

State space models, emergence, and ergodicity: How many parameters are needed for stable predictions?

Ingvar Ziemann

University of Pennsylvania

INGVARZ@SEAS.UPENN.EDU

Nikolai Matni

University of Pennsylvania

NMATMI@SEAS.UPENN.EDU

George J. Pappas

University of Pennsylvania

PAPPASG@SEAS.UPENN.EDU

Editors: A. Abate, L. Balzano, N. Ozay, D. Panagou

Abstract

How many parameters are required for a model to execute a given task? It has been argued that large language models, pre-trained via self-supervised learning, exhibit emergent capabilities such as multi-step reasoning as their parameter count reaches a critical scale. In the present work, we explore whether this phenomenon can analogously be replicated in a simple theoretical model. We show that the problem of learning linear dynamical systems—a simple instance of self-supervised learning—exhibits a corresponding phase transition. Namely, for every non-ergodic linear system there exists a critical threshold such that a learner using fewer parameters than said threshold cannot achieve bounded error for large sequence lengths. Put differently, in our model we find that tasks exhibiting substantial long-range correlation require a certain critical number of parameters—a phenomenon akin to emergence. We also investigate the role of the learner’s parametrization and consider a simple version of a linear dynamical system with hidden state—an imperfectly observed random walk in \mathbb{R} . For this situation, we show that there exists no learner using a linear filter which can successfully learn the random walk unless the filter length exceeds a certain threshold depending on the effective memory length and horizon of the problem.

1. Introduction

Consider a pre-trained large language model (LLM) obtained via self-supervised learning by predicting the next word or token. While the performance on pre-training loss exhibits rather predictable behavior (Kaplan et al., 2020), Wei et al. (2022) observe that such models often exhibit a phase transition in their downstream capabilities as the number of trainable parameters (or training FLOPs) reaches a critical scale—they exhibit emergent capabilities such as successful in-context learning (Brown et al., 2020). While these models are typically extremely large in terms of their number of parameters, a recent line of work has shown that such behavior can also be recovered in smaller models by considering appropriately simplified tasks (Allen-Zhu and Li, 2024). Here, we offer a possible mechanistic explanation for this phenomenon by restricting to a simple class of auto-regressive learning models.

Namely, we point out that certain tasks—or more precisely, predicting in certain generative models—exhibiting long-range correlations and a lack of ergodicity can only be executed successfully once model scale reaches a certain critical threshold. One may think of our result as the bias term in the bias-variance trade-off exhibiting a sharp jump—a phase transition—depending on whether the model class is rich enough to be fully descriptive of this lack of stochastic stability. We illustrate

this phenomenon by a simple problem: learning a linear dynamical system. Incidentally, such linear systems are also fundamental building blocks in the increasingly popular state space model architectures for sequence modeling—an alternative to the popular transformer architecture (Vaswani et al., 2017; Gu et al., 2022).

Here and in the sequel we study generative modelling of tasks P_Z corresponding to distributions over sequences of tokens $Z_{1:T}$. A learner has pre-trained a (compressed) generative model Q_Z using data not necessarily coming from P_Z . The performance of such a model Q_Z on a task P_Z will be measured by its divergence from the ground truth:

$$d_{\text{KL}}(P_Z \| Q_Z) = \mathbb{E}_P \log \frac{dP_Z}{dQ_Z}. \quad (1)$$

We ask the following question:

Q: *Suppose that Q comes from a parametric hypothesis class. Does there exist a critical threshold in terms of the number of parameters such that $T^{-1}d_{\text{KL}}(P_Z \| Q_Z) \rightarrow \infty$ as $T \rightarrow \infty$ unless the parameter count exceeds said threshold?*

In other words, we ask whether a given task-hypothesis class combination admits *stable learners*—learners for which the KL-risk does not diverge as the sequence length T becomes long (notice that the normalization T^{-1} is necessary to avoid trivial behavior for product measures). Our view here is that language, arriving in discrete packages such as articles and books, is non-ergodic when viewed at the package level. In this view, a single book forms a single trajectory of data in which the first word (or token) is the first data point and the last word the last data point. The distribution of words in the beginning of the book (introducing the suspects) may well be quite different from the distribution at the end of the book (who did it?)—there is different meaning to be conveyed.

It is our hypothesis that it is exactly this lack of ergodicity that leads to emergent behavior. Our main simplifying assumption in relating non-ergodicity to model complexity is that the task P_Z has a latent state space model representation.

Assumption 1.1 *The data $Z_{1:T}$ is in bijection with a state space model. More precisely, there exists a bijection g such that $Z_t = g(Y_t)$ for $t \in [T]$ where $Y_{1:T}$ is generated by:*

$$X_{t+1} = A_* X_t + W_t, \quad X_1 = W_0 \quad Y_t = C_* X_t + V_t. \quad (2)$$

where $A_* \in \mathbb{R}^{d_X \times d_X}$, $C_* \in \mathbb{R}^{d_Y \times d_X}$. Here, $W_{0:T}$ and $V_{1:T}$ are jointly Gaussian, mutually independent with block-diagonal covariance $(\Sigma_{W_1}, \Sigma_W \otimes I_T, \Sigma_V \otimes I_T)$ and mean zero.

Models of this form are standard in time series prediction tasks and systems modelling, but have also recently been popularized as building blocks in LLMs (Gu et al., 2022).

Under Assumption 1.1, a version of the maximum entropy principle yields the following. For every nondegenerate distribution Q_Z over $Z_{1:T}$ under Assumption 1.1 the following are true:

- For $Y_{1:T} \sim P_Y = P_{g^{-1}(Z)}$ then:

$$d_{\text{KL}}(P_Z \| Q_Z) = d_{\text{KL}}(P_Y \| Q_{g^{-1}(Z)}). \quad (3)$$

- The Gaussian measure Q_Y with the same mean and covariance as $Q_{g^{-1}(Z)}$ satisfies

$$d_{\text{KL}}(P_Y \| Q_{g^{-1}(Z)}) \geq d_{\text{KL}}(P_Y \| Q_Y). \quad (4)$$

The first statement follows by bijection and the second statement is simply observing that Gaussian measures minimize KL subject to constraints on the first two moments. Our next observation is the standard (trivial yet powerful!) equivalence between generative modeling and next-token-prediction. Namely the generative modelling error on the left hand side below can be expanded in terms of the KL divergence chain rule:

$$\begin{aligned} d_{\text{KL}}(P_Y \| Q_Y) &= \sum_{t=1}^T \mathbf{E}_P \log \frac{dP_Y^{t|1:t-1}}{dQ_Y^{t|1:t-1}} \\ &= \frac{1}{2} \sum_{t=1}^T \left[\|\mathbf{E}_P^{t-1} Y_t - \mathbf{E}_Q^{t-1} Y_t\|_{\Sigma_{Q_t}^{-1/2}}^2 + \text{tr} \left(\Sigma_{Q_t}^{-1} \Sigma_{P_t} \right) - \log \det \left(\Sigma_{Q_t}^{-1} \Sigma_{P_t} \right) - d_Y \right]. \end{aligned} \quad (5)$$

It is reasonable to assume that $mI \preceq \Sigma_{Q_t} \preceq MI$ for some universal constants m, M . Otherwise, either the term $\text{tr} \left(\Sigma_{Q_t}^{-1} \Sigma_{P_t} \right)$ grows unbounded (as we will see that Σ_{P_t} is well-conditioned in our examples), or the variance of the predictor becomes arbitrarily large. Combining the above we have that

$$d_{\text{KL}}(P_Z \| Q_Z) \gtrsim \sum_{t=1}^T \mathbf{E}_P \|\mathbf{E}_P^{t-1} Y_t - \mathbf{E}_Q^{t-1} Y_t\|^2. \quad (6)$$

It will be convenient to denote

$$\ell_T(\mathcal{F}, \mathcal{P}) \triangleq \inf_{Q \in \mathcal{F}} \sup_{P \in \mathcal{P}} \mathbf{E}_P \sum_{t=1}^{T-1} \|\mathbf{E}_P^{t-1} Y_t - \mathbf{E}_Q^{t-1} Y_t\|^2. \quad (7)$$

By the above reasoning via (5)-(6), ℓ_T defined above in (7) constitutes a lower bound on the KL-divergence risk (1) in which a learner—by picking a hypothesis in \mathcal{F} —competes with an adversary selecting a generative model from \mathcal{P} . Thus, imposing these additional constraints above, an instantiation of the above question **Q** becomes as follows.

Q’: Fix a family of parametric hypothesis classes $\{\mathcal{F}_d\}_{d \in \mathbb{N}}$ and a family of possible generative models \mathcal{P} . Does there exist a critical threshold d_* in terms of the number parameters such that

$$T^{-1} \ell_T(\mathcal{F}_d, \mathcal{P}) \rightarrow \infty$$

as $T \rightarrow \infty$ unless the parameter count exceeds said threshold ($d > d_*$)?

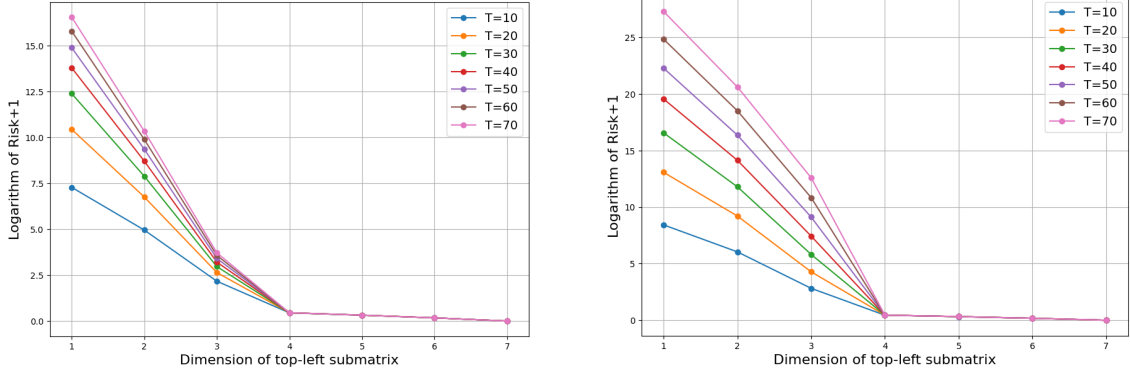
In the sequel we focus on identifying task-hypothesis pairs $(\mathcal{P}, \{\mathcal{F}_d\}_{d \in \mathbb{N}})$ where this divergence occurs. We will think of a task as exhibiting emergent behavior if it admits a nontrivial threshold d_* mentioned in **Q’** above.

Finally, before we proceed let us also remark that there is some degree of necessity to our choice of considering an adversarial model class \mathcal{P} that we use to obtain meaningful lower bounds. To make this concrete, consider a parametric class of distributions \mathcal{P} parametrized by some set of parameters, say $\theta \in \mathcal{P}$. Suppose the generative model corresponds to the parameter θ_* . As long as \mathcal{F} contains this parameter the only lower bound that can be obtained without including the supremum in (7) is 0. In other words, we need to model the fact that the learner does not have access to the parameter a priori. We accomplish this by letting an adversary pick a parameter against which the learner must compete.

2. Contribution

Our contributions can be stated informally as follows.

Theorem [Informal version of Theorem 2] *There is emergent behavior in learning non-ergodic auto-regressive models: in a simple linear dynamical system with fully observed state, there exists no successful learner without using at least as many parameters as the square of the number of (marginally) unstable eigenvalues.¹ By contrast, this is possible once the parameter count exceeds said threshold.*



(a) A_* chosen as the block-diagonal matrix consisting of first a Jordan block of size 4 with eigenvalue 1 and second the rescaled identity of size 3 with eigenvalue 0.4.

(b) A_* chosen as the block-diagonal matrix consisting of first a Jordan block of size 4 with eigenvalue 1.1 and second the rescaled identity of size 3 with eigenvalue 0.4.

Figure 1: We illustrate Theorem 2 by a simple numerical example of learning a d_X -dimensional linear system $X_{t+1} = A_* X_t + W_t$ with fewer than the required number of parameters and where $d_X = 7$. Namely, we only estimate the top-left $k \times k$ sub-matrix, $k \in [d_X]$. We run the least squares estimator for samples drawn from $m \gg 2^{d_X}$ many trajectories and vary the trajectory length, T . As predicted, as T grows the risk diverges unless the parametrization is sufficiently high-dimensional, $k = 4$, at which point the risk drops to near zero and exhibits more stable behavior (note the logarithmic scale on the y -axis).

We also prove an extension to Theorem 2 that applies to imperfect state observations but is restricted to learning in parametric classes consisting of finite-dimensional filters.

Theorem [Informal version of Theorem 4] *For an imperfectly observed random walk in \mathbb{R} , there exists no successful learner in the class of linear filters unless the filter length exceeds a certain threshold based on the detectability and horizon of the problem.*

Remark 1 *As a byproduct of our analysis, we note in passing that Theorem 4 shows that the truncation level used in Tsiamis and Pappas (2019) for improper linear system identification cannot*

1. Note that Theorem 2 gives a more nuanced statement in terms of Jordan blocks—the above statement corresponds to the worst case Jordan block structure.

be much improved in general. In particular, improper learning with a finite length filter always (unless further constraints are added to the hypothesis class) incurs an extra approximation-theoretically induced logarithmic factor as opposed to the maximum likelihood estimator.

3. Emergence in Fully Observed Systems

As a first example, let us consider a fully observed state space model. In this case, C_\star in (1.1) is simply the identity and V_t is identically zero:

$$X_{t+1} = A_\star X_t + W_t, \quad t = 1, \dots, T-1, \quad X_1 = W_0 \quad (8)$$

We consider the setting in which a learner observes the trajectory $X_{1:T}$ and seeks to learn the generative model by recovering A_\star . We suppose that each \mathcal{F}_d is given by a map $A_d : \mathcal{M} \mapsto \mathbb{R}^{d_X \times d_X}$ such that $\mathbf{E}_Q^{t-1} Y_t = A(\theta) X_t$ where \mathcal{M} is some smooth manifold of dimension d_M . In this case the prediction risk becomes

$$\sum_{t=1}^T \mathbf{E}_P \|\mathbf{E}_P^{t-1} Y_t - \mathbf{E}_Q^{t-1} Y_t\|^2 = \sum_{t=1}^{T-1} \mathbf{E} \|A_d(\theta) X_t - A_\star X_t\|^2. \quad (9)$$

Assumption 3.1 *The spectral radius of A_\star is at least unity.*

We now show that when the generative model (8) is not ergodic—Assumption 3.1 holds—the risk exhibits a phase transition in how it scales with the trajectory length T as a function of the number of trainable parameters—the dimension of \mathcal{M} , d_M . In both cases below we abuse notation and write $\ell_T(\mathcal{M}, A_\star) = \ell_T(\mathcal{M}, P_\star)$ where P_\star is the distribution of $X_{1:T}$ with the parametrizing matrix A_\star in the generative model (8).

Theorem 2 *Impose Assumption 3.1. Let d_\star^2 be the sum of the squares of the algebraic multiplicities of all eigenvalues of A_\star with magnitude at least unity.*

1. *If $d_M < d_\star^2$, then for every $\varepsilon > 0$ there exists $A_\star^\varepsilon \in \mathbb{R}^{d_X \times d_X}$ with $\|A_\star^\varepsilon - A_\star\| \leq \varepsilon$ such that:*

$$\lim_{T \rightarrow \infty} T^{-1} \ell_T(\mathcal{M}, A_\star^\varepsilon) = \infty. \quad (10)$$

2. *If $d_M < d_\star^2 - d_{\star,1}$, there exists invertible P such that:*

$$\lim_{T \rightarrow \infty} T^{-1} \ell_T(\mathcal{M}, P^{-1} A_\star P) = \infty \quad (11)$$

where $d_{\star,1}$ is the sum of the algebraic multiplicities of all eigenvalues of A_\star with magnitude at least unity.

The first part of Theorem 2 shows that unless the number of parameters is quadratic in the number of unstable modes, there exists no learner with bounded loss that is robust to infinitesimal perturbations of the generative model. It is also interesting to note that learning becomes drastically more difficult if A_\star has a single large Jordan block as opposed to being diagonal in some basis. The second part shows that this remains true even if the spectrum is fixed a priori and the perturbations are restricted to a change of basis. By contrast, if A_\star is strictly stable, it is easy to see that there always

exists $\varepsilon > 0$ such that if $\|A_\star^\varepsilon - A_\star\| \leq \varepsilon$ it holds that $\lim_{T \rightarrow \infty} \ell_M(A_\star^\varepsilon) < \infty$ —and this holding is independent of d_M .

Proof First, we observe that for some invertible matrix P_ε we may write $A_\star^\varepsilon = P_\varepsilon^{-1} J_\star^\varepsilon P_\varepsilon$ for the Jordan normal form of A_\star^ε and where J_\star^ε is block diagonal with the eigenvalues of A_\star^ε on its main diagonal. The model (8) can thus equivalently be written as

$$\underbrace{P_\varepsilon X_{t+1}}_{\triangleq H_{t+1}} = J_\star^\varepsilon \underbrace{P_\varepsilon X_t}_{\triangleq H_t} + \underbrace{P_\varepsilon W_t}_{\triangleq V_t} \quad (12)$$

and (12) can unrolled as

$$H_t = \sum_{k=0}^{t-1} (J_\star^\varepsilon)^k V_{t-k-1} \quad \text{with} \quad \mathbf{E} H_t H_t^\dagger = \sum_{k=0}^{t-1} (J_\star^\varepsilon)^k P_\varepsilon P_\varepsilon^\dagger (J_\star^\varepsilon)^{k,\dagger} \quad (13)$$

where \dagger denotes conjugate transpose.

Second, we observe that we may restrict attention without loss of generality to the situation in which A_\star has a single repeated eigenvalue with multiplicity d_\star by decomposing the system (8) into its distinct A_\star -invariant subspaces. The general lower bound then follows by summing each of the individual subspace lower bounds.

Third, we notice that

$$\begin{aligned} & \min_{A \in \mathcal{M}} \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E} \| (A - A_\star^\varepsilon) X_t \|^2 \\ &= \min_{A \in \mathcal{M}} \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E} \| (A - P_\varepsilon^{-1} J_\star^\varepsilon P_\varepsilon) X_t \|^2 \quad (A_\star^\varepsilon = P_\varepsilon^{-1} J_\star^\varepsilon P_\varepsilon) \\ &= \min_{A \in \mathcal{M}} \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E} \| (P_\varepsilon^{-1} P_\varepsilon A P_\varepsilon^{-1} P_\varepsilon - P_\varepsilon^{-1} J_\star^\varepsilon P_\varepsilon) X_t \|^2 \quad (P_\varepsilon^{-1} P_\varepsilon = I) \\ &= \min_{J \in P_\varepsilon \mathcal{M} P_\varepsilon^{-1}} \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E} \| P_\varepsilon^{-1} (J - J_\star^\varepsilon) P X_t \|^2 \quad (J \triangleq P_\varepsilon A P_\varepsilon^{-1}) \\ &= \min_{J \in P_\varepsilon \mathcal{M} P_\varepsilon^{-1}} \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E} \| P_\varepsilon^{-1} (J - J_\star^\varepsilon) H_t \|^2 \quad (P_\varepsilon X_t = H_t) \\ &\geq \lambda_{\min}^2(P^{-1}) \min_{J \in P_\varepsilon \mathcal{M} P_\varepsilon^{-1}} \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E} \| (J - J_\star^\varepsilon) H_t \|^2 \\ &\geq \lambda_{\min}^2(P^{-1}) \lambda_{\min}^2(P) \min_{J \in P_\varepsilon \mathcal{M} P_\varepsilon^{-1}} \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{E} \operatorname{tr} \left(\sum_{k=0}^{t-1} (J_\star^\varepsilon)^k (J_\star^\varepsilon)^{\dagger,k} (J - J_\star) (J - J_\star)^\dagger \right). \quad (13) \end{aligned}$$

(14)

Now the d_\star -many of the diagonal elements of each $(J_\star^\varepsilon)^k (J_\star^\varepsilon)^{\dagger,k}$ are at least unity. Consequently all the d_\star -many of the diagonal elements of $\sum_{k=0}^{t-1} (J_\star^\varepsilon)^k (J_\star^\varepsilon)^{\dagger,k}$ are larger than or equal to t . Indeed for some $|\lambda| \geq 1$ we have $J_\star^\varepsilon = (\lambda I_{d_\star} + N)$ for some nilpotent matrix N and consequently $\lim_{t \rightarrow \infty} \frac{1}{t} \lambda_{\min} \left(\sum_{k=0}^{t-1} (J_\star^\varepsilon)^k (J_\star^\varepsilon)^{\dagger,k} \right) > 0$. Since both this matrix and $(J - J_\star^\varepsilon)(J - J_\star^\varepsilon)^\dagger$ are positive semi-definite, it follows that $\ell_M < \infty$ if and only if there exists $J \in P_\varepsilon \mathcal{M} P_\varepsilon^{-1}$ with $J = J_\star^\varepsilon$.

To finish the proof notice that:

$$\begin{aligned}
 & \exists J \in P_\varepsilon \mathbf{M} P_\varepsilon^{-1} : & J &= J_\star^\varepsilon, \\
 & \Leftrightarrow \exists A \in \mathbf{M} : & P_\varepsilon A P_\varepsilon^{-1} &= J_\star^\varepsilon, \\
 & \Leftrightarrow \exists A \in \mathbf{M} : & A &= P_\varepsilon^{-1} J_\star^\varepsilon P_\varepsilon = A_\star^\varepsilon.
 \end{aligned} \tag{15}$$

The first part of the result follows since A_\star^ε varies over a d_\star^2 -dimensional manifold and A varies over a d_M -dimensional manifold. Hence, for (15) to have a solution for every A_\star^ε we require that $d_M \geq d_\star^2$.

Now for the second part, let instead P vary over the general linear group. In this case, $P^{-1} J_\star P$ varies over a $(d_\star^2 - d_\star)$ -dimensional manifold whereas $A \in \mathbf{M}$ only varies over a d_M -dimensional manifold. To see that the degrees of freedom of $P^{-1} J_\star P$ are indeed $d_\star^2 - d_\star$, invoke the Orbit-Stabilizer Theorem and notice that the dimension of the orbit of J_\star under conjugation by the general linear group is equal to the dimension of the quotient space of the general linear group modulo the centralizer of J_\star . The general linear group has dimension d_\star^2 and the centralizer of a Jordan block under this action has dimension d_\star . Consequently, this equation cannot have a solution for every admissible choice of right hand side unless $d_M \geq d_\star^2 - d_\star$. \blacksquare

4. Hidden States and the Role of the Parametrization

In Theorem 2 we saw that we require a quadratic amount of parameters in the number of unstable modes. However, this was assuming direct access to the internal system state. If instead the state is hidden, the observations are no longer Markovian and exhibit longer range memory. We will now turn to investigating the appearance of such memory interacts with the potential instability (non-ergodicity) of A_\star . Let us also restrict attention to hypothesis classes consisting of finite-dimensional filters of the form $f_t(Y_{1:t-1}) = \sum_{k=1}^h F_k Y_{t-k}$ for every t (where F_k is the decision-variable that does not depend on t). Finite memory of this type is present in many popular architectures, including transformers, where it is referred to as the context length (Vaswani et al., 2017). We denote these classes \mathbf{M}_h . In this setting, for a fixed integer h and hypothesis $f \in \mathbf{M}_h$, with representation $F_{1:h}$, we have that:

$$\sum_{t=1}^T \mathbf{E}_P \|\mathbf{E}_P^{t-1} Y_t - \mathbf{E}_Q^{t-1} Y_t\|^2 = \sum_{t=1}^{T-1} \mathbf{E} \left\| \sum_{k=1}^h F_k Y_{t-k} - \mathbf{E}[Y_t | Y_{1:t-1}] \right\|^2. \tag{16}$$

At this stage it must be pointed out that it is not just the dimensionality of the parametrization that matters but also the parametrization itself. There certainly exists a hypothesis class using no more than $d_X(d_X + d_Y)$ -many parameters rendering (16) null. On the other hand, the dimension of the internal state may be large or not even known a priori in which case it is appropriate to approximate (2) by a finite-dimensional filter—the question then becomes: *what is the minimal filter length such that (16) remains stable?*

The analysis in the sequel passes via the Kalman filter. The next assumption guarantees that this can be represented by a linear time-invariant system. The part of the assumption dealing with time-invariance does not meaningfully restrict the generality of our results since the filter parameters convergence to their steady-state values at a super-exponential rate.

Assumption 4.1 *The pair (C, A) is observable and $\Sigma_W, \Sigma_V \succ 0$. Moreover, the covariance of the initial state satisfies $\Sigma_{W_1} = \Sigma_{ss}$, where Σ_{ss} solves the filter discrete algebraic Riccati equation.*

Under Assumption 4.1 we have that

$$\mathbf{E}[Y_t|Y_{1:t-1}] = \sum_{k=1}^{t-1} M_k^* Y_{t-k} = M_* \mathbf{Z}^{T-t} Y_{1:T-1} \quad (17)$$

where $M_k^* = C_*(A_* - L_* C_*)^k L_*$ for some matrix L_* known as the *Kalman gain* and accordingly $M_* = C_* [(A_* - L_* C_*) \cdots (A_* - L_* C_*)^{T-1}]$ and \mathbf{Z} is the downshift operator. Similarly

$$\sum_{k=1}^h F_k Y_{t-k} = F \begin{bmatrix} 0_{t-h-1} & I_h & 0_{T-t-1} \end{bmatrix} Y_{1:T-1} = F \begin{bmatrix} 0_{T-h-1} & I_h \end{bmatrix} \mathbf{Z}^{T-t} Y_{1:T-1} = F E_h \mathbf{Z}^{T-t} Y_{1:T-1} \quad (18)$$

where $F = [F_h \cdots F_1]$ and $E_h = [0_{T-h-1} \ I_h]$. This conveniently allows us to lower-bound the prediction risk via the following closed form.

$$\begin{aligned} \min_{F_{1:h}} \sum_{t=1}^{T-1} \mathbf{E} \left\| \sum_{k=1}^h F_k Y_{t-k} - \mathbf{E}[Y_t|Y_{1:t-1}] \right\|^2 &\geq \sum_{t=1}^{T-1} \min_{F_{1:h}} \mathbf{E} \left\| \sum_{k=1}^h F_k Y_{t-k} - \mathbf{E}[Y_t|Y_{1:t-1}] \right\|^2 \\ &= \sum_{t=1}^{T-1} \mathbf{E} \min_F \left\| (F E_h - M_*) \mathbf{Z}^{T-t} Y_{1:T-1} \right\|^2 \\ &\geq \sum_{t=1}^{T-1} \mathbf{E} \min_F \left\| (F E_h - M_*) \mathbf{Z}^{T-t} \mathbf{C} X_{1:T-1} \right\|^2 \\ &= \sum_{t=1}^{T-1} (\text{vec } M_*)_1^\top [\mathbf{R}_{11} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}](t) (\text{vec } M_*)_1 \end{aligned} \quad (19)$$

where \mathbf{R} and \mathbf{C} are as in (20). Henceforth, we fix a single possible generative model (1.1) and drop the dependency on \mathcal{P} in $\ell_T(\mathbf{M}_h) = \ell_T(\mathbf{M}_h, \mathcal{P})$ with \mathcal{P} described by (1.1). We have established the following.

Proposition 3 *Impose Assumption 4.1, and let*

$$\begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} (t) = \mathbf{R}(t) = \mathbf{Z}^{T-t} \mathbf{C} \left(\mathbf{E} \begin{bmatrix} X_{1:T-1} X_{1:T-1}^\top \end{bmatrix} \right) \mathbf{C}^\top \mathbf{Z}^{T-t, \top} \quad \text{with} \quad \mathbf{C} = \text{blkdiag}(C_*). \quad (20)$$

For every class of linear filters \mathbf{M}_h we have that:

$$\ell(\mathbf{M}_h) \geq \sum_{t=1}^{T-1} (\text{vec } M_*)_1^\top [\mathbf{R}_{11} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}](t) (\text{vec } M_*)_1. \quad (21)$$

The question now is whether the quadratic form $(\text{vec } M_*)_1^\top [\mathbf{R}_{11} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}](t) (\text{vec } M_*)_1$ is uniformly bounded in time or not. We shall see that there are simple examples in which it is not unless the history h is allowed to grow sufficiently rapidly. Namely, let us consider noisy observations of the following scalar random walk model:

$$X_{t+1} = X_t + W_t, \quad Y_t = X_t + V_{t+1}. \quad (22)$$

Our next result shows that it is not just the number of unstable modes that matter in determining how many parameters are required, but also the memory length of the process $Y_{1:T}$.

Theorem 4 *Impose Assumption 4.1 and suppose that $A_\star = C_\star = 1$ as in (22). Let $\rho = A_\star - L_\star C_\star = 1 - L_\star$. For every $h = o\left(\frac{\log T}{\log(1-\rho)}\right)$ we have that*

$$\lim_{T \rightarrow \infty} T^{-1} \ell(M_h) = \infty. \quad (23)$$

The result states that we require a context length or history at least of order $\frac{\log T}{|\log(1-\rho)|}$ for a length T task with $\rho \in (0, 1)$. It is interesting to note that when the variance of the V_t grows large, it can be analytically verified that ρ tends to 1. This offers the following interpretation: a poor signal to noise ratio in the filtering task corresponding to the generative model appearing in Assumption 1.1 leads to a large required parameter dimension (context length).

Proof Via (3) and ?? we have that:

$$\begin{aligned} & \min_{F_{1:h}} \frac{1}{T} \sum_{t=1}^{T-1} \mathbf{E} \left\| \sum_{k=1}^h F_k Y_{t-k} - \mathbf{E}[Y_t | Y_{1:t-1}] \right\|^2 \\ & \geq \frac{1}{T} \sum_{t=1}^{T-1} (\text{vec } M_\star)_1^\top [\mathbf{R}_{11} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}](t) (\text{vec } M_\star)_1 \quad (\text{Theorem 3}) \\ & \gtrsim \frac{1}{T} \sum_{t=1}^{T-1} \sum_{l=1}^{t-k} \left(\sum_{j=1}^{t-k-l} \rho^{j-1} \right)^2 - \frac{1}{h+1} \left(\sum_{j=1}^{t-h} j \rho^{t-k-j} \right)^2 \quad (??) \\ & \gtrsim T^{-1} \left(\frac{T^2}{(1-\rho)^2} - O(1) \right) \\ & \asymp \frac{T}{(1-\rho)^{2h}}. \end{aligned} \quad (24)$$

Note that the RHS of (24) diverges for $h = o\left(\frac{\log T}{|\log(1-\rho)|}\right)$. ■

The step denoted ?? is available in the online appendix, which can be found at <https://arxiv.org/abs/2409.13421>.

5. Discussion

We have proposed a mechanistic explanation of emergence in a relatively simple class of autoregressive learning models. Crucially, and somewhat in parallel to empirical observation (Wei et al., 2022), we find that tasks requiring long-range prediction (put differently: multi-step reasoning) are precisely those which *emerge* at a critical model scale. We also note that our findings are not at all in contrast with the recent theoretical model offered by Arora and Goyal (2023). They take scaling laws for loss functions as a given (Kaplan et al., 2020), and illustrate how such scaling laws can naturally lead to the emergence of more complex reasoning. In the present work we argue directly about the loss. Consequently, we offer a complementary perspective to theirs and try rather to understand whether certain tasks intrinsically require a critical scale.

Our work also begs a number of further interesting questions and future directions are abound. We believe that there are many opportunities in exploring LLM related phenomena through the lens of systems modelling. This has also been pointed out by e.g., Soatto et al. (2023) and Alonso et al.

(2024). It would certainly be interesting to study more concrete emergent skills from this lens, such as in-context learning. Garg et al. (2022) show that standard transformer models—such as the GPT-2 family (Radford et al., 2019)—can perform linear regression from iid examples without explicit supervision. How does the situation change when the examples are drawn sequentially and possibly lack ergodicity? Another interesting phenomenon in which one may want to understand the role of ergodicity, and in which sequence modelling may help, are language model "hallucinations". Kalai and Vempala (2024) find that there is no necessary statistical reason for these to occur in an iid generative model—does this change if we adopt a structured sequential perspective?

Our study also has a number of interesting extensions to other model classes. It may for instance be worthwhile to instantiate the Markovianesque model of Ildiz et al. (2024b) and see if similar results can be derived. It may also be interesting to consider other function classes allowing for some degree of nonlinearity. Goel and Bartlett (2024) prove that an attention-style architecture can approximate a stabilizing Kalman filter with sufficient context length—can we find corresponding lower bounds? Arguably, one would also like to incorporate some degree of representation learning into the present analysis. Ildiz et al. (2024a) study how multiple tasks compete for "representation capacity" via the spectral properties of certain tasks. It is natural to ask how phenomena such as lack of ergodicity and instability affect this competition.

Acknowledgements

The authors acknowledge support from a Swedish Research Council international postdoc grant, NSF award SLES-2331880, AFOSR Award FA9550-24-1-0102, NSF CAREER award ECCS-2045834 and NSF award EnCORE-2217062.

References

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*, 2024.
- Carmen Amo Alonso, Jerome Sieber, and Melanie N Zeilinger. State space models as foundation models: A control theoretic overview. *arXiv preprint arXiv:2403.16899*, 2024.
- Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.

- Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.
- Gautam Goel and Peter Bartlett. Can a transformer represent a kalman filter? In *6th Annual Learning for Dynamics & Control Conference*, pages 1502–1512. PMLR, 2024.
- Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.
- Muhammed E Ildiz, Zhe Zhao, and Samet Oymak. Understanding inverse scaling and emergence in multitask representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 4726–4734. PMLR, 2024a.
- Muhammed Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. In *Forty-first International Conference on Machine Learning*, 2024b.
- Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, 2024.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Stefano Soatto, Paulo Tabuada, Pratik Chaudhari, and Tian Yu Liu. Taming ai bots: Controllability of neural states in large language models. *arXiv preprint arXiv:2305.18449*, 2023.
- Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.