How do Transformers Perform In-Context Autoregressive Learning?

Michael E. Sander[1], Raja Giryes[2], Taiji Suzuki[3], Mathieu Blondel[4], Gabriel Peyré[1]

(1) Ecole Normale Supérieure and CNRS, France. (2) Tel Aviv University, Israel. (3) University of Tokyo and RIKEN AIP, Japan. (4) Google DeepMind.

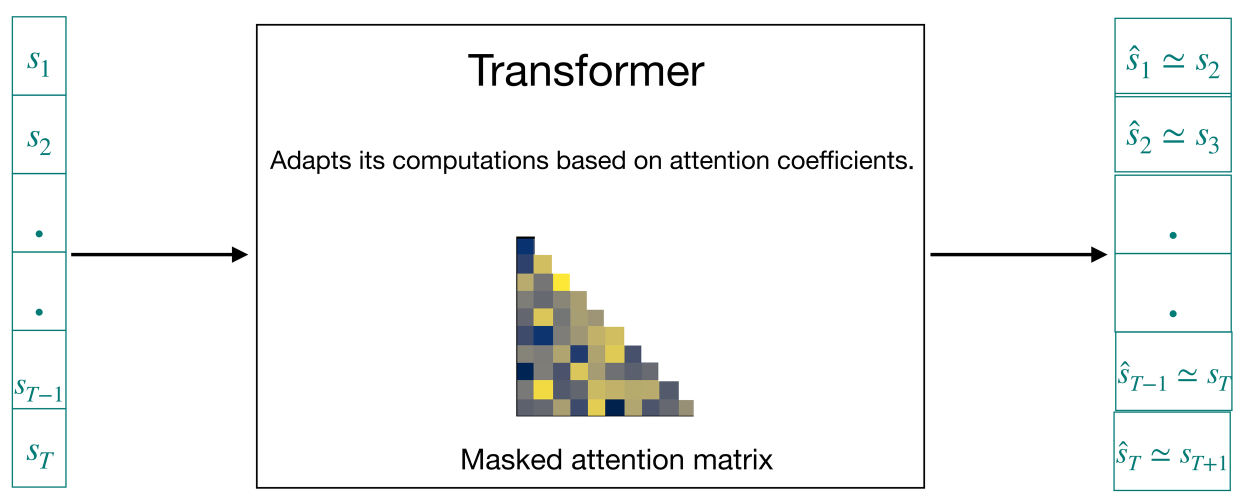
Abstract

We consider the training of Transformers on a simple next token prediction task for the autoregressive process $s_{t+1} = W s_t$. We show how a trained Transformer predicts the next token by first learning W in-context, and then applying a prediction mapping. We call the resulting procedure *in-context autoregressive learning*.

Notations. $\|.\|$ is the ℓ_2 norm. $O(d)$ (resp $U(d)$) is the orthogonal (resp unitary) manifold: $O(d) \coloneqq \{W \in \mathbb{R}^{d\times d} | W^\top W = I_d\}$ and $U(d) \coloneqq \{W \in \mathbb{C}^{d\times d} | W^* W = I_d\}$.

Transformers for next-token prediction

- Given a sequence of tokens $(s_1, \dots, s_T, \dots)$, Transformers are trained to match $s_{1:T} \coloneqq (s_1, \dots, s_T)$ to s_{T+1} for all T.



Goal

We want to show that, assuming the tokens satisfy $s_{T+1} = \phi_W(s_{1:T})$, with W varying with each sequence, the trained Transformer decomposes its prediction into 2 steps: first, estimating W (in-context mapping) and then applying a simple prediction mapping.

We focus on the autoregressive process of order 1: $s_{t+1} = W s_t$.

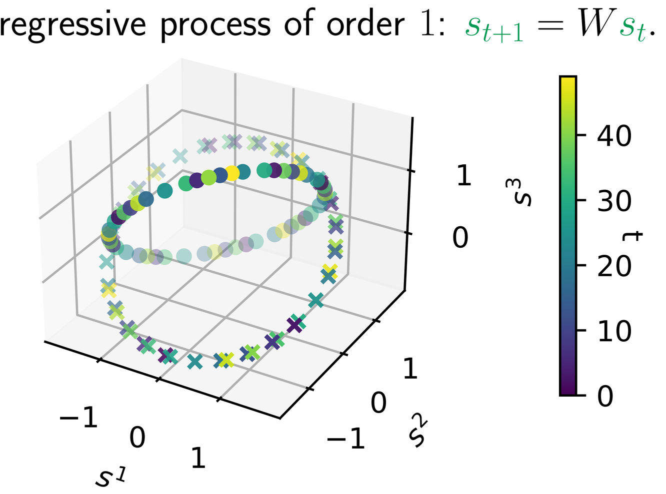


Figure 1: **Two autoregressive processes of order 1 in $\mathbb{R}^3$.**

Token Encoding

Each sequence begins with an initial token $s_1 = 1_d$. The subsequent states are generated according to $s_{t+1} = W s_t$. W is the **context matrix**, sampled uniformly from a subset $\mathcal{C}_O$ (respectively, $\mathcal{C}_U$) of $O(d)$ (respectively, $U(d)$): $W \sim \mathcal{W} \coloneqq \mathcal{U}(\mathcal{C})$. We consider two settings in which the sequence $s_{1:T}$ is first mapped to a new sequence $e_{1:T}$.

- *Augmented Setting*: the tokens are defined as $e_t \coloneqq (0, s_t, s_{t-1})$, aligning with the setup used by Von Oswald et al. (2023).
- *Non-Augmented Setting*: the tokens are simply $e_t \coloneqq s_t$.

Commutativity assumption. The matrices W commute. Hence, they are co-diagonalizable in a unitary basis of $\mathbb{C}^{d\times d}$. Up to a change of basis, we suppose $\mathcal{C}_U = \{\mathrm{diag}(\lambda_1, \cdots, \lambda_d), |\lambda_i| = 1\}$, $\mathcal{C}_O = \{(\lambda_1, \bar{\lambda}_1, \cdots, \lambda_\delta, \bar{\lambda}_\delta), |\lambda_i| = 1\}$, with $d = 2\delta$.

Causal Linear Multi-Head Attention

We consider a model $\mathcal{T}_\theta$ involving Causal Linear Multi-Head Attention:

$$e_{1:T} \mapsto \Big(\sum_{h=1}^{H}\sum_{t'=1}^{t} \mathcal{A}^h_{t,t'} B^h e_{t'}\Big)_{t\in\{1,\cdots,T\}}. \tag{1}$$

$\mathcal{A}^h$ is the attention matrix:

$$\mathcal{A}^h_{t,t'} = P_{t,t'} \langle A^h e_t | e_{t'} \rangle.$$

with $P \in \mathbb{R}^{T_{\max}\times T_{\max}}$ is an optionally trainable positional encoding. The trainable parameters are $\theta = ((A^h, B^h)_{1\le h\le H}, P)$

- We focus on the population loss, defined as:

$$\ell(\theta) \coloneqq \sum_{T=2}^{T_{\max}} \mathbb{E}_{W\sim\mathcal{W}} \|\mathcal{T}_\theta(e_{1:T}) - s_{T+1}\|^2, \tag{2}$$

indicating the model's objective to predict s_{T+1} given $e_{1:T}$.

In-Context Autoregressive Learning

Contributions:

- Theoretically characterize θ^* that minimize ℓ.
- Discuss the convergence of gradient descent to these minima.
- Characterize the in-context autoregressive learning process of the model.

In-Context Autoregressive Learning

We say that $\mathcal{T}_{\theta^*}$ *learns autoregressively in-context* the AR process $s_{t+1} = W s_t$ if $\mathcal{T}_{\theta^*}(e_{1:T})$ can be decomposed in two steps:

- First applying an in-context mapping $\gamma = \Gamma_{\theta^*}(e_{1:T})$
- Then using a prediction mapping $\mathcal{T}_{\theta^*}(e_{1:T}) = \psi_\gamma(e_{1:T})$. This prediction mapping should be of the form $\psi_\gamma(e_{1:T}) = \gamma s_\tau$ for some shift $\tau \in \{1, \cdots, T\}$.

With such a factorization, in-context learning arises when the training loss $\ell(\theta^*)$ is small. This corresponds to having $\Gamma_{\theta^*}(e_{1:T}) \approx W^{T+1-\tau}$ when applied to data $e_{1:T}$ exactly generated by the AR process with matrix W.

In-Context Mapping with Gradient Descent

- **Augmented** tokens $e_t \coloneqq (0, s_t, s_{t-1})$ and $\mathcal{W} = \mathcal{U}(\mathcal{C}_U)$.
- **Model** $\mathcal{T}_\theta(e_{1:T}) = \Big(e_T + \sum_{t=1}^{T} \langle A e_T | e_t \rangle_{\mathbb{C}} B e_t\Big)_{1:d}$.
- **Parametrization**: we take A and B as

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & a_1 I & a_2 I \\ 0 & a_3 I & a_4 I \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & b_1 I & b_2 I \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Proposition (In-context autoregressive learning with gradient-descent)

Loss (2) is minimal for θ^* such that $a_1^* + a_4^* = a_2^* = b_2^* = 0$ and $a_3^* b_1^* = \frac{\sum_{T=2}^{T_{\max}} T}{\sum_{T=2}^{T_{\max}} (T^2 + (d-1)T)}$. Furthermore, an optimal in-context mapping Γ_{θ^*} is one step of gradient descent on the loss $L(W, e_{1:T}) = \frac{1}{2}\sum_{t=1}^{T-1} \|s_{t+1} - W s_t\|^2$ starting from the initialization $W = 0$, with a step size asymptotically equivalent to $\frac{3}{2T_{\max}}$ with respect to $T_{\max}$.

In-Context Mapping as a Geometric Relation

- **Non-augmented** tokens $e_t \coloneqq s_t$.
- **Model** $\mathcal{T}_\theta(e_{1:T}) = \sum_{h=1}^{H}\sum_{t=1}^{T} P_{T-1,t} \langle e_t | A^h e_{T-1} \rangle_{\mathbb{C}} B^h e_t$.
- **Parametrization:** we take $A^h = \mathrm{diag}(a_h)$ and $B^h = \mathrm{diag}(b_h)$.

Then there exists $\mathtt{A}$ and $\mathtt{B} \in \mathbb{R}^{H\times d}$ such that one has for $e_{1:T} = (1_d, \lambda, \cdots, \lambda^{T-1})$:

$$\mathcal{T}_\theta(e_{1:T}) = \sum_{t=1}^{T} P_{T-1,t} [\mathtt{B}^\top \mathtt{A}] \lambda^{t-T+1} \odot \lambda^{t-1}.$$

Proposition (Unitary optimal in-context mapping)

- Any $\theta^* = (\mathtt{A}^*, \mathtt{B}^*, P^*)$ achieving zero of the loss (2) satisfies $P^*_{T-1,t} = 0$ if $t \neq T$, $P^*_{T-1,T}(\mathtt{B}^{*\top}\mathtt{A}^*)_{ii} = 1$, and $(\mathtt{B}^{*\top}\mathtt{A}^*)_{ij} = 0$ for $i \neq j$. Therefore, one must have $H \geq d$. An optimal in-context mapping satisfies $\Gamma_{\theta^*}(e_{1:T}) = \bar{e}_{T-1} \odot e_T$ and the predictive mapping $\psi_\gamma(e_{1:T}) = \gamma \odot e_T$.
- Loss (2) reads $\ell(\mathtt{A}, \mathtt{B}, P) = \sum_{T=2}^{T_{\max}} l(\mathtt{B}^\top\mathtt{A}, P_{T-1})$ with $l(\mathtt{C}, p) = \|p\|_2^2 \|\mathtt{C}\|_F^2 + {p_{T-1}}^2 S(\mathtt{C}^\top\mathtt{C}) - 2\mathrm{Tr}(\mathtt{C}) p_T + d$, where S is the sum of all coefficients operator.

- The equality $(\mathtt{B}^\top\mathtt{A})_{ij} = 0$ for $i \neq j$ corresponds to an orthogonality property between heads. When there are more than d heads, some can be pruned.
- Even for $T_{\max} = 2$ convergence of gradient descent in $(\mathtt{A}, \mathtt{B}, P)$ on ℓ to a global minimum is an open problem (matrix factorization).

Proposition (Orthogonal optimal in-context mapping)

Any $\theta^* = (\mathtt{A}^*, \mathtt{B}^*, P^*)$ with $\ell(\theta^*) = 0$ in (2) satisfies, denoting $\mathtt{C}^* = \mathtt{B}^{*\top}\mathtt{A}^*$ and $p^* = P^*_{T-1}$: $p_t^* = 0$ if $t < T-1$, $p_T^* \mathtt{C}^*_{i,i} = 1$, $p_T^* \mathtt{C}^*_{2i-1,2i} + (\mathtt{C}^*_{2i-1,2i-1} + \mathtt{C}^*_{2i-1,2i}) p^*_{T-1} = 0$, $p_T^* \mathtt{C}^*_{2i,2i-1} + (\mathtt{C}^*_{2i,2i} + \mathtt{C}^*_{2i,2i-1}) p^*_{T-1} = 0$, $\mathtt{C}^*_{2i-1,j} = \mathtt{C}^*_{2i,j} = 0$ for $j \neq 2i-1, 2i$. An optimal in-context mapping is then, for $e_t = \lambda^{t-1}$: $\Gamma_{\theta^*}(e_{1:T}) = \lambda^2$, and the corresponding predictive mapping $\psi_{\Gamma_{\theta^*}(e_{1:T})}(e_{1:T}) = \lambda^2 \odot e_{T-1} = \lambda^T$.

Interpretation: The relation implemented by Γ_{θ^*} is an extension of a known formula in trigonometry: $2\cos\rho R_\rho - I_2 = R_{2\rho}$, with R_ρ the rotation of parameter ρ in $\mathbb{R}^2$.

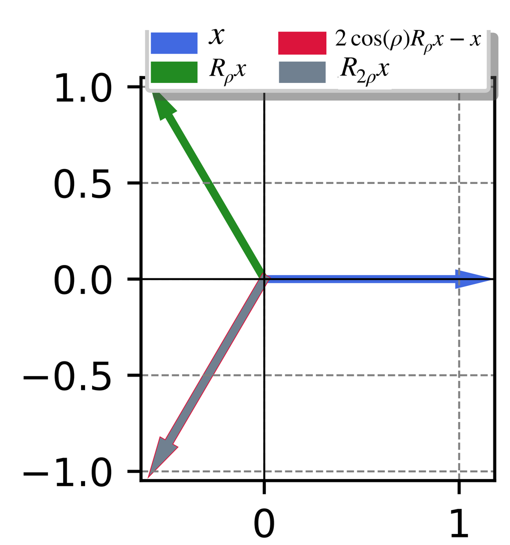


Figure 2: **Trigonometric formula** implemented by the Transformer in-context. The minima of the training loss correspond to implementing, up to multiplying factors: $2\cos\rho R_\rho - I_2 = R_{2\rho}$.

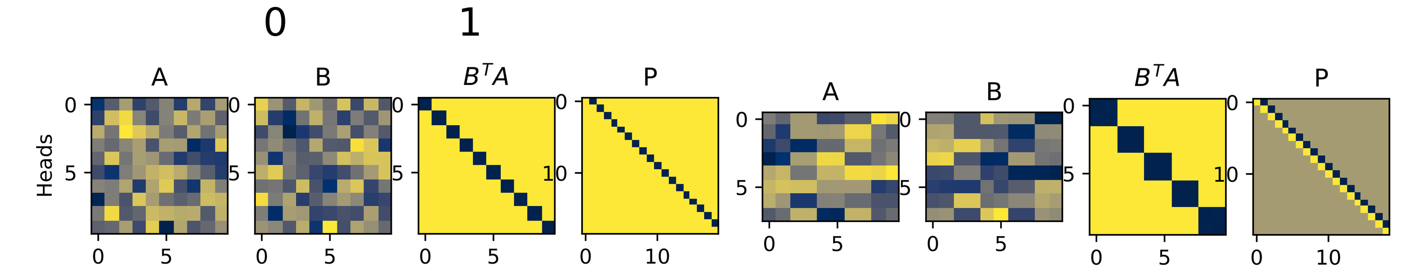


Figure 3: **Matrices $\mathtt{A}$, $\mathtt{B}$, $\mathtt{B}^\top\mathtt{A}$ and P after training on loss (2) with random initialization. Left:** Unitary context case with $H = 10$. **Right:** Orthogonal context case with $H = 8 < d$, which leads to low rank $\mathtt{B}^\top\mathtt{A}$.

Change in the Context Distribution

- **Goal:** Impact of the context distribution on the optimizaztion landscape. We break the symmetry of the context distribution.
- **Non-augmented** tokens and $d = 1$: $s_{t+1} = \lambda s_t$ for $|\lambda| = 1$. For $\mu \geq 1$ and $\rho \sim \mathcal{U}(0, 2\pi)$, we define $\lambda = e^{i\rho/\mu}$.
- **Parametrization:** positional encoding-only attention, we take $\mathtt{B}^\top\mathtt{A} = 1$.

Proposition (Conditioning)

The Hessian $H \in \mathbb{R}^{T\times T}$ of $l(p) \coloneqq \mathbb{E}_{\lambda\sim\mathcal{W}(\mu)} |\sum_{t=1}^{T} p_t \lambda^{2t-T} - \lambda^T|^2$ is

$$H_{t,t'} = \frac{\mu}{4\pi(t'-t)} \sin(4(t'-t)\frac{\pi}{\mu}).$$

With eigenvalues $\sigma_1(\mu) \geq \cdots \geq \sigma_T(\mu)$, $\sigma_1(\mu) \to T$ and $\sigma_{t>1}(\mu) \to 0$ as $\mu \to +\infty$.

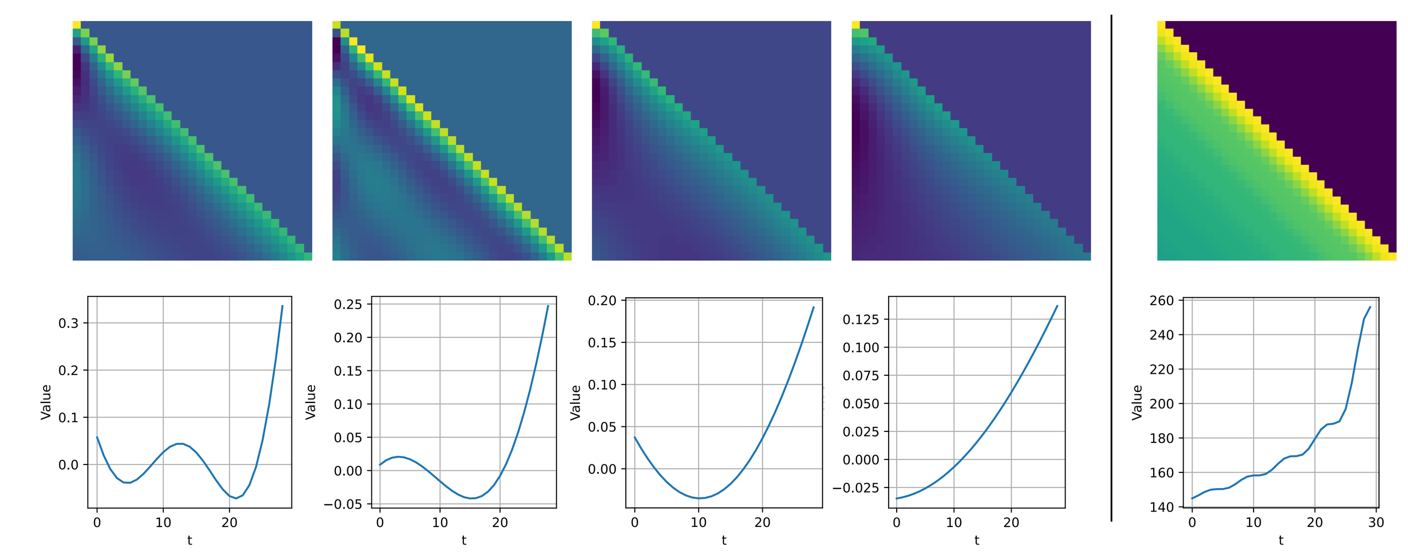


Figure 4: **Left:** Positional encodings after training for $\mu \in \{50, 100, 200, 300\}$. First raw: P. Second raw: plot of its last raw. **Right:** Comparison with cosine absolute PE used in machine translation.

Experiments

Validation of the token encoding choice.

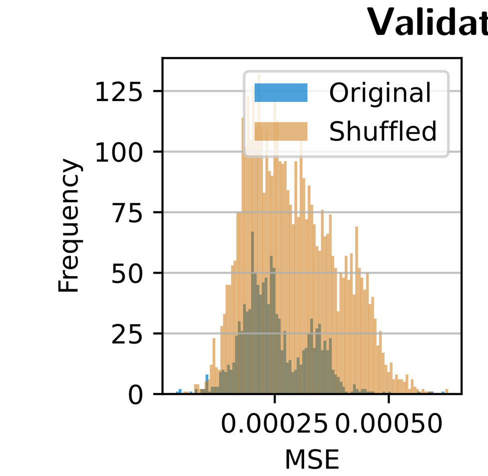


Figure 5: **Setup:** Create a dataset D with 'Moby Dick' from `nltk` package using tokenizer and word embedding of pre-trained GPT-2 model. Also form D_{shuffle} by permuting the tokens. **Plot:** histograms of the mean squared errors (MSE) when fitting an AR process to sequences in D (original, in blue) or D_{shuffle} (shuffled, in orange). We only display MSEs bigger than a threshold of 10^{-12}.

Augmented setting: In-Context Mapping with Gradient Descent

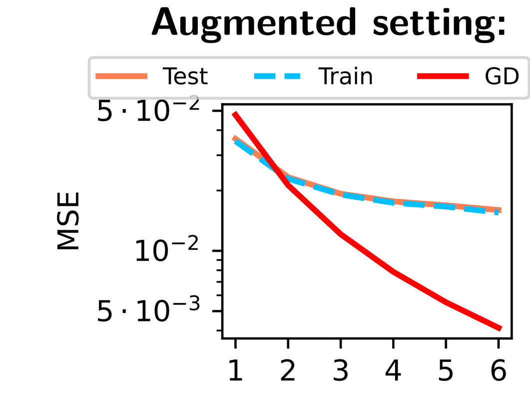


Figure 6: **Setup:** We investigate whether the results of In-Context Mapping with Gradient Descent still hold without assumptions the commutativity and parametrization assumptions. **Plot:** evolution of the MSE with depth L. We compare with L steps of gradient descent on the inner loss. At initialization, the MSE is between 1 and 2.

Non-Augmented setting: In-Context Mapping as a Geometric Relation

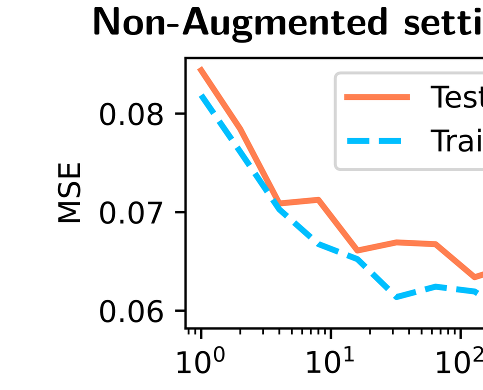


Figure 7: **Setup:** We investigate whether the results of In-Context Mapping as a Geometric Relation still hold without assumptions the commutativity and parametrization assumptions. **Plot:** evolution of the MSE with the number of heads. At initialization, the MSE is between 0.35 and 1.