# Can LLMs predict the convergence of Stochastic Gradient Descent?

**Oussama Zekri**[1,2], **Abdelhakim Benechehab**[1,3], **Ievgen Redko**[1]

[1]Huawei Noah's Ark Lab, [2]ENS Paris-Saclay, [3]EURECOM

## Contributions

- **In-context** understanding of SGD dynamics **with LLMs**.
- Estimation of the SGD **transition kernel** seen as a **Markov chain**.
- Prediction of the SGD convergence from **new random initializations** in convex and non-convex settings.

## SGD as a Markov chain

Given a training set of $N$ i.i.d. samples $(x_i) \in \mathbb{R}^d$, we solve the following optimization problem,

$$\min_{\theta} F(\theta), \quad F(\theta) = \frac{1}{N}\sum_{i=1}^{N} f(x_i, \theta)$$

with the Stochastic Gradient Descent,

$$\theta^{t+1} = \theta^t - \gamma_t \nabla \tilde{f}_t(\theta^t)$$

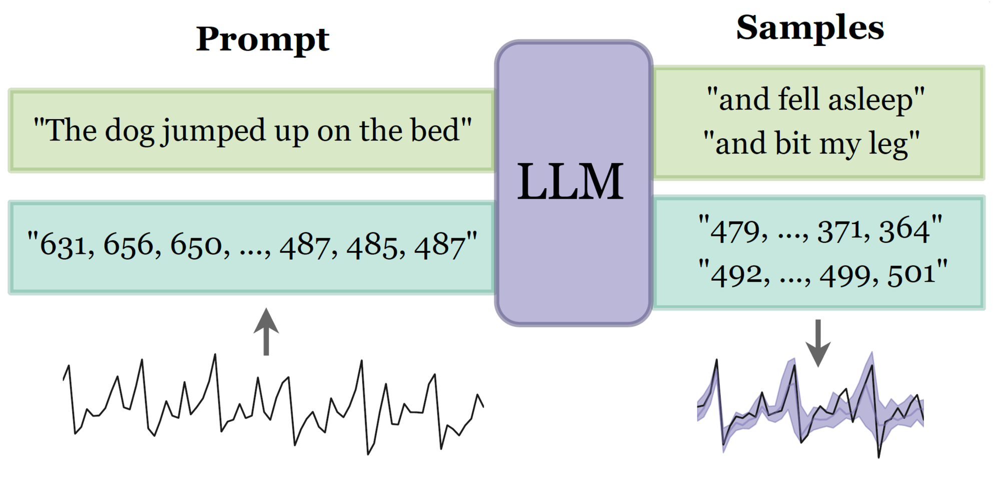
SGD updates form a **multivariate** Markov chain, which is homogeneous for constant stepsize.

**Its transition kernel can be discretized into a block matrix of size $d \times d$.**

$$Q = \begin{pmatrix} \lambda_{1,1}P^{(1,1)} & \dots & \lambda_{1,d}P^{(1,d)} \\ \vdots & \ddots & \vdots \\ \lambda_{d,1}P^{(d,1)} & \dots & \lambda_{d,d}P^{(d,d)} \end{pmatrix}$$

## Time series tokenization

Starting point: LLMTime: LLMs are zero-shot time series forecasters.

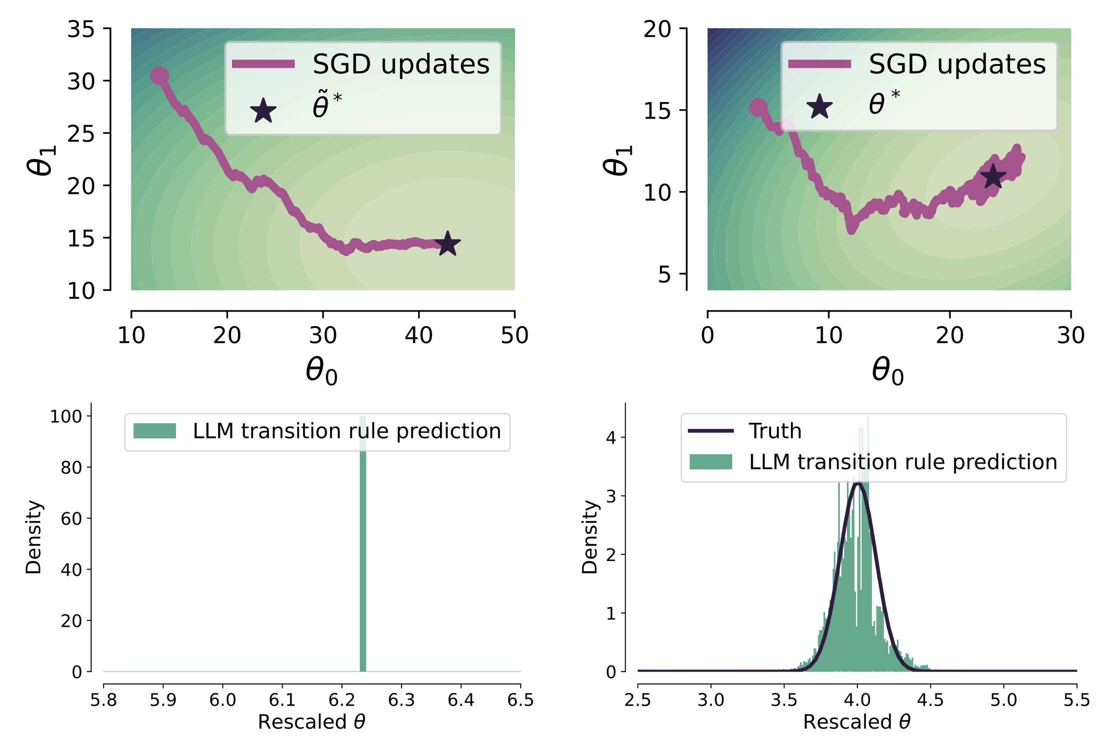



## Main References

- **Gruver et al.** - NeurIPS 2023
  *Large Language Models Are Zero-Shot Time Series Forecasters*
- **Liu et al.** - ICML 2024 ICL Workshop
  *LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law*
- **Dieuleveut et al.** - Annals of Statistics 2020
  *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains*

## Extraction of probabilities

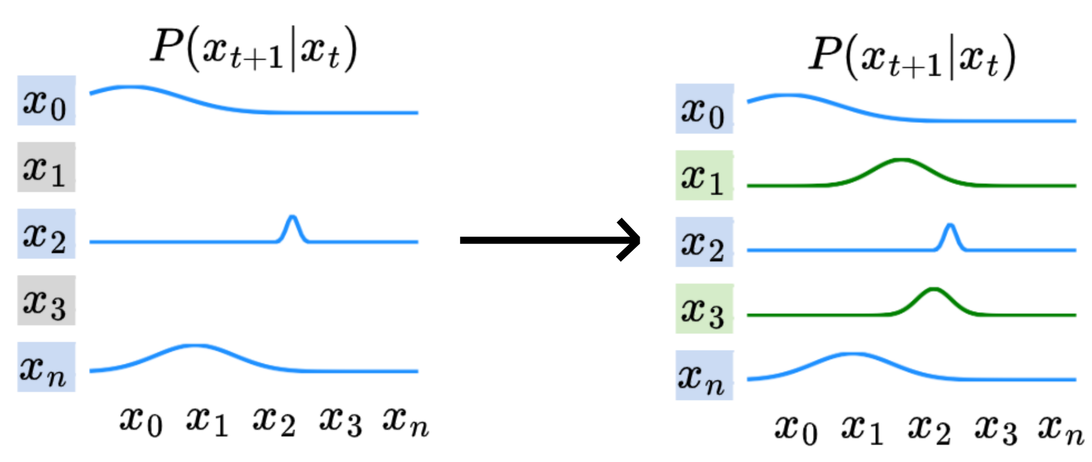
- **Prompt** the LLM with a tokenized time series $(z_t)_{0 \leq t \leq T}$
- **Extract** $\mathbb{P}(Z_{t+1}|Z_t = z_t)$ from the softmax output layer.




**LLMs identify the stationary distribution in both over-parametrized ($d >> N$) and under-parametrized ($d << N$) cases.**

## Imputation of missing values

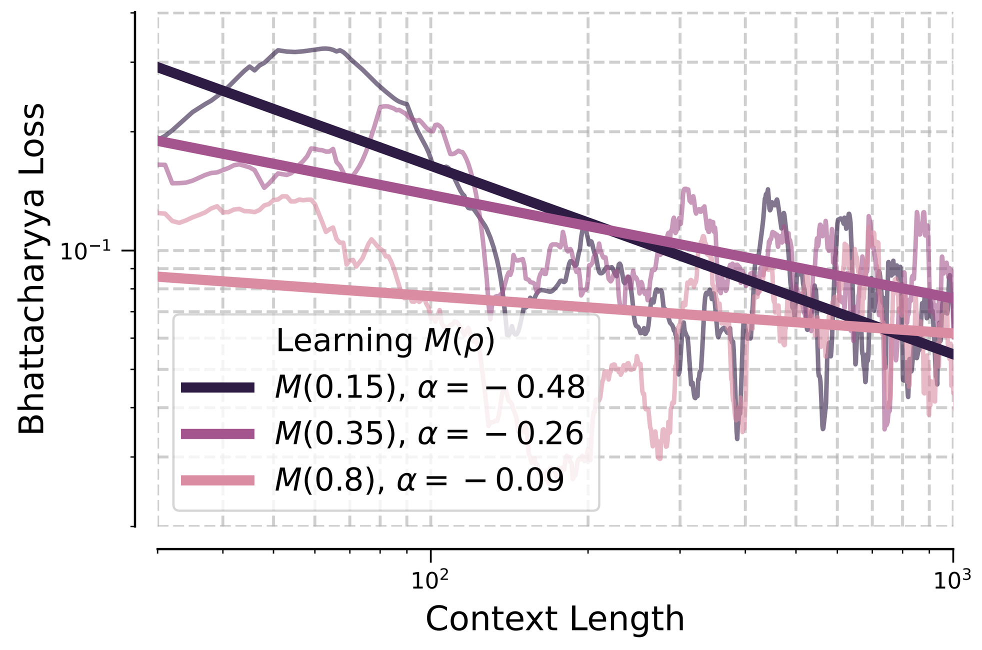
Problem: Few visited states → **sparse** transition matrices.




► Empty rows are filled in by computing the **optimal transport barycenters** between the observed states.

## Neural scaling laws

For spectral gap $\rho$, speed of convergence is given by $d_{TV}(\pi_t, \pi) \leq C_\pi \exp(-\rho t)$.

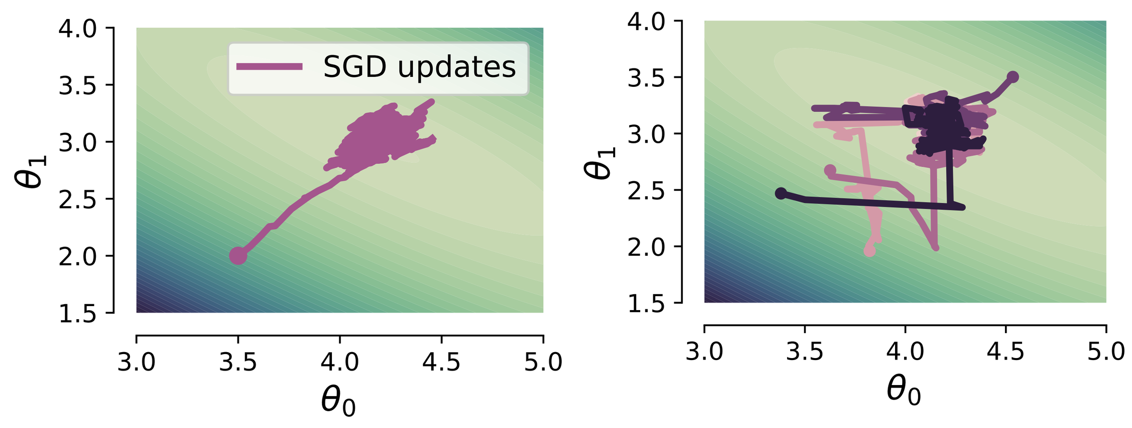



**LLMs are (in-context) Markov chains learners.**

► Neural scaling laws for in-context learning.

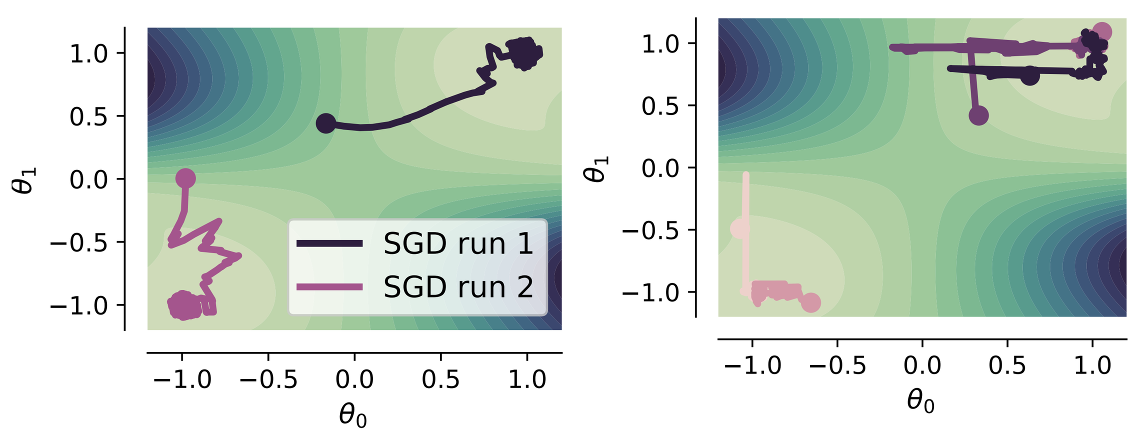
► Influence of the spectral gap on the power law coefficient.

## Predicting the SGD

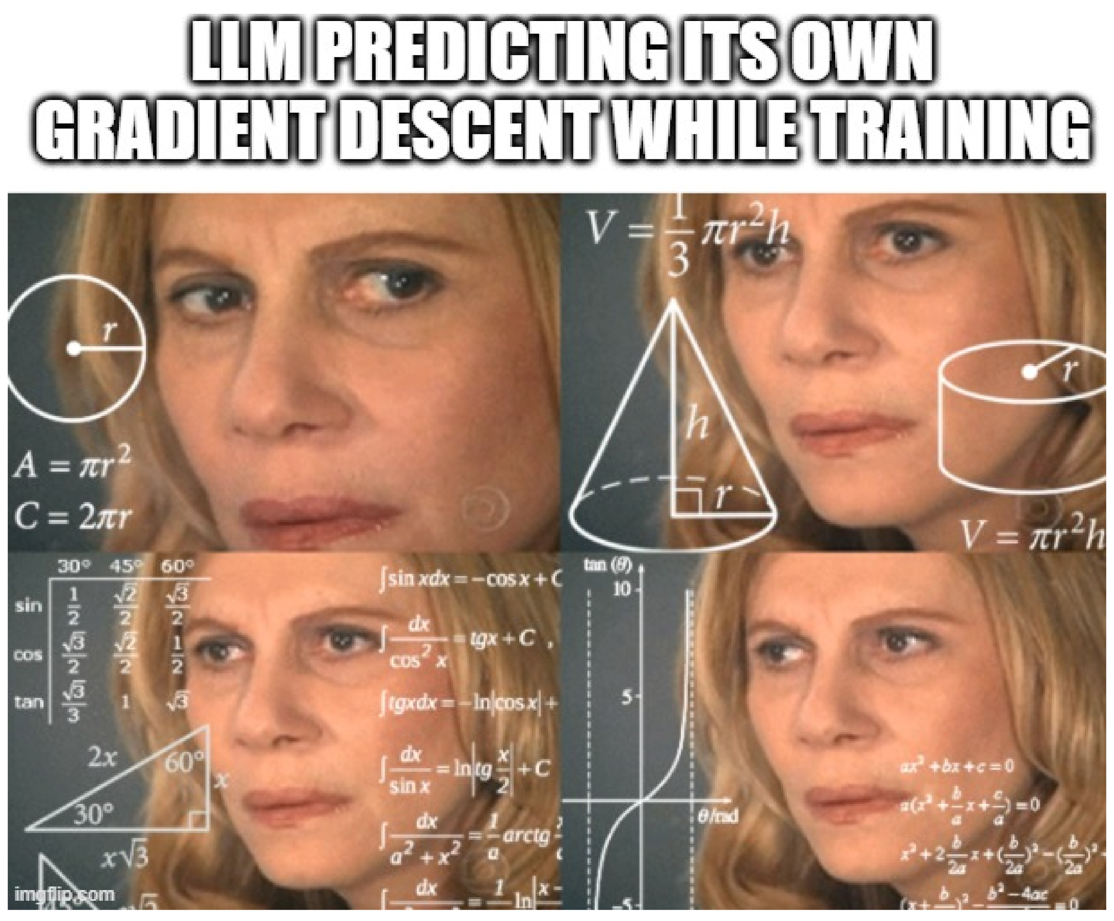



**The transition matrix $Q$ is estimated from an SGD run.**

Cheap matrix products can then be used, rather than accessing gradients.




In the non-convex case, several runs are needed to correctly identify the behavior of the SGD.




## Take Home Message

- LLMs are efficient (in-context) Markov chains learners.
- They can be used to understand SGD from a transition probability point of view.
- Matrix multiplication is cheaper than forward & backward propagation!

## Want to Know More?

paper

author