GRADIENT-BASED LEARNING OF MINIMAL PREDICTIVE REALIZATIONS FROM FINITE HANKEL BLOCKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study gradient-based learning of linear recurrent models in an overparameterized hidden space from finite Hankel matrices, whose entries are scores of concatenated prefixes and suffixes. We assume that the target has finite Hankel rank and observe a base matrix attaining this rank together with its one-symbol shifts. Exact fitting of the base and one-symbol Hankel blocks through a decoder does not guarantee correct multistep predictions with unconstrained transitions. We construct a training procedure that updates state, decoder, and transition parameters concurrently, with each loss restricted to its corresponding parameter group. We prove global convergence from any full-rank state initialization and arbitrary finite initializations of the remaining parameters. The limiting state matrix has the target Hankel rank, and removing redundant dimensions yields a model with the minimum possible state dimension that predicts every finite sequence exactly. For every fixed positive feedback strength, we exhibit a normally attracting manifold of non-global local minima with an open basin containing full-rank initializations. A computable switching certificate then combines global triangular training with convergent fully coupled updates. We also prove deterministic fixed-step convergence under sufficient step-size conditions and derive fixed-length prediction-error bounds for least-squares reconstructions from perturbed Hankel matrices. Experiments on finite-rank weighted sequence systems recover the target rank in all 90 state-formation runs, demonstrate failures of one-step fitting, verify local attraction and perturbation stability, reproduce positive-feedback collapse with matched triangular success, and show that 480 of 540 fixed-step runs meet the numerical criterion within budget.

1 INTRODUCTION

Sequential learning systems must compress an expanding observation history into an internal state that can be updated as new inputs arrive. Recurrent neural networks carry past information through recurrent connections, while long short-term memory networks are designed to improve information retention and gradient propagation over long temporal dependencies (Elman, 1990; Graves, 2012). However, one-step prediction error alone cannot show that an internal state is sufficient for multistep prediction. The state must retain the information needed for future prediction and remain correct under repeated updates. Predictive state representations and observable operator models define state in terms of predictions of future observations, providing an observable basis for state sufficiency and recursive updating in sequential systems (Littman & Sutton, 2001; Jaeger, 2000). This raises a fundamental question: can gradient-based training jointly learn an internal state, a readout map, and transition operators from a general initialization while retaining the information needed for future prediction and remaining exact under repeated transition composition?

Predictive state representations, observable operator models, rational series, and weighted automata provide closely related finite-dimensional descriptions of sequential systems through predictive quantities and finite-rank Hankel structure (Littman & Sutton, 2001; Rudary & Singh, 2003; Wingate & Singh, 2007; Jaeger, 2000; Thon & Jaeger, 2015; Balle et al., 2019). Spectral learning and system-identification methods exploit this structure to recover weighted automata, predictive-state models, and low-dimensional dynamical realizations from finite Hankel matrices or tensors (Balle et al., 2013;

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

Denis et al., 2014; Balle & Maillard, 2017; Siddiqi et al., 2010; Rabusseau et al., 2019; Tadipatri et al., 2025). Gradient-based predictive-state methods provide a complementary line of work in which predictive representations are learned directly rather than reconstructed solely through spectral factorization (McCracken & Bowling, 2005; Li et al., 2018a; Hefny et al., 2018). Our analysis is also related to quotient formulations of low-rank positive-semidefinite factors (Bonnabel & Sepulchre, 2010) and to matrix factorization and overparameterized linear networks, where gradient dynamics have been studied through convergence, implicit low-rank structure, balancing, conservation laws, and initialization dependence (Du et al., 2018; Li et al., 2018b; Valavi et al., 2020; Ye & Du, 2021; Tarmoun et al., 2021; Min et al., 2021; Xu et al., 2023; Min & Vidal, 2025; Chen et al., 2022). Related work also studies the formation and interpretation of internal computation in Transformers and recurrent networks (Huang et al., 2025; Nichani et al., 2024; van Rossem & Saxe, 2025; Proca et al., 2025; Dave et al., 2025; Geiger et al., 2021; 2025); classical normal-hyperbolicity theory provides tools for analyzing attracting invariant manifolds under dynamical perturbations (Fenichel, 1971; 1974).

Balle et al. (2012) describe exact reconstruction from complete Hankel blocks through rank factorization and prove that global minimizers of a constrained local loss yield exact realizations when the chosen state dimension is at least the Hankel rank. Tarmoun et al. (2021, Section 2) derive the Riccati representation and establish convergence of symmetric matrix-factorization gradient flow under nonspectral initialization. Our autonomous state flow is an instance of that system with a shifted target. Our analysis instead concerns concurrent updates of an overparameterized state factor, a decoder, and symbol-specific transitions, for which we establish global convergence of the triangular flow and local convergence of the fully coupled flow. Gradient-based system identification already has global guarantees under different system assumptions (Hardt et al., 2018); the nonconvex method of Tadipatri et al. (2025) uses stationary-point tests and, when necessary, additional factors. Our fixed-width dynamics require neither stationary-point tests nor factor augmentation. Stop-gradient analyses in self-supervised learning also establish collapse under specified feedback conditions (Tian et al., 2021). Here the target blocks are fixed and nonzero: positive feedback creates non-global attracting minima, and an explicit certificate permits eventual fully coupled training. The local analysis uses established Morse–Bott principles (Rebjock & Boumal, 2025).

Our contributions are fourfold. **First**, we construct a triangular training scheme that converges from any full-rank state initialization and arbitrary finite auxiliary initializations to an exact minimal recurrent realization. Its proof controls decoder and transition trajectories while the state Gram matrix approaches a singular limit. **Second**, we characterize when one-step fitting on the learned state subspace implies exact recurrent prediction, and prove that every fixed positive feedback strength admits a non-global attracting equilibrium manifold with an open basin containing full-rank initializations. Thus the triangular scheme's global guarantee cannot extend to fixed positive feedback over the same initialization class. **Third**, we establish the Morse–Bott structure of the exact solutions and local attraction throughout the interpolation to fully coupled training. We give a computable certificate for switching to the fully coupled system and prove that triangular training reaches it after finitely many checks. **Fourth**, we prove fixed-step convergence, including a certified joint-gradient step after switching, and finite-block perturbation bounds for fixed-length prediction. Six controlled experiments examine state formation, recurrent compatibility, local attraction under coupled gradients, static perturbation stability, initialization-dependent failure under positive feedback, and fixed-step triangular training.

2 Predictive State Space and Finite Hankel Coordinates

This section recalls standard Hankel realization results and fixes the coordinates used below. The finite-block reconstruction follows the rank-factorization construction described by Balle et al. (2012).

Let $\mathcal{A}$ be a finite alphabet. For each integer $\ell \geq 0$, let $\mathcal{A}^\ell$ denote the set of strings of length ℓ over $\mathcal{A}$, with $\mathcal{A}^0 = \{\epsilon\}$, where ϵ is the empty string. Define $\mathcal{A}^* = \bigcup_{\ell=0}^{\infty} \mathcal{A}^\ell$ as the set of all finite strings over $\mathcal{A}$. For $u, v \in \mathcal{A}^*$, write uv for their concatenation. For an integer $n \geq 1$, write $[n] = \{1, \ldots, n\}$. Let $f : \mathcal{A}^* \to \mathbb{R}$ be a real-valued sequence function. Its infinite Hankel matrix is $\mathsf{H}_f(u, v) = f(uv),\ u, v \in \mathcal{A}^*$. We define $\mathrm{rank}(\mathsf{H}_f)$ as the dimension of the linear span of its rows and assume $1 \leq r = \mathrm{rank}(\mathsf{H}_f) < \infty$.

For every prefix $u \in \mathcal{A}^*$, let h_u denote the *row of the infinite Hankel matrix* indexed by u, equivalently the function $h_u : \mathcal{A}^* \to \mathbb{R}, h_u(v) = f(uv)$. We refer to h_u as the *full future-score row* of u, since it records the score of every continuation $v \in \mathcal{A}^*$. Define the *predictive row space* by $\mathcal{R}_f = \mathrm{span}\{h_u : u \in \mathcal{A}^*\}$.

The predictive row space has dimension r. Right concatenation induces linear maps $\mathsf{S}_x h_u = h_{ux}$, and evaluation at the empty string recovers f after composing these maps. Proposition 15 in Appendix A states this standard realization result in full.

The predictive row space is the intrinsic finite-dimensional state space. Every matrix realization introduced below is a coordinate representation obtained after choosing a basis.

Definition 1 (Exact linear recurrent realization)**.** Let $d \geq 1$ be an integer, and let $(\boldsymbol{\alpha}, \{\boldsymbol{M}_x\}_{x\in\mathcal{A}}, \boldsymbol{c})$ be a triple with $\boldsymbol{\alpha} \in \mathbb{R}^{1\times d}$, $\boldsymbol{M}_x \in \mathbb{R}^{d\times d}$ for every $x \in \mathcal{A}$, and $\boldsymbol{c} \in \mathbb{R}^{d\times 1}$. We call this triple a d-dimensional exact linear recurrent realization of f if, for every $w = x_1 \cdots x_L \in \mathcal{A}^*$, $f(w) = \boldsymbol{\alpha}\boldsymbol{M}_{x_1} \cdots \boldsymbol{M}_{x_L}\boldsymbol{c}$, where the matrix product is interpreted as $\boldsymbol{I}_d$ when $L = 0$.

Choose prefix and suffix collections $\mathcal{U} = (u_1, \ldots, u_m),\ \mathcal{V} = (v_1, \ldots, v_p)$, with $m, p \geq r$, $\epsilon \in \mathcal{U} \cap \mathcal{V}$, and define $\mathsf{H}_{ij} = f(u_i v_j),\ (\mathsf{H}_x)_{ij} = f(u_i x v_j)$.

Assumption 2 (Complete Hankel block)**.** *The finite Hankel block* $\mathsf{H} \in \mathbb{R}^{m\times p}$ *satisfies* $\mathrm{rank}(\mathsf{H}) = r$.

Assumption 2 implies that the restriction map $\rho_{\mathcal{V}} : \mathcal{R}_f \to \mathbb{R}^{1\times p},\ \rho_{\mathcal{V}}(h) = \big(h(v_1), \ldots, h(v_p)\big)$, is injective. Moreover, the restricted rows associated with the selected prefixes $u_1, \ldots, u_m$ span $\mathrm{im}(\rho_{\mathcal{V}})$. Let $\mathsf{H} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^\top$ be the compact singular value decomposition of H, where $\boldsymbol{P} \in \mathbb{R}^{m\times r},\ \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r},\ \boldsymbol{Q} \in \mathbb{R}^{p\times r}$, and $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}_r,\ \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_r,\ \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. The singular values satisfy $\sigma_1 \geq \cdots \geq \sigma_r > 0$. The orthogonal projectors onto the column and row spaces of H are $\boldsymbol{\Pi}_{\mathrm{L}} = \boldsymbol{P}\boldsymbol{P}^\top, \boldsymbol{\Pi}_{\mathrm{R}} = \boldsymbol{Q}\boldsymbol{Q}^\top$.

For any prefix u, define its restricted future-score row $\boldsymbol{g}(u) = \big(f(uv_1), \ldots, f(uv_p)\big) \in \mathbb{R}^{1\times p}$ and its coordinate in the fixed SVD gauge by $\boldsymbol{z}_\star(u)^\top = \boldsymbol{g}(u)\boldsymbol{Q} \in \mathbb{R}^{1\times r}$. For the selected prefixes, stack these rows as $\boldsymbol{Z}_\star = \boldsymbol{P}\boldsymbol{\Sigma} = \mathsf{H}\boldsymbol{Q}$. For each $x \in \mathcal{A}$, define $\boldsymbol{B}_x^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{P}^\top\mathsf{H}_x\boldsymbol{Q}$. The following theorem shows that $\boldsymbol{B}_x^\star$ is the state transition induced by appending the symbol x.

Theorem 3 (Minimal exact predictive realization in an SVD gauge)**.** *Under Assumption 2, for every $x \in \mathcal{A}$, the shifted block satisfies $\mathsf{H}_x = \boldsymbol{Z}_\star\boldsymbol{B}_x^\star\boldsymbol{Q}^\top$. For every prefix u and symbol x, $\boldsymbol{z}_\star(ux)^\top = \boldsymbol{z}_\star(u)^\top\boldsymbol{B}_x^\star$, and $\boldsymbol{g}(u) = \boldsymbol{z}_\star(u)^\top\boldsymbol{Q}^\top$. Let j_ϵ satisfy $v_{j_\epsilon} = \epsilon$ and set $\boldsymbol{c}_\epsilon^\star = \boldsymbol{Q}^\top\boldsymbol{e}_{j_\epsilon}$. Then for $w = x_1 \cdots x_L$, $f(w) = \boldsymbol{z}_\star(\epsilon)^\top\boldsymbol{B}_{x_1}^\star \cdots \boldsymbol{B}_{x_L}^\star\boldsymbol{c}_\epsilon^\star$. Every exact linear recurrent realization has state dimension at least r.*

Any two minimal exact realizations are related by a unique invertible change of state coordinates (Proposition 16, Appendix A). The group $\mathrm{GL}(r)$ therefore describes the coordinate freedom of a minimal predictive realization. Its action changes the state coordinates, conjugates the transition matrices, and transforms the initial and readout vectors, while leaving every scalar sequence score invariant. The SVD construction above selects one representative from this orbit. The learner below receives only the scalar entries of H and H_x; it never receives $\boldsymbol{z}_\star(u)^\top$ or $\boldsymbol{B}_x^\star$.

3 State Formation on the Orthogonal Quotient

We associate each selected prefix u_i with the ith row of a learnable matrix $\boldsymbol{W} \in \mathbb{R}^{m\times m}$: $\boldsymbol{h}_i(\boldsymbol{W}) = \boldsymbol{e}_i^\top\boldsymbol{W},\ \boldsymbol{W} \in \mathbb{R}^{m\times m}$. The width satisfies $m \geq r$. The observable future Gram matrix is $\boldsymbol{G} = \mathsf{H}\mathsf{H}^\top = \boldsymbol{P}\boldsymbol{\Sigma}^2\boldsymbol{P}^\top$. Fix $0 < \lambda < \sigma_r^2$ and define $\mathcal{L}_{\mathrm{state}}(\boldsymbol{W}) = \frac{1}{4}\left\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{G}\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2$. Its gradient flow is $\dot{\boldsymbol{W}} = \big(\boldsymbol{G} - \lambda\boldsymbol{I}_m - \boldsymbol{W}\boldsymbol{W}^\top\big)\boldsymbol{W}$. This is the shifted-target symmetric matrix-factorization flow of Tarmoun et al. (2021, Section 2); we use their Riccati representation and nonspectral convergence in the downstream analysis.

The state objective depends only on $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{W}^\top$ and is invariant under $\boldsymbol{W} \mapsto \boldsymbol{W}\boldsymbol{R}$ for $\boldsymbol{R} \in \mathrm{O}(m)$. At finite time, $\boldsymbol{W} \in \mathrm{GL}(m)$, and two factors have the same Gram matrix if and only if they differ by right multiplication with an orthogonal matrix. Thus $\mathrm{GL}(m)/\mathrm{O}(m) \cong \mathbb{S}_{++}^m$ describes the finite-time quotient; Appendix B gives the orbit calculation.

Assumption 4 (Full-rank initialization)**.** *The state flow is initialized at an arbitrary matrix $\boldsymbol{W}(0) = \boldsymbol{W}_0 \in \mathrm{GL}(m)$.*

For a symmetric matrix $\boldsymbol{M}$, let $\boldsymbol{M}_+$ replace its negative eigenvalues by zero. The quotient state satisfies $\dot{\boldsymbol{X}} = \boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M} - 2\boldsymbol{X}^2$ with $\boldsymbol{M} = \boldsymbol{G} - \lambda \boldsymbol{I}_m$. Its inverse obeys a linear Sylvester equation, giving convergence to $\boldsymbol{M}_+$; Proposition 17 in Appendix B states the general result.

Define $\beta_j = \sigma_j^2 - \lambda, \boldsymbol{\Lambda} = \mathrm{diag}(\sqrt{\beta_1}, \ldots, \sqrt{\beta_r})$, and $\boldsymbol{T} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$. Since $0 < \lambda < \sigma_r^2 \le \sigma_j^2$ for every $j \in [r]$, each β_j is strictly positive. Hence both $\boldsymbol{\Lambda}$ and $\boldsymbol{T}$ are invertible. Moreover, $\boldsymbol{P}\boldsymbol{\Lambda} = \boldsymbol{Z}_\star \boldsymbol{T}$, so $\boldsymbol{T}$ rescales the canonical SVD state coordinates to the weight-decayed limiting scales.

Theorem 5 (Predictive-subspace formation from arbitrary full-rank initialization)**.** *Under Assumptions 2 and 4, the state flow converges to a limit $\boldsymbol{W}_\infty$ satisfying $\boldsymbol{W}_\infty \boldsymbol{W}_\infty^\top = (\boldsymbol{G} - \lambda \boldsymbol{I}_m)_+ = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$. Define $\boldsymbol{J} = \boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top \boldsymbol{W}_\infty \in \mathbb{R}^{r\times m}$. Then $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$ and $\boldsymbol{W}_\infty = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J} = \boldsymbol{Z}_\star \boldsymbol{T}\boldsymbol{J}$. The limiting state matrix has rank r, and its selected-prefix rows satisfy $\boldsymbol{h}_\infty(u_i) = \boldsymbol{z}_\star(u_i)^\top \boldsymbol{T}\boldsymbol{J}$. Thus the quotient dynamics determine a unique limiting Gram matrix, while the factor $\boldsymbol{W}_\infty$ retains an initialization-dependent right-isometric gauge.*

The quotient state converges exponentially, with $\|\boldsymbol{W}(t)\boldsymbol{W}(t)^\top - \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top\|_{\mathrm{op}} \le c_X e^{-\rho_X t}$ for some $c_X, \rho_X > 0$ (Proposition 18, Appendix B). The resulting integrable error is used in the cascade analysis.

4 Gauge-Compatible Triangular Cascade Formation

We first show that unconstrained base and one-symbol block fitting need not be recurrently exact, then characterize the support-compatible operators for which one-step fitting implies exact recurrent composition.

4.1 One-step fitting does not imply recurrent compatibility

Let $\boldsymbol{C} \in \mathbb{R}^{m\times p}$ denote a hidden-state decoder, and for each $x \in \mathcal{A}$, let $\boldsymbol{A}_x \in \mathbb{R}^{m\times m}$ denote the hidden-state transition associated with the symbol x.

Proposition 6 (Failure of unconstrained one-step block fitting)**.** *There exist a rank-one sequence function, a state-loss minimizer $\boldsymbol{W}_\infty$, and unconstrained parameters $\boldsymbol{C}$ and $\{\boldsymbol{A}_z\}_{z\in\mathcal{A}}$ such that $\boldsymbol{W}_\infty \boldsymbol{C} = \mathsf{H}, \boldsymbol{W}_\infty \boldsymbol{A}_z \boldsymbol{C} = \mathsf{H}_z$ for every $z \in \mathcal{A}$, but at least one two-step recurrent score is incorrect.*

The failure arises because unconstrained transitions can leave and re-enter the learned state subspace. Here the fitted identities are $\boldsymbol{W}_\infty \boldsymbol{C} = \mathsf{H}$ and $\boldsymbol{W}_\infty \boldsymbol{A}_x \boldsymbol{C} = \mathsf{H}_x$. By contrast, the local loss of Balle et al. (2012) includes transition-consistency residuals $\mathsf{H}\boldsymbol{V}\boldsymbol{D}_x - \mathsf{H}_x\boldsymbol{V}$ with $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_n$. Proposition 6 therefore concerns a different fitting problem. We next characterize the support restriction that removes this ambiguity.

4.2 Recurrent compatibility on the learned active subspace

For a word $w = x_1 \cdots x_L$, define $(\mathsf{H}_w)_{ij} = f(u_i w v_j), \mathsf{H}_w \in \mathbb{R}^{m\times p}$. Write $\boldsymbol{A}_w = \boldsymbol{A}_{x_1} \cdots \boldsymbol{A}_{x_L}$. For the empty word ϵ, set $\mathsf{H}_\epsilon = \mathsf{H}, \boldsymbol{A}_\epsilon = \boldsymbol{I}_m$. Let $\boldsymbol{W} \in \mathbb{R}^{m\times m}$ have rank r and satisfy $\mathrm{col}(\boldsymbol{W}) = \mathrm{col}(\boldsymbol{Z}_\star)$. Define $\boldsymbol{R}_{\mathrm{act}} = \boldsymbol{Z}_\star^\dagger \boldsymbol{W}, \boldsymbol{P}_{\mathrm{act}} = \boldsymbol{W}^\dagger \boldsymbol{W}$. Then $\boldsymbol{P}_{\mathrm{act}}$ is the orthogonal projector onto $\mathrm{row}(\boldsymbol{W})$.

Definition 7 (Support-compatible operators)**.** A decoder $\boldsymbol{C} \in \mathbb{R}^{m\times p}$ and a family of transitions $\{\boldsymbol{A}_x\}_{x\in\mathcal{A}}$ are support-compatible with $\boldsymbol{W}$ if $\boldsymbol{C} = \boldsymbol{P}_{\mathrm{act}}\boldsymbol{C}$ and $\boldsymbol{A}_x = \boldsymbol{P}_{\mathrm{act}}\boldsymbol{A}_x\boldsymbol{P}_{\mathrm{act}}$ for every $x \in \mathcal{A}$.

Theorem 8 (Supported one-step fitting and recurrent compatibility)**.** *Under Assumption 2, let $\boldsymbol{W}$ have rank r and satisfy $\mathrm{col}(\boldsymbol{W}) = \mathrm{col}(\boldsymbol{Z}_\star)$. Let $\boldsymbol{C}$ and $\{\boldsymbol{A}_x\}_{x\in\mathcal{A}}$ be support-compatible with $\boldsymbol{W}$. The following statements are equivalent.*

- *(i) The base and one-symbol blocks are fitted exactly, so $\boldsymbol{W}\boldsymbol{C} = \mathsf{H}$ and $\boldsymbol{W}\boldsymbol{A}_x\boldsymbol{C} = \mathsf{H}_x$ for every $x \in \mathcal{A}$.*
- *(ii) The supported operators satisfy $\boldsymbol{C} = \boldsymbol{R}_{\mathrm{act}}^\dagger \boldsymbol{Q}^\top$ and $\boldsymbol{A}_x = \boldsymbol{R}_{\mathrm{act}}^\dagger \boldsymbol{B}_x^\star \boldsymbol{R}_{\mathrm{act}}$ for every $x \in \mathcal{A}$.*
- *(iii) Every shifted block is realized exactly, so $\boldsymbol{W}\boldsymbol{A}_w\boldsymbol{C} = \mathsf{H}_w$ for every $w \in \mathcal{A}^*$.*

Moreover, the decoder condition $\boldsymbol{C} = \boldsymbol{P}_{\mathrm{act}}\boldsymbol{C}$ holds iff $\boldsymbol{C} = \boldsymbol{W}^\top\boldsymbol{\Gamma}$ for some auxiliary matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{m\times p}$. Likewise, the transition condition $\boldsymbol{A}_x = \boldsymbol{P}_{\mathrm{act}}\boldsymbol{A}_x\boldsymbol{P}_{\mathrm{act}}$ holds iff $\boldsymbol{A}_x = \boldsymbol{W}^\top\boldsymbol{\Omega}_x\boldsymbol{W}$ for some $\boldsymbol{\Omega}_x \in \mathbb{R}^{m\times m}$. If $\boldsymbol{W} = \boldsymbol{Z}_\star\boldsymbol{T}\boldsymbol{J}$ with $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$, then the unique support-compatible operators satisfying the equivalent conditions above are $\boldsymbol{C} = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{Q}^\top$ and $\boldsymbol{A}_x = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J}$.

4.3 TRIANGULAR CASCADE DYNAMICS

Theorem 8 identifies the target operator class. It remains to show that these operators can be learned concurrently while the state factor is still evolving toward its rank-r limit. The following triangular cascade constructs the decoder and transition operators without waiting for the state factor to converge. We now let the auxiliary matrices evolve as $\boldsymbol{\Gamma}(t) \in \mathbb{R}^{m\times p}, \boldsymbol{\Omega}_x(t) \in \mathbb{R}^{m\times m}$ and define $\boldsymbol{X}(t) = \boldsymbol{W}(t)\boldsymbol{W}(t)^\top, \boldsymbol{R}_{\mathrm{cas}}(t) = \boldsymbol{X}(t)\boldsymbol{\Gamma}(t)$. Here $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}_x$ induce the hidden-space decoder $\boldsymbol{C}_{\mathrm{cas}}(t) = \boldsymbol{W}(t)^\top\boldsymbol{\Gamma}(t)$, and symbol-indexed transition operators $\boldsymbol{A}_{x,\mathrm{cas}}(t) = \boldsymbol{W}(t)^\top\boldsymbol{\Omega}_x(t)\boldsymbol{W}(t)$. The matrix $\boldsymbol{R}_{\mathrm{cas}} = \boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{W}\boldsymbol{C}_{\mathrm{cas}}$ is the decoder block realized on the selected prefix states. Similarly, $\boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{R}_{\mathrm{cas}} = \boldsymbol{W}\boldsymbol{A}_{x,\mathrm{cas}}\boldsymbol{C}_{\mathrm{cas}}$ is the corresponding realized one-symbol transition block. Thus the decoder and transition losses fit H and H_x, respectively.

Fix learning-rate multipliers $\eta_d, \eta_a > 0$. Given initial values for all variables at $t = 0$, consider

$$\dot{\boldsymbol{W}} = (\boldsymbol{G} - \lambda\boldsymbol{I}_m - \boldsymbol{X})\,\boldsymbol{W}. \tag{1}$$

$$\dot{\boldsymbol{\Gamma}} = -\eta_d\boldsymbol{X}\,(\boldsymbol{X}\boldsymbol{\Gamma} - \mathsf{H})\,. \tag{2}$$

For every $x \in \mathcal{A}$, let

$$\dot{\boldsymbol{\Omega}}_x = -\eta_a\boldsymbol{X}\,(\boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{R}_{\mathrm{cas}} - \mathsf{H}_x)\,\boldsymbol{R}_{\mathrm{cas}}^\top. \tag{3}$$

The decoder flow is gradient flow in $\boldsymbol{\Gamma}$ for the time-dependent loss $\frac{1}{2}\left\|\boldsymbol{X}(t)\boldsymbol{\Gamma} - \mathsf{H}\right\|_{\mathrm{F}}^2$, while, for every $x \in \mathcal{A}$, the transition flow is gradient flow in $\boldsymbol{\Omega}_x$ for $\frac{1}{2}\left\|\boldsymbol{X}(t)\boldsymbol{\Omega}_x\boldsymbol{R}_{\mathrm{cas}}(t) - \mathsf{H}_x\right\|_{\mathrm{F}}^2$. The dynamics are triangular. The state flow is autonomous, the decoder flow uses the evolving state Gram matrix, and each transition flow uses the evolving state and decoder blocks. No downstream loss feeds back into an upstream equation. The next section restores these downstream-to-upstream gradient terms and analyzes the resulting joint-gradient dynamics.

The parameterization is compatible with the right orthogonal gauge. If $\boldsymbol{W}(t)$ is replaced by $\boldsymbol{W}(t)\boldsymbol{R}$ for a fixed $\boldsymbol{R} \in \mathrm{O}(m)$, then $\boldsymbol{X}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega}_x$ are unchanged, while $\boldsymbol{C}_{\mathrm{cas}}(t) \longmapsto \boldsymbol{R}^\top\boldsymbol{C}_{\mathrm{cas}}(t)$, $\boldsymbol{A}_{x,\mathrm{cas}}(t) \longmapsto \boldsymbol{R}^\top\boldsymbol{A}_{x,\mathrm{cas}}(t)\boldsymbol{R}$. For a hidden row state $\boldsymbol{h} \in \mathbb{R}^{1\times m}$, the product $\boldsymbol{h}\boldsymbol{C}_{\mathrm{cas}}$ is its predicted future-score row, while $\boldsymbol{h}\boldsymbol{A}_{x,\mathrm{cas}}$ is the hidden state obtained after appending the symbol x. Hence $\boldsymbol{W}\boldsymbol{C}_{\mathrm{cas}} = \boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{R}_{\mathrm{cas}}$, and $\boldsymbol{W}\boldsymbol{A}_{x,\mathrm{cas}}\boldsymbol{C}_{\mathrm{cas}} = \boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{X}\boldsymbol{\Gamma}$. Thus the first realized block is fitted to H, while the second is fitted to H_x. Define the active auxiliary targets

$$\boldsymbol{\Gamma}_\star = \boldsymbol{P}\boldsymbol{\Lambda}^{-2}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \in \mathbb{R}^{m\times p}, \tag{4}$$

$$\boldsymbol{\Omega}_x^\star = \boldsymbol{P}\boldsymbol{\Lambda}^{-2}\boldsymbol{\Sigma}\boldsymbol{B}_x^\star\boldsymbol{\Sigma}^{-1}\boldsymbol{P}^\top \in \mathbb{R}^{m\times m}. \tag{5}$$

Theorem 9 (Global gauge-compatible triangular-cascade formation)**.** *Under Assumptions 2 and 4, let $\eta_d, \eta_a > 0$ and choose arbitrary finite initial values $\boldsymbol{\Gamma}(0) = \boldsymbol{\Gamma}_0,\ \boldsymbol{\Omega}_x(0) = \boldsymbol{\Omega}_{x,0}$ for every $x \in \mathcal{A}$. The cascade in equation 1–equation 3 has a unique global solution. The auxiliary trajectories converge to limits $\boldsymbol{\Gamma}_\infty$ and $\boldsymbol{\Omega}_{x,\infty}$ satisfying $\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Gamma}_\infty = \boldsymbol{\Gamma}_\star$, $\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Omega}_{x,\infty}\boldsymbol{\Pi}_{\mathrm{L}} = \boldsymbol{\Omega}_x^\star$. Their inactive components may depend on initialization, but the realized blocks satisfy $\boldsymbol{X}(t)\boldsymbol{\Gamma}(t) \longrightarrow \mathsf{H}$, $\boldsymbol{X}(t)\boldsymbol{\Omega}_x(t)\boldsymbol{X}(t)\boldsymbol{\Gamma}(t) \longrightarrow \mathsf{H}_x$. Let $\boldsymbol{W}_\infty$, $\boldsymbol{J}$, and $\boldsymbol{T}$ be defined by Theorem 5. The induced hidden-space operators converge to $\boldsymbol{C}_{\mathrm{cas},\infty} = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{Q}^\top$, $\boldsymbol{A}_{x,\mathrm{cas},\infty} = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J}$. In particular, these induced limits do not depend on $\boldsymbol{\Gamma}_0$ or $\boldsymbol{\Omega}_{x,0}$.*

Remark 10 (Scope of the cascade theorem). Theorem 9 establishes global convergence for the triangular cascade, not for the fully coupled joint-gradient system.

Corollary 11 (Minimal recurrent realization formed by the cascade)**.** *Let i_ϵ and j_ϵ index the empty prefix and suffix. Define $\boldsymbol{h}_{\mathrm{cas},\infty}(\epsilon) = \boldsymbol{e}_{i_\epsilon}^\top\boldsymbol{W}_\infty$ and recursively set $\boldsymbol{h}_{\mathrm{cas},\infty}(ux) = \boldsymbol{h}_{\mathrm{cas},\infty}(u)\boldsymbol{A}_{x,\mathrm{cas},\infty}$. Then, for every prefix u, $\boldsymbol{h}_{\mathrm{cas},\infty}(u) = \boldsymbol{z}_\star(u)^\top\boldsymbol{T}\boldsymbol{J}$, $\boldsymbol{g}(u) = \boldsymbol{h}_{\mathrm{cas},\infty}(u)\boldsymbol{C}_{\mathrm{cas},\infty}$. With $\boldsymbol{c}_{\epsilon,\mathrm{cas}} = \boldsymbol{C}_{\mathrm{cas},\infty}\boldsymbol{e}_{j_\epsilon}$, every word $w = x_1\cdots x_L$ satisfies $f(w) = \boldsymbol{h}_{\mathrm{cas},\infty}(\epsilon)\boldsymbol{A}_{x_1,\mathrm{cas},\infty}\cdots\boldsymbol{A}_{x_L,\mathrm{cas},\infty}\boldsymbol{c}_{\epsilon,\mathrm{cas}}$. The active compressed parameters $\boldsymbol{\alpha}_{\mathrm{cas}} = \boldsymbol{h}_{\mathrm{cas},\infty}(\epsilon)\boldsymbol{J}^\top$, $\boldsymbol{M}_{x,\mathrm{cas}} = \boldsymbol{J}\boldsymbol{A}_{x,\mathrm{cas},\infty}\boldsymbol{J}^\top$, and $\boldsymbol{c}_{\mathrm{cas}} = \boldsymbol{J}\boldsymbol{c}_{\epsilon,\mathrm{cas}}$ form an r-dimensional exact linear recurrent realization and are therefore minimal.*

5 JOINT FEEDBACK AND CERTIFIED SWITCHING

The cascade theorem gives the limiting operators selected by each initialization. We now study the geometry of the full set of exact limits and then restore all cross-level derivatives of downstream losses. Define the state space $\mathcal{Z} = \mathbb{R}^{m\times m} \times \mathbb{R}^{m\times p} \times \prod_{x\in\mathcal{A}} \mathbb{R}^{m\times m}$. For $\boldsymbol{z} = (\boldsymbol{W}, \boldsymbol{\Gamma}, \{\boldsymbol{\Omega}_x\}_{x\in\mathcal{A}})$, use the product norm $\|\boldsymbol{z}\|_{\mathcal{Z}}^2 = \|\boldsymbol{W}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 + \sum_{x\in\mathcal{A}} \|\boldsymbol{\Omega}_x\|_{\mathrm{F}}^2$. For $\boldsymbol{R} \in \mathrm{O}(m)$, define the right gauge action by

$$\rho_{\boldsymbol{R}}\left(\boldsymbol{W}, \boldsymbol{\Gamma}, \{\boldsymbol{\Omega}_x\}_{x\in\mathcal{A}}\right) = \left(\boldsymbol{W}\boldsymbol{R}, \boldsymbol{\Gamma}, \{\boldsymbol{\Omega}_x\}_{x\in\mathcal{A}}\right). \tag{6}$$

Write F_0 for the triangular vector field in equation 1–equation 3. Define $\mathrm{St}(r,m) = \{\boldsymbol{J} \in \mathbb{R}^{r\times m} : \boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r\}$, $\mathcal{K}_{\mathrm{d}} = \{\boldsymbol{N} \in \mathbb{R}^{m\times p} : \boldsymbol{P}^\top\boldsymbol{N} = \boldsymbol{0}\}$, and $\mathcal{K}_{\mathrm{a}} = \{\boldsymbol{N} \in \mathbb{R}^{m\times m} : \boldsymbol{P}^\top\boldsymbol{N}\boldsymbol{P} = \boldsymbol{0}\}$. The exact-realization manifold is $\mathcal{E}_0 = \{(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}, \boldsymbol{\Gamma}_\star + \boldsymbol{N}_{\mathrm{d}}, \{\boldsymbol{\Omega}_x^\star + \boldsymbol{N}_x\}_{x\in\mathcal{A}}) : \boldsymbol{J} \in \mathrm{St}(r,m), \boldsymbol{N}_{\mathrm{d}} \in \mathcal{K}_{\mathrm{d}}, \boldsymbol{N}_x \in \mathcal{K}_{\mathrm{a}}\}$. Every point of $\mathcal{E}_0$ is an equilibrium of F_0. Its inactive auxiliary coordinates are unbounded, so uniform geometric statements are made on compact subsets.

Theorem 20 in Appendix D gives the tangent spaces and the triangular flow's normal spectral gap $\min\{\lambda, 2\beta_r, \eta_d\beta_r^2, \eta_a\beta_r^2\sigma_r^2\}$. In particular, $\mathcal{E}_0$ is normally attracting on compact subsets.

We now restore the cross-level gradient terms omitted by the triangular cascade. In particular, the decoder loss is allowed to update the state factor, while each transition loss is allowed to update both the state factor and the decoder variable. Define $\mathcal{L}_{\mathrm{d}}(\boldsymbol{W}, \boldsymbol{\Gamma}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\Gamma} - \mathsf{H}\|_{\mathrm{F}}^2$. For every $x \in \mathcal{A}$, define $\mathcal{L}_{\mathrm{a},x}(\boldsymbol{W}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_x) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{X}\boldsymbol{\Gamma} - \mathsf{H}_x\|_{\mathrm{F}}^2$. For $\varepsilon \in [0,1]$, consider

$$\dot{\boldsymbol{W}} = -\nabla_{\boldsymbol{W}}\mathcal{L}_{\mathrm{state}} - \varepsilon\nabla_{\boldsymbol{W}}\mathcal{L}_{\mathrm{d}} - \varepsilon^2 \sum_{x\in\mathcal{A}} \nabla_{\boldsymbol{W}}\mathcal{L}_{\mathrm{a},x}. \tag{7}$$

$$\dot{\boldsymbol{\Gamma}} = -\eta_d\nabla_{\boldsymbol{\Gamma}}\mathcal{L}_{\mathrm{d}} - \varepsilon\eta_d \sum_{x\in\mathcal{A}} \nabla_{\boldsymbol{\Gamma}}\mathcal{L}_{\mathrm{a},x}. \tag{8}$$

$$\dot{\boldsymbol{\Omega}}_x = -\eta_a\nabla_{\boldsymbol{\Omega}_x}\mathcal{L}_{\mathrm{a},x}. \tag{9}$$

Write F_ε for the resulting vector field. The parameter ε controls only the feedback of downstream losses into upstream variables. The direct decoder and transition updates remain active throughout the interpolation.

At $\varepsilon = 0$, the system reduces exactly to the triangular cascade. For $\varepsilon > 0$, define $\mathcal{J}_\varepsilon = \mathcal{L}_{\mathrm{state}} + \varepsilon\mathcal{L}_{\mathrm{d}} + \varepsilon^2\sum_{x\in\mathcal{A}}\mathcal{L}_{\mathrm{a},x}$ and let $\mathcal{D}_\varepsilon$ be the positive block-diagonal operator $\mathcal{D}_\varepsilon = \mathrm{diag}\left(\boldsymbol{I}_{\boldsymbol{W}}, \frac{\eta_d}{\varepsilon}\boldsymbol{I}_{\boldsymbol{\Gamma}}, \left\{\frac{\eta_a}{\varepsilon^2}\boldsymbol{I}_{\boldsymbol{\Omega}_x}\right\}_{x\in\mathcal{A}}\right)$. Then the vector field can be written as $F_\varepsilon(\boldsymbol{z}) = -\mathcal{D}_\varepsilon\nabla\mathcal{J}_\varepsilon(\boldsymbol{z})$. The mobility $\mathcal{D}_\varepsilon$ is defined only for $\varepsilon > 0$; the triangular equations directly define the zero-feedback endpoint. At $\varepsilon = 1$, the objective becomes $\mathcal{J}_1 = \mathcal{L}_{\mathrm{state}} + \mathcal{L}_{\mathrm{d}} + \sum_{x\in\mathcal{A}}\mathcal{L}_{\mathrm{a},x}$, and the system is the block-preconditioned joint-gradient flow of the unweighted state, decoder, and transition losses.

For every $\varepsilon \in (0,1]$, $\mathcal{E}_0$ is exactly the global minimizer set of $\mathcal{J}_\varepsilon$. Its Hessian is positive semidefinite with kernel $T_{\boldsymbol{z}_\star}\mathcal{E}_0$. Proposition 21 in Appendix D proves this Morse–Bott property and the strictly negative normal spectrum of DF_ε.

For a word $w = x_1\cdots x_L$, define the instantaneous recurrent score $\mathsf{s}_w(\boldsymbol{z}) = \boldsymbol{e}_{i_\epsilon}^\top\boldsymbol{W}\boldsymbol{A}_{x_1}(\boldsymbol{z})\cdots\boldsymbol{A}_{x_L}(\boldsymbol{z})\boldsymbol{C}(\boldsymbol{z})\boldsymbol{e}_{j_\epsilon}$, where $\boldsymbol{C}(\boldsymbol{z}) = \boldsymbol{W}^\top\boldsymbol{\Gamma}$, $\boldsymbol{A}_x(\boldsymbol{z}) = \boldsymbol{W}^\top\boldsymbol{\Omega}_x\boldsymbol{W}$. The product is interpreted as $\boldsymbol{I}_m$ when $L = 0$.

Theorem 22 in Appendix D proves local exponential convergence to $\mathcal{E}_0$, uniformly over $\varepsilon \in [0,1]$ near any compact gauge-invariant subset. Its asymptotic-phase map is gauge equivariant, and recurrent scores converge at the same exponential rate for every fixed maximum word length.

This local result does not imply convergence from arbitrary full-rank state initializations under positive feedback, as the next theorem shows.

5.1 FAILURE UNDER FIXED POSITIVE FEEDBACK

Theorem 12 (Attracting nonoptimal equilibria under positive feedback)**.** *Under Assumption 2, fix $\varepsilon \in (0,1]$ and define $\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma}) = \lambda\boldsymbol{I}_m - \boldsymbol{G} - \varepsilon(\boldsymbol{\Gamma}\mathsf{H}^\top + \mathsf{H}\boldsymbol{\Gamma}^\top)$ and $\mathcal{E}_{\mathrm{bad},\varepsilon} = \{(\boldsymbol{0}, \boldsymbol{\Gamma}, \{\boldsymbol{\Omega}_x\}_{x\in\mathcal{A}}) : \boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma}) \succ 0\}$. This is a nonempty smooth manifold of non-global local minimizers of $\mathcal{J}_\varepsilon$, normally*

attracting on compact subsets. Every point has an open neighborhood whose trajectories converge exponentially to this manifold; that neighborhood contains a nonempty open set of full-rank state initializations. All induced recurrent scores at these equilibria are zero.

Thus no fixed positive feedback strength retains the triangular flow's global convergence guarantee for arbitrary finite auxiliary initializations. This statement holds separately for each fixed $\varepsilon > 0$; it does not imply that a single initialization fails for every positive feedback strength. The proof gives an explicit nonempty family, $\mathbf{\Gamma} = -c\mathsf{H}$ with $c \geq (2\varepsilon)^{-1}$, and shows that the failure persists under sufficiently small decoder perturbations.

5.2 A COMPUTABLE SWITCH TO FULLY COUPLED TRAINING

Using an orthonormal vectorization, define

$$\boldsymbol{y} = \mathcal{D}_1^{-1/2}\boldsymbol{z},\ \Psi(\boldsymbol{y}) = \mathcal{J}_1(\mathcal{D}_1^{1/2}\boldsymbol{y}),\ \widetilde{\mathcal{E}} = \mathcal{D}_1^{-1/2}\mathcal{E}_0. \tag{10}$$

In these coordinates, fully coupled training is the Euclidean gradient flow $\dot{\boldsymbol{y}} = -\nabla\Psi(\boldsymbol{y})$. Appendix E constructs an explicit nearest-point projection $\boldsymbol{q}$ of the current $\boldsymbol{y}$ onto $\widetilde{\mathcal{E}}$ and sets $\delta = \|\boldsymbol{y} - \boldsymbol{q}\|_2$. It also gives the smallest positive Hessian eigenvalue μ at $\boldsymbol{q}$ and explicit bounds M and L for the Hessian norm and its Lipschitz constant on the closed unit ball centered at $\boldsymbol{q}$. The switching test is

$$\left(1 + \frac{4M}{\mu}\right)\delta < b,\ b = \min\left\{\frac{1}{2}, \frac{\beta_r}{2\sqrt{\beta_1}}, \frac{\mu}{8L}\right\}. \tag{11}$$

The test depends only on the observed finite blocks, the current parameters, and finitely many matrix operations.

Theorem 13 (Finite switching to fully coupled training)**.** *Under the assumptions of Theorem 9, run the triangular flow and inspect equation 11 at times $n\Delta$, where $\Delta > 0$ is fixed and $n \geq 0$. At the first successful inspection, set $\varepsilon = 1$ thereafter. A successful inspection occurs after finitely many checks. The resulting trajectory converges to a point of $\mathcal{E}_0$ and hence induces an exact minimal recurrent realization after compression. If time is reset to zero at the switch, its transformed parameters satisfy*

$$\|\boldsymbol{y}(t) - \boldsymbol{y}_\infty\|_2 \leq \frac{2M}{\mu}\delta e^{-\mu t/2}. \tag{12}$$

The same conclusion holds for a convergent fixed-step triangular iteration from Corollary 14, inspected at every iteration and replaced, after the first successful test, by

$$\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} - h\mathcal{D}_1\nabla\mathcal{J}_1(\boldsymbol{z}^{(k)}),\ 0 < h \leq \frac{\mu}{2M^2}, \tag{13}$$

where h is selected at the switch and thereafter held fixed. With $k = 0$ at the switch and $\rho_h = \sqrt{1 - h\mu/2}$,

$$\|\boldsymbol{y}^{(k)} - \boldsymbol{y}_\infty\|_2 \leq \frac{4M}{\mu}\delta\rho_h^k. \tag{14}$$

Every fixed finite word inherits the corresponding exponential or geometric prediction rate.

The number of inspections is finite; no polynomial upper bound is asserted.

6 FIXED-STEP TRIANGULAR CASCADE EXTENSION

Continuous-time dynamics make the formation mechanism transparent, while the same triangular dependency structure is preserved under deterministic fixed-step updates. Appendix F gives the deterministic fixed-step updates, sufficient step-size conditions, and proofs.

Corollary 14 (Fixed-step triangular-cascade formation)**.** *Under Assumption 2, let $\boldsymbol{W}^{(0)} \in \mathrm{GL}(m)$ and let $\mathbf{\Gamma}^{(0)}$ and $\{\mathbf{\Omega}_x^{(0)}\}_{x\in\mathcal{A}}$ be arbitrary finite matrices. Under the step-size conditions of Theorems 26 and 28, every finite state iterate remains invertible and the state, decoder, and transition levels converge geometrically. The state Gram matrix converges to $\boldsymbol{P}\mathbf{\Lambda}^2\boldsymbol{P}^\top$. With $\boldsymbol{J}_{\mathrm{d}} = \mathbf{\Lambda}^{-1}\boldsymbol{P}^\top\boldsymbol{W}_\infty^{\mathrm{d}}$, we have $\boldsymbol{J}_{\mathrm{d}}\boldsymbol{J}_{\mathrm{d}}^\top = \boldsymbol{I}_r$, and the induced operators converge to $\boldsymbol{J}_{\mathrm{d}}^\top\boldsymbol{T}^{-1}\boldsymbol{Q}^\top$ and $\boldsymbol{J}_{\mathrm{d}}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J}_{\mathrm{d}}$. Their compression forms an exact r-dimensional recurrent realization and is therefore minimal. The induced limits are independent of the initial auxiliary variables.*

The step-size conditions are deterministic and sufficient. This triangular result excludes stochastic gradients, momentum, and adaptive optimizers. The certified fixed-step switch in Theorem 13 separately covers subsequent fully coupled updates.

7 Experimental Results and Discussion

Experiments used FP64 CUDA arithmetic (TF32 disabled) on an Intel Core i9-10900X, 64 GB RAM, and NVIDIA GeForce RTX 4090. Weighted sequence systems over $\mathcal{A} = \{0, 1\}$ were generated with transitions given by 0.95 times seeded random orthogonal matrices and Gaussian initial and final vectors normalized to unit Euclidean norm. We retained targets with block rank r at relative singular-value threshold 10^{-11} and condition number ≤ 3 (≤ 1.8 in Experiment 3) over the nonzero singular values, then rescaled scores to $\sigma_r = 1$. Experiments 1, 2, 4, 5, and 6 used three targets per $r \in \{3, 5, 7\}$ with all prefixes and suffixes of length ≤ 3 (15×15 blocks); Experiment 3 used three rank-three targets with length cutoff 2 (7×7 blocks). Relative Frobenius residuals compare $\boldsymbol{X}$ with $\boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$, $\boldsymbol{X}\boldsymbol{\Gamma}$ with H, and $\boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{X}\boldsymbol{\Gamma}$ with H_x (maximum over symbols). Recurrent errors compare recursive scores with the generating function, without substituting target states. Per length, we enumerated all words if their number was ≤ 512, otherwise sampled 512 uniformly with replacement. Reported maxima are therefore over the evaluated words only. Experiments 1 and 3 used classical fourth-order Runge–Kutta integration (Hairer et al., 1993). Protocols and detailed results for Experiments 5 and 6 are given in Appendix H.

Table 1: Experiment 1 at $t = 100$: mean $\pm$ sample standard deviation over 30 runs per rank. Recurrent errors are within-run maxima over evaluated words and lengths.

r	Rank recovery	State (10^{-14})	Decoder (10^{-13})	Transition (10^{-7})	Recurrent error (10^{-6})
3	30/30	0.91 ± 0.42	2.42 ± 0.23	2.81 ± 0.42	0.94 ± 0.52
5	30/30	0.80 ± 0.21	1.58 ± 0.10	1.34 ± 0.51	1.12 ± 0.96
7	30/30	1.17 ± 0.25	1.31 ± 0.07	1.28 ± 0.21	0.99 ± 0.33

Experiment 1: state formation and recurrent prediction For ten initializations per target, $\boldsymbol{W}(0)$ was $0.7\boldsymbol{I}_{15}$ plus Gaussian noise (entrywise standard deviation 0.08); Gaussian auxiliary entries had standard deviation 0.05. All variables evolved concurrently from $t = 0$ with $(\lambda, \eta_d, \eta_a) = (0.25\sigma_r^2, 0.5, 0.25)$, integration step 0.005, and final time 100. Numerical rank used a relative singular-value cutoff of 10^{-6} on the current state matrix. Test lengths $\{8, 12, 16, 24\}$ exceeded the training maximum 7. All 90 runs recovered rank r (Table 1); the overall maximum absolute recurrent error was 2.74×10^{-6}. At $r = 5$, mean state, decoder, and transition residuals decreased from $(2.58 \times 10^{-6}, 9.26 \times 10^{-4}, 1.13 \times 10^{-2})$ at $t = 20$ to $(8.01 \times 10^{-15}, 1.58 \times 10^{-13}, 1.34 \times 10^{-7})$ at $t = 100$, consistent with state formation followed by downstream tracking (Theorem 9). The continuous-time state factor remains invertible at every finite time (Proposition 17).

Table 2: Experiment 2: maxima over indicated realizations and evaluated words; one-step errors are relative, recurrent errors absolute. Words at $L = 12$ were sampled.

Realization	Count	One-step error	Recurrent error $L = 2$	Recurrent error $L = 12$
Supported	9	2.22×10^{-14}	1.42×10^{-14}	4.04×10^{-14}
Perturbed, 0.1	90	2.23×10^{-14}	2.20×10^{-2}	5.88×10^{-2}
Perturbed, 0.5	90	2.23×10^{-14}	4.99×10^{-1}	3.08×10^{0}
Perturbed, 1.0	90	2.25×10^{-14}	2.35×10^{0}	3.38×10^{2}

Experiment 2: one-step fitting and recurrent compatibility From exact support-compatible realizations, we fixed state and decoder and perturbed each transition by two rank-one maps through a common inactive direction, leaving $\boldsymbol{W}\Delta\boldsymbol{A}_x\boldsymbol{C} = \boldsymbol{0}$. Perturbation scales $\{0.1, 0.5, 1.0\}$ with ten draws per target produced 270 families; nine supported controls were retained. All perturbed families

Table 3: Experiment 3: criterion $\rho_{\mathcal{E}}(100)/\rho_{\mathcal{E}}(0) \leq 0.02$; ratios and recurrent errors are maxima over 120 initial conditions per coupling.

ε	Criterion satisfied	Minimum spectral gap	Largest residual ratio	Largest recurrent error, $L = 8$
0.00	120/120	0.2500	1.11×10^{-11}	3.16×10^{-14}
0.25	120/120	0.2245	6.34×10^{-11}	1.73×10^{-12}
0.50	120/120	0.1100	4.43×10^{-6}	9.32×10^{-8}
0.75	120/120	0.0633	5.02×10^{-4}	1.13×10^{-5}
1.00	120/120	0.0409	4.91×10^{-3}	1.01×10^{-4}

met one-step tolerance 10^{-10} (relative residuals $< 2.25 \times 10^{-14}$), but each had maximum absolute recurrent error $> 10^{-6}$ over lengths $\{2, 4, 8, 12\}$; controls remained below 4.04×10^{-14} (Table 2). This constructed test illustrates Proposition 6 and Theorem 8; it neither estimates unconstrained-training failure rates nor implies that every support violation changes scores.

Experiment 3: local attraction under joint feedback For each target, we selected one equilibrium in $\mathcal{E}_0$ with zero inactive auxiliary components and added Gaussian state and auxiliary perturbations at relative Frobenius norms $\{0.001, 0.01, 0.05, 0.10\}$ for each matrix. Ten draws per target and scale yielded 120 initial conditions, each evaluated at $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$, for 600 trajectories. Settings were $(\lambda, \eta_d, \eta_a) = (0.25\sigma_r^2, 1, 1)$, integration step 0.01, and final time 100. The residual $\rho_{\mathcal{E}}$ combines the state residual with active decoder and transition residuals and vanishes on $\mathcal{E}_0$; scores were evaluated exhaustively at lengths $\{2, 4, 8\}$. All 15 equilibrium Jacobians in dimension 196 had 123 eigenvalues of magnitude $\leq 10^{-8}$, matching the tangent dimension, and 73 with negative real parts. Minimum decay rates across targets decreased with feedback, with slower residual decay, consistent with Proposition 21 and Theorem 20. Evidence is local to these three equilibria.

Experiment 4: static reconstruction from perturbed blocks For each target, six perturbation levels up to 0.8 times the sufficient limit and ten draws per level gave 540 static reconstructions using Theorem 33. All instances satisfied the projector and observable-operator bounds, and all 2700 rollout checks at lengths $\{1, 2, 4, 8, 12\}$ satisfied the prediction bound. The largest error/bound ratios for left projectors, right projectors, operators, and scores were 0.0980, 0.1064, 0.0124, and 1.70×10^{-3}, respectively; mean maximum recurrent error increased from 7.46×10^{-4} to 3.02×10^{-2} with perturbation level.

Experiment 5: triangular training and fixed positive feedback Across 2160 matched-initialization trajectories, all 540 triangular runs met the criterion (maximum recurrent error 6.87×10^{-8}), whereas positive feedback produced 540 verified collapses in the constructed profiles and none in the ordinary profiles. This illustrates the initialization-dependent obstruction in Theorem 12, not a general failure of positive-feedback training.

Experiment 6: fixed-step triangular training Of 540 synchronous fixed-step runs, 480 met the numerical criterion within 300,000 iterations: 249/270 at step-selection factor $f = 1$ and 231/270 at $f = 0.5$. The remaining 60 exhausted the budget; no numerical failure or monitored descent or curvature violation occurred. The maximum absolute recurrent error among successful runs was 8.58×10^{-7}. Budget exhaustion alone does not refute the asymptotic convergence result of Corollary 14.

8 Limitations

Our analysis assumes exact finite Hankel rank and a fixed complete prefix–suffix table, excluding approximate low rank, nonlinear predictive states, and standard autoregressive training. Global guarantees apply to triangular training and certified switching, while fixed positive feedback may admit non-global attractors and the switching certificate can be conservative. Fixed-step results exclude stochastic or adaptive optimization, and perturbation results concern static reconstruction and fixed-length prediction rather than noisy joint training.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, to support the retrieval and discovery of related work, and to assist in checking and revising portions of the mathematical arguments and proofs, and to support the verification of portions of the experimental code. All AI-assisted mathematical content was independently verified by the authors against the stated assumptions and derivations, and all references identified with AI assistance were checked against the original publications. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Complete proofs of the theoretical results are provided in Appendices A–G. Section 7 reports the experimental setup and results, with additional protocols and numerical details for Experiments 5 and 6 in Appendix H. The supplementary material contains the source code used to reproduce the experiments and numerical evaluations reported in the paper.

REFERENCES

P. A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005. doi: 10.1137/040605266. URL https://doi.org/10.1137/040605266.

B. Balle, X. Carreras, F. M. Luque, and A.J. Quattoni. Spectral learning of weighted automata: a forward-backward perspective. *Machine learning*, (October):1–31, October 2013. ISSN 0885-6125. doi: 10.1007/s10994-013-5416-x. URL https://hdl.handle.net/2117/21075.

Borja Balle and Odalric-Ambrym Maillard. Spectral learning from a single trajectory under finite-state policies. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 361–370. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/balle17a.html.

Borja Balle, Ariadna Quattoni, and Xavier Carreras. Local loss optimization in operator models: a new insight into spectral learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML'12, pp. 1819–1826, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Borja Balle, Prakash Panangaden, and Doina Precup. Singular value automata and approximate minimization. *Mathematical Structures in Computer Science*, 29(9):1444–1478, 2019. doi: 10.1017/S0960129519000094.

Silvère Bonnabel and Rodolphe Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, 2010. doi: 10.1137/080731347. URL https://doi.org/10.1137/080731347.

Kun Chen, Dachao Lin, and Zhihua Zhang. On non-local convergence analysis of deep linear networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3417–3443. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/chen22p.html.

Neisarg Dave, Daniel Kifer, C. Lee Giles, and Ankur Mali. Bridging neural and symbolic computation: A learnability study of rnns on counter and dyck languages. In Leilani H. Gilpin, Eleonora Giunchiglia, Pascal Hitzler, and Emile van Krieken (eds.), *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, volume 284 of *Proceedings of Machine Learning Research*, pp. 86–115. PMLR, 08–10 Sep 2025. URL https://proceedings.mlr.press/v284/dave25a.html.

François Denis, Mattias Gybels, and Amaury Habrard. Dimension-free concentration bounds on hankel matrices for spectral learning. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine*

Learning Research, pp. 449–457, Beijing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/denis14.html.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: layers are automatically balanced. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 382–393, Red Hook, NY, USA, 2018. Curran Associates Inc.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Fenichel. Persistence and smoothness of invariant manifolds for flows. *Indiana University Mathematics Journal*, 21(3):193–226, 1971.

Neil Fenichel. Asymptotic stability with rate conditions for dynamical systems. *Bulletin of the American Mathematical Society*, 80:346–349, 1974. URL https://api.semanticscholar.org/CorpusID:120133838.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

Ernst Hairer, Gerhard Wanner, and Syvert P Nørsett. *Solving ordinary differential equations I: Nonstiff problems*. Springer, 1993.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Ahmed Hefny, Zita Marinho, Wen Sun, Siddhartha Srinivasa, and Geoffrey Gordon. Recurrent predictive state policy networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1949–1958. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hefny18a.html.

Ruiquan Huang, Yingbin Liang, and Jing Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=yTAR011mOF.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000. doi: 10.1162/089976600300015411.

Tianyu Li, Guillaume Rabusseau, and Doina Precup. Nonlinear weighted finite automata. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 679–688. PMLR, 09–11 Apr 2018a. URL https://proceedings.mlr.press/v84/li18a.html.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2–47. PMLR, 06–09 Jul 2018b. URL https://proceedings.mlr.press/v75/li18a.html.

Michael Littman and Richard S Sutton. Predictive representations of state. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Peter McCracken and Michael Bowling. Online discovery and learning of predictive state representations. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'05, pp. 875–882, Cambridge, MA, USA, 2005. MIT Press.

Hancheng Min and René Vidal. Understanding incremental learning with closed-form solution to gradient flow on overparamerterized matrix factorization. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 1130–1135, 2025. doi: 10.1109/CDC57313.2025.11312037.

Hancheng Min, Salma Tarmoun, Rene Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7760–7768. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/min21c.html.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38018–38070. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/nichani24a.html.

Alexandra Maria Proca, Clémentine Carla Juliette Dominé, Murray Shanahan, and Pedro A. M. Mediano. Learning dynamics in linear recurrent neural networks. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49860–49902. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/proca25a.html.

Guillaume Rabusseau, Tianyu Li, and Doina Precup. Connecting weighted automata and recurrent neural networks through spectral learning. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1630–1639. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/rabusseau19a.html.

Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent conditions for c 2 functions. *Mathematical Programming*, 213(1):151–199, 2025.

Matthew Rudary and Satinder Singh. A nonlinear predictive state representation. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/72e6d3238361fe70f22fb0ac624a7072-Paper.pdf.

Sajid Siddiqi, Byron Boots, and Geoffrey Gordon. Reduced-rank hidden markov models. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 741–748, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/siddiqi10a.html.

Gilbert W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, San Diego, 1993. ISBN 978-0-12-670230-9.

Uday Kiran Reddy Tadipatri, Benjamin D. Haeffele, Joshua Agterberg, Ingvar Ziemann, and Rene Vidal. Nonconvex linear system identification with minimal state representation. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate (eds.), *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 1286–1299. PMLR, 04–06 Jun 2025. URL https://proceedings.mlr.press/v283/tadipatri25a.html.

Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10153–10161. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tarmoun21a.html.

Michael Thon and Herbert Jaeger. Links between multiplicity automata, observable operator models and predictive state representations: a unified learning framework. *J. Mach. Learn. Res.*, 16(1): 103–147, January 2015. ISSN 1532-4435.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

Hossein Valavi, Sulin Liu, and Peter Ramadge. Revisiting the landscape of matrix factorization. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1629–1638. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/valavi20a.html.

Loek van Rossem and Andrew M Saxe. Algorithm development in neural networks: Insights from the streaming parity task. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=3go0lhfxd0.

David Wingate and Satinder Singh. Exponential family predictive representations of state. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, pp. 1617–1624, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

Ziqing Xu, Hancheng Min, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2262–2284. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/xu23c.html.

Tian Ye and Simon Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1429–1439. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0af854284f4ab0cfea8fcfd889cbb41a-Paper.pdf.

A Proofs for the Predictive Space and Finite Hankel Coordinates

Proposition 15 (Predictive row space)**.** *The predictive row space satisfies* $\dim(\mathcal{R}_f) = r$*. For every* $x \in \mathcal{A}$*, right concatenation induces a well-defined linear operator* $\mathsf{S}_x : \mathcal{R}_f \to \mathcal{R}_f,\ \mathsf{S}_x h_u = h_{ux}$*. Let* $\mathrm{ev}_\epsilon : \mathcal{R}_f \to \mathbb{R},\ \mathrm{ev}_\epsilon(h) = h(\epsilon)$*. Then* $f(x_1 \cdots x_L) = \mathrm{ev}_\epsilon\left(\mathsf{S}_{x_L} \cdots \mathsf{S}_{x_1} h_\epsilon\right)$ *for every* $x_1 \cdots x_L \in \mathcal{A}^*$*. Consequently, any choice of basis of* $\mathcal{R}_f$ *yields an exact* r*-dimensional linear recurrent realization of* f*.*

Proof of Proposition 15. By definition, $\mathcal{R}_f$ is the row space of the infinite Hankel matrix. Hence $\dim(\mathcal{R}_f) = \mathrm{rank}(\mathsf{H}_f) = r$.

Fix $x \in \mathcal{A}$. To show that the assignment $h_u \mapsto h_{ux}$ induces a well-defined linear operator on $\mathcal{R}_f$, it is enough to verify that it preserves every linear relation among the Hankel rows. Suppose $\sum_{k=1}^K a_k h_{u_k} = 0$. Then, for every $v \in \mathcal{A}^*$, $\sum_{k=1}^K a_k h_{u_k x}(v) = \sum_{k=1}^K a_k f(u_k x v) = \sum_{k=1}^K a_k h_{u_k}(xv) = 0$. Thus the assignment $h_u \mapsto h_{ux}$ respects all linear relations among the generators of $\mathcal{R}_f$ and therefore extends uniquely to a well-defined linear operator $\mathsf{S}_x : \mathcal{R}_f \to \mathcal{R}_f$.

Now choose a basis $\mathcal{B}$ of $\mathcal{R}_f$, and write $[h]_\mathcal{B}$ for the row-coordinate vector of $h \in \mathcal{R}_f$. Let $\boldsymbol{\alpha} = [h_\epsilon]_\mathcal{B}$, and, for each $x \in \mathcal{A}$, let $\boldsymbol{M}_x$ be the matrix determined by $[\mathsf{S}_x h]_\mathcal{B} = [h]_\mathcal{B} \boldsymbol{M}_x$. Let $\boldsymbol{c}$ represent evaluation at the empty suffix, so that $h(\epsilon) = [h]_\mathcal{B} \boldsymbol{c}$ for every $h \in \mathcal{R}_f$.

For $w = x_1 \cdots x_L$, repeated right shifts give $h_w = \mathsf{S}_{x_L} \cdots \mathsf{S}_{x_1} h_\epsilon$, and hence $[h_w]_{\mathcal{B}} = \boldsymbol{\alpha} \boldsymbol{M}_{x_1} \cdots \boldsymbol{M}_{x_L}$. Evaluating at the empty suffix yields $f(w) = h_w(\epsilon) = \boldsymbol{\alpha} \boldsymbol{M}_{x_1} \cdots \boldsymbol{M}_{x_L} \boldsymbol{c}$. Since $\dim(\mathcal{R}_f) = r$, this is an exact r-dimensional linear recurrent realization of f. □

Proof of Theorem 3. Let $\rho_{\mathcal{V}} : \mathcal{R}_f \to \mathbb{R}^{1\times p}$, $\rho_{\mathcal{V}}(h) = (h(v_1), \ldots, h(v_p))$ be the restriction map. Each row of H is of the form $\rho_{\mathcal{V}}(h_{u_i})$, and hence $\mathrm{row}(\mathsf{H}) \subseteq \mathrm{im}(\rho_{\mathcal{V}})$. By Assumption 2, $\dim \mathrm{row}(\mathsf{H}) = \mathrm{rank}(\mathsf{H}) = r = \dim \mathcal{R}_f$. Since $\dim \mathrm{im}(\rho_{\mathcal{V}}) \leq \dim \mathcal{R}_f$, we obtain $\mathrm{im}(\rho_{\mathcal{V}}) = \mathrm{row}(\mathsf{H})$. In particular, $\dim \mathrm{im}(\rho_{\mathcal{V}}) = \dim \mathcal{R}_f$, so $\rho_{\mathcal{V}}$ is injective.

Define $\phi : \mathcal{R}_f \to \mathbb{R}^{1\times r}$, $\phi(h) = \rho_{\mathcal{V}}(h)\boldsymbol{Q}$. Since the rows of $\boldsymbol{Q}^\top$ form an orthonormal basis of $\mathrm{row}(\mathsf{H})$, the map $y \mapsto y\boldsymbol{Q}$ is an isomorphism from $\mathrm{row}(\mathsf{H})$ to $\mathbb{R}^{1\times r}$, with inverse $z \mapsto z\boldsymbol{Q}^\top$. Hence ϕ is an isomorphism. Then $\boldsymbol{g}(u) = \boldsymbol{g}(u)\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{z}_\star(u)^\top \boldsymbol{Q}^\top$ and $\boldsymbol{Z}_\star = \mathsf{H}\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{\Sigma}$.

Conjugating the abstract right shift through ϕ gives a unique matrix $\widetilde{\boldsymbol{B}}_x \in \mathbb{R}^{r\times r}$ satisfying $\phi(\mathsf{S}_x h) = \phi(h)\widetilde{\boldsymbol{B}}_x$ for every $h \in \mathcal{R}_f$. Applying this identity to the selected prefixes yields $\boldsymbol{Z}_\star \widetilde{\boldsymbol{B}}_x = \mathsf{H}_x \boldsymbol{Q}$. Since $\boldsymbol{Z}_\star = \boldsymbol{P}\boldsymbol{\Sigma}$, left multiplication by $\boldsymbol{\Sigma}^{-1}\boldsymbol{P}^\top$ gives $\widetilde{\boldsymbol{B}}_x = \boldsymbol{\Sigma}^{-1}\boldsymbol{P}^\top \mathsf{H}_x \boldsymbol{Q} = \boldsymbol{B}^\star_x$. Thus $\phi(\mathsf{S}_x h) = \phi(h)\widetilde{\boldsymbol{B}}_x$ holds with $\boldsymbol{B}^\star_x$. The ith row of H_x is $\rho_{\mathcal{V}}(h_{u_i x})$, and therefore belongs to $\mathrm{im}(\rho_{\mathcal{V}}) = \mathrm{row}(\mathsf{H})$. Hence $\mathsf{H}_x = \mathsf{H}_x \boldsymbol{Q}\boldsymbol{Q}^\top$. Multiplying $\boldsymbol{Z}_\star \widetilde{\boldsymbol{B}}_x = \mathsf{H}_x \boldsymbol{Q}$. by $\boldsymbol{Q}^\top$ proves $\mathsf{H}_x = \boldsymbol{Z}_\star \boldsymbol{B}^\star_x \boldsymbol{Q}^\top$.

Applying $\phi(\mathsf{S}_x h) = \phi(h)\widetilde{\boldsymbol{B}}_x$ to h_u proves $\boldsymbol{z}_\star(ux)^\top = \boldsymbol{z}_\star(u)^\top \boldsymbol{B}^\star_x$. Since $v_{j_\epsilon} = \epsilon$, equation $\boldsymbol{g}(u) = \boldsymbol{z}_\star(u)^\top \boldsymbol{Q}^\top$ gives $f(u) = \boldsymbol{z}_\star(u)^\top \boldsymbol{Q}^\top \boldsymbol{e}_{j_\epsilon}$. Repeated application of $\boldsymbol{z}_\star(ux)^\top = \boldsymbol{z}_\star(u)^\top \boldsymbol{B}^\star_x$ gives $\boldsymbol{z}_\star(w)^\top = \boldsymbol{z}_\star(\epsilon)^\top \boldsymbol{B}^\star_{x_1} \cdots \boldsymbol{B}^\star_{x_L}$. Since $\boldsymbol{c}^\star_\epsilon = \boldsymbol{Q}^\top \boldsymbol{e}_{j_\epsilon}$, the readout identity above then yields $f(w) = \boldsymbol{z}_\star(\epsilon)^\top \boldsymbol{B}^\star_{x_1} \cdots \boldsymbol{B}^\star_{x_L} \boldsymbol{c}^\star_\epsilon$.

Finally, consider any d-dimensional exact realization $(\boldsymbol{\alpha}, \{\boldsymbol{M}_x\}_{x\in\mathcal{A}}, \boldsymbol{c})$. For a string $u = x_1 \cdots x_k$, write $\boldsymbol{M}_u = \boldsymbol{M}_{x_1} \cdots \boldsymbol{M}_{x_k}$, with $\boldsymbol{M}_\epsilon = \boldsymbol{I}_d$. Then $\mathsf{H}_f(u, v) = f(uv) = \boldsymbol{\alpha}\boldsymbol{M}_u\boldsymbol{M}_v\boldsymbol{c}$. Thus the infinite Hankel matrix factors through a d-dimensional state space, and hence $\mathrm{rank}(\mathsf{H}_f) \leq d$. Since $\mathrm{rank}(\mathsf{H}_f) = r$, every exact linear recurrent realization satisfies $d \geq r$. □

Proposition 16 (Similarity of minimal exact linear recurrent realizations)**.** *Assume* $\mathrm{rank}(\mathsf{H}_f) = r$. *Let* $(\boldsymbol{\alpha}, \{\boldsymbol{M}_x\}_{x\in\mathcal{A}}, \boldsymbol{c})$ *and* $\left(\widetilde{\boldsymbol{\alpha}}, \{\widetilde{\boldsymbol{M}}_x\}_{x\in\mathcal{A}}, \widetilde{\boldsymbol{c}}\right)$ *be two* r*-dimensional exact linear recurrent realizations. Then there is a unique* $\boldsymbol{S} \in \mathrm{GL}(r)$ *such that* $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}\boldsymbol{S}$, $\widetilde{\boldsymbol{M}}_x = \boldsymbol{S}^{-1}\boldsymbol{M}_x\boldsymbol{S}$, $\widetilde{\boldsymbol{c}} = \boldsymbol{S}^{-1}\boldsymbol{c}$ *for every* $x \in \mathcal{A}$.

Proof of Proposition 16. For a word u, write $\boldsymbol{r}(u) = \boldsymbol{\alpha}\boldsymbol{M}_u$, $\widetilde{\boldsymbol{r}}(u) = \widetilde{\boldsymbol{\alpha}}\widetilde{\boldsymbol{M}}_u$, and for a suffix v, write $\boldsymbol{o}(v) = \boldsymbol{M}_v\boldsymbol{c}$, $\widetilde{\boldsymbol{o}}(v) = \widetilde{\boldsymbol{M}}_v\widetilde{\boldsymbol{c}}$. Then $\mathsf{H}_f(u, v) = \boldsymbol{r}(u)\boldsymbol{o}(v) = \widetilde{\boldsymbol{r}}(u)\widetilde{\boldsymbol{o}}(v)$. Because $\mathsf{H}_f(u, v) = \boldsymbol{r}(u)\boldsymbol{o}(v)$ and $\mathrm{rank}(\mathsf{H}_f) = r$, the reachable rows $\{\boldsymbol{r}(u) : u \in \mathcal{A}^*\}$ must span $\mathbb{R}^{1\times r}$; otherwise the row rank of H_f would be smaller than r. The observable columns $\{\boldsymbol{o}(v) : v \in \mathcal{A}^*\}$ likewise span $\mathbb{R}^{r\times 1}$. The same conclusions hold for the tilded realization.

Consider the assignment $\boldsymbol{r}(u) \longmapsto \widetilde{\boldsymbol{r}}(u)$. To show that this assignment defines a linear map, assume that $\sum_i a_i \boldsymbol{r}(u_i) = 0$. Then, for every $v \in \mathcal{A}^*$, $(\sum_i a_i \widetilde{\boldsymbol{r}}(u_i))\,\widetilde{\boldsymbol{o}}(v) = \sum_i a_i f(u_i v) = 0$. Since the tilded observable columns span $\mathbb{R}^{r\times 1}$, $\sum_i a_i \widetilde{\boldsymbol{r}}(u_i)$ must be zero. Thus the assignment extends uniquely to a well-defined linear map $\mathcal{S} : \mathbb{R}^{1\times r} \to \mathbb{R}^{1\times r}$. Its image contains every tilded reachable row, and these rows span $\mathbb{R}^{1\times r}$. Hence $\mathcal{S}$ is surjective and therefore invertible. It is thus represented by right multiplication with some $\boldsymbol{S} \in \mathrm{GL}(r)$: $\widetilde{\boldsymbol{r}}(u) = \boldsymbol{r}(u)\boldsymbol{S}$. Taking $u = \epsilon$ gives $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}\boldsymbol{S}$. For every u and x, $\boldsymbol{r}(u)\boldsymbol{M}_x\boldsymbol{S} = \widetilde{\boldsymbol{r}}(ux) = \boldsymbol{r}(u)\boldsymbol{S}\widetilde{\boldsymbol{M}}_x$. Reachability gives $\boldsymbol{M}_x\boldsymbol{S} = \boldsymbol{S}\widetilde{\boldsymbol{M}}_x$. Moreover, for every prefix u, $\boldsymbol{r}(u)\boldsymbol{c} = f(u) = \widetilde{\boldsymbol{r}}(u)\widetilde{\boldsymbol{c}} = \boldsymbol{r}(u)\boldsymbol{S}\widetilde{\boldsymbol{c}}$. Since the reachable rows span $\mathbb{R}^{1\times r}$, it follows that $\boldsymbol{c} = \boldsymbol{S}\widetilde{\boldsymbol{c}}$, or equivalently, $\widetilde{\boldsymbol{c}} = \boldsymbol{S}^{-1}\boldsymbol{c}$.

Finally, suppose that $\boldsymbol{S}' \in \mathrm{GL}(r)$ also satisfies $\widetilde{\boldsymbol{r}}(u) = \boldsymbol{r}(u)\boldsymbol{S}'$ for every $u \in \mathcal{A}^*$. Then $\boldsymbol{r}(u)(\boldsymbol{S} - \boldsymbol{S}') = \boldsymbol{0}$ for every $u \in \mathcal{A}^*$. Since the reachable rows span $\mathbb{R}^{1\times r}$, it follows that $\boldsymbol{S} = \boldsymbol{S}'$. □

B Proofs for State Formation on the Orthogonal Quotient

For every $\boldsymbol{R} \in \mathrm{O}(m)$, $\mathcal{L}_{\mathrm{state}}(\boldsymbol{W}\boldsymbol{R}) = \mathcal{L}_{\mathrm{state}}(\boldsymbol{W})$, where $\mathrm{O}(m)$ denotes the orthogonal group. Indeed, the objective depends on $\boldsymbol{W}$ only through $\boldsymbol{W}\boldsymbol{W}^\top$, since $\|\boldsymbol{W}\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{W}\boldsymbol{W}^\top)$.

The map $\pi : \mathrm{GL}(m) \to \mathbb{S}^m_{++},\ \pi(\boldsymbol{W}) = \boldsymbol{W}\boldsymbol{W}^\top$ is invariant under the right action of $\mathrm{O}(m)$. Its fibers are exactly the right-orthogonal orbits. Indeed, if $\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{W}}^\top = \boldsymbol{W}\boldsymbol{W}^\top$, then $\boldsymbol{R} = \boldsymbol{W}^{-1}\widetilde{\boldsymbol{W}} \in \mathrm{O}(m),\ \widetilde{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{R}$. Conversely, right multiplication by an orthogonal matrix leaves the Gram matrix unchanged. Every matrix in $\mathbb{S}^m_{++}$ admits a factorization $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{W}^\top$ with $\boldsymbol{W} \in \mathrm{GL}(m)$. Consequently, $\mathrm{GL}(m)/\mathrm{O}(m) \cong \mathbb{S}^m_{++}$. At finite time, $\boldsymbol{W}(t) \in \mathrm{GL}(m)$; rank loss occurs only in the limit. For rank-k $\boldsymbol{W}$, the right-orthogonal orientation is a gauge choice, while $\boldsymbol{W}\boldsymbol{W}^\top$ is invariant and has rank k.

For a symmetric matrix $\boldsymbol{M}$ with eigendecomposition $\boldsymbol{M} = \boldsymbol{U}\,\mathrm{diag}(\mu_1, \dots, \mu_m)\boldsymbol{U}^\top, \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_m$, define its positive part by $\boldsymbol{M}_+ = \boldsymbol{U}\,\mathrm{diag}\big((\mu_1)_+, \dots, (\mu_m)_+\big)\boldsymbol{U}^\top, (a)_+ = \max\{a, 0\}$.

Proposition 17 (Riccati–Sylvester reduction)**.** *Let $\widetilde{\boldsymbol{G}} \in \mathbb{R}^{m\times m}$ satisfy $\widetilde{\boldsymbol{G}} = \widetilde{\boldsymbol{G}}^\top \succeq 0$, let $\lambda > 0$ satisfy $\lambda \notin \mathrm{spec}(\widetilde{\boldsymbol{G}})$, and set $\boldsymbol{M} = \widetilde{\boldsymbol{G}} - \lambda\boldsymbol{I}_m$. Consider $\dot{\boldsymbol{W}} = (\boldsymbol{M} - \boldsymbol{W}\boldsymbol{W}^\top)\boldsymbol{W}, \boldsymbol{W}(0) = \boldsymbol{W}_0 \in \mathrm{GL}(m)$. Then $\boldsymbol{W}(t)$ exists globally and remains invertible for every finite t. Define the quotient Gram variable and its inverse by $\boldsymbol{X}(t) = \boldsymbol{W}(t)\boldsymbol{W}(t)^\top, \boldsymbol{Y}(t) = \boldsymbol{X}(t)^{-1}$. They satisfy $\dot{\boldsymbol{X}} = \boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M} - 2\boldsymbol{X}^2, \dot{\boldsymbol{Y}} = -\boldsymbol{Y}\boldsymbol{M} - \boldsymbol{M}\boldsymbol{Y} + 2\boldsymbol{I}_m$, and $\boldsymbol{Y}(t) = e^{-\boldsymbol{M}t}\boldsymbol{Y}(0)e^{-\boldsymbol{M}t} + 2\int_0^t e^{-2\boldsymbol{M}s}\,\mathrm{d}s$. Equivalently, $\boldsymbol{X}(t) = e^{\boldsymbol{M}t}\left[\boldsymbol{Y}(0) + 2\int_0^t e^{2\boldsymbol{M}s}\,\mathrm{d}s\right]^{-1} e^{\boldsymbol{M}t}$. Consequently, as $t \to \infty$, $\boldsymbol{X}(t) \longrightarrow \boldsymbol{M}_+$.*

Proof of Proposition 17. The vector field in equation $\dot{\boldsymbol{W}} = (\boldsymbol{M} - \boldsymbol{W}\boldsymbol{W}^\top)\boldsymbol{W}$ is the negative gradient of $\widetilde{\mathcal{L}}(\boldsymbol{W}) = \frac{1}{4}\left\|\boldsymbol{W}\boldsymbol{W}^\top - \widetilde{\boldsymbol{G}}\right\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2$. This objective is coercive and nonincreasing along the gradient flow. Hence the trajectory remains in the bounded sublevel set $\left\{\boldsymbol{W} : \widetilde{\mathcal{L}}(\boldsymbol{W}) \le \widetilde{\mathcal{L}}(\boldsymbol{W}_0)\right\}$, and the solution exists globally. Since the equation has the form $\dot{\boldsymbol{W}} = \boldsymbol{A}(t)\boldsymbol{W},\ \boldsymbol{A}(t) = \boldsymbol{M} - \boldsymbol{W}(t)\boldsymbol{W}(t)^\top$, Liouville's formula gives $\det \boldsymbol{W}(t) = \det \boldsymbol{W}_0 \exp\left(\int_0^t \mathrm{tr}\,\boldsymbol{A}(s)\,\mathrm{d}s\right)$, so $\boldsymbol{W}(t)$ is invertible at every finite time.

Differentiating $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{W}^\top$ gives $\dot{\boldsymbol{X}} = (\boldsymbol{M} - \boldsymbol{X})\boldsymbol{X} + \boldsymbol{X}(\boldsymbol{M} - \boldsymbol{X}) = \boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M} - 2\boldsymbol{X}^2$. Since $\boldsymbol{X} \succ 0$ at every finite time, differentiation of $\boldsymbol{Y} = \boldsymbol{X}^{-1}$ yields $\dot{\boldsymbol{Y}} = -\boldsymbol{Y}\dot{\boldsymbol{X}}\boldsymbol{Y} = -\boldsymbol{Y}\boldsymbol{M} - \boldsymbol{M}\boldsymbol{Y} + 2\boldsymbol{I}_m$. Indeed, $\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\boldsymbol{M}t}\boldsymbol{Y}(t)e^{\boldsymbol{M}t}\right) = 2e^{2\boldsymbol{M}t}$, so integration from 0 to t gives equation $\boldsymbol{Y}(t) = e^{-\boldsymbol{M}t}\boldsymbol{Y}(0)e^{-\boldsymbol{M}t} + 2\int_0^t e^{-2\boldsymbol{M}s}\,\mathrm{d}s$.

It remains to identify the limit of $\boldsymbol{X} = \boldsymbol{Y}^{-1}$. Since $\lambda \notin \mathrm{spec}(\widetilde{\boldsymbol{G}})$, the matrix $\boldsymbol{M}$ has no zero eigenvalues. Diagonalize $\boldsymbol{M}$ and separate its positive and negative spectral subspaces: $\boldsymbol{M} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{B} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{A}\end{bmatrix}\boldsymbol{U}^\top$, where $\boldsymbol{B} \succ 0$ and $\boldsymbol{A} \succ 0$; either block may be absent. Write $\boldsymbol{Y}(t) = e^{-\boldsymbol{M}t}\boldsymbol{R}(t)e^{-\boldsymbol{M}t},\ \boldsymbol{R}(t) = \boldsymbol{Y}(0) + 2\int_0^t e^{2\boldsymbol{M}s}\,\mathrm{d}s$. In the orthogonal basis given by $\boldsymbol{U}$, write $\boldsymbol{Y}(0) = \begin{bmatrix}\boldsymbol{Y}_{++}(0) & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{Y}_{--}(0)\end{bmatrix}$. Then $\boldsymbol{R}(t) = \begin{bmatrix}\boldsymbol{R}_{++}(t) & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{R}_{--}(t)\end{bmatrix}$. The off-diagonal block $\boldsymbol{C}$ is constant, while $\boldsymbol{R}_{++}(t) = \boldsymbol{B}^{-1}e^{2\boldsymbol{B}t} + O(1),\ \boldsymbol{R}_{--}(t) \longrightarrow \boldsymbol{D}_\infty = \boldsymbol{Y}_{--}(0) + \boldsymbol{A}^{-1} \succ 0$. Let $\boldsymbol{S}(t) = \boldsymbol{R}_{++}(t) - \boldsymbol{C}\boldsymbol{R}_{--}(t)^{-1}\boldsymbol{C}^\top$ be the Schur complement. Since $\boldsymbol{S}(t) = \boldsymbol{R}_{++}(t) - \boldsymbol{C}\boldsymbol{R}_{--}(t)^{-1}\boldsymbol{C}^\top$, the block asymptotics above give $e^{-\boldsymbol{B}t}\boldsymbol{S}(t)e^{-\boldsymbol{B}t} = \boldsymbol{B}^{-1} + e^{-\boldsymbol{B}t}O(1)e^{-\boldsymbol{B}t} - e^{-\boldsymbol{B}t}\boldsymbol{C}\boldsymbol{R}_{--}(t)^{-1}\boldsymbol{C}^\top e^{-\boldsymbol{B}t}$. Since $\boldsymbol{B} \succ 0$ and $\boldsymbol{R}_{--}(t)^{-1}$ remains bounded, the last two terms converge to zero. Hence $e^{-\boldsymbol{B}t}\boldsymbol{S}(t)e^{-\boldsymbol{B}t} \longrightarrow \boldsymbol{B}^{-1}$. The limit is positive definite, so the matrices on the left are invertible for all sufficiently large t. By continuity of matrix inversion, $\left(e^{-\boldsymbol{B}t}\boldsymbol{S}(t)e^{-\boldsymbol{B}t}\right)^{-1} \longrightarrow \boldsymbol{B}$. Since $\left(e^{-\boldsymbol{B}t}\boldsymbol{S}(t)e^{-\boldsymbol{B}t}\right)^{-1} = e^{\boldsymbol{B}t}\boldsymbol{S}(t)^{-1}e^{\boldsymbol{B}t}$, we obtain $e^{\boldsymbol{B}t}\boldsymbol{S}(t)^{-1}e^{\boldsymbol{B}t} \longrightarrow \boldsymbol{B}$. The block inverse formula gives $\boldsymbol{R}(t)^{-1} = \begin{bmatrix}\boldsymbol{S}(t)^{-1} & -\boldsymbol{S}(t)^{-1}\boldsymbol{C}\boldsymbol{R}_{--}(t)^{-1}\\ -\boldsymbol{R}_{--}(t)^{-1}\boldsymbol{C}^\top\boldsymbol{S}(t)^{-1} & \boldsymbol{R}_{--}(t)^{-1} + \boldsymbol{R}_{--}(t)^{-1}\boldsymbol{C}^\top\boldsymbol{S}(t)^{-1}\boldsymbol{C}\boldsymbol{R}_{--}(t)^{-1}\end{bmatrix}$. Since $\boldsymbol{C}$ is constant and $\boldsymbol{R}_{--}(t)^{-1}$ remains bounded, the asymptotic behavior of these blocks is controlled by $\boldsymbol{S}(t)^{-1}$ together with the outer exponential factors. Moreover, the preceding convergence implies that $e^{\boldsymbol{B}t}\boldsymbol{S}(t)^{-1}e^{\boldsymbol{B}t}$ remains bounded. For example, the positive-negative block can be written as $-\left[e^{\boldsymbol{B}t}\boldsymbol{S}(t)^{-1}e^{\boldsymbol{B}t}\right]e^{-\boldsymbol{B}t}\boldsymbol{C}\boldsymbol{R}_{--}(t)^{-1}e^{-\boldsymbol{A}t}$, which converges to zero. The other off-diagonal block and the negative-negative block are handled in the same way. Hence, after multiplication by the

outer factors in $\boldsymbol{X}(t) = e^{\boldsymbol{M}t}\boldsymbol{R}(t)^{-1}e^{\boldsymbol{M}t}$, the positive-positive block converges to $\boldsymbol{B}$, while the cross blocks and the negative-negative block contain at least one factor $e^{-\boldsymbol{A}t}$ and converge to zero. Thus $\boldsymbol{X}(t) \to \boldsymbol{M}_+$. The same argument applies when either spectral block is absent. □

Proof of Theorem 5. By Proposition 17, $\boldsymbol{W}(t)\boldsymbol{W}(t)^\top \longrightarrow \boldsymbol{M}_+$. Since $\boldsymbol{G} = \boldsymbol{P}\boldsymbol{\Sigma}^2\boldsymbol{P}^\top$, the positive eigenvalues of $\boldsymbol{M} = \boldsymbol{G} - \lambda\boldsymbol{I}_m$ are $\beta_j = \sigma_j^2 - \lambda$ for $j \in [r]$, while every remaining eigenvalue is $-\lambda$. Therefore, $\boldsymbol{M}_+ = \boldsymbol{P}\,\mathrm{diag}(\beta_1, \ldots, \beta_r)\boldsymbol{P}^\top = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$. Hence $\boldsymbol{W}(t)\boldsymbol{W}(t)^\top \longrightarrow \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$.

The state loss is coercive and real analytic. Since it is nonincreasing along the gradient flow, the trajectory remains in the compact sublevel set $\{\boldsymbol{W} : \mathcal{L}_{\mathrm{state}}(\boldsymbol{W}) \le \mathcal{L}_{\mathrm{state}}(\boldsymbol{W}(0))\}$. The standard Łojasiewicz convergence theorem for bounded analytic gradient trajectories therefore implies that $\boldsymbol{W}(t)$ converges to a limit $\boldsymbol{W}_\infty$ (Absil et al., 2005). Combining the convergence of $\boldsymbol{W}(t)$ with the preceding Gram convergence and using continuity of matrix multiplication gives $\boldsymbol{W}_\infty\boldsymbol{W}_\infty^\top = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$.

For $\boldsymbol{J} = \boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top\boldsymbol{W}_\infty$, we have $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top\boldsymbol{W}_\infty\boldsymbol{W}_\infty^\top\boldsymbol{P}\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^2\boldsymbol{\Lambda}^{-1} = \boldsymbol{I}_r$. Moreover, $\mathrm{col}(\boldsymbol{W}_\infty) = \mathrm{col}(\boldsymbol{W}_\infty\boldsymbol{W}_\infty^\top) = \mathrm{col}(\boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top) = \mathrm{col}(\boldsymbol{P})$, where the last equality uses the invertibility of $\boldsymbol{\Lambda}$. Therefore, $\boldsymbol{W}_\infty = \boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{W}_\infty = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}$. Using $\boldsymbol{Z}_\star = \boldsymbol{P}\boldsymbol{\Sigma}$ and $\boldsymbol{T} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$, we obtain $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{T}\boldsymbol{J} = \boldsymbol{Z}_\star\boldsymbol{T}\boldsymbol{J}$. Therefore, $\boldsymbol{W}_\infty = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J} = \boldsymbol{Z}_\star\boldsymbol{T}\boldsymbol{J}$.

Since $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$ and $\boldsymbol{\Lambda}$ is invertible, $\mathrm{rank}(\boldsymbol{W}_\infty) = r$. Taking the ith row of $\boldsymbol{W}_\infty = \boldsymbol{Z}_\star\boldsymbol{T}\boldsymbol{J}$ gives $\boldsymbol{h}_\infty(u_i) = \boldsymbol{z}_\star(u_i)^\top\boldsymbol{T}\boldsymbol{J}$. □

Proposition 18 (Exponential convergence of the quotient state)**.** *Under the assumptions of Theorem 5, there exist constants $c_X > 0, \rho_X > 0$ depending on $\boldsymbol{W}_0$, $\boldsymbol{G}$, and λ such that $\left\|\boldsymbol{W}(t)\boldsymbol{W}(t)^\top - \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top\right\|_{\mathrm{op}} \le c_X e^{-\rho_X t}$ for every $t \ge 0$. In particular, $\int_0^\infty \left\|\boldsymbol{W}(t)\boldsymbol{W}(t)^\top - \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top\right\|_{\mathrm{op}} \mathrm{d}t < \infty$.*

Proof of Proposition 18. Let $\boldsymbol{M} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{B} & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{A}\end{bmatrix}\boldsymbol{U}^\top$, $\boldsymbol{X}_\infty = \boldsymbol{U}\begin{bmatrix}\boldsymbol{B} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\boldsymbol{U}^\top$, where $\boldsymbol{B} \succ 0$ and, when $r < m$, $\boldsymbol{A} \succ 0$; the negative block is absent when $r = m$. For the state problem, the eigenvalues of $\boldsymbol{B}$ are $\beta_1, \ldots, \beta_r$ and $\boldsymbol{A} = \lambda\boldsymbol{I}_{m-r}$. Set $\boldsymbol{E}(t) = \boldsymbol{X}(t) - \boldsymbol{X}_\infty$. Expanding the Riccati equation around $\boldsymbol{X}_\infty$ gives $\dot{\boldsymbol{E}} = \mathcal{L}(\boldsymbol{E}) - 2\boldsymbol{E}^2$, where $\mathcal{L}(\boldsymbol{E}) = \boldsymbol{M}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{M} - 2\boldsymbol{X}_\infty\boldsymbol{E} - 2\boldsymbol{E}\boldsymbol{X}_\infty$. In the $\boldsymbol{U}$-basis, write $\widetilde{\boldsymbol{E}} = \boldsymbol{U}^\top\boldsymbol{E}\boldsymbol{U} = \begin{bmatrix}\boldsymbol{E}_{++} & \boldsymbol{E}_{+-}\\ \boldsymbol{E}_{-+} & \boldsymbol{E}_{--}\end{bmatrix}$. Then $\boldsymbol{U}^\top\mathcal{L}(\boldsymbol{E})\boldsymbol{U} = -\begin{bmatrix}\boldsymbol{B}\boldsymbol{E}_{++} + \boldsymbol{E}_{++}\boldsymbol{B} & \boldsymbol{B}\boldsymbol{E}_{+-} + \boldsymbol{E}_{+-}\boldsymbol{A}\\ \boldsymbol{A}\boldsymbol{E}_{-+} + \boldsymbol{E}_{-+}\boldsymbol{B} & \boldsymbol{A}\boldsymbol{E}_{--} + \boldsymbol{E}_{--}\boldsymbol{A}\end{bmatrix}$. Let $\boldsymbol{D} = \begin{bmatrix}\boldsymbol{B} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A}\end{bmatrix}$. In the $\boldsymbol{U}$-basis, $\mathcal{L}(\boldsymbol{E}) = -(\boldsymbol{D}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{D})$. Since $\boldsymbol{E}$ is symmetric, $\langle\boldsymbol{E}, \mathcal{L}(\boldsymbol{E})\rangle = -2\,\mathrm{tr}(\boldsymbol{D}\boldsymbol{E}^2) \le -2\lambda_{\min}(\boldsymbol{D})\|\boldsymbol{E}\|_{\mathrm{F}}^2$. Thus the claim holds with $\rho_0 = 2\lambda_{\min}(\boldsymbol{D}) > 0$. When $r = m$, the negative spectral block is absent and $\boldsymbol{D} = \boldsymbol{B}$. Hence, for every symmetric $\boldsymbol{E}$, $\langle\boldsymbol{E}, \mathcal{L}(\boldsymbol{E})\rangle \le -\rho_0\|\boldsymbol{E}\|_{\mathrm{F}}^2$. The quotient convergence in Proposition 17 gives $\boldsymbol{E}(t) \to \boldsymbol{0}$. Hence there is $t_0 < \infty$ such that $2\|\boldsymbol{E}(t)\|_{\mathrm{op}} \le \frac{\rho_0}{2}$ for every $t \ge t_0$. Using $\left|\mathrm{tr}(\boldsymbol{E}^3)\right| \le \|\boldsymbol{E}\|_{\mathrm{op}}\|\boldsymbol{E}\|_{\mathrm{F}}^2$, together with $\dot{\boldsymbol{E}} = \mathcal{L}(\boldsymbol{E}) - 2\boldsymbol{E}^2$, we obtain $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{E}(t)\|_{\mathrm{F}}^2 = \langle\boldsymbol{E}(t), \mathcal{L}(\boldsymbol{E}(t))\rangle_{\mathrm{F}} - 2\,\mathrm{tr}(\boldsymbol{E}(t)^3)$. Therefore, $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{E}(t)\|_{\mathrm{F}}^2 \le (-\rho_0 + 2\|\boldsymbol{E}(t)\|_{\mathrm{op}})\,\|\boldsymbol{E}(t)\|_{\mathrm{F}}^2$. For every $t \ge t_0$, the choice of t_0 gives $2\|\boldsymbol{E}(t)\|_{\mathrm{op}} \le \frac{\rho_0}{2}$, and hence $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{E}(t)\|_{\mathrm{F}}^2 \le -\frac{\rho_0}{2}\|\boldsymbol{E}(t)\|_{\mathrm{F}}^2$. for $t \ge t_0$. Gronwall's inequality gives exponential decay of $\|\boldsymbol{E}(t)\|_{\mathrm{F}}$ for $t \ge t_0$. After enlarging the multiplicative constant, the estimate holds for every $t \ge 0$. Since $\|\boldsymbol{E}(t)\|_{\mathrm{op}} \le \|\boldsymbol{E}(t)\|_{\mathrm{F}}$, the operator-norm estimate follows, and its exponential decay implies the stated integrability. □

C Proofs for Triangular Cascade Formation

Proof of Proposition 6. Let $\mathcal{A} = \{x, y\}$, fix $\sigma > \sqrt{\lambda}$, and define $f(x^k) = \sigma$ for every $k \ge 0$, with $f(w) = 0$ whenever w contains y. The infinite Hankel matrix satisfies $\mathsf{H}_f(u, v) = \sigma\mathbf{1}_{\{u \in \{x\}^*\}}\mathbf{1}_{\{v \in \{x\}^*\}}$, and therefore has rank one. Choose $\mathcal{U} = (\epsilon, y)$, $\mathcal{V} = (\epsilon)$. Then $\mathsf{H} = \mathsf{H}_x = \begin{bmatrix}\sigma\\ 0\end{bmatrix}$, $\mathsf{H}_y = \begin{bmatrix}0\\ 0\end{bmatrix}$ and H has rank one. Set $\beta = \sigma^2 - \lambda > 0$, fix $\nu \ne 0$, and take $\boldsymbol{W}_\infty = \begin{bmatrix}\sqrt{\beta} & 0\\ 0 & 0\end{bmatrix}$, $\boldsymbol{C} = \begin{bmatrix}\sigma/\sqrt{\beta}\\ \nu\end{bmatrix}$, $\boldsymbol{A}_x = \begin{bmatrix}0 & \sigma/(\sqrt{\beta}\nu)\\ 0 & 0\end{bmatrix}$, $\boldsymbol{A}_y = \boldsymbol{0}$. The state Gram matrix

equals $(\mathsf{H}\mathsf{H}^\top - \lambda \boldsymbol{I}_2)_+$. Direct multiplication gives $\boldsymbol{W}_\infty \boldsymbol{C} = \begin{bmatrix}\sigma \\ 0\end{bmatrix} = \mathsf{H},\ \boldsymbol{W}_\infty \boldsymbol{A}_x \boldsymbol{C} = \begin{bmatrix}\sigma \\ 0\end{bmatrix} = \mathsf{H}_x,\ \boldsymbol{W}_\infty \boldsymbol{A}_y \boldsymbol{C} = \boldsymbol{0} = \mathsf{H}_y$. However, $\boldsymbol{A}_x^2 = \boldsymbol{0}$, so $\boldsymbol{e}_1^\top \boldsymbol{W}_\infty \boldsymbol{A}_x^2 \boldsymbol{C} = 0 \neq \sigma = f(x^2)$. □

Proof of Theorem 8. Because $\mathrm{col}(\boldsymbol{W}) = \mathrm{col}(\boldsymbol{Z}_\star)$ and $\boldsymbol{Z}_\star$ has full column rank, $\boldsymbol{W} = \boldsymbol{Z}_\star \boldsymbol{R}_{\mathrm{act}}$, where $\boldsymbol{R}_{\mathrm{act}} = \boldsymbol{Z}_\star^\dagger \boldsymbol{W}$ has full row rank. Moreover, $\boldsymbol{W} = \boldsymbol{Z}_\star \boldsymbol{R}_{\mathrm{act}}$ and $\boldsymbol{R}_{\mathrm{act}} = \boldsymbol{Z}_\star^\dagger \boldsymbol{W}$ imply $\mathrm{row}(\boldsymbol{W}) = \mathrm{row}(\boldsymbol{R}_{\mathrm{act}})$. Hence their common row-space projector is $\boldsymbol{P}_{\mathrm{act}} = \boldsymbol{R}_{\mathrm{act}}^\dagger \boldsymbol{R}_{\mathrm{act}} = \boldsymbol{W}^\dagger \boldsymbol{W}$.

Assume statement (i). Using $\boldsymbol{W} = \boldsymbol{Z}_\star \boldsymbol{R}_{\mathrm{act}}$ and $\mathsf{H} = \boldsymbol{Z}_\star \boldsymbol{Q}^\top$, the identity $\boldsymbol{W}\boldsymbol{C} = \mathsf{H}$ becomes $\boldsymbol{Z}_\star \boldsymbol{R}_{\mathrm{act}} \boldsymbol{C} = \boldsymbol{Z}_\star \boldsymbol{Q}^\top$. Left multiplication by $\boldsymbol{Z}_\star^\dagger$ gives $\boldsymbol{R}_{\mathrm{act}} \boldsymbol{C} = \boldsymbol{Q}^\top$. Support compatibility then yields $\boldsymbol{C} = \boldsymbol{R}_{\mathrm{act}}^\dagger \boldsymbol{R}_{\mathrm{act}} \boldsymbol{C} = \boldsymbol{R}_{\mathrm{act}}^\dagger \boldsymbol{Q}^\top$.

Using $\mathsf{H}_x = \boldsymbol{Z}_\star \boldsymbol{B}_x^\star \boldsymbol{Q}^\top$, the one-symbol identity becomes $\boldsymbol{R}_{\mathrm{act}} \boldsymbol{A}_x \boldsymbol{R}_{\mathrm{act}}^\dagger \boldsymbol{Q}^\top = \boldsymbol{B}_x^\star \boldsymbol{Q}^\top$. Right multiplication by $\boldsymbol{Q}$ and the identity $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_r$ give $\boldsymbol{R}_{\mathrm{act}} \boldsymbol{A}_x \boldsymbol{R}_{\mathrm{act}}^\dagger = \boldsymbol{B}_x^\star$. Since $\boldsymbol{A}_x = \boldsymbol{P}_{\mathrm{act}} \boldsymbol{A}_x \boldsymbol{P}_{\mathrm{act}}$, we obtain $\boldsymbol{A}_x = \boldsymbol{R}_{\mathrm{act}}^\dagger \boldsymbol{B}_x^\star \boldsymbol{R}_{\mathrm{act}}$. This proves statement (ii).

Assume statement (ii), and let $w = x_1 \cdots x_L$. Repeated use of $\boldsymbol{R}_{\mathrm{act}} \boldsymbol{R}_{\mathrm{act}}^\dagger = \boldsymbol{I}_r$ gives $\boldsymbol{W} \boldsymbol{A}_w \boldsymbol{C} = \boldsymbol{Z}_\star \boldsymbol{B}_{x_1}^\star \cdots \boldsymbol{B}_{x_L}^\star \boldsymbol{Q}^\top = \mathsf{H}_w$. Hence statement (iii) holds. Statement (iii) implies statement (i) by considering the empty word and words of length one.

If $\boldsymbol{C} = \boldsymbol{W}^\top \boldsymbol{\Gamma}$, then every column of $\boldsymbol{C}$ lies in $\mathrm{row}(\boldsymbol{W})$, so $\boldsymbol{C} = \boldsymbol{P}_{\mathrm{act}} \boldsymbol{C}$. Conversely, if $\boldsymbol{C} = \boldsymbol{P}_{\mathrm{act}} \boldsymbol{C}$, then $\boldsymbol{C} = \boldsymbol{W}^\top (\boldsymbol{W}^\top)^\dagger \boldsymbol{C}$, so the required representation holds with $\boldsymbol{\Gamma} = (\boldsymbol{W}^\top)^\dagger \boldsymbol{C}$.

Likewise, $\boldsymbol{A}_x = \boldsymbol{W}^\top \boldsymbol{\Omega}_x \boldsymbol{W}$ implies $\boldsymbol{A}_x = \boldsymbol{P}_{\mathrm{act}} \boldsymbol{A}_x \boldsymbol{P}_{\mathrm{act}}$. Conversely, if $\boldsymbol{A}_x = \boldsymbol{P}_{\mathrm{act}} \boldsymbol{A}_x \boldsymbol{P}_{\mathrm{act}}$, then the choice $\boldsymbol{\Omega}_x = (\boldsymbol{W}^\top)^\dagger \boldsymbol{A}_x \boldsymbol{W}^\dagger$ satisfies $\boldsymbol{A}_x = \boldsymbol{W}^\top \boldsymbol{\Omega}_x \boldsymbol{W}$.

Finally, suppose $\boldsymbol{R}_{\mathrm{act}} = \boldsymbol{T}\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$. Since $\boldsymbol{T}$ is invertible and $\boldsymbol{J}$ has orthonormal rows, $\boldsymbol{R}_{\mathrm{act}}^\dagger = \boldsymbol{J}^\top \boldsymbol{T}^{-1}$. Substitution into the formulas in statement (ii) gives the stated expressions. □

Lemma 19 (Semidefinite tracking with an explicit stable–center rate)**.** *Let $\boldsymbol{K} : [0,\infty) \to \mathbb{R}^{d\times d}$ and $\boldsymbol{b} : [0,\infty) \to \mathbb{R}^d$ be continuous, and suppose that $\boldsymbol{K}(t)$ is symmetric and positive semidefinite for every $t \geq 0$. Suppose there are $\boldsymbol{K}_\infty \succeq 0$, $\boldsymbol{b}_\infty \in \mathbb{R}^d$, and constants $c, \rho > 0$ such that $\|\boldsymbol{K}(t) - \boldsymbol{K}_\infty\|_{\mathrm{op}} + \|\boldsymbol{b}(t) - \boldsymbol{b}_\infty\|_2 \leq c e^{-\rho t}$. Assume that $\boldsymbol{z}_\star$ satisfies $\boldsymbol{K}_\infty \boldsymbol{z}_\star = \boldsymbol{b}_\infty$. Let $\boldsymbol{\Pi}$ be the orthogonal projector onto $\mathrm{range}(\boldsymbol{K}_\infty)$ and set $\boldsymbol{\Pi}_0 = \boldsymbol{I}_d - \boldsymbol{\Pi}$. If $\boldsymbol{K}_\infty \neq \boldsymbol{0}$, define $\mu_+ = \lambda_{\min}\Big(\boldsymbol{K}_\infty|_{\mathrm{range}(\boldsymbol{K}_\infty)}\Big) > 0$. If $\boldsymbol{K}_\infty = \boldsymbol{0}$, set $\mu_+ = +\infty$. For every $\eta > 0$ and $\boldsymbol{z}(0) \in \mathbb{R}^d$, the solution of $\dot{\boldsymbol{z}} = -\eta \boldsymbol{K}(t)\boldsymbol{z} + \eta \boldsymbol{b}(t)$ exists globally and converges to a limit $\boldsymbol{z}_\infty$ satisfying $\boldsymbol{\Pi}\boldsymbol{z}_\infty = \boldsymbol{\Pi}\boldsymbol{z}_\star$. Define $\rho_z = \frac{1}{2}\min\{\rho, \eta\mu_+\}$, with the convention $\min\{\rho, +\infty\} = \rho$. There are constants $C_{\mathrm{s}}, C_{\mathrm{c}} < \infty$ such that, for every $t \geq 0$, $\|\boldsymbol{\Pi}(\boldsymbol{z}(t) - \boldsymbol{z}_\star)\|_2 \leq C_{\mathrm{s}} e^{-\rho_z t}$ and $\|\boldsymbol{\Pi}_0(\boldsymbol{z}(t) - \boldsymbol{z}_\infty)\|_2 \leq C_{\mathrm{c}} e^{-\rho t}$. Consequently, there is $C_z < \infty$ such that $\|\boldsymbol{z}(t) - \boldsymbol{z}_\infty\|_2 \leq C_z e^{-\rho_z t}$ for every $t \geq 0$.*

Proof. Continuity of $\boldsymbol{K}$ and $\boldsymbol{b}$ gives a unique local solution. The exponential convergence assumption implies that both coefficient functions are bounded on $[0,\infty)$, so the continuation theorem for linear ordinary differential equations gives a unique global solution. Set $\boldsymbol{e}(t) = \boldsymbol{z}(t) - \boldsymbol{z}_\star$ and $\boldsymbol{r}(t) = \eta\big[\boldsymbol{b}(t) - \boldsymbol{K}(t)\boldsymbol{z}_\star\big]$. Since $\boldsymbol{K}_\infty \boldsymbol{z}_\star = \boldsymbol{b}_\infty$, there is $c_r = \eta c\big(1 + \|\boldsymbol{z}_\star\|_2\big)$ such that $\|\boldsymbol{r}(t)\|_2 \leq c_r e^{-\rho t}$.

The error satisfies $\dot{\boldsymbol{e}} = -\eta \boldsymbol{K}(t)\boldsymbol{e} + \boldsymbol{r}(t)$. Because $\boldsymbol{K}(t) \succeq 0$, the upper right Dini derivative of $\|\boldsymbol{e}(t)\|_2$ is at most $\|\boldsymbol{r}(t)\|_2$. Hence $\|\boldsymbol{e}(t)\|_2 \leq B_e := \|\boldsymbol{e}(0)\|_2 + \frac{c_r}{\rho}$ for every $t \geq 0$.

Write $\boldsymbol{e}_{\mathrm{s}} = \boldsymbol{\Pi}\boldsymbol{e}$ and $\boldsymbol{e}_{\mathrm{c}} = \boldsymbol{\Pi}_0 \boldsymbol{e}$. The stable component satisfies $\dot{\boldsymbol{e}}_{\mathrm{s}} = -\eta \boldsymbol{K}_\infty \boldsymbol{e}_{\mathrm{s}} + \boldsymbol{g}_{\mathrm{s}}(t)$, where $\boldsymbol{g}_{\mathrm{s}}(t) = -\eta \boldsymbol{\Pi}\big(\boldsymbol{K}(t) - \boldsymbol{K}_\infty\big)\boldsymbol{e}(t) + \boldsymbol{\Pi}\boldsymbol{r}(t)$. The preceding bounds give $\|\boldsymbol{g}_{\mathrm{s}}(t)\|_2 \leq c_g e^{-\rho t}$ with $c_g = \eta c B_e + c_r$. On $\mathrm{range}(\boldsymbol{K}_\infty)$, the semigroup satisfies $\|e^{-\eta \boldsymbol{K}_\infty t}\boldsymbol{\Pi}\|_{\mathrm{op}} \leq e^{-\eta\mu_+ t}$. Variation of constants therefore gives $\boldsymbol{e}_{\mathrm{s}}(t) = e^{-\eta \boldsymbol{K}_\infty t}\boldsymbol{e}_{\mathrm{s}}(0) + \int_0^t e^{-\eta \boldsymbol{K}_\infty (t-s)} \boldsymbol{g}_{\mathrm{s}}(s)\,\mathrm{d}s$. The elementary convolution bound for $e^{-\eta\mu_+(t-s)}e^{-\rho s}$ yields a constant $C_{\mathrm{s}} < \infty$ such that $\|\boldsymbol{e}_{\mathrm{s}}(t)\|_2 \leq C_{\mathrm{s}} e^{-\rho_z t}$. When $\boldsymbol{K}_\infty = \boldsymbol{0}$, the stable component is absent and this conclusion is immediate.

The center component satisfies $\dot{\boldsymbol{e}}_{\mathrm{c}} = -\eta \boldsymbol{\Pi}_0\big(\boldsymbol{K}(t) - \boldsymbol{K}_\infty\big)\boldsymbol{e}(t) + \boldsymbol{\Pi}_0 \boldsymbol{r}(t)$. Its derivative has norm at most $c_g e^{-\rho t}$. Therefore the improper integral $\int_0^\infty \dot{\boldsymbol{e}}_{\mathrm{c}}(s)\,\mathrm{d}s$ converges. Define $\boldsymbol{e}_{\mathrm{c},\infty} = \boldsymbol{e}_{\mathrm{c}}(0) + \int_0^\infty \dot{\boldsymbol{e}}_{\mathrm{c}}(s)\,\mathrm{d}s$. Then $\|\boldsymbol{e}_{\mathrm{c}}(t) - \boldsymbol{e}_{\mathrm{c},\infty}\|_2 \leq c_g \rho^{-1} e^{-\rho t}$. Setting $\boldsymbol{z}_\infty = \boldsymbol{z}_\star + \boldsymbol{e}_{\mathrm{c},\infty}$ gives $\boldsymbol{\Pi}\boldsymbol{z}_\infty = \boldsymbol{\Pi}\boldsymbol{z}_\star$. Combining the stable and center bounds proves the stated estimate. □

Proof of Theorem 9. The state trajectory is global by Proposition 17, and Theorem 5 gives $\boldsymbol{X}(t) \longrightarrow \boldsymbol{X}_\infty = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$. Proposition 18 gives exponential convergence.

Vectorizing equation 2 yields $\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{vec}(\boldsymbol{\Gamma}) = -\eta_d\left(\boldsymbol{I}_p \otimes \boldsymbol{X}^2\right)\operatorname{vec}(\boldsymbol{\Gamma}) + \eta_d \operatorname{vec}(\boldsymbol{X}\mathsf{H})$. The coefficient matrix is symmetric positive semidefinite and converges exponentially to $\boldsymbol{K}_{d,\infty} = \boldsymbol{I}_p \otimes \boldsymbol{X}_\infty^2$. The target $\boldsymbol{\Gamma}_\star$ defined in equation 4 satisfies $\boldsymbol{X}_\infty^2\boldsymbol{\Gamma}_\star = \boldsymbol{X}_\infty\mathsf{H}$.

Applying Lemma 19 to the vectorized decoder equation yields exponential convergence of $\boldsymbol{\Gamma}(t)$ to a limit $\boldsymbol{\Gamma}_\infty$ satisfying $\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Gamma}_\infty = \boldsymbol{\Gamma}_\star$. Since $\boldsymbol{X}(t) \to \boldsymbol{X}_\infty = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$, $\boldsymbol{X}_\infty\boldsymbol{\Gamma}_\infty = \boldsymbol{X}_\infty\boldsymbol{\Gamma}_\star = \mathsf{H}$. Consequently, $\boldsymbol{X}(t)\boldsymbol{\Gamma}(t) \to \mathsf{H}$, and $\boldsymbol{R}_{\mathrm{cas}}(t) = \boldsymbol{X}(t)\boldsymbol{\Gamma}(t)$ converges to H exponentially.

For each symbol x, vectorization of equation 3 gives $\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{vec}(\boldsymbol{\Omega}_x) = -\eta_a\left[\boldsymbol{R}_{\mathrm{cas}}\boldsymbol{R}_{\mathrm{cas}}^\top \otimes \boldsymbol{X}^2\right]\operatorname{vec}(\boldsymbol{\Omega}_x) + \eta_a \operatorname{vec}\left(\boldsymbol{X}\mathsf{H}_x\boldsymbol{R}_{\mathrm{cas}}^\top\right)$. The coefficient matrix is symmetric positive semidefinite and converges exponentially to $\boldsymbol{K}_{a,\infty} = \boldsymbol{G} \otimes \boldsymbol{X}_\infty^2$. Using $\mathsf{H}_x = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{B}_x^\star\boldsymbol{Q}^\top$, $\mathsf{H} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^\top$, direct multiplication gives $\boldsymbol{X}_\infty^2\boldsymbol{\Omega}_x^\star\boldsymbol{G} = \boldsymbol{X}_\infty\mathsf{H}_x\mathsf{H}^\top$. Applying Lemma 19 to the vectorized transition equation yields exponential convergence of $\boldsymbol{\Omega}_x(t)$ to a limit $\boldsymbol{\Omega}_{x,\infty}$. The range projector of $\boldsymbol{G} \otimes \boldsymbol{X}_\infty^2$ acts under vectorization as $\boldsymbol{\Omega}_x \longmapsto \boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Omega}_x\boldsymbol{\Pi}_{\mathrm{L}}$, and hence $\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Omega}_{x,\infty}\boldsymbol{\Pi}_{\mathrm{L}} = \boldsymbol{\Omega}_x^\star$. Multiplying this identity by $\boldsymbol{X}_\infty$ on the left and by H on the right gives $\boldsymbol{X}_\infty\boldsymbol{\Omega}_{x,\infty}\mathsf{H} = \mathsf{H}_x$. Therefore, $\boldsymbol{X}(t)\boldsymbol{\Omega}_x(t)\boldsymbol{X}(t)\boldsymbol{\Gamma}(t) \longrightarrow \mathsf{H}_x$.

The auxiliary equations are linear once their upstream trajectories are fixed. The preceding bounds show that their solutions are bounded on every finite and infinite interval, so the full cascade is global.

Finally, Theorem 5 gives $\boldsymbol{W}_\infty = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}$. Combining $\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Gamma}_\infty = \boldsymbol{\Gamma}_\star$ and $\boldsymbol{\Gamma}_\star = \boldsymbol{P}\boldsymbol{\Lambda}^{-2}\boldsymbol{\Sigma}\boldsymbol{Q}^\top \in \mathbb{R}^{m\times p}$ yield $\boldsymbol{W}_\infty^\top\boldsymbol{\Gamma}_\infty = \boldsymbol{J}^\top\boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{\Gamma}_\star = \boldsymbol{J}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^\top = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{Q}^\top$. Similarly, $\boldsymbol{W}_\infty^\top\boldsymbol{\Omega}_{x,\infty}\boldsymbol{W}_\infty = \boldsymbol{J}^\top\boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{\Omega}_x^\star\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J} = \boldsymbol{J}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}\boldsymbol{B}_x^\star\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{J} = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J}$. The inactive auxiliary components disappear because $\boldsymbol{W}_\infty^\top(\boldsymbol{I}_m - \boldsymbol{\Pi}_{\mathrm{L}}) = \boldsymbol{0}$ and $(\boldsymbol{I}_m - \boldsymbol{\Pi}_{\mathrm{L}})\boldsymbol{W}_\infty = \boldsymbol{0}$. □

Proof of Corollary 11. The state-isomorphism identity gives the base case $\boldsymbol{h}_{\mathrm{cas},\infty}(\epsilon) = \boldsymbol{z}_\star(\epsilon)^\top\boldsymbol{T}\boldsymbol{J}$. Assume that $\boldsymbol{h}_{\mathrm{cas},\infty}(u) = \boldsymbol{z}_\star(u)^\top\boldsymbol{T}\boldsymbol{J}$ holds for a prefix u. Using the recursive rollout definition, the limiting transition operator, and $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$, we obtain $\boldsymbol{h}_{\mathrm{cas},\infty}(ux) = \boldsymbol{z}_\star(u)^\top\boldsymbol{T}\boldsymbol{J}\boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J} = \boldsymbol{z}_\star(ux)^\top\boldsymbol{T}\boldsymbol{J}$. The last equality follows from the canonical predictive-state update. Induction therefore proves $\boldsymbol{h}_{\mathrm{cas},\infty}(u) = \boldsymbol{z}_\star(u)^\top\boldsymbol{T}\boldsymbol{J}$ for every prefix u.

Substituting the limiting decoder and using $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$ gives $\boldsymbol{h}_{\mathrm{cas},\infty}(u)\boldsymbol{C}_{\mathrm{cas},\infty} = \boldsymbol{z}_\star(u)^\top\boldsymbol{Q}^\top = \boldsymbol{g}(u)$. Since $v_{j_\epsilon} = \epsilon$, selecting the empty-suffix coordinate gives $\boldsymbol{g}(u)\boldsymbol{e}_{j_\epsilon} = f(u)$. Combining this identity with the recursive rollout and the definition of $\boldsymbol{c}_{\epsilon,\mathrm{cas}}$ yields $f(w) = \boldsymbol{h}_{\mathrm{cas},\infty}(\epsilon)\boldsymbol{A}_{x_1,\mathrm{cas},\infty}\cdots\boldsymbol{A}_{x_L,\mathrm{cas},\infty}\boldsymbol{c}_{\epsilon,\mathrm{cas}}$ for every word $w = x_1\cdots x_L$.

The compressed parameters satisfy $\boldsymbol{\alpha}_{\mathrm{cas}} = \boldsymbol{z}_\star(\epsilon)^\top\boldsymbol{T}$, $\boldsymbol{M}_{x,\mathrm{cas}} = \boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}$, $\boldsymbol{c}_{\mathrm{cas}} = \boldsymbol{T}^{-1}\boldsymbol{c}_\epsilon^\star$. They therefore form an exact r-dimensional linear recurrent realization. The rank lower bound in Theorem 3 proves that this realization is minimal. □

D Proofs for Joint Feedback and Certified Switching

Theorem 20 (Normally attracting exact-realization manifold)**.** *Under Assumption 2, the set $\mathcal{E}_0$ is a smooth embedded $\mathrm{O}(m)$-invariant equilibrium manifold. Every point of $\mathcal{E}_0$ with state factor $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}$, where $\boldsymbol{J} \in \mathrm{St}(r,m)$, induces $\boldsymbol{C} = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{Q}^\top$, $\boldsymbol{A}_x = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J}$. These operators realize $f(w)$ exactly for every $w \in \mathcal{A}^*$. At a point $\boldsymbol{z}_\star = (\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}, \boldsymbol{\Gamma}_\star + \boldsymbol{N}_{\mathrm{d}}, \{\boldsymbol{\Omega}_x^\star + \boldsymbol{N}_x\}_{x\in\mathcal{A}})$, its tangent space consists of perturbations satisfying $\Delta\boldsymbol{W} = \boldsymbol{P}\boldsymbol{\Lambda}\Delta\boldsymbol{J}$ and $\Delta\boldsymbol{J}\boldsymbol{J}^\top + \boldsymbol{J}\Delta\boldsymbol{J}^\top = \boldsymbol{0}$. The auxiliary tangent conditions are $\boldsymbol{P}^\top\Delta\boldsymbol{\Gamma} = \boldsymbol{0}$ and $\boldsymbol{P}^\top\Delta\boldsymbol{\Omega}_x\boldsymbol{P} = \boldsymbol{0}$. for every $x \in \mathcal{A}$. Define the normal spectral gap of the triangular flow by $\gamma_\star = \min\left\{\lambda, 2\beta_r, \eta_d\beta_r^2, \eta_a\beta_r^2\sigma_r^2\right\}$. For every compact set $K \subset \mathcal{E}_0$ and every $\gamma \in (0,\gamma_\star)$, there is a continuous DF_0-invariant splitting along K, $T_{\boldsymbol{z}}\mathcal{Z} = T_{\boldsymbol{z}}\mathcal{E}_0 \oplus E_{\boldsymbol{z}}^{\mathrm{s}}$, and a constant $C_K < \infty$ such that $\left\|e^{tDF_0(\boldsymbol{z})}\boldsymbol{v}\right\|_{\mathcal{Z}} \le C_K e^{-\gamma t}\|\boldsymbol{v}\|_{\mathcal{Z}}$ for every $\boldsymbol{z} \in K$, $\boldsymbol{v} \in E_{\boldsymbol{z}}^{\mathrm{s}}$, and $t \ge 0$. Since F_0 vanishes identically on $\mathcal{E}_0$, the derivative $DF_0(\boldsymbol{z})$ annihilates $T_{\boldsymbol{z}}\mathcal{E}_0$, and the linearized tangent flow is the identity. Hence $\mathcal{E}_0$ is normally attracting on compact subsets. Moreover, there is a neighborhood $\mathcal{U}_K$ of K in $\mathcal{Z}$ that is foliated by local stable leaves based at points of $\mathcal{E}_0$ near K.*

Proof of Theorem 20. The row Stiefel manifold $\mathrm{St}(r, m)$ is smooth, and $\mathcal{K}_\mathrm{d}$ and $\mathcal{K}_\mathrm{a}$ are linear spaces. The map $(\boldsymbol{J}, \boldsymbol{N}_\mathrm{d}, \{\boldsymbol{N}_x\}_{x\in\mathcal{A}}) \longmapsto (\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}, \boldsymbol{\Gamma}_\star + \boldsymbol{N}_\mathrm{d}, \{\boldsymbol{\Omega}_x^\star + \boldsymbol{N}_x\}_{x\in\mathcal{A}})$ is smooth. Since $\boldsymbol{P}\boldsymbol{\Lambda}$ has full column rank, its first component determines $\boldsymbol{J}$ uniquely through $\boldsymbol{J} = \boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top\boldsymbol{W}$. The remaining components determine $\boldsymbol{N}_\mathrm{d}$ and $\{\boldsymbol{N}_x\}_{x\in\mathcal{A}}$ uniquely through $\boldsymbol{N}_\mathrm{d} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}_\star, \boldsymbol{N}_x = \boldsymbol{\Omega}_x - \boldsymbol{\Omega}_x^\star$. Thus the map is injective, its differential is injective at every point, and its inverse on its image is smooth. It is therefore a smooth embedding. Since its image is $\mathcal{E}_0$, the set $\mathcal{E}_0$ is a smooth embedded manifold. Differentiating the parameterization gives the tangent conditions stated in the theorem. The manifold is invariant under the right action of $\mathrm{O}(m)$. Indeed, for every $\boldsymbol{R} \in \mathrm{O}(m)$, $(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J})\,\boldsymbol{R} = \boldsymbol{P}\boldsymbol{\Lambda}\,(\boldsymbol{J}\boldsymbol{R})$, and $(\boldsymbol{J}\boldsymbol{R})\,(\boldsymbol{J}\boldsymbol{R})^\top = \boldsymbol{J}\boldsymbol{R}\boldsymbol{R}^\top\boldsymbol{J}^\top = \boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$. Thus $\boldsymbol{J}\boldsymbol{R} \in \mathrm{St}(r, m)$, while the auxiliary coordinates are unchanged by the right gauge action. Hence the right action maps $\mathcal{E}_0$ into itself.

Fix $\boldsymbol{z}_\star \in \mathcal{E}_0$ and write $\boldsymbol{W}_\star = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}$ and $\boldsymbol{X}_\star = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$. Since $\boldsymbol{M} = \boldsymbol{G} - \lambda\boldsymbol{I}_m$ satisfies $\boldsymbol{M}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{\Lambda}^2$, $\boldsymbol{M}(\boldsymbol{I}_m - \boldsymbol{\Pi}_\mathrm{L}) = -\lambda(\boldsymbol{I}_m - \boldsymbol{\Pi}_\mathrm{L})$, we have $(\boldsymbol{M} - \boldsymbol{X}_\star)\boldsymbol{W}_\star = \boldsymbol{0}$. If $\boldsymbol{N}_\mathrm{d} \in \mathcal{K}_\mathrm{d}$, then $\boldsymbol{X}_\star\boldsymbol{N}_\mathrm{d} = \boldsymbol{0}$, and equation 4 gives $\boldsymbol{X}_\star(\boldsymbol{\Gamma}_\star + \boldsymbol{N}_\mathrm{d}) = \mathsf{H}$. Likewise, $\boldsymbol{N}_x \in \mathcal{K}_\mathrm{a}$ implies $\boldsymbol{X}_\star\boldsymbol{N}_x\mathsf{H} = \boldsymbol{0}$, and equation 5 gives $\boldsymbol{X}_\star(\boldsymbol{\Omega}_x^\star + \boldsymbol{N}_x)\mathsf{H} = \mathsf{H}_x$. Thus every point of $\mathcal{E}_0$ is an equilibrium of the triangular vector field.

At the same point, the induced decoder is $\boldsymbol{C} = \boldsymbol{W}_\star^\top(\boldsymbol{\Gamma}_\star + \boldsymbol{N}_\mathrm{d}) = \boldsymbol{J}^\top\boldsymbol{\Lambda}\boldsymbol{P}^\top(\boldsymbol{\Gamma}_\star + \boldsymbol{N}_\mathrm{d}) = \boldsymbol{J}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}\boldsymbol{Q}^\top = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{Q}^\top$, where $\boldsymbol{P}^\top\boldsymbol{N}_\mathrm{d} = \boldsymbol{0}$. Likewise, the induced transition operator is $\boldsymbol{A}_x = \boldsymbol{W}_\star^\top(\boldsymbol{\Omega}_x^\star + \boldsymbol{N}_x)\boldsymbol{W}_\star = \boldsymbol{J}^\top\boldsymbol{\Lambda}\boldsymbol{P}^\top(\boldsymbol{\Omega}_x^\star + \boldsymbol{N}_x)\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J} = \boldsymbol{J}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}\boldsymbol{B}_x^\star\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{J} = \boldsymbol{J}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J}$, where $\boldsymbol{P}^\top\boldsymbol{N}_x\boldsymbol{P} = \boldsymbol{0}$. Since $\boldsymbol{W}_\star = \boldsymbol{Z}_\star\boldsymbol{T}\boldsymbol{J}$ and $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$, Theorem 8 shows that these operators realize every shifted Hankel block and hence realize $f(w)$ exactly for every $w \in \mathcal{A}^*$.

We next compute the diagonal blocks of the linearization. Complete $\boldsymbol{P}$ to an orthogonal matrix $[\boldsymbol{P}, \boldsymbol{P}_\perp]$ and complete the orthonormal rows of $\boldsymbol{J}$ to an orthogonal matrix whose first r rows are $\boldsymbol{J}$. In these left and right coordinates, $\boldsymbol{W}_\star = \begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}$. Write a state perturbation as $\Delta\boldsymbol{W} = \begin{bmatrix}\boldsymbol{E} & \boldsymbol{F}\\ \boldsymbol{E}_{21} & \boldsymbol{H}\end{bmatrix}$ in these coordinates. Differentiating $F_{0,\boldsymbol{W}}(\boldsymbol{W}) = (\boldsymbol{M} - \boldsymbol{W}\boldsymbol{W}^\top)\boldsymbol{W}$ gives $DF_{0,\boldsymbol{W}}(\boldsymbol{W}_\star)[\Delta\boldsymbol{W}] = \begin{bmatrix}-\boldsymbol{E}\boldsymbol{\Lambda}^2 - \boldsymbol{\Lambda}\boldsymbol{E}^\top\boldsymbol{\Lambda} & \boldsymbol{0}\\ -\lambda\boldsymbol{E}_{21} - \boldsymbol{E}_{21}\boldsymbol{\Lambda}^2 & -\lambda\boldsymbol{H}\end{bmatrix}$. The $\boldsymbol{F}$ block and the matrices $\boldsymbol{E} = \boldsymbol{\Lambda}\boldsymbol{A}$ with $\boldsymbol{A}^\top = -\boldsymbol{A}$ are exactly the state-gauge tangent directions. On the complementary active block, the diagonal modes contract with rates $2\beta_i$, and each off-diagonal pair contracts in one direction with rate $\beta_i + \beta_j$. The lower-left and lower-right blocks contract with rates at least $\lambda + \beta_r$ and λ. Hence every normal state mode contracts with rate at least $\min\{2\beta_r, \lambda\}$. The derivative of the decoder equation with respect to $\boldsymbol{\Gamma}$ at $\boldsymbol{z}_\star$ is $\Delta\boldsymbol{\Gamma} \longmapsto -\eta_d\boldsymbol{X}_\star^2\Delta\boldsymbol{\Gamma}$. Its kernel is $\mathcal{K}_\mathrm{d}$, and its nonzero eigenvalues are $-\eta_d\beta_i^2$. The derivative of the xth transition equation with respect to $\boldsymbol{\Omega}_x$ is $\Delta\boldsymbol{\Omega}_x \longmapsto -\eta_a\boldsymbol{X}_\star^2\Delta\boldsymbol{\Omega}_x\boldsymbol{G}$. Its kernel is $\mathcal{K}_\mathrm{a}$, and its nonzero eigenvalues are $-\eta_a\beta_i^2\sigma_j^2$ for $i, j \in [r]$.

The full Jacobian is block lower triangular in the order $\boldsymbol{W}$, $\boldsymbol{\Gamma}$, and $\{\boldsymbol{\Omega}_x\}_{x\in\mathcal{A}}$. The state calculation above shows that zero is a semisimple eigenvalue of the state diagonal block. Its kernel consists exactly of the $\boldsymbol{F}$ block and the matrices $\boldsymbol{E} = \boldsymbol{\Lambda}\boldsymbol{A}$ with $\boldsymbol{A}^\top = -\boldsymbol{A}$. The decoder and transition diagonal blocks are symmetric negative semidefinite, so their zero eigenvalues are also semisimple, with kernels $\mathcal{K}_\mathrm{d}$ and $\mathcal{K}_\mathrm{a}$, respectively.

Since F_0 vanishes identically on $\mathcal{E}_0$, differentiation along any smooth curve in $\mathcal{E}_0$ gives $T_{\boldsymbol{z}}\mathcal{E}_0 \subseteq \ker DF_0(\boldsymbol{z})$. By block triangularity, the algebraic multiplicity of the zero eigenvalue of $DF_0(\boldsymbol{z})$ is the sum of its algebraic multiplicities on the diagonal blocks. The preceding diagonal-block calculations show that this sum equals $\dim T_{\boldsymbol{z}}\mathcal{E}_0$. It follows that $\ker DF_0(\boldsymbol{z}) = T_{\boldsymbol{z}}\mathcal{E}_0$, and zero is a semisimple eigenvalue of the full Jacobian.

All remaining diagonal eigenvalues have real part at most $-\gamma_\star$. The off-diagonal blocks do not change the spectrum, although they may produce polynomial factors in the matrix exponential. Consequently, for every $\gamma \in (0, \gamma_\star)$, the stable spectral subspace satisfies an estimate of the form in equation $\left\|e^{tDF_0(\boldsymbol{z})}\boldsymbol{v}\right\|_{\mathcal{Z}} \le C_K e^{-\gamma t}\|\boldsymbol{v}\|_{\mathcal{Z}}$.

On a compact set $K \subset \mathcal{E}_0$, the Jacobian depends continuously on the base point and the separation between the zero spectrum and the stable spectrum is uniform. The corresponding spectral projections therefore vary continuously and are uniformly bounded on K. The off-diagonal blocks are also uniformly bounded on K. This yields a continuous DF_0-invariant splitting and a uniform constant C_K.

The linearized flow is the identity on $T_{\boldsymbol{z}}\mathcal{E}_0$, while it contracts exponentially on the stable complement. The graph-transform construction of the local stable foliation is local in the base point. Covering K by finitely many adapted tubular charts and using the uniform estimates above gives a common neighborhood, a continuous family of local stable leaves, and the asymptotic-phase map in the theorem. This is the compact-subset form of the standard stable-foliation theorem for normally hyperbolic invariant manifolds (Fenichel, 1971; 1974). It does not require a global uniform bound along the unbounded inactive auxiliary fibers. □

Proposition 21 (Morse–Bott structure of the weighted joint objectives)**.** *For every $\varepsilon \in (0, 1]$, the set $\mathcal{E}_0$ is exactly the set of global minimizers of $\mathcal{J}_\varepsilon$. At every $\boldsymbol{z}_\star \in \mathcal{E}_0$, the Hessian $\mathcal{H}_\varepsilon(\boldsymbol{z}_\star) = \nabla^2\mathcal{J}_\varepsilon(\boldsymbol{z}_\star)$ is positive semidefinite and satisfies $\ker \mathcal{H}_\varepsilon(\boldsymbol{z}_\star) = T_{\boldsymbol{z}_\star}\mathcal{E}_0$. Let $\mathcal{D}_\varepsilon$ be the positive block-diagonal mobility with block scalar 1 on $\boldsymbol{W}$, η_d/ε on $\boldsymbol{\Gamma}$, and η_a/ε^2 on each $\boldsymbol{\Omega}_x$ block. Then $DF_\varepsilon(\boldsymbol{z}_\star) = -\mathcal{D}_\varepsilon\mathcal{H}_\varepsilon(\boldsymbol{z}_\star)$. This Jacobian is similar to the symmetric negative-semidefinite operator*

$$-\mathcal{D}_\varepsilon^{1/2}\mathcal{H}_\varepsilon(\boldsymbol{z}_\star)\mathcal{D}_\varepsilon^{1/2}. \tag{15}$$

Its zero eigenvalue is semisimple with kernel $T_{\boldsymbol{z}_\star}\mathcal{E}_0$, and every other eigenvalue is strictly negative.

Proof of Proposition 21. Fix $\varepsilon \in (0, 1]$ and $\boldsymbol{z}_\star \in \mathcal{E}_0$. Write $\boldsymbol{X}_\star = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$. The state objective differs by a constant from $\frac{1}{4}\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$, where $\boldsymbol{M} = \boldsymbol{G} - \lambda\boldsymbol{I}_m$. Since $\boldsymbol{W}\boldsymbol{W}^\top$ ranges over the positive-semidefinite cone, its unique optimal Gram matrix is the Frobenius projection $\boldsymbol{M}_+ = \boldsymbol{X}_\star$. Every minimizing factor therefore has the form $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}$ with $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$.

Because $\varepsilon > 0$, both weights ε and ε^2 are strictly positive, and the remaining losses are nonnegative. The decoder residual vanishes exactly when $\boldsymbol{X}_\star\boldsymbol{\Gamma} = \mathsf{H}$. Using $\mathsf{H} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^\top$ and the invertibility of $\boldsymbol{\Lambda}$ gives $\boldsymbol{P}^\top\boldsymbol{\Gamma} = \boldsymbol{\Lambda}^{-2}\boldsymbol{\Sigma}\boldsymbol{Q}^\top = \boldsymbol{P}^\top\boldsymbol{\Gamma}_\star$. Thus $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_\star + \boldsymbol{N}_{\mathrm{d}}$ with $\boldsymbol{N}_{\mathrm{d}} \in \mathcal{K}_{\mathrm{d}}$. Once this identity holds, the xth transition residual vanishes exactly when $\boldsymbol{X}_\star\boldsymbol{\Omega}_x\mathsf{H} = \mathsf{H}_x$. Left multiplication by $\boldsymbol{P}^\top$ and right multiplication by $\boldsymbol{Q}$ give $\boldsymbol{\Lambda}^2\left(\boldsymbol{P}^\top\boldsymbol{\Omega}_x\boldsymbol{P}\right)\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{B}_x^\star$. Hence $\boldsymbol{P}^\top\boldsymbol{\Omega}_x\boldsymbol{P} = \boldsymbol{\Lambda}^{-2}\boldsymbol{\Sigma}\boldsymbol{B}_x^\star\boldsymbol{\Sigma}^{-1} = \boldsymbol{P}^\top\boldsymbol{\Omega}_x^\star\boldsymbol{P}$. Therefore $\boldsymbol{\Omega}_x = \boldsymbol{\Omega}_x^\star + \boldsymbol{N}_x$ with $\boldsymbol{N}_x \in \mathcal{K}_{\mathrm{a}}$. This proves that $\mathcal{E}_0$ is exactly the set of global minimizers of $\mathcal{J}_\varepsilon$.

Define the residual maps $\boldsymbol{r}_{\mathrm{d}}(\boldsymbol{z}) = \boldsymbol{X}\boldsymbol{\Gamma} - \mathsf{H}$ and

$$\boldsymbol{r}_{\mathrm{a},x}(\boldsymbol{z}) = \boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{X}\boldsymbol{\Gamma} - \mathsf{H}_x. \tag{16}$$

Both residuals vanish at $\boldsymbol{z}_\star$. Consequently,

$$\begin{aligned}\mathcal{H}_\varepsilon(\boldsymbol{z}_\star) = \nabla^2\mathcal{L}_{\mathrm{state}}(\boldsymbol{W}_\star) + \varepsilon D\boldsymbol{r}_{\mathrm{d}}(\boldsymbol{z}_\star)^* D\boldsymbol{r}_{\mathrm{d}}(\boldsymbol{z}_\star) \\ + \varepsilon^2 \sum_{x\in\mathcal{A}} D\boldsymbol{r}_{\mathrm{a},x}(\boldsymbol{z}_\star)^* D\boldsymbol{r}_{\mathrm{a},x}(\boldsymbol{z}_\star).\end{aligned} \tag{17}$$

Each term is positive semidefinite. Since the state component of F_0 is $-\nabla\mathcal{L}_{\mathrm{state}}$, the state linearization computed in the proof of Theorem 20 is the negative state Hessian. That calculation shows that $\ker\nabla^2\mathcal{L}_{\mathrm{state}}(\boldsymbol{W}_\star) = \left\{\boldsymbol{P}\boldsymbol{\Lambda}\Delta\boldsymbol{J} : \Delta\boldsymbol{J}\boldsymbol{J}^\top + \boldsymbol{J}\Delta\boldsymbol{J}^\top = \boldsymbol{0}\right\}$. These are exactly the state tangent directions in Theorem 20. Every such perturbation satisfies $\Delta\boldsymbol{X} = \Delta\boldsymbol{W}\boldsymbol{W}_\star^\top + \boldsymbol{W}_\star\Delta\boldsymbol{W}^\top = \boldsymbol{0}$.

Let $\Delta\boldsymbol{z}$ lie in the kernel of $\mathcal{H}_\varepsilon(\boldsymbol{z}_\star)$. Taking the inner product of equation 17 with $\Delta\boldsymbol{z}$ gives $0 = \left\langle\Delta\boldsymbol{W}, \nabla^2\mathcal{L}_{\mathrm{state}}(\boldsymbol{W}_\star)\Delta\boldsymbol{W}\right\rangle + \varepsilon\left\|D\boldsymbol{r}_{\mathrm{d}}(\boldsymbol{z}_\star)[\Delta\boldsymbol{z}]\right\|_{\mathrm{F}}^2 + \varepsilon^2\sum_{x\in\mathcal{A}}\left\|D\boldsymbol{r}_{\mathrm{a},x}(\boldsymbol{z}_\star)[\Delta\boldsymbol{z}]\right\|_{\mathrm{F}}^2$. Every term on the right-hand side is nonnegative, and $\varepsilon > 0$. Hence every term vanishes. In particular, $\Delta\boldsymbol{W}$ lies in the kernel of the state Hessian, so $\Delta\boldsymbol{X} = \boldsymbol{0}$, and every residual derivative vanishes. The decoder derivative reduces to $D\boldsymbol{r}_{\mathrm{d}}(\boldsymbol{z}_\star)[\Delta\boldsymbol{z}] = \boldsymbol{X}_\star\Delta\boldsymbol{\Gamma}$. It is zero iff $\boldsymbol{P}^\top\Delta\boldsymbol{\Gamma} = \boldsymbol{0}$. Under these two conditions, the transition derivative reduces to $D\boldsymbol{r}_{\mathrm{a},x}(\boldsymbol{z}_\star)[\Delta\boldsymbol{z}] = \boldsymbol{X}_\star\Delta\boldsymbol{\Omega}_x\mathsf{H}$. Using $\boldsymbol{X}_\star = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$ and $\mathsf{H} = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^\top$, this derivative vanishes if and only if $\boldsymbol{P}^\top\Delta\boldsymbol{\Omega}_x\boldsymbol{P} = \boldsymbol{0}$. We have therefore obtained exactly the tangent conditions in Theorem 20. The converse follows by the same calculations. This proves $\ker\mathcal{H}_\varepsilon(\boldsymbol{z}_\star) = T_{\boldsymbol{z}_\star}\mathcal{E}_0$.

For $\varepsilon > 0$, the flow is the block-preconditioned gradient flow of $\mathcal{J}_\varepsilon$, so equation $DF_\varepsilon(\boldsymbol{z}_\star) = -\mathcal{D}_\varepsilon\mathcal{H}_\varepsilon(\boldsymbol{z}_\star)$ holds. Since $\mathcal{D}_\varepsilon$ is positive definite, $\mathcal{D}_\varepsilon^{-1/2}DF_\varepsilon(\boldsymbol{z}_\star)\mathcal{D}_\varepsilon^{1/2} = -\mathcal{D}_\varepsilon^{1/2}\mathcal{H}_\varepsilon(\boldsymbol{z}_\star)\mathcal{D}_\varepsilon^{1/2}$. The right-hand side is symmetric and negative semidefinite. Since $\mathcal{D}_\varepsilon^{1/2}$ is invertible, its kernel is $\mathcal{D}_\varepsilon^{-1/2}\ker\mathcal{H}_\varepsilon(\boldsymbol{z}_\star)$, and therefore has dimension $\dim T_{\boldsymbol{z}_\star}\mathcal{E}_0$. Moreover, equation $DF_\varepsilon(\boldsymbol{z}_\star) = -\mathcal{D}_\varepsilon\mathcal{H}_\varepsilon(\boldsymbol{z}_\star)$ and the invertibility of $\mathcal{D}_\varepsilon$ give $\ker DF_\varepsilon(\boldsymbol{z}_\star) = \ker\mathcal{H}_\varepsilon(\boldsymbol{z}_\star) = T_{\boldsymbol{z}_\star}\mathcal{E}_0$. The symmetric operator is diagonalizable, so its zero eigenvalue is semisimple. Similarity transfers this property to $DF_\varepsilon(\boldsymbol{z}_\star)$. Every remaining eigenvalue is strictly negative. □

Theorem 22 (Joint-feedback normal attraction and recurrent-score stability)**.** *Under Assumption 2, let $K \subset \mathcal{E}_0$ be compact and invariant under the right action of $\mathrm{O}(m)$. There are constants $c_K, C_K > 0$ and gauge-invariant neighborhoods $\mathcal{V}'_K \subset \mathcal{V}_K$ of K such that the following statements hold for every $\varepsilon \in [0,1]$.*

(i) The set $\mathcal{E}_0 \cap \mathcal{V}_K$ is an equilibrium manifold of equation 7–equation 9 and is normally attracting. Its local stable foliation admits an asymptotic-phase map $\pi_{\varepsilon,K} : \mathcal{V}'_K \to \mathcal{E}_0$.

(ii) Every trajectory starting in $\mathcal{V}'_K$ remains in $\mathcal{V}_K$ and converges to $\boldsymbol{z}_\infty = \pi_{\varepsilon,K}(\boldsymbol{z}(0)) \in \mathcal{E}_0$. More precisely, $\|\boldsymbol{z}(t) - \boldsymbol{z}_\infty\|_{\mathcal{Z}} \le C_K e^{-c_K t} \operatorname{dist}_{\mathcal{Z}}(\boldsymbol{z}(0), \mathcal{E}_0)$ for every $t \ge 0$.

(iii) The phase map is gauge equivariant. For every $\boldsymbol{R} \in \mathrm{O}(m)$ for which both sides are defined, $\pi_{\varepsilon,K}\left(\rho_{\boldsymbol{R}}(\boldsymbol{z})\right) = \rho_{\boldsymbol{R}}\left(\pi_{\varepsilon,K}(\boldsymbol{z})\right)$.

(iv) For every fixed $L_{\max} < \infty$, there is $C_{K,L_{\max}} < \infty$ such that $\max_{0\le L\le L_{\max}} \max_{w\in\mathcal{A}^L} |\mathsf{s}_w(\boldsymbol{z}(t)) - f(w)| \le C_{K,L_{\max}} e^{-c_K t} \operatorname{dist}_{\mathcal{Z}}(\boldsymbol{z}(0), \mathcal{E}_0)$ for every $t \ge 0$.

Proof of Theorem 22. Every point of $\mathcal{E}_0$ satisfies the state equilibrium equation and the zero-residual identities $\boldsymbol{X}_\star(\boldsymbol{\Gamma}_\star + \boldsymbol{N}_{\mathrm{d}}) = \mathsf{H}$. Hence every gradient in equation 7–equation 9 vanishes on $\mathcal{E}_0$ for every $\varepsilon \in [0,1]$.

Fix a compact $\mathrm{O}(m)$-invariant set $K \subset \mathcal{E}_0$. At $\varepsilon = 0$, Theorem 20 gives a continuous invariant splitting and strict normal contraction on K. Since F_ε depends smoothly on ε and the zero and stable spectra of DF_0 are uniformly separated on K, a continuous nearby invariant splitting and uniform strict normal contraction persist on K for ε in some interval $[0, \varepsilon_0]$.

It remains to cover $\varepsilon \in [\varepsilon_0, 1]$. For $\boldsymbol{z}_\star \in K$ and $\varepsilon > 0$, set $\mathcal{S}_\varepsilon(\boldsymbol{z}_\star) = \mathcal{D}_\varepsilon^{1/2} \mathcal{H}_\varepsilon(\boldsymbol{z}_\star) \mathcal{D}_\varepsilon^{1/2}$. This is a continuous family of symmetric positive-semidefinite operators on the compact set $K \times [\varepsilon_0, 1]$. Proposition 21 shows that its kernel has constant dimension and corresponds, under the mobility similarity, to the tangent space of $\mathcal{E}_0$. Its smallest positive eigenvalue therefore has a positive minimum μ_K on this compact set. By equation 15, every nonzero eigenvalue of $DF_\varepsilon(\boldsymbol{z}_\star)$ is at most $-\mu_K$ for $\varepsilon \in [\varepsilon_0, 1]$. The operators $\mathcal{D}_\varepsilon^{1/2}$ and $\mathcal{D}_\varepsilon^{-1/2}$ are uniformly bounded on this parameter range, so the same similarity gives uniform linearized-flow constants.

Combining the two parameter ranges gives a uniform spectral separation between the zero tangent spectrum and the stable normal spectrum on $K \times [0,1]$. The corresponding spectral projections depend continuously on $(\boldsymbol{z}_\star, \varepsilon)$. For some $c_K > 0$ and $C_K < \infty$, the stable linearized flow therefore satisfies $\left\| e^{tDF_\varepsilon(\boldsymbol{z}_\star)} \boldsymbol{v} \right\|_{\mathcal{Z}} \le C_K e^{-c_K t} \|\boldsymbol{v}\|_{\mathcal{Z}}$ uniformly for $\boldsymbol{z}_\star \in K$, $\varepsilon \in [0,1]$, and stable vectors $\boldsymbol{v}$. Because the family $(\boldsymbol{z}, \varepsilon) \mapsto F_\varepsilon(\boldsymbol{z})$ is smooth and $K \times [0,1]$ is compact, the required local C^1 bounds and tubular-chart estimates can be chosen uniformly. The parameter-dependent stable-foliation theorem, applied in finitely many tubular charts covering K, gives relatively compact neighborhoods $\mathcal{U}'_K \subset \mathcal{U}_K$ of K, local stable leaves, asymptotic-phase maps, and the phase-convergence estimate in part (ii) of Theorem 22. Define $\mathcal{V}'_K = \bigcup_{\boldsymbol{R}\in\mathrm{O}(m)} \rho_{\boldsymbol{R}}(\mathcal{U}'_K)$ and $\mathcal{V}_K = \bigcup_{\boldsymbol{R}\in\mathrm{O}(m)} \rho_{\boldsymbol{R}}(\mathcal{U}_K)$. Since the right action is continuous and $\mathrm{O}(m)$ is compact, $\mathcal{V}'_K$ and $\mathcal{V}_K$ are gauge-invariant open neighborhoods of K, and $\overline{\mathcal{V}_K}$ is compact. Equivariance of the flow transfers the local stable foliation, forward invariance, and the phase-convergence estimate from $\mathcal{U}'_K \subset \mathcal{U}_K$ to $\mathcal{V}'_K \subset \mathcal{V}_K$ (Fenichel, 1971; 1974).

The right gauge action in equation 6 leaves every loss invariant. Hence every F_ε is $\mathrm{O}(m)$-equivariant. The image of a local stable leaf under this action is another local stable leaf. Uniqueness of the foliation gives $\pi_{\varepsilon,K}\left(\rho_{\boldsymbol{R}}(\boldsymbol{z})\right) = \rho_{\boldsymbol{R}}\left(\pi_{\varepsilon,K}(\boldsymbol{z})\right)$.

For each fixed word w, the map $\boldsymbol{z} \mapsto \mathsf{s}_w(\boldsymbol{z})$ is polynomial and therefore continuously differentiable. Since $\overline{\mathcal{V}_K}$ is compact and $\mathcal{Z}$ is finite dimensional, its convex hull $\operatorname{conv}(\overline{\mathcal{V}_K})$ is compact. Define $L_{K,w} = \sup_{\boldsymbol{z}\in\operatorname{conv}(\overline{\mathcal{V}_K})} \|D\mathsf{s}_w(\boldsymbol{z})\|_{\mathrm{op}} < \infty$. If $\boldsymbol{z}_\infty = \pi_{\varepsilon,K}(\boldsymbol{z}(0))$, then $\boldsymbol{z}(t), \boldsymbol{z}_\infty \in \overline{\mathcal{V}_K}$ for every $t \ge 0$. Theorem 20 gives $\mathsf{s}_w(\boldsymbol{z}_\infty) = f(w)$. The mean-value inequality therefore yields $|\mathsf{s}_w(\boldsymbol{z}(t)) - f(w)| \le L_{K,w} \|\boldsymbol{z}(t) - \boldsymbol{z}_\infty\|_{\mathcal{Z}}$ for every $t \ge 0$. For fixed $L_{\max}$, the set $\bigcup_{L=0}^{L_{\max}} \mathcal{A}^L$ is finite. Taking the maximum of $L_{K,w}$ over this finite set and applying the phase-convergence estimate in part (ii) proves the recurrent-score estimate in part (iv). □

Proof of Theorem 12. Fix $\varepsilon \in (0,1]$. Write $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{W}^\top$ and set $J_{\mathrm{bad}} = \frac{1}{4}\|\boldsymbol{G}\|_{\mathrm{F}}^2 + \frac{\varepsilon}{2}\|\mathsf{H}\|_{\mathrm{F}}^2 + \frac{\varepsilon^2}{2}\sum_{x\in\mathcal{A}}\|\mathsf{H}_x\|_{\mathrm{F}}^2$. Expanding the three losses gives the exact identity $\mathcal{J}_\varepsilon - J_{\mathrm{bad}} = \frac{1}{2}\operatorname{tr}\big(\boldsymbol{W}^\top\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})\boldsymbol{W}\big) + \frac{1}{4}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\varepsilon}{2}\|\boldsymbol{X}\boldsymbol{\Gamma}\|_{\mathrm{F}}^2 - \varepsilon^2\sum_{x\in\mathcal{A}}\langle \mathsf{H}_x, \boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{X}\boldsymbol{\Gamma}\rangle_{\mathrm{F}} + \frac{\varepsilon^2}{2}\sum_{x\in\mathcal{A}}\|\boldsymbol{X}\boldsymbol{\Omega}_x\boldsymbol{X}\boldsymbol{\Gamma}\|_{\mathrm{F}}^2$. The terms after the quadratic term have degree at least four in $\boldsymbol{W}$. Thus, at $\boldsymbol{W}=\boldsymbol{0}$, every first derivative vanishes, the state Hessian is the operator $\Delta\boldsymbol{W}\mapsto\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})\Delta\boldsymbol{W}$, and all other Hessian blocks vanish. Since the mobility $\mathcal{D}_\varepsilon$ is positive definite and has identity state block, every point with $\boldsymbol{W}=\boldsymbol{0}$ is an equilibrium of F_ε.

The condition $\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})\succ 0$ is open in $\boldsymbol{\Gamma}$, so $\mathcal{E}_{\mathrm{bad},\varepsilon}$ is a smooth embedded submanifold of the linear subspace $\boldsymbol{W}=\boldsymbol{0}$. It is nonempty: for $\boldsymbol{\Gamma}=-c\mathsf{H}$ with $c\ge(2\varepsilon)^{-1}$, $\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})=\lambda\boldsymbol{I}_m+(2\varepsilon c-1)\boldsymbol{G}\succeq\lambda\boldsymbol{I}_m$. More generally, if $\boldsymbol{\Gamma}=-c\mathsf{H}+\boldsymbol{N}$ and $\|\boldsymbol{N}\|_{\mathrm{op}}<\lambda/(4\varepsilon\|\mathsf{H}\|_{\mathrm{op}})$, then $\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})\succeq(\lambda/2)\boldsymbol{I}_m$, because $\|\boldsymbol{N}\mathsf{H}^\top+\mathsf{H}\boldsymbol{N}^\top\|_{\mathrm{op}}\le 2\|\boldsymbol{N}\|_{\mathrm{op}}\|\mathsf{H}\|_{\mathrm{op}}$.

Fix $\boldsymbol{z}_{\mathrm{b}}=(\boldsymbol{0},\boldsymbol{\Gamma}_{\mathrm{b}},\{\boldsymbol{\Omega}_{x,\mathrm{b}}\}_{x\in\mathcal{A}})\in\mathcal{E}_{\mathrm{bad},\varepsilon}$. On a sufficiently small bounded auxiliary neighborhood of $\boldsymbol{z}_{\mathrm{b}}$, there are constants $b>0$ and $C<\infty$ such that $\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})\succeq b\boldsymbol{I}_m$ and the absolute value of the remainder after the quadratic term is at most $C\|\boldsymbol{W}\|_{\mathrm{F}}^4$ for $\|\boldsymbol{W}\|_{\mathrm{F}}\le 1$. Hence $\mathcal{J}_\varepsilon - J_{\mathrm{bad}}\ge(b/2)\|\boldsymbol{W}\|_{\mathrm{F}}^2 - C\|\boldsymbol{W}\|_{\mathrm{F}}^4$. This is strictly positive for sufficiently small nonzero $\boldsymbol{W}$ and vanishes when $\boldsymbol{W}=\boldsymbol{0}$. Every point of $\mathcal{E}_{\mathrm{bad},\varepsilon}$ is therefore a local minimizer. Its Hessian kernel consists exactly of the auxiliary directions, which form $T_{\boldsymbol{z}_{\mathrm{b}}}\mathcal{E}_{\mathrm{bad},\varepsilon}$.

Proposition 21 identifies $\mathcal{E}_0$ as the global minimizer set. Its decoder and transition losses vanish. The decrease in the state objective from $\boldsymbol{W}=\boldsymbol{0}$ to its optimal Gram matrix is $\frac{1}{4}\sum_{i=1}^r(\sigma_i^2-\lambda)^2$. Subtracting this optimum from J_{bad} gives

$$\mathcal{J}_\varepsilon(\boldsymbol{z}) - \min\mathcal{J}_\varepsilon = \frac{1}{4}\sum_{i=1}^r(\sigma_i^2-\lambda)^2 + \frac{\varepsilon}{2}\|\mathsf{H}\|_{\mathrm{F}}^2 + \frac{\varepsilon^2}{2}\sum_{x\in\mathcal{A}}\|\mathsf{H}_x\|_{\mathrm{F}}^2 > 0. \tag{18}$$

Thus these local minimizers are not global. At $\boldsymbol{W}=\boldsymbol{0}$, the induced decoder, transition operators, and initial state all vanish, so every recurrent score is zero. Since H has positive rank, these scores do not realize the target.

At any point of $\mathcal{E}_{\mathrm{bad},\varepsilon}$, the full linearization is

$$\Delta\dot{\boldsymbol{W}} = -\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})\Delta\boldsymbol{W},\ \Delta\dot{\boldsymbol{\Gamma}}=\boldsymbol{0},\ \Delta\dot{\boldsymbol{\Omega}}_x=\boldsymbol{0}.$$

The state directions form an invariant stable complement to the auxiliary tangent directions. On every compact subset, the smallest eigenvalue of $\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})$ has a positive minimum, so normal contraction is uniform there.

The local stable-foliation theorem therefore gives, around every $\boldsymbol{z}_{\mathrm{b}}\in\mathcal{E}_{\mathrm{bad},\varepsilon}$, an open neighborhood whose trajectories converge exponentially to $\mathcal{E}_{\mathrm{bad},\varepsilon}$. Because full-rank matrices form an open dense subset of $\mathbb{R}^{m\times m}$, every such neighborhood contains a nonempty open set of initial conditions with full-rank $\boldsymbol{W}(0)$.

□

E AN EXPLICIT CERTIFICATE FOR SWITCHING TO JOINT GRADIENTS

E.1 NEAREST-POINT PROJECTION AND FINITE DERIVATIVE BOUNDS

Let

$$N = m^2 + mp + |\mathcal{A}|m^2,\ d_{\mathcal{E}} = rm - \frac{r(r+1)}{2} + (m-r)p + |\mathcal{A}|(m^2-r^2). \tag{19}$$

We use orthonormal vectorization to identify the parameter space with $\mathbb{R}^N$. The matrix $\mathcal{D}_1$ has scalar blocks 1, η_d, and η_a, so equation 10 transforms the fully coupled vector field into $-\nabla\Psi$. Proposition 21 implies that $\widetilde{\mathcal{E}}$ is a smooth manifold of global minimizers of Ψ and

$$\nabla^2\Psi(\boldsymbol{q})\succeq 0,\ \ker\nabla^2\Psi(\boldsymbol{q}) = T_{\boldsymbol{q}}\widetilde{\mathcal{E}},\ \boldsymbol{q}\in\widetilde{\mathcal{E}}. \tag{20}$$

In particular, the Hessian has the constant nullity $d_{\mathcal{E}}$ on this manifold.

Lemma 23 (Explicit nearest-point projection). *Let $\boldsymbol{z} = (\boldsymbol{W}, \boldsymbol{\Gamma}, \{\boldsymbol{\Omega}_x\})$ and $\boldsymbol{y} = \mathcal{D}_1^{-1/2}\boldsymbol{z}$. Suppose that*

$$\boldsymbol{K} = \boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{W} \tag{21}$$

has full row rank. Define

$$\boldsymbol{J} = (\boldsymbol{K}\boldsymbol{K}^\top)^{-1/2}\boldsymbol{K}, \tag{22}$$

and let $\boldsymbol{z_q}$ have blocks

$$\begin{aligned} \boldsymbol{W_q} &= \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}, \\ \boldsymbol{\Gamma_q} &= \boldsymbol{\Gamma}_\star + (\boldsymbol{I}_m - \boldsymbol{\Pi}_{\mathrm{L}})\boldsymbol{\Gamma}, \\ \boldsymbol{\Omega}_{x,\boldsymbol{q}} &= \boldsymbol{\Omega}_x^\star + \boldsymbol{\Omega}_x - \boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Omega}_x\boldsymbol{\Pi}_{\mathrm{L}}. \end{aligned} \tag{23}$$

Then $\boldsymbol{q} = \mathcal{D}_1^{-1/2}\boldsymbol{z_q}$ is the unique nearest point in $\widetilde{\mathcal{E}}$ to $\boldsymbol{y}$. The projection is smooth wherever $\boldsymbol{K}$ has full row rank, and

$$\begin{aligned} \delta^2 = \|\boldsymbol{W} - \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}\|_{\mathrm{F}}^2 + \eta_d^{-1}\|\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_\star\|_{\mathrm{F}}^2 \\ + \eta_a^{-1}\sum_{x\in\mathcal{A}}\|\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Omega}_x\boldsymbol{\Pi}_{\mathrm{L}} - \boldsymbol{\Omega}_x^\star\|_{\mathrm{F}}^2. \end{aligned} \tag{24}$$

Moreover, for any $\boldsymbol{q} \in \widetilde{\mathcal{E}}$, this projection is smooth and unique throughout

$$\|\boldsymbol{y} - \boldsymbol{q}\|_2 < \frac{\beta_r}{2\sqrt{\beta_1}}. \tag{25}$$

Proof. The squared distance in transformed coordinates is

$$\|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\|_{\mathrm{F}}^2 + \eta_d^{-1}\|\boldsymbol{\Gamma} - \widetilde{\boldsymbol{\Gamma}}\|_{\mathrm{F}}^2 + \eta_a^{-1}\sum_x\|\boldsymbol{\Omega}_x - \widetilde{\boldsymbol{\Omega}}_x\|_{\mathrm{F}}^2. \tag{26}$$

The squared distance separates over the state, decoder, and transition coordinates. The decoder and transition formulas are their orthogonal projections onto the affine auxiliary spaces defining $\mathcal{E}_0$; multiplication of a whole block norm by a positive scalar does not change its nearest point. For the state block,

$$\|\boldsymbol{W} - \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}\|_{\mathrm{F}}^2 = \|\boldsymbol{W}\|_{\mathrm{F}}^2 + \mathrm{tr}(\boldsymbol{\Lambda}^2) - 2\langle\boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{W}, \boldsymbol{J}\rangle_{\mathrm{F}} \tag{27}$$

whenever $\boldsymbol{J}\boldsymbol{J}^\top = \boldsymbol{I}_r$. The full-row-rank orthogonal Procrustes problem has the unique maximizer equation 22. This can also be verified from a compact singular value decomposition of $\boldsymbol{K}$: equality in the nuclear-norm upper bound for the inner product forces the row polar factor. The inverse square root is smooth on the positive-definite cone, so the projection is smooth. The distance formula follows by substitution.

Finally, write the state block of a point $\boldsymbol{q} \in \widetilde{\mathcal{E}}$ as $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}_{\boldsymbol{q}}$. Its polar input is $\boldsymbol{\Lambda}^2\boldsymbol{J}_{\boldsymbol{q}}$, whose smallest singular value is β_r. Perturbing the state block by at most $\|\boldsymbol{y} - \boldsymbol{q}\|_2$ perturbs this input in operator norm by at most $\sqrt{\beta_1}\|\boldsymbol{y} - \boldsymbol{q}\|_2$. Weyl's singular-value inequality therefore gives

$$\sigma_{\min}(\boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{W}) \geq \beta_r - \sqrt{\beta_1}\|\boldsymbol{y} - \boldsymbol{q}\|_2 > \frac{\beta_r}{2}. \tag{28}$$

This proves the final claim. □

For a point $\boldsymbol{q} \in \widetilde{\mathcal{E}}$, define

$$\boldsymbol{H}_{\boldsymbol{q}} = \nabla^2\Psi(\boldsymbol{q}) = \mathcal{D}_1^{1/2}\nabla^2\mathcal{J}_1(\mathcal{D}_1^{1/2}\boldsymbol{q})\mathcal{D}_1^{1/2},\ \mu = \lambda_{d_\mathcal{E}+1}(\boldsymbol{H}_{\boldsymbol{q}}) > 0, \tag{29}$$

where the eigenvalues are listed in nondecreasing order. Thus μ is the smallest positive eigenvalue of a finite symmetric matrix; the known nullity avoids an unspecified numerical rank threshold. For numerical certification, one may instead use any certified lower bound $\underline{\mu} \in [\mu/2, \mu]$.

Set

$$B = \|\boldsymbol{q}\|_2 + 1,\ c_d = \sqrt{\eta_d},\ c_a = \sqrt{\eta_d\eta_a},\ h_0 = \|\mathsf{H}\|_{\mathrm{F}},\ h_1 = \sum_{x\in\mathcal{A}}\|\mathsf{H}_x\|_{\mathrm{F}},\ a = |\mathcal{A}|. \tag{30}$$

Let $\boldsymbol{M}_s = \boldsymbol{G} - \lambda \boldsymbol{I}_m$ and define

$$M = 3B^2 + \|\boldsymbol{M}_s\|_{\mathrm{F}} + 15\eta_d B^4 + 6c_d h_0 B + 66a\eta_d\eta_a B^{10} + 30c_a h_1 B^4, \tag{31}$$

$$L = 6B + 60\eta_d B^3 + 6c_d h_0 + 660a\eta_d\eta_a B^9 + 120c_a h_1 B^3. \tag{32}$$

Lemma 24 (Explicit derivative bounds)**.** *On the closed unit ball $\overline{B}(\boldsymbol{q}, 1)$, the constants in equation 31–equation 32 satisfy*

$$\|\nabla^2\Psi(\boldsymbol{y})\|_{\mathrm{op}} \le M,\ \|D^3\Psi(\boldsymbol{y})\|_{\mathrm{op}} \le L. \tag{33}$$

Consequently, the Hessian is L-Lipschitz on this ball.

Proof. Every transformed parameter block has Frobenius norm at most B on this ball. Any degree-d product of such blocks, allowing repeated or transposed factors and multiplied by a scalar coefficient c, has Frobenius norm at most cB^d. Its jth derivative has multilinear operator norm at most

$$\frac{d!}{(d-j)!} cB^{d-j},\ 0 \le j \le d. \tag{34}$$

Indeed, each derivative distributes its arguments among distinct product positions; there are at most $d!/(d-j)!$ resulting terms. Frobenius submultiplicativity bounds each term, and each extracted block of a unit product-norm direction has norm at most one.

If Q is a matrix-valued residual and $\ell = \frac{1}{2}\|Q\|_{\mathrm{F}}^2$, differentiation and Cauchy–Schwarz give

$$\begin{aligned} \|D^2\ell\|_{\mathrm{op}} &\le \|DQ\|_{\mathrm{op}}^2 + \|Q\|_{\mathrm{F}}\|D^2Q\|_{\mathrm{op}}, \\ \|D^3\ell\|_{\mathrm{op}} &\le 3\|DQ\|_{\mathrm{op}}\|D^2Q\|_{\mathrm{op}} + \|Q\|_{\mathrm{F}}\|D^3Q\|_{\mathrm{op}}. \end{aligned} \tag{35}$$

Up to an additive constant, the state objective is $\frac{1}{4}\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}_s\|_{\mathrm{F}}^2$. Its residual has degree two, with value, first derivative, and second derivative bounded by $B^2 + \|\boldsymbol{M}_s\|_{\mathrm{F}}$, $2B$, and 2. Its Hessian and third derivative are therefore bounded by $3B^2 + \|\boldsymbol{M}_s\|_{\mathrm{F}}$ and $6B$.

In transformed coordinates, the decoder product has degree three and coefficient c_d. Its residual, first three derivatives, and the preceding inequalities give bounds

$$15c_d^2B^4 + 6c_d h_0 B,\ 60c_d^2B^3 + 6c_d h_0 \tag{36}$$

for the decoder-loss Hessian and third derivative. Each transition product has degree six and coefficient c_a. Its corresponding bounds are

$$66c_a^2B^{10} + 30c_a\|\mathsf{H}_x\|_{\mathrm{F}}B^4,\ 660c_a^2B^9 + 120c_a\|\mathsf{H}_x\|_{\mathrm{F}}B^3. \tag{37}$$

Summation proves equation 33. Integrating the third derivative along a line segment in the convex unit ball proves Hessian Lipschitzness. □

E.2 Capture by the fully coupled dynamics

Lemma 25 (Explicit local capture)**.** *Let $\boldsymbol{y}_0$ have nearest point $\boldsymbol{q} \in \widetilde{\mathcal{E}}$ from Lemma 23, and compute δ, μ, M, L, b from equation 24, equation 29, equation 31, equation 32, and equation 11. If $(1 + 2M/\mu)\delta < b$, the gradient flow starting at $\boldsymbol{y}_0$ is global, stays in $B(\boldsymbol{q}, b)$, and converges to a point of $\widetilde{\mathcal{E}}$ with bound equation 12. If $(1 + 4M/\mu)\delta < b$ and $0 < h \le \mu/(2M^2)$, the gradient iteration with this fixed h stays in $B(\boldsymbol{q}, b)$ and satisfies equation 14.*

Proof. For $\boldsymbol{v} \in B(\boldsymbol{q}, b)$, write $\boldsymbol{p} = \mathrm{proj}_{\widetilde{\mathcal{E}}}(\boldsymbol{v})$, $\boldsymbol{n} = \boldsymbol{v} - \boldsymbol{p}$, and $d = \|\boldsymbol{n}\|_2$. Lemma 23 makes this projection unique and smooth throughout the ball. Since $\boldsymbol{q}$ is a feasible comparison point,

$$d \le \|\boldsymbol{v} - \boldsymbol{q}\|_2 < b,\ \|\boldsymbol{p} - \boldsymbol{q}\|_2 \le 2\|\boldsymbol{v} - \boldsymbol{q}\|_2 < 2b \le 1. \tag{38}$$

The segment joining $\boldsymbol{p}$ and $\boldsymbol{v}$ lies in the closed unit ball around $\boldsymbol{q}$.

The Hessians at $\boldsymbol{p}$ and $\boldsymbol{q}$ are positive semidefinite with the same nullity $d_{\mathcal{E}}$. Weyl's eigenvalue inequality and Lemma 24 give

$$\lambda_{d_{\mathcal{E}}+1}(\nabla^2\Psi(\boldsymbol{p})) \geq \mu - L\|\boldsymbol{p}-\boldsymbol{q}\|_2 \geq \mu - 2Lb \geq \frac{3\mu}{4}. \tag{39}$$

Nearest-point optimality gives $\boldsymbol{n} \perp T_{\boldsymbol{p}}\widetilde{\mathcal{E}}$. Since $\nabla\Psi(\boldsymbol{p}) = 0$, Taylor's formula gives

$$\nabla\Psi(\boldsymbol{v}) = \nabla^2\Psi(\boldsymbol{p})\boldsymbol{n} + \boldsymbol{r},\ \|\boldsymbol{r}\|_2 \leq \frac{L}{2}d^2. \tag{40}$$

Combining the normal eigenvalue bound with $Lb \leq \mu/8$ yields

$$\langle \boldsymbol{n}, \nabla\Psi(\boldsymbol{v})\rangle \geq \left(\frac{3\mu}{4} - \frac{L}{2}d\right)d^2 \geq \frac{\mu}{2}d^2. \tag{41}$$

Integration of the Hessian along the same segment also gives

$$\|\nabla\Psi(\boldsymbol{v})\|_2 \leq Md. \tag{42}$$

For the continuous dynamics, smoothness of the nearest-point projection implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\operatorname{dist}(\boldsymbol{y}(t),\widetilde{\mathcal{E}})^2 = -\langle \boldsymbol{y}(t) - \boldsymbol{p}(\boldsymbol{y}(t)), \nabla\Psi(\boldsymbol{y}(t))\rangle \leq -\frac{\mu}{2}\operatorname{dist}(\boldsymbol{y}(t),\widetilde{\mathcal{E}})^2 \tag{43}$$

as long as $\boldsymbol{y}(t) \in B(\boldsymbol{q}, b)$. Hence

$$d(t) \leq \delta e^{-\mu t/2},\ \|\dot{\boldsymbol{y}}(t)\|_2 \leq M\delta e^{-\mu t/2}. \tag{44}$$

The accumulated path length is at most $2M\delta/\mu$, and therefore

$$\|\boldsymbol{y}(t) - \boldsymbol{q}\|_2 \leq \left(1 + \frac{2M}{\mu}\right)\delta < b. \tag{45}$$

A first-exit argument extends this estimate to the entire maximal solution. This solution stays in a compact ball on which the polynomial vector field is locally Lipschitz, so it extends to every $t \geq 0$. Its finite path length gives a limit $\boldsymbol{y}_\infty$; since $d(t) \to 0$ and $\widetilde{\mathcal{E}}$ is closed, this limit lies in $\widetilde{\mathcal{E}}$. Integrating the speed bound from t to infinity proves equation 12.

For the discrete iteration, compare the next point with the old nearest point $\boldsymbol{p}$. The bounds in equation 41 and equation 42 give

$$\begin{aligned}\operatorname{dist}(\boldsymbol{v} - h\nabla\Psi(\boldsymbol{v}),\widetilde{\mathcal{E}})^2 &\leq \|\boldsymbol{n} - h\nabla\Psi(\boldsymbol{v})\|_2^2 \\ &\leq (1 - h\mu + h^2M^2)d^2 \leq (1 - h\mu/2)d^2.\end{aligned} \tag{46}$$

Since $\mu \leq M$ and $h \leq \mu/(2M^2)$, the number $\rho_h = \sqrt{1 - h\mu/2}$ belongs to $(0,1)$. Each step has norm at most hMd. Also,

$$1 - \rho_h = \frac{h\mu/2}{1+\rho_h} \geq \frac{h\mu}{4}. \tag{47}$$

If the previous iterates lie in the ball, their distance and cumulative-step estimates imply

$$\|\boldsymbol{y}^{(k+1)} - \boldsymbol{q}\|_2 \leq \delta + \sum_{j=0}^{k} hM\delta\rho_h^j \leq \left(1 + \frac{4M}{\mu}\right)\delta < b. \tag{48}$$

Induction therefore keeps every iterate in the ball and proves $\operatorname{dist}(\boldsymbol{y}^{(k)},\widetilde{\mathcal{E}}) \leq \delta\rho_h^k$. The summable steps give a limit in $\widetilde{\mathcal{E}}$, and their tail satisfies

$$\|\boldsymbol{y}^{(k)} - \boldsymbol{y}_\infty\|_2 \leq \frac{hM}{1-\rho_h}\delta\rho_h^k \leq \frac{4M}{\mu}\delta\rho_h^k. \tag{49}$$

This proves the discrete claim. □

E.3 Finite detection and proof of the switching theorem

Proof of Theorem 13. Theorem 9 implies that the triangular trajectory converges to a finite point $\boldsymbol{z}_\star \in \mathcal{E}_0$. Thus its transformed trajectory converges to $\boldsymbol{y}_\star = \mathcal{D}_1^{-1/2}\boldsymbol{z}_\star \in \widetilde{\mathcal{E}}$. Every finite triangular state factor is invertible, so $\boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{W}(t)$ has full row rank and Lemma 23 defines the projection at every inspection. At the limit its smallest singular value is $\beta_r > 0$. Smoothness of the projection therefore gives

$$\boldsymbol{q}_n \longrightarrow \boldsymbol{y}_\star,\ \delta_n \longrightarrow 0. \tag{50}$$

The finite Hessian depends continuously on $\boldsymbol{q}_n$ and has the same nullity at every projected point. Its smallest positive eigenvalue consequently converges to the strictly positive smallest positive eigenvalue at $\boldsymbol{y}_\star$. The explicit functions $M(\boldsymbol{q}_n)$ and $L(\boldsymbol{q}_n)$ in equation 31–equation 32 also converge to finite positive values. It follows that

$$\frac{b_n}{1 + 4M_n/\mu_n} \tag{51}$$

has a strictly positive limit. If certified eigenvalue lower bounds between one half and the full smallest positive eigenvalue are used, the same quantity instead has a strictly positive lower limit, which is sufficient. Since $\delta_n \to 0$, the strict test succeeds at a finite inspection. The continuous conclusion follows from Lemma 25, because the common certificate is stronger than the continuous certificate.

For a triangular fixed-step sequence, Corollary 14 gives the same convergence to a finite point of $\mathcal{E}_0$ and invertibility at every finite iteration. The preceding argument therefore applies to the integer inspection sequence. After detection, select any fixed h in the nonempty interval $(0, \mu/(2M^2)]$. The second part of Lemma 25 proves the discrete conclusion. This argument chooses the coupled step size at detection; it does not require a prescribed coupled step size to satisfy every possible local certificate.

Both limits lie in $\mathcal{E}_0$ after reversing the linear coordinate transform. The exact-realization result therefore gives a minimal exact recurrent realization after compression. For any fixed finite word w, the function $\boldsymbol{y} \mapsto \mathsf{s}_w(\mathcal{D}_1^{1/2}\boldsymbol{y})$ is a polynomial. Its derivative is bounded on the compact capture ball. Applying the mean-value formula between the trajectory and its limit transfers equation 12 or equation 14 to the prediction error. A common bound exists for all words of any fixed bounded length because there are finitely many such words. □

F Full Fixed-Step State and Triangular Cascade Theory

We first show that, with a sufficiently small fixed step size, gradient descent recovers the predictive subspace from any full-rank initialization. We then prove a discrete stable–center tracking lemma and apply it to the decoder and symbol-indexed transition iterates, which are updated from the initial iteration.

F.1 General full-rank state gradient descent

Set $\boldsymbol{M} = \boldsymbol{G} - \lambda\boldsymbol{I}_m$ and define $\Phi_s(\boldsymbol{W}) = \frac{1}{4}\left\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}\right\|_{\mathrm{F}}^2$. This objective differs from $\mathcal{L}_{\mathrm{state}}$ only by a constant and has gradient $\nabla\Phi_s(\boldsymbol{W}) = \left(\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}\right)\boldsymbol{W}$.

For an arbitrary initialization $\boldsymbol{W}^{(0)} \in \mathrm{GL}(m)$, consider $\boldsymbol{W}^{(k+1)} = \boldsymbol{W}^{(k)} + \eta_s\left(\boldsymbol{M} - \boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)\top}\right)\boldsymbol{W}^{(k)}$. Let $\mathcal{K}_0 = \left\{\boldsymbol{W} : \Phi_s(\boldsymbol{W}) \le \Phi_s(\boldsymbol{W}^{(0)})\right\}$, and define $D_0 = 2\sqrt{\Phi_s(\boldsymbol{W}^{(0)})}$, $G_0 = \max_{\boldsymbol{W}\in\mathcal{K}_0}\|\nabla\Phi_s(\boldsymbol{W})\|_{\mathrm{F}}$. Let $L_0 < \infty$ be any Lipschitz constant of $\nabla\Phi_s$ on $\mathcal{N}_0 = \{\boldsymbol{W} : \mathrm{dist}(\boldsymbol{W}, \mathcal{K}_0) \le 1\}$. The set $\mathcal{K}_0$ is compact, so G_0 and L_0 are finite. We use the convention $a^{-1} = +\infty$ when $a = 0$.

Theorem 26 (Fixed-step predictive-subspace formation from a full-rank initialization)**.** *Under Assumption 2, let $\boldsymbol{W}^{(0)} \in \mathrm{GL}(m)$ and define $\boldsymbol{W}^{(k+1)} = \boldsymbol{W}^{(k)} + \eta_s\left(\boldsymbol{M} - \boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)\top}\right)\boldsymbol{W}^{(k)}$. Assume $0 < \eta_s < \min\left\{G_0^{-1}, D_0^{-1}, L_0^{-1}, \beta_1^{-1}, \frac{2}{\sigma_1^2}, \frac{2}{\lambda}\right\}$. Then every iterate is invertible, and $\Phi_s(\boldsymbol{W}^{(k)})$ is nonincreasing. The factor sequence converges to a limit $\boldsymbol{W}_\infty^{\mathrm{d}}$, and $\boldsymbol{X}^{(k)} := \boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)\top} \longrightarrow \boldsymbol{X}_\infty := (\boldsymbol{G} - \lambda\boldsymbol{I}_m)_+ = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$. Define $q_{\mathrm{loc}} =$*

$\max\left\{\max_{i,j\in[r]}|1-\eta_s(\beta_i+\beta_j)|, \max_{i\in[r]}|1-\eta_s(\beta_i+\lambda)|, (1-\eta_s\lambda)^2\right\}$. *Terms involving the inactive subspace are omitted when $r=m$. The step-size condition implies $q_{\mathrm{loc}}<1$.*

For every $q_X\in(q_{\mathrm{loc}},1)$, there is $C_X<\infty$ such that $\|\boldsymbol{X}^{(k)}-\boldsymbol{X}_\infty\|_{\mathrm{op}}\le C_X q_X^k$ for every $k\ge 0$. Setting $q_W=\sqrt{q_X}$, there is $C_W<\infty$ such that $\|\boldsymbol{W}^{(k)}-\boldsymbol{W}_\infty^{\mathrm{d}}\|_{\mathrm{F}}\le C_W q_W^k$ for every $k\ge 0$.

Define

$$\boldsymbol{J}_{\mathrm{d}}=\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top\boldsymbol{W}_\infty^{\mathrm{d}}. \tag{52}$$

Then $\boldsymbol{J}_{\mathrm{d}}\boldsymbol{J}_{\mathrm{d}}^\top=\boldsymbol{I}_r,\ \boldsymbol{W}_\infty^{\mathrm{d}}=\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}_{\mathrm{d}}=\boldsymbol{Z}_\star\boldsymbol{T}\boldsymbol{J}_{\mathrm{d}}$.

F.2 Discrete semidefinite tracking

Lemma 27 (Discrete semidefinite tracking with an explicit stable–center rate)**.** *Let $\{\boldsymbol{K}_k\}_{k\ge0}\subset\mathbb{R}^{d\times d}$ and $\{\boldsymbol{b}_k\}_{k\ge0}\subset\mathbb{R}^d$, where every $\boldsymbol{K}_k$ is symmetric and positive semidefinite. Suppose there are $\boldsymbol{K}_\infty\succeq 0$, $\boldsymbol{b}_\infty\in\mathbb{R}^d$, $c<\infty$, and $q\in(0,1)$ such that $\|\boldsymbol{K}_k-\boldsymbol{K}_\infty\|_{\mathrm{op}}+\|\boldsymbol{b}_k-\boldsymbol{b}_\infty\|_2\le cq^k$. Assume that $\boldsymbol{z}_\star$ satisfies $\boldsymbol{K}_\infty\boldsymbol{z}_\star=\boldsymbol{b}_\infty$. Let $\overline{K}=\sup_{k\ge0}\|\boldsymbol{K}_k\|_{\mathrm{op}}<\infty$ and choose $0<\eta\le\overline{K}^{-1}$, with the convention $\overline{K}^{-1}=+\infty$ when $\overline{K}=0$. Let $\boldsymbol{\Pi}$ be the orthogonal projector onto $\mathrm{range}(\boldsymbol{K}_\infty)$ and set $\boldsymbol{\Pi}_0=\boldsymbol{I}_d-\boldsymbol{\Pi}$. If $\boldsymbol{K}_\infty\neq\boldsymbol{0}$, define $\mu_+=\lambda_{\min}\Big(\boldsymbol{K}_\infty|_{\mathrm{range}(\boldsymbol{K}_\infty)}\Big)>0,\ q_+=1-\eta\mu_+$. If $\boldsymbol{K}_\infty=\boldsymbol{0}$, set $q_+=0$. Then every sequence satisfying $\boldsymbol{z}_{k+1}=\boldsymbol{z}_k-\eta\boldsymbol{K}_k\boldsymbol{z}_k+\eta\boldsymbol{b}_k$ converges to a limit $\boldsymbol{z}_\infty$ satisfying $\boldsymbol{\Pi}\boldsymbol{z}_\infty=\boldsymbol{\Pi}\boldsymbol{z}_\star$. For every q_z satisfying $\max\{q,q_+\}<q_z<1$, there is $C_z<\infty$ such that $\|\boldsymbol{z}_k-\boldsymbol{z}_\infty\|_2\le C_z q_z^k$ for every $k\ge0$.*

F.3 Fixed-step triangular cascade

Let $\boldsymbol{X}^{(k)}$ be the state sequence from Theorem 26. Starting the downstream variables at iteration zero, define $\boldsymbol{R}_{\mathrm{cas}}^{(k)}=\boldsymbol{X}^{(k)}\boldsymbol{\Gamma}^{(k)}$ and update $\boldsymbol{\Gamma}^{(k+1)}=\boldsymbol{\Gamma}^{(k)}-\eta_d\boldsymbol{X}^{(k)}\left(\boldsymbol{X}^{(k)}\boldsymbol{\Gamma}^{(k)}-\mathsf{H}\right),\ \boldsymbol{\Omega}_x^{(k+1)}=\boldsymbol{\Omega}_x^{(k)}-\eta_a\boldsymbol{X}^{(k)}\left(\boldsymbol{X}^{(k)}\boldsymbol{\Omega}_x^{(k)}\boldsymbol{R}_{\mathrm{cas}}^{(k)}-\mathsf{H}_x\right)\boldsymbol{R}_{\mathrm{cas}}^{(k)\top}$ for every $x\in\mathcal{A}$. The induced hidden-space operators are $\boldsymbol{C}_{\mathrm{cas}}^{(k)}=\boldsymbol{W}^{(k)\top}\boldsymbol{\Gamma}^{(k)},\ \boldsymbol{A}_{x,\mathrm{cas}}^{(k)}=\boldsymbol{W}^{(k)\top}\boldsymbol{\Omega}_x^{(k)}\boldsymbol{W}^{(k)}$.

Theorem 28 (Fixed-step gauge-compatible triangular-cascade formation)**.** *Assume the conditions of Theorem 26. Fix $q_X\in(q_{\mathrm{loc}},1)$ and let C_X be a corresponding constant from that theorem. Set $B_X=\sup_{k\ge0}\|\boldsymbol{X}^{(k)}\|_{\mathrm{op}}$. Choose $0<\eta_d\le B_X^{-2}$. Define $B_\Gamma=\|\boldsymbol{\Gamma}_\star\|_{\mathrm{F}}+\|\boldsymbol{\Gamma}^{(0)}-\boldsymbol{\Gamma}_\star\|_{\mathrm{F}}+\frac{\eta_d B_X C_X\|\boldsymbol{\Gamma}_\star\|_{\mathrm{F}}}{1-q_X},\ B_R=B_X B_\Gamma$. Choose $0<\eta_a\le\left(B_X^2B_R^2\right)^{-1}$.*

For arbitrary finite initial values $\boldsymbol{\Gamma}^{(0)}$ and $\boldsymbol{\Omega}_x^{(0)}$, the downstream sequences converge exponentially to limits $\boldsymbol{\Gamma}_\infty^{\mathrm{d}}$ and $\boldsymbol{\Omega}_{x,\infty}^{\mathrm{d}}$. Their active components satisfy $\boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Gamma}_\infty^{\mathrm{d}}=\boldsymbol{\Gamma}_\star,\ \boldsymbol{\Pi}_{\mathrm{L}}\boldsymbol{\Omega}_{x,\infty}^{\mathrm{d}}\boldsymbol{\Pi}_{\mathrm{L}}=\boldsymbol{\Omega}_x^\star$. The realized blocks satisfy $\boldsymbol{X}^{(k)}\boldsymbol{\Gamma}^{(k)}\longrightarrow\mathsf{H},\ \boldsymbol{X}^{(k)}\boldsymbol{\Omega}_x^{(k)}\boldsymbol{X}^{(k)}\boldsymbol{\Gamma}^{(k)}\longrightarrow\mathsf{H}_x$. The induced operators converge to $\boldsymbol{C}_{\mathrm{cas},\infty}^{\mathrm{d}}=\boldsymbol{J}_{\mathrm{d}}^\top\boldsymbol{T}^{-1}\boldsymbol{Q}^\top,\ \boldsymbol{A}_{x,\mathrm{cas},\infty}^{\mathrm{d}}=\boldsymbol{J}_{\mathrm{d}}^\top\boldsymbol{T}^{-1}\boldsymbol{B}_x^\star\boldsymbol{T}\boldsymbol{J}_{\mathrm{d}}$. These induced limits are independent of the initial auxiliary variables. Their compression to $\mathrm{row}(\boldsymbol{W}_\infty^{\mathrm{d}})$ is an exact r-dimensional recurrent realization and is therefore minimal.

F.4 Proofs

Lemma 29 (Classification of critical Gram matrices)**.** *Let $\overline{\boldsymbol{W}}$ be a critical point of $\Phi_s(\boldsymbol{W})=\frac14\|\boldsymbol{W}\boldsymbol{W}^\top-\boldsymbol{M}\|_{\mathrm{F}}^2$, and set $\overline{\boldsymbol{X}}=\overline{\boldsymbol{W}}\,\overline{\boldsymbol{W}}^\top$. Then $\overline{\boldsymbol{X}}$ commutes with $\boldsymbol{M}$. For each positive eigenvalue μ of $\boldsymbol{M}$, let E_μ denote its eigenspace. There is an orthogonal projector $\boldsymbol{P}_\mu$ with $\mathrm{range}(\boldsymbol{P}_\mu)\subseteq E_\mu$ such that $\overline{\boldsymbol{X}}=\sum_{\mu>0}\mu\boldsymbol{P}_\mu$. In particular, $\overline{\boldsymbol{X}}=\boldsymbol{M}_+$ if and only if $\boldsymbol{P}_\mu$ is the identity on E_μ for every $\mu>0$.*

Proof. The critical-point equation is $(\overline{\boldsymbol{X}}-\boldsymbol{M})\overline{\boldsymbol{W}}=\boldsymbol{0}$. Right multiplication by $\overline{\boldsymbol{W}}^\top$ gives $(\overline{\boldsymbol{X}}-\boldsymbol{M})\overline{\boldsymbol{X}}=\boldsymbol{0}$, and transposition gives $\overline{\boldsymbol{X}}(\overline{\boldsymbol{X}}-\boldsymbol{M})=\boldsymbol{0}$. Hence $\boldsymbol{M}\overline{\boldsymbol{X}}=\overline{\boldsymbol{X}}\boldsymbol{M}=\overline{\boldsymbol{X}}^2$. Therefore $\overline{\boldsymbol{X}}$ preserves every eigenspace E_μ of $\boldsymbol{M}$. On E_μ, it satisfies $\overline{\boldsymbol{X}}^2=\mu\overline{\boldsymbol{X}}$. Since $\overline{\boldsymbol{X}}\succeq0$, it vanishes when $\mu\le0$, while for $\mu>0$ its eigenvalues are either 0 or μ. Thus its restriction to E_μ is $\mu\boldsymbol{P}_\mu$ for an orthogonal projector $\boldsymbol{P}_\mu$. □

Lemma 30 (Exclusion of an unstable limiting multiplier)**.** *Let $\boldsymbol{W}_{k+1} = \boldsymbol{A}_k \boldsymbol{W}_k$, where every $\boldsymbol{W}_k$ is invertible, $\boldsymbol{W}_k \to \boldsymbol{W}_\infty$, every $\boldsymbol{A}_k$ is symmetric and positive semidefinite, and $\boldsymbol{A}_k \to \boldsymbol{A}_\infty$. Then $\boldsymbol{A}_\infty$ has no eigenvalue strictly larger than one.*

Proof. Suppose that $\boldsymbol{A}_\infty$ has an eigenvalue larger than one. Let $\mathcal{U}$ be the direct sum of its eigenspaces with eigenvalues larger than one, and let $\mathcal{C} = \mathcal{U}^\perp$. Write $\boldsymbol{P}$ and $\boldsymbol{Q}$ for the corresponding orthogonal projectors. Since $\boldsymbol{A}_\infty$ is symmetric and positive semidefinite, both subspaces are invariant and every eigenvalue on $\mathcal{C}$ lies in $[0, 1]$. Set $a = \lambda_{\min}\left(\boldsymbol{A}_\infty|_{\mathcal{U}}\right) > 1,\ b = \|\boldsymbol{A}_\infty|_{\mathcal{C}}\|_{\mathrm{op}} \le 1$.

Choose $\varepsilon > 0$ such that $a - 2\varepsilon > 1,\ b + 2\varepsilon \le a - 2\varepsilon$. There is $K < \infty$ such that $\|\boldsymbol{A}_k - \boldsymbol{A}_\infty\|_{\mathrm{op}} \le \varepsilon$ for every $k \ge K$.

Consider the cone $\mathcal{C}_{\mathcal{U}} = \{\boldsymbol{v} : \|\boldsymbol{Q}\boldsymbol{v}\|_2 \le \|\boldsymbol{P}\boldsymbol{v}\|_2\}$. For $\boldsymbol{v} \in \mathcal{C}_{\mathcal{U}}$ and $k \ge K$, $\|\boldsymbol{P}\boldsymbol{A}_k\boldsymbol{v}\|_2 \ge a\|\boldsymbol{P}\boldsymbol{v}\|_2 - \varepsilon\|\boldsymbol{v}\|_2 \ge (a - 2\varepsilon)\|\boldsymbol{P}\boldsymbol{v}\|_2$, where we used $\|\boldsymbol{v}\|_2 \le 2\|\boldsymbol{P}\boldsymbol{v}\|_2$. Similarly, $\|\boldsymbol{Q}\boldsymbol{A}_k\boldsymbol{v}\|_2 \le b\|\boldsymbol{Q}\boldsymbol{v}\|_2 + \varepsilon\|\boldsymbol{v}\|_2 \le (b + 2\varepsilon)\|\boldsymbol{P}\boldsymbol{v}\|_2 \le \|\boldsymbol{P}\boldsymbol{A}_k\boldsymbol{v}\|_2$. Thus the cone is forward invariant, and its $\mathcal{U}$ component expands by at least $\gamma = a - 2\varepsilon > 1$.

Choose a nonzero $\boldsymbol{v}_K \in \mathcal{U}$. Since $\boldsymbol{W}_K$ is invertible, there is a vector $\boldsymbol{c}$ such that $\boldsymbol{v}_K = \boldsymbol{W}_K\boldsymbol{c}$. Define $\boldsymbol{v}_k = \boldsymbol{W}_k\boldsymbol{c}$. Then $\boldsymbol{v}_{k+1} = \boldsymbol{A}_k\boldsymbol{v}_k$. The cone invariance and expansion estimates give $\|\boldsymbol{P}\boldsymbol{v}_k\|_2 \ge \gamma^{k-K}\|\boldsymbol{v}_K\|_2$. Hence $\|\boldsymbol{v}_k\|_2 \to \infty$. This contradicts $\boldsymbol{W}_k \to \boldsymbol{W}_\infty$, which implies that $\boldsymbol{W}_k\boldsymbol{c}$ is bounded. □

Lemma 31 (Local contraction of the quotient update)**.** *Let $\mathcal{T}(\boldsymbol{X}) = [\boldsymbol{I}_m + \eta_s(\boldsymbol{M} - \boldsymbol{X})]\boldsymbol{X}[\boldsymbol{I}_m + \eta_s(\boldsymbol{M} - \boldsymbol{X})]$, and let $\boldsymbol{X}_\infty = \boldsymbol{M}_+$. Under the step-size condition in Theorem 26, the derivative of $\mathcal{T}$ at $\boldsymbol{X}_\infty$, restricted to symmetric matrices with the Frobenius norm, has operator norm $q_{\mathrm{loc}} < 1$. For every $\overline{q} \in (q_{\mathrm{loc}}, 1)$, there is $\varepsilon > 0$ such that $\|\mathcal{T}(\boldsymbol{X}) - \mathcal{T}(\boldsymbol{Y})\|_{\mathrm{F}} \le \overline{q}\|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathrm{F}}$ whenever $\boldsymbol{X}$ and $\boldsymbol{Y}$ are symmetric and satisfy $\|\boldsymbol{X} - \boldsymbol{X}_\infty\|_{\mathrm{F}} \le \varepsilon$ and $\|\boldsymbol{Y} - \boldsymbol{X}_\infty\|_{\mathrm{F}} \le \varepsilon$.*

Proof. In an eigenbasis of $\boldsymbol{M}$, the derivative multipliers are $1 - \eta_s(\beta_i + \beta_j)$ on active–active entries, $1 - \eta_s(\beta_i + \lambda)$ on active–inactive entries, and $(1 - \eta_s\lambda)^2$ on inactive–inactive entries. The step-size condition makes their absolute values strictly smaller than one, so the derivative norm is $q_{\mathrm{loc}} < 1$. Continuity of $D\mathcal{T}$ gives a neighborhood on which its norm is at most $\overline{q}$. The mean-value formula on the convex symmetric-matrix ball then gives the stated contraction. □

Proof of Theorem 26. The original state objective differs from Φ_s only by a constant, and $\nabla\Phi_s(\boldsymbol{W}) = (\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M})\boldsymbol{W}$. The objective Φ_s is coercive, so $\mathcal{K}_0$ is compact.

Assume inductively that $\boldsymbol{W}^{(k)} \in \mathcal{K}_0$. The update displacement satisfies $\|\boldsymbol{W}^{(k+1)} - \boldsymbol{W}^{(k)}\|_{\mathrm{F}} = \eta_s\|\nabla\Phi_s(\boldsymbol{W}^{(k)})\|_{\mathrm{F}} < 1$. The line segment between the two iterates therefore lies in $\mathcal{N}_0$. The descent lemma gives

$$\Phi_s(\boldsymbol{W}^{(k+1)}) \le \Phi_s(\boldsymbol{W}^{(k)}) - \eta_s\left(1 - \frac{L_0\eta_s}{2}\right)\|\nabla\Phi_s(\boldsymbol{W}^{(k)})\|_{\mathrm{F}}^2 \le \Phi_s(\boldsymbol{W}^{(k)}) - \frac{\eta_s}{2}\|\nabla\Phi_s(\boldsymbol{W}^{(k)})\|_{\mathrm{F}}^2. \tag{53}$$

Thus $\boldsymbol{W}^{(k+1)} \in \mathcal{K}_0$, and induction proves descent and boundedness for every iteration.

Set $\boldsymbol{X}^{(k)} = \boldsymbol{W}^{(k)}\boldsymbol{W}^{(k)\top},\ \boldsymbol{A}_k = \boldsymbol{I}_m + \eta_s(\boldsymbol{M} - \boldsymbol{X}^{(k)})$. Then $\boldsymbol{W}^{(k+1)} = \boldsymbol{A}_k\boldsymbol{W}^{(k)},\ \boldsymbol{X}^{(k+1)} = \boldsymbol{A}_k\boldsymbol{X}^{(k)}\boldsymbol{A}_k$. Since the objective decreases, $\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\|_{\mathrm{op}} \le \|\boldsymbol{X}^{(k)} - \boldsymbol{M}\|_{\mathrm{F}} \le D_0$. The step-size condition therefore gives $\boldsymbol{A}_k \succeq (1 - \eta_s D_0)\boldsymbol{I}_m \succ 0$. Every $\boldsymbol{A}_k$ is invertible, so every finite iterate $\boldsymbol{W}^{(k)}$ remains invertible.

The objective is real analytic, the sequence is bounded, and equation 53 supplies sufficient decrease. The exact gradient step also gives $\|\nabla\Phi_s(\boldsymbol{W}^{(k)})\|_{\mathrm{F}} = \eta_s^{-1}\|\boldsymbol{W}^{(k+1)} - \boldsymbol{W}^{(k)}\|_{\mathrm{F}}$. Thus the sufficient-decrease and relative-error hypotheses of the Łojasiewicz convergence theorem for bounded analytic descent sequences are satisfied (Absil et al., 2005). The sequence therefore has a single limit $\boldsymbol{W}^{(k)} \longrightarrow \overline{\boldsymbol{W}},\ \nabla\Phi_s(\overline{\boldsymbol{W}}) = 0$. Write $\overline{\boldsymbol{X}} = \overline{\boldsymbol{W}}\overline{\boldsymbol{W}}^\top$. Lemma 29 shows that the limiting Gram matrix has the form $\overline{\boldsymbol{X}} = \sum_{\mu>0}\mu\boldsymbol{P}_\mu$, where each $\boldsymbol{P}_\mu$ is an orthogonal projector on the positive eigenspace E_μ of $\boldsymbol{M}$. Suppose that $\boldsymbol{P}_\mu$ is not the identity on some E_μ. Choose a unit vector $\boldsymbol{u} \in E_\mu \cap \ker(\boldsymbol{P}_\mu)$. Then $\overline{\boldsymbol{X}}\boldsymbol{u} = 0$, while the limiting multiplier $\boldsymbol{A}_\infty = \boldsymbol{I}_m + \eta_s(\boldsymbol{M} - \overline{\boldsymbol{X}})$ satisfies $\boldsymbol{A}_\infty\boldsymbol{u} = (1 + \eta_s\mu)\boldsymbol{u}$. Thus $\boldsymbol{A}_\infty$ has an eigenvalue strictly larger than one. This contradicts

Lemma 30, because every finite factor iterate is invertible and $\boldsymbol{W}^{(k)} \to \overline{\boldsymbol{W}}$. Hence every $\boldsymbol{P}_\mu$ is the identity on E_μ, and therefore $\overline{\boldsymbol{X}} = \boldsymbol{M}_+ = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$.

Define $\mathcal{T}$ as in Lemma 31. Since $\boldsymbol{X}^{(k)} \to \boldsymbol{X}_\infty$, the quotient iterates eventually enter the contraction neighborhood of that lemma. Fix any $q_X \in (q_{\mathrm{loc}}, 1)$. There is $k_0 < \infty$ such that $\|\boldsymbol{X}^{(k+1)} - \boldsymbol{X}_\infty\|_{\mathrm{F}} \le q_X \|\boldsymbol{X}^{(k)} - \boldsymbol{X}_\infty\|_{\mathrm{F}}$ for every $k \ge k_0$. Iteration gives a geometric bound for $k \ge k_0$. Enlarging the multiplicative constant to cover the finite prefix $0 \le k < k_0$ yields $\|\boldsymbol{X}^{(k)} - \boldsymbol{X}_\infty\|_{\mathrm{op}} \le C_X q_X^k$ for every $k \ge 0$.

Let $\boldsymbol{\Pi}_\perp = \boldsymbol{I}_m - \boldsymbol{\Pi}_{\mathrm{L}}$. Since $\boldsymbol{\Pi}_\perp \boldsymbol{X}_\infty \boldsymbol{\Pi}_\perp = \boldsymbol{0}$, the quotient bound gives $\|\boldsymbol{\Pi}_\perp \boldsymbol{W}^{(k)}\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{\Pi}_\perp \boldsymbol{X}^{(k)} \boldsymbol{\Pi}_\perp) \le (m-r) C_X q_X^k$. Hence $\|\boldsymbol{\Pi}_\perp \boldsymbol{W}^{(k)}\|_{\mathrm{F}} \le \sqrt{(m-r)C_X}\, q_W^k$, where $q_W = \sqrt{q_X}$.

Set $B_W = \sup_{k \ge 0} \|\boldsymbol{W}^{(k)}\|_{\mathrm{F}} < \infty$. Since $\boldsymbol{M} - \boldsymbol{X}_\infty = -\lambda \boldsymbol{\Pi}_\perp$, the factor increments satisfy $\boldsymbol{W}^{(k+1)} - \boldsymbol{W}^{(k)} = -\eta_s \lambda \boldsymbol{\Pi}_\perp \boldsymbol{W}^{(k)} - \eta_s (\boldsymbol{X}^{(k)} - \boldsymbol{X}_\infty) \boldsymbol{W}^{(k)}$. Because $q_X^k \le q_W^k$, it follows that $\|\boldsymbol{W}^{(k+1)} - \boldsymbol{W}^{(k)}\|_{\mathrm{F}} \le C_\Delta q_W^k$, where $C_\Delta = \eta_s \lambda \sqrt{(m-r)C_X} + \eta_s C_X B_W$. The increments are absolutely summable, and their tail gives $\|\boldsymbol{W}^{(k)} - \boldsymbol{W}_\infty^{\mathrm{d}}\|_{\mathrm{F}} \le \frac{C_\Delta}{1-q_W} q_W^k$. This proves the stated factor rate.

Taking the limit in $\boldsymbol{W}^{(k)} \boldsymbol{W}^{(k)\top} \longrightarrow \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$ and defining $\boldsymbol{J}_{\mathrm{d}}$ by equation 52 gives $\boldsymbol{J}_{\mathrm{d}} \boldsymbol{J}_{\mathrm{d}}^\top = \boldsymbol{I}_r$. The columns of $\boldsymbol{W}_\infty^{\mathrm{d}}$ lie in $\mathrm{col}(\boldsymbol{P})$, so $\boldsymbol{W}_\infty^{\mathrm{d}} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{J}_{\mathrm{d}} = \boldsymbol{Z}_\star \boldsymbol{T} \boldsymbol{J}_{\mathrm{d}}$. □

Proof of Lemma 27. Set $\boldsymbol{e}_k = \boldsymbol{z}_k - \boldsymbol{z}_\star$. Then $\boldsymbol{e}_{k+1} = (\boldsymbol{I}_d - \eta \boldsymbol{K}_k)\boldsymbol{e}_k + \boldsymbol{r}_k$, where $\boldsymbol{r}_k = \eta(\boldsymbol{b}_k - \boldsymbol{K}_k \boldsymbol{z}_\star)$. The assumptions imply $\|\boldsymbol{r}_k\|_2 \le c_r q^k$ for some $c_r < \infty$.

Because $0 \preceq \eta \boldsymbol{K}_k \preceq \boldsymbol{I}_d$, $\|\boldsymbol{I}_d - \eta \boldsymbol{K}_k\|_{\mathrm{op}} \le 1$. Thus $\|\boldsymbol{e}_{k+1}\|_2 \le \|\boldsymbol{e}_k\|_2 + c_r q^k$, and $\{\boldsymbol{e}_k\}_{k \ge 0}$ is bounded.

Let $\boldsymbol{\Pi}$ project onto $\mathrm{range}(\boldsymbol{K}_\infty)$ and set $\boldsymbol{\Pi}_0 = \boldsymbol{I}_d - \boldsymbol{\Pi}$. Write $\boldsymbol{e}_{\mathrm{s},k} = \boldsymbol{\Pi}\boldsymbol{e}_k$, $\boldsymbol{e}_{\mathrm{c},k} = \boldsymbol{\Pi}_0 \boldsymbol{e}_k$. On $\mathrm{range}(\boldsymbol{K}_\infty)$, define $\boldsymbol{A}_{\mathrm{s}} = \boldsymbol{\Pi}(\boldsymbol{I}_d - \eta \boldsymbol{K}_\infty)\boldsymbol{\Pi}$. Its operator norm is $q_+ < 1$ when the stable subspace is nontrivial. The stable component satisfies $\boldsymbol{e}_{\mathrm{s},k+1} = \boldsymbol{A}_{\mathrm{s}} \boldsymbol{e}_{\mathrm{s},k} + \boldsymbol{g}_{\mathrm{s},k}$, where $\boldsymbol{g}_{\mathrm{s},k} = -\eta \boldsymbol{\Pi}(\boldsymbol{K}_k - \boldsymbol{K}_\infty)\boldsymbol{e}_k + \boldsymbol{\Pi}\boldsymbol{r}_k$. Boundedness of $\boldsymbol{e}_k$ and the coefficient assumptions give $\|\boldsymbol{g}_{\mathrm{s},k}\|_2 \le c_g q^k$ for some $c_g < \infty$. Iteration yields $\boldsymbol{e}_{\mathrm{s},k} = \boldsymbol{A}_{\mathrm{s}}^k \boldsymbol{e}_{\mathrm{s},0} + \sum_{j=0}^{k-1} \boldsymbol{A}_{\mathrm{s}}^{k-1-j} \boldsymbol{g}_{\mathrm{s},j}$. For every q_z with $\max\{q, q_+\} < q_z < 1$, the resulting geometric convolution is bounded by $C_{\mathrm{s}} q_z^k$ for some $C_{\mathrm{s}} < \infty$. If $\boldsymbol{K}_\infty = \boldsymbol{0}$, the stable component is absent.

For the center component, $\boldsymbol{e}_{\mathrm{c},k+1} - \boldsymbol{e}_{\mathrm{c},k} = -\eta \boldsymbol{\Pi}_0 (\boldsymbol{K}_k - \boldsymbol{K}_\infty)\boldsymbol{e}_k + \boldsymbol{\Pi}_0 \boldsymbol{r}_k$. The right-hand side has norm at most $c_g q^k$ and is absolutely summable. Hence $\boldsymbol{e}_{\mathrm{c},k}$ converges to a limit $\boldsymbol{e}_{\mathrm{c},\infty}$, and $\|\boldsymbol{e}_{\mathrm{c},k} - \boldsymbol{e}_{\mathrm{c},\infty}\|_2 \le \frac{c_g}{1-q} q^k$. Setting $\boldsymbol{z}_\infty = \boldsymbol{z}_\star + \boldsymbol{e}_{\mathrm{c},\infty}$ gives $\boldsymbol{\Pi}\boldsymbol{z}_\infty = \boldsymbol{\Pi}\boldsymbol{z}_\star$. Combining the stable and center estimates proves the result. □

Proof of Theorem 28. Theorem 26 gives exponential convergence $\boldsymbol{X}^{(k)} \longrightarrow \boldsymbol{X}_\infty = \boldsymbol{P}\boldsymbol{\Lambda}^2\boldsymbol{P}^\top$.

Vectorizing the decoder update gives $\mathrm{vec}\left(\boldsymbol{\Gamma}^{(k+1)}\right) = \left[\boldsymbol{I}_{mp} - \eta_d \left(\boldsymbol{I}_p \otimes (\boldsymbol{X}^{(k)})^2\right)\right] \mathrm{vec}\left(\boldsymbol{\Gamma}^{(k)}\right) + \eta_d\, \mathrm{vec}\left(\boldsymbol{X}^{(k)} \mathsf{H}\right)$. The coefficient matrices are positive semidefinite and converge exponentially to $\boldsymbol{K}_{d,\infty} = \boldsymbol{I}_p \otimes \boldsymbol{X}_\infty^2$. The target $\boldsymbol{\Gamma}_\star$ satisfies $\boldsymbol{X}_\infty^2 \boldsymbol{\Gamma}_\star = \boldsymbol{X}_\infty \mathsf{H}$. Since $\sup_k \left\|\boldsymbol{I}_p \otimes (\boldsymbol{X}^{(k)})^2\right\|_{\mathrm{op}} \le B_X^2$, Lemma 27 applies and yields exponential convergence of $\boldsymbol{\Gamma}^{(k)}$. Its active projector acts by left multiplication with $\boldsymbol{\Pi}_{\mathrm{L}}$, so $\boldsymbol{\Pi}_{\mathrm{L}} \boldsymbol{\Gamma}_\infty^{\mathrm{d}} = \boldsymbol{\Gamma}_\star$.

To obtain the bounds B_Γ and B_R, set $\boldsymbol{E}_d^{(k)} = \boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}_\star$. Using $\mathsf{H} = \boldsymbol{X}_\infty \boldsymbol{\Gamma}_\star$ gives $\boldsymbol{E}_d^{(k+1)} = \left[\boldsymbol{I}_m - \eta_d (\boldsymbol{X}^{(k)})^2\right] \boldsymbol{E}_d^{(k)} + \eta_d \boldsymbol{X}^{(k)} \left(\boldsymbol{X}_\infty - \boldsymbol{X}^{(k)}\right) \boldsymbol{\Gamma}_\star$. The first factor is nonexpansive. Therefore $\|\boldsymbol{E}_d^{(k)}\|_{\mathrm{F}} \le \|\boldsymbol{E}_d^{(0)}\|_{\mathrm{F}} + \eta_d B_X C_X \|\boldsymbol{\Gamma}_\star\|_{\mathrm{F}} \sum_{\ell=0}^{k-1} q_X^\ell \le \|\boldsymbol{E}_d^{(0)}\|_{\mathrm{F}} + \frac{\eta_d B_X C_X \|\boldsymbol{\Gamma}_\star\|_{\mathrm{F}}}{1-q_X}$. This proves $\|\boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}} \le B_\Gamma$, $\|\boldsymbol{R}_{\mathrm{cas}}^{(k)}\|_{\mathrm{op}} \le B_R$. It also follows that $\boldsymbol{R}_{\mathrm{cas}}^{(k)} = \boldsymbol{X}^{(k)} \boldsymbol{\Gamma}^{(k)} \longrightarrow \boldsymbol{X}_\infty \boldsymbol{\Gamma}_\infty^{\mathrm{d}} = \mathsf{H}$ exponentially.

For each symbol x, vectorizing the transition update gives $\mathrm{vec}\left(\boldsymbol{\Omega}_x^{(k+1)}\right) = \left[\boldsymbol{I}_{m^2} - \eta_a \left(\boldsymbol{R}_{\mathrm{cas}}^{(k)} \boldsymbol{R}_{\mathrm{cas}}^{(k)\top} \otimes (\boldsymbol{X}^{(k)})^2\right)\right] \mathrm{vec}\left(\boldsymbol{\Omega}_x^{(k)}\right) + \eta_a\, \mathrm{vec}\left(\boldsymbol{X}^{(k)} \mathsf{H}_x \boldsymbol{R}_{\mathrm{cas}}^{(k)\top}\right)$. The coefficient

converges exponentially to $\boldsymbol{K}_{a,\infty} = \boldsymbol{G} \otimes \boldsymbol{X}_\infty^2$. The target $\boldsymbol{\Omega}_x^\star$ satisfies $\boldsymbol{X}_\infty^2 \boldsymbol{\Omega}_x^\star \boldsymbol{G} = \boldsymbol{X}_\infty \mathsf{H}_x \mathsf{H}^\top$. Moreover, $\sup_k \left\| \boldsymbol{R}_{\mathrm{cas}}^{(k)} \boldsymbol{R}_{\mathrm{cas}}^{(k)\top} \otimes (\boldsymbol{X}^{(k)})^2 \right\|_{\mathrm{op}} \le B_R^2 B_X^2$. Lemma 27 therefore yields exponential convergence of $\boldsymbol{\Omega}_x^{(k)}$. The range projector of $\boldsymbol{G} \otimes \boldsymbol{X}_\infty^2$ acts under vectorization as $\boldsymbol{\Omega}_x \longmapsto \boldsymbol{\Pi}_{\mathrm{L}} \boldsymbol{\Omega}_x \boldsymbol{\Pi}_{\mathrm{L}}$. Consequently, $\boldsymbol{\Pi}_{\mathrm{L}} \boldsymbol{\Omega}_{x,\infty}^{\mathrm{d}} \boldsymbol{\Pi}_{\mathrm{L}} = \boldsymbol{\Omega}_x^\star$. Multiplying this identity by $\boldsymbol{X}_\infty$ on the left and by H on the right gives $\boldsymbol{X}_\infty \boldsymbol{\Omega}_{x,\infty}^{\mathrm{d}} \mathsf{H} = \mathsf{H}_x$, which proves the realized transition-block limit.

Finally, $\boldsymbol{W}_\infty^{\mathrm{d}} = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{J}_{\mathrm{d}}$. Using the active auxiliary limits gives $\boldsymbol{W}_\infty^{\mathrm{d}\top} \boldsymbol{\Gamma}_\infty^{\mathrm{d}} = \boldsymbol{J}_{\mathrm{d}}^\top \boldsymbol{\Lambda} \boldsymbol{P}^\top \boldsymbol{\Gamma}_\star = \boldsymbol{J}_{\mathrm{d}}^\top \boldsymbol{T}^{-1} \boldsymbol{Q}^\top$, and $\boldsymbol{W}_\infty^{\mathrm{d}\top} \boldsymbol{\Omega}_{x,\infty}^{\mathrm{d}} \boldsymbol{W}_\infty^{\mathrm{d}} = \boldsymbol{J}_{\mathrm{d}}^\top \boldsymbol{\Lambda} \boldsymbol{P}^\top \boldsymbol{\Omega}_x^\star \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{J}_{\mathrm{d}} = \boldsymbol{J}_{\mathrm{d}}^\top \boldsymbol{T}^{-1} \boldsymbol{B}_x^\star \boldsymbol{T} \boldsymbol{J}_{\mathrm{d}}$. Inactive auxiliary components disappear because $\boldsymbol{W}_\infty^{\mathrm{d}\top} (\boldsymbol{I}_m - \boldsymbol{\Pi}_{\mathrm{L}}) = \boldsymbol{0},\ (\boldsymbol{I}_m - \boldsymbol{\Pi}_{\mathrm{L}}) \boldsymbol{W}_\infty^{\mathrm{d}} = \boldsymbol{0}$.

Compression with $\boldsymbol{J}_{\mathrm{d}}$ gives $\boldsymbol{\alpha}_{\mathrm{d}} = \boldsymbol{z}_\star(\epsilon)^\top \boldsymbol{T},\ \boldsymbol{M}_{x,\mathrm{d}} = \boldsymbol{T}^{-1} \boldsymbol{B}_x^\star \boldsymbol{T},\ \boldsymbol{c}_{\mathrm{d}} = \boldsymbol{T}^{-1} \boldsymbol{c}_\epsilon^\star$. These parameters are similar to the canonical realization in Theorem 3. They therefore realize every sequence score exactly and have state dimension r, which proves minimality. □

G Deterministic Block Perturbation and Fixed-Length Rollout Stability

This section studies static perturbations of the observed finite blocks. It controls the predictive subspaces, a minimum-norm least-squares realization, and every fixed-length rollout. It does not analyze the triangular cascade under noisy time-dependent coefficients.

G.1 Block perturbation and static least-squares realization

Suppose finite data produce estimates $\widehat{\mathsf{H}}, \widehat{\mathsf{H}}_x$ with $\|\widehat{\mathsf{H}} - \mathsf{H}\|_{\mathrm{op}} \le \eta, \|\widehat{\mathsf{H}}_x - \mathsf{H}_x\|_{\mathrm{op}} \le \eta_x$. Define $\eta_G = (2\|\mathsf{H}\|_{\mathrm{op}} + \eta)\eta$. Then $\|\widehat{\mathsf{H}}\widehat{\mathsf{H}}^\top - \boldsymbol{G}\|_{\mathrm{op}} \le \eta_G$ and the same bound holds for the right Gram matrix.

Recall that $\boldsymbol{\Pi}_{\mathrm{L}}$ and $\boldsymbol{\Pi}_{\mathrm{R}}$ in equation $\boldsymbol{\Pi}_{\mathrm{L}} = \boldsymbol{P}\boldsymbol{P}^\top, \boldsymbol{\Pi}_{\mathrm{R}} = \boldsymbol{Q}\boldsymbol{Q}^\top$ are the true predictive left and right projectors. Let $\widehat{\Pi}_{\mathrm{L}}$ and $\widehat{\Pi}_{\mathrm{R}}$ be the leading rank-r singular projectors of $\widehat{\mathsf{H}}$.

Theorem 32 (Stable state dimension and shifted-block prediction)**.** *Assume $\eta_G \le \frac{\sigma_r^2}{4}$ and choose $\eta_G < \lambda < \sigma_r^2 - \eta_G$. For any initialization $\widehat{\boldsymbol{W}}(0) \in \mathrm{GL}(m)$, the state flow with target $\widehat{\boldsymbol{G}} = \widehat{\mathsf{H}}\widehat{\mathsf{H}}^\top$ converges to a rank-r state matrix whose column projector is $\widehat{\Pi}_{\mathrm{L}}$. Moreover, $\|\widehat{\Pi}_{\mathrm{L}} - \boldsymbol{\Pi}_{\mathrm{L}}\|_{\mathrm{op}} \le \frac{2\eta_G}{\sigma_r^2},\ \|\widehat{\Pi}_{\mathrm{R}} - \boldsymbol{\Pi}_{\mathrm{R}}\|_{\mathrm{op}} \le \frac{2\eta_G}{\sigma_r^2}$. Let $\widehat{\boldsymbol{W}}_\infty$ be the resulting state matrix. Let $\widehat{\boldsymbol{C}}_\infty$ be the minimum-Frobenius-norm solution of $\frac{1}{2}\left\|\widehat{\boldsymbol{W}}_\infty \boldsymbol{C} - \widehat{\mathsf{H}}\right\|_{\mathrm{F}}^2$, and, for every $x \in \mathcal{A}$, let $\widehat{\boldsymbol{A}}_{x,\infty}$ be the minimum-Frobenius-norm solution of $\frac{1}{2}\left\|\widehat{\boldsymbol{W}}_\infty \boldsymbol{A}_x \widehat{\boldsymbol{C}}_\infty - \widehat{\mathsf{H}}_x\right\|_{\mathrm{F}}^2$. These static least-squares solutions satisfy $\widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{C}}_\infty = \widehat{\Pi}_{\mathrm{L}} \widehat{\mathsf{H}} \widehat{\Pi}_{\mathrm{R}}$, $\widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{A}}_{x,\infty} \widehat{\boldsymbol{C}}_\infty = \widehat{\Pi}_{\mathrm{L}} \widehat{\mathsf{H}}_x \widehat{\Pi}_{\mathrm{R}}$. Consequently, $\left\|\widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{C}}_\infty - \mathsf{H}\right\|_{\mathrm{op}} \le \eta + \frac{2\eta_G}{\sigma_r^2}\|\mathsf{H}\|_{\mathrm{op}}$, $\left\|\widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{A}}_{x,\infty} \widehat{\boldsymbol{C}}_\infty - \mathsf{H}_x\right\|_{\mathrm{op}} \le \eta_x + \frac{4\eta_G}{\sigma_r^2}\|\mathsf{H}_x\|_{\mathrm{op}}$.*

G.2 Gauge-invariant fixed-length rollout stability

The preceding theorem controls the learned base block and every one-symbol shifted block. We now show that these controls extend to products of the learned transitions. The argument uses observable operators, so no alignment between exact and perturbed hidden coordinates is required. Define $\widehat{\mathsf{H}}_r = \widehat{\Pi}_{\mathrm{L}} \widehat{\mathsf{H}} \widehat{\Pi}_{\mathrm{R}}, \widetilde{\mathsf{H}}_x = \widehat{\Pi}_{\mathrm{L}} \widehat{\mathsf{H}}_x \widehat{\Pi}_{\mathrm{R}}$. The exact and perturbed observable operators are $\boldsymbol{O}_x = \mathsf{H}^\dagger \mathsf{H}_x, \widehat{\boldsymbol{O}}_x = \widehat{\mathsf{H}}_r^\dagger \widetilde{\mathsf{H}}_x$.

Theorem 33 (Gauge-invariant transition and fixed-length rollout consistency)**.** *Assume the conditions of Theorem 32. Define $d_H = \eta + \frac{2\eta_G}{\sigma_r^2}\|\mathsf{H}\|_{\mathrm{op}}$ and, for every $x \in \mathcal{A}$, $d_x = \eta_x + \frac{4\eta_G}{\sigma_r^2}\|\mathsf{H}_x\|_{\mathrm{op}}$. The condition $\eta_G \le \sigma_r^2/4$ implies $\eta \le \sigma_r/8$. Set $\chi_\dagger = d_H\left(\frac{1}{\sigma_r^2} + \frac{1}{(\sigma_r - \eta)^2} + \frac{1}{\sigma_r(\sigma_r - \eta)}\right)$ and $\epsilon_{\mathrm{obs}} = \max_{x \in \mathcal{A}}\left[\frac{d_x}{\sigma_r - \eta} + \chi_\dagger \|\mathsf{H}_x\|_{\mathrm{op}}\right]$. Then $\max_{x \in \mathcal{A}} \|\widehat{\boldsymbol{O}}_x - \boldsymbol{O}_x\|_{\mathrm{op}} \le \epsilon_{\mathrm{obs}}$.*

Let $\widehat{\boldsymbol{W}}_\infty$, $\widehat{\boldsymbol{C}}_\infty$, and $\widehat{\boldsymbol{A}}_{x,\infty}$ be the empirical state limit and static least-squares solutions from Theorem 32. For $w = x_1 \cdots x_L$, define their recurrent score by $\widehat{f}_{\mathrm{LS}}(w) = \boldsymbol{e}_{i_\epsilon}^\top \widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{A}}_{x_1,\infty} \cdots \widehat{\boldsymbol{A}}_{x_L,\infty} \widehat{\boldsymbol{C}}_\infty \boldsymbol{e}_{j_\epsilon}$. Then $\widehat{f}_{\mathrm{LS}}(w) = \boldsymbol{e}_{i_\epsilon}^\top \widehat{\mathsf{H}}_r \widehat{\boldsymbol{O}}_{x_1} \cdots \widehat{\boldsymbol{O}}_{x_L} \boldsymbol{e}_{j_\epsilon}$. Define $R_{\mathrm{obs}} = \max_{x\in\mathcal{A}} \max\left\{\|\boldsymbol{O}_x\|_{\mathrm{op}}, \|\widehat{\boldsymbol{O}}_x\|_{\mathrm{op}}\right\}$. It satisfies

$$R_{\mathrm{obs}} \le \max_{x\in\mathcal{A}} \max\left\{\frac{\|\mathsf{H}_x\|_{\mathrm{op}}}{\sigma_r}, \frac{\|\mathsf{H}_x\|_{\mathrm{op}} + d_x}{\sigma_r - \eta}\right\}. \tag{54}$$

For the empty word, $|\widehat{f}_{\mathrm{LS}}(\epsilon) - f(\epsilon)| \le d_H$. For every $L \ge 1$, $\max_{w\in\mathcal{A}^L} |\widehat{f}_{\mathrm{LS}}(w) - f(w)| \le d_H R_{\mathrm{obs}}^L + \|\mathsf{H}\|_{\mathrm{op}} L R_{\mathrm{obs}}^{L-1} \epsilon_{\mathrm{obs}}$. Consequently, if $\eta \to 0$ and $\max_x \eta_x \to 0$, then the perturbed recurrent score converges uniformly to f on words of every fixed length.

The observable operators act on future-score rows, not on a chosen hidden coordinate system. The theorem therefore gives recurrent prediction consistency without claiming entrywise convergence of hidden transition matrices in an arbitrary gauge. For a length bound L that may increase as the perturbations decrease, the displayed bound gives consistency whenever $d_H R_{\mathrm{obs}}^L \to 0$ and $L R_{\mathrm{obs}}^{L-1} \epsilon_{\mathrm{obs}} \to 0$.

The block-error bounds are assumed rather than derived from a sampling model.

G.3 Proofs

Lemma 34 (Shifted blocks share the predictive singular spaces)**.** *Under Assumption 2, $\boldsymbol{\Pi}_{\mathrm{L}} \mathsf{H}_x \boldsymbol{\Pi}_{\mathrm{R}} = \mathsf{H}_x$ for every $x \in \mathcal{A}$.*

Proof. Equation $\mathsf{H}_x = \boldsymbol{Z}_\star \boldsymbol{B}_x^\star \boldsymbol{Q}^\top$ gives $\mathsf{H}_x = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{B}_x^\star \boldsymbol{Q}^\top$, so left and right multiplication by $\boldsymbol{P}\boldsymbol{P}^\top$ and $\boldsymbol{Q}\boldsymbol{Q}^\top$ leaves the block unchanged. □

Proof of Theorem 32. First, $\widehat{\mathsf{H}}\widehat{\mathsf{H}}^\top - \mathsf{H}\mathsf{H}^\top = (\widehat{\mathsf{H}} - \mathsf{H})\mathsf{H}^\top + \mathsf{H}(\widehat{\mathsf{H}} - \mathsf{H})^\top + (\widehat{\mathsf{H}} - \mathsf{H})(\widehat{\mathsf{H}} - \mathsf{H})^\top$. Taking operator norms proves $\|\widehat{\mathsf{H}}\widehat{\mathsf{H}}^\top - \boldsymbol{G}\|_{\mathrm{op}} \le \eta_G$. The right Gram bound is identical.

Weyl's eigenvalue inequality shows that the leading r eigenvalues of $\widehat{\boldsymbol{G}}$ are at least $\sigma_r^2 - \eta_G > \lambda$, while every remaining eigenvalue is at most $\eta_G < \lambda$. Proposition 17 gives convergence of the empirical state Gram matrix to the positive part of $\widehat{\boldsymbol{G}} - \lambda \boldsymbol{I}_m$. The analytic-gradient convergence argument from Theorem 5 then gives convergence of the factor trajectory to a rank-r limit with column projector $\widehat{\Pi}_{\mathrm{L}}$.

For the left projector, suppose first that $r < m$. The true left Gram matrix $\mathsf{H}\mathsf{H}^\top$ has eigenvalues $\sigma_1^2, \ldots, \sigma_r^2$ followed by zeros, so its leading r-dimensional eigenspace is separated from its orthogonal complement by the gap σ_r^2. The Davis–Kahan theorem (Stewart & Sun, 1993) and $\eta_G \le \frac{\sigma_r^2}{4}$ therefore give $\|\widehat{\Pi}_{\mathrm{L}} - \boldsymbol{\Pi}_{\mathrm{L}}\|_{\mathrm{op}} \le \frac{2\eta_G}{\sigma_r^2}$. If $r = m$, both projectors equal $\boldsymbol{I}_m$, so the bound is immediate. The right-projector bound follows identically from the eigengap of $\mathsf{H}^\top\mathsf{H}$ when $r < p$, while the case $r = p$ is trivial.

Let the leading rank-r singular triplet of $\widehat{\mathsf{H}}$ be $\widehat{\mathsf{H}}_r = \widehat{\boldsymbol{P}}_r \widehat{\boldsymbol{\Sigma}}_r \widehat{\boldsymbol{Q}}_r^\top$, where $\widehat{\boldsymbol{\Sigma}}_r = \mathrm{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_r)$. Define $\widehat{\boldsymbol{\Lambda}}_r = \mathrm{diag}\left(\sqrt{\widehat{\sigma}_1^2 - \lambda}, \ldots, \sqrt{\widehat{\sigma}_r^2 - \lambda}\right)$. The empirical state-flow limit satisfies $\widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{W}}_\infty^\top = \widehat{\boldsymbol{P}}_r \widehat{\boldsymbol{\Lambda}}_r^2 \widehat{\boldsymbol{P}}_r^\top$. Set $\widehat{\boldsymbol{J}}_r = \widehat{\boldsymbol{\Lambda}}_r^{-1} \widehat{\boldsymbol{P}}_r^\top \widehat{\boldsymbol{W}}_\infty$. Then $\widehat{\boldsymbol{J}}_r \widehat{\boldsymbol{J}}_r^\top = \boldsymbol{I}_r$ and $\widehat{\boldsymbol{W}}_\infty = \widehat{\boldsymbol{P}}_r \widehat{\boldsymbol{\Lambda}}_r \widehat{\boldsymbol{J}}_r$. Its Moore–Penrose inverse is $\widehat{\boldsymbol{W}}_\infty^\dagger = \widehat{\boldsymbol{J}}_r^\top \widehat{\boldsymbol{\Lambda}}_r^{-1} \widehat{\boldsymbol{P}}_r^\top$. The minimum-Frobenius-norm decoder is $\widehat{\boldsymbol{C}}_\infty = \widehat{\boldsymbol{W}}_\infty^\dagger \widehat{\mathsf{H}} = \widehat{\boldsymbol{J}}_r^\top \widehat{\boldsymbol{\Lambda}}_r^{-1} \widehat{\boldsymbol{\Sigma}}_r \widehat{\boldsymbol{Q}}_r^\top$. Consequently, $\widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{C}}_\infty = \widehat{\boldsymbol{P}}_r \widehat{\boldsymbol{\Sigma}}_r \widehat{\boldsymbol{Q}}_r^\top = \widehat{\Pi}_{\mathrm{L}} \widehat{\mathsf{H}} \widehat{\Pi}_{\mathrm{R}}$.

Consider the linear realization map $\mathcal{T}(\boldsymbol{A}) = \widehat{\boldsymbol{W}}_\infty \boldsymbol{A} \widehat{\boldsymbol{C}}_\infty$. For every $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, $\mathcal{T}(\boldsymbol{A}) = \widehat{\boldsymbol{P}}_r \left[\widehat{\boldsymbol{\Lambda}}_r \widehat{\boldsymbol{J}}_r \boldsymbol{A} \widehat{\boldsymbol{J}}_r^\top \widehat{\boldsymbol{\Lambda}}_r^{-1} \widehat{\boldsymbol{\Sigma}}_r\right] \widehat{\boldsymbol{Q}}_r^\top$. Hence $\mathrm{Range}(\mathcal{T}) \subseteq \mathcal{M}_r$, where $\mathcal{M}_r = \left\{\widehat{\boldsymbol{P}}_r \boldsymbol{B} \widehat{\boldsymbol{Q}}_r^\top : \boldsymbol{B} \in \mathbb{R}^{r\times r}\right\}$. Conversely, for any $\boldsymbol{B} \in \mathbb{R}^{r\times r}$, define $\boldsymbol{A}_{\boldsymbol{B}} = \widehat{\boldsymbol{J}}_r^\top \widehat{\boldsymbol{\Lambda}}_r^{-1} \boldsymbol{B} \widehat{\boldsymbol{\Sigma}}_r^{-1} \widehat{\boldsymbol{\Lambda}}_r \widehat{\boldsymbol{J}}_r$. Then $\mathcal{T}(\boldsymbol{A}_{\boldsymbol{B}}) = \widehat{\boldsymbol{P}}_r \boldsymbol{B} \widehat{\boldsymbol{Q}}_r^\top$. Therefore $\mathrm{Range}(\mathcal{T}) = \mathcal{M}_r$. The Frobenius-orthogonal projection of an arbitrary matrix $\boldsymbol{X}$ onto $\mathcal{M}_r$ is $\widehat{\Pi}_{\mathrm{L}} \boldsymbol{X} \widehat{\Pi}_{\mathrm{R}}$. The transition least-squares regression therefore realizes $\widehat{\boldsymbol{W}}_\infty \widehat{\boldsymbol{A}}_{x,\infty} \widehat{\boldsymbol{C}}_\infty = \widehat{\Pi}_{\mathrm{L}} \widehat{\mathsf{H}}_x \widehat{\Pi}_{\mathrm{R}}$.

The parameter is chosen by the minimum-Frobenius-norm rule, while the realized projected block is uniquely determined.

Using $\mathbf{\Pi}_{\mathrm{L}}\mathsf{H} = \mathsf{H}$, $\left\|\widehat{\Pi}_{\mathrm{L}}\widehat{\mathsf{H}} - \mathsf{H}\right\|_{\mathrm{op}} \le \|\widehat{\mathsf{H}} - \mathsf{H}\|_{\mathrm{op}} + \|\widehat{\Pi}_{\mathrm{L}} - \mathbf{\Pi}_{\mathrm{L}}\|_{\mathrm{op}}\|\mathsf{H}\|_{\mathrm{op}}$, which gives $\left\|\widehat{\boldsymbol{W}}_{\infty}\widehat{\boldsymbol{C}}_{\infty} - \mathsf{H}\right\|_{\mathrm{op}} \le \eta + \frac{2\eta_G}{\sigma_r^2}\|\mathsf{H}\|_{\mathrm{op}}$.

Finally, Lemma 34 gives $\left\|\widehat{\Pi}_{\mathrm{L}}\widehat{\mathsf{H}}_x\widehat{\Pi}_{\mathrm{R}} - \mathsf{H}_x\right\|_{\mathrm{op}} \le \eta_x + \|\widehat{\Pi}_{\mathrm{L}} - \mathbf{\Pi}_{\mathrm{L}}\|_{\mathrm{op}}\|\mathsf{H}_x\|_{\mathrm{op}} + \|\widehat{\Pi}_{\mathrm{R}} - \mathbf{\Pi}_{\mathrm{R}}\|_{\mathrm{op}}\|\mathsf{H}_x\|_{\mathrm{op}}$. Substituting $\|\widehat{\Pi}_{\mathrm{L}} - \mathbf{\Pi}_{\mathrm{L}}\|_{\mathrm{op}} \le \frac{2\eta_G}{\sigma_r^2}$, $\|\widehat{\Pi}_{\mathrm{R}} - \mathbf{\Pi}_{\mathrm{R}}\|_{\mathrm{op}} \le \frac{2\eta_G}{\sigma_r^2}$ proves $\left\|\widehat{\boldsymbol{W}}_{\infty}\widehat{\boldsymbol{A}}_{x,\infty}\widehat{\boldsymbol{C}}_{\infty} - \mathsf{H}_x\right\|_{\mathrm{op}} \le \eta_x + \frac{4\eta_G}{\sigma_r^2}\|\mathsf{H}_x\|_{\mathrm{op}}$. □

Lemma 35 (A same-rank pseudoinverse perturbation bound)**.** *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same rank. Then* $\|\boldsymbol{B}^{\dagger} - \boldsymbol{A}^{\dagger}\|_{\mathrm{op}} \le \|\boldsymbol{B} - \boldsymbol{A}\|_{\mathrm{op}}\Big(\|\boldsymbol{B}^{\dagger}\|_{\mathrm{op}}\|\boldsymbol{A}^{\dagger}\|_{\mathrm{op}} + \|\boldsymbol{B}^{\dagger}\|_{\mathrm{op}}^2 + \|\boldsymbol{A}^{\dagger}\|_{\mathrm{op}}^2\Big)$.

Proof. The Moore–Penrose inverses satisfy the identity $\boldsymbol{B}^{\dagger} - \boldsymbol{A}^{\dagger} = -\boldsymbol{B}^{\dagger}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{\dagger} + \boldsymbol{B}^{\dagger}\boldsymbol{B}^{\dagger\top}(\boldsymbol{B}^{\top} - \boldsymbol{A}^{\top})(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{\dagger}) + (\boldsymbol{I} - \boldsymbol{B}^{\dagger}\boldsymbol{B})(\boldsymbol{B}^{\top} - \boldsymbol{A}^{\top})\boldsymbol{A}^{\dagger\top}\boldsymbol{A}^{\dagger}$. The two complementary factors are orthogonal projectors and therefore have operator norm at most one. Taking operator norms term by term proves equation $\|\boldsymbol{B}^{\dagger} - \boldsymbol{A}^{\dagger}\|_{\mathrm{op}} \le \|\boldsymbol{B} - \boldsymbol{A}\|_{\mathrm{op}}\Big(\|\boldsymbol{B}^{\dagger}\|_{\mathrm{op}}\|\boldsymbol{A}^{\dagger}\|_{\mathrm{op}} + \|\boldsymbol{B}^{\dagger}\|_{\mathrm{op}}^2 + \|\boldsymbol{A}^{\dagger}\|_{\mathrm{op}}^2\Big)$. □

Proof of Theorem 33. Let the leading rank-r singular value decomposition of $\widehat{\mathsf{H}}$ be $\widehat{\mathsf{H}}_r = \widehat{\boldsymbol{P}}_r\widehat{\mathbf{\Sigma}}_r\widehat{\boldsymbol{Q}}_r^{\top}$. Theorem 32 gives $\|\widehat{\mathsf{H}}_r - \mathsf{H}\|_{\mathrm{op}} \le d_H$, $\|\widetilde{\mathsf{H}}_x - \mathsf{H}_x\|_{\mathrm{op}} \le d_x$. Weyl's inequality gives $\sigma_r(\widehat{\mathsf{H}}_r) = \widehat{\sigma}_r \ge \sigma_r - \eta$. Moreover, $\eta_G = (2\|\mathsf{H}\|_{\mathrm{op}} + \eta)\eta \ge 2\sigma_r\eta$. Thus $\eta_G \le \sigma_r^2/4$ implies $\eta \le \sigma_r/8$.

Apply Lemma 35 with $\boldsymbol{A} = \mathsf{H}$ and $\boldsymbol{B} = \widehat{\mathsf{H}}_r$. Equations $\|\widehat{\mathsf{H}}_r - \mathsf{H}\|_{\mathrm{op}} \le d_H$, $\|\widetilde{\mathsf{H}}_x - \mathsf{H}_x\|_{\mathrm{op}} \le d_x$ and $\sigma_r(\widehat{\mathsf{H}}_r) = \widehat{\sigma}_r \ge \sigma_r - \eta$ give $\|\widehat{\mathsf{H}}_r^{\dagger} - \mathsf{H}^{\dagger}\|_{\mathrm{op}} \le \chi_{\dagger}$. For every symbol x, $\widehat{\boldsymbol{O}}_x - \boldsymbol{O}_x = \widehat{\mathsf{H}}_r^{\dagger}(\widetilde{\mathsf{H}}_x - \mathsf{H}_x) + (\widehat{\mathsf{H}}_r^{\dagger} - \mathsf{H}^{\dagger})\mathsf{H}_x$. Since $\|\widehat{\mathsf{H}}_r^{\dagger}\|_{\mathrm{op}} \le (\sigma_r - \eta)^{-1}$, equations $\|\widehat{\mathsf{H}}_r - \mathsf{H}\|_{\mathrm{op}} \le d_H$, $\|\widetilde{\mathsf{H}}_x - \mathsf{H}_x\|_{\mathrm{op}} \le d_x$ and $\|\widehat{\mathsf{H}}_r^{\dagger} - \mathsf{H}^{\dagger}\|_{\mathrm{op}} \le \chi_{\dagger}$ prove equation $\max_{x\in\mathcal{A}}\|\widehat{\boldsymbol{O}}_x - \boldsymbol{O}_x\|_{\mathrm{op}} \le \epsilon_{\mathrm{obs}}$.

We next identify the empirical least-squares score with the observable-operator score. Define $\widehat{\boldsymbol{B}}_x = \widehat{\mathbf{\Sigma}}_r^{-1}\widehat{\boldsymbol{P}}_r^{\top}\widehat{\mathsf{H}}_x\widehat{\boldsymbol{Q}}_r$. Let $\widehat{\mathbf{\Lambda}}_r$, $\widehat{\boldsymbol{J}}_r$, and $\widehat{\boldsymbol{T}}_r = \widehat{\mathbf{\Sigma}}_r^{-1}\widehat{\mathbf{\Lambda}}_r$ be the empirical analogues of $\mathbf{\Lambda}$, $\boldsymbol{J}$, and $\boldsymbol{T}$. The decoder formula follows from $\widehat{\boldsymbol{C}}_{\infty} = \widehat{\boldsymbol{W}}_{\infty}^{\dagger}\widehat{\mathsf{H}}$. The transition least-squares problem determines the active block uniquely, and the minimum-Frobenius-norm solution vanishes on its orthogonal complement. Therefore $\widehat{\boldsymbol{C}}_{\infty} = \widehat{\boldsymbol{J}}_r^{\top}\widehat{\boldsymbol{T}}_r^{-1}\widehat{\boldsymbol{Q}}_r^{\top}$ and $\widehat{\boldsymbol{A}}_{x,\infty} = \widehat{\boldsymbol{J}}_r^{\top}\widehat{\boldsymbol{T}}_r^{-1}\widehat{\boldsymbol{B}}_x\widehat{\boldsymbol{T}}_r\widehat{\boldsymbol{J}}_r$. All similarity and isometric factors cancel in a recurrent product. Hence $\widehat{f}_{\mathrm{LS}}(w) = \boldsymbol{e}_{i_\epsilon}^{\top}\widehat{\boldsymbol{P}}_r\widehat{\mathbf{\Sigma}}_r\widehat{\boldsymbol{B}}_{x_1}\cdots\widehat{\boldsymbol{B}}_{x_L}\widehat{\boldsymbol{Q}}_r^{\top}\boldsymbol{e}_{j_\epsilon}$. On the other hand, $\widehat{\boldsymbol{O}}_x = \widehat{\boldsymbol{Q}}_r\widehat{\boldsymbol{B}}_x\widehat{\boldsymbol{Q}}_r^{\top}$. Substituting equation $\widehat{\boldsymbol{O}}_x = \widehat{\boldsymbol{Q}}_r\widehat{\boldsymbol{B}}_x\widehat{\boldsymbol{Q}}_r^{\top}$ into the right-hand side of equation $\widehat{f}_{\mathrm{LS}}(w) = \boldsymbol{e}_{i_\epsilon}^{\top}\widehat{\mathsf{H}}_r\widehat{\boldsymbol{O}}_{x_1}\cdots\widehat{\boldsymbol{O}}_{x_L}\boldsymbol{e}_{j_\epsilon}$ proves that identity.

The norm bound in equation $R_{\mathrm{obs}} \le \max_{x\in\mathcal{A}}\max\left\{\frac{\|\mathsf{H}_x\|_{\mathrm{op}}}{\sigma_r}, \frac{\|\mathsf{H}_x\|_{\mathrm{op}} + d_x}{\sigma_r - \eta}\right\}$ follows from $\|\boldsymbol{O}_x\|_{\mathrm{op}} \le \frac{\|\mathsf{H}_x\|_{\mathrm{op}}}{\sigma_r}$ and $\|\widehat{\boldsymbol{O}}_x\|_{\mathrm{op}} \le \frac{\|\widetilde{\mathsf{H}}_x\|_{\mathrm{op}}}{\sigma_r - \eta} \le \frac{\|\mathsf{H}_x\|_{\mathrm{op}} + d_x}{\sigma_r - \eta}$.

For the exact system, $\boldsymbol{O}_x = \boldsymbol{Q}\boldsymbol{B}_x^{\star}\boldsymbol{Q}^{\top}$, so Theorem 3 gives $f(w) = \boldsymbol{e}_{i_\epsilon}^{\top}\mathsf{H}\boldsymbol{O}_{x_1}\cdots\boldsymbol{O}_{x_L}\boldsymbol{e}_{j_\epsilon}$. The empty-word bound follows directly from equation $\|\widehat{\mathsf{H}}_r - \mathsf{H}\|_{\mathrm{op}} \le d_H$, $\|\widetilde{\mathsf{H}}_x - \mathsf{H}_x\|_{\mathrm{op}} \le d_x$. For $L \ge 1$, add and subtract $\boldsymbol{e}_{i_\epsilon}^{\top}\mathsf{H}\widehat{\boldsymbol{O}}_{x_1}\cdots\widehat{\boldsymbol{O}}_{x_L}\boldsymbol{e}_{j_\epsilon}$. The initial-row error is at most $d_H R_{\mathrm{obs}}^L$. A telescoping expansion of the two operator products gives $\left\|\widehat{\boldsymbol{O}}_{x_1}\cdots\widehat{\boldsymbol{O}}_{x_L} - \boldsymbol{O}_{x_1}\cdots\boldsymbol{O}_{x_L}\right\|_{\mathrm{op}} \le LR_{\mathrm{obs}}^{L-1}\epsilon_{\mathrm{obs}}$. Since $\|\boldsymbol{e}_{i_\epsilon}^{\top}\mathsf{H}\|_2 \le \|\mathsf{H}\|_{\mathrm{op}}$, combining the two terms proves equation $\max_{w\in\mathcal{A}^L}|\widehat{f}_{\mathrm{LS}}(w) - f(w)| \le d_H R_{\mathrm{obs}}^L + \|\mathsf{H}\|_{\mathrm{op}}LR_{\mathrm{obs}}^{L-1}\epsilon_{\mathrm{obs}}$. The final consistency statement follows because d_H, every d_x, $\chi_{\dagger}$, and ϵ_{obs} converge to zero, while equation 54 keeps R_{obs} bounded for sufficiently small perturbations. □

Table 4: Experiment 5: successful runs out of 90 per initialization profile and feedback strength. Feedback settings within each row use identical initial parameters.

Initialization	$\varepsilon = 0$	$\varepsilon = 0.25$	$\varepsilon = 0.5$	$\varepsilon = 1$
Ordinary reference	90/90	90/90	90/90	40/90
Ill-conditioned state	90/90	85/90	85/90	10/90
Large auxiliary	90/90	0/90	0/90	0/90
Constructed, $\varepsilon_{\mathrm{ref}} = 0.25$	90/90	0/90	0/90	0/90
Constructed, $\varepsilon_{\mathrm{ref}} = 0.5$	90/90	40/90	0/90	0/90
Constructed, $\varepsilon_{\mathrm{ref}} = 1$	90/90	39/90	18/90	0/90

H Protocols for Experiments 5 and 6

H.1 Common protocol and numerical criteria

Both experiments used the FP64 implementation and target-generation procedure in Section 7, with three targets per rank $r \in \{3, 5, 7\}$ and 15×15 Hankel blocks. Each initialization profile used ten seeds per target (1000–1009). We set $\lambda = 0.25\sigma_r^2$ and evaluated recurrent predictions at lengths $\{2, 4, 8, 12, 16, 24\}$. All words were enumerated when their number was at most 512; otherwise, 512 words were sampled uniformly with replacement. Reported maxima are over these evaluated words.

Success required relative state, decoder, and transition residuals at most $(10^{-8}, 10^{-7}, 10^{-7})$, respectively, a normalized vector-field norm at most 10^{-7}, and a normalized recurrent error at most 10^{-5}. The vector-field diagnostic was $\|F_\varepsilon(\boldsymbol{z})\|_{\mathcal{Z}} / \max\{1, \|\boldsymbol{z}\|_{\mathcal{Z}}\}$, evaluated with unit decoder and transition rate multipliers and with $\varepsilon = 0$ in Experiment 6. At each length, recurrent error was normalized by the largest absolute target score among the evaluated words.

H.2 Experiment 5: triangular training and fixed positive feedback

We compared $\varepsilon \in \{0, 0.25, 0.5, 1\}$ using identical initial state, decoder, and transition parameters within each target, initialization profile, and seed. The ordinary reference, ill-conditioned-state, and large-auxiliary profiles used random state singular vectors and $\|\boldsymbol{W}_0\|_{\mathrm{op}} = 0.7$. Their pairs of state condition number and Gaussian auxiliary entry standard deviation were $(3, 0.05)$, $(10^4, 0.5)$, and $(10, 2)$, respectively.

Three constructed profiles used $\varepsilon_{\mathrm{ref}} \in \{0.25, 0.5, 1\}$, $\|\boldsymbol{W}_0\|_{\mathrm{F}} = 0.02$, state condition number 5, and $\boldsymbol{\Gamma}_0 = -1.25\mathsf{H}/(2\varepsilon_{\mathrm{ref}}) + \boldsymbol{N}$, where Gaussian noise was rescaled to $\|\boldsymbol{N}\|_{\mathrm{op}} = \lambda/(16\varepsilon_{\mathrm{ref}}\sigma_1)$. Transition auxiliary entries had standard deviation 0.05. These profiles follow the nonoptimal-equilibrium construction in Theorem 12.

We integrated the resulting 2160 trajectories with $\eta_d = \eta_a = 1$ using an adaptive Dormand–Prince 5(4) method, with relative and absolute tolerances $(10^{-9}, 10^{-11})$, a maximum time of 250, and at most 150,000 attempted steps per trajectory. Table 4 reports success counts across the nine targets and ten initialization seeds.

All 540 triangular runs met the success criterion, with maximum absolute recurrent error 6.87×10^{-8}. For each constructed profile, all 90 runs at every tested $\varepsilon \geq \varepsilon_{\mathrm{ref}}$ met the collapse criteria, giving 540 cases. Collapse identification required small normalized state and vector-field norms, positive definiteness of $\boldsymbol{B}_\varepsilon(\boldsymbol{\Gamma})$, and a nonzero objective gap. All 540 cases were reproduced at relative and absolute tolerances $(10^{-11}, 10^{-13})$ using the same initializations. No collapse was verified in the ordinary profiles. Failure to meet the success criterion within the numerical budget was not, by itself, classified as collapse.

H.3 Experiment 6: fixed-step triangular training

We used the three ordinary initialization profiles from Experiment 5 and the discrete triangular updates in Appendix F. Two step-selection factors, $f \in \{1, 0.5\}$, were compared from identical initial parameters, giving 540 runs. State and decoder steps used state-sublevel bounds with safety

Table 5: Experiment 6: runs meeting the numerical success criterion within $300{,}000$ iterations. The two step settings use identical initial parameters.

Initialization	$f = 1$	$f = 0.5$
Ordinary reference	$90/90$	$90/90$
Ill-conditioned state	$90/90$	$89/90$
Large auxiliary	$69/90$	$52/90$
Total	$249/270$	$231/270$

factor $0.8f$. The transition step was selected from finite-prefix curvature bounds and an analytic tail estimate using the same safety factor.

After step selection, training was restarted from the prescribed initial parameters. State, decoder, and transition parameters were then updated synchronously from iteration zero, with every update evaluated at the preceding iterate. Step sizes remained fixed throughout training, and each run had a budget of $300{,}000$ iterations.

Overall, $480/540$ runs met the criterion (Table 5). The remaining 60 exhausted the iteration budget, comprising 59 large-auxiliary runs and one ill-conditioned-state run. No numerical failure or monitored descent or curvature violation occurred. Among successful runs, the maximum absolute recurrent error was 8.58×10^{-7}. These finite-budget outcomes do not establish failure of asymptotic convergence. Neither Experiment 5 nor Experiment 6 evaluates the switching certificate.