Attention with Markov: A Framework for Principled Analysis of Transformers via Markov Chains

Ashok Vardhan Makkuva*, Marco Bondaschi*, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, Michael Gastpar

EPFL Information Processing Group

Large Language Models (LLMs)

ChatGPT

GPT-4

Transformers

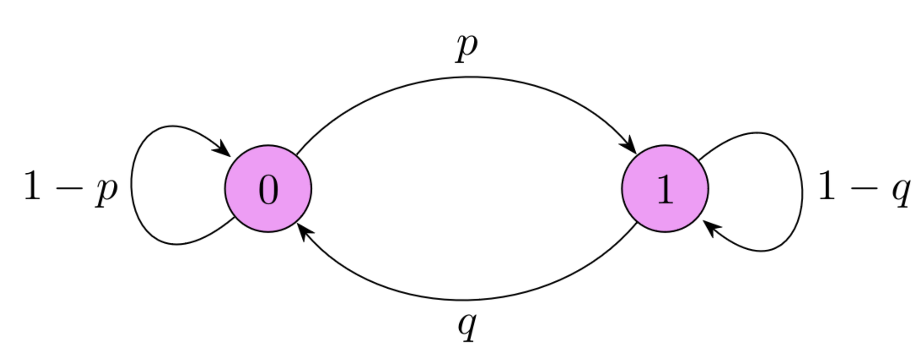
Generative pre-training

Sequential data

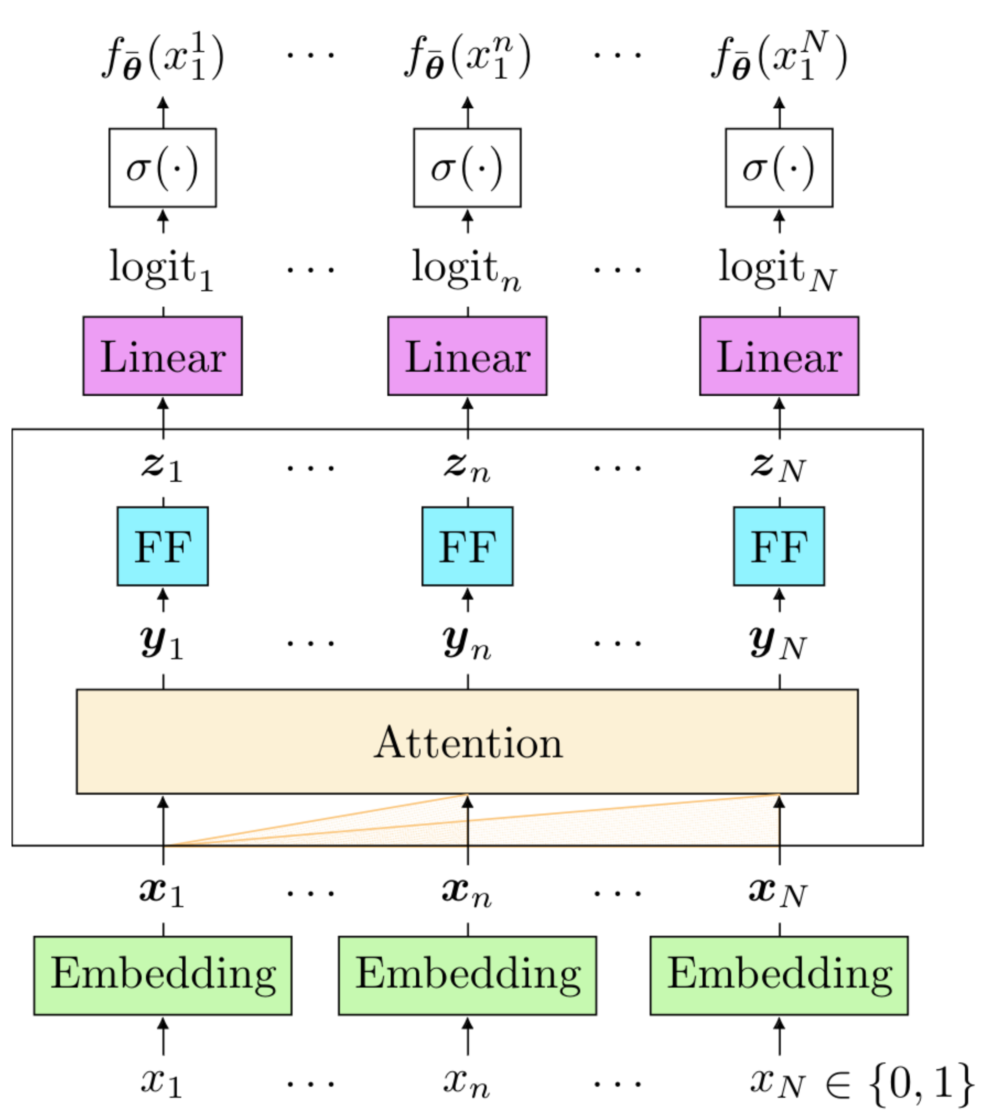
How do transformers learn from sequential data?

Transformers via Markov chains

- Data $(x_{n+1} \mid x_n)_{n \geq 1} \sim$



- Transformer



Main results

Bad local minima (weight tying)

If $p+q>1$ and the weights are tied, there exists an explicit $\boldsymbol{\theta}_{\pi}$ such that:

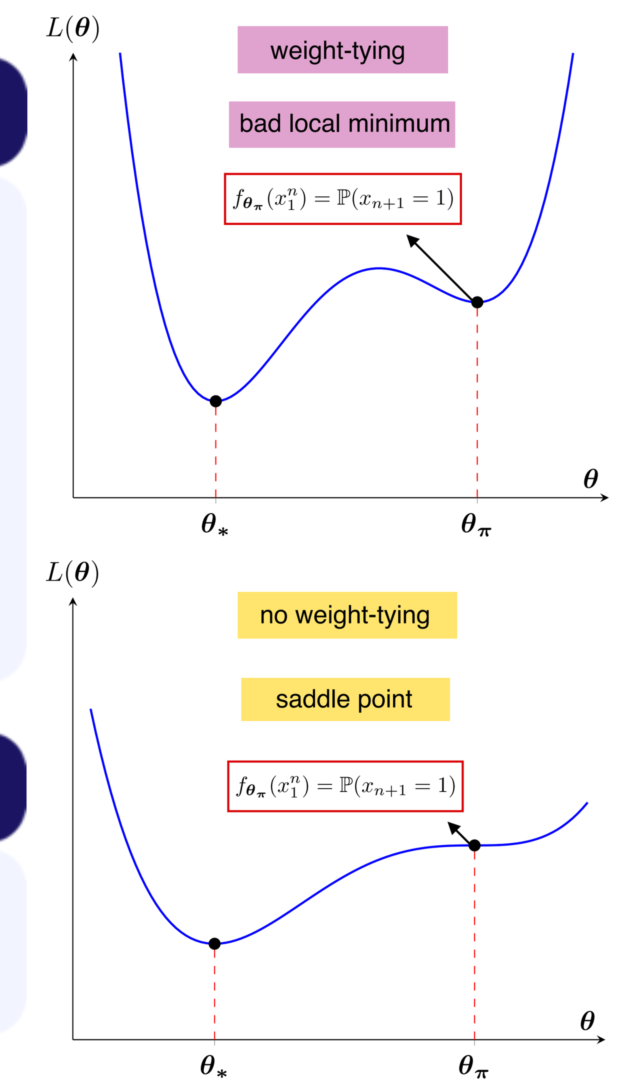
(i) $\boldsymbol{\theta}_{\pi}$ is a bad local minima for $L(\cdot)$ with $L(\boldsymbol{\theta}_{\pi}) > L(\boldsymbol{\theta}_*)$,

(ii) $f_{\boldsymbol{\theta}_{\pi}}(x_1^n) = \mathbb{P}(x_{n+1}=1)$, the marginal distribution,

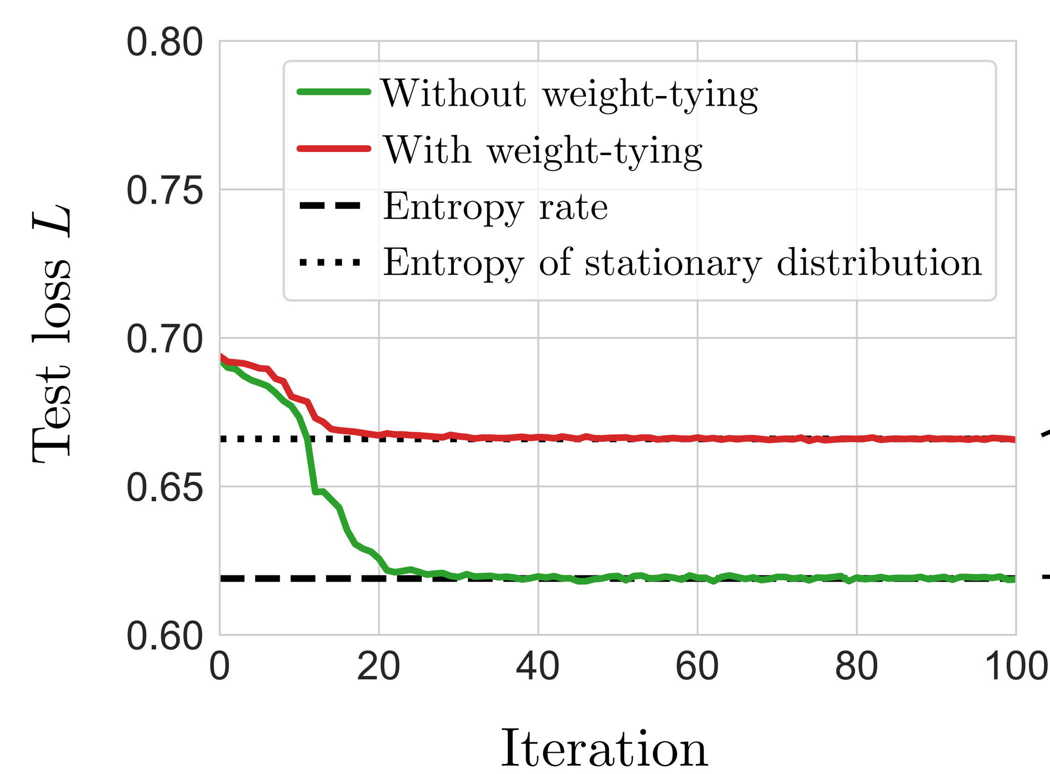
(iii) $\nabla L(\boldsymbol{\theta}_{\pi}) = 0$, i.e. $\boldsymbol{\theta}_{\pi}$ is a stationary point.

Saddle point (no weight tying)

Under the same setting as above with the weights not tied, $\boldsymbol{\theta}_{\pi}$ becomes a saddle point.

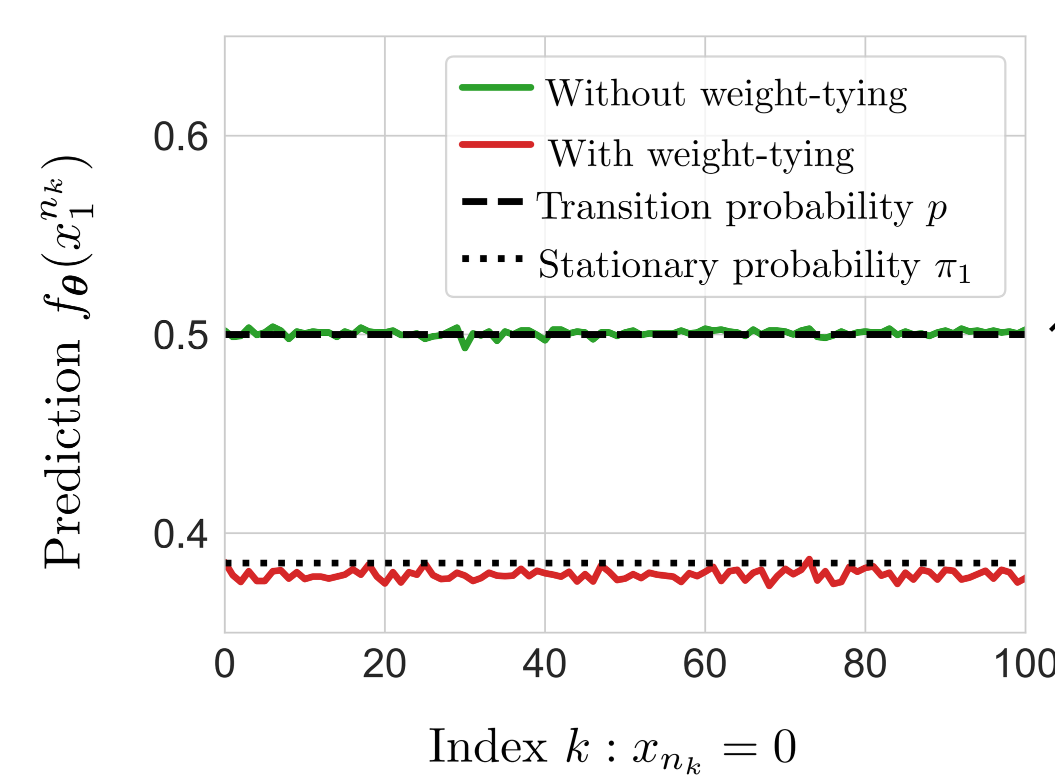


Empirical evidence



bad local minimum

global minimum

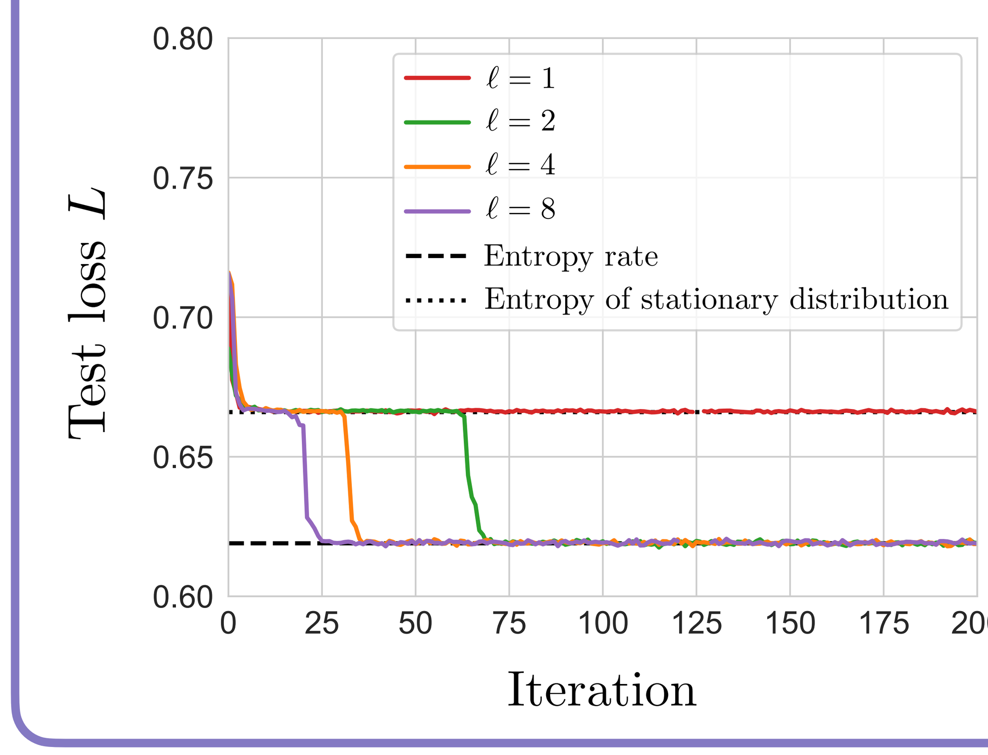


Marginal distribution

Markov kernel

Depth helps!

- First-order Markov Chain



$p+q>1$

Weight-tying

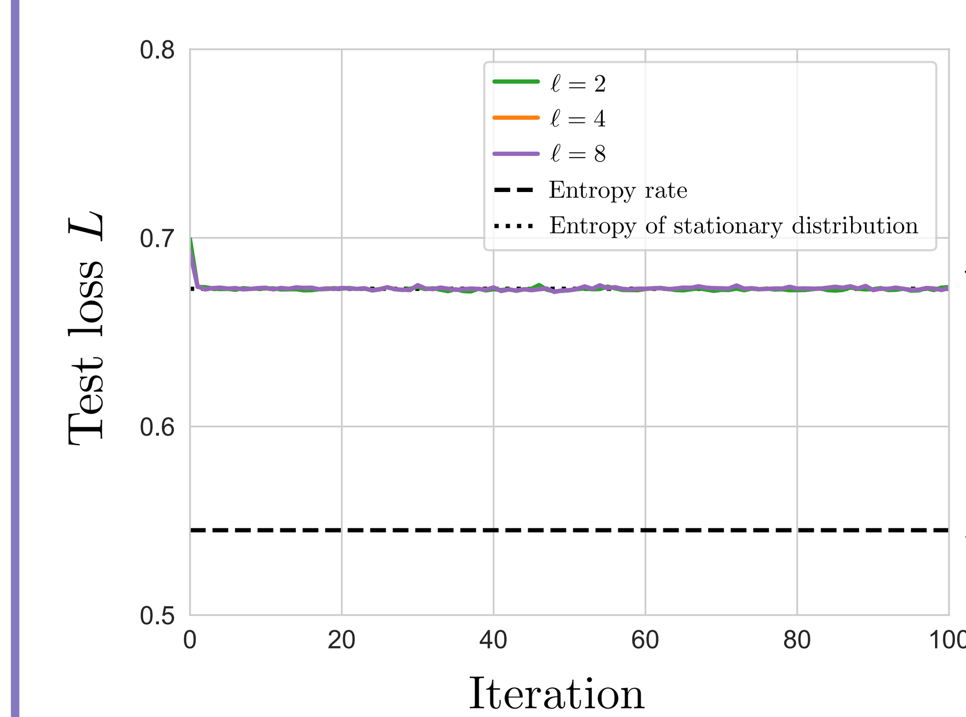
How?

Local minimum

Global minimum

Depth doesn't help!

- Second-order Markov Chain



$p+q>1$

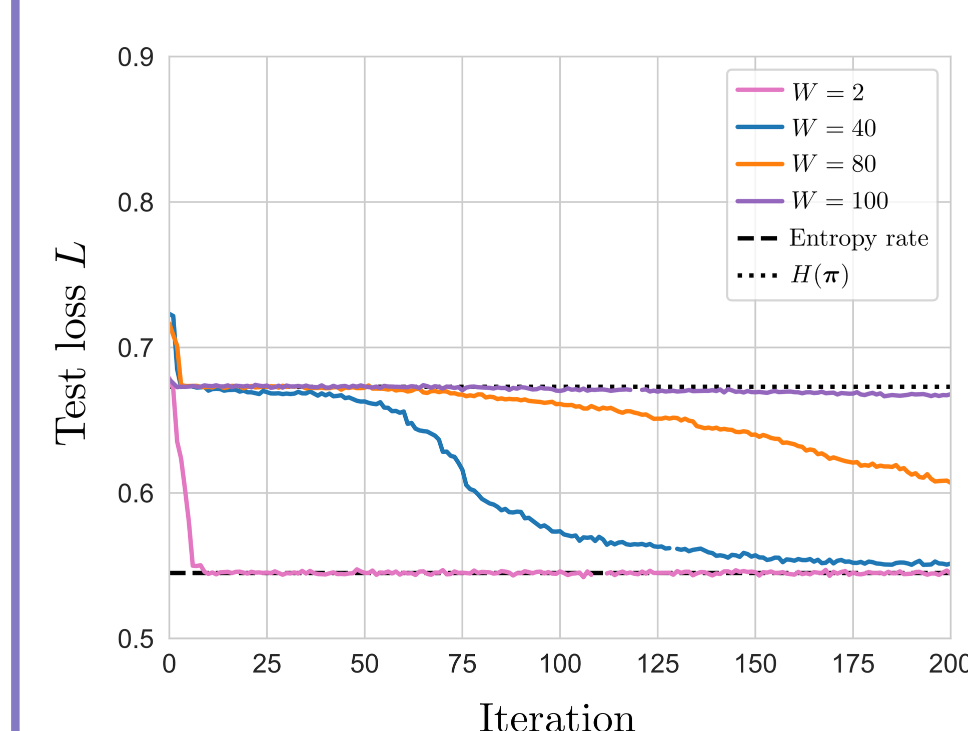
$p+q<1$

Local minimum

Why?

Global minimum

Masking helps



How does masking help?

Depth $\leftrightarrow$ Memory