Self-attention Networks Localize When QK-eigenspectrum Concentrates

June 27 (Thu.), 2024

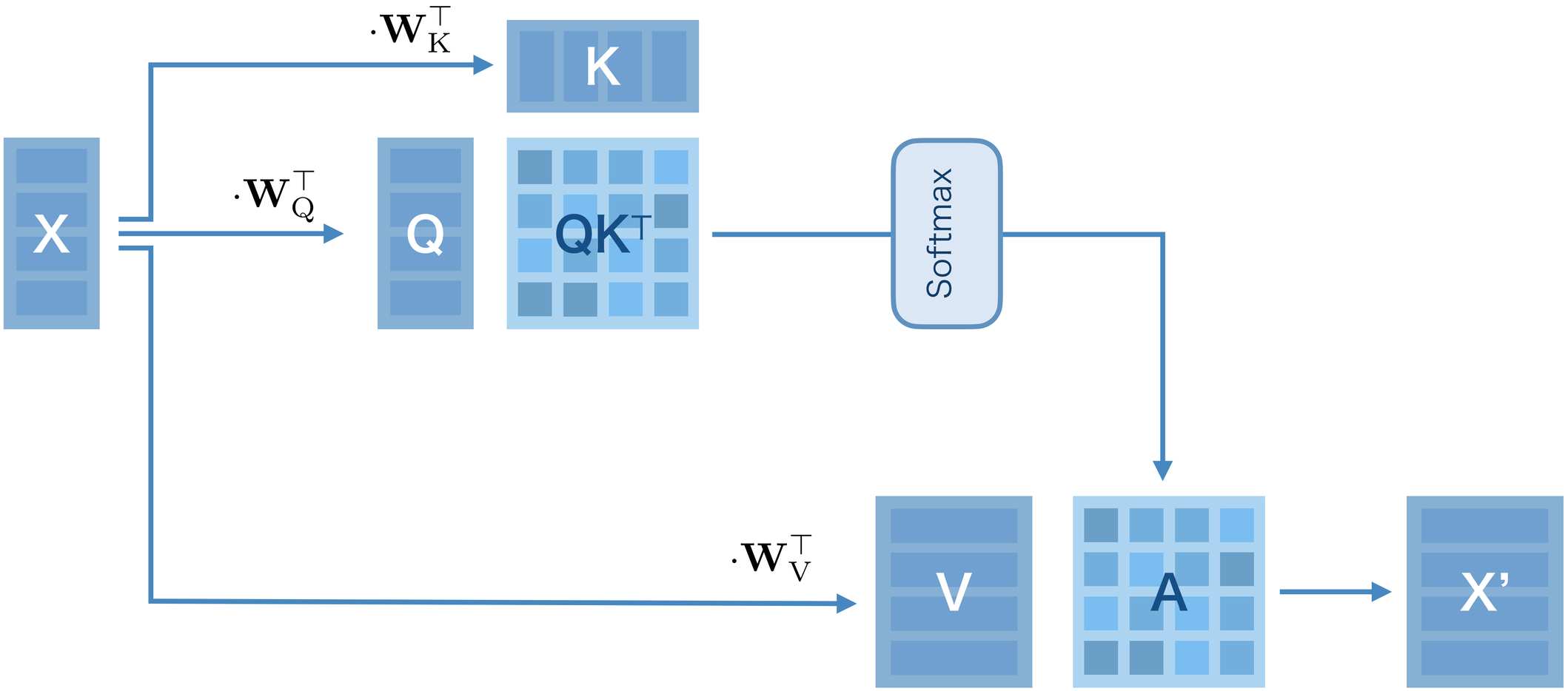
Han Bao

Great collaborators!

Ryuichiro Hataya (RIKEN)

Ryo Karakida (AIST)

Attention is prevailing …



Attention as a visualizer

[Vaswani+ 2017]

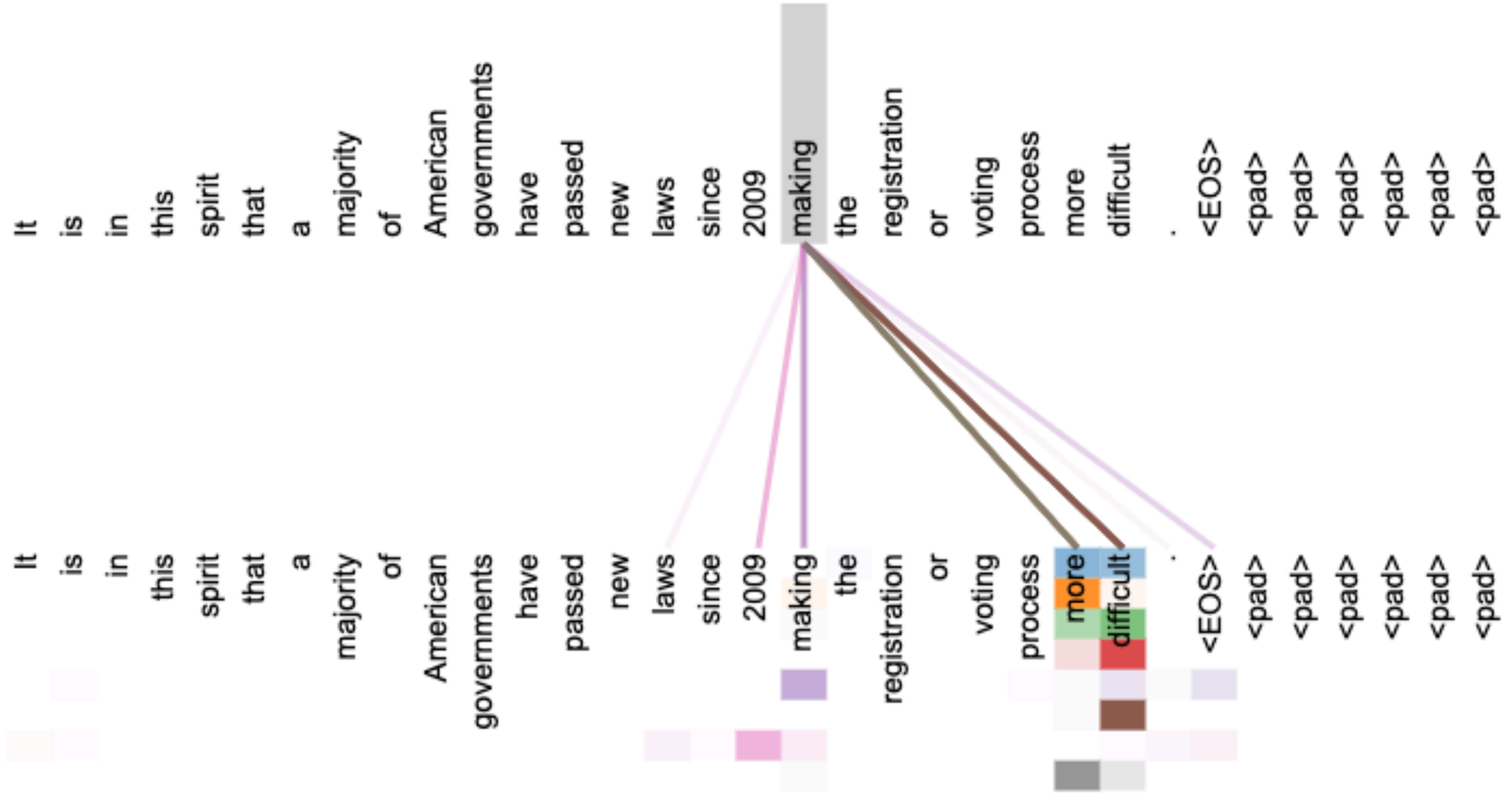


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In NeurIPS.

Attention as an aligner

[Bahdanau+ 2015]

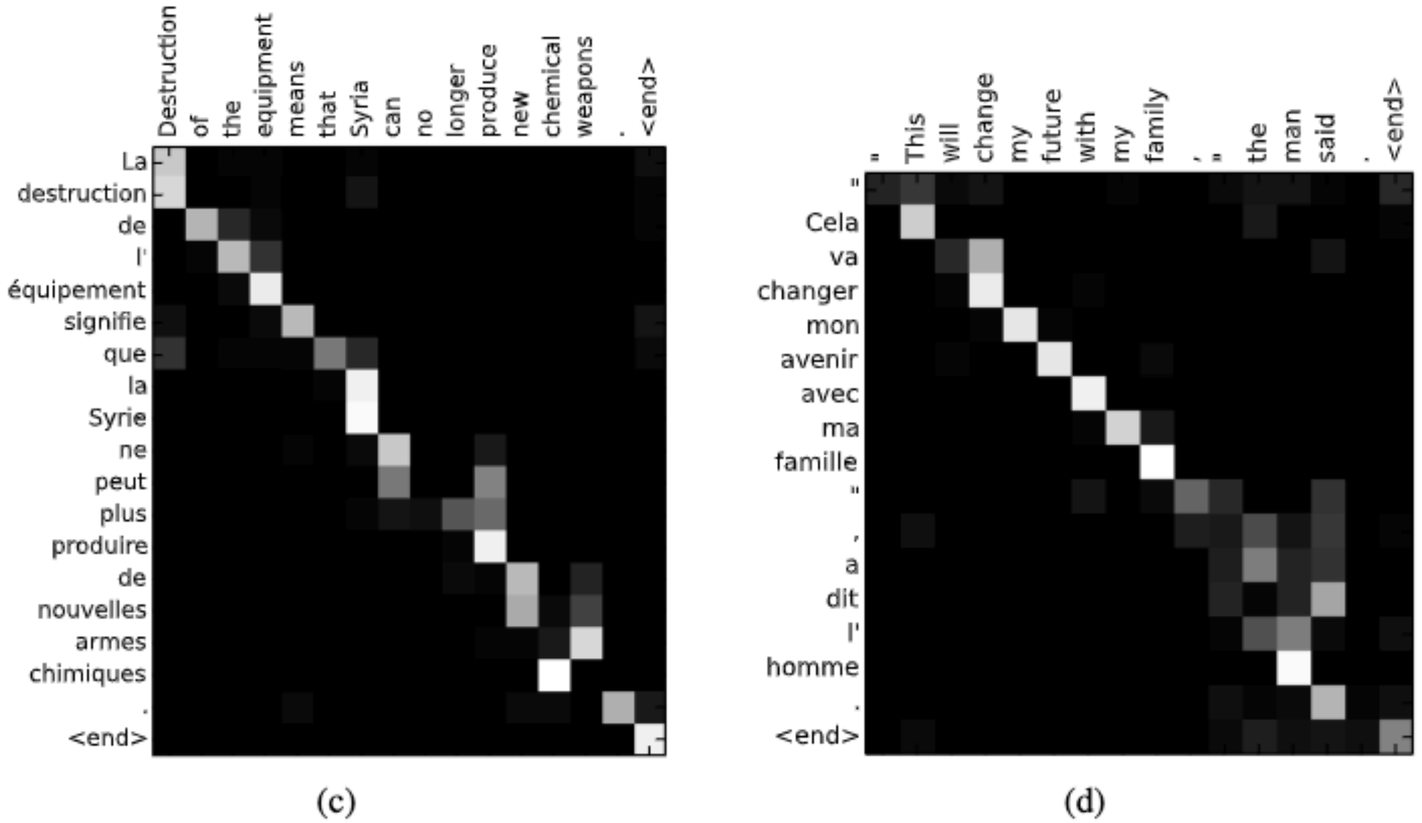


Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word (see Eq. (6)), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b–d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In ICLR.

Attention as a permutation-invariant learner

[Veličković+ 2018]

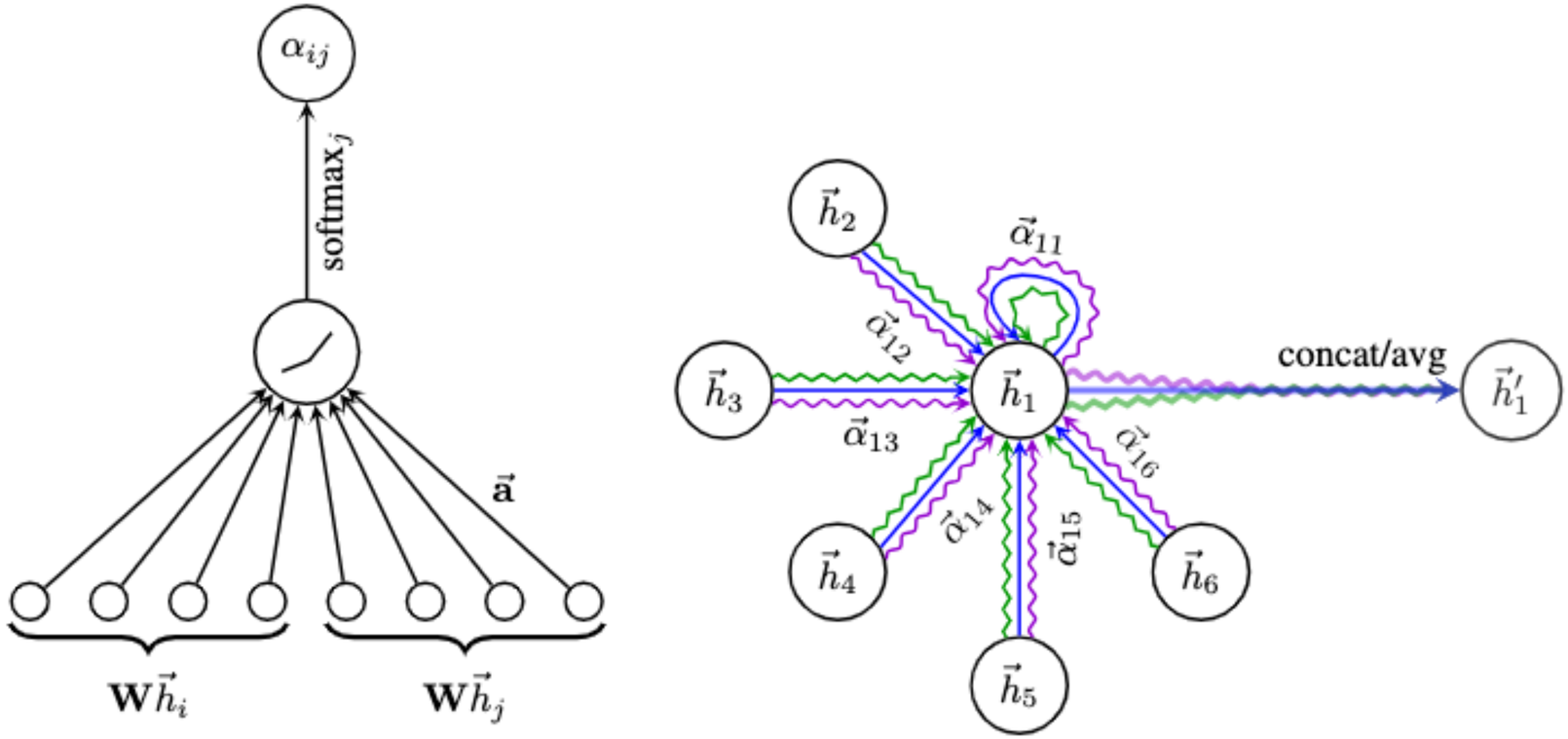


Figure 1: **Left:** The attention mechanism $a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$ employed by our model, parametrized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$, applying a LeakyReLU activation. **Right:** An illustration of multi-head attention (with $K = 3$ heads) by node 1 on its neighborhood. Different arrow styles and colors denote independent attention computations. The aggregated features from each head are concatenated or averaged to obtain $\vec{h}'_1$.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. In ICLR.

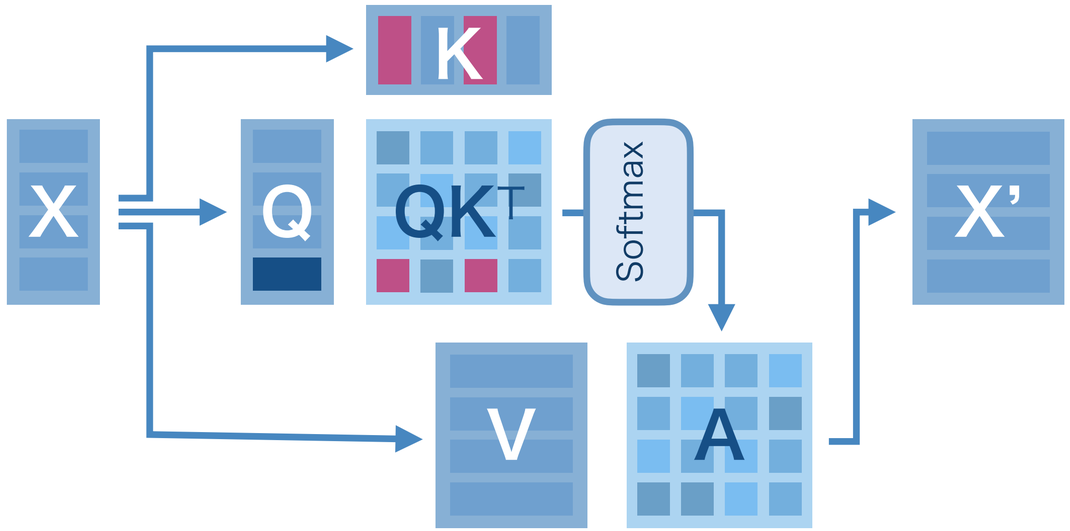
Why is attention attractive for us?

Recent research advances in machine learning community

Why attention attracts ML people? (1/3)

[Tarzanagh+ 2023]

- Mixing mechanism — How does attention selectively chooses relevant tokens?

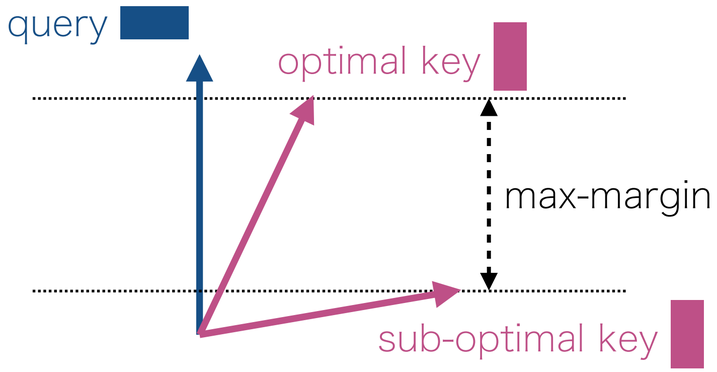


QK score selects the most relevant ■ to ■

key query

Simplification

- one-layer attention net
- key tokens are fixed
- only learns query token



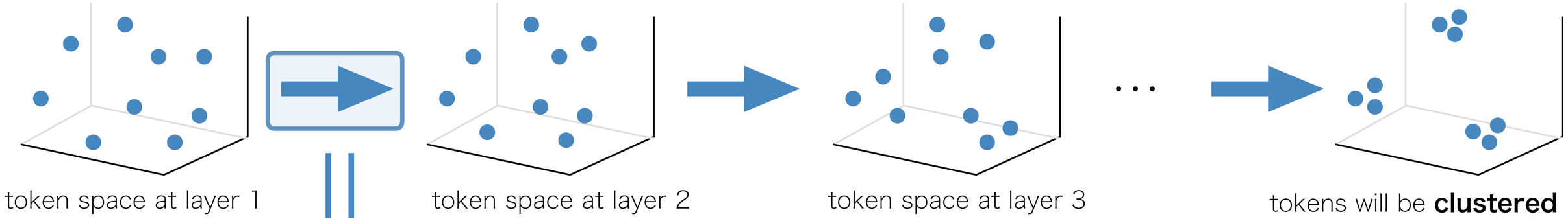
Learned query implicitly **maximize margin**

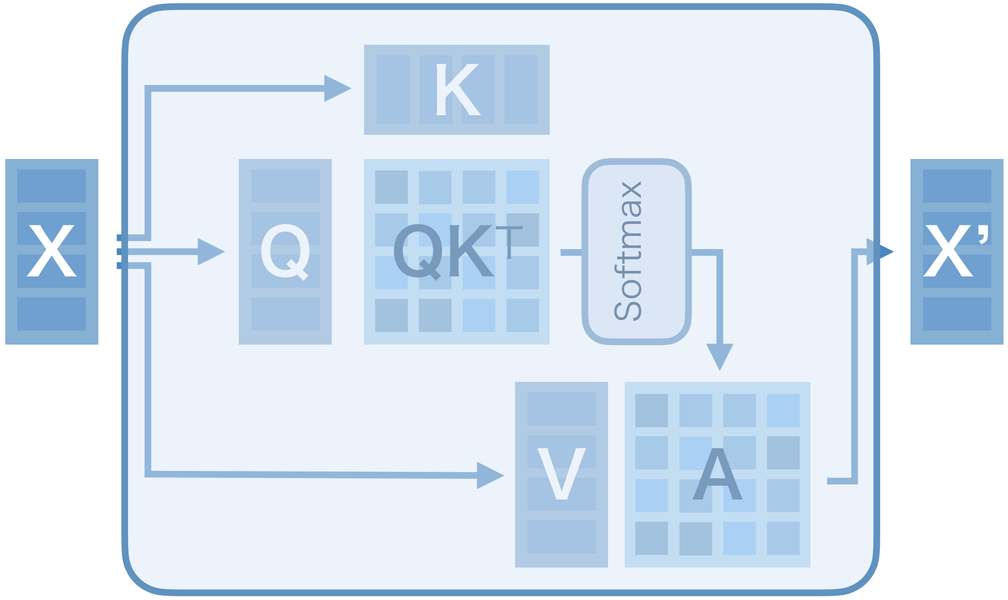
Tarzanagh, D. A., Li, Y., Zhang, X., & Oymak, S. (2023). Max-margin token selection in attention mechanism. In NeurIPS.

Why attention attracts ML people? (2/3)

[Geshkovski+ 2023]

- Representation learning — What encoded features should look like?





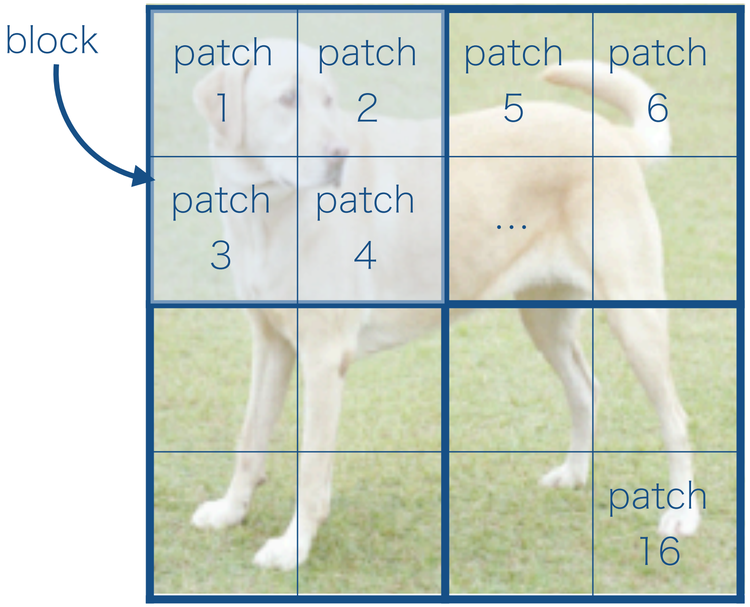
Geshkovski, B., Letrouit, C., Polyanskiy, Y., & Rigollet, P. (2023). The emergence of clusters in self-attention dynamics. In NeurIPS.

Why attention attracts ML people? (3/3)

[Jelassi+ 2022]

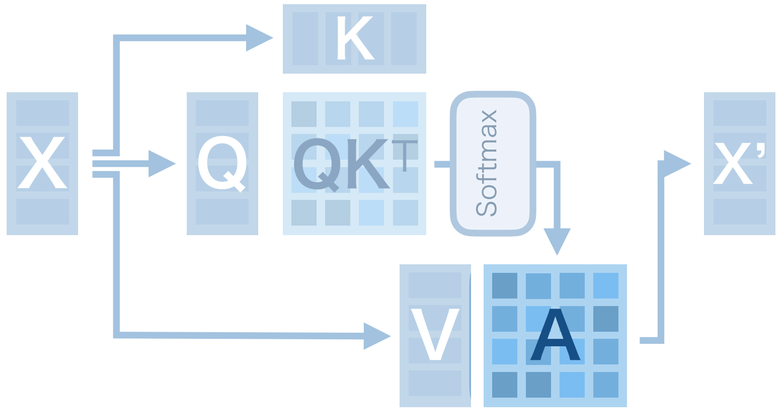
- Structure prediction — How does attention efficiently extract structured patterns?

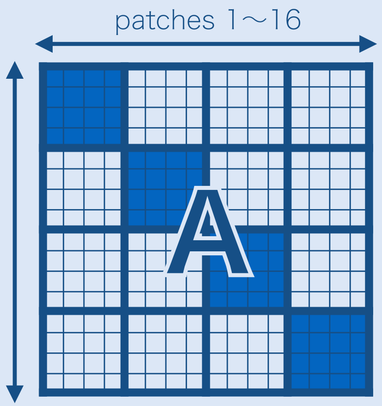
Let's consider ViT input



patches are structured within each block

after training ViT
with gradient descent





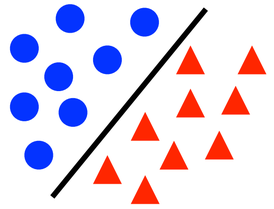
attention matrix captures
within-block association

Jelassi, S., Sander, M., & Li, Y. (2022). Vision transformers provably learn spatial structure. In NeurIPS.

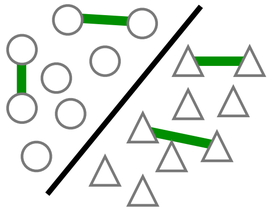
Why attention attracts ML people? (extra)

- "Similarity" among data points is interesting/cheap signal lying in data!
  - Classification with only **similarity** & unlabeled data (SU classification)
    Bao, H., Niu, G., & Sugiyama, M. (ICML2018).
    Classification from Pairwise Similarity and Unlabeled Data.
  - Classification with **similarity**, dis**similarity**, & unlabeled data
    Shimada, T., **Bao, H.**, Sato, I., & Sugiyama, M. (NeCo2021)
    Classification from Pairwise Similarities/Dissimilarities and Unlabeled Data via Empirical Risk Minimization.
  - Relationship between **metric** learning and classification
    Bao, H.*, Shimada, T.*, Xu, L., Sato, I., & Sugiyama, M. (AISTATS2022)
    Pairwise Supervision Can Provably Elicit a Decision Boundary.
  - Relationship between **contrastive** learning and linear probing
    Bao, H., Nagano, Y., & Nozawa, N. (ICML2022)
    On the Surrogate Gap between Contrastive and Supervised Losses.

Supervised Classification



SU Classification



Tons of mysteries behind attention!

- Many practical applications
  - Cross attention for multimodal inputs
  - Interpretability
  - Graph/sequence/structured inputs
- Yet our understanding is elusive
  - (Layer-/epoch-wise) dynamics
  - Representation learning
  - Expressivity

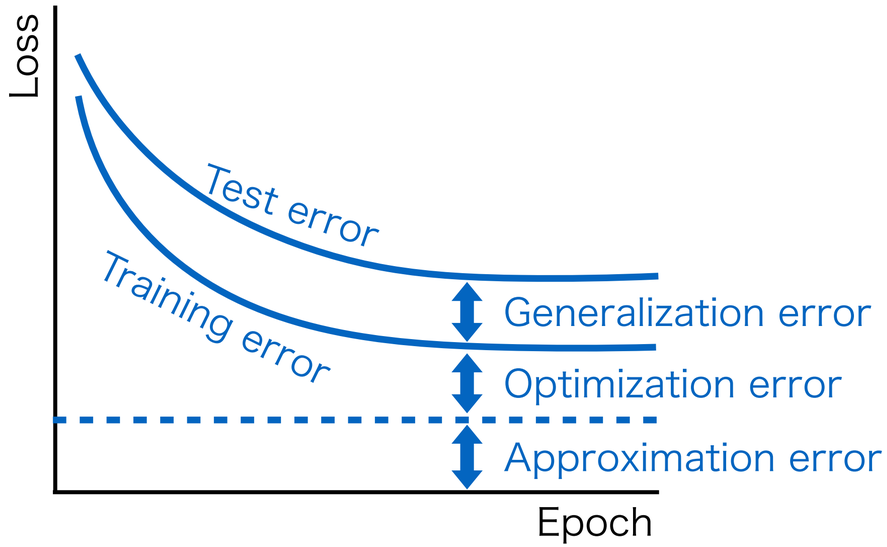
How attention matters in learning

From eigenspectrum perspective

Three factors in learning theory (in general)

- Generalization (↔ complexity error)
- Optimization (↔ optimization error)
- Model expressivity (↔ approximation error)

Error decomposition of learning curve

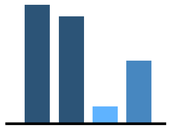


Cropped from Imaizumi-san's slides at MLSS'24

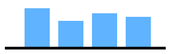
https://groups.oist.jp/sites/default/files/imce/u129210/mlss/Lecture_slide/MLSS_Masaaki_Imaizumi.pdf

Related work | Concentrated attention is better!

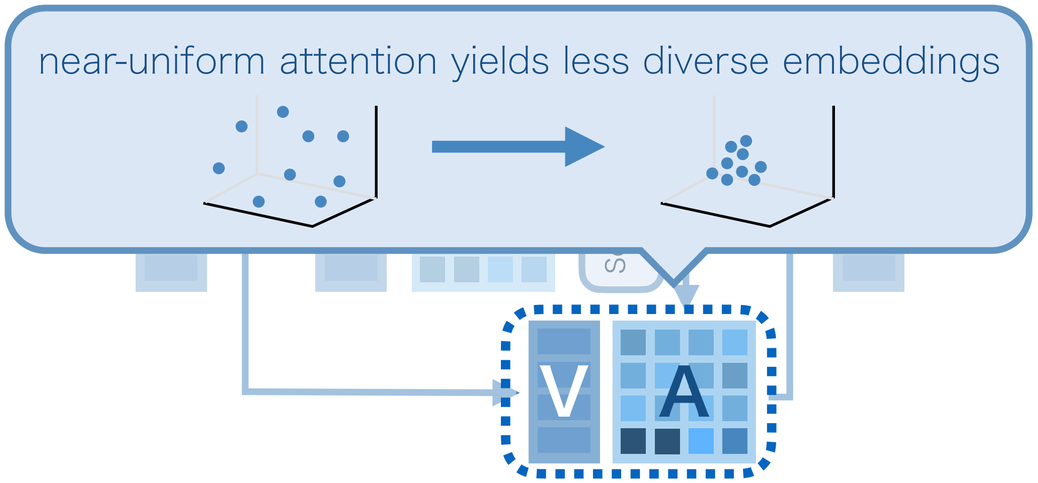
- **Rank collapse** [Dong+ 2021]: attention is close to uniform (roughly speaking)
  - induces **low model expressivity**
    (yielding large approximation error ☹)



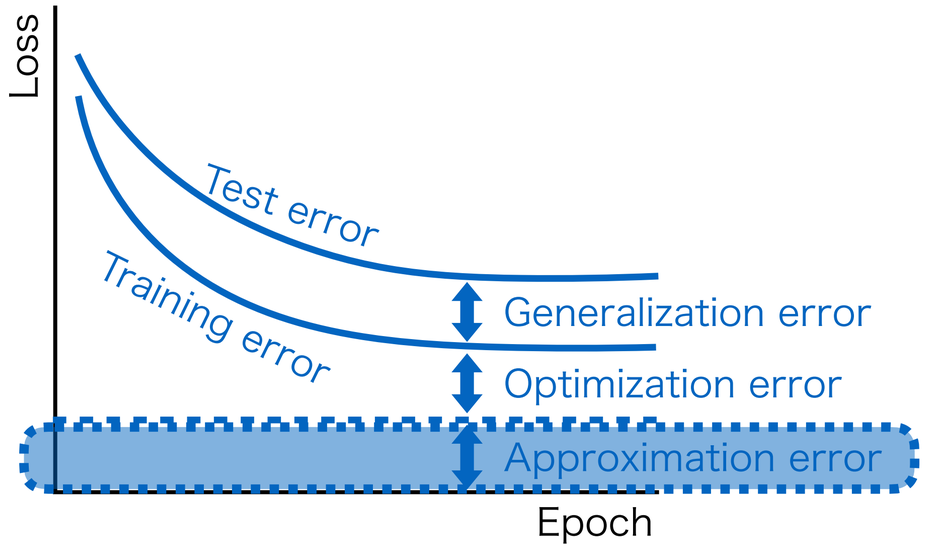
non-uniform attention
→ flexible model ☺



uniform attention
→ not flexible model ☹



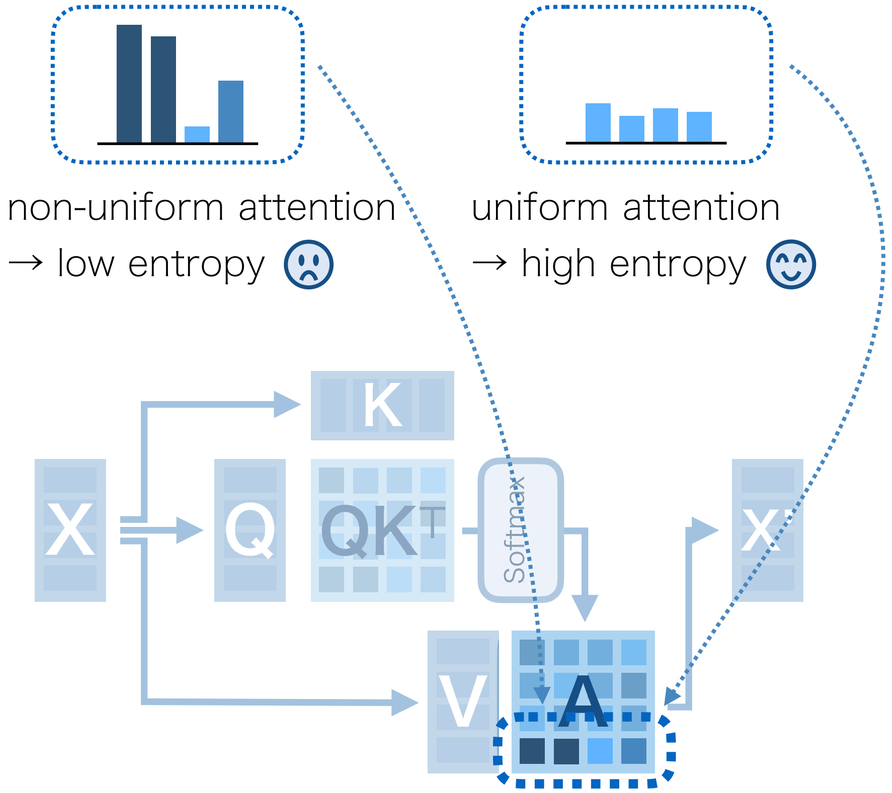
Error decomposition of learning curve



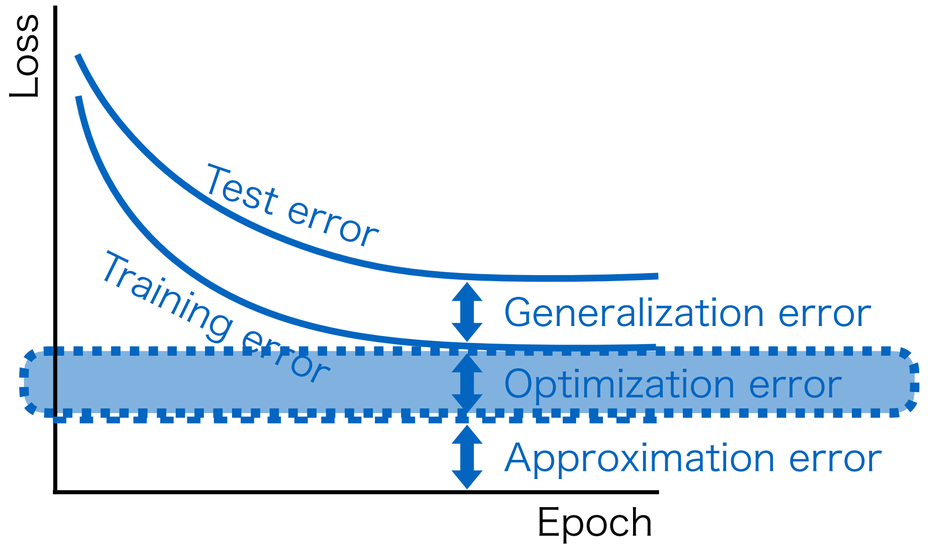
Dong, Y., Cordonnier, J. B., & Loukas, A. (2021). Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In ICML.

Related work | Uniform attention is better!

- **Entropy collapse** [Zhai+ 2023]: entropy of attention (as a probability distribution) is low
  - tends to **get stuck into plateaus** of loss landscape (yielding large optimization error ☹)

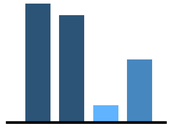


Error decomposition of learning curve



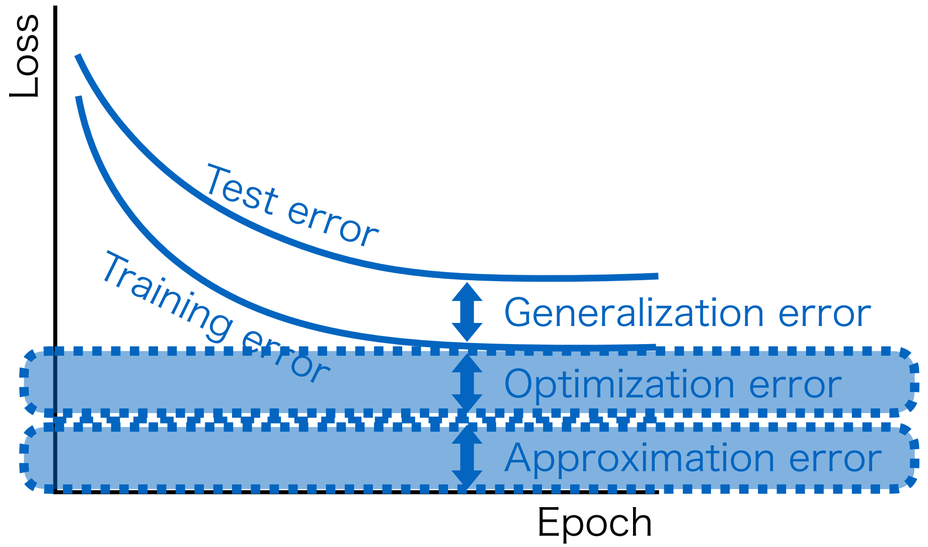
Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., ... & Susskind, J. M. (2023). Stabilizing transformer training by preventing attention entropy collapse. In ICML.

How should attention look like??

Attention is …	 concentrated	 uniform
Collapse	Entropy	Rank
Optimization error	High ☹	Low ☺
Approximation error	Low ☺	High ☹

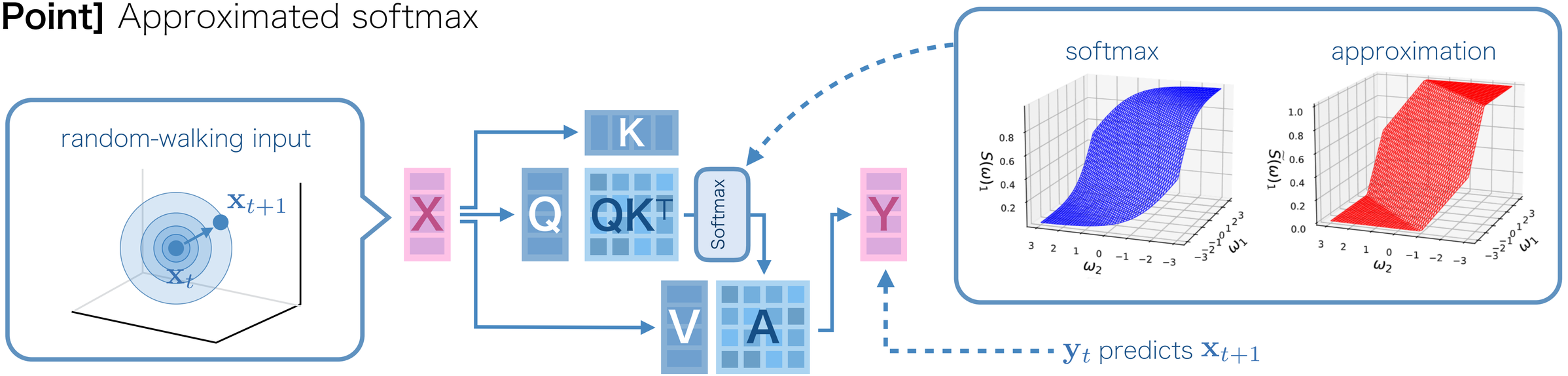
RQ. Are they reconcilable?
RQ. When concentrated/uniform?

Error decomposition of learning curve



Setup | Our analysis model

- Model: 1-layer 1-head self-attention network
  - without layer normalization & positional encoding / with feed-forward net
- Pre-training task: next token prediction
- **[Point]** Data: Gaussian random walk $\mathbf{x}_{t+1} \sim \mathcal{N}(\mathbf{x}_t, \mathbf{\Sigma})$
  - disclaimer: no embedding layer (token -> emb); $\mathbf{x}_t$ is a direct input to TF
- **[Point]** Approximated softmax



Main result | Characterize attention shape

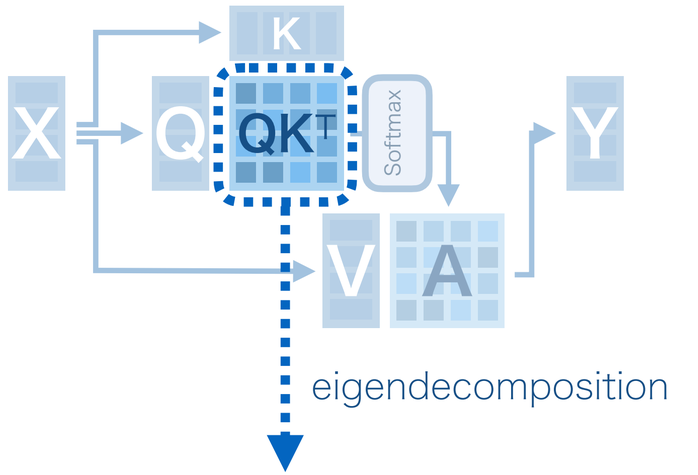
- $\mathrm{tr}(\mathbf{W})$ ($\doteq$ eigval mean) and $\mathrm{tr}(\mathbf{W}^2)$ ($\doteq$ eigval variance) determines attention shape
  - $\mathbf{W}^2 := \mathbf{W}_{\mathrm{Q}}^\top \mathbf{W}_{\mathrm{K}} \mathbf{\Sigma}$ is scaled QK parameter

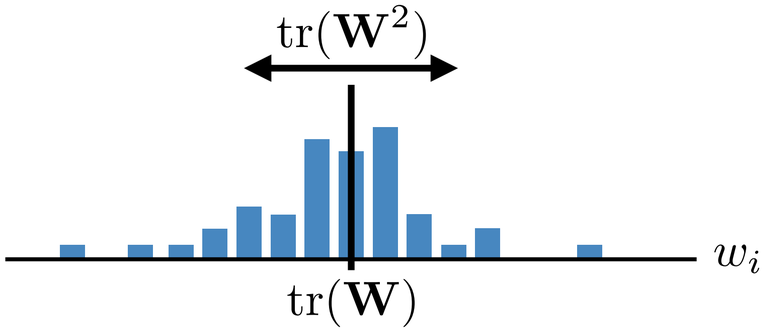
under mild assumptions:

$$\frac{1}{d}\mathrm{tr}(\mathbf{W}) = \frac{1}{d}\sum_{i=1}^{d} w_i \qquad \text{(1st moment)}$$

i-th eigenvalue

$$\frac{1}{d}\mathrm{tr}(\mathbf{W}^2) = \frac{1}{d}\sum_{i=1}^{d} w_i^2 \qquad \text{(2nd moment)}$$





Main result | Characterize attention shape

- $\mathrm{tr}(\mathbf{W})$ ($\doteqdot$ eigval mean) and $\mathrm{tr}(\mathbf{W}^2)$ ($\doteqdot$ eigval variance) determines attention shape
  - $\mathbf{W}^2 := \mathbf{W}_{\mathrm{Q}}^{\top}\mathbf{W}_{\mathrm{K}}\boldsymbol{\Sigma}$ is scaled QK parameter

analytical expression of
attention strength at token position θ

$$\rho(\theta) := \Phi\left(\left(\theta - \frac{1}{2}\right)\xi; \theta\right) - \Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}; \theta\right),$$

$$\xi := \frac{\mathrm{tr}(\mathbf{W})}{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}, \quad \eta := \frac{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}{\lambda},$$

$$\Phi(z; \theta) := \frac{1}{2}\mathrm{erf}\left(\frac{z}{\sqrt{2(2\theta^2 + \frac{7}{12})}}\right),$$

Main result | Characterize attention shape

- $\mathrm{tr}(\mathbf{W})$ ($\fallingdotseq$ eigval mean) and $\mathrm{tr}(\mathbf{W}^2)$ ($\fallingdotseq$ eigval variance) determines attention shape
  - $\mathbf{W}^2 := \mathbf{W}_{\mathrm{Q}}^\top \mathbf{W}_{\mathrm{K}} \mathbf{\Sigma}$ is scaled QK parameter

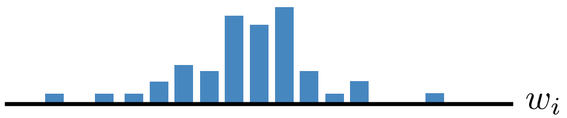
analytical expression of
attention strength at token position θ

$$\rho(\theta) := \Phi\left(\left(\theta - \frac{1}{2}\right)\xi; \theta\right) - \Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}; \theta\right),$$

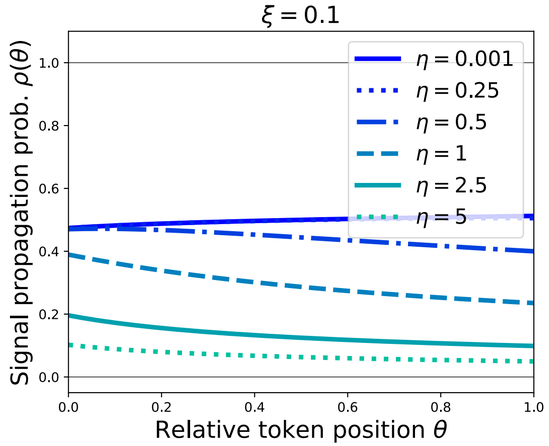
$$\xi := \frac{\mathrm{tr}(\mathbf{W})}{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}, \quad \eta := \frac{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}{\lambda},$$

$$\Phi(z; \theta) := \frac{1}{2}\mathrm{erf}\left(\frac{z}{\sqrt{2(2\theta^2 + \frac{7}{12})}}\right),$$

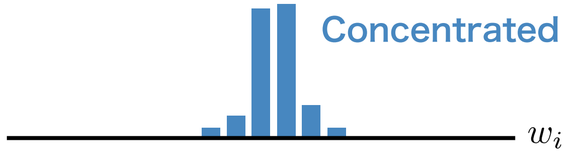
When $\mathrm{tr}(\mathbf{W}) \fallingdotseq 0$ & $\mathrm{tr}(\mathbf{W}^2) < \infty$



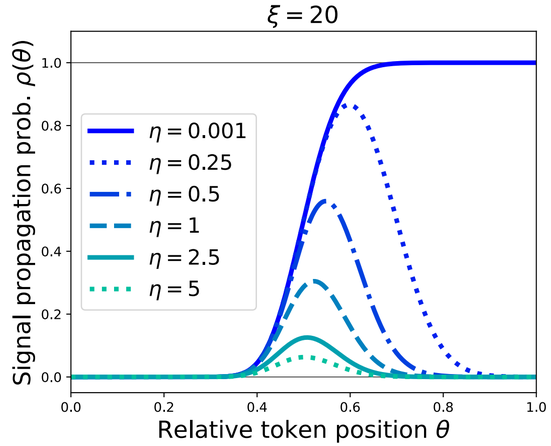
Uniform attention



When $\mathrm{tr}(\mathbf{W}) \gg 0$ & $\mathrm{tr}(\mathbf{W}^2) \fallingdotseq 0$



Non-uniform attention



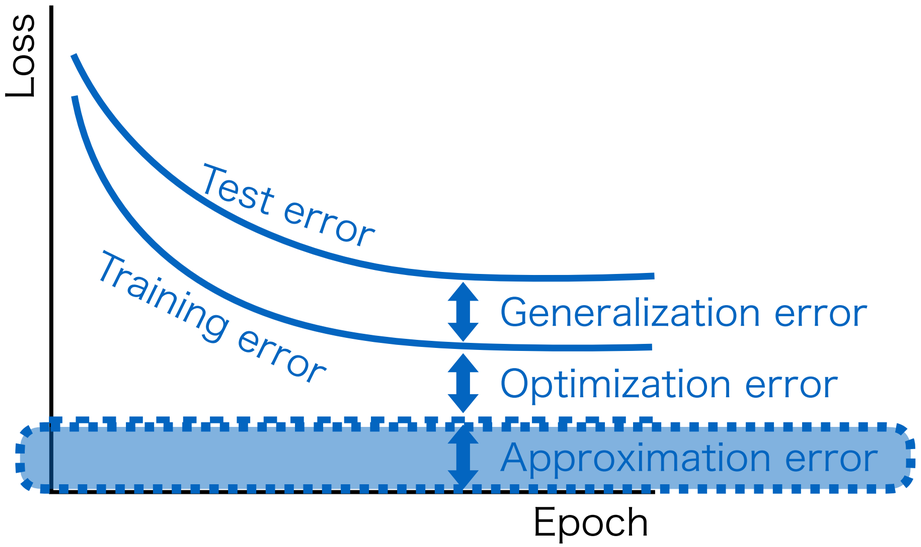
Relation to rank collapse

- **Rank collapse**: attention is close to uniform (roughly speaking)
  - [Dong+ 2021] large $\|\mathbf{W}_{\mathrm{QK}}\|_1$ (L1 norm) prevents rank collapse

- From eigenspectrum viewpoint
  - we can show $\|\mathbf{W}_{\mathrm{QK}}\|_1 \geq C|\mathrm{tr}(\mathbf{W})|$
  - smaller eigenspectrum variance ⇒ larger $\|\mathbf{W}_{\mathrm{QK}}\|_1$

$$\underbrace{\mathrm{Var}[w_i]}_{\text{decrease}} = \frac{1}{d}\underbrace{\mathrm{tr}(\mathbf{W}^2)}_{\text{fixed}} - \frac{1}{d^2}\underbrace{|\mathrm{tr}(\mathbf{W})|^2}_{\text{increase}}$$

Error decomposition of learning curve



Dong, Y., Cordonnier, J. B., & Loukas, A. (2021). Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In ICML.

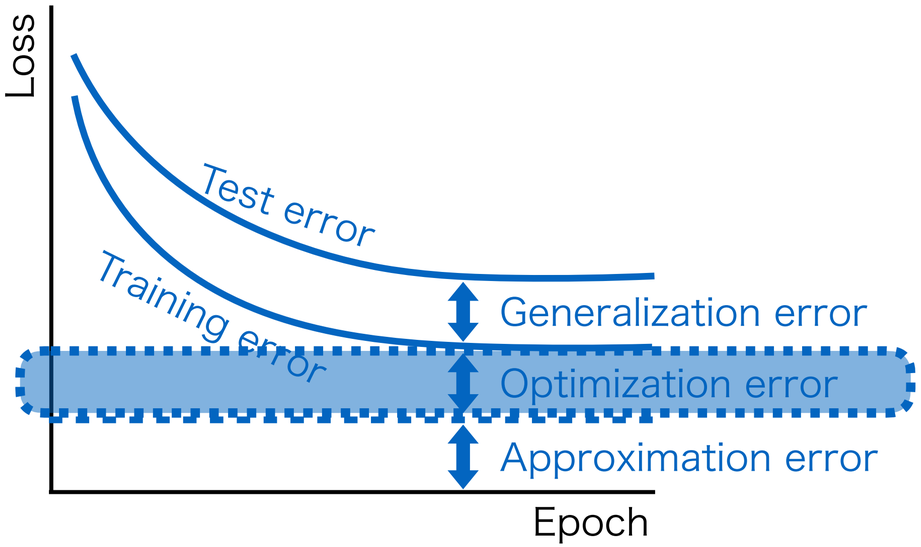
Relation to entropy collapse

- **Entropy collapse**: entropy of attention (as a probability distribution) is low
  - [Zhai+ 2023] small $\|\mathbf{W}_{\mathrm{QK}}\|_2$ (spectral norm) keeps attention entropy high

- From eigenspectrum viewpoint
  - we can show $\|\mathbf{W}_{\mathrm{QK}}\|_2 \leq C\sqrt{\mathrm{tr}(\mathbf{W}^2)}$
  - smaller eigenspectrum variance ⇒ smaller $\|\mathbf{W}_{\mathrm{QK}}\|_2$

$$\underbrace{\mathrm{Var}[w_i]}_{\text{decrease}} = \frac{1}{d}\underbrace{\mathrm{tr}(\mathbf{W}^2)}_{\text{decrease}} - \frac{1}{d^2}\underbrace{|\mathrm{tr}(\mathbf{W})|^2}_{\text{fixed}}$$

Error decomposition of learning curve



Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., ... & Susskind, J. M. (2023). Stabilizing transformer training by preventing attention entropy collapse. In ICML.

Attention shape through eigenspectrum

RQ. Are two collapse phenomena reconcilable?

A. Yes, let's decrease eigenspectrum variance!

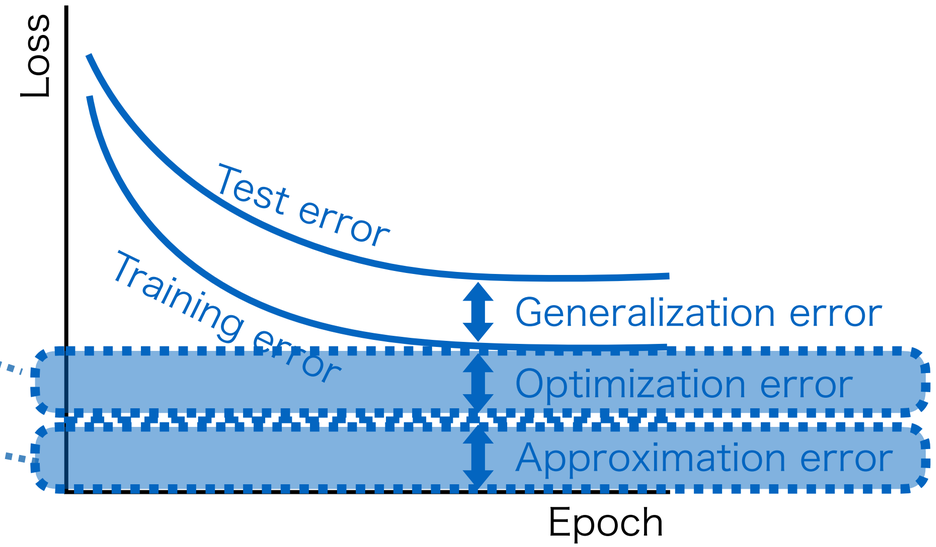
Disclaimer: this is only a **sufficient** condition!

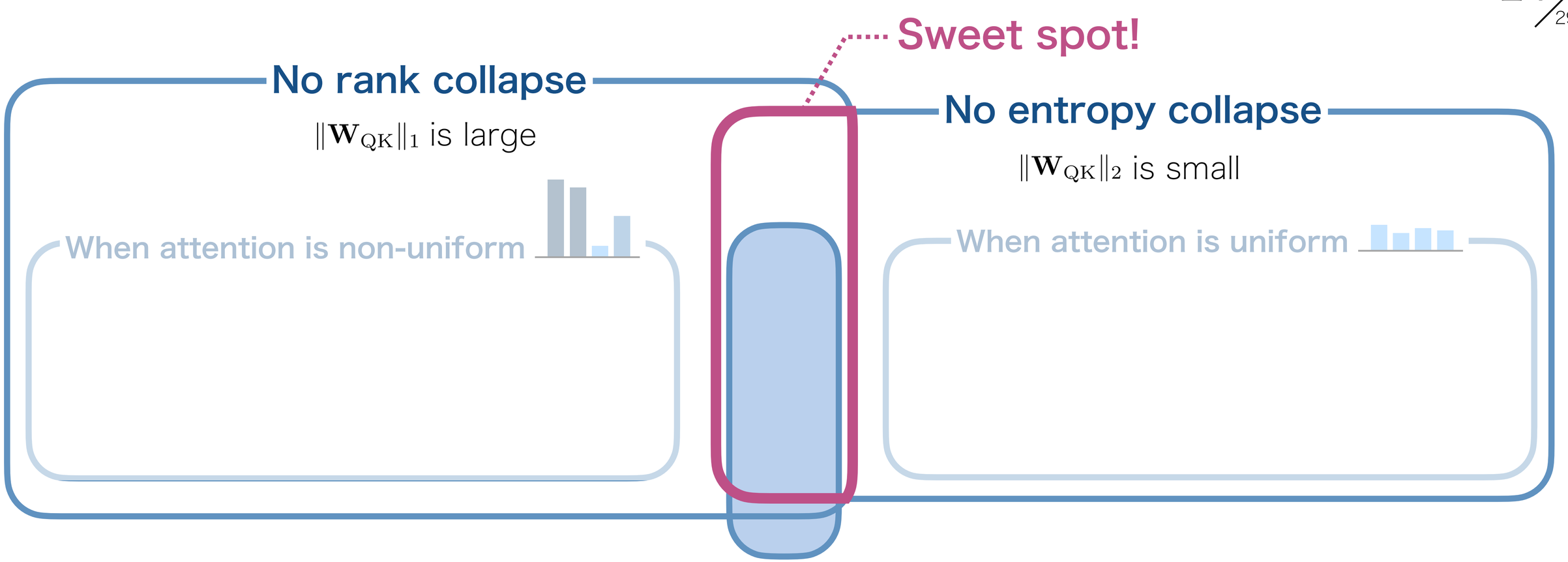
$$\underbrace{\mathrm{Var}[w_i]}_{\text{decrease}} = \frac{1}{d}\underbrace{\mathrm{tr}(\mathbf{W}^2)}_{\text{decrease}} - \frac{1}{d^2}\underbrace{|\mathrm{tr}(\mathbf{W})|^2}_{\text{increase}}$$

avoids entropy collapse

avoids rank collapse

Error decomposition of learning curve





$$\mathrm{Var}[w_i] = \frac{1}{d}\mathrm{tr}(\mathbf{W}^2) - \frac{1}{d^2}|\mathrm{tr}(\mathbf{W})|^2 \text{ is small}$$

In more depth

Signal propagation probability | Why it matters?

- What is “attention strength”?

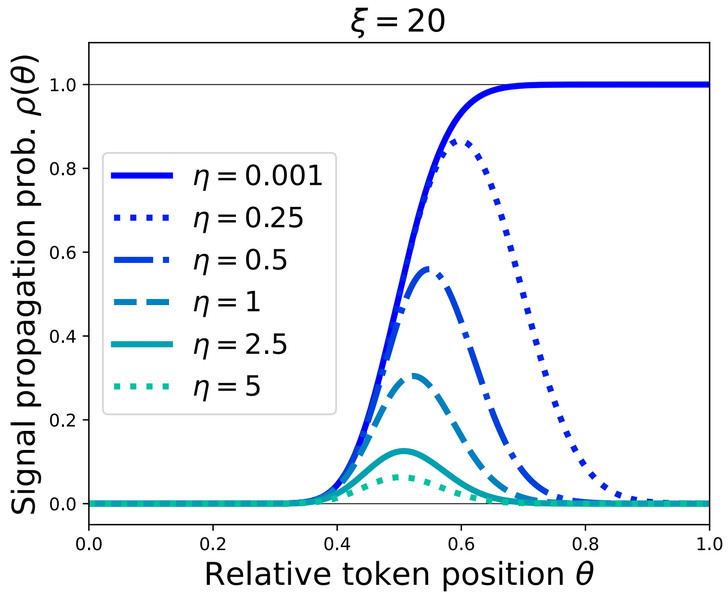
analytical expression of

attention strength at token position θ

$$\rho(\theta) := \Phi\left(\left(\theta - \frac{1}{2}\right)\xi; \theta\right) - \Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}; \theta\right),$$

$$\xi := \frac{\mathrm{tr}(\mathbf{W})}{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}, \quad \eta := \frac{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}{\lambda},$$

$$\Phi(z; \theta) := \frac{1}{2}\mathrm{erf}\left(\frac{z}{\sqrt{2(2\theta^2 + \frac{7}{12})}}\right),$$



- ρ = signal propagation probability: **“How much the token contributes to learning?”**

$$\nabla_{\mathbf{W}_{\mathrm{QK}}} \mathrm{Loss} \approx \sum_{t=1}^{T} \mathrm{signal}_t$$

gradient term depends on t-th token

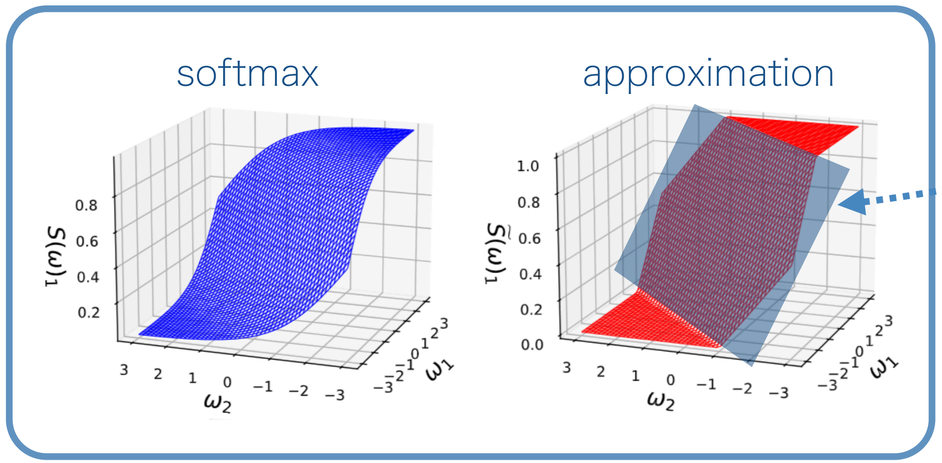
Signal propagation probability | Why it matters?

- ρ = signal propagation probability: **"How much the token contributes to learning?"**

$$\nabla_{\mathbf{W}_{\mathrm{QK}}} \mathrm{Loss} \approx \sum_{t=1}^{T} \mathrm{signal}_t$$

gradient term depends on t-th token

- Signal exists only when softmax input (=query × key) lies in "slope" area



- Specifically,

$$\rho_t := \Pr\left\{0 \le \left\langle \frac{1}{T}\mathbf{e}^t - \frac{1}{T^2}\mathbf{1}, \frac{\mathbf{X}^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T}{\sqrt{d}} \right\rangle + \frac{1}{T} \le 1\right\}$$

It's quite different from attention value itself, but matters when we consider learning dynamics

Our assumption and approximation

$$\rho_t := \Pr\left\{0 \leq \left\langle \frac{1}{T}\mathbf{e}^t - \frac{1}{T^2}\mathbf{1}, \frac{\mathbf{X}^\top \mathbf{W}_{\mathrm{QK}} \mathbf{x}_T}{\sqrt{d}} \right\rangle + \frac{1}{T} \leq 1\right\}$$

$$\langle \boldsymbol{\gamma}^t, \boldsymbol{\omega} \rangle + \gamma_0^t$$

- Assumption: Gaussian random walk $\mathbf{x}_{t+1} \sim \mathcal{N}(\mathbf{x}_t, \boldsymbol{\Sigma})$
- Approximation: Gaussian approximation $\langle \boldsymbol{\gamma}^t, \boldsymbol{\omega} \rangle + \gamma_0^t \sim \mathcal{N}(\mu^t, v^t)$

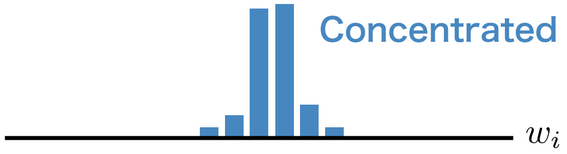
these simplifications are essential to obtain the analytical expression
(we can reduce the derivation to Gaussian moment calculation)

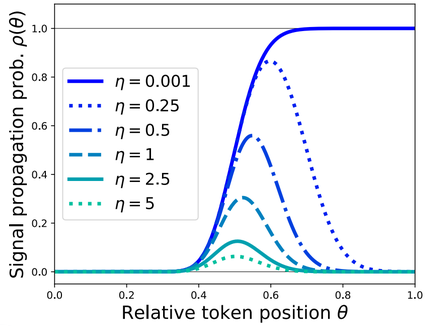
$$\rho(\theta) := \Phi\left(\left(\theta - \frac{1}{2}\right)\xi; \theta\right) - \Phi\left(\left(\theta - \frac{1}{2}\right)\xi - \frac{1}{\eta}; \theta\right),$$

Summary

RQ. When is attention concentrated/uniform?

When $\mathrm{tr}(\mathbf{W}) >> 0$ & $\mathrm{tr}(\mathbf{W}^2) \fallingdotseq 0$





RQ. Are two collapse phenomena reconcilable?

$$\underbrace{\mathrm{Var}[w_i]}_{\text{decrease}} = \frac{1}{d}\underbrace{\mathrm{tr}(\mathbf{W}^2)}_{\text{decrease}} - \frac{1}{d^2}\underbrace{|\mathrm{tr}(\mathbf{W})|^2}_{\text{increase}}$$

avoids entropy collapse

avoids rank collapse