Implicit Geometry of Next-token Prediction:

From Language Sparsity to Model Representations

Christos Thrampoulidis (UBC)

November 5, 2024

CIMI Workshop, Toulouse

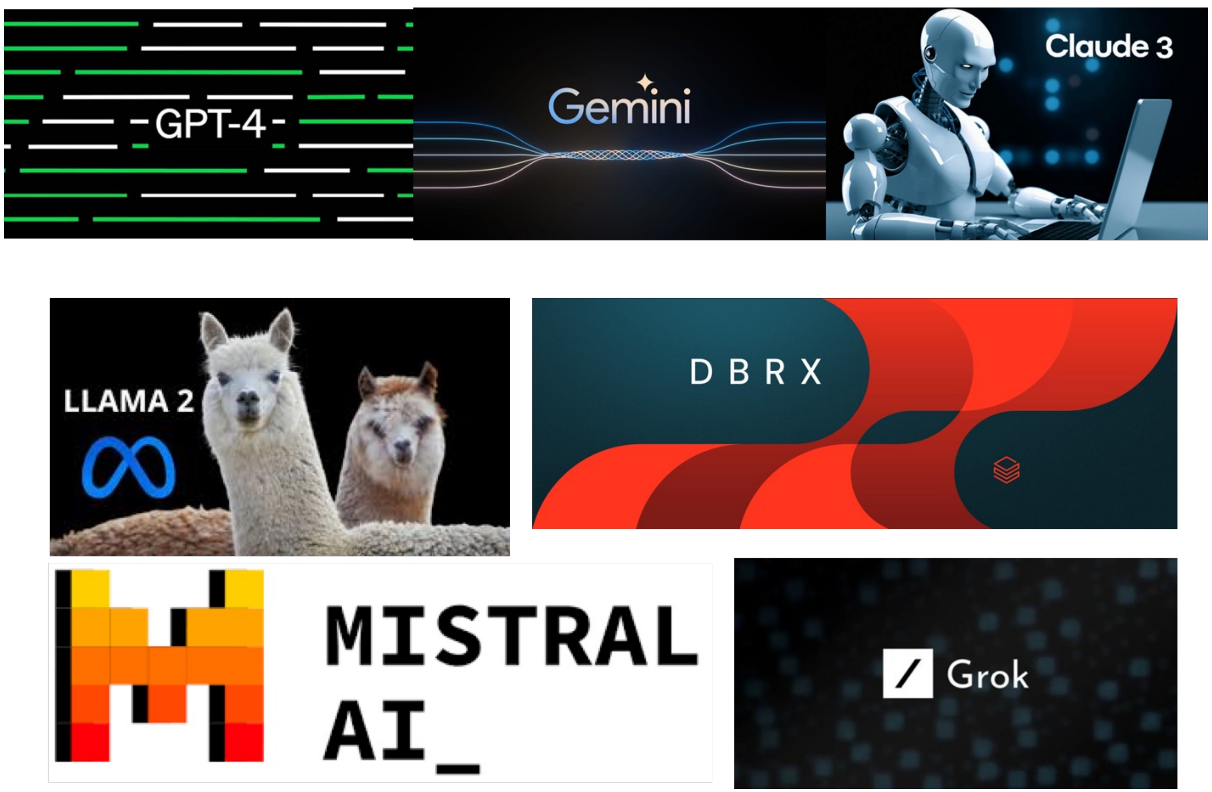
Disclaimer

- Today's talk is not explicitly about *"statistical"* ☹
- But, it might be implicitly, and is *"beyond classical regimes"* ☺

New Sheriff in town

- **DL success** “started” with **image classification** task

- Today’s “hot” topic: **Language modeling**
- LLMs: revolution in natural-language processing and generation



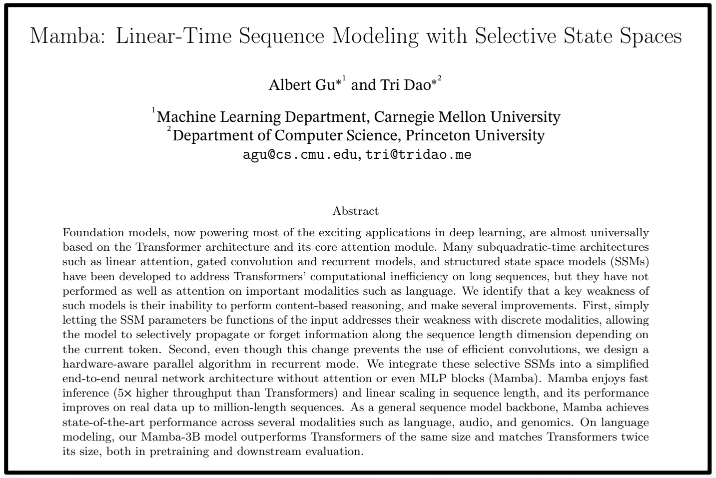
New Sheriff in town

- ❑ **DL success** "started" with **image classification** task

- ❑ Today's "hot" topic: **Language modeling**
- LLMs: revolution in natural-language processing and generation

Key ingredients:

1. Architecture: **Transformer**
   - Parallelizable + trainable to huge scale (~Θ(B) parameters)
   - Self-Attn: leverage long-range context info



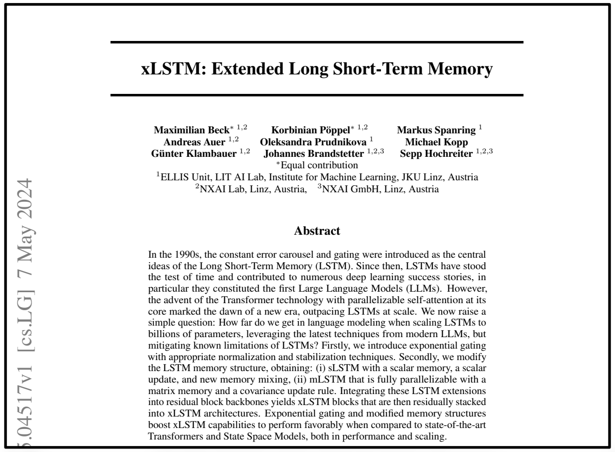
Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu*[1] and Tri Dao*[2]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Computer Science, Princeton University
agu@cs.cmu.edu, tri@tridao.me

Abstract

Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language. We identify that a key weakness of such models is their inability to perform content-based reasoning, and make several improvements. First, simply letting the SSM parameters be functions of the input addresses their weakness with discrete modalities, allowing the model to selectively propagate or forget information along the sequence length dimension depending on the current token. Second, even though this change prevents the use of efficient convolutions, we design a hardware-aware parallel algorithm in recurrent mode. We integrate these selective SSMs into a simplified end-to-end neural network architecture without attention or even MLP blocks (Mamba). Mamba enjoys fast inference (5× higher throughput than Transformers) and linear scaling in sequence length, and its performance improves on real data up to million-length sequences. As a general sequence model backbone, Mamba achieves state-of-the-art performance across several modalities such as language, audio, and genomics. On language modeling, our Mamba-3B model outperforms Transformers of the same size and matches Transformers twice its size, both in pretraining and downstream evaluation.



5.04517v1 [cs.LG] 7 May 2024

xLSTM: Extended Long Short-Term Memory

Maximilian Beck* [1,2] Korbinian Pöppel* [1,2] Markus Spanring [1]
Andreas Auer [1,2] Oleksandra Prudnikova [1] Michael Kopp
Günter Klambauer [1,2] Johannes Brandstetter [1,2,3] Sepp Hochreiter [1,2,3]
*Equal contribution
[1]ELLIS Unit, LIT AI Lab, Institute for Machine Learning, JKU Linz, Austria
[2]NXAI Lab, Linz, Austria, [3]NXAI GmbH, Linz, Austria

Abstract

In the 1990s, the constant error carousel and gating were introduced as the central ideas of the Long Short-Term Memory (LSTM). Since then, LSTMs have stood the test of time and contributed to numerous deep learning success stories, in particular they constituted the first Large Language Models (LLMs). However, the advent of the Transformer technology with parallelizable self-attention at its core marked the dawn of a new era, outpacing LSTMs at scale. We now raise a simple question: How far do we get in language modeling when scaling LSTMs to billions of parameters, leveraging the latest techniques from modern LLMs, but mitigating known limitations of LSTMs? Firstly, we introduce exponential gating with appropriate normalization and stabilization techniques. Secondly, we modify the LSTM memory structure, obtaining: (i) sLSTM with a scalar memory, a scalar update, and new memory mixing, (ii) mLSTM that is fully parallelizable with a matrix memory and a covariance update rule. Integrating these LSTM extensions into residual block backbones yields xLSTM blocks that are then residually stacked into xLSTM architectures. Exponential gating and modified memory structures boost xLSTM capabilities to perform favorably when compared to state-of-the-art Transformers and State Space Models, both in performance and scaling.

New Sheriff in town

- ❑ **DL success** “started” with **image classification** task

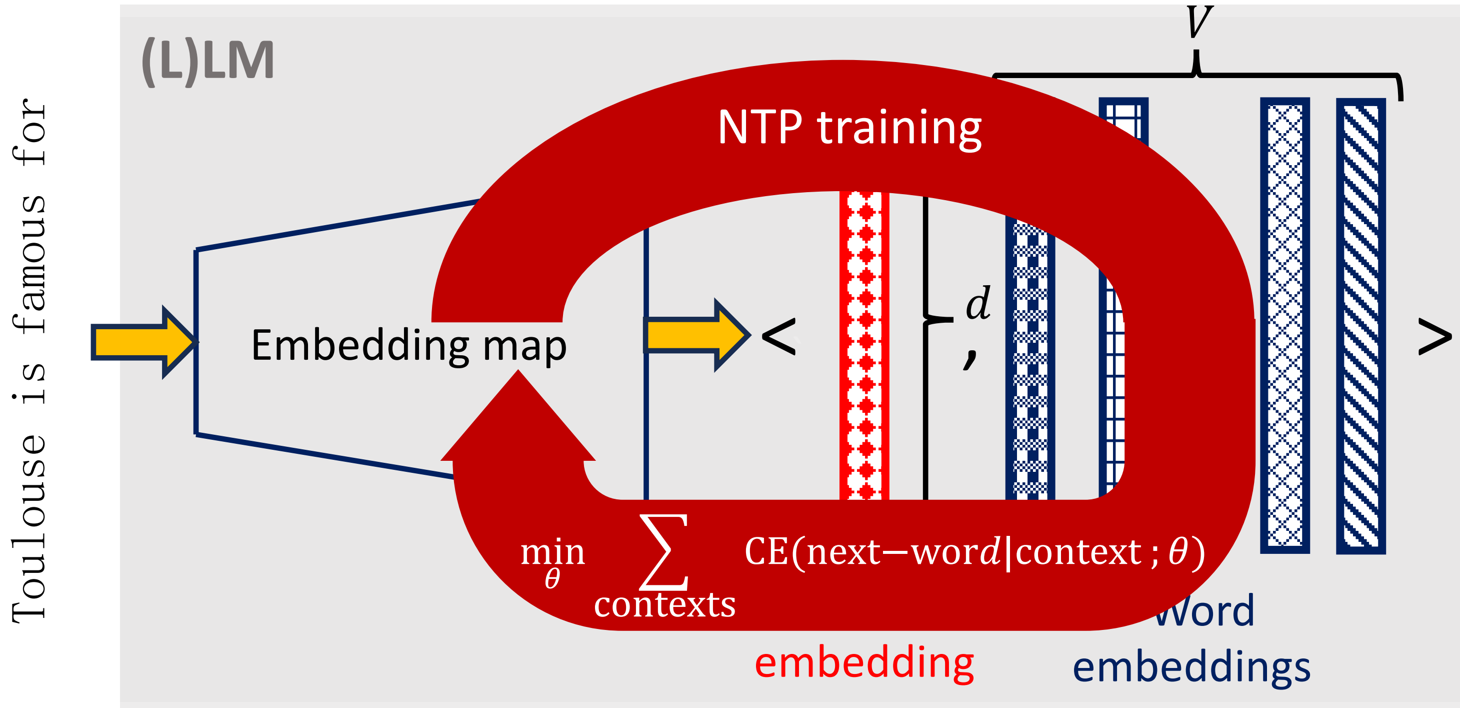
- ❑ Today’s “hot” topic: **Language modeling**
- LLMs: revolution in natural-language processing and generation

Key ingredients:

1. Architecture: **Transformer**
    - Parallelizable + trainable to huge scale (~Θ(B) parameters)
    - Self-Attn: leverage long-range context info

2. Training: Autoregressive **next-token prediction (NTP)**
    - (Pre)Train to sequentially predict **next-token** in a sequence
    - Unsupervised method with supervised flavor

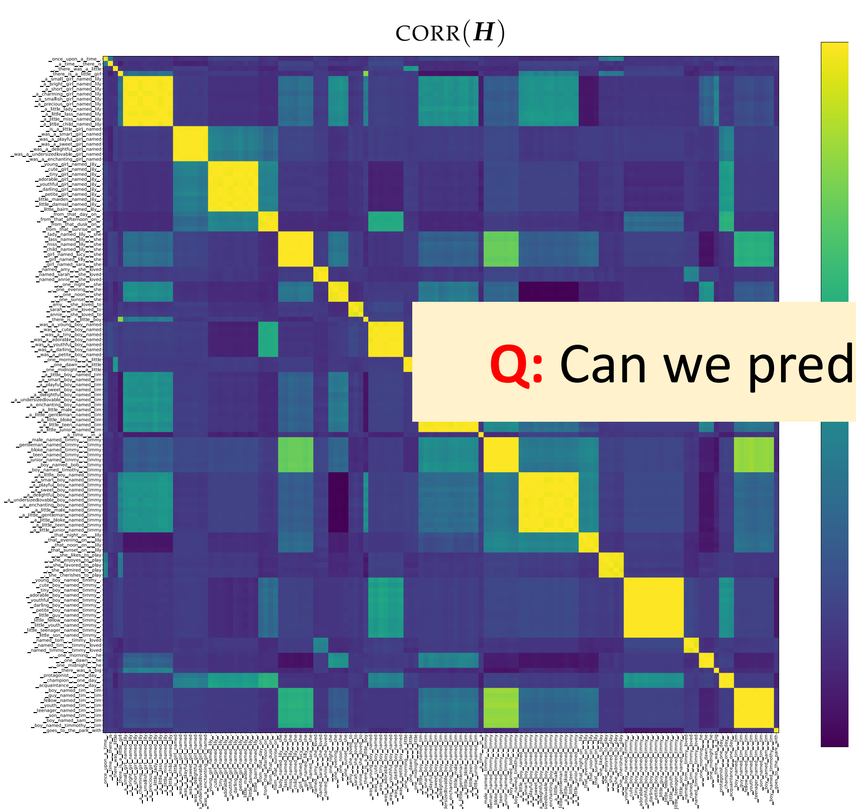
Focus: NTP

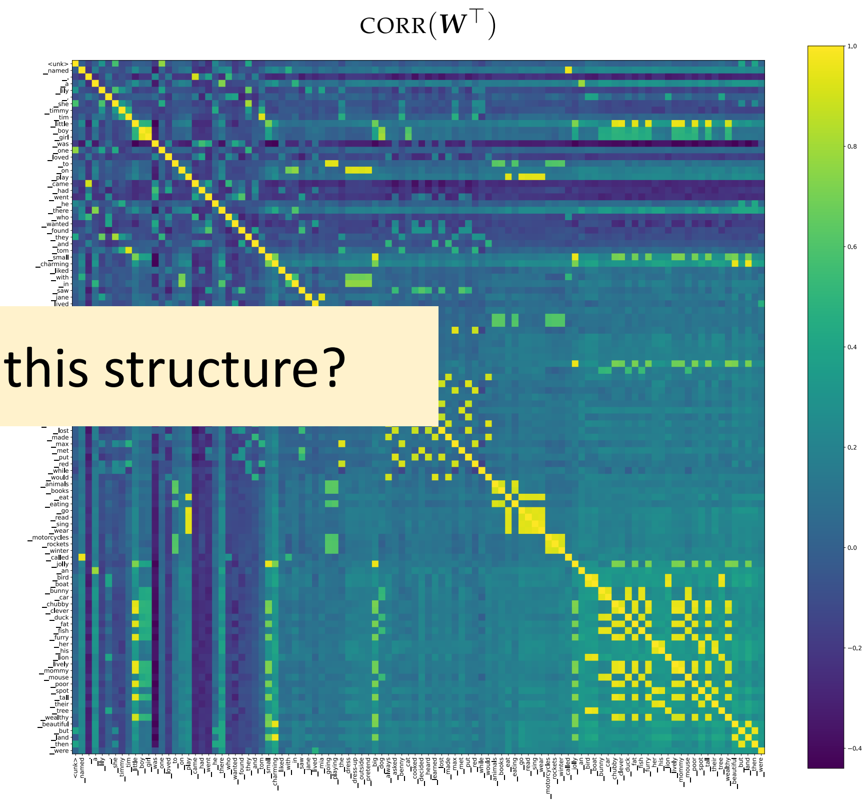


Q: How do the learnt context/word representations encode the statistics of the data they are trained on?

Implicit Geometry

Q: How distances and angles of the model representations correlate with linguistic patterns at the end of training?





Q: Can we predict this structure?

Context embeddings

Word embeddings

Why?

- Interpretability: **transparency** on the inner workings of LLMs
- Identify/mitigate **sources of errors/biases**
- **Algorithm improvements** upon vanilla NTP paradigm and optimizers
- Enhance our **grasp of language itself**

Challenges & key message

- Representations are outputs of training **complicated models** (architecture, size) over **complex datasets** (source, size, tokenization) with **varying choices of optimization hyperparameters** (learning rate, weight decay, number of iterations)

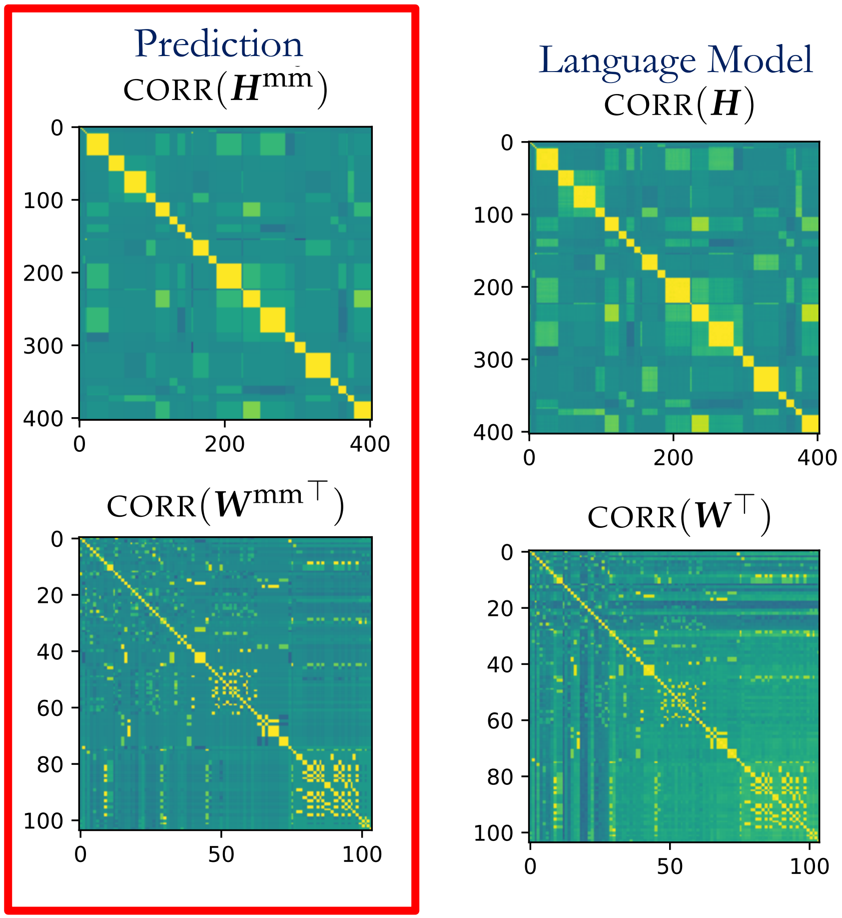
- Message:

In a doubly-asymptotic regime of
Large model + Long training

context/word embeddings are *matrix factorization* of a *logit matrix*

$$L_{\text{sparse}} + \infty \cdot L_{\text{low-rank}}$$

with components determined solely by **patterns of text training data**



Key Ingredients

1. Correctly **framing** the **next token prediction** training task

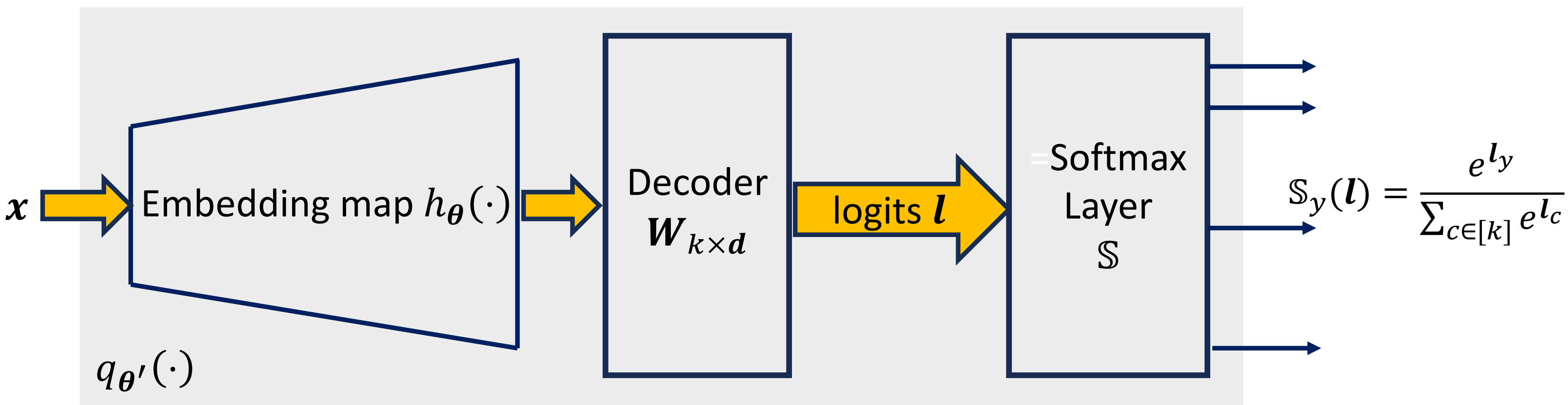
2. Leveraging the technical framework of **implicit optimization bias**

3. Assuming large model with **unconstrained features**

Traditional Setting

- One-hot multiclass classification
  - **Training Data:** $\mathcal{T}_n \triangleq (\boldsymbol{x}_i, y_i)_{i\in[n]}, \quad y_i \in [k] \triangleq \{1,2,\ldots,k\}$
  - **Training Loss:** $\min_{\boldsymbol{\theta}'} \frac{1}{n}\sum_{i\in[n]} \mathcal{L}(y_i, q_{\boldsymbol{\theta}'}(\boldsymbol{x}_i))$

$$\min_{\boldsymbol{\theta}'=(\mathbf{W},\boldsymbol{\theta})} \left\{ \mathrm{CE}(\mathbf{W},\boldsymbol{\theta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} -\log\left(\mathbb{S}_{y_i}\left(\boldsymbol{W}\, h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)\right)\right\}$$



Traditional setting

- One-hot multiclass classification
  - **Training Data:** $\mathcal{T}_n \triangleq (\boldsymbol{x}_i, y_i)_{i\in[n]}$, $y_i \in [k] \triangleq \{1,2,\dots,k\}$
  - **Training Loss:** $\min_{\boldsymbol{\theta}'} \frac{1}{n}\sum_{i\in[n]} \mathcal{L}(y_i, q_{\boldsymbol{\theta}'}(\boldsymbol{x}_i))$

$$\min_{\boldsymbol{\theta}'=(\mathbf{W},\boldsymbol{\theta})} \left\{ \mathrm{CE}(\mathbf{W},\boldsymbol{\theta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} -\log\left(\mathbb{S}_{y_i}\left(\boldsymbol{W}\, h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)\right) \right\}$$

- Trained with first-order **gradient-based optimizers,** e.g. (S)GD
- **Overparameterization** $\Longrightarrow$ Interpolation (Separability)

$$\Longrightarrow \inf_{\boldsymbol{\theta}'} \mathrm{CE}(\boldsymbol{\theta}') = 0$$

Questions:

1. Does GD lead to zero loss?
2. Among the many possible solutions, which one it “prefers”?

"Textbook" result

- **Linear model**
  - Fixed embeddings $\boldsymbol{h}_i \triangleq h_\theta(\boldsymbol{x}_i)$.
  - Trainable decoder $\boldsymbol{W} \in \mathbb{R}^{k \times d}$

- Linearly-**separable data** (e.g. $d > n$)

$$\exists \boldsymbol{W} : \ \boldsymbol{w}_{y_i}^T \boldsymbol{h}_i - \boldsymbol{w}_c^T \boldsymbol{h}_i > 0, \ \forall i \in [n], c \neq y_i \in [k]$$

$$\Longleftrightarrow \quad (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^T \boldsymbol{W} \boldsymbol{h}_i = \langle \boldsymbol{W}, (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c) \boldsymbol{h}_i^T \rangle > 0$$

$$\boldsymbol{e}_\ell = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 1 \\ 0 \\ \cdots \\ 0 \end{bmatrix} \longleftarrow \ \ell^{\text{th}} \text{ entry}$$

"Textbook" result [Soudry et al.'18]

Thm. Assume separability. Run GD with $\eta \leq 2/L$.
Then, $\lim_{k\to\infty} \mathrm{CE}(\boldsymbol{W}_k) = 0$.
Moreover, $\lim_{k\to\infty} ||\boldsymbol{W}_k|| = \infty$ and

$$\lim_{k\to\infty} \left\langle \frac{\boldsymbol{W}_k}{\|\boldsymbol{W}_k\|}, \frac{\boldsymbol{W}^{\mathbf{mm}}}{\|\boldsymbol{W}^{\mathbf{mm}}\|} \right\rangle = 1$$

Defn. (max-margin) Let $\boldsymbol{W}^{\mathbf{mm}}$ be the max-margin classifier

$$\boldsymbol{W}^{\mathbf{mm}} = \mathrm{argmin}_{\boldsymbol{W}} ||\boldsymbol{W}||$$
$$\text{subj. to } \left\langle \boldsymbol{W}, \left(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c\right)\boldsymbol{h}_i^T \right\rangle \geq 1, \ \forall c \neq y_i \in [k], \ j \in [m]$$

Why nice?

- Insights on:

  ✓ what GD learns (impact of architectures/initializations)
  [RZH03,SHN+18,JT18,GLSS18,JDST20,LL20,JT20] ++++

  ✓ role of optimizers (e.g. adaptive / mirror-descent)
  [NLG+19,ALH21,PPVF21,SA**T**A22,AF22] ++++

  ✓ **stepping stone to generalization (benign overfitting)**
  [BLLT19,MRSY19,DK**T**19,MVS19,DL20,DL21,KZSS21,W**T**21,TP**T**21,CCBG22] ++++

  ✓ **loss design and hyperparameter tuning (imbalanced data)**
  [KPO**T**21,CLB21,BKV**T**22]

  .....

Stepping stone to generalization

? How well does the solution found by GD generalize?

↕ **Implicit bias of GD**

? How well does SVM solution generalize?

$$\mathbb{P}_{(\boldsymbol{h},y)\sim D}\left(\min_{c\neq y}\left(\boldsymbol{e}_y-\boldsymbol{e}_c\right)^{\boldsymbol{T}}\boldsymbol{W}^{\mathbf{mm}}\boldsymbol{h}>0\right)$$

Stepping stone to generalization

? How well does SVM solution generalize?

- Catch: Overparameterization (d>n) makes "classical" margin-based bounds vacuous

- Rescue: modern* tools from HD-stats/RMT and universality
  - Approximate Message Passing (AMP) [DMM09,MM12++]
  - Gordon's comparison inequalities [Gor88,RV08,Sto09,CRPW'12++] [Sto13+,**T**OH15 ++]

[…]

* Developed for compressed-sensing

<?> Can we push this storyline and (eventually) its implications to “new” setting of NTP in LMs

Next-token Prediction (NTP)

❑ Training data

- Vocabulary $\mathcal{V} \triangleq [V]$ of tokens/words
- (many many) n sequences $(z_{i1}, z_{i2}, \ldots, z_{iT})_{i\in[n]}$, $z_{it} \in \mathcal{V}$

❑ Training loss

- $\min_{(\boldsymbol{W},\boldsymbol{\theta})} \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathcal{L}\left(\underbrace{z_{it}}_{\text{next-word}}, \mathbb{S}\left(\boldsymbol{W}\, h_{\boldsymbol{\theta}}(\underbrace{\boldsymbol{z}_{\boldsymbol{i},<\boldsymbol{t}}}_{\text{Context } (z_{i,1},\ldots,z_{i,t-1})})\right)\right)$

- For simplicity: focus on **last-token**

Denote $\left(\boldsymbol{z}_{\boldsymbol{i},<\boldsymbol{T}}, z_{i,T}\right) \triangleq (\boldsymbol{x}_i, z_i)$

$$\min_{(\mathbf{W},\boldsymbol{\theta})} \left\{ \mathrm{CE}(\mathbf{W}, \boldsymbol{\theta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} -\log\left(\mathbb{S}_{z_i}(\boldsymbol{W}\, h_\theta(\boldsymbol{x}_{\boldsymbol{i}}))\right) \right\}$$

NTP vs one-hot classification

- ❑ **<u>Ansatz #1:</u>**

a. Contexts repeat

b. **Multiple** possible next-tokens with **varying frequencies** after each distinct context. [Shannon48]

Example: `Vancouver is famous for its ____________`

restaurants	0.05
mountains	0.1
rain	0.4
UBC	0.01
...	
culture	0
sun	0
affordability	0

- ❑ **<u>Ansatz #2:</u> [Sparsity]**

Not all vocabulary tokens are possible next-tokens per distinct context

NTP training is sparse soft-label classification

- Data
  - $m < n$ **distinct** contexts x_j each with frequency $\hat{\pi}_j$
  - Each associated with sparse probabilistic label $\hat{\boldsymbol{p}}_{\boldsymbol{j}} \in \Delta^V$
  - support set $\mathcal{S}_j$ of $\hat{\boldsymbol{p}}_{\boldsymbol{j}}$: $|\mathcal{S}_j| < V$

- Loss wrt distinct contexts

$$\min_{(\mathbf{W},\boldsymbol{\theta})} \left\{ \mathrm{CE}(\mathbf{W},\boldsymbol{\theta}) \triangleq -\sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\mathbb{S}_z\left(\boldsymbol{W}\, h_\theta(\boldsymbol{x}_{\boldsymbol{j}})\right)\right) \right\}$$

<u>Questions:</u>

1. Does GD lead to the loss lower bound?
2. Among the many possible solutions, which one it "prefers"?

Entropy Lower-bound

- Empirical (T-gram) entropy:

$$\mathcal{H} \triangleq \mathbb{E}_{(\boldsymbol{x},z)\sim\mathcal{T}_n}[-\log(\hat{p}(z|\boldsymbol{x}))] = -\sum_{j\in[m]}\sum_{z\in\mathcal{S}_j}\hat{\pi}_j\,\hat{p}_{j,z}\log\big(\hat{p}_{j,z}\big)$$

- $\mathrm{CE}(\boldsymbol{\theta}') = \mathcal{H} + \mathrm{KL}(\hat{p}||\mathrm{q}_{\boldsymbol{\theta}'}) \Rightarrow \mathrm{CE}(\boldsymbol{\theta}') \geq \mathcal{H}$

When is the lower bound reached?

- Consider linear model (fixed embeddings)

Lemma. The NTP loss reaches its lower bound if and only if the following two conditions hold.

Defn. (NTP$_{\mathcal{H}}$—compatibility) There exists matrix $\boldsymbol{W}^{\boldsymbol{p}}$ such that for all $j \in [m]$:

$$\langle \boldsymbol{W}^{\boldsymbol{p}}, (\boldsymbol{e}_z - \boldsymbol{e}_{z'})\boldsymbol{h}_{\boldsymbol{j}}^T \rangle = \log\left(\frac{\hat{p}_{j,z}}{\hat{p}_{j,z'}}\right) \quad \forall z \neq z' \in \mathcal{S}_j$$

Defn. (NTP—separability) There exists matrix $\boldsymbol{W}^{\boldsymbol{d}}$ such that for all $j \in [m]$:

$$\langle \boldsymbol{W}^{\boldsymbol{d}}, (\boldsymbol{e}_z - \boldsymbol{e}_{z'})\boldsymbol{h}_{\boldsymbol{j}}^T \rangle = 0 \quad \forall z \neq z' \in \mathcal{S}_j$$

$$\langle \boldsymbol{W}^{\boldsymbol{d}}, (\boldsymbol{e}_z - \boldsymbol{e}_v)\boldsymbol{h}_{\boldsymbol{j}}^T \rangle \geq 1 \quad \forall z \in \mathcal{S}_j, v \notin \mathcal{S}_j$$

NTP compatibility & separability

- Need $\mathrm{KL}(\hat{p}||q_{\boldsymbol{\theta}'}) = 0$

⇕

1. $\mathbb{S}_z(\boldsymbol{W}\boldsymbol{h}_{\boldsymbol{j}}) = \hat{p}_{j,z}, \ \forall z \in \mathcal{S}_j$ ⟺ **NTP$_{\mathcal{H}}$—compatibility**: $\langle \boldsymbol{W}^{\boldsymbol{p}}, (\boldsymbol{e}_z - \boldsymbol{e}_{z'})\boldsymbol{h}_{\boldsymbol{j}}^T \rangle = \log\left(\frac{\hat{p}_{j,z}}{\hat{p}_{j,z'}}\right)$

2. $\mathbb{S}_v(\boldsymbol{W}\boldsymbol{h}_{\boldsymbol{j}}) = 0, \ \forall v \notin \mathcal{S}_j$ ⟺ **NTP—separability**: $\langle \boldsymbol{W}^{\boldsymbol{d}}, (\boldsymbol{e}_z - \boldsymbol{e}_{z'})\boldsymbol{h}_{\boldsymbol{j}}^T \rangle = 0$, $\langle \boldsymbol{W}^{\boldsymbol{d}}, (\boldsymbol{e}_z - \boldsymbol{e}_v)\boldsymbol{h}_{\boldsymbol{j}}^T \rangle \geq 1$

$$\mathbb{S}_z(\boldsymbol{a}) = \frac{\exp(\boldsymbol{e}_z^T\boldsymbol{a})}{\sum_{v\in\mathcal{V}}\exp(\boldsymbol{e}_v^T\boldsymbol{a})} = \frac{1}{\sum_{v\in\mathcal{V}}\exp(-(\boldsymbol{e}_z - \boldsymbol{e}_v)^T\boldsymbol{a})}$$

Lemma (Overparameterization). If $d > m$ and generic embeddings, then the two conditions hold.

Implicit bias

$$\mathcal{F} = \text{span}\{(\boldsymbol{e}_z - \boldsymbol{e}_{z'})\boldsymbol{h}_j^T \ : z \neq z' \in \mathcal{S}_j, j \in [m] \ \} \subseteq \mathbb{R}^{V\times d}$$

Thm. Assume NTP compatibility and separability. Run GD with $\eta \leq 2/L$.
Then, (i) $\lim_{k\to\infty} \text{CE}(\boldsymbol{W_k}) = \mathcal{H}$.

(ii) $\lim_{k\to\infty} \mathbb{P}_{\mathcal{F}}(\boldsymbol{W_k}) = \boldsymbol{W}^*$

(iii) $\lim_{k\to\infty} ||\mathbb{P}_{\perp}(\boldsymbol{W_k})|| = \infty$ with $\lim_{k\to\infty} \left\langle \frac{\boldsymbol{W_k}}{\|\boldsymbol{W_k}\|}, \frac{\boldsymbol{W}^{\mathbf{mm}}}{\|\boldsymbol{W}^{\mathbf{mm}}\|} \right\rangle = 1$

Defn. (subspace component) $\boldsymbol{W}^* \in \mathcal{F}$ is the unique solution of:

$$\left\langle \boldsymbol{W}^*, (\boldsymbol{e}_z - \boldsymbol{e}_{z'})\boldsymbol{h}_j^T \right\rangle = \log\left(\frac{\hat{p}_{j,z}}{\hat{p}_{j,z'}}\right) \quad \forall z \neq z' \in \mathcal{S}_j, j \in [m]$$

Defn. (orthogonal component) $\boldsymbol{W}^{\mathbf{mm}} \in \mathcal{F}^{\perp}$ is the unique solution of:

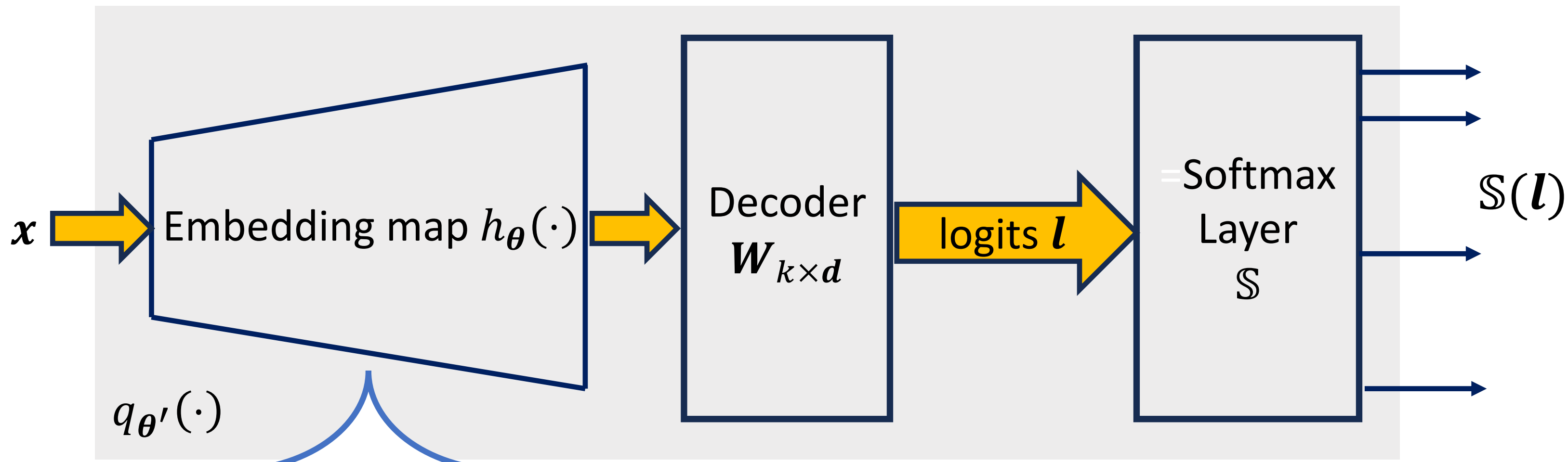
$$\boldsymbol{W}^{\mathbf{mm}} = \text{argmin}_{\boldsymbol{W}} \ ||\boldsymbol{W}|| \qquad \textbf{(NTP-SVM)}$$

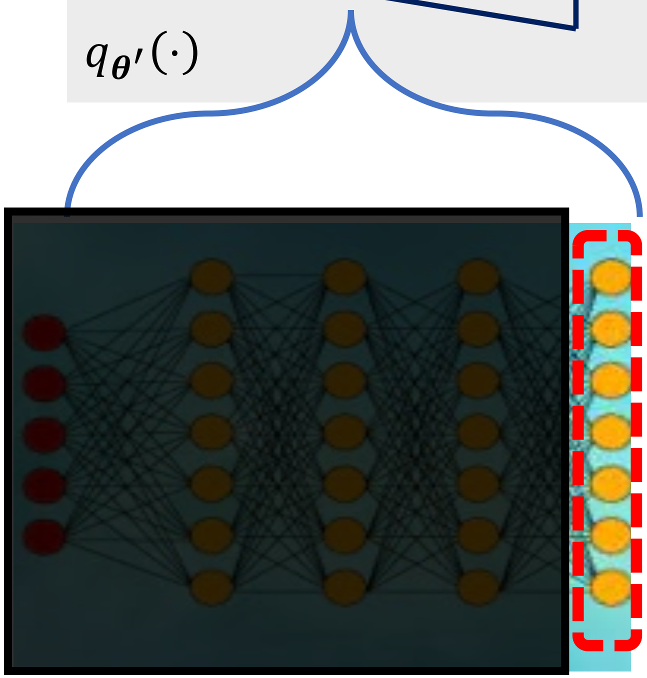
$$\text{subj. to } \left\langle \boldsymbol{W}, (\boldsymbol{e}_z - \boldsymbol{e}_{z'})\boldsymbol{h}_j^T \right\rangle = 0, \ \forall z \neq z' \in \mathcal{S}_j$$

$$\left\langle \boldsymbol{W}, (\boldsymbol{e}_z - \boldsymbol{e}_v)\boldsymbol{h}_j^T \right\rangle \geq 1, \ \forall z \in \mathcal{S}_j, v \notin \mathcal{S}_j, j \in [m]$$

<?> How to go beyond linear models?

Unconstrained features





Treat the rest of the layers as a powerful black-box that can generate **unconstrained** context embeddings $\boldsymbol{h}$

[YDSC17, MPP20, LS20]

NTP-UFM

- **Unconstrained-features model (UFM)**:

$$\min_{(\mathbf{W},\boldsymbol{H})} \left\{ \mathrm{CE}(\mathbf{W},\boldsymbol{H}) \triangleq -\sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\Big(\mathbb{S}_z(\boldsymbol{W}\,\boldsymbol{h_j})\Big) \right\}$$

- $\mathbf{W} \in \mathbb{R}^{V\times d}$: **word** embeddings
- $\boldsymbol{H} = [\boldsymbol{h_1}, \dots, \boldsymbol{h_m}] \in \mathbb{R}^{d\times m}$: **context** embeddings
- $\boldsymbol{P} = [\{\hat{p}_{j,z}\}] \in \mathbb{R}^{V\times m}$: **sparse** next-token probability matrix

NTP-UFM

- **Unconstrained-features model (UFM):**

What is the geometry of context/word embeddings in terms of the language statistics as encoded in the sparse conditional probability mtx $\boldsymbol{P}$?

$$\min_{(\mathbf{W},\boldsymbol{H})} \left\{ \mathrm{CE}(\mathbf{W},\boldsymbol{H}) \triangleq -\sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\mathbb{S}_z(\mathbf{W}\boldsymbol{h}_j)\right) \right\}$$

- $\mathbf{W} \in \mathbb{R}^{V\times d}$: **word** embeddings
- $\boldsymbol{H} \in \mathbb{R}^{d\times m}$: **context** embeddings

If I were to optimize the log-bilinear NTP-UFM model, where does GD converge?

Input: sparse conditional probabilities mtx $\boldsymbol{P} \in \mathbb{R}^{V\times m}$ and corresponding support-set mtx $\boldsymbol{S} \in \{0,1\}^{V\times m}$

Output: The implicit geometry of word/context embeddings i.e., angles between ctx-ctx / word-word / ctx-word vectors in $\mathbb{R}^d$

Proxy: Regularization path

$$\min_{(\mathbf{W},\boldsymbol{H})} \left\{ - \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\mathbb{S}_z\left(\boldsymbol{W}\,\boldsymbol{h_j}\right)\right) + \lambda(||\boldsymbol{W}||^2 + ||\boldsymbol{H}||^2) \right\}$$

- ❑ **Goal:** Compute the solution as $\lambda \to 0$

- ❑ A proxy for GD-path ("$\lambda \to 0$" ≡ "$k \to \infty$")
- Formal equivalence in linear settings [Ji et al. 20] [Rosset et al. '03]

Logit-space relaxation

$$\min_{(\mathbf{W},\boldsymbol{H})} \left\{ - \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\mathbb{S}_z(\boldsymbol{W}\,\boldsymbol{h_j})\right) + \lambda\,(||\boldsymbol{W}||^2 + ||\boldsymbol{H}||^2) \right\}$$

- ❑ $\boldsymbol{L} = \boldsymbol{W}\boldsymbol{H} \in \mathbb{R}^{V\times d}$: **logit** matrix

Lemma. The following relaxation to the $\mathbb{R}^{\boldsymbol{V}\times\boldsymbol{m}}$ **logit-space** is tight:

$$\min_{\boldsymbol{L}\,:\,\mathrm{rank}(\boldsymbol{L})\leq d} \left\{ - \sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\mathbb{S}_z(\boldsymbol{l}_j)\right) + \lambda\,||\boldsymbol{L}||_* \right\}$$

If $\boldsymbol{L}_\lambda$ has SVD $\boldsymbol{L}_\lambda = \boldsymbol{U}\boldsymbol{\Sigma}\mathbf{V}^T$, then for partially orthogonal matrix $\mathbf{R} \in \mathbb{R}^{r\times d}$

$$\boldsymbol{W}_\lambda = \boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\,\mathbf{R} \quad \text{and} \quad \boldsymbol{H}_\lambda = \boldsymbol{R}^T\sqrt{\boldsymbol{\Sigma}}\,\mathbf{V}$$

Large embedding space

$$\min_{\boldsymbol{L}\in\mathbb{R}^{V\times d}\sout{: \mathrm{rank}(\boldsymbol{L})\le d}} \left\{ -\sum_{j\in[m]} \hat{\pi}_j \sum_{z\in\mathcal{S}_j} \hat{p}_{j,z} \log\left(\mathbb{S}_z(\boldsymbol{l}_j)\right) + \lambda\, ||\boldsymbol{L}||_* \right\}$$

❑ **Assumption:** $d \geq V$

- Under this we can characterize regularization path
- Limiting but nontrivial:
  1. "# of contexts m" $\gg$ "dimension d"
  2. how geometry depends on language statistics?

Regularization-path of NTP-UFM

$$\mathcal{F} = \text{span}\left\{(\boldsymbol{e}_z - \boldsymbol{e}_{z'})\,\tilde{\boldsymbol{e}}_j^T \;: z \neq z' \in \mathcal{S}_j, j \in [m]\;\right\} \subseteq \mathbb{R}^{V\times m}$$

Thm. Assume $d \geq V - 1$.

Then, (i) $\lim_{\lambda\to 0} \mathrm{CE}(\boldsymbol{L}_\lambda) = \mathcal{H}$.

(ii) $\lim_{\lambda\to 0} \mathbb{P}_{\mathcal{F}}(\boldsymbol{L}_\lambda) = \boldsymbol{L}^*$

(iii) $\lim_{\lambda\to 0} ||\mathbb{P}_\perp(\boldsymbol{L}_\lambda)|| = \infty$ with $\lim_{\lambda\to 0} \left\langle \frac{\boldsymbol{L}_\lambda}{\|\boldsymbol{L}_\lambda\|}, \frac{\boldsymbol{L}^{\mathbf{mm}}}{\|\boldsymbol{L}^{\mathbf{mm}}\|} \right\rangle = 1$

Defn. $\boldsymbol{L}^* \in \mathcal{F}$ is the unique solution of:

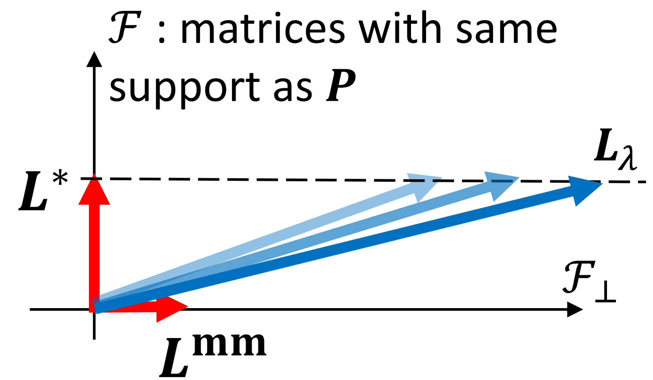
$$\boldsymbol{L}_{z,j} - \boldsymbol{L}_{z',j} = \log\left(\frac{\hat{p}_{j,z}}{\hat{p}_{j,z'}}\right) \quad \forall z \neq z' \in \mathcal{S}_j, j \in [m]$$

Defn. $\boldsymbol{L}^{\mathbf{mm}} \in \mathcal{F}^\perp$ is a solution of:

$$\begin{aligned} \min_{\boldsymbol{L}} \quad & ||\boldsymbol{L}||_* \\ \text{subj. to} \quad & \boldsymbol{L}_{z,j} - \boldsymbol{L}_{z',j} = 0, \;\forall z \neq z' \in \mathcal{S}_j \\ & \boldsymbol{L}_{z,j} - \boldsymbol{L}_{v,j} \geq 1, \;\; \forall z \in \mathcal{S}_j, v \notin \mathcal{S}_j, j \in [m] \end{aligned} \qquad \textbf{(NTP-SVM)}$$

Regularization-path of NTP-UFM

As $\lambda \to 0$, for some $\rho(\lambda) \to \infty$:
$\boldsymbol{L}_\lambda \approx \boldsymbol{L}_{\text{sparse}} + \rho(\lambda) \cdot \boldsymbol{L}_{\text{low-rank}}$



- ❑ $\boldsymbol{L}^* \stackrel{\text{def}}{=} \boldsymbol{L}_{\text{sparse}}$ inherits **sparsity** of $\boldsymbol{P}$ and depends on frequencies of in-support tokens

- ❑ $\boldsymbol{L}^{\mathbf{mm}} \stackrel{\text{def}}{=} \boldsymbol{L}_{\text{low-rank}}$ minimizes nuclear-norm promoting **low-rankness** and only depends on sparsity pattern $\boldsymbol{S}$ (not on frequencies)

Dominant as $\boldsymbol{\lambda} \to \mathbf{0}$

NTP max-margin logits

- In some special cases, can compute $\boldsymbol{L}^{\mathbf{mm}}$ in closed form

Prop. Suppose $\boldsymbol{S}$ contains all $m = \binom{V}{k}$ support sets of size k.
Then, (i) $\boldsymbol{L}^{\mathbf{mm}} = (\mathbf{I}_V - 11^T)\boldsymbol{S} \stackrel{\text{def}}{=} \overline{\boldsymbol{S}}$.
(ii) Word embeddings form equiangular tight frame
(iii) Context embeddings are equinorm and $\boldsymbol{h}_j$ is colinear to $\sum_{z \in S_j} w_z$

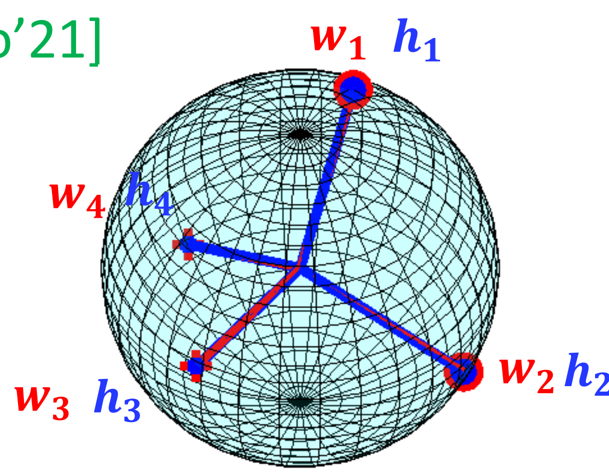
$$\boldsymbol{S} = \underbrace{\begin{bmatrix} 1 & 1 & & & & 0 \\ 1 & 0 & \dots & 1 & & 0 \\ 0 & 1 & & 1 & & 0 \\ \vdots & \vdots & & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 0 & & 1 \\ 0 & 0 & & 0 & & 1 \end{bmatrix}}_{m = \binom{V}{2}} \left.\vphantom{\begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}}\right\} V$$

NTP max-margin logits

- In some special cases, can compute $\boldsymbol{L}^{\mathbf{mm}}$ in closed form

Prop. Suppose $\boldsymbol{S}$ contains all $m = \binom{V}{k}$ support sets of size k.
Then, (i) $\boldsymbol{L}^{\mathbf{mm}} = (\mathbf{I}_V - 11^T)\boldsymbol{S} \stackrel{\text{def}}{=} \overline{\boldsymbol{S}}$.
(ii) Word embeddings form equiangular tight frame
(iii) Context embeddings are equinorm and $\boldsymbol{h}_j$ is colinear to $\sum_{z \in S_j} w_z$

- Special case $k = 1$, recovers the **Neural Collapse** geometry by [Papyan,Han,Donoho'21]



$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & & & 0 \\ 1 & 0 & \dots & 1 & 0 \\ 0 & 1 & & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ & & & \dots & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & & 0 & 1 \end{bmatrix} \} V$$

$$m = \binom{V}{2}$$

NTP max-margin logits

- ❑ In some special cases, can compute $\boldsymbol{L}^{\mathbf{mm}}$ in closed form

Prop. Suppose $\boldsymbol{S}$ contains all $m = \binom{V}{k}$ support sets of size k.
Then, (i) $\boldsymbol{L}^{\mathbf{mm}} = (\mathbf{I}_V - 11^T)\boldsymbol{S} \stackrel{\text{def}}{=} \overline{\boldsymbol{S}}$.
(ii) Word embeddings form equiangular tight frame
(iii) Context embeddings are equinorm and $\boldsymbol{h}_j$ is colinear to $\sum_{z \in S_j} w_z$

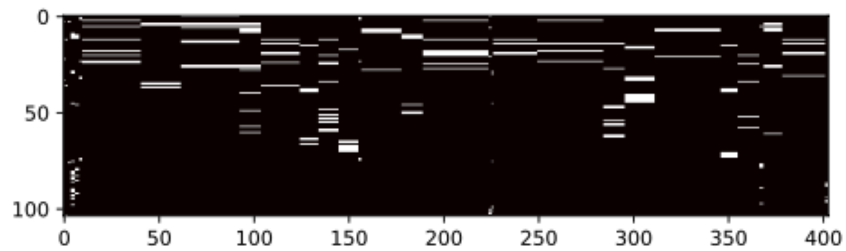
- ❑ In general, need to solve SDP.
- ❑ But, experimentally $\overline{\boldsymbol{S}}$ is a good “proxy”

Experiment

Data:

- Synthetic extracted from TinyStories*
- $n = 3050$ contexts of length $T = 5$
- $m = 400$ *distinct* contexts
- $V = 104$

S



*

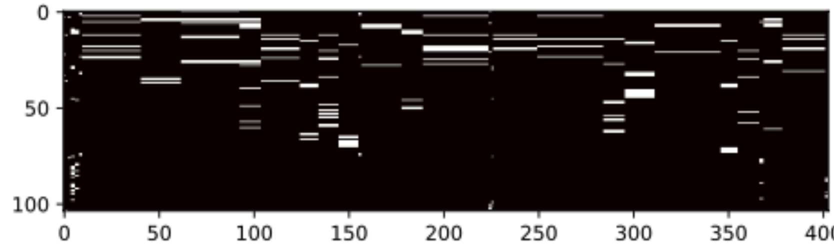
"a little girl named lily" .... {"and", "was", "found", "had", "went", "."}

"there was a little boy" .... {"named", "called", "and", ".", "who", "had", "with"}

Experiment

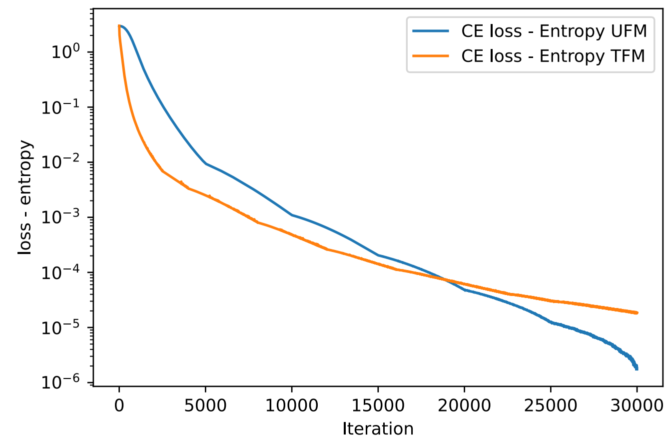
Data:
- Synthetic extracted from TinyStories
- $n = 3050$ contexts of length $T = 5$
- $m = 400$ *distinct* contexts
- $V = 104$

S



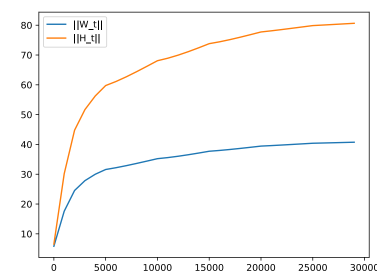
Language Model:
- 4-layer **TF**
- $d = 128$
- Trained with Adam-W for 30k epochs

$\mathcal{L}(\boldsymbol{W}_k \boldsymbol{H}_k) - \mathcal{H}$

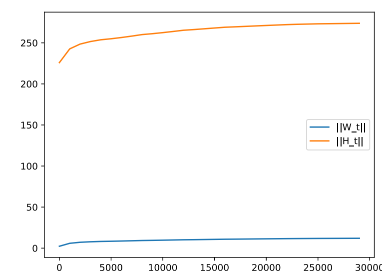


Analysis Model:
- Unconstrained features model (**UFM**)
- $d = 128$
- Trained with Adam-W for 30k epochs

UFM on Tinystory Context

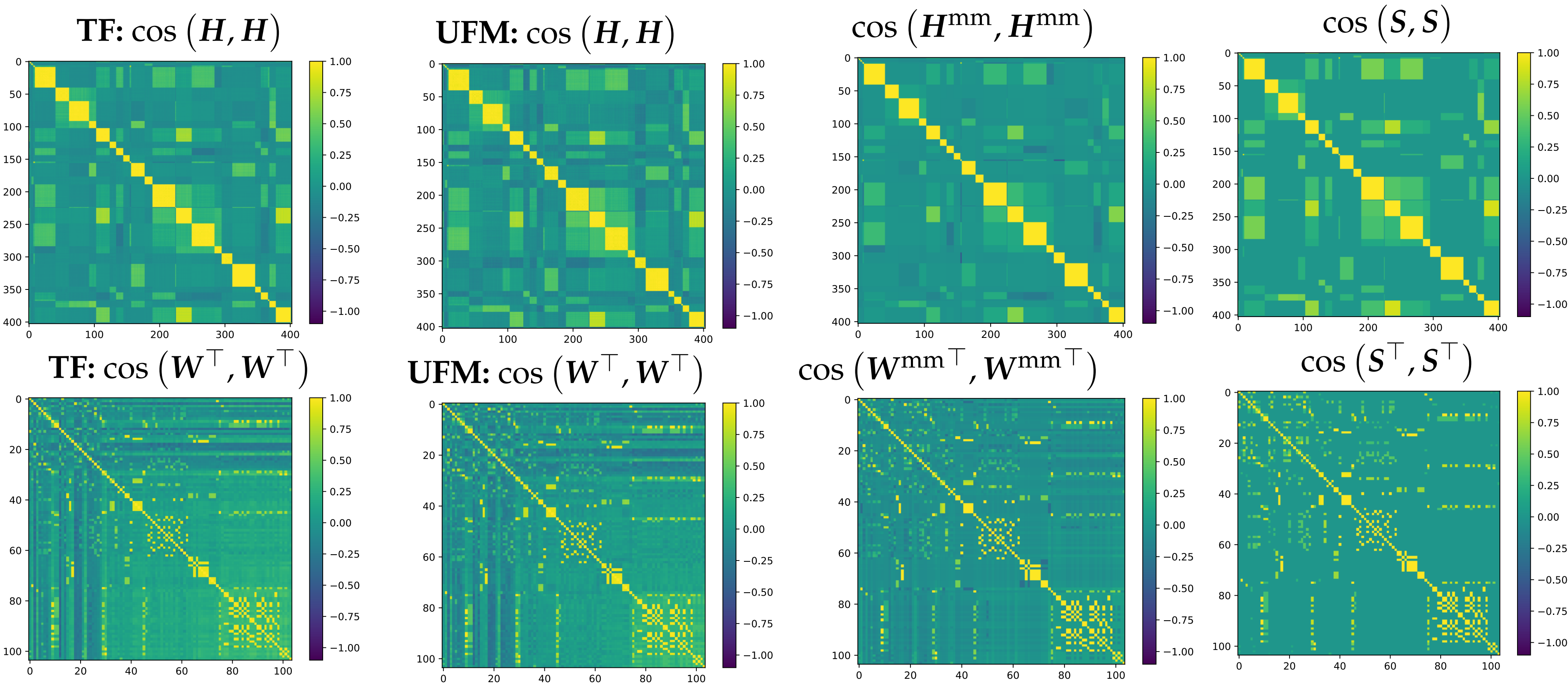


Transformer on Tinystory Context



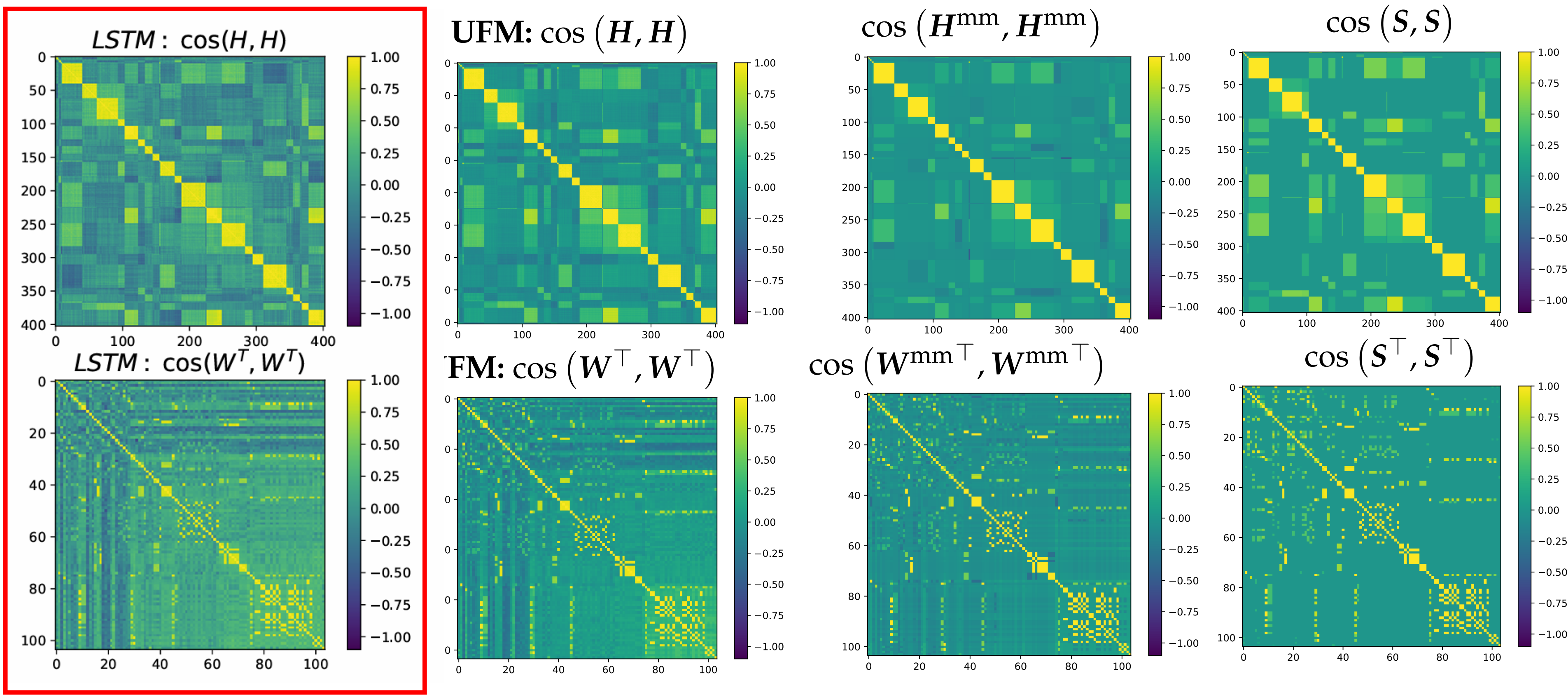
Numerical Example

- ✓ UFM is a good proxy
- ✓ Eigenfactors of NTP-max-margin L^{mm} predict geometry
- ✓ Support overlaps already give a good proxy for geometry



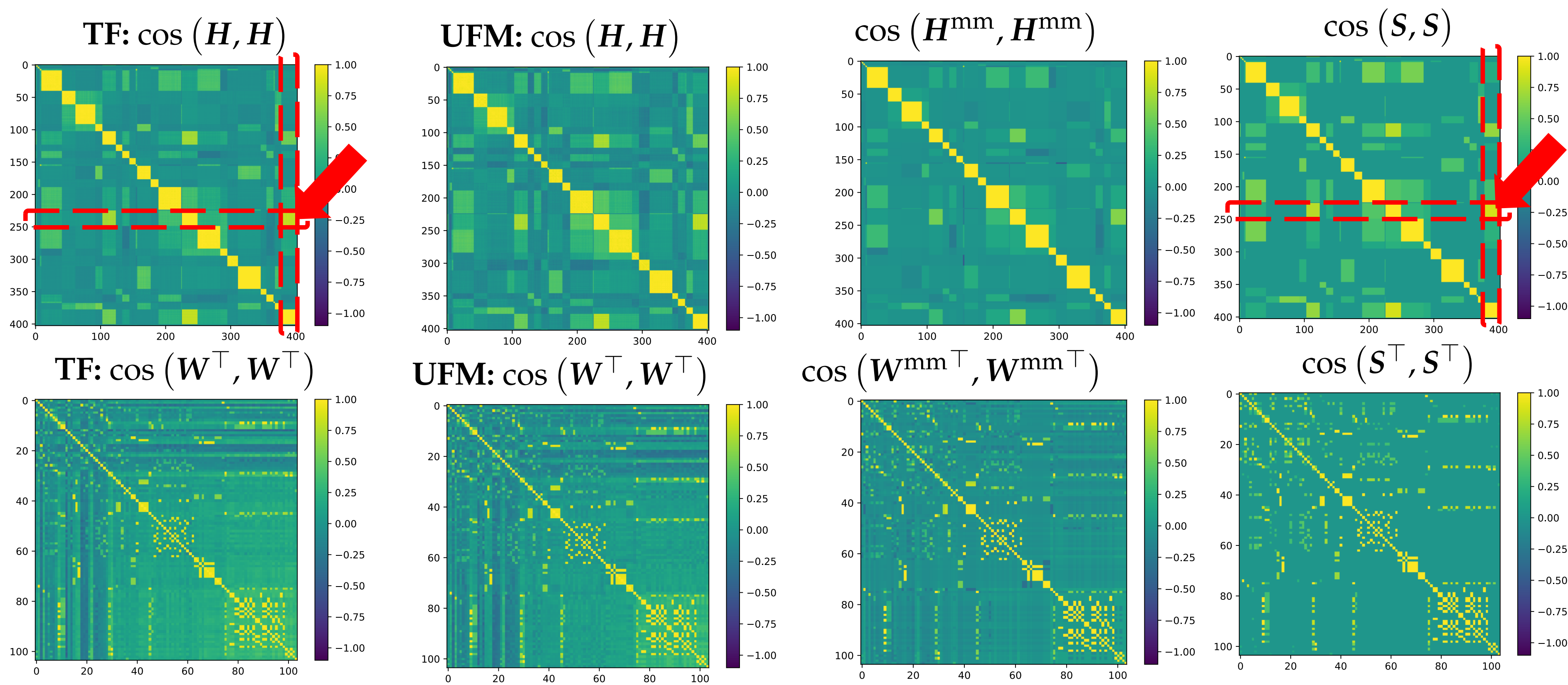
Numerical Example

- ✓ UFM is a good proxy
- ✓ Eigenfactors of NTP-max-margin L^{mm} predict geometry
- ✓ Support overlaps already give a good proxy for geometry



Numerical Example

- ✓ UFM is a good proxy
- ✓ Eigenfactors of NTP-max-margin L^{mm} predict geometry
- ✓ Support overlaps already give a good proxy for geometry



Context 1: "boy named timmy . timmy"
Context 2: "kid called lilly. she"

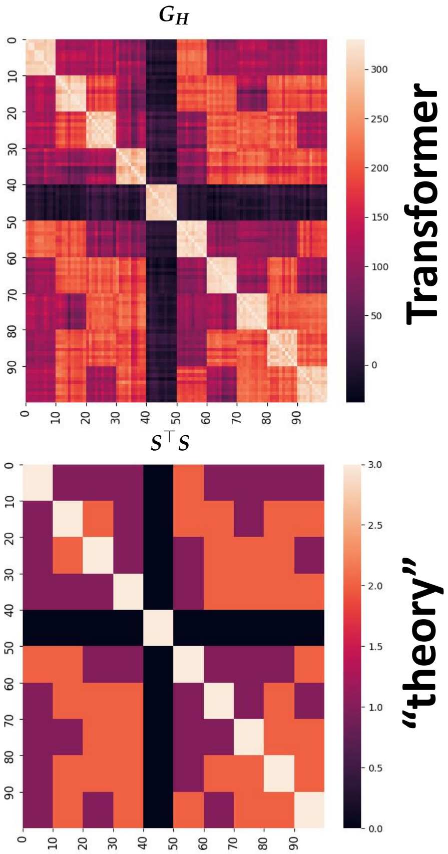
Subspace collapse: Embeddings of ctxs whose supports overlap collapse

Summary

A framework for **mapping language patterns to embeddings geometry** via:

1. Framing NTP as **sparse** **soft labels** classification
2. Applying **unconstrained features**
3. Leveraging **implicit bias** viewpoint

➔ Word/context embeddings as mtx factorization of $\boldsymbol{L}_\lambda \approx \boldsymbol{L}_{\text{sparse}} + \rho(\lambda) \cdot \boldsymbol{L}_{\text{low-rank}}$

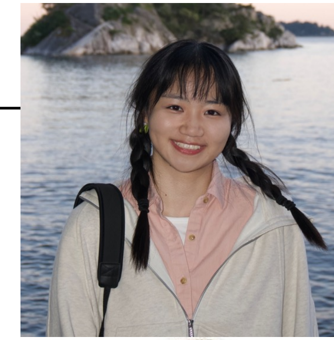


NTP: Open questions

- Directly-related questions:
  - ? **Gradient-descent** convergence
  - ? What is the impact of Zipf-law imbalances on convergence?
  - ? **d<V**: **Do linguistics sparsity patterns lead to low-rank solutions?**
  - ? Geometry at higher layers of linguistic understanding, e.g. concepts

- The setting is clearly "statistical":
  - ? How do these optimization results inform **generalization**?
  - ? What is the statistical role of margin btwn in/out-of-support tokens?
  - ? What are good data models to study these
  - ? When is it good to train long or is better to stop early?

Summary



A framework for **mapping language patterns to embeddings geometry** via:

1. Framing NTP as **sparse** **soft labels** classification
2. Applying **unconstrained features**
3. Leveraging **implicit bias** viewpoint

➔ Word/context embeddings as mtx factorization of $\boldsymbol{L}_\lambda \approx \boldsymbol{L}_{\text{sparse}} + \rho(\lambda) \cdot \boldsymbol{L}_{\text{low-rank}}$

1. CT, Implicit Bias of Next-token Prediction, NeurIPS 2024
2. Zhao, Behnia, Vakilan, CT, Implicit Geometry of Next-token Prediction: From Language Sparsity Patterns to Model Representations, COLM 2024.

Thank you!