Mamba as Measure-Valued Associative Memory: Infinite-Context Limits and Minimax-Optimal Learning

Zeyu Bao
zeyu@u.nus.edu
National University of Singapore

Tianyao Zhang
e0902048@u.nus.edu
Dunhe Fund

Abstract

As sequence models emerge as efficient architectures for long-context modeling, it becomes important to understand whether state-space models are capable of associative recall. We study the recall-predict problem, where a context is a mixture of tagged probability measures and a query specifies the component whose content distribution determines the response. First, focusing on two-layer Mamba models, we introduce the query insertion encoding and show the existence of an infinite-context measure-valued Mamba limit. Under separated tags and exponential decay assumptions, we study trained Mamba hypothesis classes and prove that approximate empirical risk minimization over these classes yields estimators with the population-risk bound at a sub-polynomial rate. Finally, we complement this upper bound with an architecture-independent minimax lower bound of comparable order, demonstrating that the exponent is statistically optimal. These results extend measure-level associative-memory theory beyond attention mechanisms and identify query insertion, recurrent stability and spectral effective dimension as the key mechanisms enabling optimal learning from infinite contexts.

1. Introduction

Modern sequence models are increasingly expected to reason over contexts whose length is large, variable, and often not naturally bounded. In such regimes, a context is better viewed not merely as a finite list of tokens, but as an empirical sample from an underlying distribution of information. This perspective is especially natural for document collections, retrieval-augmented systems and in-context learning tasks, where a model must identify the component of a large context relevant to a query and then compute a prediction from the selected component.

Transformers have provided the dominant mechanism for this type of content-addressable computation. Their attention layers can be interpreted as associative-memory modules: a query selects values from a context through similarity scores and the output is a weighted aggregation of the retrieved content (Vaswani et al., 2017; Ramsauer et al., 2020). Recent theoretical work has sharpened this viewpoint by studying memory capacity, factual recall, and the emergence of associative structure in Transformers (Bietti et al., 2023; Cabannes et al.; Kim et al., 2023; Mahdavi et al.; Jiang et al., 2024; Nichani et al., 2024).

A complementary line of work studies sequence models on infinite-dimensional inputs. When a finite context is interpreted through its empirical measure, the large-context limit becomes a problem about maps on probability measures. For attention-based models, this has led to measure-theoretic formulations of self-attention, mean-field descriptions of attention dynamics, and universality or generalization results for distributional inputs (Geshkovski et al., 2024; Furuya et al.).

However, the theoretical understanding of measure-level associative memory is still highly architecture-dependent. Existing sharp analyses are largely attention-based: softmax attention has an explicit query–key mechanism that can concentrate weight on the queried component. By contrast, Mamba and related selective state-space models process a sequence through recurrent state updates rather than through all-pairs attention (Gu & Dao, 2024; Gupta et al., 2022; Gu et al., 2021). This distinction is central. Mamba is attractive for long contexts because its computation scales linearly in

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

sequence length, but it is not obvious whether a recurrent selective state-space architecture can realize the same kind of query-conditioned associative recall that attention implements directly. Moreover, if the context is represented by a probability measure rather than a fixed finite sequence, one must first ask whether the recurrent computation even has a well-defined infinite-context limit.

This motivates our central question: can the associative-memory behavior be realized by a Mamba model, when contexts are infinite-dimensional probability measures rather than finite token sequences? In our paper, we develop a statistical theory of measure-valued associative memory for Mamba. Our main contributions are presented as follows:

- We formulate a two-layer Mamba student model for the recall-predict task and show the existence of a limit for the measure-valued Mamba operator induced by the recurrent dynamics.
- We show that a two layer Mamba model can learn the recall-predict mapping at the level of measures and we establish the sub-polynomial population risk upper bound controlled by the kernel decay.
- We establish a matching minimax lower bound over a structured coefficient model for the content distributions, which means that our population rate is statistically optimal.

2. Related Work

Associative Memory and Recall. Recent theoretical work studies how associative memory emerges in Transformers, how memory capacity scales with model and sample size and how shallow Transformers can implement factual recall through linear or MLP associative memories (Bietti et al., 2023; Cabannes et al.; Kim et al., 2023; Mahdavi et al.; Nichani et al.). Closest to our setting, measure-theoretic Transformer work formalizes recall from a mixture of probability measures and proves statistical upper and lower bounds for learned softmax attention (Kawata & Suzuki, 2026). Our work is complementary but distinct: instead of using attention as an explicit integral operator, we study whether selective state-space dynamics can realize the same measure-level recall–predict task.

State-Space Models. Structured state-space models such as S4 and Mamba introduce stable and computationally efficient parameterizations for long-range dependencies (Gu et al., 2022; 2021). Recent studies design recurrences that explicitly solve associative recall and analyze whether Mamba-like models can perform in-context learning (Arora et al.; Le Corre et al.; Huang et al., 2025), which mainly focus on finite-token retrieval or algorithmic mechanisms for in-context learning. Specifically, Le Corre et al. gives an optimization dynamics explanation for how simplified Mamba learns recall mechanisms. In contrast, our work focuses on the existence of the infinite-context limit for two-layer Mamba and derives statistically optimal rates for learning recall-predict mappings.

Approximations for Functional Mappings. Our upper-bound analysis relies on approximating nonlinear maps whose inputs are infinite-dimensional functions, distributions or probability measures. Zhou et al. (2024) studied neural-network approximation of RKHS functionals and gave quantitative bounds showing how smoothness and spectral structure control approximation complexity. For the transformer architectures, Furuya et al. proved that attention-based architectures can universally approximate continuous in-context mappings when the context is represented as a probability distribution and Takakura & Suzuki (2023) analyzed approximation and estimation for sequence-to-sequence functions with infinite-dimensional inputs. Our work uses these approximation ideas for the generalization error of Mamba recurrence under the recall-predict setting.

3. Recall–Predict Task with Measure-Valued Contexts

In this section, we formulate a distributional associative-memory problem in which the context is represented as a probability measure over tokens. This measure-valued view is intended to capture the limiting regime of infinite contexts so that we could investigate what can be learned from the information contained in the context itself.

3.1. Recall–Predict Task

The recall–predict task has two stages. First, the query must recall the component of the context associated with a particular tag. Second, after this component has been identified, the learner must predict a scalar functional of the corresponding content distribution. Now we formalize this task as follows:

Let $X_0 \subset \mathbb{R}^{d_2}$ be a compact content domain and define

$$X := \mathbb{R}^{d_1} \times X_0$$

A token is written as $x = (v, z)$, where $v \in \mathbb{R}^{d_1}$ is a tag or document feature and $z \in X_0$ is the associated content variable. Thus the first coordinate provides an address, while the second coordinate carries the information to be used for prediction. For a tag v and a content measure $\mu \in \mathcal{P}(X_0)$, we define the tagged content measure

$$\mu_v := \delta_v \otimes \mu \in \mathcal{P}(X).$$

This construction places all mass on tokens whose tag coordinate is v, while leaving the content coordinate distributed according to μ. In this sense, μ_v is a measure-valued memory cell indexed by the tag v.

Definition 3.1 (Recall-Predict Task)**.** Fix $I \geq 1$. For $i \in [I]$, let

$$v^{(i)} \in \mathbb{S}^{d_1-1}, \quad \mu_i \in \mathcal{P}(X_0), \quad \mu^{(i)}_{v^{(i)}} := \delta_{v^{(i)}} \otimes \mu_i.$$

The context measure is the balanced mixture

$$\nu := \frac{1}{I}\sum_{i=1}^{I} \mu^{(i)}_{v^{(i)}} = \frac{1}{I}\sum_{i=1}^{I} \delta_{v^{(i)}} \otimes \mu_i \in \mathcal{P}(X).$$

For a distinguished index $i_\star \in [I]$, the query is

$$q := v^{(i_\star)}.$$

The ground-truth regression function has recall–predict form if there exists a functional $\widetilde{F}_\star$ such that

$$F_\star(\nu, q) = \widetilde{F}_\star(\mu_{i_\star}, q).$$

Thus the output depends on the context only through the content measure selected by the query tag.

The key point is that the query q does not itself contain the content distribution $\mu_{i_\star}$. It only specifies an address inside the context measure ν. A successful predictor must use q to select the tagged component $\delta_{v^{(i_\star)}} \otimes \mu_{i_\star}$ from the mixture, discard irrelevant components, and then evaluate the functional $\widetilde{F}_\star$ in the recalled content distribution. This separation between ν and q is quite useful. The first is an associative-memory difficulty: the relevant content is hidden inside a mixture and can only be accessed through its tag. The second is a statistical difficulty: the target may be a nonlinear functional of a probability measure, rather than a function of a single token or a finite-dimensional feature vector.

3.2. Statistical Estimation Problem

For the recall-predict task, the statistical question is whether a learned sequence model can recover enough information about $\mu_{i_\star}$ from an arbitrarily long context and then estimate a nonlinear functional of it.

Now, we describe the supervised learning problem induced by the recall–predict model. We observe n independent training examples

$$S_n = \{(\nu_t, q_t, Y_t)\}_{t=1}^{n},$$

where

$$Y_t = F_\star(\nu_t, q_t) + \xi_t, \qquad \xi_t \sim N(0, \sigma^2),$$

and the noises are independent of the inputs. Each pair (ν_t, q_t) is a measure-valued context together with a query tag, and the response Y_t is a noisy observation of the scalar quantity determined by the recalled content distribution.

For a hypothesis class $\mathcal{H}_n$ of measurable functions $F : \mathcal{P}(X) \times \mathbb{R}^{d_1} \to \mathbb{R}$, let $\widehat{F}_n$ be a measurable empirical risk minimizer satisfying

$$\widehat{F}_n = \arg\min_{F \in \mathcal{H}_n} \frac{1}{n}\sum_{t=1}^{n}\big(Y_t - F(\nu_t, q_t)\big)^2$$

The population prediction risk is defined as follows, which measures how accurately the learned predictor recovers the noiseless recall–predict target on unseen contexts.

$$\mathcal{R}(\widehat{F}_n, F_\star) := \mathbb{E}_{(\nu,q)}\left[\left(\widehat{F}_n(\nu,q) - F_\star(\nu,q)\right)^2\right].$$

It evaluates whether $\widehat{F}_n$ has learned both parts of the task: the recall mechanism that selects the query-matched content distribution from ν and the prediction mechanism that estimates the functional of that selected distribution.

4. Two-Layer Mamba Student Model

We now define our student architecture. A selective state-space layer is well suited to the recall-predict setting because it updates a fixed-dimensional hidden state sequentially, with token-dependent transitions and readouts. Several papers have shown the potential of recurrent models, which explicitly solve associative recall and perform in-context learning (Le Corre et al.; Huang et al., 2025; Arora et al.).

4.1. Input Encoding and Query Insertion

The query is inserted twice, once before the context and once after it. The initial query token gives the recurrent state access to the address q before scanning the context, so that subsequent state updates can be query-conditioned. The final query token asks the model to read out the answer after the context has been processed.

Let $\mathcal{U} \subset \mathbb{R}^{d_{\text{in}}}$ be a compact augmented token domain. We use three fixed marker vectors $m_{\text{in}}, m_{\text{ctx}}, m_{\text{out}}$ and define measurable embeddings that make each token identifiable: $\iota_{\text{in}}(q) = (m_{\text{in}}, q, 0),\ \iota_{\text{ctx}}(v,z) = (m_{\text{ctx}}, v, z),\ \iota_{\text{out}}(q) = (m_{\text{out}}, q, 0)$. Given (ν, q), we sample $X_1, \dots, X_T \overset{\text{i.i.d.}}{\sim} \nu, X_t = (V_t, Z_t)$ and form the length-$(T+2)$ input sequence

$$U_0 = \iota_{\text{in}}(q), \quad U_t = \iota_{\text{ctx}}(X_t)(1 \le t \le T), \quad U_{T+1} = \iota_{\text{out}}(q).$$

Thus the context is observed through an exchangeable random sample from ν, while the query is treated as a control variable that conditions both the scan and the final readout.

4.2. S6 Layer

We use a vector-valued selective state-space layer. For an input sequence $u_0, \dots, u_L \in \mathcal{U}$, an S6 layer with hidden width d_h and output width d_{out} is specified by

$$\theta = \left(A, \{B^{(m)}\}_{m=0}^{d_{\text{in}}}, \{C^{(m)}\}_{m=0}^{d_{\text{in}}}, a_\Delta, b_\Delta, h_{\text{init}}\right),$$

where $A \in \mathbb{R}^{d_h \times d_h}$, $B^{(m)} \in \mathbb{R}^{d_h \times d_{\text{in}}}$ and $C^{(m)} \in \mathbb{R}^{d_h \times d_{\text{out}}}$. Each θ induces the sequence-to-sequence mapping

$$\mathsf{S6}_{\theta,L} : \mathcal{U}^{L+1} \to (\mathbb{R}^{d_{\text{out}}})^{L+1}$$

For each token u_ℓ, define

$$B_\theta(u_\ell) := B^{(0)} + \sum_{m=1}^{d_{\text{in}}} u_{\ell,m} B^{(m)}, \quad C_\theta(u_\ell) := C^{(0)} + \sum_{m=1}^{d_{\text{in}}} u_{\ell,m} C^{(m)},$$

$$\Delta_\theta(u_\ell) = \text{softplus}(a_\Delta^\top u_\ell + b_\Delta), \quad M_\theta(u_\ell) = \exp\{\Delta_\theta(u_\ell) A\},$$

and

$$N_\theta(u_\ell) := \left(\int_0^{\Delta_\theta(u_\ell)} \exp\{sA\}\, ds\right) B_\theta(u_\ell).$$

Starting from $h_{-1} = h_{\text{init}}$, the layer evolves as

$$h_\ell = M_\theta(u_\ell) h_{\ell-1} + N_\theta(u_\ell) u_\ell, \qquad o_\ell = C_\theta(u_\ell)^\top h_\ell,$$

for $0 \le l \le L$. The integral expression is used instead of $(\Delta A)^{-1}(\exp\{\Delta A\} - I)\Delta B$, because it is well-defined even when A is singular. This formulation also emphasizes that the layer is a discretized input-dependent linear dynamical system: the token controls both the transition scale and the input/readout maps.

4.3. Feedforward Layers

For a compact input domain $\mathcal{X} \subset \mathbb{R}^{p_0}$, depth L, widths $p = (p_0, \dots, p_{L+1})$, sparsity s, and envelope $B > 0$, let $\mathsf{FFN}(L, p, s, B)$ denote the class of ReLU networks:

$$\mathsf{FFN}(L,p,s,B) := \Big\{\phi : \mathcal{X} \to \mathbb{R}^{p_{L+1}} : 0 \le \ell \le L, \phi(x) = W_L\sigma(W_{L-1}\sigma(\cdots\sigma(W_0x + b_0)\cdots) + b_{L-1}) + b_L,$$
$$W_\ell \in \mathbb{R}^{p_{\ell+1}\times p_\ell},\ b_\ell \in \mathbb{R}^{p_{\ell+1}},\ \sup_{x\in\mathcal{X}} \|\phi(x)\|_\infty \le B, \sum_{\ell=0}^{L}\big(\|W_\ell\|_0 + \|b_\ell\|_0\big) \le s,\ \max_{0\le\ell\le L}\big\{\|W_\ell\|_\infty, \|b_\ell\|_\infty\big\} \le B\Big\}$$

whose total number of nonzero scalar parameters is at most s, whose scalar parameters are bounded by B and whose output is uniformly bounded on the compact domain. When ϕ is applied to a sequence, the same map acts tokenwise. When a feedforward map is applied to a sequence, we use its tokenwise lift: for $x_0, \dots, x_m \in \mathcal{X}$,

$$\Phi_{\phi,m}(x_0, \dots, x_m) := (\phi(x_0), \dots, \phi(x_m)).$$

Thus the sequence-level feedforward hypothesis class associated with $\mathsf{FFN}(L, p, s, B)$ is

$$\{\Phi_\phi = \{\Phi_{\phi,m}\}_{m\ge 0} : \phi \in \mathsf{FFN}(L,p,s,B)\}.$$

4.4. Two-Layer Mamba Class

A two-layer Mamba student is a composition

$$\mathcal{M}_\Theta = \Phi_3 \circ \mathsf{S6}_{\theta_2} \circ \Phi_2 \circ \mathsf{S6}_{\theta_1} \circ \Phi_1,$$

where Φ_1, Φ_2, Φ_3 are tokenwise feedforward maps and $\mathsf{S6}_{\theta_1}, \mathsf{S6}_{\theta_2}$ are S6 layers. For a finite input sequence $U_{0:T+1}$, the scalar output is the last-token readout

$$\mathcal{M}_\Theta^{(T)}(U_{0:T+1}) := [\mathcal{M}_\Theta(U_0, \dots, U_{T+1})]_{T+1}.$$

The finite-context regression function induced by Θ is

$$F_\Theta^{(T)}(\nu, q) := \mathbb{E}\left[\mathcal{M}_\Theta^{(T)}\big(\iota_{\rm in}(q), \iota_{\rm ctx}(X_1), \dots, \iota_{\rm out}(q)\big)\right]$$

where $X_t \overset{\text{i.i.d.}}{\sim} \nu$.

4.5. Existence of the Infinite-Context Limit

In contrast to attention, which explicitly compares tokens through pairwise interactions, the recurrent layer must compress the context into a state. To obtain a hypothesis class acting on measure-valued contexts, we have to show that the finite-context functions converge as $T \to \infty$. We therefore define

$$F_\Theta(\nu, q) := \lim_{T\to\infty} F_\Theta^{(T)}(\nu, q)$$

whenever the limit exists. The limit says that, after reading a sufficiently long context sampled from ν, the expected final prediction stabilizes and depends only on ν and q, but not on the arbitrary length of the sampled sequence. We enforce this through a stability assumption.

Assumption 4.1 (Stable Recurrent)**.** For each admissible parameter Θ, let $h_t^{(1)}$ and $h_t^{(2)}$ denote the hidden states of the two S6 layers after the t-th context token has been processed, starting from the deterministic state induced by the initial query token $\iota_{\rm in}(q)$. Define the stacked context-processing state

$$H_t := \big(h_t^{(1)}, h_t^{(2)}\big).$$

For $X_t \sim \nu$, set

$$W_t := \iota_{\rm ctx}(X_t), \quad W_t \sim \bar\nu := (\iota_{\rm ctx})_\#\nu.$$

During the context-processing part of the sequence, the state admits the Markovian representation

$$H_t = \Psi_\Theta(H_{t-1}, W_t, q), \quad W_t \overset{\text{i.i.d.}}{\sim} \bar{\nu}.$$

There exists $0 < r < 1$ such that, for all admissible ν, q, w and all stacked states h, h',

$$\|\Psi_\Theta(h, w, q) - \Psi_\Theta(h', w, q)\|_2 \le r\|h - h'\|_2.$$

All token maps, readouts and hidden-state injections are uniformly bounded and Lipschitz on the compact domains.

Proposition 4.2 (Existence of the measure-valued Mamba limit)**.** *Under Assumption 4.1, for every admissible (ν, q), the limit*

$$F_\Theta(\nu, q) = \lim_{T\to\infty} F_\Theta^{(T)}(\nu, q)$$

exists. Moreover, if $\pi_{\Theta,\nu,q}$ denotes the unique invariant law of the context-state Markov chain

$$H_t = \Psi_\Theta(H_{t-1}, W_t, q), \quad W_t \overset{\text{i.i.d.}}{\sim} \bar{\nu}, \quad \bar{\nu} := (\iota_{\text{ctx}})_\# \nu,$$

and if $G_\Theta(h, q)$ denotes the deterministic scalar output obtained by applying the final query transition with token $\iota_{\text{out}}(q)$ and the last-token readout to the stacked state h, then

$$F_\Theta(\nu, q) = \int G_\Theta(h, q)\, d\pi_{\Theta,\nu,q}(h).$$

Proposition 4.2 gives the existence of the infinite-context Mamba model. During the context-processing phase, the hidden state is a Markov chain driven by i.i.d. tokens from the context measure, and we show that this chain forgets its initialization and converges to a unique invariant law $\pi_{\Theta,\nu,q}$. Thus the infinite-context limit is not an additional modeling assumption. It is the measure-level operator induced by the recurrent computation.

For each n, let $\mathcal{A}_n^{\text{Mamba}}$ denote an admissible parameter set for the two-layer Mamba architecture defined as above, including the widths, depths, sparsities, parameter bounds and stability constants allowed at sample size n. We define the associated infinite-context Mamba hypothesis class by

$$\mathcal{H}_n^{\text{Mamba}} := \Big\{ F_\Theta : \mathcal{P}(X) \times \mathbb{R}^{d_1} \to \mathbb{R} \ : \ \Theta \in \mathcal{A}_n^{\text{Mamba}},\ F_\Theta(\nu, q) = \lim_{T\to\infty} F_\Theta^{(T)}(\nu, q) \Big\}.$$

where $F_\Theta^{(T)}$ is the finite-context regression function defined in Section 4.4.

5. RKHS Model and Target Class

We now impose a spectral smoothness on the content measures. The purpose of this assumption is to quantify the effective dimension of the infinite-dimensional measure input. Since the target is a functional of a probability measure, a rate statement requires a notion of how many degrees of freedom of the measure are statistically relevant. Thus, we introduce exponential Mercer kernel decay.

5.1. Spectral Assumptions

Let $K : X_0 \times X_0 \to \mathbb{R}$ be a continuous positive-definite kernel with Mercer expansion

$$K(z, z') = \sum_{j\ge1} \lambda_j e_j(z) e_j(z'),$$

where $(e_j)_{j\ge1}$ is an orthonormal basis of $L^2(X_0)$. For a signed measure μ, define its Mercer coefficients

$$b_j(\mu) := \int_{X_0} e_j(z)\, d\mu(z).$$

For $a \in \mathbb{R}$, define the generalized RKHS norm

$$\|\mu\|_{H_0^a}^2 := \sum_{j\ge1} \lambda_j^{-a} b_j(\mu)^2.$$

This scale should be read as a spectral smoothness scale for measures.

Assumption 5.1 (Kernel decay and eigenfunction regularity)**.** There exist constants $c_\lambda, C_\lambda, c_0, C_0 > 0$ and $\alpha > 0$ such that

$$c_\lambda e^{-C_0 j^\alpha} \le \lambda_j \le C_\lambda e^{-c_0 j^\alpha}, \qquad j \ge 1.$$

The next assumption ensures that the leading spectral feature map can be implemented by the neural student and the spectral feature map could be approximated by ReLU networks with controlled complexity.

Assumption 5.2 (Analytic Mercer eigenfunction regularity)**.** Let $X_0 \subset [-R, R]^{d_2}$ be compact and suppose that there exists $R_+ > R$ such that every Mercer eigenfunction e_j admits an absolutely convergent power-series representation

$$e_j(z) = \sum_{k \in \mathbb{N}_0^{d_2}} a_{j,k} z^k, \qquad z \in [-R_+, R_+]^{d_2},$$

where $z^k = \prod_{\ell=1}^{d_2} z_\ell^{k_\ell}$ and there exist constants $A_0, A_1, \kappa > 0$ such that

$$\sum_{k \in \mathbb{N}_0^{d_2}} |a_{j,k}| R_+^{|k|} \le A_0 \exp\{A_1 j^\kappa\}, \qquad j \ge 1,$$

where $|k| = k_1 + \cdots + k_{d_2}$.

Under this assumption, we could characterize the approximation property of the spectral feature map

$$z \mapsto (e_1(z), \ldots, e_D(z))$$

by ReLU networks. This approximation condition is plausible in standard analytic settings, for example, it is satisfied by heat-kernel-type eigenfeatures on compact domains under uniform analytic bounds (Grigor'yan, 2006).

Assumption 5.3 (Content class)**.** For some $\gamma_b > 0$ and radius $R_b > 0$, every content measure belongs to

$$\mathcal{B}_{\gamma_b}(R_b) := \left\{ \mu \in \mathcal{P}(X_0) : \|\mu\|_{H_0^{\gamma_b}} \le R_b \right\}.$$

Assumption 5.3 is the smoothness condition on the memories stored in the context. It rules out content distributions whose Mercer coefficients place too much mass on high-frequency directions.

Assumption 5.4 (Separated tags)**.** The tag vectors satisfy

$$v^{(i)} \in \mathbb{S}^{d_1 - 1}, \quad \langle v^{(i)}, v^{(j)} \rangle \le 0 \quad (i \ne j), \quad I \le d_1.$$

5.2. Target Class

Fix $\gamma_f < 0$. The prediction functional is assumed Lipschitz with respect to the weak RKHS metric $H_0^{\gamma_f}$. This choice is compatible with the assumption 5.3: the content measures are smooth enough to have rapidly decaying spectral tails while the target is stable enough that those tails affect the prediction only weakly.

Definition 5.5 (Recall–predict target class)**.** Let

$$\mathcal{Q}_I := \{v^{(1)}, \ldots, v^{(I)}\}.$$

For $L, M > 0$, define $\mathcal{G}_\star(L, M)$ to be the class of functions $F_\star$ satisfying

$$F_\star(\nu, q) = \widetilde{F}_\star(\mu_{i_\star}, q), \quad q = v^{(i_\star)},$$

for some $\widetilde{F}_\star : \mathcal{B}_{\gamma_b}(R_b) \times \mathcal{Q}_I \to [-M, M]$ obeying

$$\left| \widetilde{F}_\star(\mu, q) - \widetilde{F}_\star(\mu', q') \right| \le L \left(\|\mu - \mu'\|_{H_0^{\gamma_f}} + \|q - q'\|_2 \right).$$

6. Main Results

We now state the main statistical guarantees for the recall-predict problem. The results have two parts. First, we prove an upper bound showing that a two-layer Mamba model can learn the recall-predict target at a rate determined by the effective spectral dimension of the content measures. Second, we prove a matching minimax lower bound showing that this rate cannot be improved up to constants in the exponent by any estimator.

6.1. Upper Bound for Two-Layer Mamba

Our upper bound shows that there exists a sequence of two-layer Mamba hypothesis classes whose empirical risk minimizers uniformly learn every target in the recall–predict class $\mathcal{G}_\star(L, M)$. In other words, the architecture is expressive enough to perform the two operations required by the task: it can use the query to recall the relevant tagged component of the context, and it can then approximate the target functional of the recalled content measure.

Theorem 6.1 (Upper bound for two-layer Mamba)**.** *Under assumptions 4.1–5.4, let $S_n = \{(\nu_t, q_t, Y_t)\}_{t=1}^n$ be generated by the recall–predict model with*

$$Y_t = F_\star(\nu_t, q_t) + \xi_t, \qquad \xi_t \sim N(0, \sigma^2),$$

and $F_\star \in \mathcal{G}_\star(L, M)$. There exists a sequence of two-layer Mamba hypothesis classes $\mathcal{H}_n^{\mathrm{Mamba}}$ such that any approximate ERM $\widehat{F}_n \in \mathcal{H}_n^{\mathrm{Mamba}}$ satisfies

$$\sup_{F_\star \in \mathcal{G}_\star(L,M)} \mathbb{E}\,\mathcal{R}(\widehat{F}_n, F_\star) \le C \exp\left\{-c(\log n)^{\alpha/(\alpha+1)}\right\},$$

provided one of the following two conditions holds:

1. *Query-dependent case:*
$$I \le d_1 \le C_I (\log n)^{1/(\alpha+1)}.$$
2. *Query-independent case:*
$$\widetilde{F}_\star(\mu, q) = \widetilde{F}_\star(\mu), \quad I \le d_1 = n^{o(1)}.$$

The constants $C, c > 0$ depend on $\alpha, \gamma_b, \gamma_f, R_b, L, M, \sigma$ and the constants in Assumption 5.1, but not depend on n.

The two regimes in Theorem 6.1 separate two different roles of the query. In the query-dependent case, q is not only an address used for recall but also enters the target functional itself. The learner must therefore resolve both the recalled measure and the query value as part of the regression problem and this is the reason why the number of separated tags is restricted to the same effective scale as the spectral truncation dimension $(\log n)^{1/(\alpha+1)}$.

In the query-independent case, the query is used only to select the correct memory cell, while the prediction functional depends only on the recalled content measure. Once recall has succeeded, the identity of the tag no longer affects the scalar prediction, which removes the query from the intrinsic regression dimension and allows a much larger number of tags, i.e. $I \le d_1 = n^{o(1)}$.

6.2. Structured Minimax Lower Bound

We next show that the rate in Theorem 6.1 is sharp. The lower bound is architecture-independent: it applies to every measurable estimator, not only to Mamba or to recurrent models. Hence it identifies an intrinsic statistical limitation of the recall-predict problem.

To show the lower bound, we impose a structured coefficient model for the content distributions. Our idea is to build a rich but controlled family of probability measures whose Mercer coefficients contain many independent degrees of freedom.

Assumption 6.2 (Structured coefficient model)**.** Let p_0 be a fixed density on X_0 satisfying

$$p_0(z) \ge m_0 > 0, \quad \int_{X_0} p_0(z)\, dz = 1,$$

and assume $p_0 \in H_0^{\gamma_b}$. We also assume the eigenfunctions are centered: $\int_{X_0} e_j(z)\, dz = 0, \quad j \ge 1$. Let Z_j be independent random variables with continuous densities ρ_j such that $\sup_j \|\rho_j\|_\infty \le R$ and $|Z_j| \le 1$.

Fix $\gamma_d > \gamma_b$. For sufficiently small $a_0 > 0$, define the random density

$$p_\mu(z) := p_0(z) + a_0 \sum_{j \ge 1} \lambda_j^{\gamma_d/2} Z_j e_j(z).$$

The constant a_0 is chosen small enough that $p_\mu \ge 0$ almost surely. Then $\mu(dz) = p_\mu(z)\, dz$ is a probability measure and belongs to $\mathcal{B}_{\gamma_b}(R_b)$ almost surely, after increasing R_b if necessary.

Theorem 6.3 (Minimax lower bound)**.** *Assume Assumptions 5.1, 5.3, and 6.2, where the random content measures in Assumption 6.2 are written as $\mu(dz) = p_\mu(z)\,dz$ using the Mercer eigenfunctions $(e_j)_{j\geq 1}$ from Assumption 5.1. Let*

$$\mathfrak{M}_n := \inf_{\widehat{F}_n} \sup_{F_\star \in \mathcal{G}_\star(L,M)} \mathbb{E}\,\mathcal{R}(\widehat{F}_n, F_\star),$$

where the infimum is over all measurable estimators based on $S_n = \{(\nu_t, q_t, Y_t)\}_{t=1}^n$. Then there exist constants $c, C > 0$, independent of n, such that

$$\mathfrak{M}_n \geq c \exp\left\{-C(\log n)^{\alpha/(\alpha+1)}\right\}.$$

Theorem 6.3 shows that the logarithmic exponent in the upper bound cannot be improved uniformly over the target class. The difficulty comes from small spectral perturbations of the recalled content distribution. The lower bound also clarifies the role of the smoothness assumptions. The content smoothness condition controls how much mass the content distributions can place in high-frequency Mercer directions, while the target Lipschitz condition controls how sensitively the scalar response can depend on those directions.

We empirically verify the sub-polynomial rate predicted by Theorem 6.1 and Theorem 6.3 on a synthetic measure-valued recall–predict task. The experiment setting and results could be found in Appendix A.

7. Conclusion

We developed a statistical theory of associative recall for selective state-space models in the infinite-context, measure-valued regime. In the recall-predict problem, the context is a mixture of tagged content measures and the query specifies which component should determine the response. Our first contribution is architectural: by inserting the query before and after the context and by imposing a stable recurrent core, we show that the finite-context Mamba predictor has a well-defined infinite-context limit. This limit is expressed through the invariant law of the Markov chain generated by the context scan and therefore gives a principled measure-valued Mamba operator rather than a sequence-length-dependent heuristic.

Our second contribution is statistical. Under separated tags, exponential Mercer eigenvalue decay, smooth content measures, analytic eigenfunction approximation and recurrent dynamics, we prove that approximate empirical risk minimization over two-layer Mamba classes learns the recall–predict mapping with population risk

$$\exp\{-c(\log n)^{\alpha/(\alpha+1)}\}.$$

We further prove an architecture-independent minimax lower bound with the same logarithmic exponent over a structured coefficient model. Thus the rate achieved by the Mamba class is statistically optimal at the exponent level. This is significant because it shows that measure-level associative memory is not exclusive to all-pairs attention: a recurrent architecture can perform query-conditioned recall and prediction from arbitrarily long distributional contexts, provided that the recurrence is stable and the relevant spectral dimension is controlled.

Compared with Kawata & Suzuki (2026), who establish minimax-optimal measure-level recall for learned softmax Transformers, our work is different. Their analysis relies on the explicit query–key normalization of softmax attention to concentrate mass on the queried component and then aggregate its Mercer features. In contrast, our Mamba model has no all-pairs attention or explicit attention kernel. Recall must be realized through query insertion, selective recurrent state updates and the invariant distribution of a stable context-processing Markov chain. Thus, while both works obtain the same sub-polynomial statistical rate under closely related measure-valued recall–predict assumptions, our contribution is to show that this optimal behavior can also be achieved by a selective state-space architecture, thereby separating the statistical phenomenon of measure-level associative recall from the particular softmax-attention implementation.

Several questions remain open. Our theory assumes separated tags, stable recurrent dynamics and exponentially decaying spectra. Extending the analysis to less separated tags, polynomial spectral decay, finite-sample context noise, trained stability constraints and deeper practical Mamba variants would make the theory closer to modern long-context systems.

References

Arora, S., Eyuboglu, S., Timalsina, A., Johnson, I., Poli, M., Zou, J., Rudra, A., and Re, C. Zoology: Measuring and improving recall in efficient language models. In *The Twelfth International Conference on Learning Representations*.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36:1560–1588, 2023.

Cabannes, V., Dohmatob, E., and Bietti, A. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations*.

Furuya, T., de Hoop, M. V., and Peyré, G. Transformers are universal in-context learners. In *The Thirteenth International Conference on Learning Representations*.

Geshkovski, B., Rigollet, P., and Ruiz-Balet, D. Measure-to-measure interpolation using transformers. *arXiv preprint arXiv:2411.04551*, 2024.

Grigor'yan, A. Heat kernels on weighted manifolds and applications. *Cont. Math*, 398(2006):93–191, 2006.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., and Ré, C. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34: 572–585, 2021.

Gu, A., Goel, K., and Re, C. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Gupta, A., Gu, A., and Berant, J. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.

Huang, N., Sarabia, M., Moudgil, A., Rodriguez, P., Zappella, L., and Danieli, F. Understanding input selectivity in mamba: impact on approximation power, memorization, and associative recall capacity. *arXiv preprint arXiv:2506.11891*, 2025.

Jiang, Y., Rajendran, G., Ravikumar, P., and Aragam, B. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers. *Advances in Neural Information Processing Systems*, 37:67712–67757, 2024.

Kawata, R. and Suzuki, T. Transformers as measure-theoretic associative memory: A statistical perspective and minimax optimality. *arXiv preprint arXiv:2602.01863*, 2026.

Kim, J., Kim, M., and Mozafari, B. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Le Corre, G., Huang, N., and Bietti, A. How does mamba perform associative recall? a mechanistic study. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Mahdavi, S., Liao, R., and Thrampoulidis, C. Memorization capacity of multi-head attention in transformers. In *The Twelfth International Conference on Learning Representations*.

Nichani, E., Lee, J. D., and Bietti, A. Understanding factual recall in transformers via associative memories. In *The Thirteenth International Conference on Learning Representations*.

Nichani, E., Lee, J. D., and Bietti, A. Understanding factual recall in transformers via associative memories. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Takakura, S. and Suzuki, T. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning*, pp. 33416–33447. PMLR, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhou, T.-Y., Suh, N., Cheng, G., and Huo, X. Approximation of rkhs functionals by neural networks. *arXiv preprint arXiv:2403.12187*, 2024.

A. Experiments

We empirically verify the sub-polynomial rate predicted by Theorem 6.1 on a synthetic measure-valued recall–predict instance. The data-generating process is identical to the synthetic setup of Kawata & Suzuki (2026); the only substitution is that the recall mechanism is now realized by an S6 (Mamba) selective state-space block instead of softmax attention. Any rate of the form $\exp\{-c(\log n)^{\alpha/(\alpha+1)}\}$ that appears in the empirical risk is therefore attributable to the Mamba block, not to the data distribution.

A.1. Synthetic Recall–Predict Setup

Data-generating process. We work on the content domain $X_0 = [0, 1]$ with the trigonometric basis $e_0 \equiv 1$ and $e_j(x) = \sqrt{2}\sin(\pi j x)$ for $j \geq 1$, and Mercer eigenvalues $\lambda_j = \exp(-j^\alpha)$ truncated at $M = 16$ coordinates. For each example we draw two independent Gaussian coefficient vectors $Z^{(1)}, Z^{(2)} \sim \mathcal{N}(0, I_{M-1})$ (with the constant coordinate set to zero), form unnormalized densities

$$\widetilde{\mu}_k(x) = \sum_{j=0}^{M-1} \lambda_j Z_j^{(k)} e_j(x), \qquad k \in \{1, 2\},$$

and clamp to nonnegative values and renormalize on a uniform grid of $T = 32$ points to obtain proper probability mass functions p_1, p_2. The tag pair is $v_1 \in \{-1, +1\}$ drawn uniformly and $v_2 = -v_1$, so the two tagged measures are well separated in the sense of Assumption 5.4. The context is a length-5000 i.i.d. sample from the balanced mixture

$$\nu = \tfrac{1}{2}\big(\delta_{v_1} \otimes \mu_1\big) + \tfrac{1}{2}\big(\delta_{v_2} \otimes \mu_2\big),$$

where each token is a pair $(x, v) \in [0, 1] \times \{-1, +1\}$. The query token is $q = (0, v_1)$. The target is the quadratic functional of the tagged content measure recalled by the query,

$$Y = v_1 \cdot \sum_{j=0}^{M-1} \lambda_j (Z_j^{(1)})^2 + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2),$$

with $\sigma = 10^{-2}$. This Y depends on the context only through the Mercer coefficients of $\mu_{i_\star} = \mu_1$ and is exactly of the recall–predict form of Definition 3.1.

Student architecture. We use a minimal MLP $\to$ S6 $\to$ MLP head, instantiating the encoder, S6 block, and pointwise MLP from Section 4 with $d_{\text{model}} = 16$, $d_{\text{state}} = 32$, $d_{\text{conv}} = 4$, expansion factor 2, and Δ-rank chosen automatically from d_{model}. Following the query-insertion convention of Section 4.1, we prepend and append the query embedding around the encoded context, run a single S6 scan, and read out the last position before the head MLP. The state-transition matrix is parameterized as $A = -\exp(A_{\log})$ with diagonal $A_{\log}$, keeping the recurrence stable as required by the stable-core assumption used in Theorem 6.1.

Optimization. We train each (α, n) configuration from scratch with Adam (learning rate 2×10^{-3}, exponential decay $\gamma = 0.95$), batch size $\min(16, n)$, gradient clipping at 1.0, and 30 epochs, and report the lowest validation MSE encountered during training (early stopping). Validation is computed on $n_{\text{val}} = 2000$ held-out i.i.d. examples generated with a different seed. Across runs only the number of training contexts n varies; the per-token distribution and the target functional are held fixed at the values of α.

A.2. Empirical Risk and Predicted Rate

We sweep $\alpha \in \{0.75, 1.0\}$ and $n \in \{4, 8, 16, 32, 64\}$. Theorem 6.1 predicts that, for an ERM in a sufficiently expressive stable two-layer Mamba class, $\log \mathcal{R}(\widehat{F}_n) \leq A - c(\log n)^{\alpha/(\alpha+1)}$ for some $c > 0$. We therefore fit the linear model

$$\log L(n) = A - C\,(\log n)^{\alpha/(\alpha+1)}$$

to the observed validation MSE by ordinary least squares, separately for each α.

Figure 1 shows the result. Both curves fall in the regime predicted by the upper bound: the empirical risk decreases monotonically in n and is well-described by a linear fit on the $(\log n)^{\alpha/(\alpha+1)}$ axis. The fitted slopes ($C \approx 1.81$ for

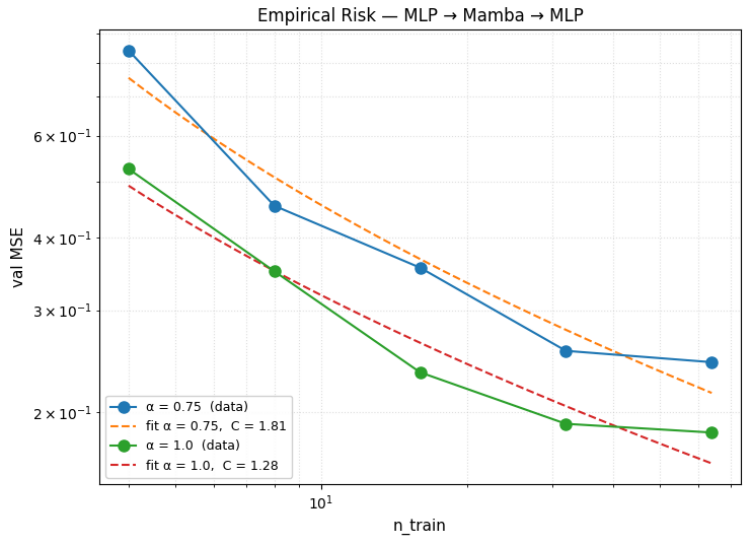

Figure 1. Validation MSE of the MLP → S6 → MLP student against the number of training examples n, for two values of the Mercer-decay parameter α. Solid curves are measured MSE; dashed curves are the rate $\exp\{A - C(\log n)^{\alpha/(\alpha+1)}\}$ fit by least squares. At fixed n, faster Mercer decay ($\alpha = 1.0$) gives a smaller risk, matching the theoretical prediction that a faster decay corresponds to a smaller effective dimension.

$\alpha = 0.75$; $C \approx 1.28$ for $\alpha = 1.0$) place the curves within the regime predicted by the proof of Theorem 6.1, where the constant in the exponent depends on γ_b, γ_f and the kernel constants but not on n. Crucially, the $\alpha = 1.0$ curve sits below the $\alpha = 0.75$ curve at every sample size, in agreement with the consequence in Section 6: at fixed n, faster spectral decay corresponds to a smaller intrinsic dimension and therefore smaller risk.

A.3. Discussion and Future Ablations

These experiments are intentionally minimal: they verify the qualitative sub-polynomial rate on a setup that matches the data-generating process used in the measure-theoretic Transformer literature (Kawata & Suzuki, 2026), with the only change being that the recall mechanism is realized by a stable S6 block instead of softmax attention. The same scaling law appearing empirically supports the theoretical message of Theorems 6.1 and 6.3: the rate is governed by the spectral structure of the content class rather than by the particular content-addressable architecture, and a stable selective state-space layer is expressive enough to realize it.

A.4. Sweep over α and T

Beyond the single (α, T) configuration reported in Section A.1, we additionally sweep

$$\alpha \in \{0.5,\ 0.75,\ 1.0,\ 2.0\}, \qquad T \in \{500,\ 1000,\ 2000\}, \qquad n_{\text{train}} \in \{4,\ 8,\ 16,\ 32,\ 64\},$$

giving $4\times3 = 12$ (α, T) regimes and 5 training-set sizes per regime (60 runs in total). All other hyper-parameters are kept exactly as in Section A.1; in particular, the two-layer Mamba student $M_\Theta = \Phi_3 \circ \mathrm{S6}_{\theta_2} \circ \Phi_2 \circ \mathrm{S6}_{\theta_1} \circ \Phi_1$ uses $d_{\text{model}} = 16$, $d_{\text{state}} = 32$, expansion factor 2 and depthwise-separable conv kernel of width 4, trained with Adam at learning rate 2×10^{-3}, exponential decay $\gamma = 0.95$, gradient clipping at $\|g\|_2 \le 1$, label noise $\sigma = 0.01$, 120 epochs and early-stopping on a held-out validation set of 500 examples. We report the *best* validation MSE attained during training.

Figures 2–4 show the validation MSE as a function of the training-set size n_{train} on a log–log scale, one panel per $T \in \{500, 1000, 2000\}$. Solid markers are measured empirical risks; dashed lines are ordinary-least-squares fits of the paper-derived rate

$$\log L(n) \;=\; A \;-\; C\,(\log n)^{\alpha/(\alpha+1)}, \tag{1}$$

which is the closed-form prediction of Theorem 6.1; the fitted constant C is reported in the legend of each panel.

The following observations hold across all three context lengths.

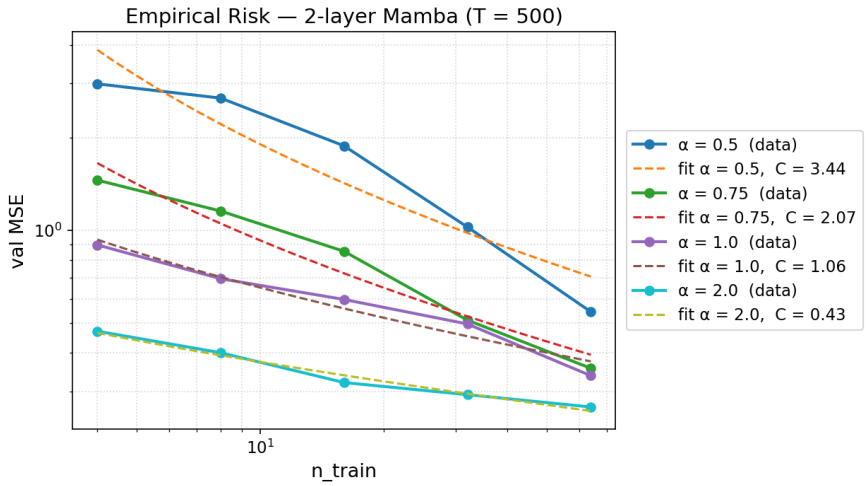

Figure 2. Empirical risk vs. n_{train} for the two-layer Mamba student at context length $T = 500$, for $\alpha \in \{0.5, 0.75, 1.0, 2.0\}$. Dashed lines are OLS fits of Eq. (1).

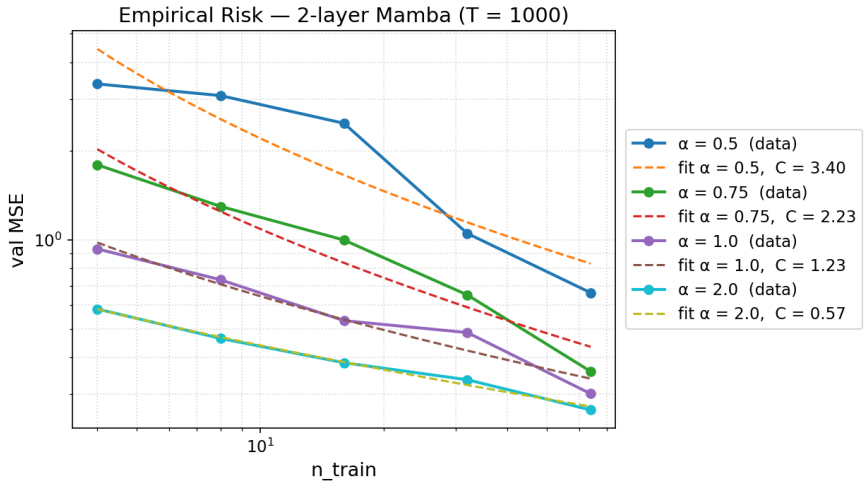

Figure 3. Same as Figure 2 but at $T = 1000$.

- **Monotone decay in n_{train}.** Every curve is monotonically decreasing in n_{train}, and the slope on log–log axes is sub-polynomial in n, consistent with Eq. (1).
- **Smoothness orders the curves.** At any fixed n_{train}, the validation MSE is monotone in α: $\alpha = 2.0 < 1.0 < 0.75 < 0.5$. Smoother targets are easier to learn from few samples, exactly as predicted by Theorem 6.1.
- **C tracks α inversely.** The fitted constant C in Eq. (1) decreases with α: approximately $C \approx 3.3$–3.4 for $\alpha = 0.5$, ≈ 2.1–2.2 for $\alpha = 0.75$, ≈ 1.1–1.2 for $\alpha = 1.0$, and ≈ 0.4–0.6 for $\alpha = 2.0$ (see legends). Rougher measures (small α) require a larger pre-factor C to fit the same asymptotic shape.
- **Stability across T.** The three figures are nearly identical up to small fluctuations: doubling or halving the context length leaves both the curve ordering and the fitted C within a few percent of each other. This matches the prediction of Theorem 6.1, in which the rate is governed by α and is essentially independent of the context length once T is large enough for the selective state-space to resolve the target measure.

Take-away. Across all 12 (α, T) configurations, both the *shape* of the predicted rate (Eq. (1)) and its *dependence on the smoothness* α are recovered, while the behaviour is robust to the choice of context length T in the regime we tested.

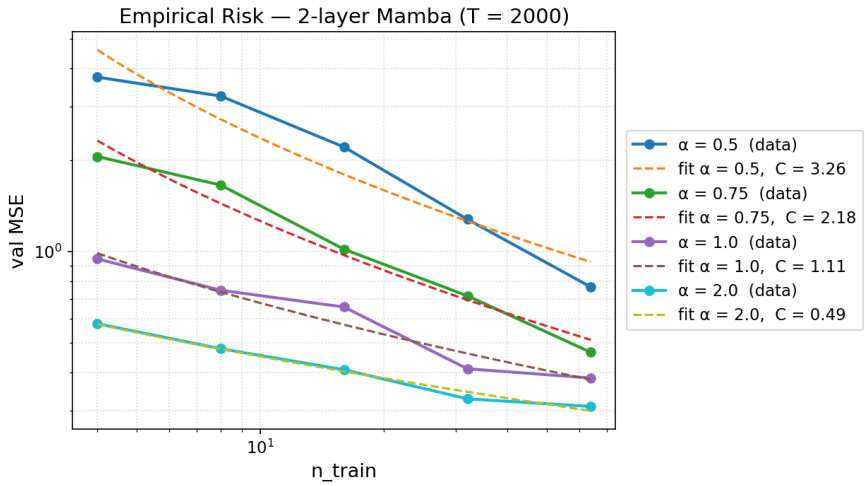

Figure 4. Same as Figure 2 but at $T = 2000$.

The residual gap that remains for $\alpha = 0.5$ at small n_{train} is consistent with the constant-order finite-sample correction discussed after Theorem 6.1, and shrinks as n_{train} grows.

B. Proof for Proposition 4.2

Proof of Proposition 4.2. Fix an admissible parameter value Θ and an admissible pair (ν, q). Write

$$\bar{\nu} := (\iota_{\text{ctx}})_{\#}\nu \in \mathcal{P}(\mathcal{U})$$

for the law of an augmented context token. Since the initial query token $\iota_{\text{in}}(q)$ is deterministic once q is fixed, the stacked state immediately after processing this token is deterministic. We write it as

$$H_0 = \chi_\Theta(q)$$

for a measurable map χ_Θ. During the context-processing phase, Assumption 4.1 gives the Markov recursion

$$H_t = \Psi_\Theta(H_{t-1}, W_t, q), \qquad W_t \overset{\text{i.i.d.}}{\sim} \bar{\nu}, \qquad t \geq 1.$$

We first restrict the chain to a compact invariant state set. Let d_H be the dimension of the stacked state H_t. Since the admissible query domain and the augmented token domain are compact, and since the relevant maps are uniformly bounded and Lipschitz on these domains, the quantities

$$B_{\Psi,\Theta} := \sup_{\substack{w \in \operatorname{supp}\bar{\nu} \\ q' \text{ admissible}}} \|\Psi_\Theta(0, w, q')\|_2, \qquad B_{\chi,\Theta} := \sup_{q' \text{ admissible}} \|\chi_\Theta(q')\|_2$$

are finite. Choose

$$R_\Theta \geq \max\left\{B_{\chi,\Theta}, \frac{B_{\Psi,\Theta}}{1-r}\right\}, \qquad \mathcal{Z}_\Theta := \{h \in \mathbb{R}^{d_H} : \|h\|_2 \leq R_\Theta\}.$$

Then $H_0 = \chi_\Theta(q) \in \mathcal{Z}_\Theta$. Moreover, for every $h \in \mathcal{Z}_\Theta$ and every $w \in \operatorname{supp}\bar{\nu}$, Assumption 4.1 yields

$$\begin{aligned}\|\Psi_\Theta(h, w, q)\|_2 &\leq \|\Psi_\Theta(0, w, q)\|_2 + \|\Psi_\Theta(h, w, q) - \Psi_\Theta(0, w, q)\|_2 \\ &\leq B_{\Psi,\Theta} + r\|h\|_2 \\ &\leq B_{\Psi,\Theta} + rR_\Theta \leq R_\Theta.\end{aligned}$$

Thus $\mathcal{Z}_\Theta$ is invariant under the context transition.

Define the Markov kernel $K_{\Theta,\nu,q}$ on $\mathcal{Z}_\Theta$ by

$$K_{\Theta,\nu,q}(h, A) := \int_{\mathcal{U}} \mathbf{1}_A(\Psi_\Theta(h, w, q))\, d\bar{\nu}(w), \qquad A \in \mathcal{B}(\mathcal{Z}_\Theta).$$

For $\rho \in \mathcal{P}(\mathcal{Z}_\Theta)$, write

$$\rho K_{\Theta,\nu,q}(A) := \int_{\mathcal{Z}_\Theta} K_{\Theta,\nu,q}(h, A)\, d\rho(h).$$

We claim that the Markov operator $\rho \mapsto \rho K_{\Theta,\nu,q}$ is a strict contraction on $(\mathcal{P}(\mathcal{Z}_\Theta), W_1)$, where W_1 is induced by the Euclidean norm. Let $\rho, \eta \in \mathcal{P}(\mathcal{Z}_\Theta)$, and let $\gamma \in \Gamma(\rho, \eta)$ be any coupling. If $(H, H') \sim \gamma$ and $W \sim \bar{\nu}$ is independent of (H, H'), then

$$(\Psi_\Theta(H, W, q), \Psi_\Theta(H', W, q))$$

is a coupling of $\rho K_{\Theta,\nu,q}$ and $\eta K_{\Theta,\nu,q}$. Therefore,

$$\begin{aligned} W_1(\rho K_{\Theta,\nu,q}, \eta K_{\Theta,\nu,q}) &\le \mathbb{E}\left[\|\Psi_\Theta(H, W, q) - \Psi_\Theta(H', W, q)\|_2\right] \\ &\le r\, \mathbb{E}\|H - H'\|_2. \end{aligned}$$

Taking the infimum over all couplings $\gamma \in \Gamma(\rho, \eta)$ gives

$$W_1(\rho K_{\Theta,\nu,q}, \eta K_{\Theta,\nu,q}) \le r\, W_1(\rho, \eta).$$

Since $\mathcal{Z}_\Theta$ is compact, $(\mathcal{P}(\mathcal{Z}_\Theta), W_1)$ is complete. Hence the Banach fixed point theorem gives a unique invariant law

$$\pi_{\Theta,\nu,q} \in \mathcal{P}(\mathcal{Z}_\Theta)$$

satisfying

$$\pi_{\Theta,\nu,q} K_{\Theta,\nu,q} = \pi_{\Theta,\nu,q}.$$

Furthermore, for every initial law $\rho_0 \in \mathcal{P}(\mathcal{Z}_\Theta)$,

$$W_1(\rho_0 K^T_{\Theta,\nu,q}, \pi_{\Theta,\nu,q}) \le r^T W_1(\rho_0, \pi_{\Theta,\nu,q}).$$

Taking $\rho_0 = \delta_{\chi_\Theta(q)}$, the law

$$\rho_T := \mathcal{L}(H_T) = \delta_{\chi_\Theta(q)} K^T_{\Theta,\nu,q}$$

therefore satisfies

$$W_1(\rho_T, \pi_{\Theta,\nu,q}) \le r^T W_1(\delta_{\chi_\Theta(q)}, \pi_{\Theta,\nu,q}) \le 2R_\Theta r^T,$$

because both measures are supported on the ball $\mathcal{Z}_\Theta$, whose diameter is at most $2R_\Theta$.

It remains to pass from convergence of context-state laws to convergence of the scalar final output. Starting from a context state h, the final query token $\iota_{\text{out}}(q)$ is deterministic. Hence applying the final query transition and then the last-token scalar readout defines a deterministic measurable map

$$G_\Theta(\cdot, q) : \mathcal{Z}_\Theta \to \mathbb{R}.$$

By Assumption 4.1, this map is bounded and Lipschitz on $\mathcal{Z}_\Theta$. Let $L_{G,\Theta}$ be a Lipschitz constant in the state variable:

$$|G_\Theta(h, q) - G_\Theta(h', q)| \le L_{G,\Theta}\|h - h'\|_2, \qquad h, h' \in \mathcal{Z}_\Theta.$$

The finite-context regression function can be written as

$$F^{(T)}_\Theta(\nu, q) = \mathbb{E}[G_\Theta(H_T, q)] = \int_{\mathcal{Z}_\Theta} G_\Theta(h, q)\, d\rho_T(h).$$

Therefore, by the Kantorovich–Rubinstein dual characterization of W_1,

$$\begin{aligned}\left|F_\Theta^{(T)}(\nu,q) - \int_{\mathcal{Z}_\Theta} G_\Theta(h,q)\, d\pi_{\Theta,\nu,q}(h)\right| &= \left|\int_{\mathcal{Z}_\Theta} G_\Theta(h,q)\, d(\rho_T - \pi_{\Theta,\nu,q})(h)\right| \\ &\le L_{G,\Theta} W_1(\rho_T, \pi_{\Theta,\nu,q}) \\ &\le 2R_\Theta L_{G,\Theta} r^T.\end{aligned}$$

Since $0 < r < 1$, the right-hand side tends to zero as $T \to \infty$. Hence the infinite-context limit exists and is given by

$$F_\Theta(\nu,q) := \lim_{T\to\infty} F_\Theta^{(T)}(\nu,q) = \int_{\mathcal{Z}_\Theta} G_\Theta(h,q)\, d\pi_{\Theta,\nu,q}(h).$$

This is exactly the invariant-law representation claimed in the proposition. □

C. Proof of Theorem 6.1

Throughout this appendix, constants denoted by $C, c, c_0, c_1, \ldots$ may change from line to line. They may depend on

$$\alpha, \gamma_b, \gamma_f, R_b, L, M, \sigma,$$

on the constants in Assumption 5.1, and on the fixed marker encoding, but not on n. We write $A \lesssim B$ if $A \le CB$ for such a constant C.

We also state the proof for exact ERM. If $\widehat{F}_n$ is only an approximate ERM with empirical excess τ_n, the final bound below has an additional $+\tau_n$ term; hence the same rate holds whenever

$$\tau_n \lesssim \exp\{-c(\log n)^{\alpha/(\alpha+1)}\}.$$

C.1. Admissible input class

Let $\mathfrak{X}_{I,d_1}$ denote the set of all pairs (ν, q) of the recall–predict form

$$\nu = \frac{1}{I}\sum_{i=1}^{I} \delta_{v^{(i)}} \otimes \mu_i, \qquad q = v^{(i_\star)},$$

where $\mu_i \in \mathcal{B}_{\gamma_b}(R_b)$, the tags satisfy Assumption 5.4, and $i_\star \in [I]$. For $\mu \in \mathcal{B}_{\gamma_b}(R_b)$, write

$$b_j(\mu) = \int_{X_0} e_j(z)\, d\mu(z), \qquad b_D(\mu) = (b_1(\mu), \ldots, b_D(\mu)).$$

C.2. Spectral truncation

Lemma C.1 (Finite-dimensional reduction)**.** *Let $F_\star \in \mathcal{G}_\star(L, M)$. For every $D \ge 1$, there exists a function*

$$f_D : \mathbb{R}^D \times \mathcal{Q}_I \to [-M, M]$$

such that, for every admissible $(\nu, q) \in \mathfrak{X}_{I,d_1}$ with $q = v^{(i_\star)}$,

$$\left|F_\star(\nu,q) - f_D(b_D(\mu_{i_\star}), q)\right| \le C\lambda_{D+1}^{(\gamma_b-\gamma_f)/2}.$$

Moreover f_D may be chosen CL-Lipschitz with respect to the metric

$$d_D\big((b,q),(b',q')\big) := \left(\sum_{j=1}^{D} \lambda_j^{-\gamma_f}(b_j - b_j')^2\right)^{1/2} + \|q - q'\|_2.$$

Since $\gamma_f < 0$, $d_D \lesssim \|(b,q) - (b',q')\|_2$, uniformly in D.

Proof. For $\mu \in \mathcal{B}_{\gamma_b}(R_b)$, define its tail seminorm

$$T_D(\mu)^2 := \sum_{j>D} \lambda_j^{-\gamma_f} b_j(\mu)^2.$$

Because $\gamma_b > 0$, $\gamma_f < 0$, and $\|\mu\|_{H_0^{\gamma_b}} \le R_b$,

$$T_D(\mu)^2 = \sum_{j>D} \lambda_j^{\gamma_b-\gamma_f} \lambda_j^{-\gamma_b} b_j(\mu)^2 \le \lambda_{D+1}^{\gamma_b-\gamma_f} R_b^2.$$

Hence

$$T_D(\mu) \le R_b \lambda_{D+1}^{(\gamma_b-\gamma_f)/2}.$$

Let

$$\mathcal{A}_D := \{(b_D(\mu), q) : \mu \in \mathcal{B}_{\gamma_b}(R_b),\ q \in \mathcal{Q}_I\}.$$

For $(b, q) \in \mathbb{R}^D \times \mathcal{Q}_I$, define the inf-extension

$$f_D(b, q) := \inf_{\mu \in \mathcal{B}_{\gamma_b}(R_b)} \left\{ \widetilde{F}_\star(\mu, q) + L \left(\sum_{j=1}^{D} \lambda_j^{-\gamma_f} (b_j - b_j(\mu))^2 \right)^{1/2} \right\},$$

and clip it to $[-M, M]$. Clipping does not increase the Lipschitz constant. The definition makes f_D L-Lipschitz in b with respect to the weighted finite-dimensional norm and L-Lipschitz in q.

Fix $\mu \in \mathcal{B}_{\gamma_b}(R_b)$. Taking the same μ in the infimum gives

$$f_D(b_D(\mu), q) \le \widetilde{F}_\star(\mu, q).$$

Conversely, for any $\nu \in \mathcal{B}_{\gamma_b}(R_b)$,

$$\begin{aligned} &\widetilde{F}_\star(\nu, q) + L \left(\sum_{j=1}^{D} \lambda_j^{-\gamma_f} (b_j(\mu) - b_j(\nu))^2 \right)^{1/2} \\ &\qquad \ge \widetilde{F}_\star(\mu, q) - L\|\mu - \nu\|_{H_0^{\gamma_f}} + L \left(\sum_{j=1}^{D} \lambda_j^{-\gamma_f} (b_j(\mu) - b_j(\nu))^2 \right)^{1/2}. \end{aligned}$$

The $H_0^{\gamma_f}$-distance is bounded by the finite-dimensional part plus the two tails:

$$\|\mu - \nu\|_{H_0^{\gamma_f}} \le \left(\sum_{j=1}^{D} \lambda_j^{-\gamma_f} (b_j(\mu) - b_j(\nu))^2 \right)^{1/2} + T_D(\mu) + T_D(\nu).$$

Therefore

$$f_D(b_D(\mu), q) \ge \widetilde{F}_\star(\mu, q) - 2LR_b \lambda_{D+1}^{(\gamma_b-\gamma_f)/2}.$$

This proves the approximation claim. The final Euclidean Lipschitz statement follows from

$$\sum_{j=1}^{D} \lambda_j^{-\gamma_f} (b_j - b'_j)^2 \le \lambda_1^{-\gamma_f} \|b - b'\|_2^2,$$

because $\gamma_f < 0$. □

By Assumption 5.1, there are constants $c_\lambda, C_\lambda > 0$ such that

$$\lambda_{D+1}^{(\gamma_b-\gamma_f)/2} \le C \exp\{-cD^\alpha\}.$$

Thus choosing

$$D_\varepsilon := \left\lceil C_D (\log \varepsilon^{-1})^{1/\alpha} \right\rceil$$

with C_D sufficiently large gives

$$\lambda_{D_\varepsilon+1}^{(\gamma_b-\gamma_f)/2} \lesssim \varepsilon.$$

C.3. A stable S6 averaging primitive

The next lemma is the architectural ingredient needed to turn the finite coefficient vector $b_D(\mu_{i_\star})$ into a stable Mamba state. It is the Mamba analogue of the recall step in the measure-theoretic Transformer proof.

Lemma C.2 (Stable query-conditioned S6 averaging primitive)**.** *Let $G : \mathbb{R}^{d_1} \times X \to \mathbb{R}^m$ be a bounded ReLU map satisfying $\|G(q,x)\|_\infty \le B_G$. Let $0 < \rho < 1$. There is a two-layer Mamba subnetwork, with the first S6 layer carrying the query register and the second S6 layer carrying an averaging state $H_t \in \mathbb{R}^m$, such that during the context-processing part,*

$$H_t = \rho H_{t-1} + (1-\rho)G(q, X_t), \qquad X_t \overset{\text{i.i.d.}}{\sim} \nu.$$

The context-state transition is ρ-contractive in H_t. Its stationary law is the law of

$$H_\rho = (1-\rho)\sum_{k=0}^{\infty} \rho^k G(q, X_{-k}),$$

where $X_0, X_{-1}, \dots$ are i.i.d. with law ν. Consequently,

$$\mathbb{E}H_\rho = \int G(q,x)\, d\nu(x),$$

and

$$\mathbb{E}\|H_\rho - \mathbb{E}H_\rho\|_2^2 \le 4mB_G^2(1-\rho).$$

Proof. Because the markers $m_{\text{in}}, m_{\text{ctx}}, m_{\text{out}}$ are fixed and take only three values, a tokenwise ReLU map can convert them into exact type flags on the augmented token domain. The first S6 layer writes the first query token into a query register. Equivalently, in the Markovian representation of Assumption 4.1, the context transition is written as

$$H_t = \Psi(H_{t-1}, U_t, q),$$

where q is an exogenous argument during the context pass. The contracted state is H_t, not the deterministic query register.

For the averaging state, choose in the second S6 layer

$$A = -I_m, \qquad \Delta_\rho = -\log\rho, \qquad a_\Delta = 0, \qquad b_\Delta = \text{softplus}^{-1}(\Delta_\rho).$$

On context tokens, choose $B_\theta(u)$ to select the $G(q,x)$-coordinates of the input to the second S6 layer. Then

$$M_\theta(u) = \exp\{\Delta_\rho A\} = \rho I_m,$$

and

$$N_\theta(u) = \left(\int_0^{\Delta_\rho} e^{-s}\, ds\right) I_m = (1-\rho)I_m.$$

Thus

$$H_t = \rho H_{t-1} + (1-\rho)G(q, X_t).$$

The map $H \mapsto \rho H + (1-\rho)G(q,x)$ is ρ-contractive.

The stationary representation follows by iterating the recursion backward:

$$H_\rho = (1-\rho)\sum_{k=0}^{\infty} \rho^k G(q, X_{-k}).$$

The expectation identity is immediate. Since the X_{-k}'s are independent,

$$\begin{aligned}
\mathbb{E}\|H_\rho - \mathbb{E}H_\rho\|_2^2 &= (1-\rho)^2 \sum_{k=0}^{\infty} \rho^{2k} \mathbb{E}\|G(q, X_{-k}) - \mathbb{E}G(q, X_{-k})\|_2^2 \\
&\le (1-\rho)^2 \sum_{k=0}^{\infty} \rho^{2k} (2B_G)^2 m \\
&= 4mB_G^2 \frac{1-\rho}{1+\rho} \le 4mB_G^2(1-\rho).
\end{aligned}$$

□

C.4. Selector and Mercer-feature approximation

Lemma C.3 (ReLU approximation of the spectral feature map)**.** *Assume Assumption 5.2. Then, for every $D \geq 1$ and every $\eta \in (0,1)$, there exists a ReLU network*

$$\psi_{D,\eta} : X_0 \to \mathbb{R}^D$$

such that

$$\sup_{z \in X_0} \|\psi_{D,\eta}(z) - (e_1(z), \dots, e_D(z))\|_\infty \leq \eta.$$

Moreover, its depth, width, and sparsity are bounded by a polynomial in D and $\log(\eta^{-1})$. More precisely, there exist constants $C, p > 0$, depending only on $d_2, R, R_+, A_0, A_1, \kappa$, such that the network can be chosen with

$$\mathrm{depth}(\psi_{D,\eta}) + \mathrm{width}(\psi_{D,\eta}) + \mathrm{sparsity}(\psi_{D,\eta}) \leq C\left(D + \log(\eta^{-1})\right)^p.$$

The parameter envelope may be chosen so that

$$\log \|\theta_{\psi_{D,\eta}}\|_\infty \leq C\left(D + \log(\eta^{-1})\right)^p.$$

If the stronger bound

$$\sum_{k \in \mathbb{N}_0^{d_2}} |a_{j,k}| R_+^{|k|} \leq A_0 j^\kappa$$

holds, then the parameter envelope itself is polynomial in D and $\log(\eta^{-1})$.

Proof. We use a standard ReLU approximation fact: for every $m \geq 1$ and $\delta \in (0,1)$, there exists a ReLU network

$$\mathcal{M}_{m,\delta} : [-1,1]^{d_2} \to \mathbb{R}^{N_m}, \qquad N_m = \#\{k \in \mathbb{N}_0^{d_2} : |k| \leq m\},$$

whose coordinates approximate all monomials $\{y^k : |k| \leq m\}$ uniformly on $[-1,1]^{d_2}$:

$$\sup_{y \in [-1,1]^{d_2}} \max_{|k| \leq m} \left|[\mathcal{M}_{m,\delta}(y)]_k - y^k\right| \leq \delta.$$

Its depth, width, and sparsity are bounded by a polynomial in m and $\log(\delta^{-1})$. This follows from the usual ReLU multiplication-network construction, combined in parallel to compute all monomials of total degree at most m.

Let

$$\rho := \frac{R}{R_+} < 1, \qquad A_D := A_0 \exp\{A_1 D^\kappa\}.$$

For $1 \leq j \leq D$, write

$$e_j(z) = \sum_{k \in \mathbb{N}_0^{d_2}} a_{j,k} z^k.$$

For $z \in X_0 \subset [-R,R]^{d_2}$, define $y = z/R \in [-1,1]^{d_2}$, so that

$$e_j(z) = \sum_{k \in \mathbb{N}_0^{d_2}} a_{j,k} R^{|k|} y^k.$$

Choose

$$m := \left\lceil \frac{\log(4A_D/\eta)}{-\log \rho} \right\rceil.$$

Then, for every $j \le D$,

$$\begin{aligned}\sup_{z\in X_0}\left|e_j(z) - \sum_{|k|\le m} a_{j,k} z^k\right| &\le \sum_{|k|>m} |a_{j,k}| R^{|k|} \\ &= \sum_{|k|>m} |a_{j,k}| R_+^{|k|} \left(\frac{R}{R_+}\right)^{|k|} \\ &\le \rho^m \sum_{k\in\mathbb{N}_0^{d_2}} |a_{j,k}| R_+^{|k|} \\ &\le \rho^m A_D \le \frac{\eta}{4}.\end{aligned}$$

Now choose

$$\delta := \frac{\eta}{4A_D}.$$

Let $\mathcal{M}_{m,\delta}$ be the monomial network above. For $1 \le j \le D$, define the network output

$$[\psi_{D,\eta}(z)]_j := \sum_{|k|\le m} a_{j,k} R^{|k|} [\mathcal{M}_{m,\delta}(z/R)]_k.$$

Then

$$\begin{aligned}\left|[\psi_{D,\eta}(z)]_j - \sum_{|k|\le m} a_{j,k} z^k\right| &\le \sum_{|k|\le m} |a_{j,k}| R^{|k|} \left|[\mathcal{M}_{m,\delta}(z/R)]_k - (z/R)^k\right| \\ &\le \delta \sum_{k\in\mathbb{N}_0^{d_2}} |a_{j,k}| R^{|k|} \\ &\le \delta \sum_{k\in\mathbb{N}_0^{d_2}} |a_{j,k}| R_+^{|k|} \\ &\le \delta A_D = \frac{\eta}{4}.\end{aligned}$$

Combining this with the Taylor-tail bound gives

$$\sup_{z\in X_0} |[\psi_{D,\eta}(z)]_j - e_j(z)| \le \frac{\eta}{2} \le \eta, \qquad 1 \le j \le D.$$

Taking the maximum over $j \le D$ proves the desired ℓ^∞-approximation.

It remains to check the network size. Since

$$m \le C\left(D^\kappa + \log(\eta^{-1})\right), \qquad \log(\delta^{-1}) \le C\left(D^\kappa + \log(\eta^{-1})\right),$$

the monomial dictionary network has depth, width, and sparsity polynomial in D and $\log(\eta^{-1})$. The final affine layer has at most

$$DN_m = D\binom{m+d_2}{d_2}$$

nonzero coefficients, which is again polynomial in D and $\log(\eta^{-1})$. The final-layer coefficients satisfy

$$|a_{j,k}| R^{|k|} \le \sum_{k\in\mathbb{N}_0^{d_2}} |a_{j,k}| R_+^{|k|} \le A_D,$$

so

$$\log \|\theta_{\psi_{D,\eta}}\|_\infty \le C\left(D^\kappa + \log(\eta^{-1})\right).$$

This proves the claimed complexity bounds. □

Lemma C.4 (ReLU selector for separated tags)**.** *For every $D \geq 1$ and $\eta \in (0,1)$, there exists a tokenwise ReLU map*

$$G_{D,\eta} : \mathbb{R}^{d_1} \times X \to \mathbb{R}^D$$

such that, for every admissible tag system, every query $q = v^{(i_\star)}$, and every $x = (v^{(i)}, z)$,

$$\Big\| G_{D,\eta}(q, v^{(i)}, z) - I\,\mathbf{1}\{i = i_\star\}(e_1(z), \dots, e_D(z)) \Big\|_\infty \leq \eta.$$

Moreover,

$$\|G_{D,\eta}\|_\infty \leq CI,$$

and its depth, width, sparsity, and parameter envelope are bounded by a polynomial in

$$D,\ d_1,\ \log(I/\eta).$$

Proof. Define the ramp function

$$\chi(t) := 2\sigma(t - 1/4) - 2\sigma(t - 3/4),$$

where $\sigma(t) = \max\{t, 0\}$. Then $\chi(t) = 0$ for $t \leq 1/4$ and $\chi(t) = 1$ for $t \geq 3/4$.

For admissible tags,

$$\langle q, v^{(i_\star)} \rangle = 1, \qquad \langle q, v^{(i)} \rangle \leq 0 \quad (i \neq i_\star).$$

A ReLU network can approximate multiplication on compact intervals. Applying this coordinatewise and summing, we obtain a ReLU approximation $\widehat{s}(q, v)$ of $\langle q, v \rangle$ satisfying

$$|\widehat{s}(q, v) - \langle q, v \rangle| \leq 1/8$$

on the compact tag domain, with size polynomial in d_1. Therefore

$$\chi(\widehat{s}(q, v^{(i)})) = \mathbf{1}\{i = i_\star\}$$

on all admissible query–tag pairs.

By Assumption 5.1, there is a ReLU network $\psi_{D,\eta'}$ such that

$$\sup_{z \in X_0} \|\psi_{D,\eta'}(z) - (e_1(z), \dots, e_D(z))\|_\infty \leq \eta',$$

with size polynomial in D and $\log((\eta')^{-1})$. Choosing $\eta' \asymp \eta/I$, and using another standard ReLU multiplication network to multiply the scalar selector by the D feature coordinates, gives

$$G_{D,\eta}(q, v, z) \approx I\,\chi(\widehat{s}(q, v))\psi_{D,\eta'}(z)$$

with uniform error at most η. Uniform boundedness of the eigenfunctions gives $\|G_{D,\eta}\|_\infty \leq CI$. The stated size bound follows from the polynomial ReLU approximation bounds for products and from Assumption 5.1. □

For $X = (V, Z) \sim \nu$ and $q = v^{(i_\star)}$, Lemma C.4 gives

$$\Big\| \int G_{D,\eta}(q, x)\, d\nu(x) - b_D(\mu_{i_\star}) \Big\|_\infty \leq \eta,$$

because

$$\int I\mathbf{1}\{V = q\} e_j(Z)\, d\nu(V, Z) = \int e_j(z)\, d\mu_{i_\star}(z) = b_j(\mu_{i_\star}).$$

C.5. Approximation of the finite-dimensional prediction map

Lemma C.5 (ReLU approximation of the finite prediction map)**.** *Let $D \geq 1$, and let $d_q = d_1$ in the query-dependent case and $d_q = 0$ in the query-independent case. For every $\varepsilon \in (0,1)$, the function f_D from Lemma C.1, restricted to the admissible coefficient domain, can be extended and approximated by a clipped ReLU network*

$$\phi_{D,\varepsilon} : \mathbb{R}^{D+d_q} \to [-M, M]$$

such that

$$\sup_{(b,q)} |\phi_{D,\varepsilon}(b,q) - f_D(b,q)| \leq \varepsilon.$$

The network can be chosen with sparsity

$$s_\phi \leq C\varepsilon^{-C(D+d_q)}$$

and with Lipschitz constant bounded by

$$\mathrm{Lip}(\phi_{D,\varepsilon}) \leq \exp\{C(D+d_q)^C(\log \varepsilon^{-1})^C\}.$$

In the query-independent case, q is omitted and $d_q = 0$.

Proof. The admissible coefficients satisfy

$$\sum_{j=1}^{D} \lambda_j^{-\gamma_b} b_j^2 \leq R_b^2,$$

hence they lie in a cube $[-B_b, B_b]^D$, where B_b depends only on R_b, γ_b, λ_1. Also $q \in \mathbb{S}^{d_1-1} \subset [-1,1]^{d_1}$. By Lemma C.1, f_D is Lipschitz with respect to the Euclidean metric on this compact set, with a Lipschitz constant independent of D. The McShane extension theorem extends f_D to the whole cube without increasing the Lipschitz constant.

Standard ReLU approximation bounds for Lipschitz functions on a compact m-dimensional cube, with $m = D + d_q$, give a ReLU network with sup-norm error ε and sparsity at most $C\varepsilon^{-Cm}$. Clipping the output to $[-M, M]$ can be implemented by ReLU operations and does not increase the sup-norm error or the Lipschitz constant by more than a universal factor. The displayed Lipschitz bound follows from the product of the layer operator norms in the standard construction. □

C.6. Approximation by a stable two-layer Mamba

Lemma C.6 (Uniform approximation by stable Mamba)**.** *For every $\varepsilon \in (0,1)$, there exists a stable two-layer Mamba regressor F_{Θ_ε} such that*

$$\sup_{F_\star \in \mathcal{G}_\star(L,M)} \sup_{(\nu,q) \in \mathfrak{X}_{I,d_1}} |F_{\Theta_\varepsilon}(\nu,q) - F_\star(\nu,q)| \leq C\varepsilon.$$

In the query-dependent case this construction uses final prediction dimension $D_\varepsilon + d_1$. In the query-independent case it uses final prediction dimension D_ε.

Proof. Choose

$$D = D_\varepsilon = \left\lceil C_D(\log \varepsilon^{-1})^{1/\alpha} \right\rceil$$

so that the spectral truncation error in Lemma C.1 is at most $C\varepsilon$.

Let $\phi_{D,\varepsilon}$ be the ReLU network from Lemma C.5, and denote

$$L_\phi := \mathrm{Lip}(\phi_{D,\varepsilon}).$$

Choose

$$\eta = \frac{\varepsilon}{8L_\phi\sqrt{D}}$$

in Lemma C.4. Then

$$\left\| \int G_{D,\eta}(q,x)\, d\nu(x) - b_D(\mu_{i_\star}) \right\|_2 \le \sqrt{D}\,\eta \le \frac{\varepsilon}{8L_\phi}.$$

Use Lemma C.2 with $G = G_{D,\eta}$. Since $\|G_{D,\eta}\|_\infty \le CI$, its stationary state H_ρ satisfies

$$\mathbb{E}\|H_\rho - \mathbb{E}H_\rho\|_2^2 \le CDI^2(1-\rho).$$

Choose

$$1-\rho \le \frac{\varepsilon^2}{CDI^2L_\phi^2}.$$

Then

$$\mathbb{E}\|H_\rho - \mathbb{E}H_\rho\|_2 \le \frac{\varepsilon}{8L_\phi}.$$

Therefore

$$\begin{aligned}\mathbb{E}\|H_\rho - b_D(\mu_{i_\star})\|_2 &\le \|\mathbb{E}H_\rho - b_D(\mu_{i_\star})\|_2 + \mathbb{E}\|H_\rho - \mathbb{E}H_\rho\|_2 \\ &\le \frac{\varepsilon}{4L_\phi}.\end{aligned}$$

The infinite-context Mamba output is

$$F_{\Theta_\varepsilon}(\nu, q) = \mathbb{E}\big[\phi_{D,\varepsilon}(H_\rho, q)\big]$$

in the query-dependent case, and

$$F_{\Theta_\varepsilon}(\nu, q) = \mathbb{E}\big[\phi_{D,\varepsilon}(H_\rho)\big]$$

in the query-independent case. Hence

$$\begin{aligned}|F_{\Theta_\varepsilon}(\nu,q) - f_D(b_D(\mu_{i_\star}),q)| &\le \mathbb{E}\,|\phi_{D,\varepsilon}(H_\rho,q) - \phi_{D,\varepsilon}(b_D(\mu_{i_\star}),q)| \\ &\quad + |\phi_{D,\varepsilon}(b_D(\mu_{i_\star}),q) - f_D(b_D(\mu_{i_\star}),q)| \\ &\le L_\phi \mathbb{E}\|H_\rho - b_D(\mu_{i_\star})\|_2 + \varepsilon \\ &\le C\varepsilon.\end{aligned}$$

Combining this with Lemma C.1 gives the claim. The constructed recurrence has contraction coefficient $\rho < 1$, so Assumption 4.1 holds. □

C.7. Covering entropy of the constructed Mamba class

Let $\mathcal{H}_\varepsilon^{\mathrm{Mam}}$ be the stable two-layer Mamba class with the architecture and parameter envelopes used in Lemma C.6. The class contains all networks of the same depth, widths, sparsities, contraction lower bound $1-\rho$, and parameter envelopes as in the construction.

Lemma C.7 (Entropy bound)**.** *For every $\delta \in (0,1)$,*

$$\log N\big(\mathcal{H}_\varepsilon^{\mathrm{Mam}}, \delta, \|\cdot\|_\infty\big) \le CS_\varepsilon \log\left(\frac{CB_\varepsilon\Lambda_\varepsilon}{\delta}\right),$$

where S_ε is the number of nonzero scalar parameters, B_ε is the parameter envelope, and Λ_ε is a uniform parameter-Lipschitz constant of the input–output map. Moreover:

$$\begin{aligned}&\textit{query-dependent case:} && S_\varepsilon \le \exp\{C(D_\varepsilon + d_1)\log(\varepsilon^{-1})\}; \\ &\textit{query-independent case:} && S_\varepsilon \le d_1^C \exp\{CD_\varepsilon \log(\varepsilon^{-1})\}.\end{aligned}$$

The logarithmic factor satisfies

$$\log(B_\varepsilon\Lambda_\varepsilon) \le C(1+\log I) + C(D_\varepsilon + d_1)^C(\log\varepsilon^{-1})^C$$

in the query-dependent case, and the same bound with D_ε replacing $D_\varepsilon + d_1$ in the final MLP part in the query-independent case.

Proof. The covering argument is by scalar-parameter discretization. For a network with S_ε nonzero parameters, each bounded by B_ε, and with input–output map Λ_ε-Lipschitz in the parameter vector, a grid of mesh

$$\frac{\delta}{CS_\varepsilon\Lambda_\varepsilon}$$

on each active scalar parameter gives a δ-cover in sup norm. Thus

$$\log N \le CS_\varepsilon \log\left(\frac{CB_\varepsilon\Lambda_\varepsilon}{\delta}\right).$$

The parameter count is obtained by adding the sizes of the selector network, the Mercer-feature network, the averaging S6 block, and the final ReLU network. By Assumption 5.1 and Lemma C.4, the selector and feature networks have size polynomial in

$$D_\varepsilon,\ d_1,\ \log(I/\eta).$$

The averaging S6 block has $O(D_\varepsilon)$ active recurrent parameters. The dominant term is the final Lipschitz-function approximator. By Lemma C.5, its sparsity is at most

$$\varepsilon^{-C(D_\varepsilon+d_1)}$$

in the query-dependent case and at most

$$\varepsilon^{-CD_\varepsilon}$$

in the query-independent case. This gives the displayed bounds for S_ε, with only a polynomial d_1^C factor in the query-independent case.

It remains to justify that $\log(B_\varepsilon\Lambda_\varepsilon)$ has the stated size. The only potentially large parameters are those used to set the contraction coefficient ρ and those in the final ReLU network. From Lemma C.6,

$$1-\rho \gtrsim \frac{\varepsilon^2}{D_\varepsilon I^2 L_\phi^2}.$$

Thus

$$\log\frac{1}{1-\rho} \le C(1+\log I) + C\log D_\varepsilon + C\log\varepsilon^{-1} + C\log L_\phi.$$

Lemma C.5 gives

$$\log L_\phi \le C(D_\varepsilon + d_q)^C(\log\varepsilon^{-1})^C,$$

where $d_q = d_1$ in the query-dependent case and $d_q = 0$ in the query-independent case.

Finally, differentiating the stable recursion with respect to any scalar parameter gives an inequality of the form

$$R_t \le \rho R_{t-1} + C_\varepsilon,$$

so that

$$\sup_t R_t \le \frac{C_\varepsilon}{1-\rho}.$$

It remains to justify the parameter-Lipschitz constant Λ_ε. We give the argument at the level of the infinite-context maps.

Let θ, θ' be two admissible parameter vectors in the same architecture class. Write $K_{\theta,\nu,q}$ and $K_{\theta',\nu,q}$ for the corresponding context-state Markov kernels, and write $\pi_{\theta,\nu,q}$ and $\pi_{\theta',\nu,q}$ for their invariant laws. The class is constructed so that every context transition is ρ_ε-contractive in the state variable, where $\rho_\varepsilon < 1$ and $1-\rho_\varepsilon$ is bounded from below by the quantity chosen in Lemma C.6.

Moreover, on the compact invariant state set and compact token domain, the finite-dimensional network maps are Lipschitz in their scalar parameters. Thus there is a constant $L_{\Psi,\varepsilon}$ such that

$$\sup_{h,w,q} \|\Psi_\theta(h,w,q) - \Psi_{\theta'}(h,w,q)\|_2 \le L_{\Psi,\varepsilon}\|\theta-\theta'\|_1.$$

Similarly, if $G_\theta(h,q)$ denotes the deterministic final query/readout map, then

$$\sup_{h,q} |G_\theta(h,q) - G_{\theta'}(h,q)| \le L_{G,\varepsilon}\|\theta-\theta'\|_1,$$

and $G_\theta(\cdot,q)$ is $L_{\mathrm{out},\varepsilon}$-Lipschitz in h.

We now compare the invariant laws. Couple the two chains using the same context token $W \sim \bar{\nu}$. For any coupling γ of $\pi_{\theta,\nu,q}$ and $\pi_{\theta',\nu,q}$, if $(H,H') \sim \gamma$, then

$$\begin{aligned}
&\mathbb{E}\,\|\Psi_\theta(H,W,q) - \Psi_{\theta'}(H',W,q)\|_2 \\
&\quad \le \mathbb{E}\,\|\Psi_\theta(H,W,q) - \Psi_\theta(H',W,q)\|_2 + \mathbb{E}\,\|\Psi_\theta(H',W,q) - \Psi_{\theta'}(H',W,q)\|_2 \\
&\quad \le \rho_\varepsilon \mathbb{E}\|H-H'\|_2 + L_{\Psi,\varepsilon}\|\theta-\theta'\|_1.
\end{aligned}$$

Taking the infimum over γ gives

$$W_1(\pi_{\theta,\nu,q}K_{\theta,\nu,q}, \pi_{\theta',\nu,q}K_{\theta',\nu,q}) \le \rho_\varepsilon W_1(\pi_{\theta,\nu,q}, \pi_{\theta',\nu,q}) + L_{\Psi,\varepsilon}\|\theta-\theta'\|_1.$$

Using invariance of both laws,

$$W_1(\pi_{\theta,\nu,q}, \pi_{\theta',\nu,q}) \le \frac{L_{\Psi,\varepsilon}}{1-\rho_\varepsilon}\|\theta-\theta'\|_1.$$

Therefore

$$\begin{aligned}
|F_\theta(\nu,q) - F_{\theta'}(\nu,q)| &= \left| \int G_\theta(h,q)\, d\pi_{\theta,\nu,q}(h) - \int G_{\theta'}(h,q)\, d\pi_{\theta',\nu,q}(h) \right| \\
&\le L_{G,\varepsilon}\|\theta-\theta'\|_1 + L_{\mathrm{out},\varepsilon} W_1(\pi_{\theta,\nu,q}, \pi_{\theta',\nu,q}) \\
&\le \left(L_{G,\varepsilon} + \frac{L_{\mathrm{out},\varepsilon} L_{\Psi,\varepsilon}}{1-\rho_\varepsilon} \right) \|\theta-\theta'\|_1.
\end{aligned}$$

Hence the infinite-context map is parameter-Lipschitz with

$$\Lambda_\varepsilon := L_{G,\varepsilon} + \frac{L_{\mathrm{out},\varepsilon} L_{\Psi,\varepsilon}}{1-\rho_\varepsilon}.$$

The constants $L_{\Psi,\varepsilon}, L_{G,\varepsilon}$, and $L_{\mathrm{out},\varepsilon}$ grow at most polynomially in the layer widths, the parameter envelope, and $(1-\rho_\varepsilon)^{-1}$. Therefore

$$\log \Lambda_\varepsilon \le C(1+\log I) + C(D_\varepsilon + d_q)^C (\log \varepsilon^{-1})^C,$$

where $d_q = d_1$ in the query-dependent case and $d_q = 0$ in the query-independent final prediction network.

Thus the parameter-Lipschitz constant grows at most polynomially in $(1-\rho)^{-1}$, in the layer widths, and in the parameter envelopes. Taking logs yields the stated bound. □

C.8. ERM oracle inequality

We use the following standard bounded-sieve least-squares inequality.

Lemma C.8 (ERM bound)**.** *Let*

$$Y = F_\star(X) + \xi, \qquad \xi \sim N(0,\sigma^2),$$

and let $\mathcal{H}$ be a class of functions uniformly bounded by B. Let $\widehat{F}$ be an exact empirical risk minimizer over $\mathcal{H}$. Then, for every $\delta \in (0,1)$,

$$\mathbb{E}\,\|\widehat{F} - F_\star\|^2_{L^2(P_X)} \le C\left[\inf_{F\in\mathcal{H}} \|F - F_\star\|^2_{L^2(P_X)} + \frac{(B^2+\sigma^2)\log N(\mathcal{H},\delta,\|\cdot\|_\infty)}{n} + (B+\sigma)\delta \right].$$

C.9. Proof of Theorem 6.1

Proof. Let

$$a := \frac{\alpha}{\alpha+1}.$$

Choose a sufficiently small constant $\kappa > 0$, to be fixed below, and set

$$\varepsilon_n := \exp\{-\kappa(\log n)^a\}.$$

Let

$$D_n := D_{\varepsilon_n} = \left\lceil C_D(\log \varepsilon_n^{-1})^{1/\alpha} \right\rceil.$$

Since

$$\log \varepsilon_n^{-1} = \kappa(\log n)^a,$$

we have

$$D_n \asymp (\log n)^{1/(\alpha+1)}.$$

Let $\mathcal{H}_n^{\mathrm{Mam}} := \mathcal{H}_{\varepsilon_n}^{\mathrm{Mam}}$. By Lemma C.6,

$$\sup_{F_\star \in \mathcal{G}_\star(L,M)} \inf_{F \in \mathcal{H}_n^{\mathrm{Mam}}} \|F - F_\star\|_\infty \le C\varepsilon_n.$$

Therefore

$$\sup_{F_\star \in \mathcal{G}_\star(L,M)} \inf_{F \in \mathcal{H}_n^{\mathrm{Mam}}} \|F - F_\star\|_{L^2(P_{\nu,q})}^2 \le C\varepsilon_n^2.$$

We now bound the entropy at scale $\delta_n = \varepsilon_n^2$.

Query-dependent case. Assume

$$I \le d_1 \le C_I(\log n)^{1/(\alpha+1)}.$$

Since $D_n \asymp (\log n)^{1/(\alpha+1)}$, we have $d_1 \lesssim D_n$. By Lemma C.7,

$$\log N(\mathcal{H}_n^{\mathrm{Mam}}, \delta_n, \|\cdot\|_\infty) \le \exp\{CD_n \log(\varepsilon_n^{-1})\}.$$

Using the definitions of D_n and ε_n,

$$D_n \log(\varepsilon_n^{-1}) \lesssim \kappa^{(\alpha+1)/\alpha} \log n.$$

Thus

$$\log N(\mathcal{H}_n^{\mathrm{Mam}}, \delta_n, \|\cdot\|_\infty) \le n^{C\kappa^{(\alpha+1)/\alpha}}.$$

Choosing $\kappa > 0$ sufficiently small gives

$$\log N(\mathcal{H}_n^{\mathrm{Mam}}, \delta_n, \|\cdot\|_\infty) \le n^{1/2}$$

for all large n.

Query-independent case. Assume

$$\widetilde{F}_\star(\mu, q) = \widetilde{F}_\star(\mu), \qquad I \le d_1 = n^{o(1)}.$$

The final prediction network now has input dimension D_n, not $D_n + d_1$. Lemma C.7 gives

$$\log N(\mathcal{H}_n^{\mathrm{Mam}}, \delta_n, \|\cdot\|_\infty) \le d_1^C \exp\{CD_n \log(\varepsilon_n^{-1})\}.$$

Since $d_1 = n^{o(1)}$ and

$$\exp\{CD_n \log(\varepsilon_n^{-1})\} \le n^{C\kappa^{(\alpha+1)/\alpha}},$$

choosing $\kappa > 0$ sufficiently small yields

$$\log N(\mathcal{H}_n^{\mathrm{Mam}}, \delta_n, \|\cdot\|_\infty) \le n^{1/2+o(1)} \le n^{2/3}$$

for all sufficiently large n. This term is still negligible compared with the target sub-polynomial rate.

Applying Lemma C.8 with $\mathcal{H} = \mathcal{H}_n^{\mathrm{Mam}}$, $B \asymp M$, and $\delta_n = \varepsilon_n^2$, we obtain in the query-dependent case

$$\sup_{F_\star \in \mathcal{G}_\star(L,M)} \mathbb{E}\,\mathcal{R}(\widehat{F}_n, F_\star) \le C\varepsilon_n^2 + C\frac{n^{1/2}}{n} + C\varepsilon_n^2$$
$$\le C\exp\{-2\kappa(\log n)^a\} + Cn^{-1/2}.$$

Since $a < 1$, $n^{-1/2} \le \exp\{-c(\log n)^a\}$ for all sufficiently large n. Therefore

$$\sup_{F_\star \in \mathcal{G}_\star(L,M)} \mathbb{E}\,\mathcal{R}(\widehat{F}_n, F_\star) \le C\exp\{-c(\log n)^{\alpha/(\alpha+1)}\}.$$

The same argument in the query-independent case gives the identical rate, because

$$n^{-1/3} \le \exp\{-c(\log n)^a\}$$

for all sufficiently large n. Enlarging C handles the finitely many small values of n. This proves the theorem. □

D. Proof of Theorem 6.3

We prove the lower bound by a finite-packing argument. The proof uses only a query-independent subclass of $\mathcal{G}_\star(L, M)$, and hence applies a fortiori to the full recall–predict target class.

Throughout this proof, write

$$s := \frac{\gamma_d - \gamma_f}{2} > 0.$$

Let $b_j^0 := \int e_j(z)p_0(z)\,dz$. Under Assumption 6.2, the Mercer coefficients of a random content measure μ have the form

$$b_j(\mu) = b_j^0 + a_0\lambda_j^{\gamma_d/2}Z_j, \qquad j \ge 1,$$

where Z_j are independent, $|Z_j| \le 1$, and Z_j has density ρ_j satisfying $\sup_j \|\rho_j\|_\infty \le R$. Let Φ_j denote the distribution function of Z_j. Since ρ_j is bounded by R, each Φ_j is R-Lipschitz. Moreover, by the probability integral transform,

$$U_j := \Phi_j(Z_j), \qquad j \ge 1,$$

are independent $\mathrm{Unif}[0,1]$ random variables.

For $d \in \mathbb{N}$, define the coefficient-to-cube map

$$H_d(\mu) := \left(\Phi_1\left(\frac{b_1(\mu) - b_1^0}{a_0\lambda_1^{\gamma_d/2}}\right), \dots, \Phi_d\left(\frac{b_d(\mu) - b_d^0}{a_0\lambda_d^{\gamma_d/2}}\right)\right) \in [0,1]^d.$$

For two admissible content measures μ, μ', using the Lipschitzness of Φ_j gives

$$\|H_d(\mu) - H_d(\mu')\|_\infty \le \max_{1\le j\le d} \frac{R}{a_0}\lambda_j^{-\gamma_d/2}|b_j(\mu) - b_j(\mu')|.$$

On the other hand,

$$|b_j(\mu) - b_j(\mu')| \le \lambda_j^{\gamma_f/2}\|\mu - \mu'\|_{H_0^{\gamma_f}},$$

because

$$\|\mu - \mu'\|_{H_0^{\gamma_f}}^2 = \sum_{k\ge1} \lambda_k^{-\gamma_f}\big(b_k(\mu) - b_k(\mu')\big)^2.$$

Since λ_j is nonincreasing and $\gamma_f - \gamma_d < 0$, for $1 \le j \le d$,

$$\lambda_j^{(\gamma_f-\gamma_d)/2} \le \lambda_d^{(\gamma_f-\gamma_d)/2} = \lambda_d^{-s}.$$

Therefore

$$\boxed{\|H_d(\mu) - H_d(\mu')\|_\infty \le \frac{R}{a_0}\lambda_d^{-s}\|\mu - \mu'\|_{H_0^{\gamma_f}}.} \tag{A.1}$$

We shall use the following elementary packing lemma for finite-dimensional Lipschitz functions.

Lemma D.1 (A packing of Lipschitz functions on the cube)**.** *There exist universal constants $c_0, c_1, c_2 > 0$ such that, for every $d \ge 1$ and every integer $m \ge 2$, there is a finite set $\Theta_{d,m}$ and functions*

$$g_\theta : [0,1]^d \to [0,1], \qquad \theta \in \Theta_{d,m},$$

such that

$$\mathrm{Lip}_{\|\cdot\|_\infty}(g_\theta) \le 1 \quad \textit{for all } \theta \in \Theta_{d,m},$$
$$\log|\Theta_{d,m}| \ge c_0 m^d,$$

and, for all distinct $\theta, \theta' \in \Theta_{d,m}$,

$$\frac{c_1}{dm} \le \|g_\theta - g_{\theta'}\|_{L^2([0,1]^d)} \le \frac{c_2}{dm}.$$

Proof. Partition $[0,1]^d$ into m^d cubes Q_k of side length m^{-1}, and let x_k be the center of Q_k. Define the tent bump

$$\psi_k(x) := \frac{1}{4m}\big(1 - 2m\|x - x_k\|_\infty\big)_+.$$

The supports of the ψ_k's are disjoint up to boundaries, and each ψ_k is $1/2$-Lipschitz with respect to $\|\cdot\|_\infty$. Hence, for every binary vector $\omega \in \{0,1\}^{m^d}$,

$$g_\omega(x) := \sum_{k=1}^{m^d} \omega_k \psi_k(x)$$

takes values in $[0,1]$ and is 1-Lipschitz with respect to $\|\cdot\|_\infty$.

By the Varshamov–Gilbert bound, there is a subset $\Theta_{d,m} \subset \{0,1\}^{m^d}$ such that

$$\log|\Theta_{d,m}| \ge c_0 m^d$$

and the Hamming distance between any two distinct elements of $\Theta_{d,m}$ is at least a fixed positive fraction of m^d. A direct calculation gives

$$\|\psi_k\|^2_{L^2([0,1]^d)} = \frac{1}{16m^2}\int_{\|u\|_\infty \le 1/(2m)} \big(1 - 2m\|u\|_\infty\big)^2\, du \asymp \frac{1}{d^2 m^{d+2}},$$

where the constants are universal. Since the bump supports are disjoint, the claimed lower and upper L^2-distance bounds follow. □

We now embed the packing from Lemma D.1 into the recall–predict class. Choose

$$\kappa := \frac{1}{2}\min\left\{\frac{La_0}{R}, \frac{M}{\max\{1, \lambda_1^s\}}\right\}, \qquad A_d := \kappa\lambda_d^s.$$

For $\theta \in \Theta_{d,m}$, define

$$\widetilde{F}_\theta(\mu, q) := A_d\, g_\theta(H_d(\mu)).$$

This functional is independent of q. By $(A.1)$,

$$\begin{aligned}
|\widetilde{F}_\theta(\mu, q) - \widetilde{F}_\theta(\mu', q')| &\le A_d\|H_d(\mu) - H_d(\mu')\|_\infty \\
&\le \kappa\lambda_d^s \frac{R}{a_0}\lambda_d^{-s}\|\mu - \mu'\|_{H_0^{\gamma_f}} \\
&\le L\|\mu - \mu'\|_{H_0^{\gamma_f}} \le L\Big(\|\mu - \mu'\|_{H_0^{\gamma_f}} + \|q - q'\|_2\Big).
\end{aligned}$$

Also $|\widetilde{F}_\theta| \le A_d \le M$. Hence each $\widetilde{F}_\theta$ defines an element of $\mathcal{G}_\star(L, M)$ by

$$F_\theta(\nu, q) := \widetilde{F}_\theta(\mu_{i_\star}, q),$$

where $\mu_{i_\star}$ is the content measure selected by the query.

Let $P_\theta^{(n)}$ denote the law of the training sample S_n when the regression function is F_θ. The covariate law of (ν, q) is the same for every θ; only the conditional mean of Y changes. Since the noise is $N(0, \sigma^2)$,

$$D_{\mathrm{KL}}\Big(P_\theta^{(n)} \,\Big\|\, P_{\theta'}^{(n)}\Big) = \frac{n}{2\sigma^2}\|F_\theta - F_{\theta'}\|^2_{L^2(P_{\nu,q})}.$$

Because $H_d(\mu_{i_\star}) \sim \mathrm{Unif}([0,1]^d)$, the isometry identity

$$\|F_\theta - F_{\theta'}\|_{L^2(P_{\nu,q})} = A_d\,\|g_\theta - g_{\theta'}\|_{L^2([0,1]^d)}$$

holds. Therefore Lemma D.1 implies that, for distinct θ, θ',

$$\underline{\delta}_{d,m} := \frac{c_1 A_d}{dm} \le \|F_\theta - F_{\theta'}\|_{L^2(P_{\nu,q})} \le \frac{c_2 A_d}{dm} =: \overline{\delta}_{d,m}. \tag{A.2}$$

Consequently,

$$D_{\mathrm{KL}}\Big(P_\theta^{(n)} \,\Big\|\, P_{\theta'}^{(n)}\Big) \le \frac{n\overline{\delta}^2_{d,m}}{2\sigma^2}. \tag{A.3}$$

We now choose d and m. Fix any $\eta > 0$ and set

$$m_d := \left\lfloor e^{\eta d^\alpha} \right\rfloor.$$

For all sufficiently large d,

$$\log|\Theta_{d,m_d}| \ge c_0 m_d^d \ge c\,\exp\Big(\frac{\eta}{2} d^{\alpha+1}\Big). \tag{A.4}$$

Let

$$d_n := \left\lfloor \left(\frac{4}{\eta}\log n\right)^{1/(\alpha+1)} \right\rfloor, \qquad m_n := m_{d_n}.$$

Then $(A.4)$ gives

$$\log|\Theta_{d_n,m_n}| \ge cn^2$$

for all sufficiently large n. On the other hand, by $(A.3)$,

$$\max_{\theta\ne\theta'} D_{\mathrm{KL}}\Big(P_\theta^{(n)} \,\Big\|\, P_{\theta'}^{(n)}\Big) \le Cn,$$

where C depends only on the fixed problem constants. Hence, for all large n,

$$\max_{\theta\ne\theta'} D_{\mathrm{KL}}\Big(P_\theta^{(n)} \,\Big\|\, P_{\theta'}^{(n)}\Big) \le \frac{1}{8}\log|\Theta_{d_n,m_n}|. \tag{A.5}$$

Fano's inequality applied to the finite family $\{F_\theta : \theta \in \Theta_{d_n,m_n}\}$ now yields

$$\begin{aligned}\inf_{\widehat{F}_n}\sup_{\theta\in\Theta_{d_n,m_n}} \mathbb{E}_\theta\|\widehat{F}_n - F_\theta\|^2_{L^2(P_{\nu,q})} &\ge \frac{\underline{\delta}^2_{d_n,m_n}}{4}\left[1 - \frac{\max_{\theta\ne\theta'} D_{\mathrm{KL}}(P_\theta^{(n)}\|P_{\theta'}^{(n)}) + \log 2}{\log|\Theta_{d_n,m_n}|}\right] \\ &\ge c\,\underline{\delta}^2_{d_n,m_n}.\end{aligned}$$

Since the finite family is contained in $\mathcal{G}_\star(L, M)$, this lower bounds the minimax risk over the full target class.

It remains to compute the order of $\underline{\delta}^2_{d_n,m_n}$. By the lower eigenvalue bound in Assumption 5.1,

$$\lambda_d \ge c_\lambda e^{-C_0 d^\alpha}.$$

Thus

$$A_d = \kappa \lambda_d^s \geq c\, e^{-sC_0 d^\alpha}.$$

Using $m_d \leq e^{\eta d^\alpha}$, we get

$$\underline{\delta}_{d,m_d}^2 = \left(\frac{c_1 A_d}{d m_d}\right)^2 \geq c\, d^{-2} \exp\{-2(sC_0 + \eta) d^\alpha\}.$$

With $d = d_n$,

$$d_n^\alpha \asymp (\log n)^{\alpha/(\alpha+1)}.$$

The factor d_n^{-2} is absorbed into the exponential because $\log d_n = o((\log n)^{\alpha/(\alpha+1)})$. Hence there exist constants $c, C > 0$, independent of n, such that

$$\underline{\delta}_{d_n,m_n}^2 \geq c \exp\Big\{-C(\log n)^{\alpha/(\alpha+1)}\Big\}.$$

Therefore

$$\mathfrak{M}_n = \inf_{\widehat{F}_n} \sup_{F_\star \in \mathcal{G}_\star(L,M)} \mathbb{E}\, \mathcal{R}(\widehat{F}_n, F_\star) \geq c \exp\Big\{-C(\log n)^{\alpha/(\alpha+1)}\Big\},$$

for all sufficiently large n. Adjusting c handles finitely many smaller values of n. This proves the theorem.