Transformers as Support Vector Machines

Davoud Ataee Tarzanagh

University of Pennsylvania

Refs:

- “Transformers as Support Vector Machines”, arXiv:2308.16898, 2023.
- “Margin Maximization in Attention Mechanism”, NeurIPS 2023.

Collaborators

Samet Oymak
University of Michigan
Ann Arbor

Yingcong Li
University of California
Riverside

Xuechen Zhang
University of California
Riverside

Christos Thrampoulidis
University of British
Columbia

What is a transformer?

A neural network architecture that:

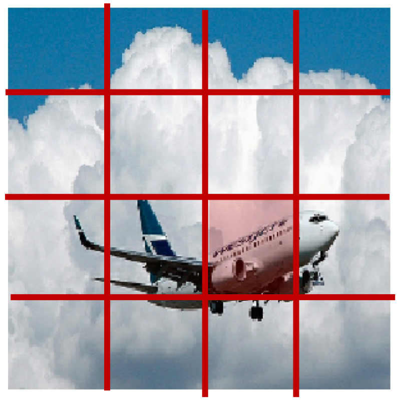
1. **Tokenization:** Input is treated as a sequence of tokens
2. **Attention mechanism:** Calculates dot-products between tokens

Text input: "This is a sample sentence."

Tokens: ["This", "is", "a", "sample", "sentence"]

Visual input:

Tokens are patches:



Transformer is introduced in

Attention is all you need

Vaswani et al, NeurIPS 2017

Revolutionized NLP
Underlies ChatGPT

This talk: Understanding transformer and attention through optimization theory.

Why attention

✓ Allows the model to **focus on relevant subset of sequence**

✓**Tokens explicitly interact!** (in contrast to traditional neural nets)

Example: English to French translation

Input: "The agreement on the European Economic Area was signed in August 1992."

Output: "L'accord sur la zone économique européenne a été signé en août 1992."

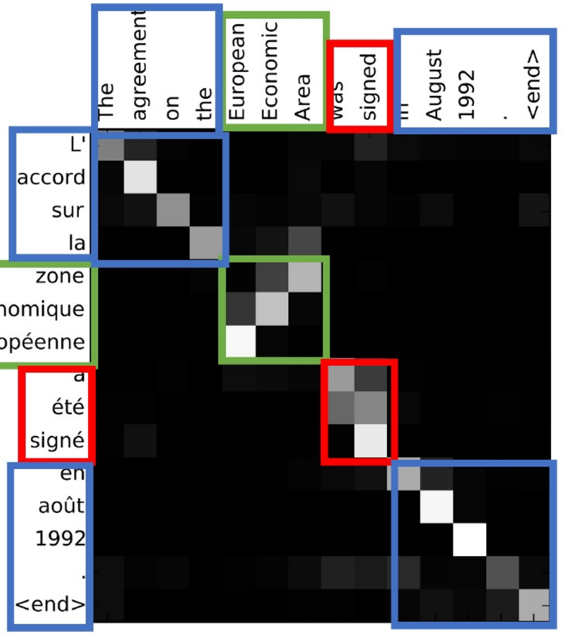
Visualize attention weights $a_{t,i}$

Diagonal attention means words correspond in order

Attention figures out different word orders

Verb conjugation

Diagonal attention means words correspond in order



Bahdanau et al, "Neural machine translation by jointly learning to align and translate", ICLR 2015

Transformer

Transformer maps sequence to sequence

Input: Sequence of tokens $X = [x_1 \dots x_T]$

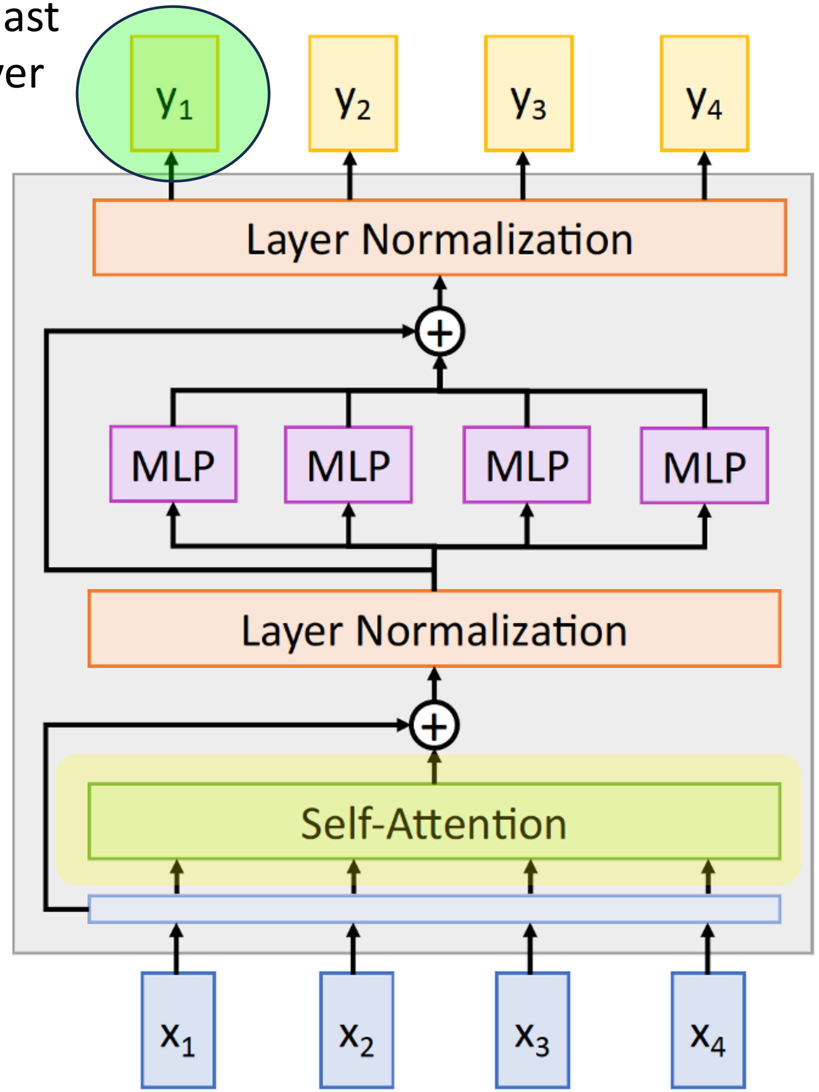
Output: Sequence of tokens $Y = [y_1 \dots y_T]$

Self-attention is the only interaction between tokens

Layer norm and MLP work independent per token

Predict via first/last token of last layer

Single transformer layer



Modern transformers stack multiple layers of Self-attention+MLP.

Self-attention Layer

Maps an input sequence to an output sequence

Let $X \in \mathbb{R}^{T\times d}$ be an input sequence of T tokens

- T: Sequence length
- d: Dimensionality of tokens

- Self-attention layer has trainable weight matrices $K, Q, V \in \mathbb{R}^{d\times d}$. Obtain
  - **keys:** $X_K = XK$
  - **queries:** $X_Q = XQ$
  - **values:** $X_V = XV$

  $\in \mathbb{R}^{T\times d}$

- It outputs the sequence

$$\mathbb{S}(X_Q X_K^\top)X_V = \mathbb{S}(XQK^\top X^\top)XV$$

Query $T\times d$ Key $d\times T$ Value $T\times d$

- $\mathbb{S}(\cdot)$ denotes the **softmax** nonlinearity

Let us focus on a clean formulation!

Optimization formulation

Classification: Map input sequence $X \in \mathbb{R}^{T\times d}$ to label $Y \in \{-1,1\}$

Original model: $f(X) = \mathbb{S}(XQK^\top X^\top)XV$ => $T\times d$ dimensional

- Read only first token's output
- Use $V = v \in \mathbb{R}^d$

Classification model: $f(X) = v^\top X^\top \mathbb{S}(XKQ^\top x_1)$ => Scalar output (transposed notation)

Training dataset: $(X_i, Y_i)_{i=1}^n$

Empirical risk minimization:

$$\mathcal{L}(K,Q) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^T X_i^T \mathbb{S}(X_i K Q^T x_{i1})\right)$$

- $\ell: \mathbb{R} \to \mathbb{R}$ is a strictly decreasing smooth loss
- We can actually allow any $z_i \leftarrow x_{i1} \in \mathbb{R}^d$

Goal: Understand what happens when we train a transformer. What does it learn to do?

Optimization formulation

Classification dataset: $(z_i, X_i, Y_i)_{i=1}^{n}$ with labels $Y_i \in \{-1,1\}$

➢Starting point $W = KQ^T$

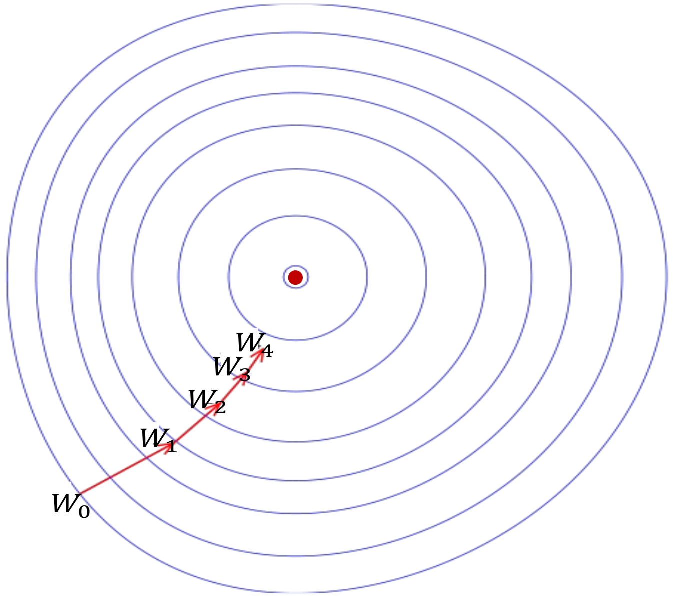
$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^T X_i^T \mathbb{S}(X_i W z_i)\right)$$

Main Q: When we solve this problem, which attention weights $\boldsymbol{W}$ we find?

Gradient descent (GD) trajectory

Given $\boldsymbol{W}(0) \in \mathbb{R}^{d\times d}$, $\eta > 0$, for $t \geq 0$ do:

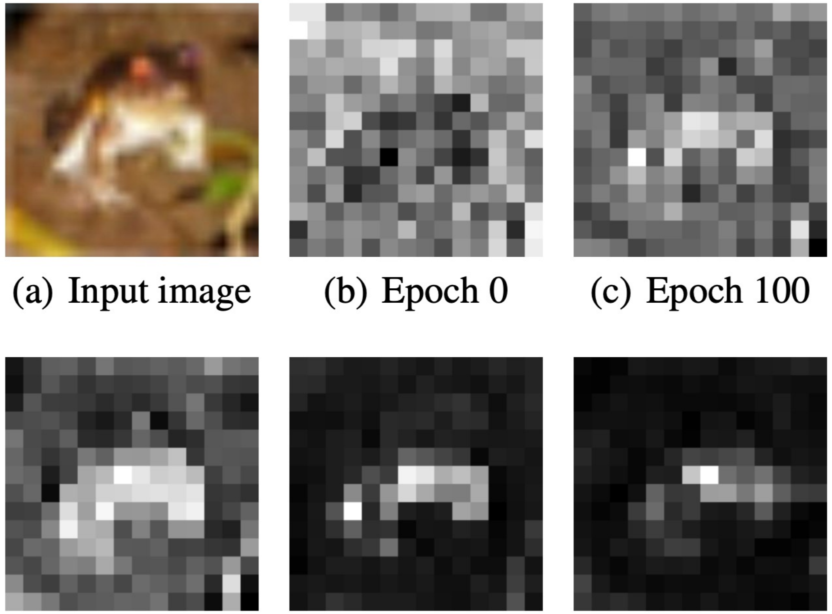
$$\boldsymbol{W}(t+1) = \boldsymbol{W}(t) - \eta \nabla \mathcal{L}(\boldsymbol{W}(t)). \qquad \text{(GD-W)}$$



Empirical insights

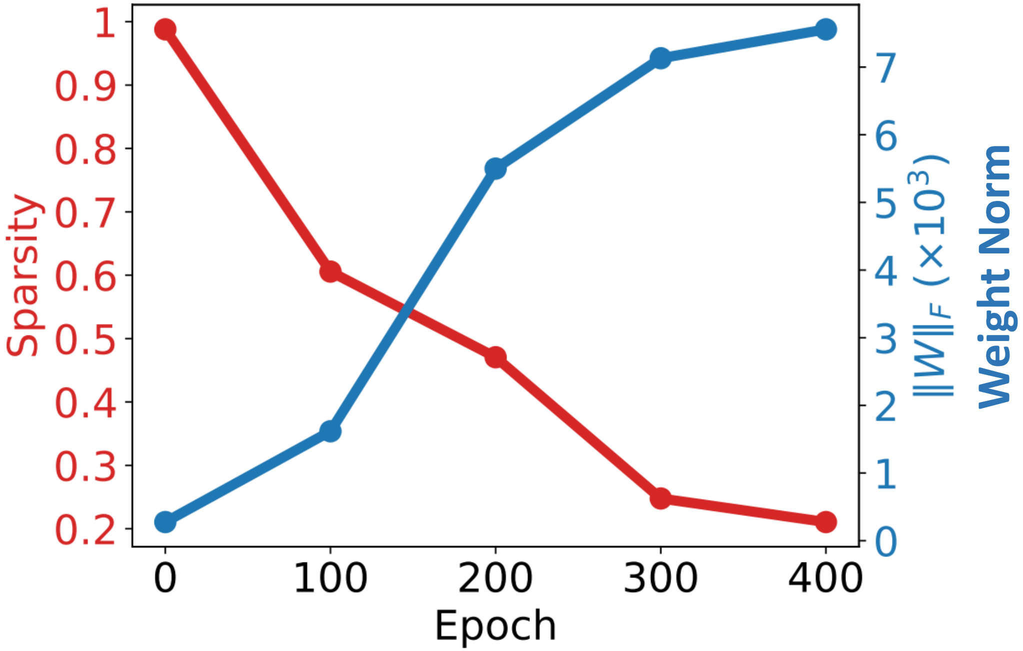
Observation: Attention mechanism selects few tokens **most relevant** for prediction. As we select fewer tokens, norm of the weights grow.

Our theory rigorizes this via "**optimal tokens**" & **Transformer-SVM** equivalence



(a) Input image (b) Epoch 0 (c) Epoch 100

(d) Epoch 200 (e) Epoch 300 (f) Epoch 400



Recap: Softmax function $\mathbb{S}$

Softmax maps a vector $v \in \mathbb{R}^T$ into probability distribution

$$\mathbb{S}(v)_t = \frac{\mathrm{e}^{v_\mathrm{t}}}{\sum_{t=1}^{T} e^{v_t}}$$

Softmax implies $\sum_{t=1}^{T} \mathbb{S}(v)_t = 1$

✓For finite v: $1 > \mathbb{S}(v)_t > 0$

✓Only way to attain $\mathbb{S}(v)_t \in \{0,1\}$ is $||v|| \to \infty$

(a.k.a. saturated softmax)

Attention outputs: $x^{\mathrm{att}} = X^\top s$ where $\mathrm{s} = \mathbb{S}(XWz)$

➢x^{att} is a **convex** combination of tokens of X

$$\boxed{\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^\top X_i^\top \mathbb{S}(X_i W z_i)\right)}$$

🤔 What if we want to output the k'th token i.e. $x^{\mathrm{att}} \leftarrow x_k$

✓Then $s_t = 1$ if and only if $t = k$

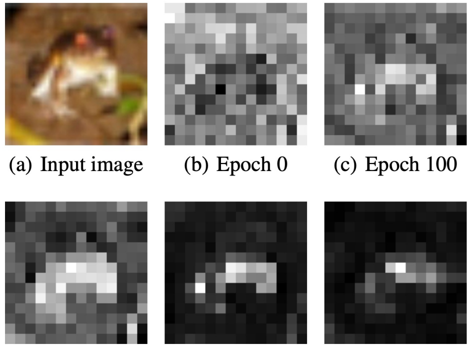
✓$||W|| \to \infty$

Contributions (high-level summary)

- **Main contribution:** We characterize the optimization geometry of self-attention layer.

- Attention weights converge towards an **SVM solution** that separates *optimal* tokens within each input sequence from *non-optimal* tokens. Attention's SVM serves as a **good-token-selector**.
  - ✓ SVM bias arises because *gradient descent saturates softmax to select optimal tokens*

- **How attention induces sparsity:** *Non-optimal* tokens that fall on the wrong side of the SVM decision boundary are suppressed by the softmax function, while *optimal* tokens receive nonzero probability.

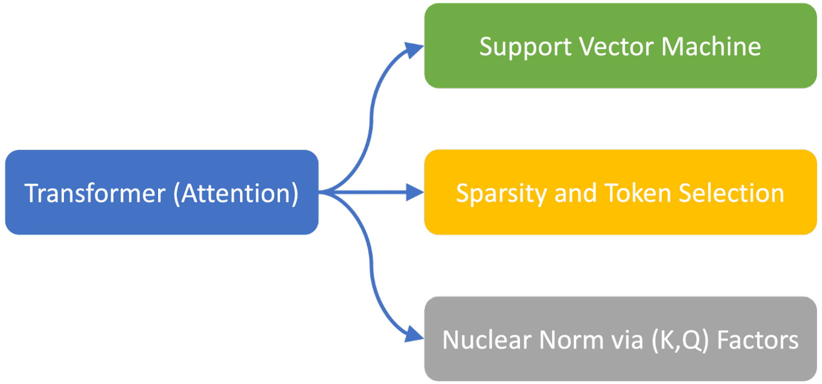


(a) Input image (b) Epoch 0 (c) Epoch 100
(d) Epoch 200 (e) Epoch 300 (f) Epoch 400

Attention focusing on fewer tokens over time

- **Connections to Core ML:** Our results reveal transformers integrate 3 core ML themes:

1. SVMs and margin maximization
2. Token selection and sparsity (↔ feature selection, lasso…)
3. Low-rank factorization and nuclear norm
   - **Why?** (K, Q) in $\mathbb{S}(XQK^TX^T)XV$ is factorization of $W = QK^T$



- **Further discussion…**
  1. *Locally- vs globally-optimal* SVMs
  2. Role of *overparameterization*
  3. *Generalized SVM* equivalence for *MLP nonlinearities*
- Many open problems ☺

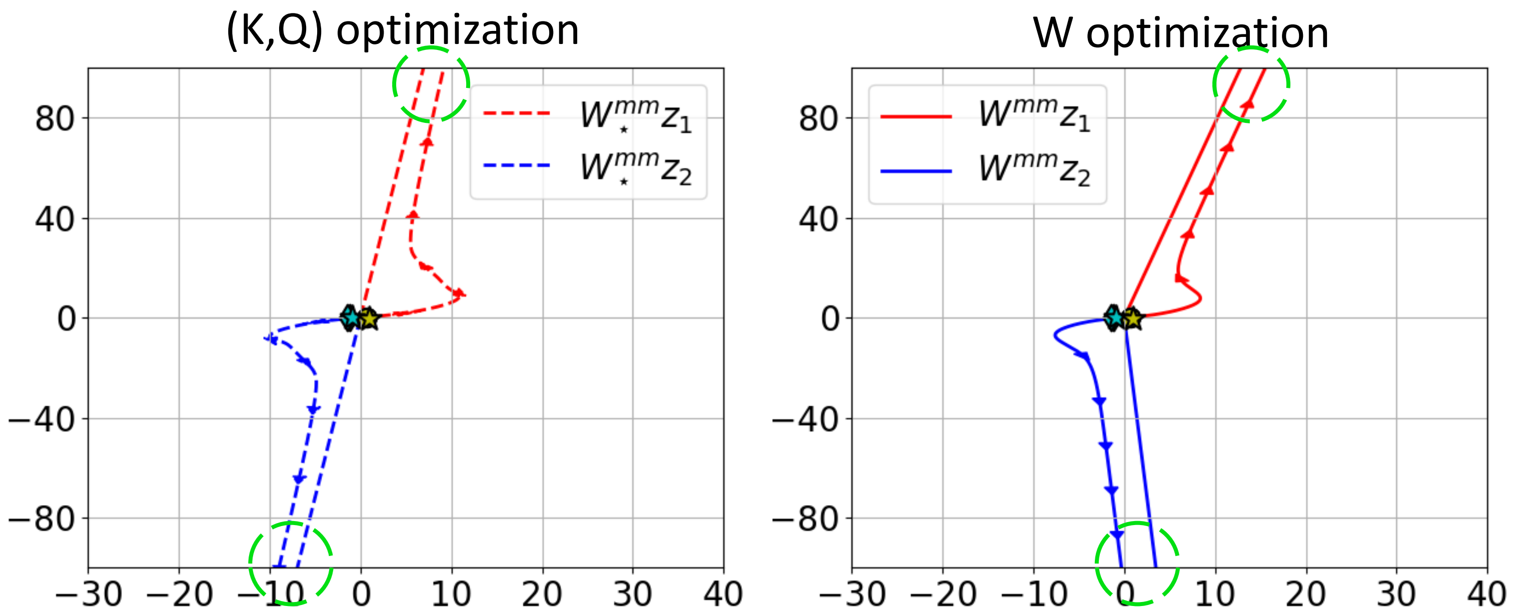
Numerical example: $n = 2$ inputs each with T=3 tokens. Token dim d=2

$W \in \mathbb{R}^{2\times 2}$

1. **W-ERM:** $\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^T X_i^T \mathbb{S}(X_i W z_i)\right)$
2. **KQ-ERM:** $\mathcal{L}(K,Q) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^T X_i^T \mathbb{S}(X_i K Q^T z_i)\right)$

Arrows: Trajectory of gradient descent
Straight lines: Direction of the SVM solutions

- Display 2D projections of W: (Wz_1, Wz_2)
- For (K,Q) optimization, we show $W \leftarrow KQ^T$

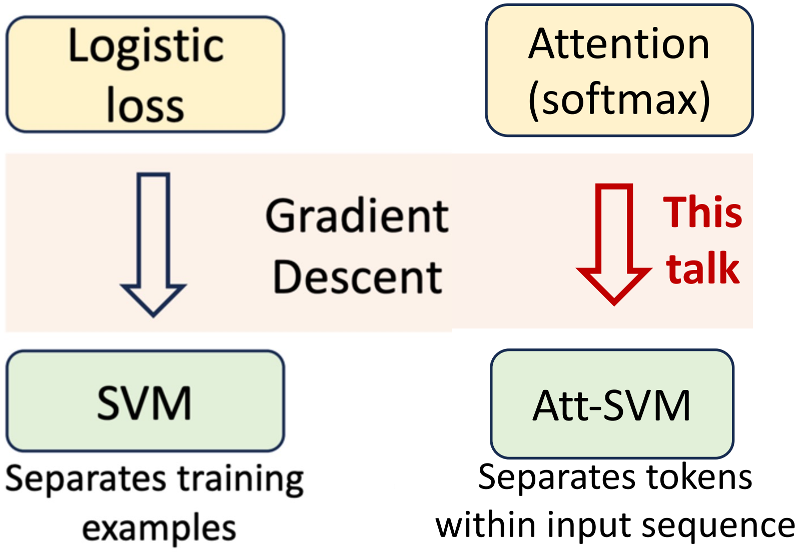


Teal and yellow markers represent tokens from X_1 and X_2. **Green circles** denote GD $\leftrightarrow$ SVM pairings.

Connection to prior work (high-level)

➢ Gradient-methods under exponential or logistic loss minimization are biased towards maximum-margin solutions [Ji and Telgarsky'18, Soudry et al.'18, Gunasekar et al'18]. Goes back to [Rosset et al'03]

➢ Softmax within attention layer has exponential nature



Key differences from prior works:

1. Nonconvex loss ℓ + nonlinear softmax
2. Complex problem geometry:
   - ➢ SGD can converge to one of many SVMs
3. Att-SVM is different from vanilla SVM

Intuition

$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^\top X_i^\top \mathbb{S}(X_i W z_i)\right)$$

Suppose ℓ is decreasing: W should maximize inner sum $\sum_{i=1}^{T} \mathbb{S}_t \cdot \boxed{Y_i \cdot v^\top x_{it}}$

token score

- Input sequence $X_i = [x_{i1} \dots x_{iT}]$ have T tokens
- Fortunately, we can define **optimal token** which minimizes the training loss $\mathcal{L}(W)$

Definition 1 (Optimal token) *Given $\boldsymbol{v} \in \mathbb{R}^d$, the optimal token for $\boldsymbol{X}_i$ is the index $\mathbf{opt}_i \in \arg\max_{t\in[T]} Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{it}$.*

Lemma 2 (Optimal tokens minimize training risk) *Suppose ℓ is strictly decreasing and smooth. Then, training risk obeys $\mathcal{L}(\boldsymbol{W}) > \mathcal{L}_\star := \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{i\mathrm{opt}_i})$.*

Training loss at optimal tokens

WHY: Because the best we can do is setting $\mathbb{S}_{opt_i} = 1$

Question: Can we ever achieve the optimal loss $\mathcal{L}_\star$? 😢

Answer: Yes, if softmax selects "optimal tokens". But we have to let $||W||_F \to \infty$

$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^\top X_i^\top \mathbb{S}(X_i W z_i)\right)$$

Attention SVM

Our theory: 1-layer attention is biased towards a hard-margin *Att-SVM*.

Att-SVM separates "optimal tokens from non-optimal tokens".

SVM for W-ERM

$$\boldsymbol{W}^{mm} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{subj. to} \quad (\boldsymbol{x}_{i\mathrm{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1 \quad \text{for all} \quad t \neq \mathrm{opt}_i, \quad i \in [n]. \qquad \text{(Att-SVM)}$$

Max-margin solution

Theorem 2 (TLTO'23, Regularization Path→Att-SVM) *Suppose optimal indices* $(\mathrm{opt}_i)_{i=1}^n$ *are unique and* (Att-SVM) *is feasible. Let* $\boldsymbol{W}^{mm}$ *be the unique solution of* (Att-SVM) *with Frobenius norm. Then,*

Weights go to ∞, but the direction converges to SVM solution!

$$\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}$$

Regularization path

$$\bar{\boldsymbol{W}}_R = \arg\min_{\|\boldsymbol{W}\|_F \leq R} \mathcal{L}(\boldsymbol{W}).$$

$$\mathcal{L}(K,Q) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^T X_i^T \mathbb{S}(X_i K Q^T z_i)\right)$$

Attention SVM: (K,Q)-ERM

SVM for (K, Q)-ERM

$$W_\star^{mm} \in \arg\min_{W} \|W\|_\star \quad \text{subj. to} \quad (x_{i\mathrm{opt}_i} - x_{it})^\top W z_i \geq 1 \quad \text{for all} \quad t \neq \mathrm{opt}_i, \quad i \in [n]. \qquad \text{(Att-SVM}^\star\text{)}$$

Nuclear norm

Theorem 3 (Regularization Path→Att-SVM*) *Suppose ℓ is smooth and decreasing, optimal indices $(\mathrm{opt}_i)_{i=1}^n$ are unique, and (Att-SVM) is feasible. Let $\mathcal{W}_\star^{mm}$ be the solution set of (Att-SVM*) achieving objective $C_\star$. Then,*

$$\lim_{R\to\infty} dist\left(\frac{\bar{K}_R \bar{Q}_R^\top}{R}, \frac{\mathcal{W}_\star^{mm}}{C_\star}\right) = 0$$

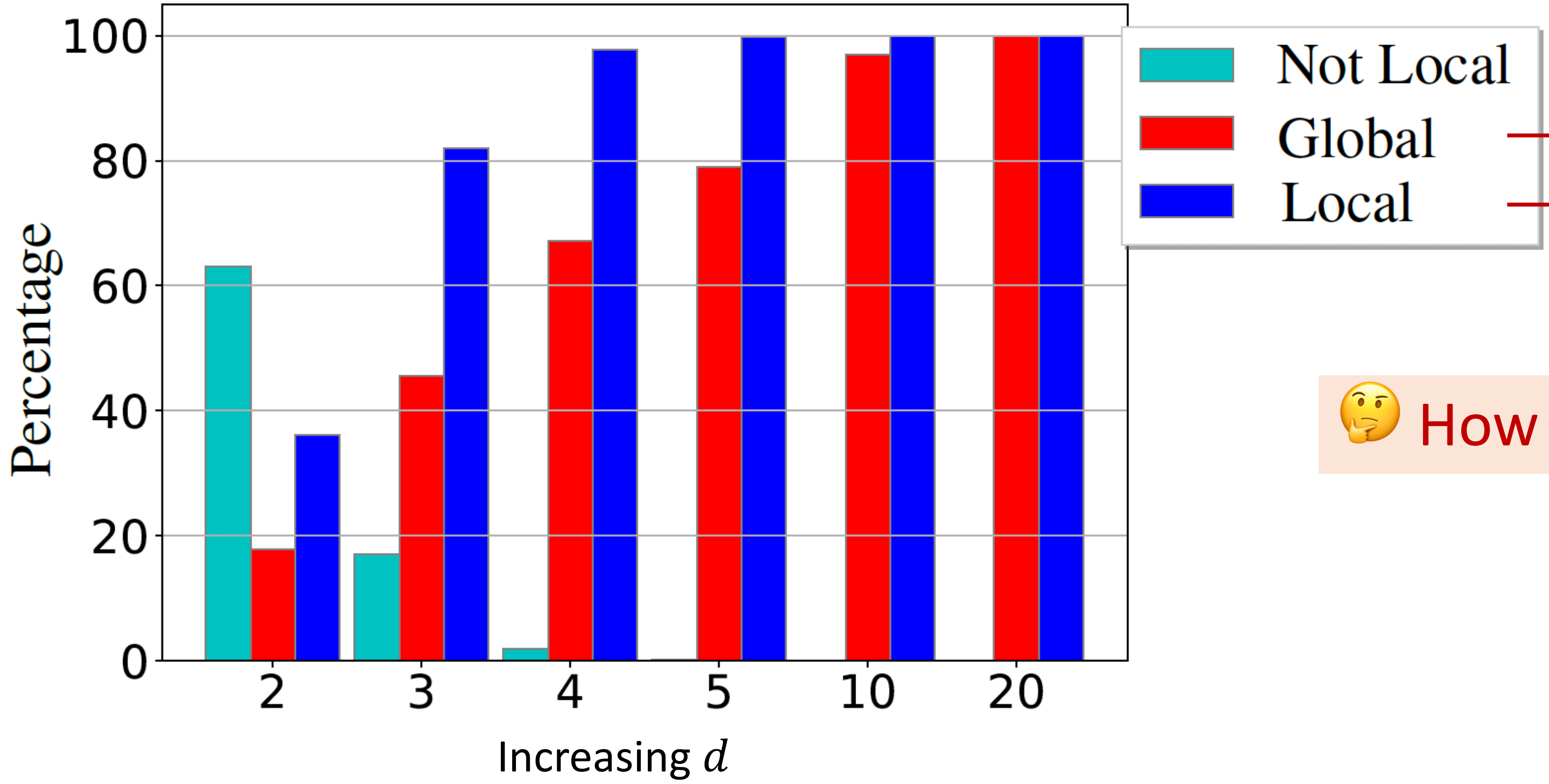
$$(\bar{K}_R, \bar{Q}_R) = \arg\min_{\|K\|_F^2 + \|Q\|_F^2 \leq 2R} \mathcal{L}(K, Q).$$

$$\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^T X_i^T \mathbb{S}(X_i W z_i)\right)$$

Gradient descent theory

So far: Regularization path selects optimal token opt_i from input sequence X_i

Q: Does GD follow regularization path for self-attention?



Global → Convergence to opt_i

Local → Convergence to $\alpha_i \neq \mathrm{opt}_i$

🤔 How to explain this?

Larger $W \in \mathbb{R}^{d\times d}$
Fixed (n, T)

Optimization geometry of attention

GD can select locally-optimal tokens!

$$\boldsymbol{W}^{\mathrm{mm}}(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{subj. to} \quad (\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1 \quad \text{for all} \quad t \neq \alpha_i, \quad i \in [n]. \qquad \text{(Local-SVM)}$$

Definition 2 (Support indices and locally-optimal direction) *Fix token indices* $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$. *Solve* (Att-SVM) *with*
$(\texttt{opt}_i)_{i=1}^n$ *replaced with* $\boldsymbol{\alpha} = (\alpha$ … $T]$ *such that* $(\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{\mathrm{mm}}(\boldsymbol{\alpha}) \boldsymbol{z}_i = 1$
for all $t \in \mathcal{T}_i$. *We refer to* $(\mathcal{T}_i)_{i=1}^n$ … *all* $i \in [n]$ *and* $t \in \mathcal{T}_i$ *scores per Def.* 1
obey $\boldsymbol{\gamma}_{i\alpha_i} > \boldsymbol{\gamma}_{it}$, *indices* $\boldsymbol{\alpha} = (\alpha$ … *alled a* locally-optimal direction.

See the paper ☺

Originally developed in [TLZO, NeurIPS’23] for prompt-tuning. [TLTO’23] adapts to self-attention.

Gradient descent theory

GD can select locally-optimal tokens!

$$\boldsymbol{W}^{mm}(\alpha) = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{subj. to} \quad (\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1 \ \ \forall \ t \neq \alpha_i. \qquad \text{(Local-SVM)}$$

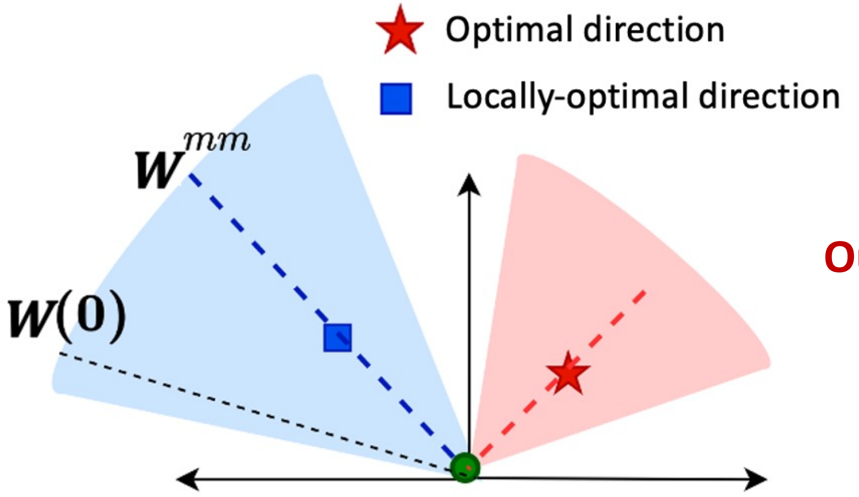


Figure 2: Gradient descent initialization $\boldsymbol{W}(\boldsymbol{0})$ inside the cone containing the locally-optimal solution $\boldsymbol{W}^{mm}$

Our GD theory

Main results (simplified)

Gradient descent: Given $\boldsymbol{W}_0 \in \mathbb{R}^{d\times d}$, $\eta > 0$, for $k \geq 0$ do:

$$\boldsymbol{W}_{k+1} = \boldsymbol{W}_k - \eta \nabla \mathcal{L}(\boldsymbol{W}_k).$$

Theorem (local conv): For locally-optimal $\boldsymbol{\alpha}$, if GD is initialized in the local cone with large $||\boldsymbol{W}_0||$ then $\frac{\boldsymbol{W}_k}{||\boldsymbol{W}_k||_F} \to \frac{\boldsymbol{W}^{mm}(\boldsymbol{\alpha})}{||\boldsymbol{W}^{mm}(\boldsymbol{\alpha})||_F}$

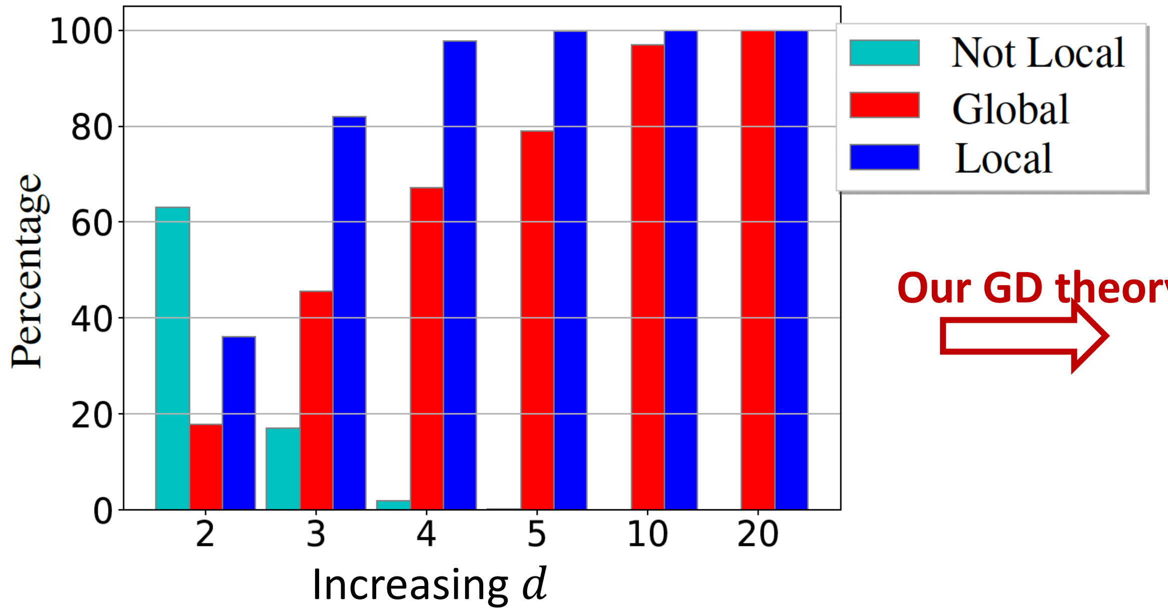
🤔 **When do we converge to global optimum?**

Gradient descent theory

Gradient descent: Given $W_0 \in \mathbb{R}^{d\times d}$, $\eta > 0$, for $k \geq 0$ do:

$$W_{k+1} = W_k - \eta \nabla \mathcal{L}(W_k).$$

$$W^{mm}(\mathrm{opt}) = \arg\min_{W} \|W\|_F \quad \text{subj. to} \quad (x_{i\alpha_i} - x_{it})^\top W z_i \geq 1 \;\; \forall \; t \neq \mathrm{opt}_i. \quad \text{(Att-SVM)}$$



Our GD theory

Main results on large d

Theorem: If *all tokens are support vectors of Att-SVM* (i.e. SVM margin constraints are tight), then

- **No stationary points:** $\nabla\mathcal{L}(W) \neq 0$ for all W
- **GD diverges:** $\|W_k\|_F \to \infty$

✓ This condition holds as d grows (explains blue bars → 1)

Lemma: If all tokens are support vectors in all Local-SVM's, then $(\mathrm{opt_i})_{i=1}^{n}$ is the **only feasible locally-optimal solution.**

✓ Holds as d becomes even larger (explains red bars → 1)

✓ Culminates in our global convergence conjecture (see paper)

Global conv with alternative criteria

Theorem: We have that $\frac{W_k}{\|W_k\|_F} \to \frac{W^{mm}(\mathrm{opt})}{\|W^{mm}(\mathrm{opt})\|_F}$, if

✓ Scores of non-optimal tokens are ≈equal

✓ Initial gradient $\nabla\mathcal{L}(W_0)$ is favorable.

Can the theory account for MLP layers?

So far: $\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot v^\top X_i^\top \mathbb{S}(X_i W z_i)\right) \Rightarrow$ Attention selects 1-token α_i

$$\boldsymbol{W}^{mm}(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{subj. to} \quad (\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1 \;\; \forall \;\; t \neq \alpha_i. \qquad \text{(Local-SVM)}$$

How about: $\mathcal{L}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \cdot h(X_i^\top \mathbb{S}(X_i W z_i))\right)$ for nonlinear h?

In a nutshell: Nonlinearity is key to selecting >1 token from input seqs

Question: How should this SVM theory be generalized?

Generalized SVM↔Attention Equivalence

Suppose GD solution "selects" a token set $\mathcal{O}_i \subseteq [T]$ for X_i for $1 \leq i \leq n$

Claim: $W_{GD} \approx W_{\text{fin}} + W_{\text{svm}}$

- Job of W_{svm}: Select $\mathcal{O}_i$ and suppress $\bar{\mathcal{O}}_i = T - \mathcal{O}_i$ for all $1 \leq i \leq n$
- Job of W_{fin}: Allocate the nonzero softmax probabilities within tokens $\mathcal{O}_i$
- $||W_{\text{svm}}||_F \rightarrow \infty$, $||W_{\text{fin}}||_F \rightarrow$ bounded

For W_{fin}: See TLTO'23

$$\boldsymbol{W}^{mm} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{subj. to} \quad \begin{cases} \forall\, t \in \mathcal{O}_i, \tau \in \bar{\mathcal{O}}_i : & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1, \\ \forall\, t, \tau \in \mathcal{O}_i : & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i = 0, \end{cases} \quad \forall 1 \leq i \leq n. \quad \text{(Gen-SVM)}$$

Generalized SVM↔Attention Equivalence

Claim: GD with MLP should select >1 tokens.

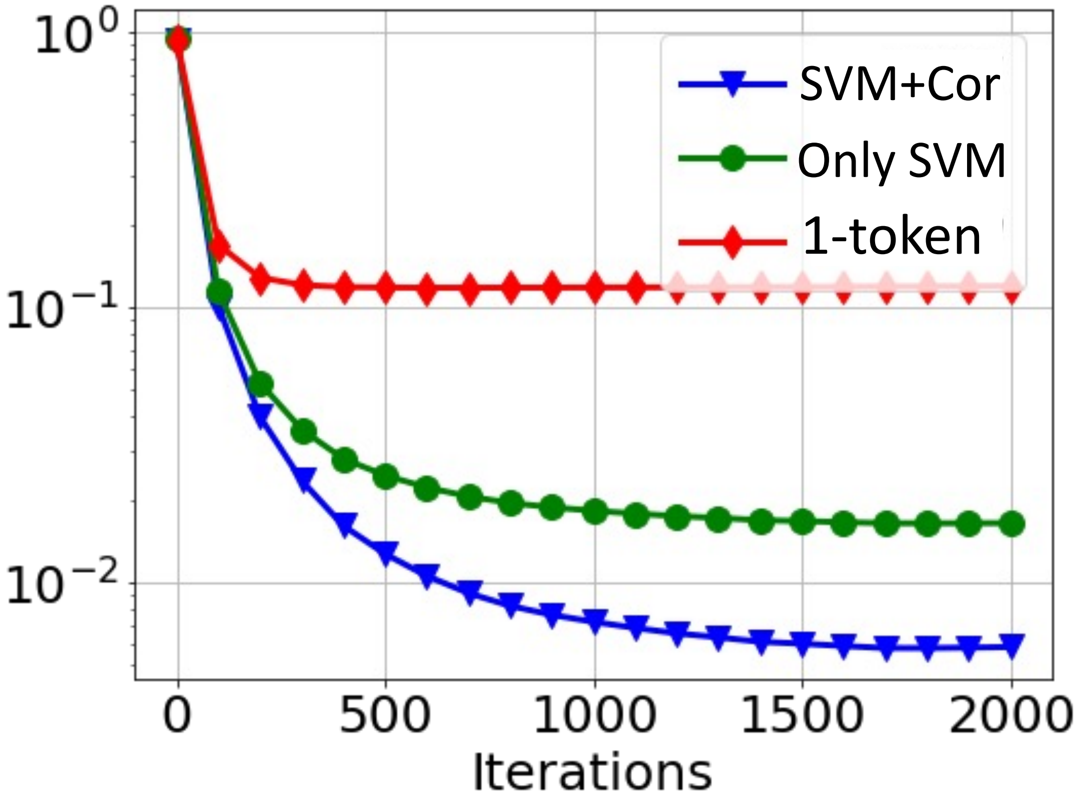
General form: $W_{GD} \approx W_{\text{cor}} + W_{\text{svm}}$

$$W^{mm} = \arg\min_{W} \|W\|_F \quad \text{subj. to} \quad \begin{cases} \forall\, t \in O_i, \tau \in \bar{O}_i: & (x_{it} - x_{i\tau})^\top W z_i \geq 1, \\ \forall\, t, \tau \in O_i: & (x_{it} - x_{i\tau})^\top W z_i = 0, \end{cases} \quad \forall 1 \leq i \leq n. \quad \text{(Gen-SVM)}$$

Q: Do these actually work in experiments?

$1 - \text{corr\_coef}$
$(W_{\text{GD}}, W_{\text{theory}})$

>99% corr



Summary

- **This talk:** Optimization theory for attention and transformers
  - ✓ Fundamental connections to **support vector machines**
  - ✓ **Attention** is a max-margin ~~classifier~~ **token selector**
  - ✓ **Parameterization matters:** $\boldsymbol{W}$ →min_Frob_norm, $(\boldsymbol{K}, \boldsymbol{Q})$ →min_Nuclear_norm bias
  - ✓ **A new perspective:** Can we interpret multilayer transformers as an SVM hierarchy?
  - ✓ **MLP nonlinearity** is key to selecting and composing multiple tokens
    - Results in a richer SVM equivalence (no rigorous theory yet!)

- Some future directions
  - Optimization meets Generalization
  - Gradient descent on (K,Q)
  - Convergence rates
  - Demystifying wide/narrow cone
  - MLP and Generalized SVM
  - Resolving global convergence of GD
  - Multilayer/Multihead architectures
  - Jointly optimizing (W, V)