

Markov Categorical Deep Learning with Coalgebraic Final Semantics for Transformer, State Space Models and Linear Attention

Yoshihiro Maruyama

Department of Mathematical and Information Sciences
Kyoto University

Ryo Nasu

Graduate School of Informatics
Nagoya University

Abstract

We develop a unified categorical theory of sequence architectures under the formalism of Markov categories combined with coalgebraic final semantics, which organizes Transformer, linear attention, and Mamba/SSM (state-space models) architectures through memory geometries, and supports exact realization theorems and architecture-level comparison principles.

1 Introduction

Sequence modeling is simultaneously a problem of semantics, memory, and architecture. On the semantic side, one wants to specify a causal input-output behavior and determine which aspects of the past are relevant for predicting the future. On the architectural side, one wants to understand which hidden-state mechanisms can realize that behavior exactly. In this paper, we systematize them using Markov categories and coalgebraic final semantics.

Markov categories provide a compositional foundation for stochastic maps, conditioning, sufficient statistics, and probabilistic state transformations [6, 11, 4, 14, 12, 18, 13]. Coalgebra and automata theory provide canonical notions of state, behavior, residualization, and minimal realization [25, 24, 17].

Predictive-state and observable-operator approaches study dynamical systems through observable summaries of future behavior rather than latent generative states [19, 27, 2, 21, 1]. Meanwhile, modern neural sequence architectures instantiate sharply different memory mechanisms: full-history attention [28, 7, 3], additive linear-attention summaries [20, 5], structured state-space dynamics [16, 15, 8], sparse routed experts [26, 22, 10, 9], and hybrids that combine several of these mechanisms [23, 8].

These viewpoints can be organized by a single question:

Given a causal behavior, what is the minimal predictive memory object required to realize it, and which architecture classes can represent that object?

From this perspective, the primary distinction between Transformer, linear attention, state-space models (SSMs) such as Mamba, and mixture-of-experts (MoE) architectures is not their implementation-level branding, but the geometry of the memory objects they admit and the way those objects arise as quotients of free history. Attention-based models preserve large portions of ordered history; additive models preserve commutative sufficient statistics; state-space models preserve low-dimensional affine dynamical states; MoE modifies routing and readout while leaving the underlying memory geometry unchanged. This perspective can be categorically formalized as follows.

We model sequence behaviors as points of a final state-deterministic stochastic Mealy object, combining coalgebraic final semantics [24, 17] with Markov categories [11, 4, 13]. Concretely, fix input and output objects X and Y . Write X^* for the object of finite input histories and I for the monoidal unit. We use the final state-deterministic stochastic Mealy object

$$(\Omega, a_\Omega, \lambda)$$

as a universal object of predictive behaviors: informally, Ω collects future predictive behaviors, $a_\Omega : \Omega \otimes X^* \rightarrow \Omega$ updates such a behavior after reading a finite history, and $\lambda : \Omega \otimes X \rightarrow Y$ is the one-step readout. A behavior is a point $f : I \rightarrow \Omega$. The key construction is the residual map $r_f : X^* \rightarrow \Omega$, given by $r_f = a_\Omega \circ (f \otimes \text{id}_{X^*})$, which sends each finite history to the predictive behavior induced after reading that history. The residual map is the semantic backbone of the theory. Its image $\Sigma_f := \text{Im}(r_f)$ is the syntactic predictive-state object of f . Intuitively, Σ_f is the minimal exact memory that distinguishes exactly those histories with different predictive futures. In the concrete category **Set**, Σ_f is the quotient of the free monoid of histories by predictive equivalence, $\Sigma_f \cong X^*/\sim_f$, where $w \sim_f w' \iff r_f(w) = r_f(w')$. This makes Σ_f a stochastic and sequential analogue of a syntactic automaton

state space [25] and, at the same time, a categorical predictive-state object in the spirit of observable-operator and predictive-state models [19, 27, 2].

We then show that exact reachable realizations of f are classified by factorizations of the residual map: $\mathbf{Reach}(f) \simeq \mathbf{Fact}_{\text{repi}}(r_f)$. Thus the exact realization problem reduces to a purely categorical factorization problem. The free-history realization on X^* is initial, the syntactic predictive-state realization on Σ_f is terminal, and every other exact realization lies between them as an intermediate quotient of history. We formalize an architecture class as a *memory geometry* $\mathcal{G}$, namely an image-closed full subcategory of history actions. This yields the realizability criterion that exact realizability of f in $\mathcal{G}$ is equivalent to the condition $\Sigma_f \in \mathcal{G}$. In other words, the architecture question is a representability question for Σ_f .

The contributions of the paper can be summarized as follows.

- **A Markov-categorical exact realization theory.** We first formulate sequence models as state-deterministic stochastic Mealy objects in a Markov category and prove that exact reachable realizations of a behavior f are equivalent to regular-epimorphic factorizations of its residual map $r_f : X^* \rightarrow \Omega$. This gives a categorical realization theorem in the spirit of classical syntactic-state constructions [25, 24, 17], but adapted to causal stochastic semantics.
- **A minimal predictive memory object.** We define the syntactic predictive-state object $\Sigma_f = \text{Im}(r_f)$ and show that it is the terminal exact reachable realization of f . It therefore plays the role of minimal exact predictive memory. Conceptually, this places predictive-state ideas [19, 27, 2] inside a categorical realization theory.
- **A geometry criterion for architecture classes.** We introduce memory geometries as image-closed subcategories of history actions and prove that exact realizability in a geometry $\mathcal{G}$ is equivalent to representability of Σ_f inside $\mathcal{G}$. This turns architecture design into a problem of admissible predictive quotients.
- **A categorical comparison of sequence architectures.** We identify Transformer-type memory with the free-history action [28, 7], linear attention with additive commutative-memory geometry [20, 5], and SSM/Mamba-type models with affine-module geometry [16, 15, 8]. We also identify MoE with a routing constructor acting on readout rather than a primitive memory geometry [26, 10, 9], and hybrids of memory geometries via product/image closure, together with a separate

universality observation for hybrids retaining an explicit free-history branch [23, 8].

- **Concrete examples and separation results.** We provide concrete examples exhibiting finite exact memory, exact additive memory, exact affine memory, strict separation between additive and affine geometries, and routing/hybrid constructions. These examples make explicit which predictive distinctions are preserved or destroyed by each memory mechanism.

These results support the following interpretation: the relevant invariant of a sequence task is neither the architecture label nor the raw hidden-state dimension, but the exact predictive quotient of free history determined by the residual map.

The organization of the paper is as follows. Section 2 reviews the Markov-categorical and sequential notions required for the paper, including deterministic morphisms, state-deterministic stochastic Mealy objects, and final predictive behavior objects. Section 3 introduces residual maps, reachable realizations, the free-history realization, and the syntactic predictive-state object. Section 4 proves the exact realization theorem and classifies exact models by intermediate quotients of history. Section 5 introduces memory geometries and establishes the geometry criterion for exact realizability. Section 6 specializes the general framework to Transformer, linear attention, Mamba/SSM, MoE, and hybrid architectures. The appendices then provide supplementary material: Appendix A gives deferred proofs, Appendix B presents concrete examples and separation results, Appendix C specializes the framework to Borel spaces and Markov kernels, and Appendix D explains additional architectural implications for memory compression, routing, hybrid design, and long-context inference.

2 Preliminaries on Markov Categories and Sequential Semantics

We work in a Markov category $\mathcal{C}$ in the sense of [11]. Its monoidal unit is denoted by I , the tensor product by $\otimes$, and the canonical copy and discard morphisms by

$$\Delta_A : A \rightarrow A \otimes A, \quad !_A : A \rightarrow I.$$

For the purposes of sequence semantics, we fix an input object X and an output object Y .

A morphism $f : A \rightarrow B$ in $\mathcal{C}$ is called *deterministic* if it preserves copying, that is,

$$\Delta_B \circ f = (f \otimes f) \circ \Delta_A.$$

The deterministic morphisms form a subcategory $\mathcal{C}_{\text{det}} \subseteq \mathcal{C}$, and $\mathcal{C}_{\text{det}}$ is cartesian monoidal. In particular, tensor products of deterministic morphisms are deterministic, and the canonical projections, diagonals, and finite products live in $\mathcal{C}_{\text{det}}$. Intuitively, a morphism in $\mathcal{C}$ is a channel, while a deterministic morphism is a state update or data-processing map. This distinction is central for sequence modeling: in the architectures considered here, hidden-state evolution is deterministic, while one-step outputs may remain stochastic.

Throughout Sections 2-4, we assume that the deterministic subcategory $\mathcal{C}_{\text{det}}$ is regular and that the object X admits a free monoid object

$$(X^*, \eta, \mu, \iota), \quad \eta : I \rightarrow X^*, \quad \mu : X^* \otimes X^* \rightarrow X^*, \quad \iota : X \rightarrow X^*.$$

in $\mathcal{C}_{\text{det}}$. The free monoid object X^* is the internal analogue of the set of finite words over the alphabet X . We interpret X^* as the object of finite input histories. A *right X^* -action* in $\mathcal{C}_{\text{det}}$ is a deterministic morphism $a_Q : Q \otimes X^* \rightarrow Q$ satisfying the unit and associativity axioms $a_Q \circ (\text{id}_Q \otimes \eta) = \text{id}_Q$ and $a_Q \circ (a_Q \otimes \text{id}_{X^*}) = a_Q \circ (\text{id}_Q \otimes \mu)$. A morphism of right actions $h : (Q, a_Q) \rightarrow (Q', a_{Q'})$ is a deterministic morphism $h : Q \rightarrow Q'$ such that

$$h \circ a_Q = a_{Q'} \circ (h \otimes \text{id}_{X^*}).$$

We write $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ for the category of right X^* -actions in $\mathcal{C}_{\text{det}}$. The regular action $\mu : X^* \otimes X^* \rightarrow X^*$ is the free right X^* -action on the monoidal unit I : for every action (Q, a_Q) and every point $q_0 : I \rightarrow Q$, there is a unique action morphism

$$\alpha_{q_0} := a_Q \circ (q_0 \otimes \text{id}_{X^*}) : X^* \rightarrow Q$$

with $\alpha_{q_0} \circ \eta = q_0$. We will use this uniqueness repeatedly below.

Proposition 2.1. *The category $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ is regular. Moreover, regular epimorphisms and image factorizations are created in $\mathcal{C}_{\text{det}}$.*

Note that every right X^* -action restricts along $\iota : X \rightarrow X^*$ to a deterministic one-step update

$$u_Q := a_Q \circ (\text{id}_Q \otimes \iota) : Q \otimes X \rightarrow Q.$$

We now define the basic notion of sequential system that will internalize the architectures of interest.

Definition 2.2. A *state-deterministic stochastic Mealy object* on input X and output Y is a triple

$$(Q, a_Q, \kappa)$$

consisting of a state object Q , a deterministic right X^* -action $a_Q : Q \otimes X^* \rightarrow Q$, and an arbitrary output channel $\kappa : Q \otimes X \rightarrow Y$. Its induced one-step update is $u_Q := a_Q \circ (\text{id}_Q \otimes \iota) : Q \otimes X \rightarrow Q$. A morphism

$$h : (Q, a_Q, \kappa) \rightarrow (Q', a_{Q'}, \kappa')$$

is a deterministic morphism $h : Q \rightarrow Q'$ such that $h \circ a_Q = a_{Q'} \circ (h \otimes \text{id}_{X^*})$ and $\kappa = \kappa' \circ (h \otimes \text{id}_X)$. We denote the resulting category by $\mathbf{SMealy}_{X,Y}^{\text{sd}}(\mathcal{C})$.

The adjective “state-deterministic” emphasizes that randomness is confined to the readout channel κ ; the hidden-state action a_Q remains deterministic. This exactly matches the architectural pattern of interest in this paper: recurrence, memory compression, and routing occur through deterministic state evolution, whereas the output semantics may be stochastic.

If $h : (Q, a_Q, \kappa) \rightarrow (Q', a_{Q'}, \kappa')$ is a morphism of state-deterministic stochastic Mealy objects, then the induced one-step updates satisfy $h \circ u_Q = u_{Q'} \circ (h \otimes \text{id}_X)$. Indeed,

$$\begin{aligned} h \circ u_Q &= h \circ a_Q \circ (\text{id}_Q \otimes \iota) = a_{Q'} \circ (h \otimes \text{id}_{X^*}) \circ (\text{id}_Q \otimes \iota) \\ &= a_{Q'} \circ (\text{id}_{Q'} \otimes \iota) \circ (h \otimes \text{id}_X) = u_{Q'} \circ (h \otimes \text{id}_X). \end{aligned}$$

We assume that the category $\mathbf{SMealy}_{X,Y}^{\text{sd}}(\mathcal{C})$ admits a final object, denoted by $(\Omega, a_\Omega, \lambda)$. Thus, for every state-deterministic stochastic Mealy object (Q, a_Q, κ) , there exists a unique deterministic morphism $b_Q : Q \rightarrow \Omega$ such that

$$b_Q \circ a_Q = a_\Omega \circ (b_Q \otimes \text{id}_{X^*}), \quad \kappa = \lambda \circ (b_Q \otimes \text{id}_X).$$

We call b_Q the *semantics map* of Q . The object Ω is the internal object of predictive behaviors. A point of Ω should be read as a complete causal behavior from future inputs to future output laws, while the transition $t : \Omega \otimes X \rightarrow \Omega$ updates the predictive state after reading one more input symbol, where $t := a_\Omega \circ (\text{id}_\Omega \otimes \iota)$.

For readers who prefer a concrete model, it is useful to keep in mind the case $\mathcal{C} = \mathbf{Set}$. Then every morphism is deterministic, and a state-deterministic stochastic Mealy object is a set-based action/readout pair $a_Q : Q \times X^* \rightarrow Q$ and $\kappa : Q \times X \rightarrow Y$. Its induced one-step update is $u_Q(q, x) = a_Q(q, \iota(x))$. In this case one may take Ω to be the set of causal

functions $F : X^\omega \rightarrow Y^\omega$. Here X^ω and Y^ω denote infinite input and output streams, respectively, and causality means that agreement of two input streams on the first n symbols implies agreement of the corresponding output streams on the first n symbols. The one-step transition $t : \Omega \times X \rightarrow \Omega$ is given by residualization, $t(F, x) = \partial_x F$, and the one-step readout $\lambda : \Omega \times X \rightarrow Y$ returns the first output symbol after reading x : $\lambda(F, x) = F(xu)_1$, which is independent of the tail $u \in X^\omega$ by causality. In this model the abstract object Ω is literally the set of causal future behaviors, and the action a_Ω sends a behavior F and a finite history w to the residual behavior $\partial_w F$. This set-based picture is a useful guide for intuition, but the remainder of the paper is formulated internally in a Markov category so that stochastic readouts and probabilistic semantics are handled uniformly.

3 Residual Semantics and Predictive Memory

We now fix a deterministic point $f : I \rightarrow \Omega$, which represents the behavior to be realized exactly. The goal of this section is to extract from f a canonical memory object that records exactly the predictive information needed to continue the behavior after an arbitrary finite history.

3.1 Residual maps and pointed reachable realizations

Definition 3.1. The *residual map* of f is the deterministic morphism $r_f : X^* \rightarrow \Omega$ defined by $r_f := a_\Omega \circ (f \otimes \text{id}_{X^*})$.

Intuitively, r_f sends a finite history to the predictive behavior obtained after reading that history. In the set-based model, this is precisely

$$r_f(w) = \partial_w f.$$

The residual map is the central semantic object of the paper. All exact realizations of f will be shown to arise from factorizations of r_f .

Proposition 3.2. *The residual map $r_f : X^* \rightarrow \Omega$ is a morphism of right X^* -actions from the regular action $\mu : X^* \otimes X^* \rightarrow X^*$ to the final action $a_\Omega : \Omega \otimes X^* \rightarrow \Omega$. Equivalently, $r_f \circ \mu = a_\Omega \circ (r_f \otimes \text{id}_{X^*})$.*

We next introduce the notions of pointed realization and reachability.

Definition 3.3. A *pointed realization* of f is a state-deterministic stochastic Mealy object (Q, a_Q, κ) equipped with a deterministic initial state $q_0 : I \rightarrow$

Q such that $b_Q \circ q_0 = f$. Its *reachability map* is the induced deterministic morphism

$$\alpha_Q : X^* \rightarrow Q, \quad \alpha_Q := a_Q \circ (q_0 \otimes \text{id}_{X^*}).$$

We say that the pointed realization is *reachable* if α_Q is a regular epimorphism.

Thus reachability means that every state of Q is represented by some finite input history from the initial state. In the concrete case $\mathcal{C} = \mathbf{Set}$, a regular epimorphism is just a surjection, so this says that every state of Q is reached as $\alpha_Q(w)$ for some $w \in X^*$.

Proposition 3.4. *For every pointed realization (Q, q_0, a_Q, κ) of f , one has $b_Q \circ \alpha_Q = r_f$.*

Hence every reachable realization of f yields a factorization of the residual map through its state object.

3.2 Free-history realization and syntactic predictive-state object

The most obvious exact realization stores the entire finite history as its state.

Definition 3.5. The *free-history realization* of f is the pointed realization $H_f := (X^*, \eta, \mu, \kappa_H)$ defined by $\kappa_H := \lambda \circ (r_f \otimes \text{id}_X) : X^* \otimes X \rightarrow Y$, with initial state $\eta : I \rightarrow X^*$.

Its induced one-step update is $u_H := \mu \circ (\text{id}_{X^*} \otimes \iota) : X^* \otimes X \rightarrow X^*$, so the update appends one more input symbol to the stored history. The readout κ_H reads the next output directly from the residual behavior determined by that history.

Proposition 3.6. *The free-history realization H_f is reachable, and its semantics map is exactly the residual map $b_{H_f} = r_f$.*

The free-history realization is exact but maximally redundant: it remembers the entire input history, regardless of whether two histories induce the same predictive behavior. The syntactic predictive-state object introduced next is the canonical quotient removing exactly this redundancy.

By Proposition 3.2, the residual map is a morphism in $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$. Since $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ is regular by Proposition 2.1, the residual map admits an image factorization

$$X^* \xrightarrow{e_f} \Sigma_f \xrightarrow{m_f} \Omega$$

in $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$, where e_f is a regular epimorphism and m_f is monic.

Definition 3.7. The object $\Sigma_f := \text{Im}(r_f)$ is called the *syntactic predictive-state object* of f .

Because the factorization is taken in the action category, Σ_f carries a canonical right X^* -action $a_{\Sigma_f} : \Sigma_f \otimes X^* \rightarrow \Sigma_f$. We also define the readout $\kappa_{\Sigma_f} := \lambda \circ (m_f \otimes \text{id}_X) : \Sigma_f \otimes X \rightarrow Y$, and the initial state $q_0^\Sigma := e_f \circ \eta : I \rightarrow \Sigma_f$.

Proposition 3.8. *The tuple $(\Sigma_f, q_0^\Sigma, a_{\Sigma_f}, \kappa_{\Sigma_f})$ is a reachable pointed realization of f , and its semantics map is m_f .*

Thus Σ_f is a canonical exact realization obtained by quotienting free history by the observational equivalence encoded in the residual map.

The image map $e_f : X^* \rightarrow \Sigma_f$ has a kernel pair $R_f \rightrightarrows X^*$ in $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$, which we interpret as the *predictive congruence* associated to f .

Proposition 3.9. *The kernel pair $R_f \rightrightarrows X^*$ is an internal right congruence on the free monoid object X^* .*

In the special case $\mathcal{C} = \mathbf{Set}$, the predictive congruence is simply the equivalence relation $w \sim_f w' \iff r_f(w) = r_f(w')$. Equivalently, $w \sim_f w'$ if and only if the two histories determine the same residual causal behavior. The syntactic predictive-state object is then the quotient $\Sigma_f \cong X^*/\sim_f$.

4 Exact Realization Theory

We now show that the exact realizations of f are classified by regular-epi factorizations of the residual map. This gives a categorical realization theorem, identifies the free-history and syntactic realizations as universal extremal objects, and yields a Nerode-type classification of all exact memories.

4.1 Reachable realizations as regular-epi factorizations

We begin by packaging exact realizations into a category.

Definition 4.1. Let $\mathbf{Reach}(f)$ denote the category whose objects are reachable pointed realizations (Q, q_0, a_Q, κ) of f , and whose morphisms are deterministic morphisms $h : Q \rightarrow Q'$ of the form

$$h : (Q, q_0, a_Q, \kappa) \rightarrow (Q', q'_0, a_{Q'}, \kappa')$$

such that $h \circ q_0 = q'_0$, $h \circ a_Q = a_{Q'} \circ (h \otimes \text{id}_{X^*})$, and $\kappa = \kappa' \circ (h \otimes \text{id}_X)$.

Definition 4.2. Let $\mathbf{Fact}_{\text{repi}}(r_f)$ be the category whose objects are factorizations $X^* \xrightarrow{e} Q \xrightarrow{m} \Omega$ in $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ such that e is a regular epimorphism and $m \circ e = r_f$. A morphism $v : (e, m) \rightarrow (e', m')$ is an action morphism $v : Q \rightarrow Q'$ such that $e' = v \circ e$ and $m = m' \circ v$.

Thus $\mathbf{Fact}_{\text{repi}}(r_f)$ is the category of all regular-epi factorizations of the residual map through deterministic X^* -actions.

Theorem 4.3 (Categorical exact realization theorem). *There is an equivalence of categories $\mathbf{Reach}(f) \simeq \mathbf{Fact}_{\text{repi}}(r_f)$.*

The equivalence above immediately identifies the extremal realizations.

Corollary 4.4. *The category $\mathbf{Reach}(f)$ has an initial object and a terminal object. Its initial object is the free-history realization $H_f = (X^*, \eta, \mu, \kappa_H)$, corresponding to the factorization $X^* \xrightarrow{\text{id}_{X^*}} X^* \xrightarrow{r_f} \Omega$. Its terminal object is the syntactic predictive-state realization $(\Sigma_f, q_0^\Sigma, a_{\Sigma_f}, \kappa_{\Sigma_f})$, corresponding to the image factorization $X^* \xrightarrow{e_f} \Sigma_f \xrightarrow{m_f} \Omega$.*

Thus H_f stores all history, while Σ_f stores exactly the predictive information that survives quotienting by observational equivalence.

4.2 Classification by intermediate quotients

The universal properties above imply that every reachable realization sits canonically between free history and syntactic predictive state.

Corollary 4.5. *For every reachable realization (Q, q_0, a_Q, κ) of f , there exist unique morphisms*

$$X^* \xrightarrow{\alpha_Q} Q \xrightarrow{\pi_Q} \Sigma_f \xrightarrow{m_f} \Omega$$

in $\mathbf{Act}_{X^}(\mathcal{C}_{\text{det}})$ such that*

$$\pi_Q \circ \alpha_Q = e_f, \quad m_f \circ \pi_Q = b_Q.$$

Moreover, both α_Q and π_Q are regular epimorphisms.

Corollary 4.5 shows that every exact realization is an *intermediate quotient* of free history. In the concrete case of $\mathbf{Set}$, this becomes the following classification.

Corollary 4.6 (Set-based classification). *Assume $\mathcal{C} = \mathbf{Set}$. Then isomorphism classes of reachable exact realizations of f are in bijection with right congruences $\theta \subseteq \sim_f$ on X^* . The realization corresponding to θ has state set $Q_\theta \cong X^*/\theta$.*

Note that this is a Nerode-type classification: exact realizations are precisely quotients of free history by predictive right congruences.

5 Memory Geometries and Architecture Classes

Section 4 identifies the exact behavior of a sequence model with a factorization of the residual map $r_f : X^* \rightarrow \Omega$. The remaining question is architectural: which state objects are regarded as admissible? We formalize such admissibility as a *memory geometry*. The guiding principle is that exact realizability should depend only on the predictive quotient Σ_f , and hence any admissible class of memories must be stable under passage to exact images.

5.1 Memory geometries as image-closed subcategories

Definition 5.1. A *memory geometry* is a replete full subcategory $\mathcal{G} \subseteq \mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ which is closed under regular images. Equivalently, whenever $Q \in \mathcal{G}$ and $Q \twoheadrightarrow Q'$ is a regular epimorphism in $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$, then $Q' \in \mathcal{G}$.

The repleteness condition means that isomorphic actions are identified, while image-closure expresses the principle that forgetting redundant state information should preserve admissibility.

Definition 5.2. Let $\mathcal{G}$ be a memory geometry. We say that $f : I \rightarrow \Omega$ is *exactly realizable in $\mathcal{G}$* if there exists a reachable pointed realization $(Q, q_0, a_Q, \kappa) \in \mathbf{Reach}(f)$ whose state action (Q, a_Q) belongs to $\mathcal{G}$.

Thus the state object is constrained by the chosen geometry, while the readout channel remains free.

5.2 The geometry criterion for exact realizability

The exact realization theorem of Section 4 immediately implies that exact realizability in a geometry depends only on the syntactic predictive-state object.

Theorem 5.3 (Geometry criterion). *Let $\mathcal{G}$ be a memory geometry. Then the following are equivalent.*

$$f \text{ is exactly realizable in } \mathcal{G} \iff \Sigma_f \in \mathcal{G}.$$

Theorem 5.3 says that the true architecture-invariant semantic object is not a particular implementation, but the minimal predictive memory Σ_f . A geometry can realize f exactly precisely when it can represent Σ_f .

5.3 Representation versus compression

The residual map $r_f : X^* \rightarrow \Omega$ encodes the exact predictive semantics of all finite histories. Any exact realization in a geometry amounts to compressing the free-history state X^* into an admissible action object without collapsing histories that remain predictively distinct.

Proposition 5.4. *Let $q : X^* \twoheadrightarrow Q$ be a regular epimorphism in $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$. The following are equivalent.*

- (1) *There exists an action morphism $m : Q \rightarrow \Omega$ such that $m \circ q = r_f$.*
- (2) *There exists a reachable pointed realization of f whose reachability map is q .*
- (3) *The syntactic quotient map $e_f : X^* \twoheadrightarrow \Sigma_f$ factors through q , i.e. there exists a unique action morphism $\pi : Q \rightarrow \Sigma_f$ such that $e_f = \pi \circ q$.*

If moreover $\mathcal{C} = \mathbf{Set}$, these are also equivalent to $\ker(q) \subseteq \sim_f$.

Proposition 5.4 formalizes the distinction between *representation* and *compression*. The map q is an admissible compression of history, while the existence of m means that no predictive information required by f has been lost. The minimal exact compression is the syntactic quotient e_f , whereas the trivial compression is the identity on X^* .

5.4 Intermediate quotients as architecture-level memory choices

The exact quotients of f form a bounded compression hierarchy. The finest exact memory stores the entire history, while the coarsest exact memory stores only the predictive state.

Definition 5.5. Let $\mathfrak{Q}(f)$ denote the class of regular epimorphisms of actions $q : X^* \twoheadrightarrow Q$ such that the equivalent conditions of Proposition 5.4 hold. We preorder $\mathfrak{Q}(f)$ by

$$q \preceq q' \iff \exists h : Q \twoheadrightarrow Q' \text{ in } \mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}}) \text{ with } q' = h \circ q.$$

Here $q \preceq q'$ means that q' is a coarser exact compression than q .

Proposition 5.6. *The preordered class $\mathfrak{Q}(f)$ has a least element and a greatest element: id_{X^*} and $e_f : X^* \twoheadrightarrow \Sigma_f$. Moreover, every exact reachable realization of f determines an element of $\mathfrak{Q}(f)$, and every element of $\mathfrak{Q}(f)$ determines an exact reachable realization of f .*

In $\mathbf{Set}$, Proposition 5.6 identifies architecture-level exact memory choices with right congruences θ such that $\Delta_{X^*} \subseteq \theta \subseteq \sim_f$, ordered by reverse inclusion. The expressive question for any concrete architecture class is therefore: which of these exact quotients can its state geometry represent?

6 Transformer, Linear Attention, Mamba, and MoE

We now interpret the main sequence-model families inside the framework developed above.

6.1 Transformer as free-history action

The defining semantic feature of causal self-attention is that the current token may attend to an ordered collection of all previous key-value pairs [28, 7, 3]. At the level of exact memory, this is the *free-history action*:

$$u_H = \mu \circ (\text{id}_{X^*} \otimes \iota) : X^* \otimes X \rightarrow X^*.$$

Accordingly, the canonical abstract Transformer backbone is the free-history realization

$$H_f = (X^*, \eta, \mu, \kappa_H)$$

from Section 3.

Proposition 6.1. *Every behavior $f : I \rightarrow \Omega$ admits an exact free-history realization. Equivalently, the Transformer-type strategy of storing full ordered history is universally exact.*

Note that the Transformer should be regarded here as the *extremal non-compressing realization*, not as an image-closed geometry. Indeed, the image-closed hull of the single free action X^* would already contain all exact quotients of history and would therefore collapse the distinction between Transformer and compressed architectures. Thus, the Transformer is understood as the distinguished choice $q = \text{id}_{X^*}$ inside the compression hierarchy of Proposition 5.6.

6.2 Linear attention as additive commutative-memory geometry

Linear-attention models replace explicit history storage by a finite collection of additive sufficient statistics [20, 5]. Abstractly, this corresponds to a commutative-memory geometry.

The image-based criterion below is categorical. To recover an explicit initialized additive statistic from a chosen regular-image presentation, we use one additional lifting hypothesis here and in the next subsection: the

monoidal unit I is projective with respect to regular epimorphisms in $\mathcal{C}_{\text{det}}$. Equivalently, for every regular epimorphism

$$p : A \twoheadrightarrow Q, \quad q_0 : I \rightarrow Q \implies \exists \tilde{q}_0 : I \rightarrow A \text{ with } p \circ \tilde{q}_0 = q_0.$$

Let $(A, +, 0)$ be a commutative monoid object in $\mathcal{C}_{\text{det}}$, and let $\phi : X \rightarrow A$ be a deterministic feature map. By freeness of the monoid object X^* , there is a unique monoid morphism $\bar{\phi} : X^* \rightarrow A$ with $\bar{\phi} \circ \iota = \phi$. Define the additive action $a_\phi := + \circ (\text{id}_A \otimes \bar{\phi}) : A \otimes X^* \rightarrow A$. Its induced one-step update is

$$u_\phi := a_\phi \circ (\text{id}_A \otimes \iota) : A \otimes X \rightarrow A, \quad u_\phi(a, x) = a + \phi(x).$$

The associated history statistic $\sigma_\phi := a_\phi \circ (0 \otimes \text{id}_{X^*}) = \bar{\phi} : X^* \rightarrow A$ is uniquely characterized by $\sigma_\phi \circ \eta = 0$ and $\sigma_\phi(wx) = \sigma_\phi(w) + \phi(x)$.

More generally, for any initial point $a_0 : I \rightarrow A$, we write $\sigma_{\phi, a_0} := a_\phi \circ (a_0 \otimes \text{id}_{X^*}) : X^* \rightarrow A$ for the initialized additive history statistic. Thus $\sigma_{\phi, 0} = \sigma_\phi$.

Definition 6.2. Let $\mathbf{AddAct}_X(\mathcal{C}_{\text{det}}) \subseteq \mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ be the replete full subcategory consisting of regular images of additive actions (A, a_ϕ) as above.

By construction, $\mathbf{AddAct}_X(\mathcal{C}_{\text{det}})$ is a memory geometry.

Theorem 6.3 (Exact additive criterion). *The following are equivalent.*

- (1) f is exactly realizable in $\mathbf{AddAct}_X(\mathcal{C}_{\text{det}})$.
- (2) $\Sigma_f \in \mathbf{AddAct}_X(\mathcal{C}_{\text{det}})$.
- (3) There exist a commutative monoid object A , a feature map $\phi : X \rightarrow A$, and a regular epimorphism of actions $p : (A, a_\phi) \twoheadrightarrow (\Sigma_f, a_{\Sigma_f})$.

Under the lifting assumption above, these equivalent conditions imply that there exist a point $a_0 : I \rightarrow A$ and an action morphism $m : A \rightarrow \Omega$ such that $r_f = m \circ \sigma_{\phi, a_0}$.

Theorem 6.3 identifies the exact linear-attention regime with additive sufficiency at the geometry level: the predictive state must be a regular image of a commutative accumulated memory action. The residual map itself factors through an initialized additive statistic. This is the abstract content of kernelized linear attention, where a history is replaced by a sum of feature vectors and feature-value tensors.

6.3 Mamba and selective SSMs as affine-module geometry

State-space models and selective SSMs compress history into a finite-dimensional state updated by an affine recursion [16, 15, 8]. Their abstract geometry is

therefore affine rather than purely additive. We thus make the following affine-endomorphism assumption. Fix an internal semiring R in $\mathcal{C}_{\text{det}}$. Assume that for every internal R -module object M , there exists an object

$$\text{Aff}_R(M) \in \mathcal{C}_{\text{det}}$$

equipped with deterministic morphisms $c_M : \text{Aff}_R(M) \otimes \text{Aff}_R(M) \rightarrow \text{Aff}_R(M)$ and $j_M : I \rightarrow \text{Aff}_R(M)$, making $\text{Aff}_R(M)$ into a monoid object, together with a deterministic evaluation morphism $\text{ev}_M : M \otimes \text{Aff}_R(M) \rightarrow M$ which is a right action of this monoid object on M . This package $(\text{Aff}_R(M), c_M, j_M, \text{ev}_M)$ is taken as additional primitive structure. Note that the adjective “affine” below is always relative to this chosen monoid of affine endomorphisms and its evaluation action on M .

An *affine X -update* on an internal R -module object M is a deterministic morphism

$$\psi : X \rightarrow \text{Aff}_R(M).$$

By freeness of the monoid object X^* , there is a unique monoid morphism $\bar{\psi} : X^* \rightarrow \text{Aff}_R(M)$ with $\bar{\psi} \circ \iota = \psi$. Define the induced affine history action $a_\psi := \text{ev}_M \circ (\text{id}_M \otimes \bar{\psi}) : M \otimes X^* \rightarrow M$. Since ev_M is a right action and $\bar{\psi}$ is a monoid morphism, a_ψ is a right X^* -action. Its induced one-step update is

$$u_\psi := a_\psi \circ (\text{id}_M \otimes \iota) : M \otimes X \rightarrow M.$$

Definition 6.4. Let $\mathbf{AffAct}_{X,R}(\mathcal{C}_{\text{det}}) \subseteq \mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ be the replete full subcategory consisting of regular images of actions (M, a_ψ) induced by affine X -updates on internal R -module objects.

Again this is a memory geometry by construction.

If a_ψ is the X^* -action induced by an affine X -update $\psi : X \rightarrow \text{Aff}_R(M)$ on an internal R -module object M , and if $m_0 : I \rightarrow M$ is a point, we write $s_{\psi, m_0} := a_\psi \circ (m_0 \otimes \text{id}_{X^*}) : X^* \rightarrow M$ for the corresponding initialized affine history statistic.

Theorem 6.5 (Exact affine criterion). *The following are equivalent.*

- (1) f is exactly realizable in $\mathbf{AffAct}_{X,R}(\mathcal{C}_{\text{det}})$.
- (2) $\Sigma_f \in \mathbf{AffAct}_{X,R}(\mathcal{C}_{\text{det}})$.
- (3) There exist an internal R -module object M , an affine X -update $\psi : X \rightarrow \text{Aff}_R(M)$, and a regular epimorphism of actions $p : (M, a_\psi) \rightarrow (\Sigma_f, a_{\Sigma_f})$.

Under the additional lifting assumption above, these equivalent conditions imply that there exist a point $m_0 : I \rightarrow M$ and an action morphism $m : M \rightarrow \Omega$ such that

$$r_f = m \circ s_{\psi, m_0}.$$

In the concrete **Set** semantics examples, we fix the standard choice

$$\text{Aff}_R(M) := \text{End}_R(M) \times M,$$

with multiplication $(A, b) \cdot (A', b') := (A' \circ A, A'(b) + b')$ and unit $(\text{id}_M, 0)$. The evaluation action is $\text{ev}_M(m, (A, b)) := A(m) + b$. Indeed,

$$\text{ev}_M(\text{ev}_M(m, (A, b)), (A', b')) = (A' \circ A)(m) + A'(b) + b' = \text{ev}_M(m, (A, b) \cdot (A', b')),$$

so this is a right action. Under this standard instantiation, an affine X -update $\psi : X \rightarrow \text{Aff}_R(M)$ is exactly a family of affine maps $u_x(m) = A_x(m) + b_x$, $x \in X$, with $A_x \in \text{End}_R(M)$ and $b_x \in M$. The initialized statistic then satisfies $s_{\psi, m_0}(\epsilon) = m_0$ and $s_{\psi, m_0}(wx) = A_x(s_{\psi, m_0}(w)) + b_x$. This is the concrete interpretation used below when discussing SSM/Mamba-type affine dynamics in **Set**. Affine dynamics therefore generalize additive accumulation once the additive state is viewed inside a fixed module geometry: if an additive state object is equipped with an R -module structure and one takes $A_x = \text{id}_M$, the affine recurrence reduces to pure additive update. Thus the passage from linear attention to SSM/Mamba is, in this framework, the passage from commutative accumulation to affine state dynamics.

6.4 MoE as routing constructor rather than primitive memory geometry

Mixture-of-experts architectures do not, by themselves, specify a new state geometry; rather, they introduce a routing mechanism that mixes or selects readouts on top of an existing state space [26, 10, 9]. This can be expressed internally in the Markov category.

Let (Q, a_Q) be a deterministic state action, let E be a finite expert-label object in $\mathcal{C}_{\text{det}}$, and let $g : Q \otimes X \rightarrow E$ be a channel (the gate). Let $k : E \otimes Q \otimes X \rightarrow Y$ be a family of expert readout channels. Define the mixed readout

$$\kappa_{\text{MoE}} := k \circ (g \otimes \text{id}_{Q \otimes X}) \circ \Delta_{Q \otimes X} : Q \otimes X \rightarrow Y.$$

In **Stoch** this is exactly the conditional mixture

$$\kappa_{\text{MoE}}(dy \mid q, x) = \sum_{e \in E} g(e \mid q, x) k(dy \mid e, q, x).$$

Proposition 6.6. *For every deterministic state action (Q, a_Q) , the triple $(Q, a_Q, \kappa_{\text{MoE}})$ is again a state-deterministic stochastic Mealy object with the same state action.*

A more elaborate MoE architecture may endow each expert with its own state action (Q_e, a_{Q_e}) and initial state $q_{0,e} : I \rightarrow Q_e$. In that case the natural combined backbone is the product $Q_{\text{prod}} := \prod_{e \in E} Q_e$ equipped with the componentwise action and initial point $q_{0,\text{prod}} := \langle q_{0,e} \rangle_{e \in E} : I \rightarrow Q_{\text{prod}}$. Any stateful MoE system built from these expert states, together with a gate and expert readouts on Q_{prod} , is therefore a pointed realization whose state action is Q_{prod} . For whatever behavior f this combined system realizes, Theorem 4.3 implies that its syntactic predictive-state object Σ_f is a regular image of the product action Q_{prod} . Thus even in the stateful case, MoE is a *routing constructor on pre-existing geometries*, not a new geometry in its own right.

6.5 Hybrid architectures, including attention-SSM compositions

Hybrid architectures combine several memory mechanisms, typically by taking products, serial compositions, or alternating blocks [23, 8]. When the constituent classes are themselves memory geometries, such combinations are modeled by closure operations on memory geometries.

Definition 6.7. Let $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ be memory geometries. Their *hybrid closure* $\langle \mathcal{G}_1, \mathcal{G}_2 \rangle_{\text{prod, img}}$ is the smallest replete full subcategory of this category that contains $\mathcal{G}_1$ and $\mathcal{G}_2$, is closed under finite products, and is closed under regular images.

The product closure models parallel memories updated from the same input stream, while image closure accounts for exact quotienting of redundant combined state.

Proposition 6.8. Let $\mathcal{G}_{\text{hyb}} = \langle \mathcal{G}_1, \mathcal{G}_2 \rangle_{\text{prod, img}}$. Then f is exactly realizable by a hybrid architecture built from $\mathcal{G}_1$ and $\mathcal{G}_2$ if and only if $\Sigma_f \in \mathcal{G}_{\text{hyb}}$.

Proposition 6.9. Let (Q, a_Q) be a right X^* -action in $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$, and let $q_0 : I \rightarrow Q$ be a point. Then for every behavior $f : I \rightarrow \Omega$, there exists a reachable exact realization of f whose state action is a reachable subaction of the product action $X^* \times Q$. In fact, this reachable subaction is canonically isomorphic to the free-history action X^* .

Consequently, an attention-SSM architecture that retains an uncompressed free-history branch is universally exact.

7 Conclusion

We have given a categorical exact-realization theory for sequence behaviors in Markov categories, organized with final predictive semantics and the residual map $r_f : X^* \rightarrow \Omega$. Reachable exact realizations of a behavior f are classified by regular-epimorphic factorizations of r_f . Consequently, the image $\Sigma_f = \text{Im}(r_f)$ is the terminal reachable exact realization of f , hence the minimal exact predictive memory object.

The architectural consequence is that exact sequence modeling is fundamentally a representability problem. By formalizing architecture classes as memory geometries (image-closed full subcategories of history actions), we obtain the criterion that f is exactly realizable in a geometry $\mathcal{G}$ if and only if $\Sigma_f \in \mathcal{G}$. This gives architectural comparison at the semantic level: Transformer-type models correspond to free-history memory, linear attention to additive commutative statistics, and Mamba/SSM-type models to affine dynamical state. MoE modifies routing and readout over a chosen backbone. Hybrid architectures are governed by product formation together with exact quotienting of redundant joint state, and hybrids that retain an explicit free-history branch remain universally exact.

The examples and separation results in the appendix show that these distinctions are mathematically substantive. Finite-memory, additive, and affine behaviors arise as different kinds of exact predictive quotients, and the discounted-sum example yields a strict separation between additive and affine geometries. From this perspective, the relevant obstacle in long-context modeling is not raw sequence length by itself, but whether free history admits an exact predictive quotient representable in a compressed geometry. When such a quotient exists, recurrent compressed-state architectures are semantically justified; when it does not, explicit history or a hybrid memory becomes unavoidable.

References

- [1] Borja Balle, Xavier Carreras, Franco M. Luque & Ariadna Quattoni (2014): *Spectral Learning of Weighted Automata: A Forward-Backward Perspective*. *Machine Learning* 96(1–2), pp. 33–63.
- [2] Byron Boots, Sajid M. Siddiqi & Geoffrey J. Gordon (2011): *Closing the Learning-Planning Loop with Predictive State Representations*. *The International Journal of Robotics Research* 30(7), pp. 954–966.

- [3] Rewon Child, Scott Gray, Alec Radford & Ilya Sutskever (2019): *Generating Long Sequences with Sparse Transformers*. *arXiv:1904.10509*.
- [4] Kenta Cho & Bart Jacobs (2019): *Disintegration and Bayesian Inversion via String Diagrams*. *Mathematical Structures in Computer Science* 29(7), pp. 938–971.
- [5] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Łukasz Kaiser, David Belanger, Lucy Colwell & Adrian Weller (2020): *Rethinking Attention with Performers*. *arXiv:2009.14794*.
- [6] Jared Culbertson & Kirk Sturtz (2014): *A Categorical Foundation for Bayesian Probability*. *Applied Categorical Structures* 22(4), pp. 647–662.
- [7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le & Ruslan Salakhutdinov (2019): *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. *arXiv:1901.02860*.
- [8] Tri Dao & Albert Gu (2024): *Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality*. *arXiv:2405.21060*.
- [9] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen & Claire Cui (2022): *GLaM: Efficient Scaling of Language Models with Mixture-of-Experts*. In: *Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research* 162, pp. 5547–5569.
- [10] William Fedus, Barret Zoph & Noam Shazeer (2022): *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. *Journal of Machine Learning Research* 23(120), pp. 1–39.
- [11] Tobias Fritz (2020): *A Synthetic Approach to Markov Kernels, Conditional Independence and Theorems on Sufficient Statistics*. *Advances in Mathematics* 370, p. 107239.
- [12] Tobias Fritz, Tomáš Gonda, Paolo Perrone & Eigil Fjeldgren Rischel (2023): *Representable Markov Categories and Comparison of Statistical*

- Experiments in Categorical Probability. Theoretical Computer Science* 961, p. 113896.
- [13] Tobias Fritz, Andreas Klingler, Drew McNeely, Areeb Shah-Mohammed & Yuwen Wang (2025): *Hidden Markov Models and the Bayes Filter in Categorical Probability. IEEE Transactions on Information Theory* 71(9), pp. 7052–7075.
 - [14] Tobias Fritz & Wendong Liang (2023): *Free gs -Monoidal Categories and Free Markov Categories. Applied Categorical Structures* 31. Article 21.
 - [15] Albert Gu & Tri Dao (2023): *Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.*
 - [16] Albert Gu, Karan Goel & Christopher Ré (2022): *Efficiently Modeling Long Sequences with Structured State Spaces. In: International Conference on Learning Representations.*
 - [17] Bart Jacobs (2016): *Introduction to Coalgebra: Towards Mathematics of States and Observation. Cambridge Tracts in Theoretical Computer Science* 59, Cambridge University Press.
 - [18] Bart Jacobs (2022): *Sufficient Statistics and Split Idempotents in Discrete Probability Theory. arXiv:2212.09191.*
 - [19] Herbert Jaeger (2000): *Observable Operator Models for Discrete Stochastic Time Series. Neural Computation* 12(6), pp. 1371–1398.
 - [20] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas & François Fleuret (2020): *Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. arXiv:2006.16236.*
 - [21] Alex Kulesza, Nan Jiang & Satinder Singh (2015): *Spectral Learning of Predictive State Representations with Insufficient Statistics. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2715–2721.
 - [22] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer & Zhifeng Chen (2020): *GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. arXiv:2006.16668.*
 - [23] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Haim Meirom, Yonatan Belinkov, Shai

- Shalev-Shwartz et al. (2024): *Jamba: A Hybrid Transformer-Mamba Language Model*. arXiv:2403.19887.
- [24] J. J. M. M. Rutten (2000): *Universal Coalgebra: A Theory of Systems*. *Theoretical Computer Science* 249(1), pp. 3–80.
- [25] Marcel-Paul Schützenberger (1961): *On the Definition of a Family of Automata*. *Information and Control* 4(2–3), pp. 245–270.
- [26] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc V. Le, Geoffrey Hinton & Jeff Dean (2017): *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*. In: *International Conference on Learning Representations*.
- [27] Satinder Singh, Michael R. James & Matthew R. Rudary (2004): *Predictive State Representations: A New Theory for Modeling Dynamical Systems*. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 512–519.
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin (2017): *Attention Is All You Need*. In: *Advances in Neural Information Processing Systems* 30.

A Deferred Proofs

This appendix provides the proofs omitted above.

A.1 Proofs for Section 2

Proof of Proposition 2.1. The free monoid object X^* induces the endofunctor

$$T(A) = A \otimes X^*$$

on $\mathcal{C}_{\text{det}}$, and right X^* -actions are precisely the Eilenberg-Moore algebras for the monad determined by (X^*, η, μ) . The forgetful functor from this Eilenberg-Moore category creates finite limits. It therefore remains to control regular epimorphisms.

Let

$$e : A \twoheadrightarrow B$$

be a regular epimorphism in $\mathcal{C}_{\text{det}}$. Since $\mathcal{C}_{\text{det}}$ is cartesian monoidal, the map $e \otimes \text{id}_{X^*} : A \otimes X^* \rightarrow B \otimes X^*$ is the pullback of e along the projection

$B \otimes X^* \rightarrow B$. Regular epimorphisms are pullback-stable in a regular category, so

$$T(e) = e \otimes \text{id}_{X^*}$$

is again a regular epimorphism. Standard regularity results for Eilenberg-Moore categories of regular-epi-preserving monads therefore imply that the action category $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ is regular and that the forgetful functor creates regular epimorphisms. Since it also creates kernel pairs, the regular-image factorization of an action morphism is computed on the underlying morphism in $\mathcal{C}_{\text{det}}$. $\square$

A.2 Proofs for Section 3

Proof of Proposition 3.2. Using the definition of r_f and the associativity axiom for the action a_Ω , we compute

$$r_f \circ \mu = a_\Omega \circ (f \otimes \text{id}_{X^*}) \circ \mu = a_\Omega \circ (\text{id}_\Omega \otimes \mu) \circ (f \otimes \text{id}_{X^*} \otimes \text{id}_{X^*})$$

and

$$a_\Omega \circ (r_f \otimes \text{id}_{X^*}) = a_\Omega \circ (a_\Omega \otimes \text{id}_{X^*}) \circ (f \otimes \text{id}_{X^*} \otimes \text{id}_{X^*}).$$

Since

$$a_\Omega \circ (a_\Omega \otimes \text{id}_{X^*}) = a_\Omega \circ (\text{id}_\Omega \otimes \mu),$$

the two composites agree. $\square$

Proof of Proposition 3.4. Using the defining equations of the final object and the condition $b_Q \circ q_0 = f$, we compute

$$b_Q \circ \alpha_Q = b_Q \circ a_Q \circ (q_0 \otimes \text{id}_{X^*}) = a_\Omega \circ (b_Q \otimes \text{id}_{X^*}) \circ (q_0 \otimes \text{id}_{X^*}) = a_\Omega \circ (f \otimes \text{id}_{X^*}) = r_f.$$

$\square$

Proof of Proposition 3.6. The state action of H_f is the regular right action of X^* on itself, so its reachability map is the identity on X^* , hence a regular epimorphism. By Proposition 3.2, the map

$$r_f : X^* \rightarrow \Omega$$

is action-equivariant, and by definition $\kappa_H = \lambda \circ (r_f \otimes \text{id}_X)$. Therefore r_f is a morphism

$$H_f \rightarrow (\Omega, a_\Omega, \lambda)$$

in $\mathbf{SMealy}_{X,Y}^{\text{sd}}(\mathcal{C})$, so finality identifies b_{H_f} with r_f . $\square$

Proof of Proposition 3.8. Since m_f is an action morphism and $\kappa_{\Sigma_f} = \lambda \circ (m_f \otimes \text{id}_X)$, m_f is a morphism from $(\Sigma_f, a_{\Sigma_f}, \kappa_{\Sigma_f})$ to the final object $(\Omega, a_\Omega, \lambda)$, so by finality it is the semantics map. Furthermore,

$$m_f \circ q_0^\Sigma = m_f \circ e_f \circ \eta = r_f \circ \eta = f,$$

so this is a pointed realization of f . Finally, its reachability map is e_f , hence a regular epimorphism. $\square$

Proof of Proposition 3.9. Since e_f is a morphism of right X^* -actions, its kernel pair is stable under the action of X^* . Equivalently, if two histories are identified by e_f , then appending any further history identifies them again. This is exactly the internal right-congruence condition. $\square$

A.3 Proofs for Section 4

Proof of Theorem 4.3. We construct quasi-inverse functors.

Step 1: from reachable realizations to factorizations. Let (Q, q_0, a_Q, κ) be a reachable pointed realization of f . By Proposition 3.4 its semantics map and reachability map satisfy

$$b_Q \circ \alpha_Q = r_f.$$

Since the realization is reachable, α_Q is a regular epimorphism. Therefore

$$X^* \xrightarrow{\alpha_Q} Q \xrightarrow{b_Q} \Omega$$

is an object of $\mathbf{Fact}_{\text{repi}}(r_f)$.

Now let $h : (Q, q_0, a_Q, \kappa) \rightarrow (Q', q'_0, a_{Q'}, \kappa')$ be a morphism in $\mathbf{Reach}(f)$. Moreover,

$$h \circ \alpha_Q = a_{Q'} \circ (h \otimes \text{id}_{X^*}) \circ (q_0 \otimes \text{id}_{X^*}) = a_{Q'} \circ (q'_0 \otimes \text{id}_{X^*}) = \alpha_{Q'}.$$

Since both b_Q and $b_{Q'} \circ h$ are morphisms from (Q, a_Q, κ) to the final object $(\Omega, a_\Omega, \lambda)$, finality gives $b_Q = b_{Q'} \circ h$. Hence h defines a morphism of factorizations.

This yields a functor

$$\Phi : \mathbf{Reach}(f) \rightarrow \mathbf{Fact}_{\text{repi}}(r_f).$$

Step 2: from factorizations to reachable realizations. Conversely, let

$$X^* \xrightarrow{e} Q \xrightarrow{m} \Omega$$

be an object of $\mathbf{Fact}_{\text{repi}}(r_f)$, where Q carries an action $a_Q : Q \otimes X^* \rightarrow Q$. Define $\kappa_Q := \lambda \circ (m \otimes \text{id}_X) : Q \otimes X \rightarrow Y$, and $q_0 := e \circ \eta : I \rightarrow Q$. Then m is an action-equivariant map and, by definition of κ_Q , $\kappa_Q = \lambda \circ (m \otimes \text{id}_X)$. Hence m is a morphism $(Q, a_Q, \kappa_Q) \rightarrow (\Omega, a_\Omega, \lambda)$ in $\mathbf{SMealy}_{X,Y}^{\text{sd}}(\mathcal{C})$, so by finality m is the semantics map of (Q, a_Q, κ_Q) .

Furthermore, $m \circ q_0 = m \circ e \circ \eta = r_f \circ \eta = f$, so (Q, q_0, a_Q, κ_Q) is a pointed realization of f .

Its reachability map is $\alpha_Q = a_Q \circ (q_0 \otimes \text{id}_{X^*})$. Both α_Q and e are action morphisms from the free action X^* to Q and they agree on the generator $\eta : I \rightarrow X^*$, hence $\alpha_Q = e$. Since e is a regular epimorphism, the realization is reachable.

On morphisms, an equivariant map $v : (e, m) \rightarrow (e', m')$ satisfies $e' = v \circ e$ and $m = m' \circ v$. The first equality implies $v \circ q_0 = q'_0$, equivariance gives preservation of the state action, and the second equality gives preservation of readout:

$$\kappa_Q = \lambda \circ (m \otimes \text{id}_X) = \lambda \circ (m' \circ v \otimes \text{id}_X) = \kappa_{Q'} \circ (v \otimes \text{id}_X).$$

Thus v is a morphism of reachable realizations. This yields a functor $\Psi : \mathbf{Fact}_{\text{repi}}(r_f) \rightarrow \mathbf{Reach}(f)$.

Step 3: the two constructions are inverse. Starting from a reachable realization, applying Φ and then Ψ recovers the same action, readout, and initial state, and the semantics map remains b_Q . Conversely, starting from a factorization and applying Ψ then Φ recovers the same factorization, because the reconstructed reachability map is e and the reconstructed semantics map is m .

Therefore Φ and Ψ are quasi-inverse equivalences. $\square$

Proof of Corollary 4.4. In $\mathbf{Fact}_{\text{repi}}(r_f)$, the identity factorization of r_f is initial, and the image factorization is terminal. Transport these objects along the equivalence of Theorem 4.3. $\square$

Proof of Corollary 4.5. The existence and uniqueness of π_Q follow from the terminality of Σ_f in $\mathbf{Reach}(f)$, equivalently from the universal property of the image factorization of r_f . It remains to show that π_Q is a regular epimorphism.

Factor π_Q in the regular category $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$ as $Q \xrightarrow{p} \tilde{\Sigma} \xrightarrow{j} \Sigma_f$, where p is regular epi and j is mono. Then

$$r_f = m_f \circ \pi_Q \circ \alpha_Q = m_f \circ j \circ p \circ \alpha_Q.$$

Since both p and α_Q are regular epimorphisms, so is $p \circ \alpha_Q$. Therefore

$$X^* \xrightarrow{p \circ \alpha_Q} \tilde{\Sigma} \xrightarrow{m_f \circ j} \Omega$$

is another image factorization of r_f . By uniqueness of images, j is an isomorphism, hence π_Q is regular epi. $\square$

Proof of Corollary 4.6. Given a reachable realization Q , its reachability map $\alpha_Q : X^* \rightarrow Q$ is surjective. Hence $Q \cong X^* / \ker(\alpha_Q)$. Because $b_Q \circ \alpha_Q = r_f$, one has $\ker(\alpha_Q) \subseteq \sim_f$. The kernel of an action morphism is a right congruence, hence every reachable exact realization determines such a θ .

Conversely, let θ be a right congruence with $\theta \subseteq \sim_f$. The quotient set X^* / θ inherits a right X^* -action by concatenation, because θ is a right congruence. The containment $\theta \subseteq \sim_f$ implies that the residual map r_f is constant on θ -classes, hence factors through a well-defined map $X^* / \theta \rightarrow \Omega$. The induced action and readout therefore define a reachable exact realization.

These assignments are inverse up to isomorphism: a reachable realization Q is isomorphic to the quotient by its kernel congruence $\ker(\alpha_Q)$, and the quotient realization attached to θ has kernel congruence exactly θ . $\square$

A.4 Proofs for Section 5

Proof of Theorem 5.3. Assume first that f is exactly realizable in $\mathcal{G}$. Then there exists a reachable realization (Q, q_0, a_Q, κ) with $(Q, a_Q) \in \mathcal{G}$. By Corollary 4.5, there is a regular epimorphism of actions $\pi_Q : Q \twoheadrightarrow \Sigma_f$. Since $\mathcal{G}$ is closed under regular images, it follows that $\Sigma_f \in \mathcal{G}$.

Conversely, if $\Sigma_f \in \mathcal{G}$, then Proposition 3.8 gives a reachable pointed realization of f on Σ_f , hence f is exactly realizable in $\mathcal{G}$. $\square$

Proof of Proposition 5.4. The equivalence of (1) and (2) is exactly the construction used in the proof of Theorem 4.3: an action morphism m with $m \circ q = r_f$ determines a pointed realization on Q , and conversely any reachable realization yields such a factorization.

Assume (1). Since $r_f = m \circ q$ and q is a regular epimorphism, the universal property of the image factorization of r_f yields a unique action morphism $\pi : Q \rightarrow \Sigma_f$ such that $e_f = \pi \circ q$ and $m = m_f \circ \pi$. Thus (3) holds.

Conversely, if (3) holds, then setting $m := m_f \circ \pi$ gives $m \circ q = m_f \circ \pi \circ q = m_f \circ e_f = r_f$, so (1) holds.

In **Set**, a map factors through the quotient $q : X^* \twoheadrightarrow Q$ if and only if it is constant on q -fibers, equivalently if and only if the kernel relation of q is contained in the kernel relation of e_f , namely $\sim_f$. $\square$

Proof of Proposition 5.6. The map id_{X^*} is exact because it is the reachability map of the free-history realization H_f . The map e_f is exact by Proposition 3.8. Every exact reachable realization yields an exact quotient by Theorem 4.3, and every exact quotient yields a realization by Proposition 5.4.

The universal factorization through e_f from Corollary 4.5 shows that every exact quotient lies between id_{X^*} and e_f . $\square$

A.5 Proofs for Section 6

Proof of Proposition 6.1. This is exactly Proposition 3.6. $\square$

Proof of Theorem 6.3. The equivalence of (1) and (2) is the Geometry Criterion, Theorem 5.3. The equivalence of (2) and (3) is exactly the definition of $\mathbf{AddAct}_X(\mathcal{C}_{\text{det}})$, using repleteness to replace a chosen regular image isomorphic to Σ_f by Σ_f itself.

For the final claim, let $p : (A, a_\phi) \twoheadrightarrow (\Sigma_f, a_{\Sigma_f})$ be as in (3). By the additional lifting assumption, the initial state $q_0^\Sigma : I \rightarrow \Sigma_f$ lifts along p to a point $a_0 : I \rightarrow A$ with $p \circ a_0 = q_0^\Sigma$. Since p is an action morphism, one has

$$\begin{aligned} p \circ \sigma_{\phi, a_0} &= p \circ a_\phi \circ (a_0 \otimes \text{id}_{X^*}) = a_{\Sigma_f} \circ (p \otimes \text{id}_{X^*}) \circ (a_0 \otimes \text{id}_{X^*}) \\ &= a_{\Sigma_f} \circ (q_0^\Sigma \otimes \text{id}_{X^*}) = e_f. \end{aligned}$$

Therefore, with $m := m_f \circ p$, we obtain $m \circ \sigma_{\phi, a_0} = m_f \circ e_f = r_f$. $\square$

Proof of Theorem 6.5. The equivalence of (1) and (2) is the Geometry Criterion, Theorem 5.3. The equivalence of (2) and (3) is exactly the definition of $\mathbf{AffAct}_{X,R}(\mathcal{C}_{\text{det}})$, again using repleteness to replace a chosen regular image isomorphic to Σ_f by Σ_f itself.

For the final claim, let $p : (M, a_\psi) \twoheadrightarrow (\Sigma_f, a_{\Sigma_f})$ be as in (3). By the additional lifting assumption, the initial state $q_0^\Sigma : I \rightarrow \Sigma_f$ lifts along p to a point $m_0 : I \rightarrow M$ with $p \circ m_0 = q_0^\Sigma$. Since p is an action morphism, one has

$$\begin{aligned} p \circ s_{\psi, m_0} &= p \circ a_\psi \circ (m_0 \otimes \text{id}_{X^*}) = a_{\Sigma_f} \circ (p \otimes \text{id}_{X^*}) \circ (m_0 \otimes \text{id}_{X^*}) \\ &= a_{\Sigma_f} \circ (q_0^\Sigma \otimes \text{id}_{X^*}) = e_f. \end{aligned}$$

Therefore, with $m := m_f \circ p$, we obtain $m \circ s_{\psi, m_0} = m_f \circ e_f = r_f$. $\square$

Proof of Proposition 6.6. The only condition for a state-deterministic stochastic Mealy object is that the hidden-state action be deterministic. In the MoE construction the action remains a_Q ; only the readout channel is replaced by κ_{MoE} . $\square$

Proof of Proposition 6.8. The hybrid closure is again a memory geometry, so the statement follows immediately from Theorem 5.3. $\square$

Proof of Proposition 6.9. Let $j := \langle \text{id}_{X^*}, \alpha_{q_0} \rangle : X^* \rightarrow X^* \times Q$, where $\alpha_{q_0} := a_Q \circ (q_0 \otimes \text{id}_{X^*})$. Both components are action morphisms, hence so is j . Since $\pi_1 \circ j = \text{id}_{X^*}$, the morphism j is monic. Factor j as $X^* \xrightarrow{e} H \xrightarrow{m} X^* \times Q$ with e a regular epimorphism and m monic. Because j is monic, e is monic as well, hence e is an isomorphism in the regular category $\mathbf{Act}_{X^*}(\mathcal{C}_{\text{det}})$. Thus $H \cong X^*$ as an action. Transporting the free-history realization of Proposition 6.1 across this isomorphism yields a reachable exact realization of f on H . $\square$

B Concrete Examples and Separation Results

In this section we work in the concrete semantics of Section 2 with $\mathcal{C} = \mathbf{Set}$. Thus a behavior is a causal map $F : X^\omega \rightarrow Y^\omega$, its residual map is $r_F : X^* \rightarrow \Omega$, with $r_F(w) = \partial_w F$, and the syntactic predictive-state object is the quotient $\Sigma_F \cong X^* / \sim_F$, where $w \sim_F w' \iff r_F(w) = r_F(w')$. Whenever we refer to affine-module geometry in this appendix, we use the standard $\mathbf{Set}$ instantiation from Section 6, namely $\text{Aff}_R(M) = \text{End}_R(M) \times M$ with evaluation $\text{ev}_M(m, (A, b)) = A(m) + b$. The examples below illustrate, respectively, finite exact memory, exact additive memory, exact affine memory, separation between additive and affine geometries, and the role of routing and hybrid product geometries.

B.1 A finite-memory example

We begin with a simple behavior whose predictive state space is finite and transparent.

Let $X = \{0, 1\}$ and $Y = \{\perp, 0, 1\}$, where $\perp$ denotes a distinguished initial symbol. Define the *last-symbol predictor* $F^{\text{last}} : X^\omega \rightarrow Y^\omega$ by

$$F^{\text{last}}(x_1 x_2 x_3 \cdots)_1 = \perp, \quad F^{\text{last}}(x_1 x_2 x_3 \cdots)_{t+1} = x_t \quad (t \geq 1).$$

Thus the output at time $t + 1$ is the input seen at time t .

For a finite word $w \in X^*$, define

$$\ell(w) = \begin{cases} \perp, & w = \epsilon, \\ \text{the last symbol of } w, & w \neq \epsilon. \end{cases}$$

Proposition B.1. *For the behavior F^{last} , one has $w \sim_{F^{\text{last}}} w' \iff \ell(w) = \ell(w')$. Consequently, $\Sigma_{F^{\text{last}}} \cong \{\perp, 0, 1\}$, and F^{last} admits an exact 3-state realization.*

Proof. Fix a history w and a continuation $u = u_1u_2u_3\cdots$. The residual behavior $\partial_w F^{\text{last}}$ has first output equal to $\ell(w)$, and thereafter outputs the continuation symbols: $\partial_w F^{\text{last}}(u) = \ell(w)u_1u_2u_3\cdots$. Hence the residual depends only on $\ell(w)$. Therefore $\ell(w) = \ell(w') \implies r_{F^{\text{last}}}(w) = r_{F^{\text{last}}}(w')$.

Conversely, if $\ell(w) \neq \ell(w')$, then the first output symbols of $\partial_w F^{\text{last}}(u)$ and $\partial_{w'} F^{\text{last}}(u)$ differ for every continuation u . Therefore the residuals are distinct. This proves the equivalence and identifies the three predictive states. $\square$

This example shows that the abstract quotient Σ_F can recover the usual notion of a finite exact memory state. Here the entire future behavior is determined by one bit of remembered content together with an initial marker.

B.2 An exact additive-statistic example

We now give a behavior whose predictive state is an additive commutative statistic.

Let $X = Y = \{0, 1\}$, and define the *running-parity behavior* $F^{\text{par}} : X^\omega \rightarrow Y^\omega$ by $F^{\text{par}}(x_1x_2x_3\cdots)_t = \sum_{i=1}^t x_i \pmod{2}$. For a finite word $w = x_1\cdots x_n$, set $\sigma_{\text{par}}(w) := \sum_{i=1}^n x_i \pmod{2} \in \mathbb{Z}_2$.

Proposition B.2. *For the running-parity behavior, one has $w \sim_{F^{\text{par}}} w' \iff \sigma_{\text{par}}(w) = \sigma_{\text{par}}(w')$. Hence $\Sigma_{F^{\text{par}}} \cong \mathbb{Z}_2$, and F^{par} is exactly realizable in additive commutative-memory geometry.*

Proof. Let $p = \sigma_{\text{par}}(w)$, and let $u = u_1u_2u_3\cdots$. After the history w , the future outputs are

$$\partial_w F^{\text{par}}(u)_1 = p + u_1, \quad \partial_w F^{\text{par}}(u)_2 = p + u_1 + u_2, \quad \dots$$

with all sums taken modulo 2. Thus the residual depends only on p , so equal parity implies equal residual.

Conversely, if $\sigma_{\text{par}}(w) \neq \sigma_{\text{par}}(w')$, choose the continuation $u = 000\cdots$. Then the residual outputs are constant sequences with different first symbols, so the residuals are distinct. Therefore the predictive state is exactly the parity class.

The statistic σ_{par} is additive: $\sigma_{\text{par}}(wx) = \sigma_{\text{par}}(w) + x \pmod{2}$, so the exact realization lies in additive commutative-memory geometry. $\square$

This is the prototypical linear-attention situation: all predictive information is contained in a commutative sufficient statistic, and no order-sensitive history beyond that statistic survives in Σ_F .

B.3 An exact affine-state example

We next consider an example whose predictive quotient evolves by a one-dimensional affine recurrence but is no longer commutative-additive.

Fix a parameter $0 < \lambda < 1$. Let $X = \{0, 1\}$ and $Y = \mathbb{R}$, and define the *discounted-sum behavior* $F^\lambda : X^\omega \rightarrow \mathbb{R}^\omega$ by the state recursion $s_0 = 0$, $s_t = \lambda s_{t-1} + x_t$, and $F^\lambda(x_1 x_2 x_3 \cdots)_t = s_t$. For a finite word $w = x_1 \cdots x_n$, define the associated state $s(w) := \sum_{i=1}^n \lambda^{n-i} x_i$. Equivalently, $s(\epsilon) = 0$ and $s(wx) = \lambda s(w) + x$.

Proposition B.3. *For the discounted-sum behavior F^λ , one has $w \sim_{F^\lambda} w' \iff s(w) = s(w')$. Consequently, $\Sigma_{F^\lambda} \cong S_\lambda := \{s(w) : w \in X^*\} \subseteq \mathbb{R}$, and F^λ is exactly realizable in the affine-module geometry $\mathbf{AffAct}_{X, \mathbb{R}}(\mathbf{Set})$.*

Proof. Fix a history w with state $s = s(w)$, and let $u = u_1 u_2 u_3 \cdots$. The continuation behavior is obtained by restarting the same affine recursion from initial state s : $\partial_w F^\lambda(u)_1 = \lambda s + u_1$, $\partial_w F^\lambda(u)_2 = \lambda(\lambda s + u_1) + u_2$, and so on. Hence the residual depends only on the scalar state $s(w)$.

Conversely, suppose $s(w) \neq s(w')$. Take the continuation $u = 000 \cdots$. Then $\partial_w F^\lambda(u)_1 = \lambda s(w)$ and $\partial_{w'} F^\lambda(u)_1 = \lambda s(w')$. Since $0 < \lambda < 1$, these are distinct, so the residuals differ. Thus the predictive equivalence classes are exactly the level sets of s .

Let $X^+ := X^* \setminus \{\epsilon\}$ and $W := \mathbb{R}^{(X^+)}$, the free $\mathbb{R}$ -vector space on basis $\{e_w : w \in X^+\}$. For each $x \in X = \{0, 1\}$, define the linear endomorphism $A_x : W \rightarrow W$ by $A_x(e_w) := e_{wx}$, and the affine self-map $U_x : W \rightarrow W$ by $U_x(v) := A_x(v) + e_x$. Equivalently, with the standard $\mathbf{Set}$ affine package above, this is the affine update $\psi(x) := (A_x, e_x) \in \mathbf{Aff}_{\mathbb{R}}(W)$.

Choose the shortlex order on X^+ with $0 \prec 1$. It is preserved by right concatenation: if $u \prec v$, then $ux \prec vx$ for each $x \in X$. For $v \neq 0$, let $\mu(v)$ denote the maximal word in the finite support of v , and set $\mu(0) := \epsilon$. Define $p : W \rightarrow S_\lambda$ by $p(v) := s(\mu(v))$. This map is surjective because $p(0) = s(\epsilon) = 0$ and $p(e_w) = s(w)$ for every $w \in X^+$.

Let $u_\lambda : S_\lambda \times X \rightarrow S_\lambda$ be the one-step update induced by the discounted-sum recurrence on S_λ , given by $u_\lambda(\sigma, x) := \lambda\sigma + x$. For $v = 0$, one has

$$p(U_x(0)) = p(e_x) = s(x) = x = \lambda \cdot 0 + x = u_\lambda(p(0), x).$$

If $v \neq 0$, then $\text{supp}(U_x(v)) = \{wx : w \in \text{supp}(v)\} \cup \{x\}$. Because shortlex is preserved by right concatenation and every word wx with $w \neq \epsilon$ is longer than x , the maximal support word of $U_x(v)$ is $\mu(v)x$. Therefore

$$p(U_x(v)) = s(\mu(v)x) = \lambda s(\mu(v)) + x = \lambda p(v) + x = u_\lambda(p(v), x).$$

So p intertwines the one-step updates, hence also the induced X^* -actions. In **Set**, surjective maps are regular epimorphisms. Therefore $p : (W, a_\psi) \rightarrow (S_\lambda, a_\lambda)$ is a regular epimorphism of actions, where a_λ is the X^* -action induced by u_λ . Thus $(S_\lambda, a_\lambda) \in \mathbf{AffAct}_{X, \mathbb{R}}(\mathbf{Set})$. Since $\Sigma_{F^\lambda} \cong S_\lambda$, Theorem 5.3 implies that F^λ is exactly realizable in affine-module geometry. $\square$

This example captures the core semantic pattern of selective SSM and Mamba-type models: the predictive quotient is governed by the one-dimensional affine recurrence $s \mapsto \lambda s + x$ on $S_\lambda \subseteq \mathbb{R}$, and this quotient belongs to affine-module geometry.

B.4 A non-representability or separation example

The previous example is affine. We now show that it is *not* commutative-additive. This gives a clean separation between additive and affine geometries.

Proposition B.4. *For every $0 < \lambda < 1$, the discounted-sum behavior F^λ is not exactly realizable in additive commutative-memory geometry.*

Proof. Suppose, toward a contradiction, that F^λ were exactly realizable in additive commutative-memory geometry. Since this section works in **Set**, the lifting hypothesis used in Theorem 6.3 is automatic. Therefore Theorem 6.3 supplies a commutative monoid A , a feature map $\phi : X \rightarrow A$, an initial point $a_0 \in A$, and a map $m : A \rightarrow \Omega$ such that $r_{F^\lambda} = m \circ \sigma_{\phi, a_0}$.

By additivity and commutativity,

$$\sigma_{\phi, a_0}(01) = a_0 + \phi(0) + \phi(1) = a_0 + \phi(1) + \phi(0) = \sigma_{\phi, a_0}(10).$$

Hence $r_{F^\lambda}(01) = m(\sigma_{\phi, a_0}(01)) = m(\sigma_{\phi, a_0}(10)) = r_{F^\lambda}(10)$.

But the affine states after the two prefixes are different: $s(01) = 1$ and $s(10) = \lambda$. Now choose the continuation $u = 000 \dots$. Then the first residual outputs are $\partial_{01} F^\lambda(u)_1 = \lambda$ and $\partial_{10} F^\lambda(u)_1 = \lambda^2$. Since $0 < \lambda < 1$, these are distinct, and so the residuals are different. This contradicts exact realizability through an additive commutative statistic. $\square$

Corollary B.5. *In the concrete semantics of **Set**, additive commutative-memory geometry does not capture all behaviors exactly realizable in the affine-module geometry $\mathbf{AffAct}_{X, \mathbb{R}}(\mathbf{Set})$: there exist behaviors that are exactly realizable in $\mathbf{AffAct}_{X, \mathbb{R}}(\mathbf{Set})$ but not in additive geometry.*

Proof. The discounted-sum behavior F^λ is exactly realizable in the affine-module geometry $\mathbf{AffAct}_{X, \mathbb{R}}(\mathbf{Set})$ by Proposition B.3 and not exact additive by Proposition B.4. $\square$

The obstruction is conceptual rather than dimensional: a commutative additive statistic cannot distinguish the ordered histories 01 and 10, whereas the affine dynamics can. Thus order sensitivity alone already forces one beyond purely additive memory.

B.5 A routing and hybrid example

We finally exhibit an example combining additive memory, affine memory, and routing.

Fix $0 < \lambda < 1$, and let $X = \{P, S\} \times \{0, 1\}$ and $Y = \mathbb{R}$. We interpret an input letter (m, b) as a mode m together with a bit b . Define state variables $p_t \in \mathbb{Z}_2$ and $s_t \in \mathbb{R}$, initialized by $p_0 = 0$ and $s_0 = 0$. Given input $x_t = (m_t, b_t)$, update by

$$(p_t, s_t) = \begin{cases} (p_{t-1} + b_t \pmod{2}, s_{t-1}), & m_t = P, \\ (p_{t-1}, \lambda s_{t-1} + b_t), & m_t = S. \end{cases}$$

The output is

$$y_t = \begin{cases} p_t, & m_t = P, \\ s_t, & m_t = S, \end{cases}$$

where $p_t \in \{0, 1\} \subset \mathbb{R}$ is viewed as a real number.

Denote the resulting behavior by $F_\lambda^{\text{hyb}} : X^\omega \rightarrow \mathbb{R}^\omega$.

Proposition B.6. *The behavior F_λ^{hyb} admits a reachable exact realization on the product state space $\mathbb{Z}_2 \times S_\lambda$, with initial state $(0, 0)$, where $\mathbb{Z}_2$ carries additive memory and S_λ carries affine memory. Moreover, its readout is a hard-routed mixture of two experts: one expert reads the additive component and one expert reads the affine component.*

Proof. For a history w , let $p(w) \in \mathbb{Z}_2$ be the parity accumulated over the letters with mode P , and let $s(w) \in S_\lambda$ be the affine state accumulated over the letters with mode S . Then the update from state (p, s) on input (m, b) is

$$u((p, s), (m, b)) = \begin{cases} (p + b, s), & m = P, \\ (p, \lambda s + b), & m = S, \end{cases}$$

where $p + b$ is taken modulo 2.

The residual after history w depends only on the pair $(p(w), s(w))$, since future outputs are computed solely from the current parity component and the current affine component according to the same mode-dependent rule. Hence the behavior factors through the product state.

To see exactness, suppose $(p, s) \neq (p', s')$. If $p \neq p'$, choose a continuation consisting only of mode- P inputs with bit 0. The first output then reveals the parity component, so the residuals differ. If instead $p = p'$ but $s \neq s'$, choose a continuation consisting only of mode- S inputs with bit 0. The first output is then λs versus $\lambda s'$, so the residuals again differ. Therefore distinct product states induce distinct residuals.

It remains to check reachability from the initial state $(0, 0)$. Let $(p, s) \in \mathbb{Z}_2 \times S_\lambda$. Choose a word $v \in \{0, 1\}^*$ with $s(v) = s$, and read it as an S -mode input history

$$((S, v_1), \dots, (S, v_n)).$$

Starting from $(0, 0)$, this reaches $(0, s)$, because P -mode updates do not affect the affine component and the S -mode recurrence is exactly the discounted-sum update on S_λ . If $p = 1$, append one more input $(P, 1)$; if $p = 0$, append nothing. Since S -mode updates do not affect parity and P -mode updates do not affect the affine component, this reaches (p, s) . Hence every state in $\mathbb{Z}_2 \times S_\lambda$ is reachable.

Finally, define the gate and expert readouts by

$$\begin{aligned} g((p, s), (m, b)) &:= m, \\ k_P((p, s), (m, b)) &:= p + b \pmod{2}, \\ k_S((p, s), (m, b)) &:= \lambda s + b. \end{aligned}$$

Then the actual readout is obtained by selecting the appropriate expert according to the mode $m \in \{P, S\}$. This is exactly a hard-routed mixture of experts on top of a product memory state. $\square$

This example separates three roles cleanly: the parity component is additive memory; the discounted-sum component is affine memory; and the gate chooses which readout to expose at each step. Hence routing is layered *on top of* memory geometry, while the hybrid product carries the true predictive state.

C Specialization to Borel Spaces and Markov Kernels

This appendix specializes Sections 2–6 to the canonical standard-Borel Markov category $\mathcal{C} = \text{Stoch}$ of standard Borel spaces and measurable Markov kernels. This is the concrete kernel semantics already implicit in the discussion of Section 6. A morphism $k : X \rightarrow Y$ is a kernel $k(dy \mid x)$, composition is

given by integration,

$$(\ell \circ k)(C \mid x) = \int_Y \ell(C \mid y) k(dy \mid x),$$

and tensor product is given by independent product of kernels. The monoidal unit I is the one-point measurable space. Deterministic morphisms are precisely the Dirac kernels induced by Borel maps, so $\mathbf{Stoch}_{\text{det}}$ identifies with the category of standard Borel spaces and Borel maps [11, 12].

To keep the image and quotient constructions of Sections 3–6 literally inside standard Borel spaces without adding further descriptive-set-theoretic hypotheses, we work in the discrete-history regime relevant to token sequences: throughout this appendix, X is assumed to be a finite or countably infinite discrete standard Borel space, while Y is an arbitrary standard Borel space. Then

$$X^* \cong \bigsqcup_{n \geq 0} X^n$$

is again a countable discrete standard Borel space, with free monoid structure given by the empty word, concatenation, and one-letter insertion. We also retain the standing assumption from Section 2 that $\mathbf{SMealy}_{X,Y}^{\text{sd}}(\mathbf{Stoch})$ admits a final object $(\Omega, a_\Omega, \lambda)$. Whenever a concrete presentation of Ω by causal kernels $X^\omega \rightarrow Y^\omega$ is chosen, a point $f : I \rightarrow \Omega$ may be read exactly as a causal stochastic input-output law.

C.1 Concrete stochastic semantics and predictive quotients

In $\mathbf{Stoch}$, a pointed state-deterministic stochastic Mealy object (Q, q_0, a_Q, κ) is exactly a standard Borel state space Q , a Borel state update

$$u_Q := a_Q \circ (\text{id}_Q \otimes \iota) : Q \otimes X \rightarrow Q,$$

an initial state $q_0 : I \rightarrow Q$, and a Markov kernel $\kappa : Q \otimes X \rightarrow Y$, which in measure-theoretic notation we write as $\kappa(dy \mid q, x)$. Thus, given an input word $x_1 \cdots x_n$, one recursively forms states

$$q_t = u_Q(q_{t-1}, x_t), \quad q_0 \text{ given,}$$

and emits outputs according to the kernels

$$y_t \sim \kappa(- \mid q_{t-1}, x_t).$$

In measure-theoretic terms, exact realizations in **Stoch** are therefore controlled hidden-state models with deterministic transition and stochastic read-out.

The discrete-history hypothesis has an immediate structural consequence: every reachable exact realization is automatically countable-state. This turns the abstract exact realization theorem of Section 4 into a genuine probabilistic Nerode theorem.

Theorem C.1 (Probabilistic Nerode theorem in **Stoch**). *Assume that X is finite or countably infinite discrete. For every behavior $f : I \rightarrow \Omega$, the following hold.*

1. *The syntactic predictive-state object Σ_f is a countable standard Borel space.*
2. *Every reachable pointed realization (Q, q_0, a_Q, κ) of f in **Stoch** has countable standard Borel state space Q .*
3. *Isomorphism classes of reachable exact realizations of f are in bijection with right congruences $\theta \subseteq \sim_f$ on the free monoid X^* , where*

$$w \sim_f w' \iff r_f(w) = r_f(w').$$

The realization corresponding to θ has state space X^/θ , with deterministic update induced by concatenation and readout kernel induced from r_f .*

4. *The syntactic predictive-state realization is the minimal exact measurable realization:*

$$\Sigma_f \cong X^*/\sim_f,$$

and every reachable exact realization of f is an intermediate measurable quotient of free history.

Proof. Since X is countable discrete, X^* is countable discrete. If (Q, q_0, a_Q, κ) is reachable, then its reachability map $\alpha_Q : X^* \rightarrow Q$ is surjective, so the underlying set of Q is countable. Every countable standard Borel space is Borel-isomorphic to a countable discrete measurable space. Hence both Q and Σ_f lie in the countable discrete fragment of **Stoch**_{det}. The classification statement is therefore exactly Corollary 4.6, and the minimality statement is Proposition 3.8 specialized to this situation. $\square$

Theorem C.1 says that, for discrete token alphabets, every exact stochastic sequence behavior has a minimal countable predictive state space, regardless of whether one chooses to embed that state in a continuous ambient

memory object. Equivalently, every reachable exact realization in *Stoch* may be viewed as a countable-state probabilistic predictive automaton with deterministic transition and Markov-kernel emission law. The continuous geometry of a model may still be essential for obtaining a natural update law, but the exact reachable predictive quotient itself is countable.

Corollary C.2 (Predictive sufficient statistics in *Stoch*). *Let $q : X^* \rightarrow Q$ be a surjective measurable map into a standard Borel space Q . The following are equivalent.*

1. *The map q is the reachability map of a reachable exact realization of f in *Stoch*.*
2. *There exist a measurable right X^* -action $a_Q : Q \otimes X^* \rightarrow Q$ and a measurable morphism $m : Q \rightarrow \Omega$ such that $q : (X^*, \mu) \rightarrow (Q, a_Q)$ is an action morphism and $r_f = m \circ q$.*
3. *Whenever $q(w) = q(w')$, one has*

$$r_f(w) = r_f(w') \quad \text{and} \quad q(wx) = q(w'x) \text{ for every } x \in X.$$

If Ω is concretely realized as a space of causal kernels $X^\omega \rightarrow Y^\omega$, then these conditions say exactly that q is a measurable sufficient statistic of the finite history for the full future output law, together with a deterministic measurable update induced by appending one more input symbol.

Proof. Since q is a surjective map out of the countable space X^* , the space Q is countable standard Borel and therefore discrete. The equivalence of (1) and (2) is Proposition 5.4, because in the countable discrete fragment surjective measurable maps are exactly regular epimorphisms and the forgetful functor from actions creates them. Assume (2). If $q(w) = q(w')$, then $r_f(w) = m(q(w)) = m(q(w')) = r_f(w')$. Writing

$$u_Q := a_Q \circ (\text{id}_Q \otimes \iota) : Q \otimes X \rightarrow Q,$$

the action-morphism property gives

$$q(wx) = u_Q(q(w), x) = u_Q(q(w'), x) = q(w'x)$$

for every $x \in X$. Hence (3) holds.

Conversely, assume (3). By induction on $v \in X^*$, the one-step stability in (3) implies

$$q(w) = q(w') \implies q(wv) = q(w'v).$$

Therefore

$$a_Q(q(w), v) := q(wv)$$

is a well-defined right X^* -action on Q , and $q : (X^*, \mu) \rightarrow (Q, a_Q)$ is an action morphism by construction. Likewise, the first clause in (3) makes

$$m(q(w)) := r_f(w)$$

well defined. Since Q is countable discrete, both a_Q and m are measurable, and $r_f = m \circ q$. Thus (2) holds. $\square$

Corollary C.2 identifies Σ_f as the terminal exact predictive sufficient statistic of the history among those admitting a deterministic measurable update. Equivalently, exact realizations in **Stoch** are precisely those deterministic-state stochastic systems whose state is a measurable quotient of free history by a right congruence that preserves the full predictive kernel.

C.2 Additive and affine memory in measure-theoretic form

The abstract additive and affine criteria of Section 6 become concrete quotient-and-factorization statements for measurable history statistics. Because the monoidal unit in **Stoch**_{det} is the one-point space, the lifting hypothesis used in Theorems 6.3 and 6.5 is automatic.

Theorem C.3 (Exact additive criterion in **Stoch**). *The following are equivalent.*

1. f is exactly realizable in $\mathbf{AddAct}_X(\mathbf{Stoch}_{\text{det}})$.
2. $\Sigma_f \in \mathbf{AddAct}_X(\mathbf{Stoch}_{\text{det}})$.
3. There exist a commutative monoid object $(A, +, 0)$ in $\mathbf{Stoch}_{\text{det}}$, a measurable feature map $\phi : X \rightarrow A$, a regular epimorphism of actions $p : (A, a_\phi) \twoheadrightarrow (\Sigma_f, a_{\Sigma_f})$.

Under these equivalent conditions, there also exist an initial point $a_0 : I \rightarrow A$, and a measurable action morphism $m : (A, a_\phi) \rightarrow (\Omega, a_\Omega)$ such that

$$r_f = m \circ \sigma_{\phi, a_0}, \quad \sigma_{\phi, a_0}(\varepsilon) = a_0, \quad \sigma_{\phi, a_0}(wx) = \sigma_{\phi, a_0}(w) + \phi(x).$$

*In particular, if $\sigma_{\phi, a_0}(w) = \sigma_{\phi, a_0}(w')$, then the future predictive kernels after w and w' are identical. Thus exact linear-attention-type realizability in **Stoch** is equivalent to the predictive state being a measurable quotient of an additive memory action; in that case the full predictive kernel factors through an initialized additive statistic of the history.*

Proof. This is Theorem 6.3 specialized to $\mathcal{C} = \mathbf{Stoch}$, together with the observation that the singleton object I lifts along every surjective Borel map. The final implication is immediate from the factorization $r_f = m \circ \sigma_{\phi, a_0}$. $\square$

Theorem C.3 makes the exact restriction of additive memory completely explicit: no predictive distinction may survive beyond what is encoded by the accumulated commutative statistic. Hence additive exactness forces order information to disappear unless it is first encoded into the chosen feature map.

Theorem C.4 (Exact affine criterion in $\mathbf{Stoch}$). *Assume that the affine-endomorphism package of Section 6 has been chosen in $\mathbf{Stoch}_{\text{det}}$ for an internal semiring R . Then the following are equivalent.*

1. f is exactly realizable in $\mathbf{AffAct}_{X,R}(\mathbf{Stoch}_{\text{det}})$.
2. $\Sigma_f \in \mathbf{AffAct}_{X,R}(\mathbf{Stoch}_{\text{det}})$.
3. There exist an internal R -module object M , an affine X -update $\psi : X \rightarrow \mathbf{Aff}_R(M)$, a regular epimorphism of actions $p : (M, a_\psi) \twoheadrightarrow (\Sigma_f, a_{\Sigma_f})$.

Under these equivalent conditions, there also exist an initial point $m_0 : I \rightarrow M$, and a measurable action morphism $m : (M, a_\psi) \rightarrow (\Omega, a_\Omega)$ such that

$$r_f = m \circ s_{\psi, m_0}, \quad s_{\psi, m_0}(\varepsilon) = m_0, \quad s_{\psi, m_0}(wx) = u_\psi(s_{\psi, m_0}(w), x).$$

Thus exact SSM/Mamba-type realizability in $\mathbf{Stoch}$ is equivalent to the predictive state being a measurable quotient of an affine recursion state; in that case the full predictive kernel factors through an initialized affine statistic.

Proof. This is Theorem 6.5 specialized to $\mathcal{C} = \mathbf{Stoch}$, again using that the singleton object I lifts along surjective Borel maps. $\square$

In concrete subclasses of standard Borel R -modules where the chosen affine-endomorphism package is presented by a standard Borel parameter space of pairs (A, b) with measurable evaluation, one may write this package in the familiar form $\mathbf{Aff}_R(M) \cong \mathbf{End}_R(M) \times M$ and an affine update as $\psi(x) = (A_x, b_x)$. In such cases the recursion reads

$$s_{\psi, m_0}(\varepsilon) = m_0, \quad s_{\psi, m_0}(wx) = A_x(s_{\psi, m_0}(w)) + b_x.$$

No such closed-form measurable presentation is needed for Theorem C.4; in particular, the theorem does not assert that every standard Borel R -module admits one. It is only relative to the chosen affine-endomorphism package in $\mathbf{Stoch}_{\text{det}}$.

Corollary C.5 (Continuous ambient memories have countable reachable image). *Under the discrete-history hypothesis, every exact additive or affine realization of f in **Stoch** has countable reachable subset inside its ambient memory object. More precisely, $\sigma_{\phi,a_0}(X^*) \subseteq A$ and $s_{\psi,m_0}(X^*) \subseteq M$ are countable, and Σ_f is a countable measurable quotient of either reachable subset whenever the corresponding exact realization exists.*

Proof. Both reachable subsets are images of the countable set X^* . The quotient statement is Proposition 5.4 applied to the factorizations $r_f = m \circ \sigma_{\phi,a_0}$ and $r_f = m \circ s_{\psi,m_0}$. $\square$

Corollary C.5 is important conceptually: in exact stochastic sequence modeling over discrete alphabets, continuous state spaces serve as organizing geometries for update rules and readout parameterizations, not as sources of uncountably many exact reachable predictive states.

C.3 MoE, hybrids, and architecture comparison

The routing and hybrid constructions of Section 6 also admit a direct measure-theoretic reading.

Proposition C.6 (MoE and hybrids in **Stoch**). *Let E be a finite standard Borel space. For every deterministic state action (Q, a_Q) , gate kernel $g : Q \otimes X \rightarrow E$, and expert family $k : E \otimes Q \otimes X \rightarrow Y$, the induced MoE readout is the ordinary mixture*

$$\kappa_{\text{MoE}}(dy \mid q, x) = \sum_{e \in E} g(e \mid q, x) k(dy \mid e, q, x).$$

Hence routing changes only the readout kernel and leaves the hidden-state action unchanged. Consequently:

1. *exact MoE realizability of f over a backbone geometry $\mathcal{G}$ is equivalent to $\Sigma_f \in \mathcal{G}$;*
2. *exact hybrid realizability in the product/image closure $\langle \mathcal{G}_1, \mathcal{G}_2 \rangle_{\text{prod,img}}$ is equivalent to $\Sigma_f \in \langle \mathcal{G}_1, \mathcal{G}_2 \rangle_{\text{prod,img}}$;*
3. *under the discrete-history hypothesis, every exact MoE or hybrid realization may be chosen with countable standard Borel backbone state space.*

Proof. The mixture formula is exactly Proposition 6.6 specialized to **Stoch**. The geometric criteria are Theorem 5.3 and Proposition 6.8. The countability statement follows from Theorem C.1. $\square$

Proposition C.6 shows that the abstract architectural comparison developed in the main text becomes completely concrete for Markov kernels: one compares models by asking which measurable quotients of the countable history object they can represent exactly. MoE modifies only stochastic readout on top of a chosen measurable backbone, while hybrid architectures correspond to measurable product combinations followed by exact quotienting of redundant joint state.

Corollary C.7 (Major consequences for stochastic sequence modeling). *In the Markov-kernel setting with discrete token alphabet X , the general framework of this paper yields the following concrete consequences.*

1. *Every exact stochastic sequence behavior admits a minimal exact predictive state space Σ_f , and this state space is countable.*
2. *Exact memory compression is equivalent to measurable predictive sufficiency of a history statistic $q : X^* \rightarrow Q$ whose fibers are stable under one-step extension, equivalently to such a statistic carrying a deterministic measurable update.*
3. *Exact linear-attention-type models are exactly those whose predictive state is a measurable quotient of an additive memory action, and exact SSM/Mamba-type models are exactly those whose predictive state is a measurable quotient of an affine recursion state. In each case, the full predictive kernel factors through the corresponding initialized statistic.*
4. *MoE is orthogonal to memory geometry, and hybrid exactness is governed by product/image closure of measurable state geometries.*

*Therefore, in **Stoch**, architecture comparison reduces to the measurable representability problem for the countable predictive quotient Σ_f .*

Proof. Combine Theorem C.1, Corollary C.2, Theorems C.3 and C.4, and Proposition C.6. $\square$

Remark C.8. The discrete-history hypothesis is used only to ensure that all reachable exact quotients of free history remain standard Borel automatically. If one works with a non-discrete standard Borel input space X , the same formulas remain valid whenever the relevant measurable images and quotients of X^* exist in the chosen measurable subcategory.

D Architectural and Conceptual Implications

We now extract the architectural meaning of the theory. The preceding sections imply that the right primitive object is not an architecture label, but the syntactic predictive-state object $\Sigma_f = \text{Im}(r_f)$. The discussion below makes this precise.

D.1 What is actually being compressed?

The exact realization theorem shows that every exact architecture compresses histories only through their predictive content.

Proposition D.1. *Let $q : X^* \rightarrow Q$ be the reachability map of an exact realization of f . Then there exists a unique regular epimorphism $\pi_Q : Q \rightarrow \Sigma_f$ such that $e_f = \pi_Q \circ q$. In particular, $q(w) = q(w') \implies r_f(w) = r_f(w')$.*

Proof. This is exactly Corollary 4.5 specialized to the reachability map of the realization. $\square$

Proposition D.1 has a simple interpretation:

an exact model does not compress the past as a raw token sequence; it compresses the past only up to future predictive equivalence.

The free-history realization stores the entire past, the syntactic realization stores only the irreducible predictive content, and all other exact models lie between these two extremes.

Thus the phrase “memory compression” should be read literally as history $\rightsquigarrow$ predictive state, not as history $\rightsquigarrow$ arbitrary latent code. The latent code is exact only insofar as it refines the syntactic predictive quotient.

D.2 Why MoE is orthogonal to backbone choice

The above examples already suggest that routing and memory are orthogonal design axes. This becomes immediate from the categorical definitions.

Proposition D.2. *Let $\mathcal{G}$ be any memory geometry, let $Q \in \mathcal{G}$, and let $a_Q : Q \otimes X^* \rightarrow Q$ be its deterministic state action. For every gate $g : Q \otimes X \rightarrow E$ and every expert family $k : E \otimes Q \otimes X \rightarrow Y$, the MoE readout $\kappa_{\text{MoE}} = k \circ (g \otimes \text{id}_{Q \otimes X}) \circ \Delta_{Q \otimes X}$ produces a state-deterministic stochastic Mealy object with the same state action (Q, a_Q) . Hence routing modifies readout semantics but does not change the underlying backbone geometry.*

Proof. The hidden-state action remains a_Q . Only the readout channel is replaced. $\square$

If each expert has its own state geometry, then the combined backbone is their product, as in Proposition B.6. Again, the role of routing is to combine or select outputs, not to create a new primitive memory type.

This explains why MoE can be superimposed on Transformer blocks, SSM blocks, linear-attention blocks, or hybrids of these: it is not a competing memory geometry, but a constructor acting on the readout side of an existing geometry.

D.3 Design principles for hybrid architectures

The examples above suggest a practical reading of the theory. A hybrid architecture should be introduced precisely when the predictive state decomposes into several qualitatively different components.

Proposition D.3. *Suppose there exist exact history statistics $q_i : X^* \rightarrow Q_i$ for $i = 1, \dots, k$ such that each Q_i belongs to a memory geometry $\mathcal{G}_i$, and $r_f = m \circ \langle q_1, \dots, q_k \rangle$ for some morphism $m : Q_1 \times \dots \times Q_k \rightarrow \Omega$. Then f is exactly realizable in the product/image closure of the geometries $\mathcal{G}_1, \dots, \mathcal{G}_k$.*

Proof. Let $\gamma := \langle q_1, \dots, q_k \rangle : X^* \rightarrow Q_1 \times \dots \times Q_k$ and factor it in $\mathbf{Act}_{X^*}(\mathcal{C}_{\det})$ as $X^* \xrightarrow{e} Q_{\text{im}} \xrightarrow{j} Q_1 \times \dots \times Q_k$, where e is a regular epimorphism and j is monic. Since each Q_i belongs to $\mathcal{G}_i$, the product $Q_1 \times \dots \times Q_k$ lies in the product/image closure of the $\mathcal{G}_i$, and by image-closure so does Q_{im} . Because $r_f = m \circ \gamma = m \circ j \circ e$, Proposition 5.4 shows that e is an exact quotient of free history. Therefore the induced realization on Q_{im} is exact and belongs to the product/image closure of the $\mathcal{G}_i$. $\square$

Proposition D.3 leads to four concrete design principles.

- **Use free-history components for order-sensitive residuals.** If the residual map remains close to injective on long prefixes, then explicit history storage or history-like attention is unavoidable.
- **Use additive components for permutation-invariant accumulation.** If predictive behavior depends only on commutative sufficient statistics, then linear-attention-style additive memory is the correct exact geometry.

- **Use affine components for driven dynamical summaries.** If predictive behavior evolves by low-dimensional controlled linear dynamics, then SSM/Mamba-style affine states are the natural exact representation.
- **Use products when these mechanisms coexist.** If distinct predictive factors coexist, the correct abstraction is not a single compromise geometry but a product followed by exact quotienting.

The hybrid example of Section B is the simplest manifestation of this principle: the parity factor is additive, the discounted-sum factor is affine, and the gate only decides which factor is read at each step.

D.4 Implications for long-context and efficient inference

The exact theory identifies long-context difficulty with the growth of exact predictive memory, not with raw sequence length by itself. If the syntactic predictive-state object Σ_f is representable inside an additive or affine geometry by a context-independent state presentation, then inference memory can remain independent of context length. This is the semantic reason why linear-attention-style and SSM-style backbones can support long contexts when the predictive quotient is genuinely compressible. Conversely, if exact predictive distinctions continue to accumulate with longer histories, then no bounded exact recurrent state can suffice. In that regime one must either retain explicit history, as in free-history or attention-like memories, or use a hybrid architecture that isolates the compressible and non-compressible parts of the predictive state. Thus the semantic question behind efficient long-context inference is:

Does free history admit an exact predictive quotient that is representable in the geometry of interest?

If yes, compressed-state architectures are appropriate; if not, explicit history or hybrid memory becomes unavoidable. The theory suggests the following architecture-level dichotomy: behaviors with small syntactic predictive state favor compressed recurrent geometries, whereas behaviors whose predictive state grows with context length favor attention-like or hybrid geometries. What matters is not the nominal context window, but the exact predictive quotient that the task induces.