# Attention Layers Provably Solve Single-Location Regression

## ICLR 2025

**Pierre Marion**, Raphaël Berthier, Gérard Biau, Claire Boyer

# Motivation: why does attention work well for NLP tasks?

Can we devise a task and Transformer-like model that features:

- token-wise sparsity
- randomness in the position of the relevant information

```
What have I become my sweetest friend?
Everyone I know goes ill in the end.
```

```
What have I become my sweetest friend?
Everyone I know goes well in the end.
```

```
Birds flying high, you know how I feel?
And, I'm feeling good.
```

```
Birds flying high, you know how I feel?
And, I'm feeling awful.
```

# Motivation: why does attention work well for NLP tasks?

Can we devise a task and Transformer-like model that features:

- token-wise sparsity
- randomness in the position of the relevant information
- linear structure in the representations

```
What have I become my sweetest friend?
Everyone I know goes ill in the end.
```

```
What have I become my sweetest friend?
Everyone I know goes well in the end.
```

```
Birds flying high, you know how I feel?
And, I'm feeling good.
```

```
Birds flying high, you know how I feel?
And, I'm feeling awful.
```

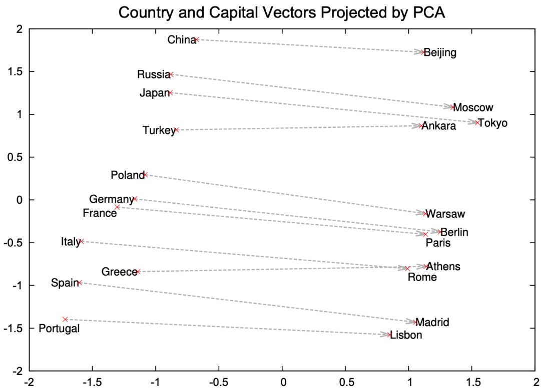


Figure from Mikolov, Sutskever, Chen, Corrado, Dean, 2013

# Statistical task: single-location regression

- Input: $L$ random tokens $(X_1, \dots, X_L)$ taking values in $\mathbb{R}^d$.
- Output: $Y \in \mathbb{R}$ given by

$$Y = X_{J_0}^\top v^\star + \xi,$$

# Statistical task: single-location regression

- Input: $L$ random tokens $(X_1, \dots, X_L)$ taking values in $\mathbb{R}^d$.
- Output: $Y \in \mathbb{R}$ given by

$$Y = X_{J_0}^\top v^\star + \xi,$$

where $J_0$ is a latent discrete random variable on $\{1, \dots, L\}$ and, conditionally on $J_0$,

$$\begin{cases} X_{J_0} & \sim \quad \mathcal{N}\Big(\sqrt{\tfrac{d}{2}} k^\star, \gamma^2 I_d\Big) \\ X_\ell & \sim \quad \mathcal{N}(0, I_d) \quad \text{for} \quad \ell \neq J_0 \,. \end{cases}$$

- $\xi \sim \mathcal{N}(0, \varepsilon^2)$ independent of everything else.
- Conditionally on $J_0$, the tokens $(X_j)_{1 \le j \le L}$ are independent.

# Statistical task: single-location regression

- Input: $L$ random tokens $(X_1, \ldots, X_L)$ taking values in $\mathbb{R}^d$.
- Output: $Y \in \mathbb{R}$ given by

$$Y = X_{J_0}^\top v^\star + \xi,$$

where $J_0$ is a latent discrete random variable on $\{1, \ldots, L\}$ and, conditionally on $J_0$,

$$\begin{cases} X_{J_0} & \sim & \mathcal{N}\left(\sqrt{\frac{d}{2}} k^\star, \gamma^2 I_d\right) \\ X_\ell & \sim & \mathcal{N}(0, I_d) \quad \text{for} \quad \ell \neq J_0 \,. \end{cases}$$

- $\xi \sim \mathcal{N}(0, \varepsilon^2)$ independent of everything else.
- Conditionally on $J_0$, the tokens $(X_j)_{1 \leq j \leq L}$ are independent.
- Notation: $(\mathbb{X}, Y) \sim \mathrm{SLR}(k^\star, v^\star)$

# From attention to our predictor

- Attention layer with a single head:

$$T_{\lambda}^{(Q,K,V,O)}(\mathbb{X}) = \operatorname{softmax}\Big(\lambda \underbrace{\mathbb{X}Q}_{L\times p} \underbrace{K^\top \mathbb{X}^\top}_{p\times L}\Big) \underbrace{\mathbb{X}V}_{L\times p} \underbrace{O^\top}_{p\times o} .$$

# From attention to our predictor

- Attention layer with a single head:

$$T_{\lambda}^{(Q,K,V,O)}(\mathbb{X}) = \mathrm{softmax}\Big(\lambda \underbrace{\mathbb{X}Q}_{L\times p} \underbrace{K^{\top}\mathbb{X}^{\top}}_{p\times L}\Big) \underbrace{\mathbb{X}V}_{L\times p} \underbrace{O^{\top}}_{p\times o} .$$

- Consider the output for the first token ([CLS] token):

$$T_{\lambda}^{(Q,K,V,O)}(\mathbb{X})_1 = \mathrm{softmax}\left(\lambda \, a \, K^{\top}\mathbb{X}^{\top}\right) \mathbb{X}VO^{\top} .$$

# From attention to our predictor

- Attention layer with a single head:

$$T_\lambda^{(Q,K,V,O)}(\mathbb{X}) = \mathrm{softmax}\Big(\lambda \underbrace{\mathbb{X}Q}_{L\times p} \underbrace{K^\top \mathbb{X}^\top}_{p\times L}\Big) \underbrace{\mathbb{X}V}_{L\times p} \underbrace{O^\top}_{p\times o} \,.$$

- Consider the output for the first token ([CLS] token):

$$T_\lambda^{(Q,K,V,O)}(\mathbb{X})_1 = \mathrm{softmax}\left(\lambda\, a\, K^\top \mathbb{X}^\top\right) \mathbb{X} V O^\top \,.$$

- Take $o = 1$ and $p = 1$:

$$T_\lambda^{(Q,K,V,O)}(\mathbb{X})_1 = \mathrm{softmax}\left(\lambda\, k^\top \mathbb{X}^\top\right) \mathbb{X} v \,.$$

# From attention to our predictor

- Attention layer with a single head:

$$T_\lambda^{(Q,K,V,O)}(\mathbb{X}) = \mathrm{softmax}\Big(\lambda\, \underbrace{\mathbb{X}Q}_{L\times p}\, \underbrace{K^\top \mathbb{X}^\top}_{p\times L}\Big)\, \underbrace{\mathbb{X}V}_{L\times p}\, \underbrace{O^\top}_{p\times o}\ .$$

- Consider the output for the first token ([CLS] token):

$$T_\lambda^{(Q,K,V,O)}(\mathbb{X})_1 = \mathrm{softmax}\left(\lambda\, a\, K^\top \mathbb{X}^\top\right) \mathbb{X} V O^\top\ .$$

- Take $o = 1$ and $p = 1$:

$$T_\lambda^{(Q,K,V,O)}(\mathbb{X})_1 = \mathrm{softmax}\left(\lambda\, k^\top \mathbb{X}^\top\right) \mathbb{X} v\,.$$

- Replace softmax by a component-wise nonlinearity:

$$T_\lambda^{(k,v)}(\mathbb{X}) = \sigma\big(\lambda k^\top \mathbb{X}^\top\big)\mathbb{X} v = \sum_{\ell=1}^{L} \sigma\big(\lambda X_\ell^\top k\big) X_\ell^\top v\,.$$

# Our result

For $(k, v) \in \mathbb{S}^{d-1}$, let

$$\mathcal{R}_\lambda(k, v) = \mathbb{E}_{(\mathbb{X}, Y) \sim \mathsf{SLR}(k^\star, v^\star)} \left[ \left( Y - \sigma\left(\lambda \mathbb{X} k\right)^\top \mathbb{X} v \right)^2 \right].$$

# Our result

For $(k, v) \in \mathbb{S}^{d-1}$, let

$$\mathcal{R}_\lambda(k, v) = \mathbb{E}_{(\mathbb{X}, Y) \sim \mathsf{SLR}(k^\star, v^\star)}\left[\left(Y - \sigma\left(\lambda \mathbb{X} k\right)^\top \mathbb{X} v\right)^2\right].$$

## Theorem (Informal)

With proper choice of $\sigma$, initialization and stepsize, projected gradient descent on $\mathcal{R}_\lambda$ satisfies

$$(k_t, v_t) \xrightarrow[t \to \infty]{} \pm(k^\star, v^\star).$$

Furthermore, in the asymptotic regime $d \to \infty$, $\lambda\sqrt{d} \to \infty$, $\lambda\sqrt{L} \to 0$, the predictor $T_\lambda^{(k^\star, v^\star)}$ is asymptotically Bayes optimal.

# Takeaways

- A (reasonable?) model for how attention layers learn to encode linear representations of the data, while handling token-wise sparsity and randomness in the token positions.
- Open questions: impact of the temperature $\lambda$, practical initialization schemes, multiple heads.