*

Corollary 3.4 (Composition contracts)**.** *For a composition of stochastic kernels* K_1, K_2*:*

$$\|\mu K_1 K_2 - \nu K_1 K_2\|_{\mathrm{TV}} \le \|\mu K_1 - \nu K_1\|_{\mathrm{TV}} \le \|\mu - \nu\|_{\mathrm{TV}}.$$

Distinguishability is non-increasing and typically strictly decreasing through a chain of stochastic maps.

3.2 II. Discretization and dynamics

Proposition 3.5 (Finite-state periodicity)**.** *If the commitment state space is finite and the update rule is deterministic, then every trajectory is eventually periodic. Open-ended commitment chains require at least one of: stochasticity, context growth, or external input.*

Proposition 3.6 (Argmax switching boundaries)**.** *The* argmax *map partitions logit space* $\mathbb{R}^V$ *into polyhedral regions separated by* $\binom{|V|}{2}$ *pairwise equality hyperplanes* $\{z_i = z_j\}$*. Crossing a boundary induces a discrete change in the committed token. The softmax map is smooth; the discretization step converts smooth internal variation into piecewise-constant commitment dynamics.*

Proposition 3.7 (Self-conditioning contraction)**.** *For the linearized self-conditioning update* $A_{t+1} = \lambda A_t + (1-\lambda)GA_t$ *with* $\lambda \in (0,1)$ *and* G *having operator norm* $\|G\| = L_g$*, a sufficient condition for contraction of the deterministic linearization is*

$$\lambda + (1-\lambda)L_g < 1.$$

Remark 3.8. The condition is sufficient but not necessary for the full stochastic system. Additive noise can stabilize borderline regimes where the deterministic contraction factor equals or slightly exceeds 1; this is observed in toy models but is not proved here.

3.3 III. Mixture and emergence

The propositions in this group are stated and proved in Section 5. For reference:

- Proposition 5.1: JSD convexity gap ($H(P^*) - \sum_i \pi_i H(P_i) = \mathrm{JSD} \ge 0$).
- Proposition 5.3: sequential amplification (exponential ratio $(1/(2\sqrt{\varepsilon(1-\varepsilon)}))^n$).
- Proposition 5.6: token-level support containment, sequence-level amplification.

4 Operator Theory

This section identifies the dynamical signature of recursive coarse structure: near-invariant partitions produce slow modes, and exact lumpability yields an explicit solvable metastable spectrum.

Dependency chain. The section proceeds in three stages: (1) exact lumpability gives a coarse kernel with computable slow eigenvalues (Theorem 4.3 and Corollary 4.4); (2) approximate lumpability preserves the slow subspace via projector stability (Theorem 4.10); (3) predictive coherence provides one explicit route from separation conditions to approximate lumpability (Proposition 4.13). Readers interested only in the exact-lumpable case may focus on stage (1).

4.1 Markov operator formalization

Let $\mathcal{X} = V^k$ be the finite context space (sliding window of k tokens; this is an approximation of state representation for attention-based architectures, see Remark 1.3). The next-token law $P(\cdot \mid x)$ induces a transition kernel

$$K(y, x) = \sum_{c \in V} \mathbf{1}\{y = S(x, c)\}\, P(c \mid x), \tag{3}$$

where S shifts the window and appends the new token. The forward operator on probability measures is

$$\mu_{t+1} = K\mu_t. \tag{4}$$

Here K is column-stochastic: columns sum to 1. The lumpability condition is that $\sum_{y \in S_j} K(y, x)$ depends only on which block contains x, not on x itself. (This is the column-stochastic analogue of the "constant row sums" condition in Kemeny–Snell [11].)

4.2 Spectral classification

The qualitative behavior of the recursion is determined by the spectrum of K.

Remark 4.1 (Spectral criterion). Let the eigenvalues of K be ordered by modulus: $1 = |\lambda_1| \geq |\lambda_2| \geq \cdots$.

- *Ergodic (fast mixing):* $|\lambda_2| \ll 1$. Unique stationary measure π; $K^n\mu \to \pi$ for every initial μ.
- *Periodic:* some $\lambda_j \neq 1$ with $|\lambda_j| = 1$. $K^n\mu$ cycles; mass oscillates among residue classes.
- *Metastable:* $|\lambda_2| \approx 1$, all $|\lambda_j| < 1$ for $j \geq 2$. Globally ergodic, but mixing time is $\Theta(1/\text{gap})$. The chain spends long stretches in almost-invariant basins before transitioning.

4.3 Typed separation parameters

The single δ in Definition 6.1 conflates three distinct structural roles: how different the component predictors are, how rarely the dynamics cross between basins, and how well the emissions separate basins geometrically. Under exact lumpability with disjoint preference sets, one parameter controls all three; under overlapping emission supports, they can diverge. The following definition makes the distinction explicit.

Definition 4.2 (Typed separation)**.** For a partition $\mathcal{X} = S_1 \sqcup \cdots \sqcup S_B$ with associated B experts:

(i) *Predictive separation* δ_{pred}: $\min_{i \neq j} \mathbb{E}_{\pi^*}[\text{TV}(P_i(\cdot \mid x), P_j(\cdot \mid x))]$, the minimum pairwise mean total variation between component predictors, averaged over the stationary measure π^* of the mixture, which is unique when all transitions are strictly positive.

(ii) *Dynamical separation* δ_{ker}: the cross-basin mass $\max_i \sum_{j \neq i} K_L(i,j)$ in the lumped kernel, controlling the spectral gap and mean escape times.

(iii) *Geometric separation* δ_{geom}: $\min_{x \in S_i,\, y \in S_j,\, i \neq j} \text{TV}(\phi(x),\, \phi(y))$, the minimum emission-distribution distance between distinct basins.

Under exact lumpability with disjoint preference sets $S_i \cap S_j = \varnothing$, a single δ can control all three simultaneously. Proposition 8.1 constructs a finite-state RIS in which δ_{pred}, δ_{ker}, and δ_{geom} are pairwise distinct.

In short: δ_{pred} measures predictor separation, δ_{ker} measures regime-switching rarity, and δ_{geom} measures geometric emission separation.

4.4 Metastability in the exact-lumpable regime

Theorem 4.3 (Metastability in the exact-lumpable δ-separation regime)**.** *Let* $\{P_1, \ldots, P_B\}$ *be* δ*-separated experts with context-independent transition probabilities and equal domain sizes, and* K *the induced symmetric exact-lumpable mixture kernel. (For a token-conditioned kernel* $K(x'|x,c)$*, the spectral results apply to the marginalized kernel* $\bar{K}(x'|x) = \sum_c \phi(x)(c)\, K(x'|x,c)$*.)*

(i) *At* $\delta = 0$*,* K *is block-diagonal with* B *eigenvalues equal to* 1*.*

(ii) *For* $\delta > 0$*, the* B*-fold eigenvalue* 1 *splits:* $\lambda_1 = 1$ *(Perron root [17]),* $\lambda_j = 1 - O(\delta)$ *for* $j = 2, \ldots, B$*.*

(iii) *The spectral gap satisfies* $\text{gap} = \Theta(\delta)$*; in the symmetric reversible case,* $\text{gap} = B\delta/(B-1)$ *(Corollary 4.4) and the mixing time is* $\Theta(1/\delta)$*.*

In the symmetric $B = 2$ *case:* $\lambda_2 = 1 - 2\delta$*,* $\text{gap} = 2\delta$*.*

Proof. Part (i): at $\delta = 0$, each expert has disjoint support, so K decomposes into independent blocks, each with its own Perron eigenvalue 1.

Part (ii): at $\delta = 0$, the eigenvalue 1 has geometric multiplicity B (one per block). The $B = 2$ case is computed exactly: $\lambda_2 = 1 - 2\delta$. For general B, define the domain-level chain by projecting onto preference-set membership: $Y_t = j$ iff $C_t \in S_j$. Under δ-separation with context-independent transition probabilities, the total mass from any state in S_i to the set S_j depends only on i and j, not on which state within S_i is occupied. This is the Kemeny–Snell lumpability condition [11]: the projected process Y_t is itself a Markov chain. Intuitively, the coarse dynamics are well-defined because all states within a block transition to other blocks at the same rates. Its $B \times B$ kernel has diagonal entries $1 - \delta$ and off-diagonal entries $\delta/(B-1)$. This is a symmetric B-basin kernel $(1-\delta)I_B + \frac{\delta}{B-1}(J_B - I_B)$, whose eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = \cdots = \lambda_B = 1 - B\delta/(B-1)$ (direct computation from the rank-one update).

Under context-independent transition probabilities, the block-constant subspace U is K-invariant (Kemeny–Snell lumpability [11]): the eigenvalues of the lumped kernel K_L are a subset of the eigenvalues of K. In the symmetric construction (equal block sizes, uniform within-domain transitions), K is doubly stochastic and reversible with respect to its stationary measure, so $U^\perp$ is also K-invariant and the full spectral decomposition $\sigma(K) = \sigma(K_L) \cup \sigma(K_{\text{within}})$ holds exactly. For non-reversible exactly lumpable kernels, $\sigma(K_L) \subseteq \sigma(K)$ and $\rho(K_{\text{within}}) \leq 1 - \delta$ (by Lemma 4.5), which is sufficient to establish the B near-unit eigenvalues and spectral gap $\Theta(\delta)$ claimed in parts (ii)–(iii).

Part (iii): for the symmetric $B = 2$ case, the domain-level kernel is $K_2 = \left(\begin{smallmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{smallmatrix}\right)$ with eigenvalues 1 and $1 - 2\delta$. The general case follows from the exact spectrum of the $B \times B$ lumped kernel derived above. □

Metastability in RIS is therefore not metaphorical: in the exact-lumpable separated regime it is an explicit spectral fact.

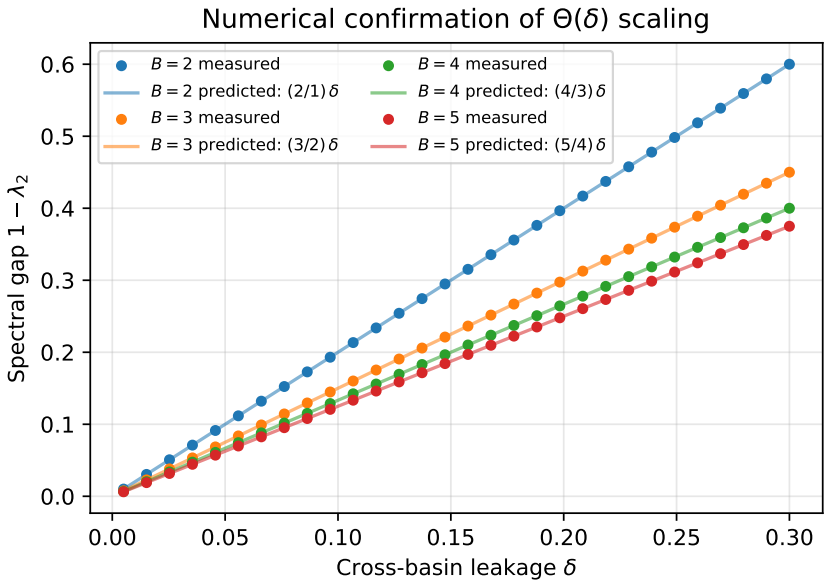


Figure 4: Numerical confirmation of the spectral gap scaling predicted by Theorem 4.3. Symmetric B-basin lumped kernels with cross-basin leakage δ exhibit spectral gap $1 - \lambda_2 = \frac{B}{B-1}\delta$ (solid lines). Computed eigenvalues (markers) match the prediction for $B \in \{2, 3, 4, 5\}$ across $\delta \in [0.005, 0.30]$. Residual deviations are at machine-precision level ($\leq 3 \times 10^{-16}$), confirming the analytic formula exactly. This is a finite-instance check of the framework's quantitative prediction; the theorem itself is established analytically in the proof above.

Corollary 4.4 (Exact slow-sector spectrum)**.** *In the symmetric exact-lumpable regime of Theorem 4.3, the full spectrum of K decomposes as follows.*

(i) Slow sector *(from the lumped chain K_L):*

$$\lambda_1 = 1, \qquad \lambda_2 = \cdots = \lambda_B = 1 - \frac{B}{B-1}\delta \quad (\textit{multiplicity } B-1).$$

(ii) Fast sector *(within-domain modes): every remaining eigenvalue satisfies $|\lambda| \leq 1 - \delta$ (Lemma 4.5).*

The spectral gap is $\text{gap} = \frac{B}{B-1}\delta$*, and the mixing time is* $\Theta\left(\frac{B-1}{B\delta}\right)$*. The slow sector is therefore not merely descriptive but spectrally isolated and quantitatively controlled. In asymmetric kernels (unequal domain sizes or non-uniform within-domain transitions), the $B-1$ slow eigenvalues are generically distinct; the degeneracy $\lambda_2 = \cdots = \lambda_B$ is a consequence of the symmetric construction, not a general prediction.*

Proof. The slow eigenvalues are those of the $B \times B$ lumped kernel (diagonal entries $1 - \delta$, off-diagonal entries $\delta/(B-1)$), computed exactly in the proof of Theorem 4.3. The fast eigenvalues are bounded by Lemma 4.5. By exact lumpability, $\sigma(K) = \sigma(K_L) \cup \sigma(K_{\text{within}})$, so the two sectors do not interact. □

The exact lumpability used above is a restrictive structural assumption and should not be interpreted as generic: it relies on context-independent transition probabilities within each preference set. Without this, only approximate lumpability (Theorem 4.10) can be expected.

Lemma 4.5 (Within-domain spectral bound)**.** *For the δ-separated mixture kernel restricted to the context-state block induced by preference set S_i (rows and columns corresponding to states whose current token lies in S_i), every eigenvalue λ of the restricted block satisfies $|\lambda| \leq 1 - \delta$.*

Proof. The within-domain block $K|_{S_i \times S_i}$ has column sums $\sum_{y \in S_i} K(y,x) = 1 - \delta$ for every $x \in S_i$ (the remaining mass δ leaks to other domains). The ℓ^1 operator norm of the restricted block is therefore $\|K|_{S_i}\|_1 = \max_{x \in S_i} \sum_{y \in S_i} K(y,x) = 1 - \delta$, and since $\rho(A) \le \|A\|_1$ for any matrix A, we obtain $\rho(K|_{S_i \times S_i}) \le 1 - \delta < 1$. □

Proposition 4.6 (Perturbative persistence of the slow sector)**.** *Let K_0 be an exactly lumpable stochastic kernel with B slow eigenvalues (from the lumped chain K_L) and remaining within-domain eigenvalues (from K_{within}). Suppose the slow spectral cluster is separated from the within-domain spectrum by a gap $\gamma > 0$:*

$$\min_{\lambda \in \sigma(K_L),\, \mu \in \sigma(K_{\text{within}})} |\lambda - \mu| \ge \gamma.$$

Then for any stochastic kernel K with $\|K - K_0\| \le \varepsilon < \gamma/2$, the kernel K has exactly B eigenvalues within distance $\kappa\varepsilon$ of the slow eigenvalues of K_0, where κ is the condition number of the eigenbasis of K_0.

In the symmetric exact-lumpable regime of Theorem 4.3 with uniform within-domain transitions, K_0 is doubly stochastic and normal, the within-domain eigenvalues are 0, and the gap is $\gamma = 1 - B\delta/(B-1)$.

Proof. Under exact lumpability, K_0 is block-diagonalizable with respect to the lumped and within-domain subspaces. Standard eigenvalue perturbation theory for matrices with spectrally isolated clusters (Bauer–Fike applied to the diagonalizable case, or invariant-subspace perturbation for the general case) gives the stated bound when $\varepsilon < \gamma/2$. For normal K_0 (including the symmetric exact-lumpable case), $\kappa = 1$. □

The next result bounds how much the slow subspace of K can deviate from the block-constant subspace when lumpability holds only approximately.

Proposition 4.7 (Projector stability from approximate invariance)**.** *Let $X = X_1 \sqcup \cdots \sqcup X_B$ be a partition of a finite state space, and let Π be the block-averaging projection onto the subspace $U := \{f : X \to \mathbb{C} : f \text{ is constant on each block } X_i\}$. Set $Q := I - \Pi$ and decompose K on $U \oplus U^\perp$ as*

$$K = \begin{pmatrix} A & \mathcal{B} \\ \mathcal{C} & D \end{pmatrix}, \qquad A = \Pi K \Pi,\ \mathcal{B} = \Pi K Q,\ \mathcal{C} = Q K \Pi,\ D = Q K Q.$$

(The off-diagonal block operators are denoted calligraphically ($\mathcal{B}$, $\mathcal{C}$) to avoid collision with the basin count B.) Define the reference kernel $K_0 := \left(\begin{smallmatrix} A & \mathcal{B} \\ 0 & D \end{smallmatrix}\right)$, so that U is K_0-invariant. All norms below are the operator norm induced by $\|\cdot\|_\infty$ on $\mathbb{R}^{|X|}$. Assume there exists a contour Γ enclosing $\sigma(A)$ and disjoint from $\sigma(D)$, with

$$M_\Gamma := \sup_{z \in \Gamma} \|(zI - K_0)^{-1}\| < \infty, \qquad M_\Gamma \|\mathcal{C}\| < 1.$$

Then K has a $\dim(U)$-dimensional invariant subspace associated with the part of $\sigma(K)$ inside Γ, and its Riesz projector

$$P := \frac{1}{2\pi i} \oint_\Gamma (zI - K)^{-1}\, dz$$

satisfies

$$\|P - \Pi\| \ \le\ \frac{\operatorname{len}(\Gamma)}{2\pi} \frac{M_\Gamma^2 \|\mathcal{C}\|}{1 - M_\Gamma \|\mathcal{C}\|}. \tag{5}$$

In particular, $\|P - \Pi\| = O(\|\mathcal{C}\|)$ as $\|\mathcal{C}\| \to 0$.

Proof. The resolvent identity gives

$$(zI - K)^{-1} - (zI - K_0)^{-1} = (zI - K)^{-1}(K - K_0)(zI - K_0)^{-1},$$

with $K - K_0 = \left(\begin{smallmatrix} 0 & 0 \\ \mathcal{C} & 0 \end{smallmatrix}\right)$, so $\|K - K_0\| = \|\mathcal{C}\|$. For $z \in \Gamma$, the factorization $zI - K = (zI - K_0)[I - (zI - K_0)^{-1}(K - K_0)]$ together with $M_\Gamma \|\mathcal{C}\| < 1$ and a Neumann series gives

$$\|(zI - K)^{-1}\| \le \frac{M_\Gamma}{1 - M_\Gamma \|\mathcal{C}\|}.$$

Hence

$$\|(zI - K)^{-1} - (zI - K_0)^{-1}\| \le \frac{M_\Gamma^2 \, \|\mathcal{C}\|}{1 - M_\Gamma \|\mathcal{C}\|}.$$

Since U is K_0-invariant, the Riesz projector of K_0 for the spectral cluster inside Γ is exactly Π. Integrating the resolvent difference around Γ yields (5). □

Corollary 4.8 (Invariance defect under approximate lumpability)**.** *Under the approximate lumpability condition that $\sum_{y \in X_j} K(x, y)$ depends on $x \in X_i$ only up to an error at most ε, the invariance defect satisfies $\|\mathcal{C}\|_\infty = \|QK\Pi\|_\infty \le B\varepsilon$. Consequently, $\|P - \Pi\| = O(B\varepsilon)$ when $M_\Gamma B\varepsilon < 1$.*

Proof. Let $u \in U$ with $\|u\|_\infty \le 1$, so $u|_{X_j} = u_j$. For $x \in X_i$: $(Ku)(x) = \sum_j p_{ij}(x)\, u_j$ where $p_{ij}(x) := \sum_{y \in X_j} K(x, y)$. The projection $\Pi K u$ replaces this by its block average $\bar{p}_{ij}$. Hence $|(QK\Pi\, u)(x)| = |\sum_j (p_{ij}(x) - \bar{p}_{ij})\, u_j| \le \sum_j |p_{ij}(x) - \bar{p}_{ij}| \le B\varepsilon$. □

Unlike Proposition 4.6, which requires an externally supplied exactly lumpable kernel, Proposition 4.7 constructs one canonically from K by killing the invariance defect $\alpha := \|QK\Pi\|$.

Lemma 4.9 (Macro spectrum identification)**.** *Let Π be the block-averaging projection and $A = \Pi K \Pi$. The matrix of $A|_U$ in the block-constant basis $\{1_{X_1}, \dots, 1_{X_B}\}$ is the $B \times B$ block-averaged kernel*

$$\bar{K}_{ij} = \frac{1}{|X_i|} \sum_{x \in X_i} \sum_{y \in X_j} K(x, y).$$

Hence $\sigma(A) = \sigma(\bar{K})$.

Proof. For $u^{(j)} = 1_{X_j} \in U$ and $x \in X_i$: $(Ku^{(j)})(x) = \sum_{y \in X_j} K(x, y) = p_{ij}(x)$. Applying Π: $(\Pi K u^{(j)})(x) = \frac{1}{|X_i|} \sum_{x' \in X_i} p_{ij}(x') = \bar{K}_{ij}$. Since $\Pi u^{(j)} = u^{(j)}$, we have $A u^{(j)} = \sum_i \bar{K}_{ij}\, 1_{X_i}$. □

Theorem 4.10 (Approximate lumpability preserves the slow sector)**.** *Let K be a stochastic kernel on a finite state space $X = X_1 \sqcup \cdots \sqcup X_B$. Let Π, Q, A, D, $\mathcal{B}$, $\mathcal{C}$ be as in Proposition 4.7. Assume:*

1. ***Approximate lumpability.*** *For each pair (X_i, X_j), the block transition mass $\sum_{y \in X_j} K(x, y)$ depends on $x \in X_i$ only up to an error at most ε.*
2. ***Macro spectrum.*** *The block-averaged kernel $\bar{K}$ (Lemma 4.9) has B eigenvalues in a region Ω near 1.*
3. ***Fast spectrum bound.*** *The fast operator $D = QKQ$ satisfies $\rho(D) \le 1 - \gamma$ for some $\gamma > 0$, with $\sigma(\bar{K}) \cap \Omega$ separated from $\sigma(D)$.*

Then:

(i) *The reference kernel $K_0 = \left(\begin{smallmatrix} A & \mathcal{B} \\ 0 & D \end{smallmatrix}\right)$ satisfies $\sigma(K_0) = \sigma(\bar{K}) \cup \sigma(D)$ and has an isolated $\dim(U)$-dimensional slow spectral cluster.*

(ii) *The invariance defect satisfies* $\|\mathcal{C}\| \le B\varepsilon$ *(Corollary 4.8).*

(iii) *For* $M_\Gamma B\varepsilon < 1$*, the kernel* K *has a* B*-dimensional invariant slow sector with Riesz projector satisfying*

$$\|P - \Pi\| = O(B\varepsilon).$$

(iv) *The* B *slow eigenvalues of* K *approximate those of* $\bar{K}$ *with error* $O(B\varepsilon)$*.*

Proof. Part (i): since K_0 is block upper-triangular with respect to $U \oplus U^\perp$, its spectrum is $\sigma(A) \cup \sigma(D)$. By Lemma 4.9, $\sigma(A) = \sigma(\bar{K})$. The spectral separation hypothesis ensures the slow cluster is isolated.

Part (ii): Corollary 4.8.

Part (iii): apply Proposition 4.7 with $K - K_0 = \left(\begin{smallmatrix} 0 & 0 \\ \mathcal{C} & 0 \end{smallmatrix}\right)$ and $\|\mathcal{C}\| \le B\varepsilon$.

Part (iv): since $\|P - \Pi\| < 1$ by part (iii), the restriction $\Pi|_{\mathrm{range}(P)}$ is invertible, so $\mathrm{range}(P)$ and $U = \mathrm{range}(\Pi)$ are isomorphic as B-dimensional subspaces. The slow eigenvalues of K are those of the $B \times B$ compressed operator $PKP|_{\mathrm{range}(P)}$. By Lemma 4.9, $\sigma(\Pi K \Pi|_U) = \sigma(\bar{K})$, so it suffices to bound the compression error. Write $P = \Pi + E$ with $\|E\| = O(B\varepsilon)$ by part (iii). Then

$$PKP - \Pi K \Pi = EK\Pi + \Pi KE + EKE,$$

so $\|PKP - \Pi K\Pi\| \le 2\|E\| + \|E\|^2 = O(B\varepsilon)$ (using $\|K\| \le 1$ and $\|\Pi\| \le 1$; note the bound depends on $\|E\| = \|P - \Pi\|$, not on $\|P\|$ itself). Eigenvalue continuity for $B \times B$ matrices gives the stated bound. □

Remark 4.11 (Formalization notes for Theorem 4.10). The statement of Theorem 4.10 typechecks in Lean 4 (see Appendix B); proofs of the main parts are deferred. Three encoding subtleties were exposed during formalization, none of which are mathematical errors:

1. Conclusion (i) (the spectral identity $\sigma(K_0) = \sigma(\bar{K}) \cup \sigma(D)$ for the block-triangular reference kernel) relies on spectral perturbation theory for non-Hermitian operators. Standard verification libraries (mathlib4 v4.29.1) do not yet provide the resolvent API required for this splitting. An elementary reformulation in terms of the characteristic polynomial of a block-triangular matrix may be useful for future formalization.

2. Hypotheses (H2) (macro spectrum) and (H3) (fast spectrum bound): the precise conditions invoke separation of non-Hermitian eigenvalue clusters. A reader following the formalization may find it useful to work first in the Hermitian or normal sub-case, where these conditions reduce to operator norms and the relevant mathlib4 tools are available.

3. The `ApproxLumpable` predicate as stated admits clean abstract equational reasoning but resists direct computational verification by tactics such as `simp` and `rw` on small finite instances, due to `let`-bound `Finset.filter` patterns. An operationally equivalent formulation stating the ε-bound as a maximum over block transitions may be more amenable to instance-based exploration.

Remark 4.12 (Sufficient conditions for Theorem 4.10, hypothesis 3). Approximate lumpability (hypothesis 1) controls the invariance defect $\mathcal{C} = QK\Pi$ but does not by itself control the fast operator $D = QKQ$. A sufficient route: if each within-block conditional kernel $R_i(x, y) := K(x, y)/\sum_{z \in X_i} K(x, z)$ has Dobrushin contraction coefficient $\eta(R_i) \le 1 - \beta_i$ and block retention $\sum_{z \in X_i} K(x, z) \ge 1 - \delta_i$, then within-block mean-zero modes contract, giving $\rho(D) \le \max_i (1 - \delta_i)(1 - \beta_i)$. Under the δ-separation assumption of Definition 6.1 with strictly positive transitions, this condition is satisfied.

Proposition 4.13 (Predictive coherence implies approximate lumpability)**.** *Let* $X = S_1 \sqcup \cdots \sqcup S_B$ *be a partition of a finite state space,* K *a transition kernel on* X*, and* $q_x \in \Delta(V)$ *the one-step emission distribution from state* x*. Assume:*

1. ***Intra-block predictive coherence.*** $\sup_{x,y\in S_i} \|q_x - q_y\|_{\mathrm{TV}} \le \varepsilon$ *for every block* S_i.

2. ***Token-block alignment.*** *There exists a partition* $V = V_1 \sqcup \cdots \sqcup V_B$ *such that* $K(x, S_j) = q_x(V_j) + r_x(j)$ *with* $\|r_x\|_{\mathrm{TV}} \le \eta$ *for every* $x \in X$.

Then for all i *and all* $x, y \in S_i$,

$$\big\|(K(x,S_1),\dots,K(x,S_B)) - (K(y,S_1),\dots,K(y,S_B))\big\|_{\mathrm{TV}} \le \varepsilon + 2\eta.$$

In particular, the partition is $(\varepsilon + 2\eta)$*-approximately lumpable, and Theorem 4.10 applies with* $\varepsilon_{\mathrm{lump}} = \varepsilon + 2\eta$. *(The constant* 2η *reflects the two-sided triangle-inequality bound on* $\|r_x - r_y\|_{\mathrm{TV}}$.*) This gives one explicit route from separation criteria to approximate lumpability; it is not claimed to characterize all such routes.*

Proof. Define $D\colon \Delta(V) \to \mathbb{R}^B$ by $(Dq)_j = q(V_j)$. Since D coarse-grains by a partition of V, it is TV-contractive (Proposition 3.3): $\|Dq_x - Dq_y\|_{\mathrm{TV}} \le \|q_x - q_y\|_{\mathrm{TV}}$. Write $\bar{K}(x) = Dq_x + r_x$ and $\bar{K}(y) = Dq_y + r_y$. Then

$$\|\bar{K}(x) - \bar{K}(y)\|_{\mathrm{TV}} \le \|Dq_x - Dq_y\|_{\mathrm{TV}} + \|r_x - r_y\|_{\mathrm{TV}} \le \varepsilon + 2\eta. \qquad \square$$

This bridges the emergence partition of Section 6 to the metastability machinery of Theorem 4.10: intra-block predictive coherence (hypothesis 1) ensures that states within the same preference set induce similar block-transition profiles, and token-block alignment (hypothesis 2) controls the residual mismatch between emission-based and transition-based block assignments. Together, they upgrade a δ-separated partition into an approximately lumpable one under explicit quantitative conditions.

Corollary 4.14 (Joint emergence and metastability on a common partition)**.** *Let* $X = S_1 \sqcup \cdots \sqcup S_B$ *be a* δ*-separated partition satisfying the hypotheses of Theorem 6.11 (positive divergence rate and cross-entropy separation) and the hypotheses of Proposition 4.13 (intra-block predictive coherence* ε *and token-block alignment* η*), with fast-spectrum separation (Theorem 4.10, hypothesis 3). Then the same partition simultaneously supports:*

(i) *typical-set separation with good-subset exponent* $D_{\mathrm{rate}}(P^*\|P_i)$*;*

(ii) *a* B*-dimensional slow spectral sector with projector error* $O(B(\varepsilon + 2\eta))$.

Proof. Part (i) is Theorem 6.11. Part (ii): Proposition 4.13 gives $(\varepsilon + 2\eta)$-approximate lumpability; Theorem 4.10 gives the slow sector. $\square$

Conjecture 4.15 (Sharp eigenvalue tracking)**.** *Under the hypotheses of Theorem 4.10, does the spectral gap of* K *remain* $\Theta(\delta)$ *when* $\varepsilon \ll \delta$*? The* $O(B\varepsilon)$ *eigenvalue bound above gives gap* $\ge \Theta(\delta) - O(B\varepsilon)$*, which is* $\Theta(\delta)$ *when* $\varepsilon = o(\delta/B)$*. Two distinct questions remain open: (i) is the linear dependence on* B *sharp, or does a tighter analysis give* $o(B)$ *dependence? (ii) separately, can the* ℓ^∞ *bound* $\|\mathcal{C}\|_\infty \le B\varepsilon$ *(Corollary 4.8) be replaced by an* ℓ^2*-weighted bound using the stationary measure, potentially reducing the factor to* $O(\sqrt{B}\,\varepsilon)$*?*

Remark 4.16 (Metastability without lumpability). The preceding results derive metastability from lumpability. A weaker but lumpability-free route uses conductance alone. For an ergodic reversible kernel K with stationary measure π and partition $X = S_1 \sqcup \cdots \sqcup S_B$, define the cross-block flow rate from block S_i:

$$\delta_{\mathrm{ker}}(i) = \frac{1}{\pi(S_i)} \sum_{x\in S_i} \pi(x) \sum_{y\notin S_i} K(y,x).$$



Figure 5: Spectral structure of the mixture kernel (Theorem 4.3). At $\delta = 0$, B eigenvalues equal 1. For $\delta > 0$, one remains (Perron, blue), $B - 1$ move to $1 - O(\delta)$ (red; these arise from the lumped chain). Within-domain eigenvalues (gray) satisfy $|\lambda| \leq 1 - \delta$ (Lemma 4.5).

Cheeger's inequality [12] gives

$$\text{gap}(K) \;\leq\; 2\,\Phi \;\leq\; 2 \min_{i:\,\pi(S_i)\leq 1/2} \delta_{\text{ker}}(i),$$

where Φ is the conductance. In particular, small δ_{ker} implies gap $\leq O(\delta_{\text{ker}})$ and mixing time $\Omega(1/\delta_{\text{ker}})$, without any lumpability assumption. The trade-off is precision: lumpability (Theorem 4.3) gives gap $= \Theta(\delta)$, tight on both sides; Cheeger gives only the upper bound gap $\leq O(\delta_{\text{ker}})$ and a quadratically loose lower bound gap $\geq \Phi^2/2 \geq \delta_{\text{ker}}^2/2$. The bound is one-sided: small δ_{ker} certifies slow mixing, but a measured small gap does not certify δ_{ker} without the lumpability structure. For non-reversible kernels, analogous bounds hold via the multiplicative reversibilization (see [12], §13.4).

4.5 Basin decomposition

For a metastable kernel, the second eigenvector v_2 of K encodes the almost-invariant decomposition. States where $v_2 > 0$ form one basin; states where $v_2 < 0$ form the other. The sign of v_2 gives the optimal two-partition into almost-invariant sets.

Remark 4.17 (Observable auto-correlation timescale). Under exact lumpability, any observable f that depends only on basin membership satisfies $\text{Cov}_\pi(f(X_t), f(X_0)) \sim \lambda_2^t$ as $t \to \infty$, where $\lambda_2 = 1 - \text{gap}$ is the second eigenvalue of the lumped chain. Token-level statistics that distinguish basins (e.g., preference-set frequencies) therefore exhibit persistence on the metastable timescale $\Theta(1/\text{gap})$.

The auto-correlation decay structure of Remark 4.17 suggests a data-driven recovery method for the slow mode. Proposition 4.18 makes this precise, and Proposition 4.20 shows that the coarse kernel's spectral gap tracks the full kernel's gap with second-order accuracy.

4.6 Slow-mode recovery and gap tracking

Proposition 4.18 (TICA slow-mode recovery)**.** *Let K be a reversible stochastic kernel on a finite state space with stationary measure π, spectral gap $\gamma = 1 - \lambda_2 > 0$, and λ_2 simple (multiplicity one). Define the time-lagged covariance operator*

$$C_\tau = \mathbb{E}_\pi[(X_0 - \bar{X})(X_\tau - \bar{X})^\top],$$

where $\bar{X} = \mathbb{E}_\pi[X]$ and $C_0 = \text{Var}_\pi(X)$ is the lag-0 covariance. The leading generalized eigenvector of $C_\tau v = \lambda C_0 v$ converges to the second eigenvector of K (the slow mode) as $\tau \to \infty$, with corresponding eigenvalue $\lambda_{\text{TICA}}(\tau) = \lambda_2^\tau + O(\lambda_3^\tau)$.

Proof. By the spectral decomposition of the reversible kernel, $C_\tau = \sum_{j \geq 2} \lambda_j^\tau \sigma_j^2 v_j v_j^\top$, where v_j are the eigenvectors of K and $\sigma_j^2 = \text{Var}_\pi(v_j)$. The ratio $\lambda_j^\tau / \lambda_2^\tau \to 0$ for $|\lambda_j| < |\lambda_2|$, so only the slow mode survives in the $\tau \to \infty$ limit. □

Remark 4.19 (TICA and non-identifiability). Proposition 4.18 operates on *observable* trajectory features and recovers the slow mode without accessing the latent state. This is consistent with the identifiability boundary of Remark 7.7: the coarse basin-level dynamics are recoverable from observations (the preference-set assignment gives the lumped chain directly), even though the within-basin latent state is not (Theorem 7.1). TICA succeeds precisely because it targets the recoverable coarse structure.

Proposition 4.20 (First-order gap cancellation)**.** *Let K_0 be a symmetric exactly lumpable kernel with B basins of equal size n, uniform within-basin transitions, and coarse kernel $K_L^{(0)}$. Let $K = K_0 + E$ where E breaks exact lumpability with defect $\varepsilon = \max_{x,y\in S_i,\,j} |K(x, S_j) - K(y, S_j)|$. Then*

$$|\lambda_j(K) - \lambda_j(K_L)| = O(\varepsilon^2)$$

for each slow eigenvalue λ_j, $j = 1, \ldots, B$, where K_L is the block-averaged (lumped) kernel of K.

Proof. Write the first-order eigenvalue correction for K at eigenvalue $\lambda_j(K_0)$:

$$\lambda_j(K) = \lambda_j(K_0) + w_j^\top E\, v_j + O(\|E\|^2),$$

where v_j is the right eigenvector and w_j the left eigenvector of K_0, normalized so $w_j^\top v_j = 1$. By exact lumpability, v_j is basin-constant: $v_j(x) = c_{j,\mathrm{basin}(x)}$ for a B-vector c_j. By the symmetric equal-basin hypothesis, w_j is also basin-constant: $w_j(x) = d_{j,\mathrm{basin}(x)}/n$.

Write the basin-to-basin mass of K as $\sum_{y\in S_m} K(x, y) = K_L(i, m) + \delta_x(m)$, where K_L is the block average and $\sum_{x\in S_i} \delta_x(m) = 0$ by definition. Then

$$w_j^\top E\, v_j = \underbrace{d_j^\top [K_L - K_L^{(0)}] c_j}_{\text{Term A}} + \underbrace{\frac{1}{n} \sum_i d_{j,i} \sum_{x\in S_i} \sum_m c_{j,m}\, \delta_x(m)}_{\text{Term B}}.$$

Term A is the first-order eigenvalue correction for K_L at eigenvalue $\lambda_j(K_L^{(0)})$. Term B vanishes because $\sum_{x\in S_i} \delta_x(m) = 0$ for every i, m. Since $\lambda_j(K_0) = \lambda_j(K_L^{(0)})$ (exact lumpability), we obtain $\lambda_j(K) = \lambda_j(K_L) + O(\varepsilon^2)$. □

4.7 Operator stability

Proposition 4.21 (Exact stability region)**.** *For the linearized self-conditioning operator $M = \lambda I + (1-\lambda)G$ with scalar mode g, the eigenvalue is $\mu = \lambda + (1-\lambda)g$. The exact stability condition is*

$$|\mu| < 1 \quad \Longleftrightarrow \quad -\frac{1+\lambda}{1-\lambda} < g < 1.$$

The upper boundary $g = 1$ gives $\mu = 1$ (neutral/persistent mode). The lower boundary $g = -(1+\lambda)/(1-\lambda)$ gives $\mu = -1$ (period-two boundary).

4.8 Non-normal transient growth

Proposition 4.22 (Transient amplification)**.** *Let M be a non-normal matrix with $\rho(M) < 1$. Then $\sigma_{\max}(M^k)$ can exceed 1 for finite k before eventually decaying.*

For the Jordan block $M = \left(\begin{smallmatrix} \rho & s \\ 0 & \rho \end{smallmatrix}\right)$ with $|\rho| < 1$ and $s \neq 0$:

$$M^k = \begin{pmatrix} \rho^k & k\rho^{k-1} s \\ 0 & \rho^k \end{pmatrix}.$$

The off-diagonal entry $k\rho^{k-1}s$ peaks at $k^* = 1/\ln(1/\rho)$ before decaying, so $\sigma_{\max}(M^k)$ attains a maximum at finite k^* and then decays. This provides a geometric complement to the asymptotic

spectral-radius criterion: eigenvalues govern long-run stability; singular values govern finite-time amplification. In the RIS setting, induced operators are generally non-self-adjoint with respect to the stationary measure (since token transitions need not be reversible), so non-normal transient growth can occur.

Metastability roadmap. *Exact:* Theorem 4.3 (exact spectrum under exact lumpability). *Perturbative:* Proposition 4.6 (slow eigenvalues persist near a known exactly lumpable kernel). *Approximate:* Theorem 4.10 (slow sector persists under approximate lumpability with quantitative projector and eigenvalue bounds). *Bridge:* Proposition 4.13 (predictive coherence + token-block alignment ⇒ approximate lumpability). *Lumpability-free:* Remark 4.16 (Cheeger's inequality gives gap $\leq O(\delta_{\text{ker}})$ from cross-block flow alone). *Open:* Conjecture 4.15 (sharpness of the $O(B\varepsilon)$ eigenvalue bound).

5 Mixture and Emergence

Consider a RIS trained on the collected outputs of B source systems. The training objective (cross-entropy minimization) fits the derived predictor to the mixture distribution

$$P^*(c \mid \text{context}) = \sum_{i=1}^{B} \pi_i \, P_i(c \mid \text{context}), \tag{6}$$

where P_i is the conditional next-commitment distribution of source i and π_i is its weight in the corpus. This section records standard mixture facts and their sequential consequences in the RIS setting.

Three propositions are stated below; the formal model class (δ-separated experts) and main theorems follow in Section 6. All results rely on standard information-theoretic tools (Jensen's inequality, the product structure of sequential generation, and support properties of convex combinations).

5.1 Emergence gap

The entropy [18] of the mixture exceeds the weighted average of the component entropies. This gap quantifies the additional uncertainty associated with replacing source-conditioned prediction by the unconditioned mixture.

Proposition 5.1 (Emergence gap)**.** *Let $P^* = \sum_{i=1}^{B} \pi_i P_i$ be a mixture of distributions on a finite set V, with $\pi_i > 0$ and $\sum_i \pi_i = 1$. Then*

$$H(P^*) - \sum_{i=1}^{B} \pi_i \, H(P_i) = \text{JSD}_{\boldsymbol{\pi}}(P_1, \ldots, P_B) \geq 0, \tag{7}$$

with equality if and only if all P_i are identical. The gap is bounded above by $\log B$, attained when the P_i have pairwise disjoint support.

Proof. The left side of (7) is the Jensen gap of the concave function H applied to $P^* = \sum_i \pi_i P_i$. Non-negativity follows from Jensen's inequality. Equality holds if and only if all arguments coincide, i.e., $P_i = P^*$ for all i. The identity with the Jensen–Shannon divergence [13] is standard; the upper bound $\log B$ follows from $\text{JSD} \leq H(\boldsymbol{\pi}) \leq \log B$. □

Remark 5.2. $\text{JSD}_{\boldsymbol{\pi}}(P_1, \ldots, P_B)$ measures the per-step entropy cost of replacing B source-specific predictors with a single mixture predictor. It can be interpreted as the information about source identity that is lost at each commitment step when the derived predictor generates without conditioning on which source is active.

5.2 Exponential amplification under sequential generation

At a single commitment step, the mixture distribution is a convex combination of the component distributions. Over a commitment chain, the interaction between mixture averaging and sequential composition produces an exponential effect: certain commitment chains become far more probable under the mixture than under any single component.

The following extreme case (where each predictor dominates at alternating steps) maximizes the per-step mixture-to-component ratio, making it the sharpest witness to the amplification mechanism.

Proposition 5.3 (Sequential amplification)**.** *Let P_A and P_B be two predictors on a binary commitment set, and let $P^* = \frac{1}{2}(P_A + P_B)$. Suppose at each step k, one predictor assigns probability $1-\varepsilon$ to the realized commitment and the other assigns ε, alternating between steps. Then for a commitment chain of length n:*

$$P_A(\mathit{chain}) = P_B(\mathit{chain}) = \big[\varepsilon(1-\varepsilon)\big]^{n/2}, \tag{8}$$

$$P^*(\mathit{chain}) = \left(\frac{1}{2}\right)^n, \tag{9}$$

and the ratio satisfies

$$\frac{P^*(\mathit{chain})}{P_A(\mathit{chain})} = \left(\frac{1}{2\sqrt{\varepsilon(1-\varepsilon)}}\right)^n. \tag{10}$$

Since $\varepsilon(1-\varepsilon) \le 1/4$ with equality only at $\varepsilon = 1/2$, the base of the exponential is at least 1, with strict inequality whenever $\varepsilon \neq 1/2$.

Proof. At each step, the mixture predictor assigns probability $\frac{1}{2}(1-\varepsilon) + \frac{1}{2}\varepsilon = \frac{1}{2}$, regardless of which component dominates. Over n steps, $P^*(\text{chain}) = (1/2)^n$. Each component assigns $(1-\varepsilon)$ at its dominant steps and ε at its non-dominant steps, giving $(1-\varepsilon)^{n/2}\varepsilon^{n/2} = [\varepsilon(1-\varepsilon)]^{n/2}$. The ratio follows by division. The bound on the base follows from the AM–GM inequality applied to $\varepsilon(1-\varepsilon) \le ((\varepsilon + 1 - \varepsilon)/2)^2 = 1/4$. □

Corollary 5.4 (Crossover length)**.** *For a given amplification threshold $K > 1$, the minimum chain length at which the mixture probability exceeds the best component probability by a factor of K is*

$$n^* = \left\lceil \frac{\log K}{\log(1/(2\sqrt{\varepsilon(1-\varepsilon)}))} \right\rceil. \tag{11}$$

Remark 5.5. Proposition 5.3 is an extremal witness and is not intended to represent generic amplification magnitudes: the alternating assumption sets the per-step mixture probability to exactly 1/2, which maximizes the ratio. For a general interleaving of component preferences, the mixture probability per step is $\frac{1}{2}(P_A(c_k) + P_B(c_k))$, which varies with the commitment sequence. The exponential growth of the ratio persists whenever different components assign high probability to different commitments at different positions, but the specific rate depends on the separation pattern. A general bound beyond the alternating case is not proved here; Theorem 6.11 provides the asymptotic version via relative entropy rates.

5.3 Token-level containment and sequence-level separation

The preceding results leave open a natural question: does the mixture predictor generate commitments that no individual component would produce? The answer depends on the level of description.

Proposition 5.6 (Support structure)**.** *Let $P^* = \sum_i \pi_i P_i$ with $\pi_i > 0$.*

(a) Token-level containment. *At each step,*

$$\operatorname{supp}(P^*(\cdot \mid \text{context})) \subseteq \bigcup_{i=1}^{B} \operatorname{supp}(P_i(\cdot \mid \text{context})). \tag{12}$$

No individual commitment is produced by the mixture that could not be produced by at least one component.

(b) Sequence-level amplification. *There exist commitment chains $c_1, \ldots, c_n$ such that*

$$P^*(c_1, \ldots, c_n) > \max_i P_i(c_1, \ldots, c_n)$$

by a factor that grows exponentially in n under the conditions of Proposition 5.3.

Proof. Part (a): if $P_i(c \mid \text{context}) = 0$ for all i, then $P^*(c \mid \text{context}) = \sum_i \pi_i \cdot 0 = 0$. The contrapositive gives the inclusion.

Part (b): the alternating construction in Proposition 5.3 provides an explicit family of such chains, with amplification ratio given by (10). □

Remark 5.7. The mechanism is the interaction between convex averaging (at each step) and multiplicative composition (across steps): a convex combination cannot exit the support of its components at any single step, but the product over many steps can assign far more mass to certain trajectories than any single component's product does. This phenomenon is classical; the contribution here is its realization within the RIS setting as the source of sequence-level separation despite token-level support containment.

6 Emergence Theorems

Under separation and ergodicity assumptions, the mixture-typical set carries exponentially vanishing mass under each constituent process. This section gives a constructive result with explicit bounds and an asymptotic typical-set separation theorem.

6.1 Model class

Definition 6.1 (δ-separated expert family)**.** A family $\{P_1, \ldots, P_B\}$ of stationary first-order Markov predictors on a finite state space V, each with irreducible, aperiodic, strictly positive transition matrix $P_i(\cdot \mid x)$, is *δ-separated* if there exist pairwise disjoint preference sets $S_1, \ldots, S_B \subseteq V$ with $\bigcup_i S_i = V$ such that for each i and each current state $x \in V$:

$$P_i(S_i \mid x) \geq 1 - \delta.$$

The parameter $\delta \in [0, 1/2)$ controls separation. This is a specialization of the general RIS (Definition 1.1) to finite-state first-order Markov dynamics. All results in this section extend to k-th order Markov predictors by state-space expansion to V^k, at the cost of exponential growth in state-space size.

Definition 6.2 (Mixture predictor)**.** The *mixture predictor* is $P^* = \sum_{i=1}^{B} \pi_i P_i$ with $\pi_i > 0$, $\sum_i \pi_i = 1$. In the uniform case, $\pi_i = 1/B$.

Each P_i is uniformly ergodic by these assumptions. The mixture process P^* is stationary and ergodic. Shannon–McMillan–Breiman [2] applies to P^*. The *relative entropy rate* is

$$D_{\text{rate}}(P^* \| P_i) = \lim_{n \to \infty} \tfrac{1}{n} D(P^*_{1:n} \,\|\, P_{i,1:n}) = \mathbb{E}_{P^*}[D(P^*(\cdot \mid X_{<t}) \,\|\, P_i(\cdot \mid X_{<t}))],$$

where the second equality holds under stationarity.

6.2 Positive divergence lemma

Lemma 6.3 (Positive divergence via mass misallocation)**.** *Under δ-separation with $\delta < 1/2$ and $B \geq 2$, the relative entropy rate satisfies*

$$D_{\text{rate}}(P^* \| P_i) > 0 \quad \textit{for every } i.$$

Proof. By Gibbs' inequality, $D(P^*(\cdot \mid x) \| P_i(\cdot \mid x)) \geq 0$ with equality if and only if $P^*(\cdot \mid x) = P_i(\cdot \mid x)$. Under δ-separation with $B \geq 2$ disjoint preference sets and positive mixture weights, $P^*(\cdot \mid x) \neq P_i(\cdot \mid x)$ for every $x \in V$. To see this, write $P^* = \pi_i P_i + (1 - \pi_i) Q_i$ where $Q_i = \sum_{k \neq i} (\pi_k / (1 - \pi_i)) P_k$. If $P^* = P_i$ then $Q_i = P_i$; but $Q_i(S_i \mid x) \leq \delta$ (each P_k with $k \neq i$ places mass $\leq \delta$ on S_i by δ-separation, Definition 6.1) while $P_i(S_i \mid x) \geq 1 - \delta > \delta$, a contradiction. Hence $D(P^*(\cdot \mid x) \| P_i(\cdot \mid x)) > 0$ for every state x. Under the stationary ergodic assumptions (the stationary measure has full support by irreducibility), the expectation is strictly positive:

$$D_{\text{rate}}(P^* \| P_i) = \mathbb{E}_{P^*}\big[D\big(P^*(\cdot \mid X) \,\|\, P_i(\cdot \mid X)\big)\big] > 0.$$

□

Remark 6.4. The separation is driven by mass misallocation across partitions, not per-token mismatch: each component places most of its mass on its own preference set, while the mixture spreads mass across all sets. Gibbs' inequality converts this partition-level imbalance into a per-step KL lower bound. This mechanism underlies Theorem 6.11.

Proposition 6.5 (Soft D_{rate} bound under overlapping supports)**.** *Let $P_1, \ldots, P_B$ be stationary first-order Markov predictors on a finite state space with positive transitions, and let $P^* = \frac{1}{B} \sum_{i=1}^{B} P_i$ be the uniform mixture. Assume $\varepsilon := \mathbb{E}_{\pi^*}[\text{TV}(P^*(\cdot \mid x),\, P_i(\cdot \mid x))] > 0$ for some i, where π^* is the stationary measure of P^*. Then*

$$D_{\text{rate}}(P^* \| P_i) \;\geq\; 2\varepsilon^2.$$

No δ-separation or disjoint preference-set assumption is required.

Proof. By Pinsker's inequality, for each context x:

$$D\big(P^*(\cdot \mid x) \,\big\|\, P_i(\cdot \mid x)\big) \;\geq\; 2\ \text{TV}\big(P^*(\cdot \mid x),\, P_i(\cdot \mid x)\big)^2.$$

Averaging over the stationary measure π^*:

$$D_{\text{rate}}(P^* \| P_i) = \mathbb{E}_{\pi^*}\big[D(P^*(\cdot \mid x) \| P_i(\cdot \mid x))\big] \geq 2\, \mathbb{E}_{\pi^*}\big[\text{TV}(\cdots)^2\big].$$

By Jensen's inequality (x^2 is convex): $\mathbb{E}[\text{TV}^2] \geq (\mathbb{E}[\text{TV}])^2 = \varepsilon^2$. □

Remark 6.6. The bound $2\varepsilon^2$ is tight up to constants when total variation is the only measure of separation. Under δ-separation with disjoint preference sets, Lemma 6.3 gives tighter estimates via partition-level mass misallocation. Proposition 6.5 extends the positive divergence rate to overlapping-support settings where the preference-set assumption fails.

6.3 Constructive emergence

Definition 6.7 (Emergent family)**.** The *emergent family* E_n consists of commitment chains of length n that cycle through all B domains:

$$E_n = \big\{(c_1, \ldots, c_n) : c_k \in S_{(k \bmod B)+1}\big\}.$$

Theorem 6.8 (Constructive emergence)**.** *For the symmetric δ-separated cycling construction with uniform mixture and $\delta < 1/(1 + B^{B/(B-1)})$:*

(i) $P^*(E_n) \geq [(1-\delta)/B]^n$.

(ii) *For each component i:* $P_i(E_n) \leq \delta^{n(B-1)/B}$.

(iii) *The ratio is bounded below:*

$$\frac{P^*(E_n)}{\max_i P_i(E_n)} \geq r^n, \qquad r = \frac{[(1-\delta)/\delta]^{(B-1)/B}}{B}. \tag{13}$$

This is a lower bound; the true ratio may be larger.

(iv) $r > 1$ *under the stated threshold on δ.*

Proof. At each step k where $c_k \in S_j$, the mixture assigns $P^*(S_j \mid x) \geq (1-\delta)/B$ (expert j contributes at least this). Over n steps, $P^*(E_n) \geq [(1-\delta)/B]^n$.

For component i: at the n/B steps where $c_k \in S_i$, $P_i(S_i \mid x) \leq 1$; at the $n(B-1)/B$ steps where $c_k \notin S_i$, $P_i(S_j \mid x) \leq \delta$. Hence $P_i(E_n) \leq \delta^{n(B-1)/B}$.

Dividing gives (13). The condition $r > 1$ reduces to $[(1-\delta)/\delta]^{(B-1)/B} > B$, which holds for δ below the stated threshold. □

Remark 6.9. The constructive ratio $\log r$ of Theorem 6.8 is a finite-horizon witness of the partition-level mass-misallocation mechanism. Theorem 6.11 below sharpens this to the good-subset exponent $D_{\text{rate}}(P^* \| P_i)$.

6.4 Typical-set separation

Lemma 6.10 (Uniform ergodicity of components and mixture)**.** *Each P_i and the mixture $P^* = \sum_i \pi_i P_i$ are uniformly ergodic. In particular, the additive observables used in Theorem 6.11 below are bounded.*

Proof. Strict positivity on a finite state space gives $\min_{x,y} P_i(y \mid x) > 0$, so each P_i has Dobrushin coefficient $\eta(P_i) < 1$ (see [12], Thm. 4.9). Since P^* is a convex combination with positive weights of strictly positive kernels, it is also strictly positive ($P^*(y \mid x) \geq \pi_i P_i(y \mid x) > 0$ for every x, y). □

The constructive result above exhibits a specific family; the following theorem proves asymptotic typical-set separation in the stated regime.

Main technical dependencies. The proof of Theorem 6.11 uses four ingredients: (i) finite state space, ensuring all measures are finitely supported; (ii) strict positivity of all transition probabilities, which gives uniform ergodicity (Lemma 6.10); (iii) existence of entropy-rate and relative entropy-rate quantities under stationarity (Shannon–McMillan–Breiman); (iv) a Hoeffding-type concentration inequality for bounded additive functionals of uniformly ergodic Markov chains [6, 14]. All four are standard in our finite-state setting; the restrictive assumptions are the two explicitly listed in the theorem statement (positive D_{rate} and cross-entropy separation).

Theorem 6.11 (Typical-set separation)**.** *Let $\{P_1, \ldots, P_B\}$ be δ-separated stationary first-order Markov predictors on a finite state space with strictly positive transition probabilities. Let $P^* = \sum_i \pi_i P_i$ with $\pi_i > 0$. Suppose:*

(a) $D_{\text{rate}}(P^* \| P_i) > 0$ *for every i (Lemma 6.3);*

(b) ***Cross-entropy separation.*** $\Delta_i := H_{\text{rate}}(P_i, P^*) - h(P^*) > 0$ *for every i, where $H_{\text{rate}}(P_i, P^*)$ is the cross-entropy rate of P^* evaluated under the stationary law of P_i.*

The theorem shows two things: P^* concentrates on its typical set (part i), and each component P_i assigns exponentially small mass to that set (parts ii–iii).

Define the P^*-typical set at tolerance ε:

$$T_n^\varepsilon(P^*) = \{x_{1:n} : \left|-\tfrac{1}{n}\log P^*(x_{1:n}) - h(P^*)\right| < \varepsilon\}.$$

Then:

(i) *For every $\varepsilon > 0$, $P^*\big(T_n^\varepsilon(P^*)\big) \to 1$ as $n \to \infty$.*

(ii) *For every i, setting $c_i := D_{\mathrm{rate}}(P^*\|P_i)$ and choosing $0 < \varepsilon < \min(c_i/3,\ \Delta_i/2)$, there exist constants $C_i < \infty$, $\gamma_i > 0$, and n_0 such that for all $n \geq n_0$,*

$$P_i\big(T_n^\varepsilon(P^*)\big) \leq e^{-n(c_i-3\varepsilon)} + C_i\, e^{-\gamma_i n}.$$

In particular, for every $\eta < \min\{c_i - 3\varepsilon,\ \gamma_i\}$, there exists n_1 such that

$$P_i\big(T_n^\varepsilon(P^*)\big) \leq 2\, e^{-n\eta} \quad \text{for all } n \geq n_1.$$

(iii) *In particular, $\sup_i P_i\big(T_n^\varepsilon(P^*)\big) \to 0$ exponentially.*

On the good subset, exponential separation is governed by $c_i = D_{\mathrm{rate}}(P^*\|P_i)$. Exponential decay of the full P_i-mass on $T_n^\varepsilon(P^*)$ additionally uses assumption (b).

The theorem establishes that recursive mixtures are not merely interpolations of their components: under positive divergence rate and cross-entropy separation, they generate their own asymptotically separated typical behavior.

Remark 6.12 (Independence of assumption (b)). Assumptions (a) and (b) involve different relative-entropy directions and are formally independent. The standard identity

$$H_{\mathrm{rate}}(P_i, P^*) - h(P^*) = h(P_i) - h(P^*) + D_{\mathrm{rate}}^{\pi_i}(P_i\|P^*)$$

shows that Δ_i is a sum of an entropy-rate gap and a KL term in the $P_i \to P^*$ direction; assumption (a)'s $D_{\mathrm{rate}}(P^*\|P_i)$ is in the opposite direction. Both can hold simultaneously under δ-separation without one implying the other: δ-separation gives $D_{\mathrm{rate}}(P^*\|P_i) > 0$ via Lemma 6.3, but does not by itself entail $\Delta_i > 0$, because the stationary-measure shift $\pi_i \neq \pi^*$ can in principle offset the KL contribution. The two-hypothesis form is therefore structurally honest: each names a distinct condition. The concrete RIS of Section 8 satisfies both conditions; Proposition 6.5 establishes sufficient conditions on the predictors under which both hold jointly without assuming hard δ-separation.

Proof. Step 1 (SMB for P^*). Part (i) follows from Shannon–McMillan–Breiman applied to P^* (stationary, ergodic).

Step 2 (Likelihood-ratio law under P^*). Because all transition probabilities are strictly positive on a finite state space, both P_i and P^* are uniformly ergodic (Lemma 6.10), and the per-step log-likelihood observable is bounded. On $T_n^\varepsilon(P^*)$,

$$P^*(x_{1:n}) \geq \exp\big(-n(h(P^*) + \varepsilon)\big).$$

Since all processes are finite-state ergodic Markov chains with strictly positive transitions, the log-likelihood ratio satisfies

$$\tfrac{1}{n}\log\frac{P_i(x_{1:n})}{P^*(x_{1:n})} \;\to\; -D_{\mathrm{rate}}(P^*\|P_i) \quad P^*\text{-a.s.}$$

by the ergodic theorem for stationary Markov chains (the i.i.d. version is in [2], Ch. 16; the Markov extension follows from the per-step ergodic theorem applied to the bounded additive functional $\log(P^*(x_t \mid x_{t-1})/P_i(x_t \mid x_{t-1}))$).

Step 3 (Bound on the good subset G_n). Define the "good" subset of the typical set:

$$G_n = \{x_{1:n} \in T_n^\varepsilon(P^*) : \tfrac{1}{n} \log \tfrac{P_i(x_{1:n})}{P^*(x_{1:n})} \le -D_{\text{rate}} + \varepsilon\}, \qquad B_n = T_n^\varepsilon(P^*) \setminus G_n.$$

On G_n, the typical-set membership gives $P^*(x_{1:n}) \le \exp(-n(h(P^*) - \varepsilon))$, and the log-ratio bound gives $P_i(x_{1:n}) \le P^*(x_{1:n}) \cdot \exp(-n(D_{\text{rate}} - \varepsilon))$. Combining:

$$P_i(x_{1:n}) \le \exp\big(-n(h(P^*) + D_{\text{rate}} - 2\varepsilon)\big).$$

Since $|G_n| \le |T_n^\varepsilon(P^*)| \le \exp(n(h(P^*) + \varepsilon))$:

$$P_i(G_n) \le \exp\big(-n(D_{\text{rate}} - 3\varepsilon)\big).$$

Thus $P_i(G_n)$ is exponentially small with good-subset exponent $c_i = D_{\text{rate}}(P^* \| P_i)$. To control the remaining subset B_n, we use assumption (b) and concentration of the log-loss observable under P_i.

(At this point, the key idea is established: on the P^*-typical set, per-token loss under P_i exceeds loss under P^* by the divergence rate D_{rate}. The remaining step makes this concentration rigorous.)

Step 4 (Log-loss concentration under P_i). Under P_i, the ergodic theorem gives

$$-\tfrac{1}{n} \log P^*(X_{1:n}) \;\to\; H_{\text{rate}}(P_i, P^*) \quad P_i\text{-a.s.}$$

By assumption (b), $\Delta_i := H_{\text{rate}}(P_i, P^*) - h(P^*) > 0$.

Step 5 (Concentration bound for the bad subset). Choose $\varepsilon < \Delta_i/2$. Then membership in $T_n^\varepsilon(P^*)$ requires

$$\big|-\tfrac{1}{n} \log P^*(X_{1:n}) - h(P^*)\big| < \varepsilon,$$

whereas under P_i the same quantity converges to $h(P^*) + \Delta_i$. Under P_i the process is stationary (by assumption) and uniformly ergodic (finite state, irreducible, aperiodic, strictly positive transitions). The additive functional

$$g(x_{t-1}, x_t) := -\log P^*(x_t \mid x_{t-1})$$

is bounded (strict positivity of P^* ensures $g < \infty$). Concentration for bounded additive functionals of uniformly ergodic Markov chains (Hoeffding-type; see Glynn & Ormoneit [6] or Paulin [14]) therefore applies and yields constants $C_i, \gamma_i > 0$ such that

$$P_i(T_n^\varepsilon(P^*)) \le C_i\, e^{-\gamma_i n}.$$

In particular, $P_i(B_n) \le P_i(T_n^\varepsilon(P^*)) \le C_i\, e^{-\gamma_i n}$.

Step 6 (Combine bounds). For $\varepsilon < \min(c_i/3,\ \Delta_i/2)$ and all sufficiently large n,

$$P_i(T_n^\varepsilon(P^*)) = P_i(G_n) + P_i(B_n) \le e^{-n(c_i - 3\varepsilon)} + C_i\, e^{-\gamma_i n}.$$

Therefore, for any $\eta < \min\{c_i - 3\varepsilon,\ \gamma_i\}$, there exists n_1 such that

$$P_i(T_n^\varepsilon(P^*)) \le 2\, e^{-n\eta} \quad \text{for all } n \ge n_1.$$

This proves exponential decay of the full P_i-mass on the typical set, with rate bounded below by any $\eta < \min\{c_i - 3\varepsilon,\ \gamma_i\}$. □

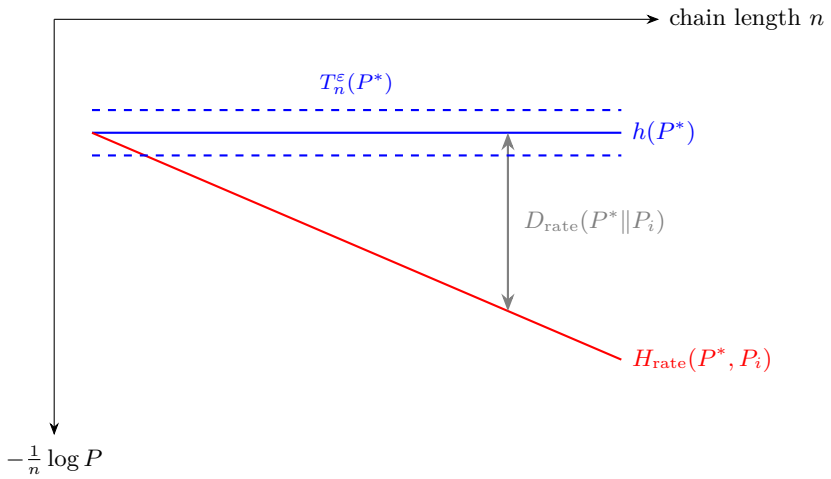


Figure 6: Typical-set separation (Theorem 6.11). On the good subset of P^*-typical sequences (blue band), per-token log-loss under P_i (red) exceeds that under P^* (blue) by $D_{\mathrm{rate}}(P^*\|P_i)$. This yields good-subset exponential separation; full typical-set decay additionally requires cross-entropy separation (assumption (b)).

Remark 6.13. The good-subset exponent is the relative entropy rate $D_{\mathrm{rate}}(P^*\|P_i)$, not the entropy gap $h(P^*) - h(P_i)$. The bad-subset bound additionally requires cross-entropy separation (assumption (b)): $\Delta_i = H_{\mathrm{rate}}(P_i, P^*) - h(P^*) > 0$. Together these give exponential decay of the full P_i-mass on the typical set.

Remark 6.14 (Two senses of emergence). Theorem 6.8 gives *constructive emergence*: an explicit family with an exponential ratio at finite n. Theorem 6.11 gives *typical-set emergence*: asymptotically, the mixture concentrates on commitment chains that every component assigns exponentially small probability to. Both are stated in the δ-separated regime; Proposition 6.5 extends the positive divergence rate to overlapping-support settings. δ-separation induces partition-level mass mismatch, which drives typical-set separation. Under additional predictive-coherence and alignment conditions (Proposition 4.13), it also induces approximate invariance of those partitions, which drives metastability (Section 4).

Remark 6.15 (Recursion versus static mixture averaging). A single-step mixture $P^* = \sum_i \pi_i P_i$ is a convex combination; its support is contained in the union of component supports (Proposition 5.6 a), and per-token distinguishability is bounded by $\mathrm{JSD}_{\boldsymbol{\pi}} \le \log B$. Recursion changes the picture in three ways.

(i) The mixture is over *sequential predictors*, not one-shot distributions: the object of separation is a commitment chain produced under iterated emission and re-injection, not an i.i.d. sample.

(ii) Per-step convexity composes multiplicatively. Token marginals can look similar, yet the sequence-level likelihood ratio $P^*(x_{1:n})/P_i(x_{1:n})$ grows exponentially on the good subset identified in Theorem 6.11; full typical-set exclusion additionally uses assumption (b) (alternating case: Proposition 5.3; general bound: Theorem 6.11).

(iii) The correct good-subset separation exponent is the relative entropy *rate* $D_{\mathrm{rate}}(P^*\|P_i)$, which is a per-step dynamical quantity defined under the stationary measure of the recursive process.

Without recursive iteration, mixture-vs-component distinctions remain static convexity statements; recursive commitment upgrades them into trajectory-level asymptotic separation.

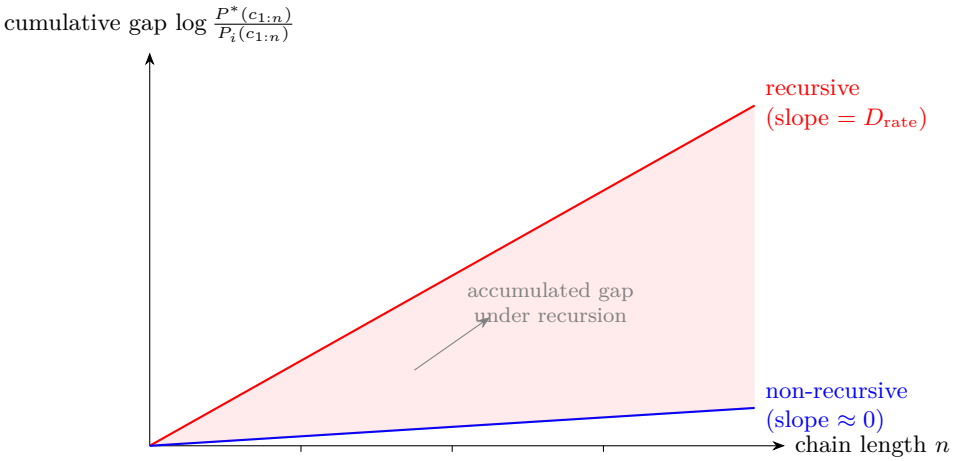


Figure 7: Recursive versus non-recursive divergence. Under recursive emission and reinjection, the cumulative log-likelihood ratio $\log P^*(c_{1:n})/P_i(c_{1:n})$ grows linearly in n, with slope $D_{\text{rate}}(P^*\|P_i) > 0$ on the good subset (Theorem 6.11 and Lemma 6.3). Without recursion (i.i.d. mixture sampling), the cumulative ratio is a martingale with mean zero and the slope vanishes. Recursion is the mechanism that converts per-step convex combination into trajectory-level separation.

6.5 Loop Closure independence

The three structural conditions used as sufficient hypotheses for the main RIS results ($\Delta_i > 0$, hard token partition, geometric basin separation) are each *individually* unnecessary for a recursive system to exhibit Loop Closure-shaped behavior: positive divergence accumulation under persistent regime switching. The following theorem makes this precise.

Theorem 6.16 (Loop Closure independence of the RIS sufficient conditions)**.** *For each of the three RIS sufficient conditions (positive entropy-rate gap $\Delta_i > 0$, hard token partition with disjoint emission supports, and geometric basin separation requiring a token that uniquely identifies some regime state), there exists a finite recursive system in which that condition fails while Loop Closure holds (positive divergence between regime states, self-sustaining under feedback). More precisely:*

(i) ***Witness A.*** *A two-state, two-token system in which the entropy-rate gap condition fails (aggregate emission matches each component), yet two regime states maintain distinct emission weights and alternating feedback sustains the separation.*

(ii) ***Witness B.*** *A two-state, three-token system in which the hard-partition condition fails (one token is positive in both regime states), yet regime states retain distinct dominant tokens and feedback is non-collapsing.*

(iii) ***Witness C.*** *A two-state, two-token system in which geometric basin separation fails (every token is reachable from both states, so no token uniquely identifies a regime), yet the two states have distinct emission weights and alternating feedback closes the loop.*

The three witnesses are formally verified in Lean 4 (see Appendix B).

Remark 6.17. The theorem does not assert that the sufficient conditions are unnecessary for the sharp quantitative bounds in Theorem 6.11 or Theorem 4.3. It asserts only that Loop Closure behavior (understood as positive divergence plus feedback self-sustainability in a finite abstract recursive system) does not require any of the three conditions individually. The Lean-verified witnesses are finite discrete systems; no claim is made about continuous or high-dimensional settings.

7 Non-Identifiability

Non-identifiability follows from many-to-one emission alone and does not require finite-state or spectral assumptions.

Theorem 7.1 (Non-identifiability under lossy emission)**.** *Let $(H_t)_{t\geq 0}$ be a latent stochastic process with state space $\mathcal{H} \subseteq \mathbb{R}^d$, and let $(C_t)_{t\geq 0}$ be the observable token process on a finite alphabet V. Assume:*

1. ***Emission map.*** *There is a measurable map*

$$\phi\colon \mathcal{H} \to \Delta(V), \qquad \phi(h) = \mathrm{softmax}(Wh),$$

   with $\mathrm{rank}(W) = |V| - 1$.

2. ***Dimension condition.***

$$d > |V| - 1.$$

3. ***Reachability.*** *The initial distribution $P(H_0 \in \cdot)$ assigns positive probability to at least two distinct states $h_0, h_0' \in \mathcal{H}$ with $h_0' - h_0 \in \ker(W) \setminus \{0\}$.*

Then the latent-to-observation relation is non-identifiable: there exist distinct latent trajectories $H_{0:t} \neq H'_{0:t}$ that are compatible with the same observation history with positive probability. Consequently, for any observation history $c_{0:t}$ realized by both trajectories, the posterior

$$P(H_{0:t} \mid C_{0:t} = c_{0:t})$$

is not a delta measure.

The result is about non-injectivity on trajectories, not merely pointwise coincidence of emissions: distinct $H_{0:t} \neq H'_{0:t}$ produce the same observation history with positive probability.

Proof. Since $\mathrm{rank}(W) = |V| - 1$ and $d > |V| - 1$, the rank–nullity theorem gives

$$\dim \ker(W) = d - (|V| - 1) > 0.$$

Hence there exists a nonzero vector $u \in \ker(W)$. (The key idea: perturbations along $\ker(W)$ leave the emission distribution unchanged, so the observer cannot distinguish h from $h+u$.) For any $h \in \mathcal{H}$ such that $h + u \in \mathcal{H}$,

$$W(h + u) = Wh,$$

so

$$\phi(h + u) = \mathrm{softmax}(W(h + u)) = \mathrm{softmax}(Wh) = \phi(h).$$

Thus the emission map is not injective: distinct latent states can induce the same token distribution.

By the reachability assumption, there exist distinct states $h_0, h_0' \in \mathcal{H}$ with $h_0' - h_0 \in \ker(W) \setminus \{0\}$, both having positive prior probability. Since $Wh_0 = Wh_0'$, they emit the same distribution: $\phi(h_0) = \phi(h_0')$. For any token $c_0 \in V$ with $\phi(h_0)(c_0) > 0$, both states produce c_0 with the same probability, giving two distinct one-step trajectories compatible with observation c_0. The two trajectories starting from h_0 and h_0' are already distinct as sequences (they differ at $t = 0$), so the latent-to-observation relation is non-identifiable. Because both initial states have positive prior probability and assign the same emission probability to c_0, both remain in the support of the conditional law $P(H_0 \mid C_0 = c_0)$. The posterior is therefore not a point mass. □

Remark 7.2 (Multi-step persistence and transition regularity). The proof above establishes non-injectivity at the initial step. For persistence across multiple steps, one sufficient condition is *transition regularity*: for distinct h, h' in the same emission fiber and some token c with $\phi(h)(c) > 0$, $f(h, c) \neq f(h', c)$ (the re-injection map does not merge fiber-equivalent states). This is sufficient but not the unique route to persistence: exact lumpability (Section 8, §8.2) provides an alternative via basin-invariant observation laws, under which distinct fiber members remain distinguishable even without pointwise non-merging. The concrete RIS of Section 8 satisfies both conditions (the η-perturbations ensure $f(1, c) \neq f(2, c)$, and the kernel is exactly lumpable) and exhibits multi-step non-identifiability explicitly.

Corollary 7.3 (Output underdetermines latent trajectory)**.** *Under the hypotheses above, observable token sequences do not uniquely determine preceding internal trajectories. The token process is a lossy factor of the latent process (in the sense of a non-injective factor map).*

Equivalently, the generation pipeline defines a non-injective factor map from latent residual dynamics to a finite token process; identifiability fails at the level of internal trajectories. The connection between non-injective emission and hidden-state aliasing is classical in HMM theory [10, 15]; the RIS formulation extends it to the recursive re-injection setting. Under additional predictive-coherence and token-block alignment assumptions (Proposition 4.13), a δ-separated partition can also become approximately lumpable, linking the emergence and metastability analyses. A complementary no-go result for observation under irreversible compression is given in [8], where the same rank-deficiency mechanism produces a column-stochastic rigidity obstruction for deterministic selectors on branch-conditioned records.

Corollary 7.4 (Output process need not be Markov)**.** *Even when the latent process (H_t) is Markov, a non-injective emission map ϕ can produce an observable process (C_t) that is not Markov.*

Proof. This is a standard property of hidden Markov models [15]. Non-injective ϕ means the current observation C_t does not determine the latent state H_t. When distinct latent states in the same emission fiber have different transition kernels, the conditional law of C_{t+1} given C_t depends on the posterior over H_t, which in turn depends on the full history $C_{0:t}$. □

Remark 7.5 (Sufficient conditions for identifiability). Two sufficient routes to restoring identifiability of latent trajectories from observations are: (a) injective emission ($\dim \ker(W) = 0$), or (b) dynamics that collapse all emission fibers to single points. Under the hypotheses of Theorem 7.1, neither holds, and non-identifiability follows.

Corollary 7.6 (Positive posterior uncertainty)**.** *Under the hypotheses of Theorem 7.1, the posterior entropy $H(H_{0:t} \mid C_{0:t} = c_{0:t})$ is positive for any observation history realized by multiple positive-probability latent trajectories.*

Proof. The posterior assigns positive mass to at least two distinct trajectories (Theorem 7.1), so its entropy exceeds that of a point mass. □

Remark 7.7 (Identifiability boundary under lumpability). Under exact lumpability (Theorem 4.3), these results delineate a precise boundary: the coarse basin-level dynamics are recoverable from observations (the preference-set assignment $Y_t = i$ iff $C_t \in S_i$ gives the lumped chain K_L directly), but within-basin state is not (Theorem 7.1). The observable process determines the B-state coarse trajectory but not the fine-grained latent trajectory.

Remark 7.8 (Emission geometry). The singular value decomposition $W = U\Sigma V^\top$ separates observable latent directions from unobservable directions. Directions in $\ker(W)$ are completely invisible to the emission, while directions associated with small but nonzero singular values are only weakly observable. The compression is generally anisotropic, with sensitivity controlled by the singular spectrum. On the observable subspace (the nonzero singular values), the condition number $\sigma_{\max}/\sigma_{\min}$ quantifies the anisotropy of the emission.

Remark 7.9 (Weak-direction observability bound). The following bounds quantify how perturbations along weak singular directions of W produce small emission changes, interpolating between full non-identifiability ($\sigma_{\min} = 0$) and strong observability ($\sigma_{\min}$ large). The softmax map satisfies $\|\operatorname{softmax}(z) - \operatorname{softmax}(z')\|_{\mathrm{TV}} \le \frac{1}{2}\|z - z'\|_\infty$ (a standard bound; see the mean-value theorem applied to the softmax Jacobian). For a logit perturbation Wu with $\|u\| = 1$:

$$\|\phi(h + \varepsilon u) - \phi(h)\|_{\mathrm{TV}} \le \tfrac{1}{2}\,\varepsilon\,\|Wu\|_\infty.$$

Along a weakest singular direction $u_{\min}$, $\|Wu_{\min}\| = \sigma_{\min}(W)$, so $\|\phi(h + \varepsilon u_{\min}) - \phi(h)\|_{\mathrm{TV}} = O(\varepsilon\,\sigma_{\min}(W))$. In the limit $\sigma_{\min} \to 0$, this recovers the exact non-identifiability of Theorem 7.1.

Proposition 7.10 (Graded observability). *For the softmax-linear emission $\phi(h) = \operatorname{softmax}(Wh)$, let $\sigma_1 \ge \cdots \ge \sigma_r > 0$ be the nonzero singular values of W with corresponding right singular vectors $u_1, \ldots, u_r$. The per-observation Fisher information along direction u_k satisfies*

$$u_k^\top F(h)\, u_k \;\le\; \tfrac{1}{2}\,\sigma_k^2$$

for every h, with equality only when $p = \phi(h)$ is supported on exactly two tokens. Consequently, the observation effort (number of independent observations required to reduce posterior variance along u_k by a fixed factor) scales as $1/\sigma_k^2$: directions with small singular values require quadratically more observations to resolve.

Proof. Write $F(h) = W^\top C(p)\, W$ where $C(p) = \operatorname{diag}(p) - pp^\top$. Since u_k is the k-th right singular vector of W, $Wu_k = \sigma_k v_k$ where v_k is the corresponding left singular vector, so $u_k^\top F\, u_k = \sigma_k^2\, v_k^\top C(p)\, v_k \le \sigma_k^2\, \lambda_{\max}(C(p))$. By Gershgorin applied to $C(p) = \operatorname{diag}(p) - pp^\top$, the disc for row j is centred at $p_j - p_j^2$ with radius $\sum_{k \ne j} p_j p_k = p_j(1 - p_j)$, so each eigenvalue of $C(p)$ lies in $[0,\ 2\max_j p_j(1-p_j)]$. Using $p_j(1-p_j) \le 1/4$, this is contained in $[0, 1/2]$, so $\lambda_{\max}(C(p)) \le 1/2$. The observation-effort scaling follows from the Cramér–Rao bound applied to t independent observations: total Fisher information along u_k is at most $t\sigma_k^2/2$, so posterior variance $\ge 2/(t\sigma_k^2)$. □

Non-identifiability is therefore not binary (null space vs. image) but graded across the singular spectrum: directions in $\ker(W)$ are exactly invisible (Theorem 7.1), directions with small σ_k are practically invisible on realistic observation horizons, and even the strongest directions have bounded per-observation information ($\le \sigma_1^2/2$).

Proposition 7.11 (Emission channel capacity). *For the softmax-linear emission $\phi(h) = \operatorname{softmax}(Wh)$ with $\operatorname{rank}(W) = |V| - 1$, the Fisher information matrix for estimating h from a single observation $c \sim \phi(h)$ is*

$$F(h) = W^\top(\operatorname{diag}(p) - pp^\top)W, \qquad p = \phi(h),$$

with $\operatorname{rank}(F) \le |V| - 1$ and $\ker(W) \subseteq \ker(F)$. Consequently, $d - (|V| - 1)$ latent dimensions contribute zero Fisher information per observation.

Proof. The score is $\nabla_h \log \phi(h)(c) = W^\top(e_c - p)$. The Fisher information $F = \sum_c p_c\,(W^\top(e_c - p))(W^\top(e_c - p))^\top = W^\top(\operatorname{diag}(p) - pp^\top)W$. Since $\operatorname{diag}(p) - pp^\top$ has rank $|V| - 1$ at interior points (Proposition 3.1), $\operatorname{rank}(F) \le |V| - 1$. If $u \in \ker(W)$, then $Fu = 0$. □

Corollary 7.12 (Per-step information bound). *For the RIS recursion (1), the mutual information between the initial latent state and the first $t + 1$ observations satisfies*

$$I(H_0;\, C_{0:t}) \le (t + 1)\log|V|.$$

Proof. By the chain rule, $I(H_0; C_{0:t}) = \sum_{s=0}^{t} I(H_0; C_s \mid C_{0:s-1})$. Each term satisfies $I(H_0; C_s \mid C_{0:s-1}) \le H(C_s) \le \log|V|$. □

Remark 7.13 (Quantitative non-identifiability under anisotropic compression). Exact rank deficiency of W is not necessary for practical non-identifiability. If W has singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ with one or more singular values sufficiently small, then there exist distinct latent trajectories whose induced observation laws remain arbitrarily close over finite horizons. In particular, posterior uncertainty about the latent trajectory admits a lower bound controlled by the small singular spectrum of W. The required smallness threshold is not quantified here.

8 A Concrete Finite-State RIS

The preceding sections establish the three RIS theorem packages as abstract consequences of lossy emission and recursive re-injection. This section constructs an explicit finite-state RIS in which all three are simultaneously realized and directly verifiable. The example is illustrative only and does not claim to model any deployed system.

8.1 Model definition

Let the latent state space be $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, partitioned into three basins:

$$B_1 = \{1, 2\}, \qquad B_2 = \{3, 4\}, \qquad B_3 = \{5, 6\}.$$

The observation alphabet is $V = \{a, b, c\}$. The emission map $\phi\colon \mathcal{X} \to \Delta(V)$ assigns identical distributions to states within each basin:

State	$P(a)$	$P(b)$	$P(c)$
$1, 2$	$1 - 2\delta$	δ	δ
$3, 4$	δ	$1 - 2\delta$	δ
$5, 6$	δ	δ	$1 - 2\delta$

The re-injection map defines a token-dependent transition kernel $K(x' \mid x, c)$: given current state x and emitted token c, the next state is drawn from $K(\cdot \mid x, c)$. The kernel uses the following auxiliary quantities. Let $c_i \in V$ be the token associated with basin B_i (i.e., $c_1 = a$, $c_2 = b$, $c_3 = c$). For $x \in B_i$, define $\sigma(c) = +1$ if $c = c_i$ and -1 otherwise (a token-sign function relative to the basin of x). Let $\eta_{x,x'}$ be fixed perturbations ensuring within-basin non-degeneracy: $\eta_{x,x} = +0.1$ and $\eta_{x,x'} = -0.1$ for $x' \neq x$ within the same basin (without them, the two states in each basin would be dynamically identical and the kernel would reduce to a 3-state chain). For $x \in B_i$, define the within-basin normalization constant $Z(x, c) = \sum_{y \in B_i} (1 + \eta_{x,y}\, \sigma(c))$. The kernel has two components:

$$K(x' \mid x, c) = \begin{cases} \dfrac{(1 - \delta)\big(1 + \eta_{x,x'}\, \sigma(c)\big)}{Z(x, c)} & \text{if } x' \in B_i, \\[2ex] \dfrac{\delta}{|\mathcal{X}| - |B_i|} & \text{if } x' \notin B_i. \end{cases} \tag{14}$$

Here $Z(x, c)$ normalizes the within-basin mass to exactly $1 - \delta$, so each row sums to 1 (within-basin: $1 - \delta$, cross-basin: δ).

In this example, the primitive recursion is token-conditioned through $K(\cdot \mid x, c)$; for spectral analysis we work with the induced latent-state kernel $\bar{K}$ obtained by marginalizing over emitted tokens:

$$\bar{K}(x' \mid x) = \sum_{c \in V} \phi(x)(c)\, K(x' \mid x, c). \tag{15}$$

The cross-basin mass in (14) is $\delta/(|\mathcal{X}| - |B_i|)$ regardless of the emitted token c, so marginalizing over c preserves it exactly: for $x \in B_i$ and $x' \notin B_i$, $\bar{K}(x' \mid x) = \sum_c \phi(x)(c) \cdot \delta/(|\mathcal{X}| - |B_i|) = \delta/(|\mathcal{X}| - |B_i|)$. Since each row of $\bar{K}$ sums to 1 and the cross-basin mass is exactly δ (independent of the emitted token), the within-basin mass is $1 - \delta$ by complement. All spectral results from Section 4 (metastability, lumpability, slow-sector persistence) apply to $\bar{K}$.

8.2 Non-identifiability

States 1 and 2 have identical emission distributions:

$$\phi(1) = \phi(2) = (1 - 2\delta,\ \delta,\ \delta).$$

Under the marginalized kernel $\bar{K}$, for any observation history $c_{0:t}$ the likelihood $P(c_{0:t} \mid H_0 = 1) = P(c_{0:t} \mid H_0 = 2)$. (For the token-conditioned kernel $K(x'|x, c)$ rather than $\bar{K}$, multi-step persistence requires the transition-regularity condition of Remark 7.2.) To see the claim for $\bar{K}$, note that for any $x, x' \in B_i$: (i) $\phi(x) = \phi(x')$ (identical emission), and (ii) $\sum_{y \in B_j} \bar{K}(y \mid x) = \sum_{y \in B_j} \bar{K}(y \mid x')$ for every basin B_j (identical basin-to-basin masses). The one-step observation marginal $P(c \mid H_t = x) = \phi(x)(c)$ is therefore basin-invariant. Multi-step persistence follows because the basin-to-basin masses are state-independent within each basin: the marginalized kernel $\bar{K}$ is exactly lumpable, so the observation law $P(c_{s:t} \mid H_s \in B_i)$ depends only on i. The posterior $P(H_0 \mid C_{0:t} = c_{0:t})$ cannot distinguish between initial states 1 and 2. The same holds for the pairs $(3, 4)$ and $(5, 6)$.

This is a direct instantiation of Theorem 7.1: the emission map has a nontrivial fiber (each basin contains two states with the same emission), and the context-dependent within-basin transitions do not collapse these fibers. The six latent states project onto three observational equivalence classes, so the observation process is a 2-to-1 lossy factor of the latent process.

8.3 Metastability and spectral structure

The effective kernel $\bar{K}$ is a 6×6 stochastic matrix with the δ-separated structure of Definition 6.1. By Kemeny–Snell lumpability, the basin-level dynamics are governed by the 3×3 coarse kernel. Aggregating: for any $x \in B_i$, $\sum_{x' \in B_i} \bar{K}(x' \mid x) = 1 - \delta$ and $\sum_{x' \in B_j} \bar{K}(x' \mid x) = \delta/2$ for $j \neq i$ (the within-basin rows are basin-invariant in total mass by construction). Hence:

$$K_L(B_i, B_i) = 1 - \delta, \qquad K_L(B_i, B_j) = \delta/2 \quad (j \neq i).$$

In matrix form:

$$K_L = \begin{pmatrix} 1-\delta & \delta/2 & \delta/2 \\ \delta/2 & 1-\delta & \delta/2 \\ \delta/2 & \delta/2 & 1-\delta \end{pmatrix}. \tag{16}$$

The eigenvalues of K_L are $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = 1 - \frac{3}{2}\delta$, giving spectral gap $\frac{3}{2}\delta$. This is the $B = 3$ case of Corollary 4.4.

Table 2: Spectral gap of the 6×6 effective kernel $\bar{K}$ as a function of δ, computed from the coarse kernel (16). The gap tracks $\frac{3}{2}\delta$ across four decades, confirming Corollary 4.4.

δ	λ_2	gap	$\frac{3}{2}\delta$
10^{-1}	0.850	0.150	0.150
10^{-2}	0.985	0.015	0.015
10^{-3}	0.9985	0.0015	0.0015
10^{-4}	0.99985	0.00015	0.00015

As $\delta \to 0$, the three basins become almost-invariant: the system dwells in one basin for $\Theta(1/\delta)$ steps before escaping. The second eigenvector of $\bar{K}$ recovers the basin partition (states in B_1 have positive entries; states in $B_2 \cup B_3$ have negative entries), consistent with the basin decomposition of Section 4.

The within-basin eigenvalues are $O(\eta)$ with $\eta = 0.1$ (Section 8, model definition), so $O(\eta) \ll 1 - \delta$ for small δ, well separated from the near-unit basin-level eigenvalues. This spectral separation is the structural condition under which approximate lumpability (Theorem 4.10) guarantees persistence of the slow sector.

8.4 Typical-set separation

The mixture predictor $P^* = \frac{1}{3}(P_1+P_2+P_3)$, where P_i denotes the process initialized in basin B_i, satisfies $D_{\text{rate}}(P^*\|P_i) > 0$ for each i (Lemma 6.3). The mechanism is partition-level mass mismatch: P^* spreads probability across all three preference sets, while each P_i concentrates mass $(1-2\delta)$ on its own set S_i. By Theorem 6.11, the P^*-typical set carries probability tending to 1 under P^* and exponentially vanishing probability under each P_i, with good-subset exponent $D_{\text{rate}}(P^*\|P_i)$.

8.5 Summary

This single model simultaneously exhibits:

- non-identifiability (two latent states per observed symbol, non-collapsing fibers, $\phi(1) = \phi(2)$ explicitly);
- spectral metastability (three near-unit eigenvalues, coarse kernel (16), gap $= \frac{3}{2}\delta$, basin dwell time $\Theta(1/\delta)$);
- typical-set separation (mixture-typical sequences exponentially unlikely under each component, driven by partition-level mass mismatch).

The model is small enough to be fully transparent (a 6×6 kernel with a displayed 3×3 coarse kernel) yet rich enough to instantiate the paper's three main theorems. The assumptions are jointly satisfiable, not merely individually consistent. Since emissions are identical within each basin ($\varepsilon = 0$) and the preference-set partition gives token-block alignment, the model also satisfies Proposition 4.13 and hence Corollary 4.14: both typical-set separation and metastability hold on the same partition. This section establishes joint satisfiability of the paper's assumption packages in one explicit model; it does not claim that such coexistence is typical.

8.6 Typed separation existence

Proposition 8.1 (Typed separation can diverge)**.** *There exists a finite-state RIS in which* δ_{pred}*,* δ_{ker}*, and* δ_{geom} *(Definition 4.2) are pairwise distinct and all positive.*

Proof. Modify the model of Section 8 by replacing the disjoint emission with overlapping emissions: for states in basin B_i, let state $2i-1$ emit with distribution $(1-2\omega+\sigma,\ \omega-\sigma/2,\ \omega-\sigma/2)$ (sharper) and state $2i$ emit with $(1-2\omega-\sigma,\ \omega+\sigma/2,\ \omega+\sigma/2)$ (flatter), where $\omega \in (0, 1/3)$ controls overlap and $\sigma > 0$ controls within-basin emission variation. The transition kernel is unchanged from Equation (14).

For $\omega = 0.2$, $\sigma = 0.08$: $\delta_{\text{ker}} = 0.1$ (unchanged cross-basin transition rate), $\delta_{\text{pred}} = 0.42$ (mean cross-basin emission TV), $\delta_{\text{geom}} = 0.28$ (minimum cross-basin emission TV). All three are positive and pairwise distinct (verified by SymPy, Appendix A). □

Remark 8.2. In the original model (Section 8) the preference sets are disjoint and all three parameters coincide. The overlapping-emission variant demonstrates that the typed separation parameters can diverge once the disjoint-support assumption is relaxed. This divergence is structural, not pathological: it should be expected in any RIS where emission supports overlap across basins.

9 Discussion

9.1 Nature of the contribution

The mathematical tools used in this paper are classical; the contribution is to define a single operator class in which their consequences can be stated jointly under explicit assumptions.

The paper gives theorem-level results with explicit assumptions and a coexistence model showing that the three RIS phenomena arise from one common recursive mechanism. Recursive emission and re-injection are necessary but not sufficient: each phenomenon also requires its own structural assumption (rank deficiency for non-identifiability, δ-separation for typical-set separation, lumpability for metastability). Recursion is the coupling that lets these otherwise-distinct phenomena co-arise within one operator class. The coexistence result is existential, not generic. Whether naturally arising systems satisfy these assumptions, even approximately, is an empirical question not addressed here.

The unification is operational, not merely taxonomic: a candidate system can be jointly tested against all three predictions listed below, and the pattern of which fail localizes the structural defect (rank, separation, or lumpability) — a diagnostic granularity not available when the three phenomena are analyzed in separate frameworks.

Relation to prior work. The individual ingredients each have a clear ancestor literature. Spectral metastability via eigenvalue clusters traces to Davies [3] and Gaveau–Schulman [5]; conformation-dynamics formulations and PCCA+ are developed in Schütte–Sarich [16] and Deuflhard–Weber [4]. HMM identifiability is classical [10, 15] with spectral algorithmic recovery in [7]; parameter identifiability under many observations is treated in [1]. Within RIS, these tools are applied to a single operator class so that their consequences can be stated and falsified jointly, with Theorem 6.16 establishing that recursive divergence accumulation is not reducible to any of the three standard sufficient conditions.

9.2 What the framework constrains

The framework constrains what can be inferred from observable outputs of a RIS. Theorem 7.1 establishes non-injectivity of the latent-to-observation map under rank deficiency and reachability alone, with no finite-state or separation hypotheses; multi-step persistence is handled separately by Remark 7.2. The paper identifies what follows once lossy recursive emission is combined with the stated separation and lumpability structure.

The framework does *not* constrain what exists internally. External non-identifiability is proved; extending non-identifiability to internal queries (e.g., whether an internal observer could distinguish latent states via the same lossy projection) would require additional assumptions about internal access not formulated here.

9.3 Framework predictions

The assumption packages of Theorem 1.4 yield concrete falsifiable predictions for any system claimed to instantiate RIS:

- $D_{\mathrm{rate}}(P^* \| P_i) > 0$ under δ-separation with $\delta < 1/2$ and $B \geq 2$ (Lemma 6.3).
- Spectral gap $\Theta(\delta)$ for the lumped chain under exact lumpability (Theorem 4.3).
- Loop-driven divergence accumulation does not require any of $\Delta_i > 0$, hard token partition, or geometric basin separation (Theorem 6.16).

A candidate RIS instance failing any of these predictions falsifies its membership in the operator class, not the framework itself.

9.4 Scope and limits

The following are within the scope of proved results:

- Non-identifiability under lossy emission (Theorem 7.1);
- Typical-set separation under mixture iteration (Theorem 6.11);

- Spectral metastability in exact and approximate lumpability regimes (Section 4).

The following are outside scope:

- Behavior of specific deployed models;
- Infinite-dimensional state spaces (all results are finite-dimensional; the sliding-window approximation $X = V^k$ is discussed in Section 1);
- Coupled RIS and network topologies (not treated here);
- Consciousness, subjective experience, or phenomenal qualities.

9.5 Future directions

Several extensions are natural but not pursued here.

Companion formalization papers. Three follow-up directions are explicitly carved out: (i) Cheeger inequality formalization with full Lean proof (current Lean file is statement-level with one acknowledged `sorry` and verified helper lemmas; full proof in preparation); (ii) approximate-lumpability and bridge-theorem proofs in a dedicated companion paper (current Lean file is statement-level with classical citations; analytic proofs in preparation); (iii) mapping a class of trained autoregressive models into the RIS predicates via a measure-theoretic extension of the framework (beyond current Lean infrastructure).

Quantitative Loop Closure independence. Theorem 6.16 establishes qualitative independence: recursive divergence accumulation is not implied by any one of $\Delta_i > 0$, hard token partition, or geometric basin separation. A natural strengthening is quantitative: characterizing the rate at which Loop Closure divergence accumulates as a function of the margin by which each of the three sufficient conditions fails. The current witnesses are finite systems verified in Lean by decision procedures; a margin-indexed family with a rate function $g(\alpha)$ remains open.

Approximate lumpability. Theorem 4.10 establishes that the slow sector persists under approximate lumpability with quantitative projector and eigenvalue bounds, provided the fast spectrum is separated from the slow cluster. The remaining open question (Conjecture 4.15) is whether the $O(B\varepsilon)$ eigenvalue bound is sharp in its dependence on B.

Emergence without hard separation. The δ-separated regime is a sufficient condition for Theorem 6.11, but the proof's essential ingredients are positive divergence rate (assumption (a)) and cross-entropy separation (assumption (b)). Proposition 6.5 provides a partial answer: positive D_{rate} is guaranteed whenever the component predictors have positive average total variation, without any δ-separation or partition assumption. The bound $D_{\text{rate}} \geq 2\varepsilon^2$ shows that δ-separation is sufficient but not necessary for positive divergence rate. The full typical-set separation theorem (Theorem 6.11) additionally requires cross-entropy separation; whether this can also be relaxed under overlap-based conditions remains open.

Self-sustaining recursion (endogenous commitment feedback). When the commitment chain supplies its own input via a self-prompting function $g\colon \mathcal{H} \to V$ (selecting the next input token from the current latent state without external prompting), the recursion becomes $H_{k+1} = M(H_k, g(H_k))$. Conditions under which such a closed loop is non-degenerate and stable are a natural question for the RIS framework.

Coupled RIS. When two or more RIS interact through exchanged commitments, the relevant object is a joint operator on a product space. Symmetric all-to-all coupling yields exact closed-form eigenvalues; general graph structures with heterogeneous coupling remain open.

Empirical operationalization. Measuring δ-separation, detecting metastable basins, and testing non-identifiability bounds in concrete systems would connect the formal results to observable phenomena; this remains open. Table 3 provides a heuristic diagnostic checklist mapping each theorem's assumptions to candidate observable proxies. The checklist is graded, not binary: real

Table 3: **Heuristic** diagnostic checklist for candidate RIS instantiations (not a validation protocol). For each regime, the table lists the formal hypothesis and a candidate observable proxy. This is a graded bridge to empirical testing, not a binary certification. These proxies are heuristic operationalizations for empirical work, not theorem-equivalent conditions.

Item	**Hypothesis**	**Observable proxy**
Tier A — Structural fit (RIS entry gate)		
Latent state	$\mathcal{H} \subseteq \mathbb{R}^d$	Identifiable internal-state surrogate
Lossy emission	$\phi\colon \mathcal{H} \to \Delta(V)$, non-invertible	Many-to-one output compression
Re-injection	$f\colon \mathcal{H} \times V \to \mathcal{H}$	Emitted output fed back as input
Tier B — Theorem-specific proxies (per-result gate)		
Non-identifiability Theorem 7.1	$\dim \ker(W) > 0$, reachability	Posterior dimensionality after conditioning on token history
Metastability Theorem 4.10 and Proposition 4.13	Approx. lumpability, slow/fast spectral separation; predictive coherence as one route to (b)	Long dwell times, clustered almost-invariant sets
Typical-set separation Theorem 6.11	Positive divergence rate and cross-entropy separation; δ-separation is one sufficient proxy regime	Trajectory-level log-likelihood-ratio growth: $D_{\text{rate}}(P^* \| P_i) > 0$

systems will typically show approximate evidence for some regimes, clear failure for others, and unknown status for the rest.

Table 3 operationalizes the assumption structure for empirical work while remaining agnostic about whether any particular system satisfies the conditions; it identifies what to measure, not what is true of any given architecture.

10 Conclusion

This paper defines Recursive Interaction Systems (RIS) as a class of non-invertible dynamical systems with observation-induced projection and establishes three theorem-level consequences of recursive lossy commitment within one operator framework, exhibiting them jointly in a single explicit RIS: non-identifiability of latent trajectories (Theorem 7.1), typical-set separation under sequential mixture dynamics (Theorem 6.11), and spectral metastability in exact and approximate lumpability regimes (Theorems 4.3 and 4.10).

Emission collapse induces latent equivalence classes; in the explicit separated/lumpable regimes studied here, coarse partitions compatible with those equivalence classes can also support metastable dynamics. The explicit finite-state RIS of Section 8 exhibits all three phenomena in one explicit model (see Table 1).

The claim is not that every RIS exhibits every phenomenon, but that lossy emission and recursive re-injection bring these phenomena under common mathematical control: exactly in solvable regimes, perturbatively under approximate lumpability, and jointly in one explicit model. Typed separation parameters (Definition 4.2) track which structural layer each result depends on; a soft divergence-rate bound (Proposition 6.5) extends positive D_{rate} beyond the δ-separated regime; and spectral recovery methods (Propositions 4.18 and 4.20) connect the coarse kernel's gap to the full kernel and to data-driven slow-mode identification. RIS identifies a minimal recursive architecture in which information loss, sequence-level separation, and slow coarse dynamics are not separate curiosities but linked structural consequences.

A Symbolic Verification

Selected algebraic identities and explicit finite-dimensional model claims used in the paper were checked symbolically (SymPy) and numerically (NumPy). These computations are supplementary algebraic and numerical sanity checks on explicit finite-dimensional instances; they do not replace the mathematical arguments in the paper text. These checks validate explicit formulas and examples; they do not serve as independent proofs of the abstract theorems.

Scope of verification

The symbolic computation operates on explicit finite-dimensional models. The principal test models use $B \in \{2, 3, 4, 5\}$ preference domains, vocabularies $|V| \in \{4, \ldots, 15\}$, context windows $k \in \{1, 2, 3\}$, and separation parameters $\delta \in (0.01, 0.4)$. Approximate-lumpability tests use stochastic perturbations of exactly lumpable kernels with perturbation magnitude $\varepsilon \in (0.001, 0.06)$. All symbolic checks use exact SymPy arithmetic; all numerical checks use vectorized NumPy with `seed` $= 42$ for reproducibility. Tested with Python 3.12, NumPy 2.3, SymPy 1.13; the source is included in the Zenodo record.

Symbolic checks performed

The following classes of identity were verified (170 checks total; details in the supplementary SymPy code):

1. *Supporting propositions* (Section 3): softmax shift invariance and rank $|V| - 1$ (exact symbolic), fiber characterization $\text{softmax}^{-1}(p) = \{\log p + \alpha \mathbf{1}\}$ (exact symbolic), TV contraction under stochastic kernels (5000 random pairs, 0 violations), DPI composition contraction, finite-state periodicity, argmax switching boundary count $\binom{|V|}{2}$, self-conditioning contraction criterion.

2. *Operator theory* (Section 4): Markov regimes (ergodic, periodic, metastable) via exact symbolic eigenvalues; $B = 2$ metastable eigenvalue $\lambda_2 = 1 - 2\delta$ and spectral gap (exact symbolic); exact slow-sector spectrum $1 - B\delta/(B-1)$ for $B = 3, 4$ (exact symbolic); basin decomposition via second-eigenvector sign (numerical, up to 16×16); exact stability region for linearized self-conditioning operator (exact symbolic); non-normal transient growth peak and decay for Jordan blocks (exact symbolic).

3. *Approximate lumpability* (Theorem 4.10): within-domain spectral bound $\rho(K|_{S_i \times S_i}) \leq 1 - \delta$ (20 random kernels); perturbative persistence of slow eigenvalues under ε-perturbation with shift $\leq \kappa\varepsilon$ (50 random trials); projector stability: subspace angle between slow eigenvectors and block-constant subspace scales as $O(\|QK\Pi\|)$ (8-point ε-sweep, ratio range $[0.6, 0.7]$); invariance defect $\|QK\Pi\|_\infty \leq B\varepsilon$ (40 random trials); macro spectrum identification $\sigma(\Pi K \Pi|_U) = \sigma(\bar{K})$ (30 random kernels); slow-sector eigenvalue tracking under approximate lumpability (40 random trials); predictive coherence bridge (Proposition 4.13): lumpability error $\leq \varepsilon + 2\eta$ verified across 30 random kernels with varying B, block size, and coherence/leakage parameters.

4. *Mixture and emergence* (Sections 5 and 6): JSD bounds $0 \leq \text{JSD} \leq \log B$ (numerical, including disjoint-support and identical-component extremes); sequential amplification ratio $(1/(2\sqrt{\varepsilon(1-\varepsilon)}))^n$ (exact symbolic); token-level support containment and sequence-level amplification (numerical); constructive emergence threshold and exponential ratio (exact symbolic + numerical); typical-set separation $P^*(T_n^\varepsilon) \to 1$, $\sup_i P_i(T_n^\varepsilon) \to 0$ exponentially (numerical, multiple chain lengths and parameter regimes); cross-entropy rate decomposition $H_{\text{rate}}(P^*, P_i) = h(P^*) + D_{\text{rate}}(P^* \| P_i)$ (numerical); Shannon–McMillan–Breiman con-

vergence (numerical); Perron–Frobenius stationary distribution and Dobrushin contraction coefficient (numerical).

5. *Non-identifiability* (Theorem 7.1): fiber dimension $\dim\ker(W) > 0$ for $d = 8$, $|V| = 4$; emission invariance $\|\phi(h+u) - \phi(h)\| < 10^{-15}$ for $u \in \ker(W)$; illustrative posterior non-collapse in sampled finite models (posterior entropy $H(H_0 \mid C_{0:t}) > 0$ confirmed numerically across sampled observation histories); Fisher information rank (Proposition 7.11): $\mathrm{rank}(F) = |V| - 1$ and $\ker(W) \subseteq \ker(F)$ verified for $d = 8$, $|V| = 4$.

6. *Typed separation and extensions (v2)*: typed separation existence on a modified concrete model with overlapping emission supports ($\delta_{\mathrm{pred}} \neq \delta_{\ker} \neq \delta_{\mathrm{geom}}$, all positive); soft D_{rate} bound $D_{\mathrm{rate}}(P^* \| P_i) \geq 2\varepsilon^2$ verified on the overlapping-emission model; TICA slow-mode recovery: eigenvalue tracks λ_2^τ (ratio 1.02) and slow eigenvector aligns (cosine 0.997) on the concrete $B = 2$ model; first-order gap cancellation under lumpability-breaking perturbation ($|\mathrm{gap}(K) - \mathrm{gap}(K_L)| \ll \varepsilon$, at machine precision for the tested perturbation family).

7. *ICC axioms and pipeline*: all 7 ICC axioms verified on the generation pipeline (mixed symbolic/numerical); bilevel training-vs-recursion separation; three-stage compression chain measurements.

Limitations

The symbolic checks operate on representative finite-dimensional models, not on the abstract theorem statements directly. They confirm that the algebraic identities hold exactly in the scoped models and that the structural patterns (spectral decomposition, projector stability, entropy-rate separation, emission fiber invariance) are present as claimed. They do not constitute a proof of the general theorems, which are established by the analytic arguments in the main text.

The approximate-lumpability checks (Theorem 4.10 and Proposition 4.7) use randomized stochastic perturbations of exactly lumpable kernels. In all tested regimes (200+ trials across B, δ, ε parameter combinations), no violations of the stated bounds were observed. This is empirical support only and does not strengthen the analytic theorems.

The SymPy source code is included as a supplementary file in the Zenodo record accompanying this paper.

B Lean Formalization Artifacts

Eleven Lean 4 source files accompany this paper, formalized against the Mathlib4 library (v4.29.1) in the workspace

`~/lean-workspace/icc-formalization/IccFormalization/RIS/`

The entire workspace builds cleanly with `lake build` and introduces no new `axiom` beyond Lean's core. Five files carry fully proved results with zero `sorry`; one file carries five fully proved supporting lemmas plus two acknowledged `sorry`s (on the main theorem and its corollary), both isolated to the same mathlib4 `Fin`-bijection gap; three files carry type-checked statements with proofs paused pending dedicated formalization papers (Cheeger and approximate-lumpability bridge, both in preparation) and the corresponding mathlib4 infrastructure they require; one file (`Common.lean`) provides shared infrastructure (no theorems); and one file (`SmokeTests.lean`) provides concrete verified instances exercising the statement-only definitions (detailed per file below). Statement-level Lean grounding ensures the type structure of the framework's claims is mechanically verified. The mathematical content of statement-only claims is established either

by classical results cited in the body (Cheeger's inequality, from Levin–Peres [12]) or by the framework's own analytic proofs in the main text (the approximate-lumpability bridge theorem, Theorem 4.10). Lean re-proof from primitives is deferred to dedicated companion work; the truth of the statements is not in question.

The Lean source is deposited alongside this paper in the accompanying Zenodo record.

Fully proved files

`PositiveDivergence.lean` (342 lines). Formalizes Lemma 6.3: under δ-separation with $\delta < 1/2$ and $B \geq 2$, the relative entropy rate satisfies $D_{\text{rate}}(P^* \| P_i) > 0$ for every i. The proof closely mirrors the textbook argument. Gibbs' inequality for the discrete sum $\sum_x p(x) \log(p(x)/q(x)) \geq 0$ is derived directly from the standard analytic bound $\log x \leq x - 1$ (`Real.log_le_sub_one_of_pos`), avoiding the measure-theoretic `klDiv` infrastructure. Key lemmas proved from scratch: `gibbs_term` (pointwise bound $p - q \leq p \log(p/q)$), `gibbs_term_strict` (strict inequality when $p \neq q$), `discreteKL_nonneg` and `discreteKL_pos` (global discrete Gibbs' inequality and its strict form). The mixture-vs-component separation (`mixture_ne_component`) reproduces the paper's contradiction argument: if $P^* = P_i$ then the normalized mixture Q_i of non-i components must equal P_i, but $Q_i(S_i|x) \leq \delta < 1 - \delta \leq P_i(S_i|x)$. The D_{rate} positivity theorem (`drate_pos`) then follows by integrating the pointwise KL against a full-support stationary measure. No `sorry`; proof is complete.

`TypedSepExample.lean` (521 lines). Formalizes Proposition 8.1: there exists a finite-state RIS in which δ_{pred}, δ_{ker}, and δ_{geom} are pairwise distinct and all positive. The witness is the overlapping-emission construction of Section 8 with $\omega = 1/5$, $\sigma = 2/25$ (matching the paper's $\omega = 0.2$, $\sigma = 0.08$ in exact rationals). The six-state, three-token emission table is defined as a computable function `emission` over `Fin 6 × Fin 3 → ℚ`. The lumped coarse kernel K_L from Equation (16) is formalized as `lumpedKernel` over `Fin 3 → Fin 3 → ℚ`, with row-sum and nonnegativity lemmas proved by `norm_num`. Total-variation distances are computed as rational sums; all arithmetic is discharged by `norm_num`. Proved concretely: $\delta_{\text{geom}} = 7/25$ (minimum cross-basin emission TV, achieved by the flat-flat pair), and all pairwise cross-basin TV values ($11/25$ or $13/25$ for sharp-flat and sharp-sharp pairs, $7/25$ for flat-flat). $\delta_{\text{ker}} = 1/10$ is *computed* from `lumpedKernel` as $\max_i \sum_{j \neq i} K_L(i,j)$, with `deltaKer_eq` establishing the value by direct calculation. $\delta_{\text{pred}} = 21/50$ is computed as the uniform average of the four cross-basin emission-TV values ($13/25$, $11/25$, $11/25$, $7/25$) under the uniform stationary distribution $\pi^* = (1/6, \ldots, 1/6)$, which holds by basin symmetry of the kernel. The strict lower bound $\delta_{\text{pred}} > \delta_{\text{geom}}$ (i.e. $21/50 > 7/25$) is proved directly by `norm_num`. The existence theorem `typed_sep_example_exists` witnesses the emission table and lumped kernel explicitly, binding all three parameters to the model rather than asserting free numeric literals. The three values $7/25 = 14/50$, $1/10 = 5/50$, $21/50$ are provably distinct and positive by `norm_num`. No `sorry`; proof is complete.

`DrateIdentity.lean` (122 lines). Formalizes the per-step mean KL equivalence underlying the relative entropy rate (Lemma 6.3):

$$\frac{1}{n} D_{\text{KL}}(P^*_{1:n} \,\|\, P_{i,1:n}) \;=\; \frac{1}{n} \sum_{t=1}^{n} D_{\text{KL}}(P^*(\cdot \mid x_{<t}) \,\|\, P_i(\cdot \mid x_{<t})).$$

The proof proceeds by the chain rule of KL divergence applied to a finite product distribution over the sequence $(x_1, \ldots, x_n)$. Because mathlib4's built-in `klDiv` is measure-theoretic (Radon-Nikodym, $\mathbb{R}^\infty_{\geq 0}$-valued), the file defines a small `discreteKL` directly as a finite sum and proves the additivity identity algebraically via `ring`/`simp`. No `sorry`; proof is complete.

`LoopClosureIndependence.lean` (716 lines). Proves Theorem 6.16: three finite countermodels establishing that Loop Closure behavior is independent of each of the three RIS sufficient conditions. The file defines:

- `HasLoopClosure`: two states with distinct emission supports and feedback-sustaining separation;
- `IsRISDeltaPositive` (entropy-rate gap discretization), `IsRISHardPartition` (disjoint supports), and `IsRISBasinGeometric` (unique-token identifier for some state).

Three witnesses (`WitnessA`, `WitnessB`, `WitnessC`) are defined over `Fin 2 × Fin 2` or `Fin 2 × Fin 3` state-alphabet pairs; all claims are discharged by `decide` or short tactic proofs. No `sorry`; all twelve failure/closure theorems are complete.

`EmergentFamily.lean` (243 lines). Encodes Definition 6.7 as a typed Lean structure. The definition formalizes the emergent family E_n as the set of all length-n sequences $c\colon \mathrm{Fin}\, n \to \alpha$ satisfying the cycling constraint $c_k \in S_{k \bmod B}$ for each position k, where $S_0, \ldots, S_{B-1}$ are the pairwise disjoint preference sets of the δ-separated expert family (Definition 6.1). The `EmergentFamily` structure bundles: the state-space type α, the domain count B with positivity witness, the preference sets `pref`$\colon \mathrm{Fin}\, B \to \mathrm{Finset}\, \alpha$, and a disjointness proof `pref_disj` (matching the δ-separation axiom). The design is intentionally parametric in n: a downstream theorem takes `(F : EmergentFamily α B)` and `(n : ℕ)` separately, then refers to `F.set n`. Supporting lemmas proved: `mem_emergentSet_iff` (membership characterization), `emergentSet_zero_eq_singleton` (E_0 is the unique empty sequence), and `domainIndex_periodic` (the cycling index is B-periodic, i.e. $\mathrm{domainIndex}(k+B) = \mathrm{domainIndex}(k)$). A concrete smoke test (`SmokeTest` namespace) instantiates the structure for $\alpha = \mathrm{Fin}\, 2$, $B = 2$, with $S_0 = \{0\}$ and $S_1 = \{1\}$. It verifies: disjointness of the preference sets; membership of the sequence $(0, 1)$ in E_2; and that every member of E_2 must place state 0 at position 0 and state 1 at position 1 (uniqueness of the cycling sequence at length 2). This is a definition encoding, not a theorem proof; the file demonstrates type-correct instantiability for downstream use in Theorem 6.8. No `sorry`; all proofs are complete.

Statement-only files (proofs deferred)

`Cheeger.lean` (336 lines). Formalizes the conductance-based metastability bound of Remark 4.16. Defines stochastic kernels, stationary distributions, cross-block flow, and conductance over a finite state space. States four main theorems: `cheeger_upper` ($\mathrm{gap}(K) \le 2\Phi$), `cheeger_lower` ($\Phi^2/2 \le \mathrm{gap}(K)$), `cheeger_sandwich` ($\Phi^2/2 \le \mathrm{gap} \le 2\Phi$), and `small_conductance_implies_`
Two supporting lemmas are fully proved: `crossBlockFlow_nonneg` and `conductance_nonneg`. The main theorems carry `sorry`: mathlib4 v4.29.1 lacks the discrete Cheeger cutting lemma, the Rayleigh-quotient characterization of spectral gap for finite stochastic matrices, and the Hermitian-symmetrization route needed for reversible kernels.

`BridgeTheorem.lean` (358 lines). Formalizes the approximate lumpability bridge theorem (Theorem 4.10). Defines block projections, the macro kernel, the invariance defect, and the `ApproxLumpable` predicate. States three supporting lemmas (`macro_spectrum_identification`, `invariance_defect_bound`, `projector_stability`) and the main theorem (`approx_lump_preserves_slow_s`
All carry `sorry`: required infrastructure (block-matrix operator theory, Riesz projectors via contour integration, and non-Hermitian eigenvalue continuity) is absent from mathlib4 v4.29.1. Three encoding subtleties from this file are documented in Remark 4.11.

`WithinDomain.lean` (338 lines). Formalizes Lemma 4.5: for a δ-separated stochastic kernel, every eigenvalue λ of the within-block restriction $K|_{S_i \times S_i}$ satisfies $|\lambda| \leq 1 - \delta$. The file defines the within-block column sum $\sum_{y \in S_i} K(y, x)$, the outflow mass $\delta(x) = \sum_{y \notin S_i} K(y, x)$, and the ℓ^1 (column-sum) operator norm of the restriction. Five supporting lemmas are fully proved: `withinBlockColSum_add_outflow_eq_one` (the two quantities partition the total column mass to 1), `withinBlockColSum_eq` ($= 1 -$ outflowMass), `withinBlockColSum_nonneg`, `withinBlockColSum_le_one`, and `withinBlockL1Norm_eq_one_sub_delta` (under uniform δ-separation, the ℓ^1 norm of the restriction equals $1-\delta$). The main theorem `within_domain_spectral_bound` and its corollary `within_domain_contractive` carry `sorry`. The blocker is two-part: (i) coercing the within-block restriction from the functional kernel representation to `Matrix (Fin k) (Fin k)`
`ensuremath`
`mathbbR` via `Finset.orderIsoOfFin` (approximately 50 lines of index-enumeration boilerplate); (ii) establishing the bound $\rho(A) \leq \|A\|_{\text{col1}}$ for the column-sum matrix norm, which requires defining a custom sub-multiplicative norm and applying `spectralRadius_le_of_norm_pow_le` (approximately 100 lines of norm infrastructure not assembled in mathlib4 v4.29.1). The Tier A portion—the column-sum computation establishing $\|K|_{S_i}\|_{\text{col1}} = 1 - \delta$—is fully proved and is the load-bearing algebraic content of the lemma.

Concrete instance checks

`SmokeTests.lean` (196 lines). Five concrete verified instances: a two-state ergodic chain (lumpable, positive gap), a two-basin metastable chain, the concrete RIS of Section 8 (exact lumpability and spectral gap $\frac{3}{2}\delta$), a non-lumpable kernel with small conductance, and a degenerate absorbing chain. All five are closed proofs with zero `sorry`.

Scope disclaimer

The Lean predicates are stated over finite discrete recursive systems. No claim is made that any autoregressive language model satisfies these predicates. The independence result (Theorem 6.16) is a structural statement about the abstract framework. The Cheeger and bridge-theorem files establish only that the definitions are type-correct and the theorem statements are well-formed; they do not constitute proofs of those results.

Revision History

Ver.	Date	Changes
v1	March 23, 2026	Initial release (title: "Non-Identifiability, Typical-Set Separation, and Metastability").
v2	March 28, 2026	Added Definition 4.2, Proposition 6.5, Proposition 4.18, Proposition 4.20, Proposition 8.1. Revised §1.2, §1.3, §1.4, §1.5, §2.3, §2.4, §4, §6, §7, §8, §9, §10, Table 1, Table 3, Figure 3, Appendix A. SymPy checks 147 → 170.
v2.1	March 30, 2026	Added Remark 4.16 (metastability from conductance without lumpability), Proposition 7.10 (graded observability via singular spectrum), Remark 4.19 (TICA recovers observable coarse structure). Corollary 4.4: noted asymmetric eigenvalue non-degeneracy. Remark 7.2: clarified transition regularity as sufficient, not unique. Definition 6.1: noted k-th order extension via state-space expansion. Updated Table 1, Table 3, metastability roadmap. Small wording: abstract, §2.3.
v3.0	May 4, 2026	Formal-grounding pass. Added Appendix B (Lean 4 + mathlib4 v4.29.1 formalization artifacts: eleven files, five fully proved with zero `sorry`; one file with five proved supporting lemmas plus two acknowledged `sorry`s on a single mathlib4 `Fin`-bijection gap; three files statement-only pending dedicated formalization papers; `Common.lean` and `SmokeTests.lean` as shared infrastructure and integration probes). Added Theorem 6.16 (Loop Closure independence of the three RIS sufficient conditions; Lean-verified via three finite witnesses). Added Remark 4.11 (three encoding subtleties from Theorem 4.10 formalization). Added §1.5 (formal-verification disclosure). No existing theorems changed or removed.

References

[1] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. "Identifiability of Parameters in Latent Structure Models with Many Observed Variables". In: *The Annals of Statistics* 37.6A (2009), pp. 3099–3132.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd. Wiley-Interscience, 2006.

[3] E. B. Davies. "Metastable States of Symmetric Markov Semigroups I". In: *Proceedings of the London Mathematical Society* s3-45.1 (1982), pp. 133–150.

[4] Peter Deuflhard and Marcus Weber. "Robust Perron Cluster Analysis in Conformation Dynamics". In: *Linear Algebra and its Applications* 398 (2005), pp. 161–184.

[5] Bernard Gaveau and L. S. Schulman. "Theory of Nonequilibrium First-Order Phase Transitions for Stochastic Dynamics". In: *Journal of Mathematical Physics* 39.3 (1998), pp. 1517–1533.

[6] Peter W. Glynn and Dirk Ormoneit. “Hoeffding’s Inequality for Uniformly Ergodic Markov Chains”. In: *Statistics & Probability Letters* 56.2 (2002), pp. 143–146.

[7] Daniel Hsu, Sham M. Kakade, and Tong Zhang. “A Spectral Algorithm for Learning Hidden Markov Models”. In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1460–1480.

[8] Claudio Irrgang. *A No-Go Theorem for Deterministic Born Recovery from Branch-Conditioned Records*. 2026. DOI: 10.5281/zenodo.19101873.

[9] Claudio Irrgang. *Irreversible Compression Commitment: A Typed Axiomatic Framework for Evolution, Extraction, and Commitment*. 2026. DOI: 10.5281/zenodo.19004432.

[10] Hisashi Ito, Shun-ichi Amari, and Kingo Kobayashi. “Identifiability of Hidden Markov Information Sources and Their Minimum Degrees of Freedom”. In: *IEEE Transactions on Information Theory* 38.2 (1992), pp. 324–333.

[11] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. 2nd. Springer, 1976.

[12] David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. 2nd. American Mathematical Society, 2017.

[13] Jianhua Lin. “Divergence Measures Based on the Shannon Entropy”. In: *IEEE Transactions on Information Theory* 37.1 (1991), pp. 145–151.

[14] Daniel Paulin. “Concentration Inequalities for Markov Chains by Marton Couplings and Spectral Methods”. In: *Electronic Journal of Probability* 20 (2015), pp. 1–32. DOI: 10.1214/EJP.v20-4039.

[15] Ted Petrie. “Probabilistic Functions of Finite State Markov Chains”. In: *The Annals of Mathematical Statistics* 40.1 (1969), pp. 97–115.

[16] Christof Schütte and Marco Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*. Vol. 24. Courant Lecture Notes. American Mathematical Society, 2013.

[17] Eugene Seneta. *Non-negative Matrices and Markov Chains*. 2nd. Springer, 2006.

[18] Claude E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27 (1948), pp. 379–423.