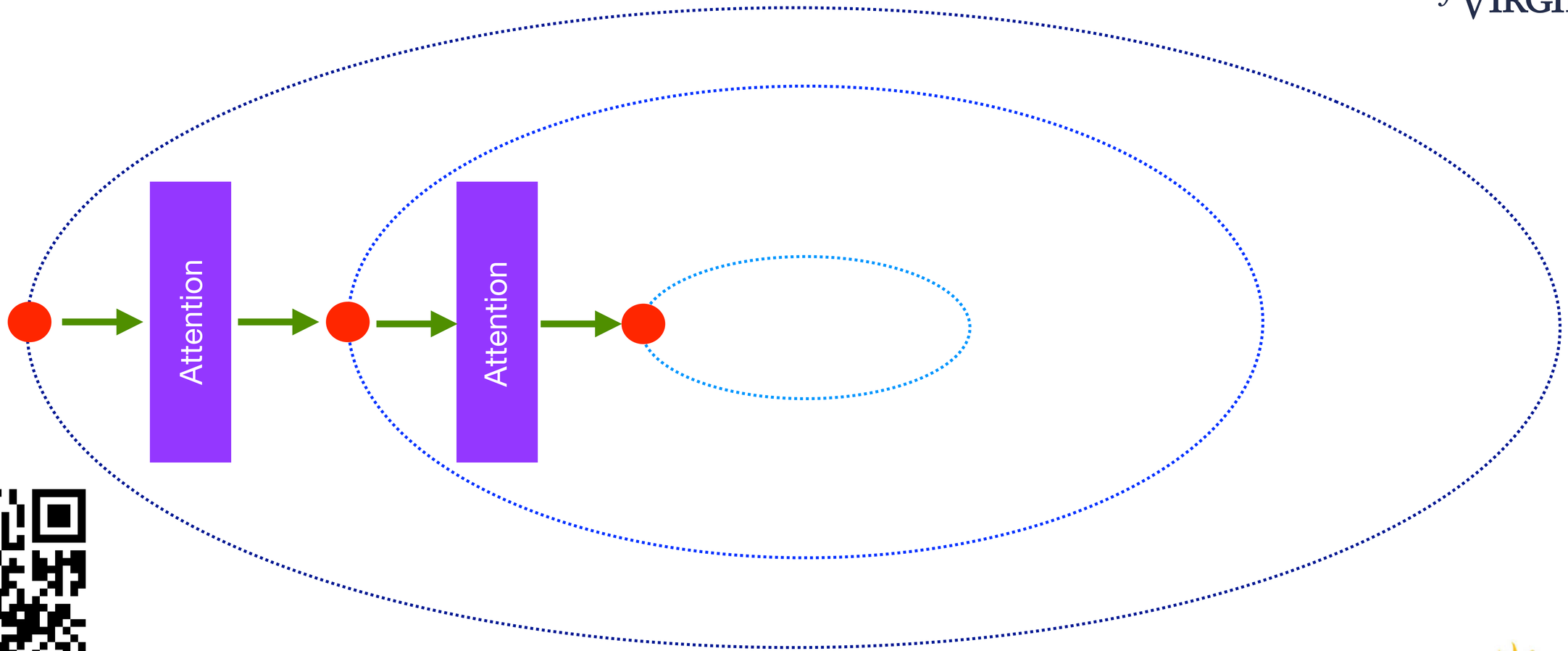
In-Context Learning for Discrete OptimalTransport: Can Transformers Sort?

Hadi Daneshmand
University of Virginia



How do transformers translate?

OT Problem

- Optimal transport:

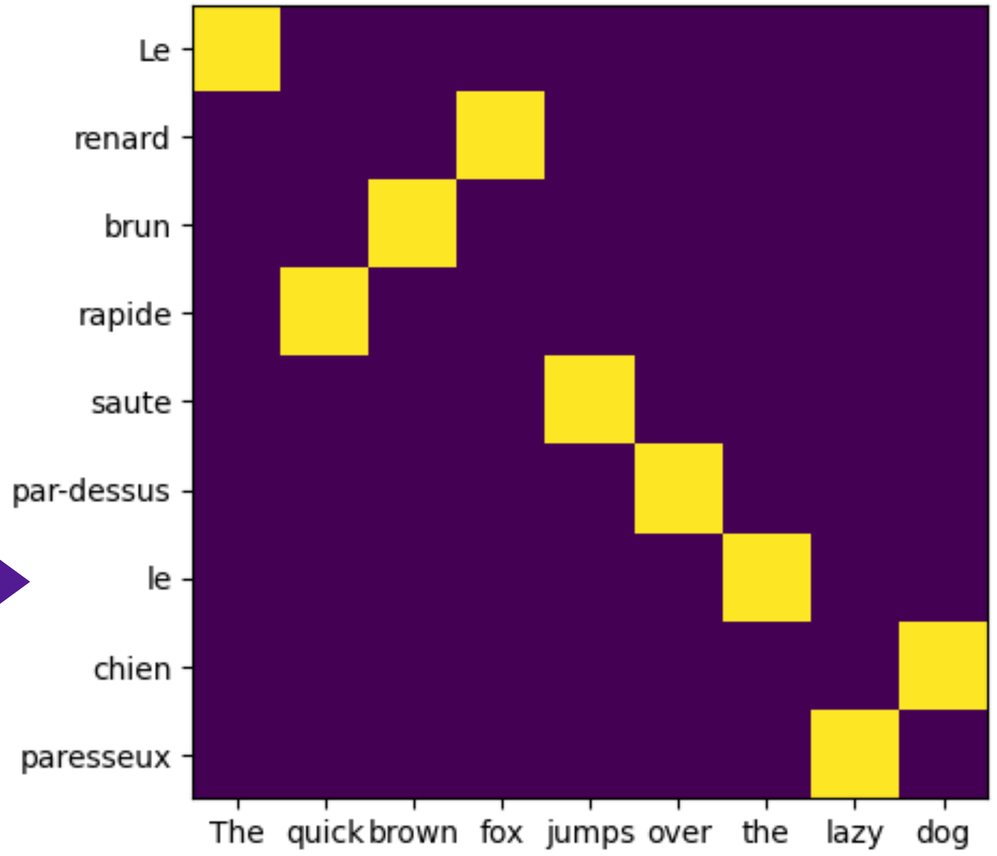
$$P^* := \arg\min_P \sum_{ij} \|x_i - y_j\|^2 P_{ij} \text{ subject to } P \text{ is a permutation matrix}$$

- Permutation matrix: $P \in \mathbb{R}^{n \times n}$, $P_{ij} \in \{0,1\}$, $\sum_{j=1}^{n} P_{ij} = 1$
- Applications: sorting, cell perturbation analysis, reviewer/paper matching, …

OT and Language Translation

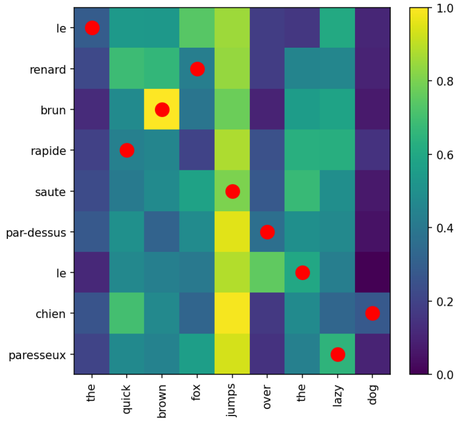
- Prompt: { The quick brown fox jumps over the lazy dog
  Le renard brun rapide saute par-dessus le chien paresseux .
- OT solution of work embeddings
  - x_i : word embeddings of English words
  - y_i : word embeddings of French words

$$P^* := \arg\min_P \sum_{ij} \|x_i - y_j\|^2 P_{ij} \text{ subject to } \ldots$$

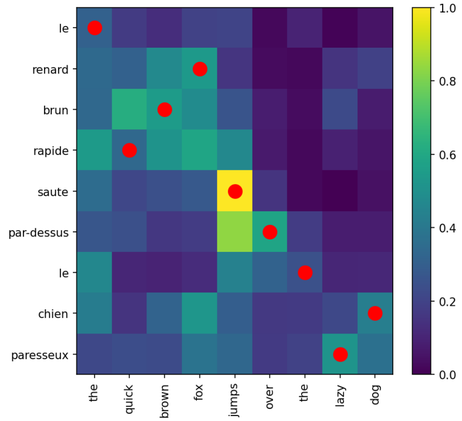


Attention, OT, translation

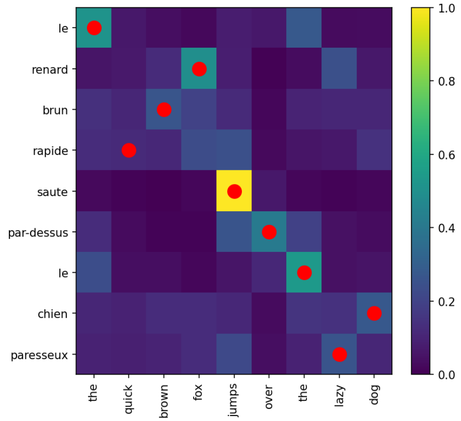
- Attention weights converge toward the OT solution across layers



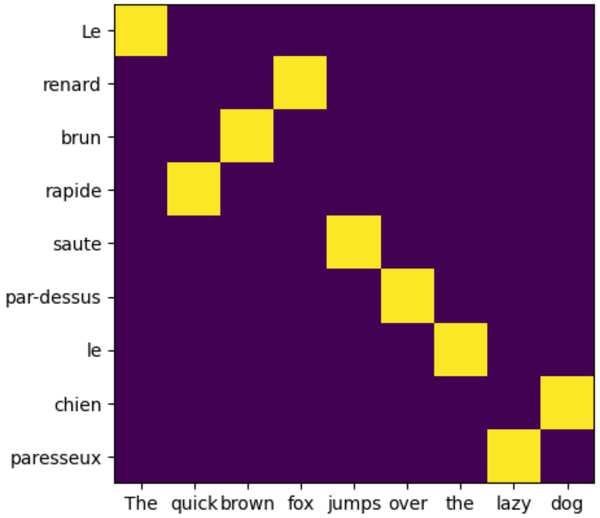
- Layer 1



- Layer 6



- Layer 12



- OT of embeddings

Main Result

Theorem (informal statement)

A transformer with ℓ softmax attention layers can approximate OT up to error

$$\frac{\sqrt{n}}{\sqrt{\ell}} \times \text{ data dependent and regularization constants}$$

Dual objective of OT

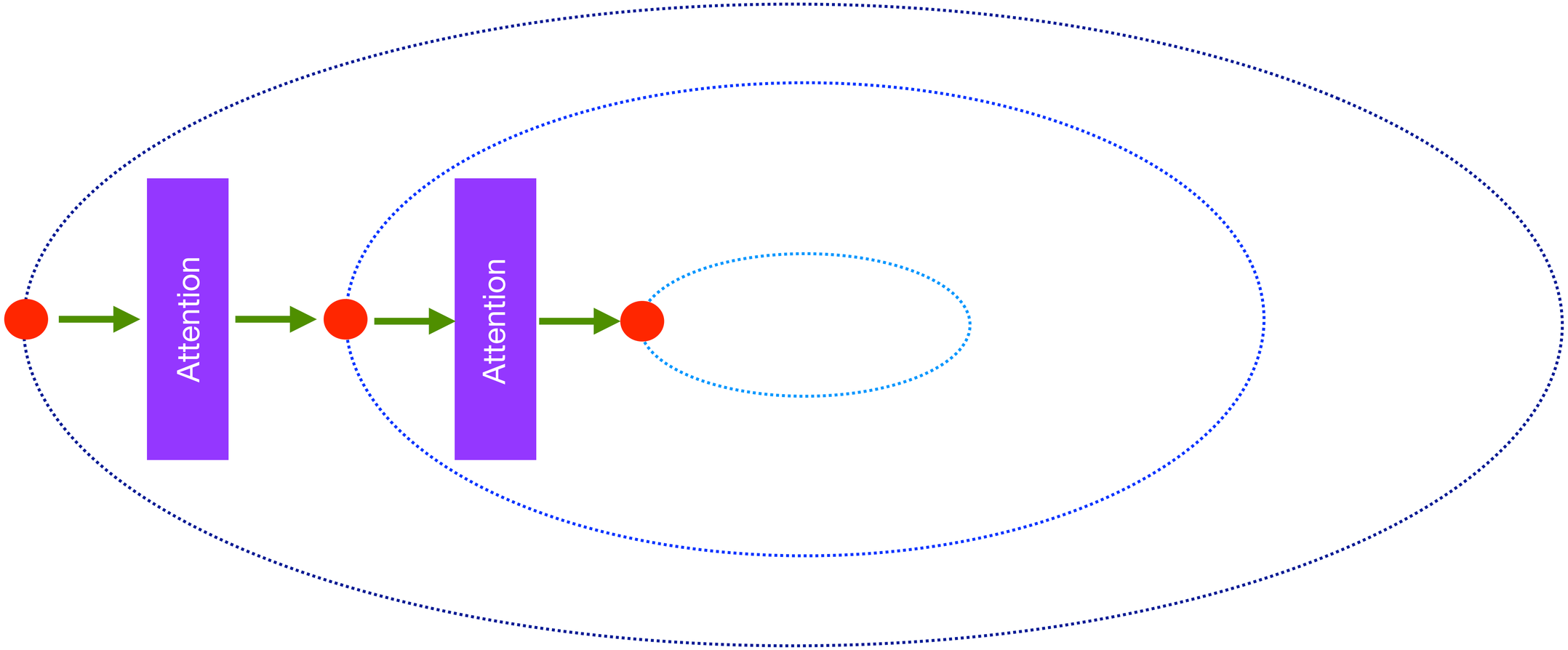
- Entropy Regularized OT: $P_\lambda = \arg\min_P \sum_{ij} \|x_i - y_j\|^2 P_{ij} + \underbrace{\lambda \sum_i \sum_j P_{ij} \log(P_{ij})}_{-entropy}$

- Dual: $\min_{u_i, v_j} L(u, v) = \lambda \sum_{ij} e^{\frac{-\|x_i - y_j\|^2 + u_i + v_j}{\lambda} - 1} - \sum_i v_i - \sum_j u_j$

Reference: Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances

Mechanism of Attention for Translation

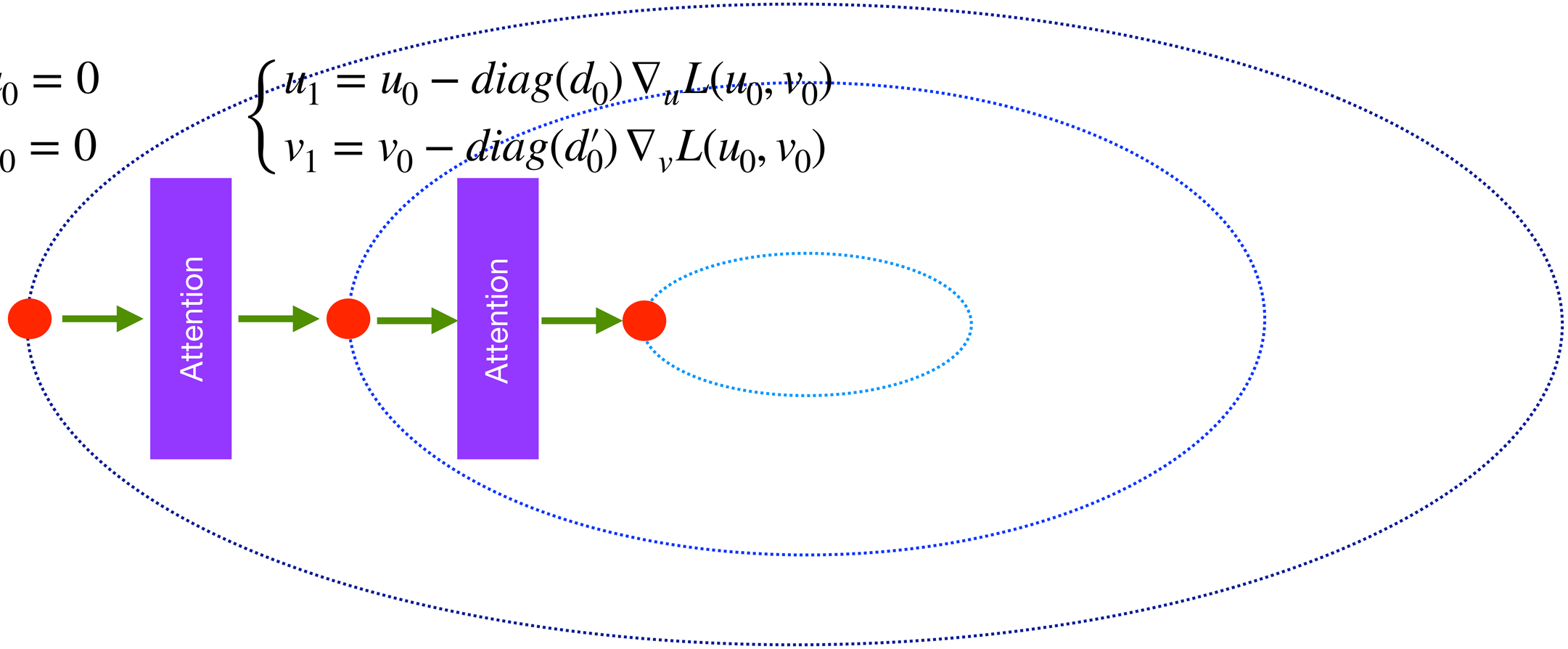
- Attention layers can simulate GD with adaptive step sizes on the dual function L



Mechanism of Attention for Translation

- Dual: $\min_{u_i, v_j} L(u, v) = \lambda \sum_{ij} e^{\frac{-\|x_i - y_j\|^2 + u_i + v_j}{\lambda} - 1} - \sum_i v_i - \sum_j u_j$

$$\begin{cases} u_0 = 0 \\ v_0 = 0 \end{cases} \qquad \begin{cases} u_1 = u_0 - diag(d_0) \nabla_u L(u_0, v_0) \\ v_1 = v_0 - diag(d_0') \nabla_v L(u_0, v_0) \end{cases}$$



Related works

- Results on linear attention (not softmax):
  - Linear regression: [Ahn et al., 2023, Von Oswald et al., 2023]
  - Policy evaluation for RL [Wang et al., 2025]
  - Low-rank matrix completion [Lutz et al., 2025]
- Mechanistic analysis of In-context learning [Garg et al., 2022, Akyürek et al., 2023]

Thank you very much for your attention

- Website: https://hadidaneshmand.github.io/papers/aistats26.html
- Github: https://github.com/hadidaneshmand/aistats26-incontextOT
- Contact: dhadi at virginia dot edu