A SECOND-ORDER COMPLEXITY ANALYSIS OF ATTENTION AND STATE-SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard analyses of sequence models characterize efficiency primarily in terms of sequence length, such as quadratic-time attention or linear-time state-space models. We show that sequence-length complexity alone does not determine the resources needed to evaluate a model at a prescribed accuracy. We develop a second-order complexity analysis of attention and selective state-space operators, explicitly accounting for approximation accuracy, input regularity, representation, and effective memory. For representative attention and selective state-space operators, we derive matching deterministic sampling bounds that distinguish the cost of resolving a signal from the amount of history required. Even a stable state-space operator requires exponentially many point queries in the requested accuracy bits on bounded Lipschitz inputs, while quantitative smoothness and stability certificates enable polynomial-time evaluation under fixed resource bounds. We introduce precision locality: the shortest recent context sufficient to recover the full-history output without a cached summary of earlier inputs. On arbitrary bounded token sequences, certified contractive selective scans admit context lengths logarithmic in inverse error at fixed contraction and magnitude bounds, whereas uniform attention requires nearly the entire sequence at high accuracy. We construct evaluators with certified context lengths and numerical precision budgets, including explicit conditions for Mamba's selective recurrence. We further prove that approximating persistent statistics over growing horizons requires weaker contraction, larger state magnitudes, or stronger amplification. These results connect precision-dependent evaluation costs with long-range approximation limits and provide a mathematical basis for analyzing the balance between forgetting and persistent memory.

1 INTRODUCTION

Long-context learning requires models to use information spread across a sequence while keeping computation manageable. Which information must survive depends on the learning target: recent observations may suffice for some predictions, whereas cumulative statistics retain dependence on distant inputs. Memory therefore connects computational efficiency with approximation capability. Forgetting can reduce the history needed to reproduce a model's output, while also limiting the persistent dependencies that the model can represent. We use second-order complexity theory (Kawamura & Cook, 2012; Kawamura et al., 2012; Neumann & Steinberg, 2017) to analyze evaluation costs, accounting explicitly for requested accuracy and input representation.

Attention provides direct access to earlier representations, while selective state-space models (SSMs) such as Mamba propagate an input-dependent summary through a recurrence (Vaswani et al., 2017; Gu & Dao, 2023). Their usual quadratic-versus-linear comparison measures the work of processing all positions in a supplied sequence. We ask how much recent context suffices to recover one full-history output of a fixed model to a prescribed accuracy, when no cached state or summary of earlier inputs is available. This setting includes evaluating an isolated window and restarting a recurrent block whose earlier state is unavailable. Building on classical approximately-finite-memory theory (Sandberg, 1991; 1992; Hanson & Raginsky, 2019), we study *precision locality*: the shortest recent context from which the full-history output can be recovered within error ε for every allowed history.

Let us consider a simple example. For $10{,}000$ values in $[-1, 1]$, recovering their uniform average to absolute error 0.01 from a suffix alone requires at least $9{,}900$ values. The normalized scan

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

$h_i = 0.9h_{i-1} + 0.1v_i$, with $h_0 = 0$ and $|v_i| \le 1$, needs only the last 44: resetting its incoming state to zero introduces error at most $0.9^{44} < 0.01$ (Corollary 6.1). Yet histories with prefixes of all $+1$ or all -1 and a common 44-token zero suffix have global means close to $+1$ and -1, while both scan outputs lie within 0.01 of zero. This example makes the tradeoff explicit: contraction enables short-context recovery of the scan, while approximating the global mean requires retaining information from the distant prefix.

The example isolates how much history must be retained; evaluation also depends on how densely and accurately that history must be sampled. This second question arises explicitly in scientific machine learning, where neural operators act on functions and continuum attention and state-space neural operators extend sequence architectures to that setting (Kovachki et al., 2023; Calvello et al., 2025; Song & Jiang, 2026; Koren & Lanthaler, 2026). Without a prescribed token grid, the evaluator chooses sampling locations and the accuracy of each value. We write the requested output error as 2^{-n}, so each additional accuracy bit halves the error. Even on a fixed interval, a stable filter can require exponentially many deterministic point queries in n on bounded Lipschitz inputs (Proposition 3.3). Thus memory decay controls how far into the past evaluation must reach, while quantitative smoothness controls the work needed to resolve that history. We account for input representation, sampling, and arithmetic through the Kawamura-Cook framework for second-order complexity (Kawamura & Cook, 2012; Kawamura et al., 2012; Neumann & Steinberg, 2017). Our discrete context guarantees apply to arbitrary bounded tokens and require no smoothness assumption.

This perspective connects dependence on past inputs, as studied in fading-memory and echo-state theory (Boyd & Chua, 1985; Bainier et al., 2026; Ortega & Rossmannek, 2026), to explicit evaluation resources. Attention-recurrence connections (Katharopoulos et al., 2020; Dao & Gu, 2024; Zimerman et al., 2025; Wang et al., 2026) and finite-precision expressivity results (Merrill et al., 2024; Chen et al., 2024) provide complementary accounts of sequence computation. Contractive recurrent models already admit logarithmic truncation windows (Miller & Hardt, 2019, Proposition 3). We build on this principle with sharp sampling laws, second-order bit-cost guarantees, and finite-precision context certificates for selective scans, including Mamba-3. We quantify the connection between context recovery and persistent memory through four results:

1. **Context and numerical precision for a chosen error.** We construct evaluators that select sufficient context and arithmetic precision from model certificates: supplied quantitative bounds on memory decay, magnitudes, and, for continuous inputs, regularity. Under the stated assumptions, our general theorem establishes second-order polynomial-time evaluation for selective dynamics (Theorem 4.2). For arbitrary bounded tokens, contractive selective scans need context logarithmic in inverse error at fixed model bounds.

2. **Certified context for Mamba and Mamba-3.** For Mamba's real negative-diagonal recurrence, accumulated contraction and supplied state/readout bounds give an adaptive rule: read backward until the omitted history fits its error budget. The guarantee covers both discarded history and numerical error, connecting receptive-field analysis (Ben-Kish et al., 2025) to certified context selection. We extend these context guarantees to Mamba-3 blocks, including rotations, two-token forcing, and multiple-input/multiple-output readouts, and establish logarithmic context bounds at fixed contraction and magnitude certificates (Corollary 6.2).

3. **Provably optimal sampling costs.** We prove matching upper and lower bounds on deterministic sampling for representative uniform and nonuniform attention operators and stable and input-selective SSMs (Theorem 3.1). Under the fixed smoothness bounds defined below, these attention operators incur an unavoidable sampling cost proportional to the full history, while the SSM cost remains bounded as history grows at fixed accuracy. The bounds also determine the optimal dependence on accuracy: adaptive sampling and exact input values cannot improve these rates. Our finite-precision algorithms attain the optimal sampling rates and establish second-order polynomial-time evaluation for these SSMs.

4. **Memory resources needed for persistent targets.** We show which memory resources must change for the studied contractive selective SSMs to approximate the full-sequence mean (Theorem 5.1). Even on a fixed family of smooth inputs, maintaining a sufficiently small worst-case error as histories grow requires slower forgetting or larger bounds on state magnitude, readout gain, or transient amplification. With these magnitude and amplification bounds fixed, the decay rate must satisfy $\gamma = O(1/T)$, regardless of state dimension. This

connects long-memory approximation (Wang & Li, 2024) to evaluation cost through the same decay and magnitude parameters.

These results link context selection and persistent-target approximation through the same memory parameters. Precision locality specifies the context needed to preserve a model's output, while the approximation bound identifies when persistent targets force its memory resources to change. The organizing properties are memory decay and the information available to the evaluator. Recurrent summaries can preserve global statistics, and attention with recency decay can admit short contexts. Our results provide a quantitative basis for assessing how a model's evaluation savings relate to the dependence on history required by a learning problem.

2 Precision Locality and the Evaluation Model

Precision locality measures information sufficiency, while query and bit complexity measure computational work. Vector norms are Euclidean, matrix norms are induced operator norms, and $\|f\|_\infty = \sup_t \|f(t)\|$ over the stated interval. Fix an ambient class $\mathcal{B}_T$ of bounded histories and write $R_H x = x|_{[T-H,T]}$ for $0 \le H \le T$. For a discrete sequence, R_H instead retains its last H tokens, and T is replaced by N.

Definition 2.1 (Precision locality)**.** The suffix recovery error and precision radius of an endpoint map F are

$$E_{\mathsf{F}}^{\mathcal{B}}(T,H) = \inf_{G} \sup_{x\in\mathcal{B}_T} \|(\mathsf{F}x)(T) - G(R_H x)\|, \tag{1}$$

$$H_{\mathsf{F}}^{\mathcal{B}}(T,n) = \inf\{H : E_{\mathsf{F}}^{\mathcal{B}}(T,H) \le 2^{-n}\}. \tag{2}$$

Here the input-dependent information available to G consists exactly of $R_H x$; the model, prescribed initial condition, and endpoint are known. Discrete radii use the minimum over integer $0 \le H \le N$. A family is *uniformly precision-local* if one polynomial P satisfies $H_{\mathsf{F}}^{\mathcal{B}}(T,n) \le P(n,\mathbf{c})$ for every member, horizon, and accuracy index, where $\mathbf{c}$ lists the supplied certificate resources, including inverse contraction margins when applicable.

Precision locality links information sufficiency to certified computation: the infimum over G allows any recovery map using the retained suffix, and our constructive theorems provide finite-precision evaluators with explicit sampling and arithmetic budgets at certified sufficient window lengths. Classical approximately-finite-memory theory compares a system's output on the full input with its output on an input restricted to a recent window, uniformly over admissible inputs and evaluation times (Sandberg, 1991; 1992); Hanson & Raginsky (2019, Definition 2.1) define the smallest window sufficient for a prescribed error. Our formulation explicitly records the horizon dependence of the radius. The parameter n specifies the requested output error.

For the continuous locality examples, we use the ambient class $\mathcal{B}_T = \{x \in C([0,T]) : \|x\|_\infty \le 1\}$. Smoothness bounds enter separately when we study finite sampling. This distinction matters for analytic inputs: knowing their exact values throughout any suffix interval of positive length uniquely determines the full input. Finitely many point observations do not supply that information. Our query bounds concern this finite-observation setting (Appendix A.3).

On this continuous unit ball, the uniform average and stable filter defined below have

$$E_{\mathsf{U}}(T,H) = \frac{T-H}{T}, \qquad E_{\mathsf{S}_0}(T,H) = e^{-H} - e^{-T}. \tag{3}$$

Thus the former has radius $(1-2^{-n})T$, while the latter admits a radius of at most $\min\{T, n\log 2\}$. The proof is in Appendix C.3. Corollary 6.1 gives their finite-sequence counterparts on arbitrary bounded tokens.

2.1 Representations and Quantitative Regularity

Point access and query cost. Let $T \in \mathbb{N}_{\ge 1}$ be binary encoded, and write $\mathrm{len}(T)$ for its encoding length. A point oracle lets the evaluator request an input value at a chosen time and accuracy. Write $\mathbb{D}$ for the dyadic rationals, whose denominators are powers of two. A query at $t \in \mathbb{D} \cap [0,T]$ with precision $p \in \mathbb{N}$ returns a dyadic number satisfying

$$|\psi_x(t,p) - x(t)| \le 2^{-p}, \qquad t \in \mathbb{D}\cap[0,T], \quad p \in \mathbb{N}. \tag{4}$$

For an operator F, let $Q_{\mathsf{F}}(T, n)$ be the smallest worst-case number of point queries needed to return $(\mathsf{F}x)(T)$ within 2^{-n}, uniformly over the stated input class and all valid oracle answers. Algorithms may adapt both query locations and precisions. Query complexity ignores other computation; our upper bounds separately control its bit cost. For Q^{ran}, success probability is at least $3/4$ for every input and valid oracle, and the query budget is worst-case. Lower bounds allow even *exact* real point answers.

Quantitative smoothness. Derivative bounds control how much the signal can vary between sampled points. For fixed rational $s \geq 1$ and $C, R \geq 1$, define

$$\mathcal{G}_s(C, R; T) = \left\{ x \in C^\infty([0, T]) : \left\| x^{(k)} \right\|_\infty \leq C R^k (k!)^s \text{ for every } k \geq 0 \right\}. \tag{5}$$

Here C bounds the signal magnitude, R sets the derivative scale, and s controls the permitted growth of higher derivatives. At fixed C, R, increasing s allows larger higher derivatives; $s = 1$ gives a quantitatively bounded analytic class. The order s is fixed throughout, and constants and polynomial degrees may depend on it. The input representation consists exactly of the point evaluator (4) and the promised bounds. These bounds also supply a common continuity bound, so no additional input oracle is needed; Appendix A gives the formal representation and time-rescaling conventions.

Bit cost. A *function name* consists of a point evaluator and a modulus: a quantitative continuity bound specifying how close input locations must be to guarantee a prescribed difference in function values. Bit complexity additionally charges query writing, answer reading, and arithmetic. We use the Kawamura-Cook (KC) model (Kawamura & Cook, 2012), which accounts for the lengths of oracle answers through a size function $\ell_x(k)$: the maximum answer length for requests of at most k bits. KC polynomial time bounds the running time by a second-order polynomial, formed from ordinary polynomials in finite input lengths and applications of ℓ_x. The bounded classes above admit encodings with $\ell_x(k) = O(k + 1)$ uniformly in T; substituting this bound yields an ordinary polynomial in the remaining resources. Appendix A specifies the length-monotone encodings and formal bit-complexity conventions.

For finite attention matrices, Alman & Song (2023) identify entry-magnitude regimes governing fast additive approximation and conditional hardness. Fixed-precision transformer theory characterizes recognizer expressivity and the complementary roles of global and local attention under specified architectural assumptions (Li & Cotterell, 2025; 2026). Our point-query model instead measures the input observations needed to approximate one endpoint output of a fixed operator. Our variable-precision bit algorithms explicitly budget numerical error and evaluation cost; mixed-precision analysis addresses conditioning and rounding-error propagation (Budzinskiy et al., 2025).

Attention and state-space operators for the sharp comparison. For scalar $x \in \mathcal{G}_s(1, R; T)$, we fix four operators:

$$(\mathsf{U}x)(t) = \frac{1}{t} \int_0^t x(u)\, du, \qquad (\mathsf{S}_0 x)(t) = \int_0^t e^{-(t-u)} x(u)\, du, \tag{6}$$

$$(\mathsf{A}_* x)(t) = \frac{\int_0^t e^{z(t)z(u)} x(u)\, du}{\int_0^t e^{z(t)z(u)}\, du}, \qquad z(t) = 1 + \tfrac{1}{2} \sin t, \tag{7}$$

$$(\mathsf{S}_{\mathrm{sel}} x)(t) = h(t), \qquad h'(t) = -(1 + x(t)^2) h(t) + x(t), \quad h(0) = 0. \tag{8}$$

Set $(\mathsf{U}x)(0) = (\mathsf{A}_* x)(0) = x(0)$. The SSM S_0 satisfies $h' = -h + x$, $h(0) = 0$. The operator A_* is a softmax-attention head on the two-channel signal (z, x) with linear query/key projections selecting z and value projection selecting x. The known channel z is fixed independently of x. Its logits are bounded, nonconstant functions of the past, and query dependent. The decay rate in $\mathsf{S}_{\mathrm{sel}}$ depends on the varying input and lies in $[1, 2]$.

3 Sharp query laws: history, precision, and randomization

These four operators isolate persistent aggregation and decaying memory, including nonuniform attention weights and an input-dependent decay rate. Their sharp query laws quantify how the history requirement combines with the cost of resolving a smooth signal. Section 4 develops the general

Table 1: Sharp point-query costs at error 2^{-n} on fixed quantitative regularity classes. The four fixed operators exhibit the roles of history, regularity, and randomization. The horizon is binary encoded; the randomized row uses a sufficiently large fixed R_s with $s > 1$.

Input class and regime	Operator	Query complexity
Gevrey-s, $s \geq 1$	U, nonuniform A_*	$\Theta(Tn^s)$
Gevrey-s, $s \geq 1$	S_0, selective $\mathsf{S}_{\mathrm{sel}}$	$\Theta(n^s \min\{T, n\})$
Gevrey-s, $s > 1$, $T \geq 4^n$	U, randomized	$\Theta(4^n)$
Lipschitz, $T = 1$	U, S_0, deterministic	$\Theta(2^n)$

certified-evaluation guarantees for parameterized attention and selective dynamics, and Section 6 gives context-selection guarantees for arbitrary bounded token sequences.

Theorem 3.1 (Deterministic separation)**.** *Fix rational $s \geq 1$ and $R \geq 1$. On $\mathcal{G}_s(1, R; T)$, uniformly for integers $T \geq 1$ and sufficiently large n,*

$$Q_{\mathsf{U}}(T, n),\ Q_{\mathsf{A}_*}(T, n) = \Theta_{s,R}(Tn^s), \tag{9}$$

$$Q_{\mathsf{S}_0}(T, n),\ Q_{\mathsf{S}_{\mathrm{sel}}}(T, n) = \Theta_{s,R}(n^s \min\{T, n\})\,. \tag{10}$$

The lower bounds hold with exact point evaluations. The upper bounds use $O_{s,R}(n)$ accuracy bits per value query. For a second-order polynomial $P_{s,R}$ and $k = n + \mathrm{len}(T)$, the SSM evaluations take time at most $P_{s,R}(\ell_x, k)$, while the attention evaluations take at most $TP_{s,R}(\ell_x, k)$.

The upper bounds distinguish the history needed from the sampling density within it. Quantitative Gevrey regularity permits high-order local interpolation using $O_{s,R}(n^s)$ samples per unit interval at error 2^{-n} (Appendix B.1). The attention operators U, A_* retain the full interval, whereas exponential decay lets $\mathsf{S}_0, \mathsf{S}_{\mathrm{sel}}$ use a window of length $O(n)$, capped by T. Normalization and stability control interpolation error in the four finite-precision evaluators (Appendix B.2.1).

The matching lower bounds use smooth functions that vanish at every sampled point while retaining enough integral mass to change the output. Such functions make opposite inputs indistinguishable to an evaluator. A periodic trigonometric construction respects the derivative bounds even for analytic inputs, so the lower bounds do not rely on compact support (Appendix B.2.2).

Corollary 3.2 (Binary-horizon feasibility)**.** *With fixed s, R, both S_0 and $\mathsf{S}_{\mathrm{sel}}$ are KC polynomial-time computable on this represented family of intervals. The query costs of U and A_* grow exponentially in the binary horizon length at fixed sufficiently large precision, and both admit precision-polynomial evaluation on every fixed horizon.*

The distinction is with respect to the encoded horizon: a cost linear in T can be exponential in $\mathrm{len}(T)$, while a polynomial in the retained window and endpoint encoding remains feasible. The proof extends endpoint evaluation to function-valued outputs (Appendix B.2.3).

Evaluation cost depends jointly on retained history, signal regularity, requested accuracy, and the sampling strategy.

Proposition 3.3 (Sharp precision costs on Lipschitz inputs)**.** *On $\{x : \|x\|_\infty \leq 1,\ \mathrm{Lip}(x) \leq 1\}$ over $[0, 1]$, deterministic evaluation of either U or S_0 needs $\Theta(2^n)$ point queries. On general continuous-function names, integration reduces in KC polynomial time to either endpoint evaluation.*

A uniform-grid algorithm attains the bound, highlighting the computational value of the quantitative regularity in Theorem 3.1. The proof, including the integration reductions, is in Appendix B.2.4.

Theorem 3.4 (Randomized long-horizon separation)**.** *For every fixed rational $s > 1$, there is a fixed $R_s \geq 1$ such that, on $\mathcal{G}_s(1, R_s; T)$ and for sufficiently large n,*

$$T \geq 4^n \quad \Longrightarrow \quad Q_{\mathsf{U}}^{\mathrm{ran}}(T, n) = \Theta_s(4^n). \tag{11}$$

On the same input class, S_0 and $\mathsf{S}_{\mathrm{sel}}$ retain deterministic polynomial-time evaluation in $n + \mathrm{len}(T)$.

The upper bound uses bounded-variance sampling. For the lower bound, the Gevrey regime $s > 1$ supplies smooth, compactly supported bumps with disjoint supports, each encoding an independent sign. Each observation accesses at most one sign, so accurate averaging requires distinguishing

two small opposite sign biases. This construction complements the periodic argument establishing deterministic optimality throughout $s \geq 1$. With $\varepsilon = 2^{-n}$ and $T \geq \varepsilon^{-2}$, randomization yields the sharp uniform-average query cost $\Theta(\varepsilon^{-2})$, removing linear horizon dependence with optimal inverse-square dependence on the target error. The proof and finite-bit sampling algorithm are in Appendix B.3.

4 Certified precision-local evaluation

A *certificate* supplies quantitative bounds on regularity, contraction, and magnitudes over the allowed input domain; the following results use these bounds to determine context windows and working precision for evaluating parameterized attention and selective dynamics. For learned models, including the certificates' numerical values and coefficient-evaluation costs in the resource bounds makes dependence on model parameters explicit.

4.1 Attention: regularity and normalization

Let $x : [0, T] \to \mathbb{R}^d$, $q_\theta, k_\theta : \mathbb{R}^d \to \mathbb{R}^{d_k}$, and $v_\theta : \mathbb{R}^d \to \mathbb{R}^m$. Write $\operatorname{size}\theta$ for the total bit length of the supplied parameters, evaluator programs, and certificates. Set $\lambda_x(t, u) = \langle q_\theta(x(t)), k_\theta(x(u))\rangle$ and $v_x(u) = v_\theta(x(u))$. Define

$$(\mathsf{A}_\theta x)(t) = \frac{\int_0^t e^{\lambda_x(t,u)} v_x(u)\, du}{\int_0^t e^{\lambda_x(t,u)}\, du}, \quad (\mathsf{A}_\theta x)(0) = v_\theta(x(0)). \tag{12}$$

Proposition 4.1 (Regularity-conditioned attention). *Suppose $|\lambda_x| \leq B$, $\|v_x\| \leq V$, and, uniformly in t, both integrands in (12), as functions of u, have Gevrey-s bounds (M, R). Suppose their point evaluations admit polynomial-time realizers (oracle algorithms returning values to any requested accuracy). Then endpoint evaluation has second-order polynomial cost in*

$$n + \operatorname{len}(T) + T + B + \log M + R + \log(1 + V) + \operatorname{size}\theta + d + m,$$

in addition to the stipulated evaluation resources. If $|\partial_1\lambda_x|, |\partial_2\lambda_x| \leq L_\lambda$ and $\|v_x'\| \leq L_v$, the output has Lipschitz constant at most $L_v + 4VL_\lambda$.

The logit bound keeps the normalized denominator away from zero, while regularity controls the samples needed to approximate numerator and denominator. The proof, including the output-modulus and endpoint claims, is in Appendix C.1.

4.2 Selective dynamics: a quantitative stability certificate

Consider state dimension r, input dimension d, output dimension m, and

$$h' = A_x(t)h + b_x(t), \quad y = C_x(t)h, \quad \|h(0)\| \leq M, \tag{13}$$

where the coefficients are pointwise functions of $x(t)$ and θ. The initial state is a supplied rational vector; $\operatorname{size}\theta$ includes its encoding and the finite descriptions of the coefficient realizers and certificates. Assume $M, R \geq 1$ and

$$\left\|A_x^{(k)}\right\|_\infty, \left\|b_x^{(k)}\right\|_\infty, \left\|C_x^{(k)}\right\|_\infty \leq MR^k(k!)^s \quad (k \geq 0). \tag{14}$$

The transition matrix satisfies $\partial_t\Phi_x(t, u) = A_x(t)\Phi_x(t, u)$ and $\Phi_x(u, u) = I$; it propagates state differences under the same input. The stability assumption is

$$\|\Phi_x(t, u)\| \leq Ke^{-\gamma(t-u)}, \qquad t \geq u, \quad K \geq 1,\ \gamma > 0, \tag{15}$$

uniformly over inputs. Here γ is the decay rate and K allows transient amplification relative to the exponential envelope. The condition $(A_x + A_x^\top)/2 \preceq -\gamma I$ supplies it with $K = 1$.

Theorem 4.2 (Precision-local, second-order feasible selective dynamics). *Assume (14)-(15) and local coefficient point realizers sharing a fixed second-order polynomial time bound in precision, time encoding length, $\operatorname{size}\theta$, and input-name size, uniformly over parameters and supplied programs. Put*

$D = KM(1+\gamma^{-1})$ and $\rho = MKD$. Here D bounds the state magnitude and ρ is the output-tail prefactor. An output at time t can be approximated within 2^{-n} using only the suffix of length

$$H_n = \min\left\{t, \frac{(n+2)\log 2 + \log\rho}{\gamma}\right\}, \tag{16}$$

or a computably rounded-up window, together with the initial state when the window reaches zero. Its bit cost is bounded by a second-order polynomial in

$$n + \mathrm{len}(t) + \mathrm{size}\,\theta + d + r + m + M + R + \gamma^{-1} + \log K, \tag{17}$$

including coefficient-realizer resources. Here M, R, γ^{-1} enter through their numerical values. The output has Lipschitz constant at most $MRD + M(MD + M)$.

The certificate bounds the discarded-prefix contribution by $\rho e^{-\gamma H}$ after a reset at suffix length $H < t$. This exponential tail converts the requested accuracy into a context window; the numerical evaluator then divides the remaining error budget among coefficient approximation, integration, and rounding. The proof and error budgets are in Appendix C.2.

For fixed certificate values, the bound is polynomial in $n + \mathrm{len}(T)$, enabling evaluation over horizons exponentially larger than their encodings. An input-dependent negative diagonal transition with a uniform margin supplies a simple architectural realization. Fixed-depth compositions also admit feasible evaluation when intermediate regularity, magnitude, and stability certificates are supplied, extending the construction to certified stacks.

4.3 The decay margin determines evaluation cost

A matching query law quantifies how the numerical stability margin sets the retained history and evaluation cost. Consider the normalized exponential filter

$$(\mathsf{S}_\gamma x)(T) = \int_0^T \gamma e^{-\gamma(T-u)} x(u)\,du, \qquad h' = -\gamma h + \gamma x, \quad h(0) = 0. \tag{18}$$

Normalization keeps the state uniformly bounded by one even as $\gamma \downarrow 0$.

Proposition 4.3 (Sharp decay-margin cost)**.** *Fix $s \geq 1$ rational and $R \geq 1$. Uniformly for rational $0 < \gamma \leq 1$, sufficiently large n, and integer $T \geq n/\gamma$, evaluation on $\mathcal{G}_s(1, R; T)$ satisfies*

$$Q_{\mathsf{S}_\gamma}(T, n) = \Theta_{s,R}(n^{s+1}/\gamma). \tag{19}$$

The matching bounds combine a decay-controlled window with the interpolation and periodic constructions (Appendix B.2.5). For $\gamma = 2^{-\kappa}$, the sharp cost becomes $\Theta_{s,R}(n^{s+1}2^{\kappa})$ while the state magnitude remains bounded by one. This exact dependence makes the stability margin a quantitative design parameter: stronger contraction reduces the retained history, and smaller margins support longer memory with a corresponding acquisition cost.

5 Precision-local memory and approximation resources

Context selection preserves the output of a given model. Learning also requires that this output approximate the desired target, which may depend on distant inputs. Uniform averaging supplies a simple such target: changing the first half of a sequence can change its global mean by a constant even when the recent half is unchanged. Approximation theory for layered SSMs identifies exponential memory decay (Wang & Xue, 2023), and StableSSM studies how parameterization affects stable approximation of long-memory targets (Wang & Li, 2024). The following theorem connects this issue to precision locality by quantifying the contraction and state/readout resources required to approximate averaging inside a fixed Gevrey-s ball with $s > 1$.

Theorem 5.1 (Stability-memory approximation tradeoff)**.** *Fix rational $s > 1$. There are constants $R_s \geq 1$ and $\beta_s > 0$ such that the following holds for every even integer T. Let F_θ be a scalar-output selective SSM of the form (13), with pointwise input-dependent coefficients, $\|h_x(t)\| \leq D$, $\|C_x(t)\| \leq M$, and (15) on $\mathcal{G}_s(1, R_s; T)$. Then*

$$\sup_{x \in \mathcal{G}_s(1,R_s;T)} |(\mathsf{F}_\theta x)(T) - (\mathsf{U}x)(T)| \geq \frac{\beta_s}{2} - MKDe^{-\gamma T/2}. \tag{20}$$

In particular, error at most $\beta_s/4$ requires $\gamma T \leq 2\log(4MKD/\beta_s)$.

The lower bound follows from histories with identical recent suffixes but different global averages. A contractive model suppresses the effect of their different prefixes, even though the target retains it. The quantitative proof is in Appendix C.3.2. The theorem permits increasing state dimension; what must change is the decay margin or the state/readout/transient-amplification budget. For fixed γ, maintaining uniform approximation error at most $\beta_s/4$ requires MKD to grow exponentially with T. Its logarithm then reappears in the precision-local window. Alternatively, reducing γ enlarges the retained horizon. Recent recall scaling laws relate Mamba's embedding and state dimensions to associative-recall capacity (Koren et al., 2026); the bound here concerns worst-case uniform-average approximation under supplied decay and magnitude certificates.

Memory structure across architectures. An integrator $h' = x$ followed by $h(t)/t$ implements U with one real state, showing how an accumulated state supports persistent memory. HiPPO (high-order polynomial projection operators) gives a broader construction of recurrent summaries through online polynomial projection (Gu et al., 2020). Attention with an exponential recency factor can share an SSM's precision locality; Forgetting Transformer introduces data-dependent forgetting into softmax attention (Lin et al., 2025). Gated Delta Networks combine memory erasure gates with targeted delta-rule updates (Yang et al., 2025). Prefix summaries likewise provide a route to reusing earlier computation. These examples identify decay structure and the available information as the organizing principles of precision-aware evaluation across architectures.

For hybrid architectures that combine fading and persistent memory (Zancato et al., 2024), our results motivate distinguishing their context requirements: certified contractive components admit context windows logarithmic in inverse error at fixed resource bounds, whereas persistent aggregation can require histories that grow with the sequence length.

6 Discrete corollary: softmax heads and selective scans

For token-based models, the same reasoning gives a context guarantee without a continuous-signal or smoothness assumption. Inputs are arbitrary bounded vectors $u_1, \dots, u_N$ with random access. The evaluator reconstructs one block output from retained tokens and known parameters. A selective scan uses a reset recurrence on the suffix (Miller & Hardt, 2019); an attention head with known weights is governed by the mass assigned to omitted values. The resulting bounds express both mechanisms in units of tokens and output error.

Corollary 6.1 (Discrete precision locality)**.** *For $n \geq 1$, the discrete version of Definition 2.1 has the following consequences.*

(i) Selective scans. *Consider the token-local recurrence*

$$h_i = \bar{A}_i h_{i-1} + b_i, \quad h_0 = 0, \qquad F_N(u) = c_N^\top h_N + d_N, \tag{21}$$

where $\bar{A}_i, b_i, c_i, d_i$ depend on u_i and fixed parameters. For $0 < \varrho < 1$, suppose uniformly over allowed tokens that $\|\bar{A}_i\| \leq \varrho$, $\|b_i\| \leq B_d$, and $\|c_i\| \leq C_d$. Set $D_d = B_d/(1-\varrho)$ and $\mathcal{R}_d = \max\{1, C_d D_d\}$. For $1 \leq H \leq N$,

$$E_F(N, H) \leq C_d D_d \varrho^H, \tag{22}$$

$$H_F(N, n) \leq H_n := \min\left\{N, \max\left\{1, \left\lceil \frac{n \log 2 + \log \mathcal{R}_d}{-\log \varrho} \right\rceil\right\}\right\}. \tag{23}$$

With local polynomial-time coefficient realizers, a variable-precision reset scan on a certified $O(H_{n+2})$-token suffix evaluates F_N to error 2^{-n} in $\mathrm{poly}(n, \mathrm{len}(N))$ bit time at fixed dimensions and certificates, using bounded-size token names.

(ii) Softmax heads. *For a standard head with known logits ℓ_{Nj} independent of freely varying values $v_j \in [-1, 1]$, let $\alpha_{Nj} = e^{\ell_{Nj}} / \sum_{k=1}^N e^{\ell_{Nk}}$. Then*

$$E_{\mathrm{Att}}(N, H) = \sum_{j=1}^{N-H} \alpha_{Nj}. \tag{24}$$

In particular, the valid parameter choice $W_Q = W_K = 0$, $W_V = 1$ implements $\mathsf{U}_N(v) = N^{-1} \sum_{j=1}^N v_j$, giving

$$E_{\mathsf{U}_N}(N, H) = \frac{N-H}{N}, \qquad H_{\mathsf{U}_N}(N, n) = \left\lceil (1 - 2^{-n}) N \right\rceil. \tag{25}$$

The uniform lower bound already holds on $v_j \in \{-1, 0, 1\}$. For bounded nonconstant logits $|\ell_{Nj}| \le B_a$, the same construction gives $E_{\mathrm{Att}}(N, H) \ge e^{-2B_a}(N - H)/N$. A nonconstant-logit witness is realized with ordinary query/key/value projections on (z_j, v_j), selecting a known bounded z_j for queries/keys and v_j for values.

For the scan, contraction controls the contribution of the state entering the retained suffix. For attention with a known logit row, the missing prefix weight is exactly the uncertainty that remains after observing the suffix. The proof and finite-bit implementation are in Appendix D.

The Mamba certificate and its units. Mamba's real negative-diagonal selective scan has entries $\exp(-a_k\Delta_{ik})$ (Gu & Dao, 2023). Here i indexes tokens, k indexes flattened state/channel coordinates, $-a_k$ are the continuous transition's diagonal entries, and $\Delta_{ik} > 0$ are input-dependent discretization steps. A supplied bound $a_k \ge a > 0$, $\Delta_{ik} \ge \delta > 0$ gives $\varrho = e^{-a\delta}$. More precisely, the tail certificate is

$$C_d D_d \exp\left(- \sum_{i=N-H+1}^{N} \min_k a_k \Delta_{ik} \right), \tag{26}$$

which permits an input-adaptive backward stopping rule. The step-size margin converts cumulative contraction into a uniform token bound. DeciMamba analyzes effective receptive fields and extends context by selective decimation (Ben-Kish et al., 2025); the stopping rule here certifies a suffix for a prescribed absolute error. A width-w causal preprocessing map adds at most $w - 1$ raw tokens; bounded pointwise gates and residuals can be absorbed into c_N, d_N. These formulas provide explicit certificates for selective blocks and their local preprocessing and readout operations.

We can extend this to Mamba-3's rotations, two-token forcing, and multiple-input/multiple-output readouts (Lahoti et al., 2026), with one additional input supplying the boundary forcing.

Corollary 6.2 (Certified Mamba-3 context)**.** *For the token-local Mamba-3 recurrence (71) with orthogonal rotations, assume uniform bounds $a_t \ge a_* > 0$, $0 < \delta \le \Delta_t \le \bar{\Delta}$, $0 \le \lambda_t \le 1$, $\|V_t\|_F \le M$, and a C_{out}-Lipschitz readout in the Frobenius state norm. Set $D_* := M\bar{\Delta}/(1-e^{-a_*\bar{\Delta}})$. For $1 \le \ell < N$, reset $S_{N-\ell}$ to zero and run the last ℓ updates using the $\ell + 1$ inputs $u_{N-\ell}, \dots, u_N$. The output $\widetilde{F}_N$ satisfies*

$$\left\| F_N - \widetilde{F}_N \right\| \le C_{\mathrm{out}} D_* \exp\left(- \sum_{t=N-\ell+1}^{N} a_t \Delta_t \right) \le C_{\mathrm{out}} D_* e^{-a_*\delta\ell}. \tag{27}$$

Consequently, for $n \ge 1$,

$$H_F(N, n) \le \min\{N, 1 + \ell_n\}, \qquad \ell_n := \max\left\{1, \left\lceil \frac{\log \max\{1, 2^n C_{\mathrm{out}} D_*\}}{a_*\delta} \right\rceil \right\}. \tag{28}$$

Appendix D.6 gives the full statement and proof. These bounds show that Mamba-3's rotations and multiple input/output channels retain precision locality under the stated certificates, while two-token forcing costs only one boundary input. Consequently, a block can be restarted without its earlier state and its full-history output recovered to a chosen accuracy using context that grows only logarithmically with inverse error, independently of total history length.

For the scalar scan, continuous-time decay and the sampling interval jointly determine contraction per token, connecting the continuous and discrete context bounds. Appendix D derives this correspondence for inputs held constant between samples and gives the exact suffix recovery error.

7 Conclusion

Precision locality links recovery of a model's output from recent context to evaluation cost. For the attention and state-space operators, we have proven sharp query bounds; under regularity and stability certificates, we have established second-order polynomial-time evaluation for selective dynamics. Our approximation theorem identifies the memory resources needed to approximate global averages over growing horizons: the studied contractive models must forget more slowly or increase state magnitude, readout gain, or transient amplification. These results provide a quantitative basis for selecting context and motivate architectures that combine efficient local processing with mechanisms for retaining distant information.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of the paper.

ETHICS STATEMENT

This work consists of theoretical analysis and proofs; there is no particular ethical concern involved.

REPRODUCIBILITY STATEMENT

All proofs are in the appendix. Appendix B proves the deterministic and randomized query bounds, including the interpolation lemma, periodic lower-bound construction, and precision-cost results. Appendix C proves the general evaluation guarantees, exact suffix errors, and memory tradeoff. Appendix D proves the discrete results. The evaluation proofs include finite-bit execution.

REFERENCES

Josh Alman and Zhao Song. Fast attention requires bounded entries. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2302.13214.

Gustave Bainier, Antoine Chaillet, Rodolphe Sepulchre, and Alessio Franci. State-space fading memory. arXiv:2603.23814, 2026.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. DeciMamba: Exploring the length extrapolation potential of Mamba. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.14528.

Stephen Boyd and Leon O. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150-1161, 1985. doi: 10.1109/TCS.1985.1085649. URL https://web.stanford.edu/~boyd/papers/fading_volterra.html.

Stanislav Budzinskiy, Wenyi Fang, Longbin Zeng, and Philipp Petersen. Numerical stability analysis of large language models. arXiv:2503.10251, 2025.

Edoardo Calvello, Nikola B. Kovachki, Matthew E. Levine, and Andrew M. Stuart. Continuum attention for neural operators. *Journal of Machine Learning Research*, 26(300):1-52, 2025. URL https://www.jmlr.org/papers/v26/24-0879.html.

Yifang Chen, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. The computational limits of state-space models and Mamba via the lens of circuit complexity. arXiv:2412.06148, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of ICML*, volume 235 of *PMLR*, pp. 10041-10071, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752, 2023.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2008.07669.

Joshua Hanson and Maxim Raginsky. Universal approximation of input-output maps by temporal convolutional nets. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.09211.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of ICML*, volume 119 of *PMLR*, pp. 5156-5165, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Akitoshi Kawamura and Stephen Cook. Complexity theory for operators in analysis. *ACM Transactions on Computation Theory*, 4(2):5, 2012. doi: 10.1145/2189778.2189780. URL https://arxiv.org/abs/1305.0453.

Akitoshi Kawamura, Norbert Th. Müller, Carsten Rösnick, and Martin Ziegler. Parameterized uniform complexity in numerics: From smooth to analytic, from NP-hard to polytime. arXiv:1211.4974, 2012.

Nodens F. Koren and Samuel Lanthaler. Merging memory and space: A state space neural operator. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=SwLxxz0x58.

Yuval Koren, Assaf Ben-Kish, Raja Giryes, Lior Wolf, and Itamar Zimerman. On the recall scaling laws in Mamba: A theoretical and mechanistic study via hashing. arXiv:2609.07681, 2026.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1-97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.15569.

Jiaoda Li and Ryan Cotterell. Characterizing the expressivity of fixed-precision transformer language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.23623.

Jiaoda Li and Ryan Cotterell. Characterizing the expressivity of local attention in transformers. In *Proceedings of ACL (Long Papers)*, pp. 37485-37507, 2026. URL https://aclanthology.org/2026.acl-long.1739/.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=q2Lnyegkr8.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. arXiv:2404.08819, 2024.

John Miller and Moritz Hardt. Stable recurrent models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hygxb2CqKm.

Eike Neumann and Florian Steinberg. Parametrised second-order complexity theory with applications to the study of interval computation. arXiv:1711.10530, 2017.

Juan-Pablo Ortega and Florian Rossmannek. Stochastic dynamics learning with state-space systems. *Mathematical Models and Methods in Applied Sciences*, 36(8):1783-1829, 2026. doi: 10.1142/S0218202526420066. URL https://arxiv.org/abs/2508.07876.

Irwin W. Sandberg. Structure theorems for nonlinear systems. *Multidimensional Systems and Signal Processing*, 2(3):267-286, 1991. URL https://doi.org/10.1007/BF01952236.

Irwin W. Sandberg. Approximately-finite memory and input-output maps. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 39(7):549-556, 1992. URL https://doi.org/10.1109/81.257287.

Zeyuan Song and Zheyu Jiang. Adaptive Mamba neural operators. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OenyzvFZPs.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: A unifying framework for designing sequence models with associative memory. *Journal of Machine Learning Research*, 27 (169):1-41, 2026. URL https://www.jmlr.org/papers/v27/25-0903.html.

Shida Wang and Qianxiao Li. StableSSM: Alleviating the curse of memory in state-space models through stable reparameterization. In *Proceedings of ICML*, volume 235 of *PMLR*, pp. 50766-50793, 2024. URL https://proceedings.mlr.press/v235/wang24ag.html.

Shida Wang and Beichen Xue. State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2309.13414.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.06464.

Luca Zancato, Arjun Seshadri, Yonatan Dukler, Aditya Golatkar, Yantao Shen, Benjamin Bowman, Matthew Trager, Alessandro Achille, and Stefano Soatto. B'MOJO: Hybrid state space realizations of foundation models with eidetic and fading memory. arXiv:2407.06324, 2024.

Itamar Zimerman, Ameen Ali, and Lior Wolf. Explaining modern gated-linear RNNs via a unified implicit attention formulation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.16504.

A Representation conventions

This section makes explicit the evaluation model introduced in Section 2: point-query access, quantitative regularity, and second-order bit cost. It specifies input and output names, physical-time scaling, and the distinction between exact suffix information and finite sampling. These conventions support Corollary 3.2 and the local coefficient access assumed by the certified evaluators in Section 4.

A.1 Names, encodings, and size functions

All logarithms without a subscript are natural. Accuracy indices are nonnegative integers supplied in unary. A dyadic $t = j2^{-k}$ is encoded by a signed binary numerator and its denominator exponent in unary, with redundant encodings allowed. The length includes both the integer and fractional parts. The horizon T in the sharp results is a positive integer supplied in binary. Arbitrary dyadic output locations use their own encoding lengths in place of $\mathrm{len}(T)$.

A name of a continuous input is a pair consisting of a modulus and a point evaluator. A modulus $\mu : \mathbb{N} \to \mathbb{N}$ satisfies

$$|t - u| \le 2^{-\mu(p)} \quad \Longrightarrow \quad |x(t) - x(u)| \le 2^{-p}.$$

The evaluator answers $(t, 1^p)$ by a dyadic within 2^{-p} of $x(t)$. Requests outside $[0, T]$ or syntactically invalid requests carry no information and can be answered by a fixed string. Norms and errors for vector-valued functions are Euclidean, and matrix norms are induced operator norms. Component evaluations to an additional $O(\log(d + r + m + 1))$ bits suffice to enforce the corresponding vector or matrix error.

For a quantitative Gevrey class the supplied certificate replaces a variable modulus with the common modulus

$$\mu(p) = p + \lceil \log_2(CR) \rceil .$$

The certificate is a supplied promise shared by every function in the fixed ball. Algorithms use it directly, and the lower-bound constructions satisfy the same common promise.

A standard tagged, padded encoding produces a length-monotone string function φ_x, meaning $|u| \le |v| \Rightarrow |\varphi_x(u)| \le |\varphi_x(v)|$, with size function $\ell_x(k) = \max_{|u|\le k} |\varphi_x(u)|$. For completeness, tag one kind of request as a value query and another as a modulus query. Encode the dyadic answer at a prescribed precision, together with a delimiter, and pad it with zeros. In the bounded classes of Theorem 3.1, every valid request of length k needs only $O(k+1)$ output bits: precision is unary, the magnitude is at most one, and the common modulus is linear. Padding every response of a given input length to the same fixed linear length gives

$$\ell_x(k) \le c(k+1) \tag{29}$$

with a constant independent of both x and T. Values can, for example, be rounded to a grid of mesh 2^{-p-1} before padding. Invalid queries have the same padding convention. In particular, the names of a fooling function and of zero can be chosen to have identical sizes and identical answers at all queried zeros.

Second-order polynomials are expressions generated from nonnegative constants, ordinary input-length variables, addition, multiplication, and application of a size function ℓ to such an expression. An oracle machine must write queries, read answers, and perform finite-bit arithmetic within this bound. Substituting the linear bound (29) into any fixed second-order polynomial leaves an ordinary polynomial. This transfers the query separation to the bit-complexity separation in Corollary 3.2. Conversely, a polynomial number of queries of polynomial length, a polynomial amount of arithmetic, and polynomial access to the returned names give the upper bounds claimed here. For nested coefficient realizers, the usual substitution of second-order bounds accounts for oracle access.

Precision convention. A value oracle specifies an absolute-error guarantee. All upper bounds hold for every sequence of answers satisfying that guarantee. All deterministic lower bounds work with exact values and therefore also with the particular finite-precision names described above. The query model permits arbitrary answer precision; the lower-bound constructions give exactly zero at every queried point, establishing the bounds throughout this precision range.

Locality convention for general coefficients. A local coefficient realizer for $A_\theta(x(t))$, $b_\theta(x(t))$, or $C_\theta(x(t))$ obtains its input-dependent information at the requested time t, at any required value precision. It also has access to known finite parameters and known time-dependent functions. For attention, the analogous realizer may access the two indicated input times. The realizer costs and finite descriptions are included in $\operatorname{size}\theta$ and in the stated second-order evaluation bounds. This gives the suffix access assertion in Theorem 4.2 an explicit algorithmic realization.

Certificates and implementations. Coefficient-realizer costs, intermediate-layer bounds, and stability margins connect architectural structure to quantitative computation. Our constructive high-order and finite-bit algorithms provide certified evaluators with explicit resource budgets. Matching exact-oracle lower bounds establish the optimal acquisition rates, giving a reference for implementations that exploit regularity and memory decay.

A.2 Output names and physical time

A continuous-function output name consists of a modulus and an evaluator at dyadic times. For the four witnesses, values lie in $[-1,1]$ and the following Lipschitz bounds are uniform in T:

$$\operatorname{Lip}(\mathsf{U}x) \le R/2, \quad \operatorname{Lip}(\mathsf{S}_0 x) \le 2, \quad \operatorname{Lip}(\mathsf{S}_{\mathrm{sel}} x) \le 3, \quad \operatorname{Lip}(\mathsf{A}_* x) \le R+3. \tag{30}$$

Indeed, $\mathsf{U}x(t) = \int_0^1 x(tr)\,dr$, and differentiation gives the first bound, including the continuous extension at zero. For either SSM, variation of constants and decay at least one imply $|h(t)| \le 1$; its defining differential equation gives the second and third bounds. For A_*, $|z| \le 3/2$, $|z'| \le 1/2$, hence each partial derivative of its score is at most $3/4$. Proposition 4.1 gives the last bound. The output modulus can thus be chosen as $p + \lceil \log_2 \max\{1, L\} \rceil$ for the relevant constant L.

At a dyadic location t, the same upper algorithms apply on $[0,t]$. For $0 < t < 1$, retain the additive one in Lemma B.1; the sharp rates are stated for integer horizons $T \ge 1$. For the SSMs, a dyadically

rounded window of length $O(n)$ suffices; its starting point has bit length polynomial in $n + \mathrm{len}(t)$. The continuous upper bounds therefore define function-valued realizers that answer queries at any output time. On a fixed horizon, the attention witnesses are precision-polynomial on the stated Gevrey balls. The lower bounds apply at the permitted endpoint query $t = T$ and consequently hold for full output names as well.

The representation uses physical time. For $\widetilde{x}(u) = x(Tu)$ on $[0, 1]$, the bound becomes $\left\|\widetilde{x}^{(k)}\right\|_\infty \leq C(TR)^k(k!)^s$. A unit-interval representation therefore records the horizon through the rescaled radius TR. The query laws also hold with unary horizon encoding; binary encoding additionally makes the horizon dependence an exponential resource in Corollary 3.2.

A.3 Information locality, finite access, and analyticity

Definition 2.1 refers to an explicitly chosen history class. For the architecture-level continuous examples this is $\mathcal{B}_T = \{x \in C([0, T]) : \|x\|_\infty \leq 1\}$; for the discrete corollary it is an arbitrary bounded token product space. The suffix recovery map receives exactly the retained suffix, known model parameters, endpoint, and prescribed initial condition. The superscript $\mathcal{B}$ denotes this ambient class and is suppressed when the class is clear. Quantitative Gevrey balls specify the additional regularity used in the sampling bounds.

The endpoint and input interface. The discrete corollary provides endpoint evaluation from input access without a precomputed prefix state or summary. In both settings, specifying the output query makes the accuracy-dependent history resource explicit.

The infimum in (1) ranges over all recovery maps and measures information sufficiency. Our results realize this information bound constructively: zero-state reset supplies an explicit recovery map, and interpolation or finite-token algorithms supply its implementation. If the recovery map incurs error at most 2^{-n-2} and its realization incurs another 2^{-n-2}, their composition has error at most 2^{-n-1}, leaving a rounding reserve. This error decomposition connects locality to precision complexity through explicit algorithms.

Analytic continuation and finite sampling. If two real-analytic functions on a connected neighborhood of $[0, T]$ agree on a nonempty interval, their difference is identically zero. To see this, zeros accumulating in the interior force all Taylor coefficients there to vanish; the resulting zero neighborhood extends across overlapping convergence neighborhoods throughout the connected interval. Consequently, if the infimum in (1) were restricted to an analytic input class and given an *entire exact* suffix, it would have zero error for every positive suffix length. This characterizes the information supplied by an exact interval; the query theorems quantify acquisition from finitely many point observations.

Ambient-history locality expresses how the architecture attenuates a bounded prefix, while quantitative regularity controls how many finite observations suffice to realize a recovery map. Lemma B.2 establishes the analytic sampling law through agreement at finitely many observed points, consistently with unique continuation from an interval. Theorem 5.1 uses compact Gevrey bumps with $s > 1$ to quantify the resources for persistent memory.

Monotonicity and comparison of information interfaces. The error $E_{\mathsf{F}}^{\mathcal{B}}(T, H)$ is nonincreasing in H: every recovery map for a shorter suffix also applies to a longer one. Restricting to a subclass $\mathcal{X}_T \subseteq \mathcal{B}_T$ or supplying a prefix summary can likewise reduce the error. These monotonicity properties organize the value of additional information, including analytic interval data and previously computed recurrent states.

The computational value of additional information. An oracle for the primitive $I_x(t) = \int_0^t x(u)\,du$ makes the uniform-attention endpoint directly accessible as $I_x(T)/T$. Derivative, integral, or spectral oracles likewise offer richer information per query. Quantitative Gevrey promises and bounded-size point names provide a precise basis for the acquisition laws proved here.

B SHARP QUERY BOUNDS AND PRECISION COSTS

Section 3 clarifies the roles of history length, regularity, and randomization in input acquisition. Here interpolation and a periodic construction prove the deterministic laws in Theorem 3.1, followed by the binary-horizon and Lipschitz consequences and the decay-margin refinement of Proposition 4.3. Compact Gevrey bumps and a finite-bit sampling algorithm then prove the randomized law in Theorem 3.4.

B.1 CONSTRUCTIVE INTERPOLATION AND BIT COSTS

We first state and prove the interpolation result used by the continuous evaluators.

Lemma B.1 (Finite-precision interpolation)**.** *A function in $\mathcal{G}_s(C, R; H)$ can be uniformly approximated to error 2^{-p} by a piecewise polynomial using*

$$O_s\big((1 + HR\nu^{s-1})\nu\big)\,, \qquad \nu = p + \lceil \log_2 C \rceil + 1, \tag{31}$$

point queries, each with $O_s(\nu)$ accuracy bits. The construction and integration of the approximant have polynomial bit overhead in these parameters and endpoint encoding lengths.

Proof. Let f be defined on a dyadic interval $[a, a + H]$ and satisfy $\left\|f^{(k)}\right\|_\infty \le CR^k(k!)^s$. Non-dyadic target lengths can be replaced by dyadic partitions within the actual dyadic endpoints. Assume $p \ge 1$ and write $\nu = p + \lceil \log_2 C \rceil + 1$.

Choose m with $m - 1$ a power of two, $m \ge 2$, and $m = O(\nu)$ sufficiently large that $C4^{-m} \le 2^{-p-2}$. Choose a dyadic cell length a_0 satisfying

$$\frac{1}{8Rm^{s-1}} < a_0 \le \frac{1}{4Rm^{s-1}}. \tag{32}$$

For a fixed rational s, such a scale is computable with polynomial bit overhead; an extra factor of two may be used instead of an exact power comparison. Partition the interval into cells of length a_0, with one possibly shorter final cell. There are at most $1 + 8HRm^{s-1}$ cells. Every cell endpoint is dyadic.

On a cell of length $a_j \le a_0$, interpolate at the m equally spaced nodes. Since $m - 1$ is a power of two, these nodes are dyadic too. The exact interpolant P_j satisfies

$$\|f - P_j\|_\infty \le \frac{CR^m(m!)^s}{m!} a_j^m \le C(Ra_j m^{s-1})^m \le C4^{-m}. \tag{33}$$

This is the standard interpolation remainder on a real interval: for each evaluation point the error is $f^{(m)}(\xi)/m!$ times the product of distances to the nodes. All those distances are at most a_j.

For noise control, use local coordinates $u \in [0, 1]$. The Lagrange basis at nodes $i/(m - 1)$ obeys

$$|L_i(u)| \le \frac{(m-1)^{m-1}}{i!(m-1-i)!}.$$

Summing over i and using $(m - 1)! \ge ((m - 1)/e)^{m-1}$ gives

$$\Lambda_m := \sup_{u\in[0,1]} \sum_{i=0}^{m-1} |L_i(u)| \le \frac{2^{m-1}(m-1)^{m-1}}{(m-1)!} \le (2e)^{m-1}. \tag{34}$$

Query every value at precision

$$p' = p + 3 + \lceil (m - 1) \log_2(2e) \rceil\,.$$

The interpolant $\widetilde{P}_j$ of these values differs from P_j by at most 2^{-p-3}, so the total uniform error is below 2^{-p}. Multiplying the number of cells by m proves the query count, and $p' = O_s(\nu)$.

To bound the arithmetic overhead, we first round each approximate value once more to a fixed dyadic grid with a single guard bit. In local coordinates the node denominators have $O(\log m)$ bits. Products of m linear factors and denominators $i!(m - 1 - i)!$ yield polynomial coefficients whose rational bit lengths are polynomial in $m, p', \log C$. The cell translation and scaling require only their endpoint bit lengths. Integrating a degree-$(m - 1)$ local polynomial divides coefficients by integers at most m

and multiplies by the cell length. All cells can use a common denominator built from powers of two, $m!$, and $\mathrm{lcm}(1,\dots,m)$, giving a common representation for the sum across cells. The additional numerator length for a sum of J cells is $O(\log(J+1))$. Alternatively, each local integral can be rounded with an explicit $O(\log J)$ guard-bit budget. Thus the arithmetic overhead is polynomial in J, m, p' and the finite input lengths, as claimed.

□

Elementary known weights. On an interval of length $H = O(n)$, an exponential weight $e^{-(t-u)}$ can be approximated by a rational polynomial to accuracy $2^{-p}/(H+1)$ in time polynomial in $p + H$ and endpoint lengths. For example, truncate its Taylor series to degree $O(p + H + 1)$ with a sufficiently large absolute constant, using the bound $e^H H^{k+1}/(k+1)!$ on the remainder. Multiply this polynomial by the local interpolant and integrate rationally. The known-weight approximation uses local arithmetic with the lag $t-u \in [0, H]$ as its exponential argument, keeping both its precision and its degree controlled by the retained window.

B.2 Deterministic query bounds and precision costs

This section proves Theorem 3.1, Corollary 3.2, and Propositions 3.3 and 4.3. The upper constructions use Lemma B.1; the matching lower bounds use the periodic construction below.

B.2.1 Upper bounds in Theorem 3.1

Throughout this subsection s, R are fixed, $\|x\|_\infty \le 1$, and $\varepsilon = 2^{-n}$. Small values of n are absorbed into the constants; all constructions are uniform in the binary integer T.

Uniform attention and the stable linear filter For U, apply Lemma B.1 with uniform error at most $\varepsilon/4$ on $[0, T]$. Since $T \ge 1$, the query count is $O_{s,R}(Tn^s)$. Integrate the resulting rational piecewise polynomial and divide by T. Uniform error contributes at most $\varepsilon/4$, and rounding the final result at precision $n+2$ contributes at most another $\varepsilon/4$. Arithmetic costs at most T times a polynomial in $n + \mathrm{len}(T)$ and the input-name size. Numerical summation uses guard bits depending on $\log T$, while the *input-value* precision remains $O(n)$ uniformly in T, because the normalized integration functional has norm one in the uniform norm.

For S_0, retain $H = \min\{T, n+4\}$. If $H < T$, then

$$\left| \int_0^{T-H} e^{-(T-u)} x(u)\, du \right| \le e^{-H}(1 - e^{-(T-H)}) \le e^{-H} < \varepsilon/8. \tag{35}$$

If $H = T$, this discarded part is zero. On the remaining interval approximate x uniformly to error $\varepsilon/8$, using $O_{s,R}(Hn^s)$ queries. The retained kernel has integral at most one. Its rational polynomial approximation, integration, and final rounding may each be allocated error $\varepsilon/8$. The construction in Appendix B.1 supplies all three in polynomial time. Since $H \le n+4$, the total bit cost is polynomial in $n + \mathrm{len}(T)$ and the input-name size. The query count is $O_{s,R}(n^s \min\{T, n\})$ for $n \ge 1$, up to a fixed factor.

Nonuniform softmax attention Put $w_T(u) = e^{z(T)z(u)}$ and $B = 9/4$. Then

$$e^{-B} \le w_T(u) \le e^B, \qquad \frac{w_T(u)}{\int_0^T w_T(v)\, dv} \ge \frac{e^{-2B}}{T}.$$

The lower bound is used in Appendix B.2.2. To prove the upper bound, we first establish uniform analytic bounds for w_T independent of T. On complex disks of radius $1/2$ about the real line, $|\sin z| \le e^{1/2}$, hence $|1 + \frac{1}{2}\sin z|$ is bounded by an absolute constant. The real factor $z(T)$ lies in $[1/2, 3/2]$. Therefore $|w_T(z)|$ is bounded by an absolute constant on these disks, and Cauchy's estimate yields

$$\left\| w_T^{(k)} \right\|_\infty \le C_0 R_0^k k!$$

for absolute C_0, R_0. By Leibniz's rule, for $s \geq 1$,

$$\left\|(w_T x)^{(k)}\right\|_\infty \leq C_0 \sum_{j=0}^{k} \binom{k}{j} R_0^j j!\, R^{k-j}((k-j)!)^s \tag{36}$$

$$\leq C_0 (R_0 + R)^k (k!)^s. \tag{37}$$

Thus both w_T and $w_T x$ have fixed Gevrey parameters. Approximate their normalized integrals with error $\delta = c\varepsilon$ for a sufficiently small absolute $c > 0$. The denominator is at least e^{-B}, so Proposition 4.1 gives total error at most ε after final rounding. At each node, multiplication by the bounded known weight requires only $n + O(1)$ bits of the input value. Lemma B.1 therefore gives $O_{s,R}(Tn^s)$ input queries.

The function $\sin$ at a binary dyadic argument of magnitude at most T can be evaluated in time polynomial in $n + \mathrm{len}(T)$. One elementary implementation computes π to $n + O(\mathrm{len}(T))$ bits (for example, with Machin's formula and geometrically convergent arctangent series), performs argument reduction with sufficient guard bits, and uses a Taylor expansion on a bounded interval. The exponential is evaluated on a bounded real interval. This explicitly accounts for the known side channel within polynomial bit cost in $n + \mathrm{len}(T)$.

The input-selective witness For $\mathsf{S}_{\mathrm{sel}}$, the rate $a_x(u) = 1 + x(u)^2$ is at least one. Variation of constants gives $|h(u)| \leq 1$. Resetting at $T - H$, $H = \min\{T, n + 4\}$, changes the terminal state by at most e^{-H}, with no reset when $H = T$.

On the retained interval let $\widetilde{x}$ be a piecewise polynomial with $\|\widetilde{x} - x\|_\infty \leq \delta \leq 1$. Form exactly the piecewise polynomial coefficients

$$\widetilde{a} = 1 + \widetilde{x}^2, \qquad \widetilde{b} = \widetilde{x}.$$

Then $\widetilde{a} \geq 1$ and

$$|\widetilde{a} - a_x| \leq \delta(2 + \delta) \leq 3\delta.$$

Let h_* be the true reset solution and $\widetilde{h}$ the exact solution of the polynomial-coefficient equation, with the same starting state. Their difference satisfies

$$(\widetilde{h} - h_*)' = -\widetilde{a}(\widetilde{h} - h_*) - (\widetilde{a} - a_x)h_* + (\widetilde{x} - x).$$

Since $|h_*| \leq 1$ and $\widetilde{a} \geq 1$, variation of constants gives the especially simple estimate

$$|\widetilde{h}(T) - h_*(T)| \leq 4\delta. \tag{38}$$

Taking $\delta = \varepsilon/32$ gives an $O(n)$ input-precision budget. Lemma B.1 gives $O_{s,R}(Hn^s)$ input queries. The coefficient bounds $1 \leq \widetilde{a} \leq 5$, $|\widetilde{b}| \leq 2$ imply that the local Picard construction of Appendix C.2 solves this equation to another $\varepsilon/8$ in polynomial bit time. Its degrees are polynomial in n, the retained length is $O(n)$, and it operates directly on the sampled polynomial coefficients. Accounting for reset and final rounding errors completes the proof of the full upper bound.

This implementation exhibits input selectivity explicitly: the transition coefficient depends nonlinearly on the varying input, with uniform stability and finite-precision perturbation control throughout the input class.

B.2.2 The periodic construction and sharp lower bounds

The following lemma constructs mass invisible to any prescribed finite set of point observations, while satisfying derivative bounds that include the analytic case.

Lemma B.2 (Mass invisible to finitely many samples)**.** *Let $s \geq 1$, $L > 0$, and $z_1, \ldots, z_q \in \mathbb{R}$, with $q \geq 1$. There is a nonnegative, L-periodic analytic function f such that $f(z_i) = 0$, $\left\|f^{(k)}\right\|_\infty \leq (k!)^s$ for all $k \geq 0$, and, on every interval I of length L,*

$$\frac{1}{L}\int_I f(u)\,du \;\geq\; \frac{1}{2}\exp\left[-s(4\pi q/L)^{1/s}\right]. \tag{39}$$

Proof. Put $D = 2q + 1$ and let $v(u)$ be the vector consisting of 1 and the q sine/cosine pairs $\sqrt{2}\cos(2\pi ju/L), \sqrt{2}\sin(2\pi ju/L)$. Let P be the orthogonal projection onto the complement of the span of $v(z_1), \dots, v(z_q)$. Define

$$F(u) = D^{-1}v(u)^\top Pv(u).$$

Then $0 \le F \le 1$, $F(z_i) = 0$, and $L^{-1}\int_I F = \operatorname{tr}(P)/D \ge 1/2$. Moreover,

$$\left\|v^{(j)}(u)\right\| \le \sqrt{D}(2\pi q/L)^j, \qquad \left\|F^{(k)}\right\|_\infty \le (4\pi q/L)^k,$$

by the product rule. For $\omega = 4\pi q/L$, the inequality $\omega^k/(k!)^s \le \exp(s\omega^{1/s})$ follows from the exponential series. Hence $f = \exp(-s\omega^{1/s})F$ has all the required properties. □

Proof of the lower bounds in Theorem 3.1. Run a deterministic algorithm on the zero input and collect its query points. Construct f using Lemma B.2; the inputs f and $-f$ produce the same transcript, regardless of query precision. With no queries, constant inputs already give a constant lower bound.

For U, take $L = T$. Its two possible outputs have magnitude at least the right side of (39). Accuracy 2^{-n} therefore requires

$$q \ge \frac{T}{4\pi}\left(\frac{(n-1)\log 2}{s}\right)^s. \tag{40}$$

For A_*, $|z(T)z(u)| \le B = 9/4$ implies that the normalized weight is at least e^{-2B}/T. The same argument changes only the additive constant in n.

For S_0, put $L = \min\{T, n\log 2/2\}$. On the last L units of time, the kernel is at least e^{-L}, and contributions outside that interval are nonnegative. Thus

$$(\mathsf{S}_0 f)(T) \ge \frac{L}{2}\exp\left[-L - s(4\pi q/L)^{1/s}\right]. \tag{41}$$

For sufficiently large n, $L \ge 1$, and requiring this to be at most 2^{-n} forces $q = \Omega_s(Ln^s)$. For $\mathsf{S}_{\mathrm{sel}}$, the inputs $\pm f$ have the same decay rate $1 + f^2 \in [1, 2]$ and opposite forcing. Their outputs are opposite, with the positive one bounded below as in (41) with $2L$ replacing L. Choose $L = \min\{T, n\log 2/4\}$ to obtain the same order. Every constructed input belongs to $\mathcal{G}_s(1, 1; T) \subseteq \mathcal{G}_s(1, R; T)$. □

B.2.3 Binary-horizon feasibility

Proof of Corollary 3.2. The SSM upper bounds are polynomial in $n + \mathrm{len}(T)$ and the name size. For attention, fix any sufficiently large n and choose names with a common linear size bound. Equation (40) then contradicts every polynomial in $\mathrm{len}(T)$. Explicit output moduli in Appendix A lift the evaluation upper bounds to function-valued operators. □

B.2.4 Sharp precision costs on Lipschitz inputs

Proof of Proposition 3.3. On the zero transcript let A be the q query points. The function $f(u) = \mathrm{dist}(u, A \cup \{0, 1\})$ is 1-Lipschitz and vanishes on A. If consecutive gaps have lengths Δ_j, then $\int_0^1 f = \frac{1}{4}\sum_j \Delta_j^2 \ge 1/[4(q+1)]$. The stable kernel is at least e^{-1}. Indistinguishability of $\pm f$ gives the lower bounds; uniform-grid quadrature gives matching upper bounds. Integration equals $(\mathsf{U}f)(1)$ and also equals $(\mathsf{S}_0 x_f)(1)$ for $x_f(u) = e^{1-u}f(u)$. □

B.2.5 Sharp decay-margin cost

Proof of Proposition 4.3. The upper bound retains $H = \min\{T, (n+4)/\gamma\}$, rounded upward when necessary. The tail is at most $e^{-\gamma H}$, and the normalized retained kernel has mass at most one. Lemma B.1 therefore uses $O_{s,R}(n^{s+1}/\gamma)$ queries at $O(n)$ accuracy bits. Known weights require no additional input queries; their exponential arguments are $O(n)$ on the retained window.

For the lower bound, use Lemma B.2 with $L = n\log 2/(2\gamma) \le T$. On the final interval of length L, the kernel is at least $\gamma e^{-\gamma L}$, so the zero-transcript ambiguity is at least

$$\frac{\gamma L}{2}\exp\left[-\gamma L - s(4\pi q/L)^{1/s}\right].$$

Since $\gamma L = n\log 2/2$, accuracy 2^{-n} forces $q = \Omega_s(Ln^s) = \Omega_s(n^{s+1}/\gamma)$, uniformly in γ. □

B.3 Randomized query bounds and Gevrey bump construction

We prove Theorem 3.4. The compact bump constructed first also supplies the inputs used in the approximation tradeoff of Theorem 5.1.

B.3.1 A fixed compact Gevrey bump

Fix $s > 1$, let $\alpha = (s-1)^{-1}$, and define

$$b(u) = \begin{cases} \exp[-u^{-\alpha} - (1-u)^{-\alpha}], & 0 < u < 1, \\ 0, & u \notin (0,1). \end{cases} \tag{42}$$

This nonnegative function has $0 < \beta := \int_0^1 b(u)\,du \le 1$. It is smooth and flat at 0 and 1. The following derivative estimate places all its signed translates in a common quantitative Gevrey class.

For $g(u) = e^{-u^{-\alpha}}$ on $u > 0$, choose $\delta > 0$ depending only on α so that a complex disk of radius δu about u lies in a sector where $\Re(z^{-\alpha}) \ge cu^{-\alpha}$ for a fixed $c > 0$. Cauchy's estimate gives

$$|g^{(k)}(u)| \le k!(\delta u)^{-k} e^{-cu^{-\alpha}}.$$

For $k \ge 1$, maximizing the right side over $u > 0$ uses

$$\sup_{u>0} u^{-k} e^{-cu^{-\alpha}} = \left(\frac{k}{c\alpha e}\right)^{k/\alpha}.$$

Using $k! \ge (k/e)^k$, one obtains $|g^{(k)}(u)| \le D_0^k (k!)^{1+1/\alpha} = D_0^k (k!)^s$. Derivatives tend to zero as $u \downarrow 0$ because exponential decay dominates each fixed power. Apply the same argument to $g(1-u)$ and the product rule. Increasing a constant gives an integer $R_s \ge 1$ with

$$\left\| b^{(k)} \right\|_\infty \le R_s^k (k!)^s \quad (k \ge 0). \tag{43}$$

Disjoint integer translates can be pasted with arbitrary signs without increasing this uniform derivative bound: at a shared endpoint all derivatives are zero.

B.3.2 Two product distributions over input signs

For $T \in \mathbb{N}$ and $\sigma \in \{-1, 1\}^T$, define

$$x_\sigma(u) = \sum_{j=0}^{T-1} \sigma_j b(u - j). \tag{44}$$

Every x_σ belongs to $\mathcal{G}_s(1, R_s; T)$, and

$$(\mathsf{U} x_\sigma)(T) = \frac{\beta}{T} \sum_{j=0}^{T-1} \sigma_j. \tag{45}$$

An exact point query in the interior of a cell can reveal its sign; a boundary query reveals none. In either case it reveals no sign from any other cell. The class certificate, common modulus, and padded answer lengths are independent of σ.

Put $\varepsilon = 2^{-n}$ and $a = 8\varepsilon/\beta$. For sufficiently large n, $a \le 1/2$. Under P_+, let the signs be independent with mean a; under P_-, use mean $-a$. The corresponding output means are $\pm 8\varepsilon$. For $T \ge \varepsilon^{-2}$,

$$\mathrm{Var}_{P_\pm}((\mathsf{U} x_\sigma)(T)) \le \frac{\beta^2}{T} \le \beta^2 \varepsilon^2.$$

Chebyshev's inequality implies probability at least $15/16$ that the output lies within 4ε of its mean, and hence has the correct sign with magnitude at least 4ε.

Suppose an evaluator has error at most ε with probability at least $3/4$ for every input. Under either product distribution, the sign of its output correctly identifies the distribution with probability at least

$3/4 - 1/16 = 11/16$. This holds after averaging over the evaluator's internal randomness, because its error guarantee is pointwise in the input.

For one revealed sign, write P_1^+ and P_1^- for the Bernoulli laws on $\{-1, 1\}$. Direct calculation gives

$$\chi^2(P_1^+ \| P_1^-) = \frac{4a^2}{1-a^2} \le 6a^2. \tag{46}$$

For q independent samples, product factorization yields

$$1 + \chi^2((P_1^+)^{\otimes q} \| (P_1^-)^{\otimes q}) = (1 + \chi^2(P_1^+ \| P_1^-))^q \le e^{6qa^2}.$$

Cauchy-Schwarz then bounds total variation by

$$\mathrm{TV}((P_1^+)^{\otimes q}, (P_1^-)^{\otimes q}) \le \tfrac{1}{2}\sqrt{e^{6qa^2} - 1}. \tag{47}$$

An adaptive transcript can be simulated by drawing a fresh independent sign whenever the evaluator first enters a previously unqueried cell. Repeated queries, exact magnitudes of the known bump, and internal random bits are post-processing of at most q such signs. If fewer than q new cells are visited, add unused independent signs. Total variation cannot increase under this post-processing.

For equally likely hypotheses the largest testing success probability is $(1 + \mathrm{TV})/2$. Thus success at least $11/16$ requires $\mathrm{TV} \ge 3/8$, and (47) forces $q \ge ca^{-2} = c_s\varepsilon^{-2}$. This proves the randomized lower bound, including adaptive algorithms and exact value answers.

B.3.3 A FINITE-BIT RANDOMIZED UPPER BOUND

Uniform Monte Carlo admits an explicit finite-bit implementation. Choose an integer b_0 such that $RT2^{-b_0} \le \varepsilon/8$, and draw independent integers J_i uniformly from $\{0, \dots, 2^{b_0} - 1\}$. The dyadic locations $u_i = TJ_i2^{-b_0}$ lie in $[0, T]$. Since $\mathrm{Lip}(x) \le R$, the mean of x on this grid differs from the continuous average by at most $RT2^{-b_0} \le \varepsilon/8$.

The exact grid values have variance at most one. With $q \ge 16\varepsilon^{-2}$, their sample mean is within $\varepsilon/2$ of the grid expectation with probability at least $3/4$. Query each value to accuracy $\varepsilon/8$ and round the final average to another $\varepsilon/8$. These extra errors are bounded deterministically, even if valid oracle errors vary adversarially across queries. The resulting total error is at most $7\varepsilon/8$. Each location uses $O(n + \mathrm{len}(T) + \log R)$ random and coordinate bits, and each input value uses $n + O(1)$ accuracy bits. Thus $Q_{\mathsf{U}}^{\mathrm{ran}} = O(\varepsilon^{-2})$ in the claimed bit-compatible model.

C CERTIFIED EVALUATION AND MEMORY TRADEOFFS

This section supplies the certified evaluators of Section 4 and the approximation obstruction of Section 5. Using interpolation from Appendix B.1, we prove Proposition 4.1 and Theorem 4.2, including output moduli and finite-bit error budgets. We then prove the exact suffix errors in (3) from Section 2 and Theorem 5.1, which uses the compact bump from Appendix B.3 to quantify the resources needed to approximate persistent memory.

C.1 GENERAL ATTENTION: NORMALIZATION, ENDPOINT CONTINUITY, AND RESOURCES

We prove Proposition 4.1, including its endpoint and output-modulus assertions.

At $t > 0$, set

$$N(t) = \frac{1}{t}\int_0^t e^{\lambda_x(t,u)} v_x(u)\, du, \qquad Z(t) = \frac{1}{t}\int_0^t e^{\lambda_x(t,u)}\, du.$$

The bounds $Z(t) \ge e^{-B}$ and $\|N(t)\| \le VZ(t)$ hold uniformly, including arbitrarily small positive t. If $\left\|\widetilde{N} - N\right\| \le \delta$ and $|\widetilde{Z} - Z| \le \delta \le e^{-B}/2$, then $\widetilde{Z} \ge Z/2$ and

$$\left\|\frac{\widetilde{N}}{\widetilde{Z}} - \frac{N}{Z}\right\| \le \frac{\left\|\widetilde{N} - N\right\|}{|\widetilde{Z}|} + \frac{\|N\|\,|Z - \widetilde{Z}|}{Z|\widetilde{Z}|} \tag{48}$$

$$\le 2(1+V)e^B\delta. \tag{49}$$

Choosing a dyadic $\delta \le 2^{-n-3}e^{-B}/(1+V)$, approximating both integrands uniformly to this tolerance, and integrating their approximants as averages gives a sufficient error budget. Lemma B.1 uses precision $p = n + O(B + \log(1+V))$, with an additional $O(\log M)$ contribution to the interpolation order. The algorithm is polynomial in the *numerical* radius R and the horizon t. The supplied coefficient realizers implement the local evaluations, with their resources included explicitly in the bound.

For the output modulus, substitute $u = tr$ and define the probability density

$$\pi_t(r) = \frac{e^{\lambda_x(t,tr)}}{\int_0^1 e^{\lambda_x(t,tv)}\,dv}, \qquad 0 \le r \le 1.$$

Writing $g_t(r) = \partial_1\lambda_x(t,tr) + r\partial_2\lambda_x(t,tr)$, differentiation under the integral yields

$$\frac{d}{dt}(\mathsf{A}_\theta x)(t) = \mathbb{E}_{\pi_t}[r v_x'(tr)] + \mathbb{E}_{\pi_t}[v_x(tr)g_t(r)] - \mathbb{E}_{\pi_t}[v_x(tr)]\mathbb{E}_{\pi_t}[g_t(r)]. \tag{50}$$

The first term is at most L_v, and each of the two remaining terms is at most $2VL_\lambda$. Hence the derivative bound is $L_v + 4VL_\lambda$. The representation on $[0,1]$ has a positive bounded denominator at $t = 0$ as well; continuity and differentiation extend there under the stated derivative assumptions. The output is bounded by V. These two bounds produce a full continuous-function name.

C.2 General selective dynamics: a complete bit algorithm

We prove Theorem 4.2 with all resources explicit. Let the certificate satisfy $M, R, K \ge 1$ and $\gamma > 0$, let $s \ge 1$ be fixed rational, and include the rational initial state, finite coefficient programs, and finite encodings of certificates in $\operatorname{size}\theta$. The numerical parameters M, R, γ^{-1} in the theorem can equivalently be supplied with unary integer upper bounds. The resulting parameterized polynomial-time statement records both accuracy and certificate values as computational resources.

C.2.1 Reset error and a computable window

Variation of constants and the stability certificate give

$$\|h(t)\| \le Ke^{-\gamma t}\|h(0)\| + \int_0^t Ke^{-\gamma(t-u)}\|b_x(u)\|\,du \tag{51}$$

$$\le KM + KM/\gamma = D. \tag{52}$$

If $a = t - H > 0$, let h_* denote the solution using the same coefficients on $[a,t]$, with $h_*(a) = 0$. The difference from the original solution is $\Phi_x(t,a)h(a)$, so the terminal readout error satisfies

$$\|C_x(t)\Phi_x(t,a)h(a)\| \le \rho e^{-\gamma H}. \tag{53}$$

If $a = 0$, use the supplied initial state rather than zero, giving no reset error. The retained true solution also obeys $\|h_*\| \le D$.

Rational certificate bounds yield a computable rounded window. Let r_ρ be an integer upper bound for $\log_2 \rho$, computable from rational certificate bounds with polynomial bit cost. For example one may use

$$\widehat{H} = \min\left\{t, \left\lceil \frac{n+4+r_\rho}{\gamma} \right\rceil\right\}. \tag{54}$$

Whenever a reset occurs, $e^{-\gamma\widehat{H}}\rho \le 2^{-n-4}$. The extra unit from rounding preserves the polynomial bound. At dyadic t, the starting point is dyadic; all input evaluation points constructed below lie within the retained interval. Notice that

$$\widehat{H} \le 1 + \gamma^{-1}\big(n + 4 + O(\log K + \log M + \log(1+\gamma^{-1}))\big)$$

unless it has been truncated by t. Thus the retained length is polynomial in the resources of the theorem even when t is exponentially large in its bit length.

C.2.2 Coefficient perturbation and a precision budget

Use componentwise versions of Lemma B.1 to obtain piecewise rational polynomials $\widetilde{A}, \widetilde{b}$ satisfying operator/vector errors at most $\delta \le 1$, and evaluate the terminal $C_x(t)$ to the same tolerance. Increasing component precision by $O(\log(r+m+1))$ achieves the norm error. In particular, $\left\|\widetilde{A}\right\| \le M+1$ and $\left\|\widetilde{b}\right\| \le M+1$.

Let $\widetilde{h}$ be the exact solution for the polynomial coefficients and the same retained initial state. Set $e = \widetilde{h} - h_*$. Then

$$e' = \widetilde{A}e + (\widetilde{A} - A_x)h_* + (\widetilde{b} - b_x), \qquad e(a) = 0.$$

Gronwall's inequality gives, throughout the window of length $H = \widehat{H}$,

$$\|e\| \le \delta H(D+1)e^{(M+1)H} =: E_c. \tag{55}$$

For an approximate readout $\widetilde{C}$, with $\left\|\widetilde{C} - C_x(t)\right\| \le \delta$, the readout error due to coefficients and this approximation is at most

$$ME_c + \delta(D + E_c).$$

Choose $\delta = 2^{-p}$ so that

$$2^{-p} \le \min\left\{\frac{1}{2}, \frac{2^{-n-6}}{(M+1)(H+1)(D+1)e^{(M+1)H}}\right\}. \tag{56}$$

This ensures $E_c \le 1$ and bounds the displayed readout error by less than 2^{-n-4}. Such an integer p can be chosen with

$$p = n + O\big((M+1)H + \log(M+1) + \log(H+1) + \log(D+1) + 1\big). \tag{57}$$

Since $\log D = O(\log K + \log M + \log(1+\gamma^{-1}))$, this precision is polynomial in the specified resources, even if D itself is exponentially large. The number of coefficient sample points is polynomial by Lemma B.1; the input point costs follow by the assumed local coefficient realizers.

C.2.3 Solving the piecewise-polynomial linear equation

Refine the coefficient cells so each new subinterval has length at most $1/[4(M+1)]$. Restriction of a polynomial to a smaller dyadic interval involves rational affine substitution and requires no input queries. The number of refined cells, J, is polynomial in H, M, R, p and the dimensions. On a cell $[a_j, a_{j+1}]$, define

$$(\mathcal{P}z)(u) = \eta_j + \int_{a_j}^{u} \big(\widetilde{A}(v)z(v) + \widetilde{b}(v)\big)\, dv,$$

where η_j is the approximate state carried into the cell. In the uniform norm, $\mathcal{P}$ is a contraction with factor at most $1/4$. Starting with the constant polynomial $z_0 = \eta_j$ and iterating $z_{k+1} = \mathcal{P}z_k$ gives

$$\|z_k - z_*\|_\infty \le \frac{4^{-k}}{1 - 1/4}\, \|z_1 - z_0\|_\infty\,. \tag{58}$$

Here z_* is the exact cell solution with initial value η_j. A coarse bound for exact or carried states is polynomial in $D+1$ times $e^{(M+1)H}$; its *logarithm* is $O(\log(D+1) + (M+1)H + \log(H+1))$.

An initial-state perturbation is amplified across the full retained interval by at most $e^{(M+1)H}$. Allocate terminal state error

$$\tau = \frac{2^{-n-5}}{M+1}$$

and local truncation error at most $\tau e^{-(M+1)H}/(J+1)$ per cell. Summing the propagated local errors bounds the total by τ. Equation (58) achieves this with

$$k = O\big(n + (M+1)H + \log(D+1) + \log(J+1) + \log(M+1) + 1\big)$$

iterations per cell.

If the coefficient degree is at most d_0, then $\deg z_k \le L := k(d_0+1)$. Work in local coordinates on $[0,1]$, absorbing each cell length into its coefficients. Let Q be a common denominator of all transformed coefficients and the retained initial state; $\log Q$ is polynomial because there are polynomially many rationals of polynomial bit length. Each Picard step introduces at most one factor Q and integration divisors at most L, so all iterate and carried-state denominators divide $Q(QL!)^{Jk}$. Their bit lengths are thus $O((Jk+1)\log Q + JkL\log(L+1))$. In the coefficient ℓ^1 norm, multiplication is submultiplicative and integration does not increase the norm; iterating the resulting affine bound also gives polynomial numerator bit lengths. Each iteration performs polynomially many rational operations, so exact rational Picard iteration has polynomial bit cost in $J, k, d_0, p, \operatorname{size}\theta$ and endpoint lengths.

Multiplying the resulting state approximation by $\widetilde{C}$ contributes at most $(M+1)\tau$ additional error. Combining reset error, coefficient/readout perturbation, Picard truncation, and final rounding gives total error below 2^{-n}. Every count and bit length is polynomial in (17), with oracle costs represented by the corresponding second-order polynomial. This completes the constructive evaluation proof.

C.2.4 Output modulus and a sufficient stability structure

Differentiating $y = C_x h$ yields

$$\|y'\| \le \|C_x'\|\,\|h\| + \|C_x\|\,(\|A_x\|\,\|h\| + \|b_x\|) \le MRD + M(MD+M).$$

Combining this estimate with $\|y\| \le MD$ gives a modulus and amplitude bound whose descriptions have polynomial length in the stated resources. A requested output-name query can be answered by the evaluation algorithm, so the result is a function-space operator theorem.

To check the sufficient stability condition stated in the main text, a homogeneous solution satisfies

$$\frac{d}{dt}\|h\|^2 = h^\top(A_x + A_x^\top)h \le -2\gamma\|h\|^2.$$

Integration gives $\|\Phi_x(t,u)\| \le e^{-\gamma(t-u)}$. For example, a diagonal transition $A_\theta(x) = -\operatorname{diag}(\gamma + a_1(x), \dots, \gamma + a_r(x))$ with nonnegative a_i satisfies this condition. This explicit stability structure, combined with smoothness and computability certificates for a_i, b_θ, and C_θ, supplies all ingredients of the evaluation theorem.

C.3 Exact suffix errors and the approximation tradeoff

C.3.1 Exact suffix recovery errors

Let $\mathcal{B}_T = \{x \in C([0,T]) : \|x\|_\infty \le 1\}$. For an arbitrary, even computationally unbounded map G of the complete suffix, define

$$E_{\mathsf{F}}(T,H) = \inf_G \sup_{x\in\mathcal{B}_T} \left|(\mathsf{F}x)(T) - G(x|_{[T-H,T]})\right|, \qquad 0 \le H \le T. \tag{59}$$

For a nonnegative continuous kernel w, the linear functional $F_w(x) = \int_0^T w(u)x(u)\,du$ has

$$E_{F_w}(T,H) = \int_0^{T-H} w(u)\,du. \tag{60}$$

The upper bound follows by integrating only the known suffix and substituting zero for the unknown prefix. For the lower bound assume $T - H > 0$. For small $\eta > 0$, construct f_η equal to one on $[0, T-H-\eta]$, linearly decreasing to zero by $T-H$, and identically zero afterward. Then $\pm f_\eta \in \mathcal{B}_T$ have the same suffix, and their outputs converge to the positive and negative values of the right side of (60) as $\eta \downarrow 0$. A single suffix-based output must incur at least their half-separation in the worst case. Taking the limit proves the lower bound. When $H = T$ both sides are zero. The continuous inputs f_η approach the supremum, establishing the exact error on the stated class.

Taking $w = 1/T$ and $w = e^{-(T-u)}$ yields (3). For U, error at most ε requires $H \ge (1-\varepsilon)T$. For S_0, the exact condition is $e^{-H} - e^{-T} \le \varepsilon$, and $H \ge \log(1/\varepsilon)$ suffices whenever that length is available. These equalities characterize memory on the continuous unit ball and complement the sampling laws on quantitative Gevrey classes.

C.3.2 Proof of the stability-memory approximation tradeoff

For Theorem 5.1, choose the bump (42) for the fixed $s > 1$, set $\beta_s := \beta = \int_0^1 b(u)\,du$, and use R_s from (43). For even T, the inputs

$$x_\pm(u) = \pm \sum_{j=0}^{T/2-1} b(u-j)$$

have all derivatives bounded by the same $R_s^k(k!)^s$ and agree identically on $[T/2, T]$. Uniform attention maps them to $\pm\beta/2$.

The two SSM trajectories may differ at $T/2$ by as much as $2D$, but the coefficients A_x, b_x, C_x are identical on their common suffix because they depend pointwise on the input. The difference therefore solves the same homogeneous equation on that suffix. Its terminal output magnitude is bounded by

$$|(\mathsf{F}_\theta x_+)(T) - (\mathsf{F}_\theta x_-)(T)| \le 2MKDe^{-\gamma T/2}.$$

If the uniform approximation error is η, the triangle inequality gives $\beta \le 2\eta + 2MKDe^{-\gamma T/2}$, which is (20). Requiring $\eta \le \beta/4$ forces $MKDe^{-\gamma T/2} \ge \beta/4$, yielding the stated logarithmic condition. The necessary resource condition holds for every state dimension and for general input-dependent coefficients satisfying the stated norm and stability bounds.

How memory and regularity enter the construction. The pointwise coefficient and readout assumptions ensure that a common suffix drives the same homogeneous state-difference equation, making transition decay the controlling resource. An integrator implements the uniform average through persistent state accumulation. The Gevrey regime $s > 1$ supplies compact bumps with identical suffixes, while the analytic deterministic construction uses trigonometric functions that agree at the finitely many observed points. Each construction captures the information supplied by its corresponding interface.

D Discrete precision locality: complete proof and architecture connection

Section 6 transfers precision locality to arbitrary bounded token sequences. Here we prove Corollary 6.1 directly from the selective-scan recurrence and softmax weights, then provide the finite-bit implementation. The final subsections identify the Mamba certificate, account for local preprocessing and readout operations, extend the construction to Mamba-3, and relate contraction in physical time to context length in tokens.

D.1 Discrete access model and suffix information

Fix a bounded token set $\mathcal{V} \subset \mathbb{R}^d$ and sequences $u = (u_1, \dots, u_N) \in \mathcal{V}^N$. The endpoint $N \ge 1$ is binary encoded. Define $R_H u = (u_{N-H+1}, \dots, u_N)$ for integer $1 \le H \le N$ and let $R_0 u$ be empty. The discrete recovery error is

$$E_F(N, H) = \inf_G \sup_{u \in \mathcal{V}^N} |F_N(u) - G(N, R_H u)|\,. \tag{61}$$

The input-dependent information available to G consists exactly of $R_H u$; fixed model parameters and any stipulated data-independent positional channel are also known. Full information gives $E_F(N, N) = 0$. The radius is the minimum integer H for which this error is at most 2^{-n}.

For the finite-bit assertion, a token query $(i, 1^p)$ returns a vector within 2^{-p} of u_i and with encoding length $O_d(p+1)$ for a fixed bounded token set. This is the finite-index version of the bounded-size names in Appendix A. Requests, replies, scalar arithmetic, and local coefficient-realizer work are charged. At variable dimensions the usual $O(\log d)$ guard bits convert componentwise to Euclidean error. If an arbitrary padded token name is used instead, its size function must be included in the second-order bound. For explicit arrays, initial reading and indexing contribute their ordinary access costs; the corollary measures evaluation with the stipulated random-access interface.

D.2 THE RESET IDENTITY FOR SELECTIVE SCANS

All quantities $\bar{A}_i, b_i, c_i, d_i$ in (21) are functions of the current block input u_i and known parameters, possibly also of the known index i. Let $D_d = B_d/(1-\varrho)$. Induction yields

$$\|h_i\| \le B_d \sum_{j=0}^{i-1} \varrho^j = D_d(1-\varrho^i) \le D_d. \tag{62}$$

For $a = N - H$, let $z_a = 0$ and run the same recurrence on the retained indices $a+1, \dots, N$. These transitions and forcings are known from the suffix. Subtracting the two recurrences gives

$$h_N - z_N = \bar{A}_N \bar{A}_{N-1} \cdots \bar{A}_{a+1} h_a. \tag{63}$$

Since u_N is retained when $H \ge 1$, the terminal readout and direct term are also known. The explicit recovery map $G = c_N^\top z_N + d_N$ therefore has error at most $C_d D_d \varrho^H$. This proves (22). If the integer in (23) is at most N, its untruncated value satisfies $\mathcal{R}_d \varrho^{H_n} \le 2^{-n}$. If it exceeds N, use all inputs and the known $h_0 = 0$ for zero information error. This proves the capped radius bound in all cases, including $B_d = 0$ or $C_d = 0$.

More generally, if a supplied certificate is $\|\bar{A}_i\| \le e^{-\lambda_i}$ with $\lambda_i \ge 0$, the identical argument gives the suffix-specific bound

$$|F_N(u) - G(N, R_H u)| \le C_d D_d \exp\left(- \sum_{i=N-H+1}^{N} \lambda_i \right), \tag{64}$$

whenever D_d bounds the possible incoming prefix state. Here that state bound is supplied by the uniform $\varrho < 1$ certificate. An adaptive implementation can inspect tokens backward until certified lower bounds for the sum exceed $n \log 2 + \log \mathcal{R}_d$. Rounding numerical exponents downward and the error budget upward preserves the certificate. The fixed uniform window provides a direct alternative to adaptive stopping.

D.3 SOFTMAX ATTENTION: AN EXACT PREFIX-WEIGHT FORMULA

Fix a known logit row and vary its scalar values independently in $[-1, 1]$. The output is

$$\mathrm{Att}_N(v) = \sum_{j=1}^{N} \alpha_{Nj} v_j, \qquad \alpha_{Nj} > 0, \quad \sum_j \alpha_{Nj} = 1.$$

The sum over known suffix values has worst-case error at most $m_H = \sum_{j=1}^{N-H} \alpha_{Nj}$. For the opposite inequality, give the recovery map a suffix consisting entirely of zeros. The two prefixes consisting entirely of $+1$ and -1 produce outputs $+m_H$ and $-m_H$. The maximum error of any single returned number is therefore at least m_H. Thus $E_{\mathrm{Att}}(N, H) = m_H$, including $H = 0, N$. This lower bound already uses the three-symbol value alphabet $\{-1, 0, 1\}$.

If all logits vanish, the weights are $1/N$, proving (25). Moreover, on the zero transcript of any deterministic algorithm that queries q distinct tokens, assign $+1$ or -1 to every unqueried token. The two possible outputs have magnitude $(N-q)/N$. Hence at least $\lceil (1-2^{-n})N \rceil$ distinct token observations are necessary for arbitrary query locations and exact returned values. The bound therefore characterizes information acquisition across all deterministic point-access algorithms.

For $|\ell_{Nj}| \le B_a$, each $\alpha_{Nj} \ge e^{-2B_a}/N$, so

$$E_{\mathrm{Att}}(N, H) \ge e^{-2B_a} \frac{N-H}{N}. \tag{65}$$

At fixed B_a this implies $H \ge (1 - e^{2B_a} 2^{-n})N$ whenever the right side is positive. This gives a standard attention witness with nonconstant logits and a quantitative history requirement controlled by B_a.

To realize such a witness by standard projections (Vaswani et al., 2017), use tokens $u_j = (z_j, v_j)$, where $z_j = 1 + \frac{1}{2}\sin j$ is a known, bounded positional channel. Choose $W_Q = W_K = (1, 0)$ and

$W_V = (0, 1)$, with query dimension one. The score at the endpoint is $\ell_{Nj} = z_N z_j$, bounded in absolute value by $9/4$, and it is nonconstant in j. Values vary freely without changing the logits. The scale factor $\sqrt{d_k}$ is one in this realization. The positional sequence is fixed independently of the freely varying values. The uniform case instead uses $W_Q = W_K = 0$ and the identity value projection. Both constructions use ordinary causal softmax heads.

D.4 Finite-bit execution with a uniform error budget

For the selective scan, write $D = D_d$, $C = C_d$, and $\varepsilon = 2^{-n}$. First retain a suffix with reset error at most $\varepsilon/4$, for instance the ideal length H_{n+2}. A conservative rounded window of $O(H_{n+2})$ tokens suffices as well. Indeed, let $a = -\log \varrho > 0$ and $U = (n+2)\log 2 + \log \mathcal{R}_d$. Computable rational bounds $a/2 \le \underline{a} \le a$ and $U \le \overline{U} \le U + a/2$ give the safe length

$$\widehat{H} = \min\{N, \max\{1, \lceil \overline{U}/\underline{a} \rceil\}\}.$$

It is at most a constant multiple of $H_{n+2} + 1$. Approximating logarithms to produce these bounds has polynomial cost in the bit encodings and $\log(1/(1-\varrho))$. These rational enclosures make window selection computable with a certified rounding margin. A rational certified upper bound on ϱ may be used throughout.

Let z_i be the exact reset trajectory; it also satisfies $\|z_i\| \le D$. Obtain approximations $\widehat{A}_i, \widehat{b}_i, \widehat{c}_N, \widehat{d}_N$ with operator, vector, or scalar error at most η. Round each state update with norm error at most η, and the final readout with scalar error at most η. If $e_i = \|\widehat{z}_i - z_i\|$, then

$$e_i \le (\varrho + \eta) e_{i-1} + \eta(D+2), \qquad e_{N-\widehat{H}} = 0. \tag{66}$$

This follows by writing $\widehat{A}_i(\widehat{z}_{i-1} - z_{i-1}) + (\widehat{A}_i - \bar{A}_i) z_{i-1} + (\widehat{b}_i - b_i)$ and adding the rounding residual. Choose

$$W = \frac{2(C+1)(D+2)}{1-\varrho} + D + 2, \tag{67}$$

$$\eta \le \min\left\{1, \frac{1-\varrho}{2}, \frac{2^{-n-2}}{W}\right\}. \tag{68}$$

Then $\varrho + \eta \le (1+\varrho)/2 < 1$ and the geometric series in (66) gives

$$e_N \le \frac{2\eta(D+2)}{1-\varrho}.$$

The readout error is bounded by

$$(C+\eta) e_N + \eta D + 2\eta \le W\eta \le \varepsilon/4.$$

With reset error included, the final error is at most $\varepsilon/2$, hence within ε. In particular, the required number of accuracy bits is

$$p = n + O\left(1 + \log(1+C) + \log(1+D) + \log \frac{1}{1-\varrho}\right), \tag{69}$$

plus dimension guard bits. The numerical-error budget is uniform in N. State and readout integer parts require their ordinary magnitude bits, including the local direct term d_N; these are charged by the local-realizer resources.

There are $O(H_{n+2})$ coefficient evaluations and matrix-vector updates. With fixed dimensions, magnitude and contraction certificates, and local polynomial-time coefficient realizers, this is polynomial bit work in $n + \mathrm{len}(N)$ with the stated bounded-size token names. For variable certificates the displayed window and precision formulas, the dimension, parameter encoding lengths, and coefficient-evaluation costs must all be charged. The formulas quantify how the contraction margin $1 - \varrho$ controls both the retained window and the working precision. This proves the computational assertion of Corollary 6.1.

D.5 Direct identification with a Mamba selective scan

The selective recurrence in Gu & Dao (2023, Equations (2)-(4), Algorithm 2) has the form

$$h_i = \overline{A}_i h_{i-1} + \overline{B}_i u_i, \qquad y_i = C_i h_i,$$

with input-dependent discretization steps, forcing, and readout. In the real negative-diagonal parameterization, after flattening independent channels and state coordinates, the transition entries are

$$(\overline{A}_i)_{kk} = \exp(-a_k \Delta_{ik}), \qquad a_k > 0, \quad \Delta_{ik} > 0. \tag{70}$$

Per-channel broadcasting is included in this notation. Thus $b_i = \overline{B}_i u_i$, and the Euclidean operator norm is $\max_k \exp(-a_k \Delta_{ik})$. Uniform bounds $a_k \geq a > 0$, $\Delta_{ik} \geq \delta > 0$ give $\left\|\overline{A}_i\right\| \leq e^{-a\delta} < 1$. The argument applies directly to arbitrary tokens in the bounded input domain.

The certificate bounds the *actual* forcing b_i and readout, accommodating both zero-order-hold forcing and other local discretizations. Under these certificates, (64) with $\lambda_i = \min_k a_k \Delta_{ik}$ proves (26). The exact diagonal product can provide the sharper exponent $\min_k \sum_{i=N-H+1}^{N} a_k \Delta_{ik}$, but the sum-of-minima bound is sufficient for a scalar certificate and stopping rule.

Quantitative margins for positive gates. For fixed affine maps on a fixed compact token domain, a softplus step has a positive minimum. Certifying this minimum makes the dependence on the learned parameters explicit. The margin $a\delta = -\log \varrho$ then converts the accuracy target into a sufficient token count. Equivalently, cumulative contraction serves as the discrete memory coordinate, giving an input-adaptive counterpart to elapsed physical time.

Local preprocessing, gates, and residuals. If the coefficient maps at index i use the last w raw tokens rather than just u_i, retain $H + w - 1$ raw tokens to compute the last H coefficient tuples exactly. The same reset proof then applies, with N as the cap. This covers a finite-width causal preprocessing convolution once its coefficient and magnitude bounds are established. A bounded pointwise output gate and output projection can be included in the readout vector c_N; a known current-token residual or direct path can be included in d_N. A locally Lipschitz post-map multiplies the error budget by its certified Lipschitz constant. For fixed-depth compositions, layerwise regularity, magnitude, and stability certificates connect these block guarantees to the intermediate signals, as described after Theorem 4.2.

D.6 Mamba-3: rotations, two-token forcing, and MIMO readouts

The local preprocessing argument also accommodates Mamba-3's exponential-trapezoidal update: its forcing uses the current and previous block inputs. In real coordinates, the complex recurrence in Lahoti et al. (2026, Appendix B.3, Eq. (25)), combined with their MIMO construction, takes the form

$$S_t = \alpha_t R_t S_{t-1} + \beta_t R_t V_{t-1} + \gamma_t V_t, \qquad V_t = B_t X_t^\top, \tag{71}$$

$$\alpha_t = e^{-a_t \Delta_t}, \qquad \beta_t = (1 - \lambda_t)\Delta_t \alpha_t, \qquad \gamma_t = \lambda_t \Delta_t. \tag{72}$$

Here $S_t \in \mathbb{R}^{m \times q}$, $B_t \in \mathbb{R}^{m \times r}$, $X_t \in \mathbb{R}^{q \times r}$, and $R_t \in \mathbb{R}^{m \times m}$ is orthogonal. The rank-one case includes the SISO recurrence. The positive decay magnitude is $a_t = -A_t$, where A_t is the real scalar coefficient for the head in the cited recurrence. All coefficients at index t depend only on u_t and fixed parameters; $S_0 = 0$ and $V_0 = 0$ are prescribed. We use the Frobenius norm for states and write the terminal output as $F_N = \mathcal{O}_{u_N}(S_N)$, allowing the readout to include local gates, projections, and direct paths. For vector outputs, the recovery error in (61) uses the output norm in place of absolute value.

The following is the full statement of Corollary 6.2.

Corollary D.1 (Certified Mamba-3 context)**.** *Suppose, uniformly over the allowed block inputs,*

$$a_t \geq a_* > 0, \qquad 0 < \delta \leq \Delta_t \leq \bar{\Delta}, \qquad 0 \leq \lambda_t \leq 1, \qquad \|V_t\|_F \leq M, \tag{73}$$

and the readout satisfies

$$\left\|\mathcal{O}_u(S) - \mathcal{O}_u(\widetilde{S})\right\| \leq C_{\text{out}} \left\|S - \widetilde{S}\right\|_F.$$

Set

$$\rho = e^{-a_*\delta} < 1, \qquad D_* := \frac{M\bar{\Delta}}{1 - e^{-a_*\bar{\Delta}}} \le M(\bar{\Delta} + a_*^{-1}). \tag{74}$$

Then $\|S_t\|_F \le D_*$ *for every* t. *For* $1 \le \ell < N$, *reset the state at* $s = N - \ell$ *to zero and evaluate the last* ℓ *updates using the* $\ell + 1$ *inputs* $u_s, \dots, u_N$. *The resulting output* $\widetilde{F}_N$ *satisfies*

$$\left\| F_N - \widetilde{F}_N \right\| \le C_{\text{out}} D_* \exp\left(- \sum_{t=N-\ell+1}^{N} a_t \Delta_t \right) \le C_{\text{out}} D_* \rho^{\ell}. \tag{75}$$

Consequently, for $n \ge 1$, *the precision-local radius obeys*

$$H_F(N, n) \le \min\{N, 1 + \ell_n\}, \qquad \ell_n := \max\left\{ 1, \left\lceil \frac{\log \max\{1, 2^n C_{\text{out}} D_*\}}{a_* \delta} \right\rceil \right\}. \tag{76}$$

If the cap is N, *evaluate the full recurrence from its prescribed initial state.*

Proof. Orthogonality gives $\|\alpha_t R_t\|_2 = \alpha_t$. The two forcing coefficients are nonnegative and satisfy

$$\beta_t + \gamma_t = \Delta_t\big[(1 - \lambda_t)\alpha_t + \lambda_t\big] \le \Delta_t,$$

so $\|S_t\|_F \le \alpha_t \|S_{t-1}\|_F + M\Delta_t$. The function $x/(1 - e^{-a_* x})$ is increasing on $(0, \infty)$: the numerator of its derivative is $1 - (1 + a_* x)e^{-a_* x} \ge 0$. Hence

$$M\Delta_t \le D_*(1 - e^{-a_t \Delta_t}) = D_*(1 - \alpha_t).$$

Induction from $S_0 = 0$ proves the state bound. The upper estimate for D_* follows from $1/(1 - e^{-x}) \le 1 + 1/x$ for $x > 0$. In particular, this certificate does not depend on the minimum step δ.

Let $\widetilde{S}_s = 0$. The boundary input u_s supplies V_s, so every forcing term used in the reset scan agrees with the full recurrence. Subtraction yields

$$S_N - \widetilde{S}_N = (\alpha_N R_N) \cdots (\alpha_{s+1} R_{s+1}) S_s.$$

A product of orthogonal matrices is orthogonal, giving the exact identity

$$\left\| S_N - \widetilde{S}_N \right\|_F = \exp\left(- \sum_{t=s+1}^{N} a_t \Delta_t \right) \|S_s\|_F .$$

The readout bound proves (75). Taking $\ell = \ell_n$ makes its right-hand side at most 2^{-n} whenever a proper suffix is used; retaining the full input gives zero information error otherwise. This proves (76), including $N = 1$ and $C_{\text{out}} D_* = 0$. □

The sum in (75) also supplies an adaptive rule: read backward for at least one update and until certified lower bounds for $\sum a_t \Delta_t$ reach $\log \max\{1, 2^n C_{\text{out}} D_*\}$, retaining one further boundary input. If the beginning is reached, use the known initial state. The cumulative rotary formulation has prefix-dependent transformed B_t, C_t. The suffix-recovery proof therefore uses the equivalent recurrence with token-local rotations in real coordinates.

Constructing the certificates from parameters. The hypotheses can be checked through parameter bounds on a specified input domain. For example, on $\|u\|_2 \le U$, the step map $\Delta(u) = \text{softplus}(w_\Delta^\top u + b_\Delta)$ admits

$$\delta = \text{softplus}(b_\Delta - \|w_\Delta\|_2 U), \qquad \bar{\Delta} = \text{softplus}(b_\Delta + \|w_\Delta\|_2 U), \tag{77}$$

where $\text{softplus}(z) = \log(1 + e^z)$. A positive enforced floor $a(u) \ge a_{\text{floor}} > 0$ supplies $a_* = a_{\text{floor}}$, and a sigmoid gate supplies $0 < \lambda(u) < 1$. These parameterizations occur in the released implementation.[1]

[1] The Mamba-3 implementation uses $a(u) = \max\{f(w_a^\top u), a_{\text{floor}}\}$, where $f(z) = 1 + z$ for $z \ge 0$ and $f(z) = 1/(1 - z)$ for $z < 0$, with default $a_{\text{floor}} = 10^{-4}$. Equation (77) gives runtime step-size bounds for the fixed parameters and input domain.

If $\|B_t\|_2 \le B_*$ and $\|X_t\|_F \le X_*$, then $M = B_* X_*$ is valid for every MIMO rank. Such bounds include the normalization gains and learned biases. Indeed, for columnwise RMS normalization

$$\mathrm{RMS}_{g,\epsilon_0}(z) = \frac{g \odot z}{\sqrt{\|z\|_2^2/m + \epsilon_0}}, \qquad \epsilon_0 > 0,$$

we have $\|\mathrm{RMS}_{g,\epsilon_0}(z)\|_2 \le \sqrt{m}\,\|g\|_\infty$. Thus a matrix with r normalized columns and subsequent bias b_B satisfies

$$\|B_t\|_2 \le \|B_t\|_F \le \sqrt{mr}\,\|g_B\|_\infty + \|b_B\|_F ;$$

the analogous estimate supplies a bound C_* on $\|C_t\|_2$. If $X(u) = \mathrm{reshape}(P_X u + b_X)$, where P_X includes the MIMO expansion, take $X_* = \|P_X\|_2 U + \|b_X\|_2$.

For a gated readout

$$\mathcal{O}_u(S) = P\big[(S^\top C(u)) \odot g(u)\big] + d(u),$$

with matrix arguments of P vectorized, the certificate is $C_{\mathrm{out}} = \|P\|_2 C_* G_*$ whenever $\sup_u \|g(u)\|_\infty \le G_*$. For $g(u) = \mathrm{SiLU}(P_Z u + b_Z)$, the inequality $|\mathrm{SiLU}(z)| \le |z|$ gives $G_* = \|P_Z\|_2 U + \|b_Z\|_2$. An additional RMS normalization contributes a multiplicative factor of at most $\|g\|_\infty / \sqrt{\epsilon_0}$ to the Lipschitz gain, as follows by differentiating the displayed normalization map. A shared current-input direct term cancels from the difference at its insertion point; any subsequent normalization is covered by its Lipschitz bound. These constructions provide certificate values specific to the block's parameters and the chosen input domain.

Numerical precision and block scope. Vectorizing S_t identifies the transition as $I_q \otimes (\alpha_t R_t)$ and the forcing as $\beta_t R_t V_{t-1} + \gamma_t V_t$. Their locality and norms allow the finite-bit argument of Appendix D.4 to apply with $D = D_*$. For completeness, this also covers a nonlinear terminal readout with the displayed Lipschitz certificate. Set $\varepsilon = 2^{-n}$ and use (76) at accuracy index $n+1$ to make the truncation error at most $\varepsilon/2$. If transition and forcing approximations, and each rounded state update, have norm error at most η, then

$$e_t \le (\rho + \eta) e_{t-1} + \eta(D_* + 2), \qquad e_N \le \frac{2\eta(D_* + 2)}{1 - \rho} \quad \text{when } \eta \le (1-\rho)/2.$$

The exact reset trajectory obeys the same invariant state bound as the full trajectory. Evaluate the terminal readout, including its dependence on u_N, to error at most η. With

$$W_* = 1 + \frac{2C_{\mathrm{out}}(D_* + 2)}{1 - \rho}, \qquad \eta \le \min\left\{1, \frac{1-\rho}{2}, \frac{\varepsilon}{2W_*}\right\}, \tag{78}$$

the numerical output error is at most $C_{\mathrm{out}} e_N + \eta \le \varepsilon/2$. The total error is therefore at most ε. Sufficient coefficient accuracy is

$$p = n + O\left(1 + \log(1 + C_{\mathrm{out}}) + \log(1 + D_*) + \log\frac{1}{1-\rho}\right),$$

plus dimension guard bits. Outward-rounded rational bounds and conservative window selection give a computable, certified suffix length (Appendix D.4). With fixed dimensions, certificate values, and local polynomial-time coefficient and readout realizers, the resulting evaluator runs in $\mathrm{poly}(n, \mathrm{len}(N))$ bit time under the stated token-access model. Variable dimensions, parameter encodings, magnitude bits, and realizer costs remain charged resources.

The construction applies to one head on its supplied block-input sequence. For finitely many heads, concatenate their states, use $D = (\sum_j D_{*,j}^2)^{1/2}$ and $\rho = \max_j \rho_j < 1$, and certify the joint output map. The same reset and finite-bit arguments then apply to the block. Evaluation from the original inputs of a multilayer model additionally requires layerwise certificates and propagation of intermediate errors, as for the compositions discussed above.

D.7 An exact discrete-continuous identity and sharp token units

The preceding certificates give sufficient windows for general selective blocks. For the normalized scalar scan

$$h_i = \varrho h_{i-1} + (1 - \varrho) v_i, \qquad h_0 = 0, \quad |v_i| \le 1,$$

unrolling gives weights $(1-\varrho)\varrho^{N-i}$. The same missing-mass argument therefore gives, for every $0 \le H \le N$,

$$E_{\mathsf{S}_\varrho}(N,H) = (1-\varrho)\sum_{i=1}^{N-H}\varrho^{N-i} = \varrho^H - \varrho^N. \tag{79}$$

This identity establishes the sharp inverse log-contraction dependence. If N is long enough that $\varrho^N \le 2^{-n}$, the necessary radius lies between $((n-1)\log 2)/(-\log\varrho)$ and $n\log 2/(-\log\varrho)$, up to integer rounding. The constants may be adjusted for the endpoint and for small n.

Take a sampling interval $\Delta > 0$, put $T = N\Delta$, and extend v_i as a piecewise-constant input on $((i-1)\Delta, i\Delta]$. Solving $h' = -\gamma h + \gamma x$ exactly on each interval gives

$$h(i\Delta) = e^{-\gamma\Delta}h((i-1)\Delta) + (1-e^{-\gamma\Delta})v_i.$$

Hence (79) equals $e^{-\gamma H\Delta} - e^{-\gamma N\Delta}$ for $\varrho = e^{-\gamma\Delta}$. This exact identity connects finite sequences to piecewise-constant inputs. An $O(n/\gamma)$ window in physical time corresponds to $O(n/(\gamma\Delta))$ tokens, giving an $O(n)$ token bound at a fixed per-token contraction margin. The learned steps in (26) provide the corresponding adaptive conversion.

Precision-aware context selection. At fixed contraction and magnitude certificates, one endpoint of a selective scan can be recovered using a precision-dependent suffix with polynomial bit overhead. A standard uniform attention head has an exact suffix-length requirement of $\lceil(1-2^{-n})N\rceil$, already on a finite value alphabet. The corollary connects accuracy targets directly to context selection and quantifies the input-access requirements within the polynomial-time evaluation guarantees in (N,n) shared by both blocks under the stated arithmetic and magnitude resources.