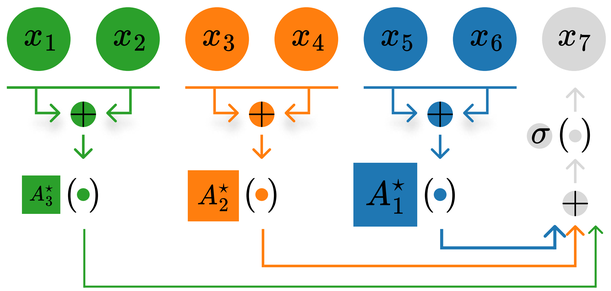
Incremental Learning of Sparse Attention Patterns in Transformers

Oğuz Kaan Yüksel · Rodrigo Alvarez Lucendo · Nicolas Flammarion

ICML International Conference On Machine Learning

Task. High-order Markov chain

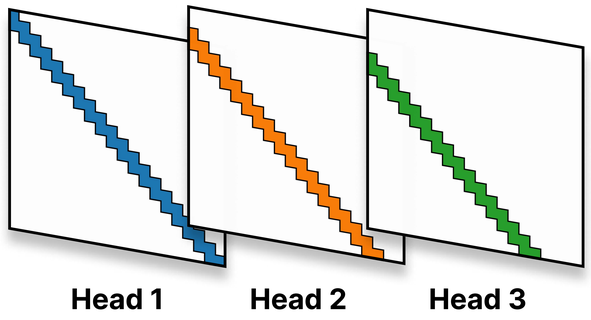


1. Past positions split into groups $I(1)$, $I(2)$, $I(3)$ with feature matrices $A_1^\star$, $A_2^\star$, $A_3^\star$
2. Generate tokens: $x_t \sim \mathrm{softmax}\left(\sum_{k=1}^{h} A_k^\star \sum_{i\in I(k)} \alpha_i x_{t-i}\right)$
3. Feature matrices have **decreasing importance**: $\|A_1^\star\| > \|A_2^\star\| > \|A_3^\star\|$

The model needs **one sparse attention pattern per group**:
do heads specialize from the start, or coordinate somehow?

$A_k^\star$ = group-k feature matrix · $I(k)$ = positions in group k · α_i = within-group weights

Mechanism. Position-only attention heads



A simplified single-layer transformer with **position-only attention**:

$$y_t = \mathrm{softmax}\left(\sum_{k=1}^{h} V_k X a_t^{(k)}\right), \quad a_t^{(k)} = \mathrm{softmax}\left(\mathrm{mask}\left(X^\top K_k^\top Q_k x_t\right)\right)$$

Each head becomes a **sparse pattern** on positions $I(k)$ as $\lambda \to \infty$:

$$K_k^\top Q_k := \lambda \sum_{i\in I(k)} \sum_j p_{j-i} p_j^\top$$

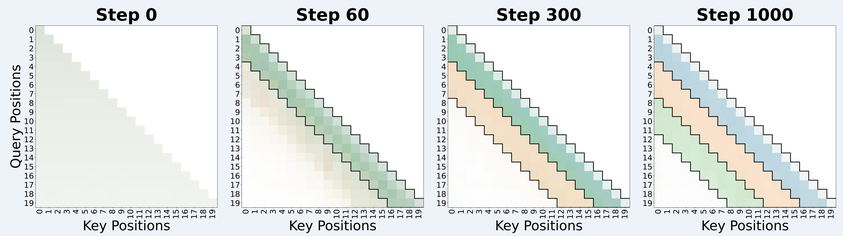
Special case: $|I(k)| = 1 \Rightarrow$ **copying circuits**, which compose into **induction heads**.

p_i = positional embedding of position i

Attention heads specialize:

first compete, then cooperate

learning features one by one



*Each head has its own color (**blue**, **orange**, **green**); overlapping heads mix colors.*

1. **Heads race to the most important pattern**
   coupling via near-symmetric init
2. **One head offshoots to the next pattern**
   saddle-to-saddle dynamics
3. **The rest cooperate with the offshoot**
   features offset transient attention shifts

Let's talk! · Paper · Code · Poster

Generalization · Learning dynamics · Interpretability

Collaborations welcome · Exploring postdoc & research scientist roles

Theory. Saddle-to-saddle dynamics

ATTENTION — **Position-only**

$$\hat{x}_{T+1} = \sum_{k=1}^{h} V_k X s_k, \quad s_k = \mathrm{softmax}(q_k)$$

TASK — **Gaussian x_t · Orthogonal $V_k^\star$**

$$x_{T+1} = \sum_{k=1}^{h} m_k^\star V_k^\star \sum_{i\in I(k)} \alpha_i x_{T+1-i} + \varepsilon$$

LOSS — **Square on the last-token**

$$\min_{V_k, q_k} \tfrac{1}{2}\mathbb{E}\left\|x_{T+1} - \hat{x}_{T+1}\right\|^2$$

DYNAMICS — **Gradient flow**

$$\dot{V}_k = -\partial_{V_k}\mathcal{L}, \quad \dot{s}_k = -\pi(s_k)\partial_{q_k}\mathcal{L}$$

Reduces to a **tensor factorization** with adaptive preconditioner:

$$\min_{V_k, s_k} \tfrac{1}{2}\|G - P\|_F^2, \quad G = \sum_{k=1}^{h} m_k^\star (V_k^\star \otimes s_k^\star), \quad P = \sum_{k=1}^{h} V_k \otimes s_k$$

1 $(V_k, s_k) \approx (V, s)$ for all k (symmetric init)

$$(V, s) \to (m_1^\star V_1^\star / h,\ s_1^\star)$$

Alignment is preserved. $\langle s, s_i^\star - s_j^\star\rangle > 0,\ \langle V, V_i^\star - V_j^\star\rangle > 0,\ m_i^\star > m_j^\star$ *is* ***forward-invariant***.

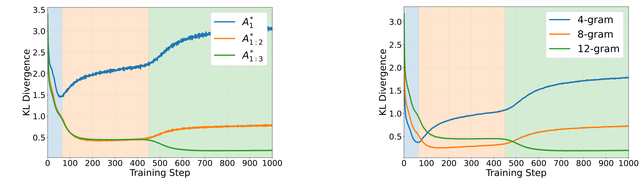
2 $(V_k, s_k) = (m_1^\star V_1^\star / h, s_1^\star)$ for $k \neq 2$; offshooting $(V_2, s_2) \approx (m_1^\star V_1^\star / h, s_1^\star)$

$$(V_2, s_2) \to (m_2^\star V_2^\star, s_2^\star), \quad (V_k, s_k) \to (m_1^\star V_1^\star/(h-1),\ s_1^\star) \text{ for } k \neq 2$$

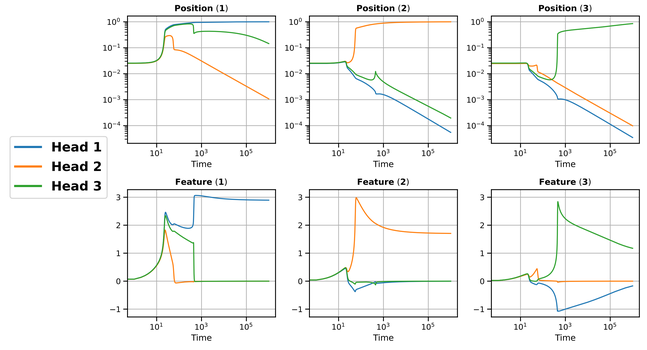
Fast-slow dynamics. *The ensemble moves faster:* $V_k = V = \mathrm{argmin}_V\, \mathcal{L}(V; V_2, s_2)$.

3 Every phase reduces to 1 + 2 on the residual $G_{(k)} := G - \sum_{j=1}^{k} m_j^\star (V_j^\star \otimes s_j^\star)$.

Experiments. Compete → cooperate transition



KL-divergence staircase — full-model training against a teacher (left) and a learned target (right).



Simulated regression ODEs — per-head attention on positions $I(1)$, $I(2)$, $I(3)$ (top) and per-head features $A_1^\star$, $A_2^\star$, $A_3^\star$ (bottom), over time. Full transformers exhibit the same **stage-wise dynamics**.