Transformers as Support Vector Machines

Davoud Ataee Tarzanagh[1] Yingcong Li[2,4] Christos Thrampoulidis[3] Samet Oymak[4]
University of Pennsylvania[1] University of California, Riverside[2] University of British Columbia[3] University of Michigan[4]

Question

Can we characterize the optimization landscape and implicit bias of Transformers' attention mechanism?

Optimization Methods

- $\boldsymbol{W}$-parameterization: **Gradient Descent** with stepsize $\eta > 0$:

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) - \eta \nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}(k)), \qquad \text{(GD-W)}$$

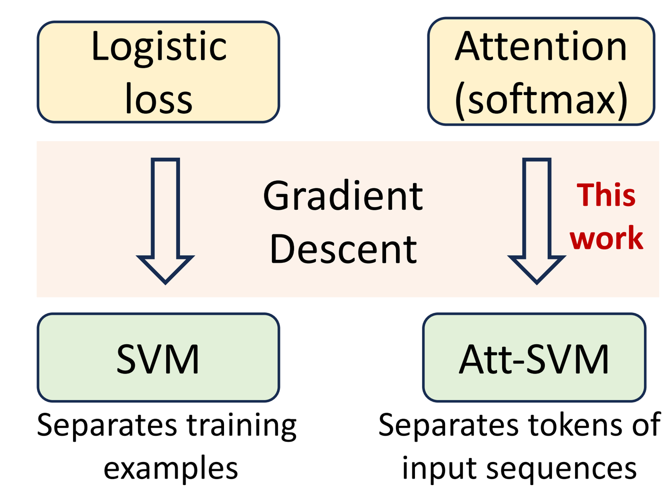
- $(\boldsymbol{K}, \boldsymbol{Q})$-parameterization: **Regularization Path** with radius $R > 0$:

$$(\boldsymbol{K}_R, \boldsymbol{Q}_R) = \underset{\|\boldsymbol{K}\|_F^2 + \|\boldsymbol{Q}\|_F^2 \le 2R}{\arg\min} \mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}). \qquad \text{(RP-KQ)}$$

Motivation

Exploring implicit bias is a key step in unraveling the generalization of the (softmax-)attention mechanism.

Conclusion



Transformers are SVMs!

Softmax-Attention

$$f(\boldsymbol{X}) = h(\boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{K}^\top \boldsymbol{X}^\top))$$

- $\boldsymbol{K}, \boldsymbol{Q} \in \mathbb{R}^{d\times m}$, $\boldsymbol{W} := \boldsymbol{K}\boldsymbol{Q}^\top$: attention weights,
- $\mathbb{S}(\cdot)$: softmax function, $h(\cdot)$: prediction head.

Problem Description: Given training dataset $(Y_i, \boldsymbol{X}_i, \boldsymbol{z}_i)_{i=1}^n$ where $Y_i \in \{-1, 1\}$, $\boldsymbol{X}_i \in \mathbb{R}^{T\times d}$ and $\boldsymbol{z}_i \in \mathbb{R}^d$, we explore the training risk with a loss ℓ as follows:

$$\mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}) = \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i \cdot f(\boldsymbol{X}_i)\right), \qquad \text{(ERM)}$$

where $f(\boldsymbol{X}_i) = h(\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{K}\boldsymbol{Q}^\top \boldsymbol{z}_i))$.

Attention SVM

For given indices of **selected tokens** $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_i)_{i=1}^n$, define

- SVM for $\boldsymbol{W}$-parameterization:

$$\boldsymbol{W}^{mm} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F$$
$$\text{s.t.} \quad (\boldsymbol{x}_{i\boldsymbol{\alpha}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}\boldsymbol{z}_i \ge 1, \ \forall i, t \ (t \neq \boldsymbol{\alpha}_i)$$

- SVM for $(\boldsymbol{K}, \boldsymbol{Q})$-parameterization ($\boldsymbol{W} := \boldsymbol{K}\boldsymbol{Q}^\top$):

$$\boldsymbol{W}_\star^{mm} \in \arg \min_{\texttt{rank}(\boldsymbol{W}) \le m} \|\boldsymbol{W}\|_\star$$
$$\text{s.t.} \quad (\boldsymbol{x}_{i\boldsymbol{\alpha}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}\boldsymbol{z}_i \ge 1, \ \forall i, t \ (t \neq \boldsymbol{\alpha}_i)$$

Convergence of Attention Weights with Linear Head

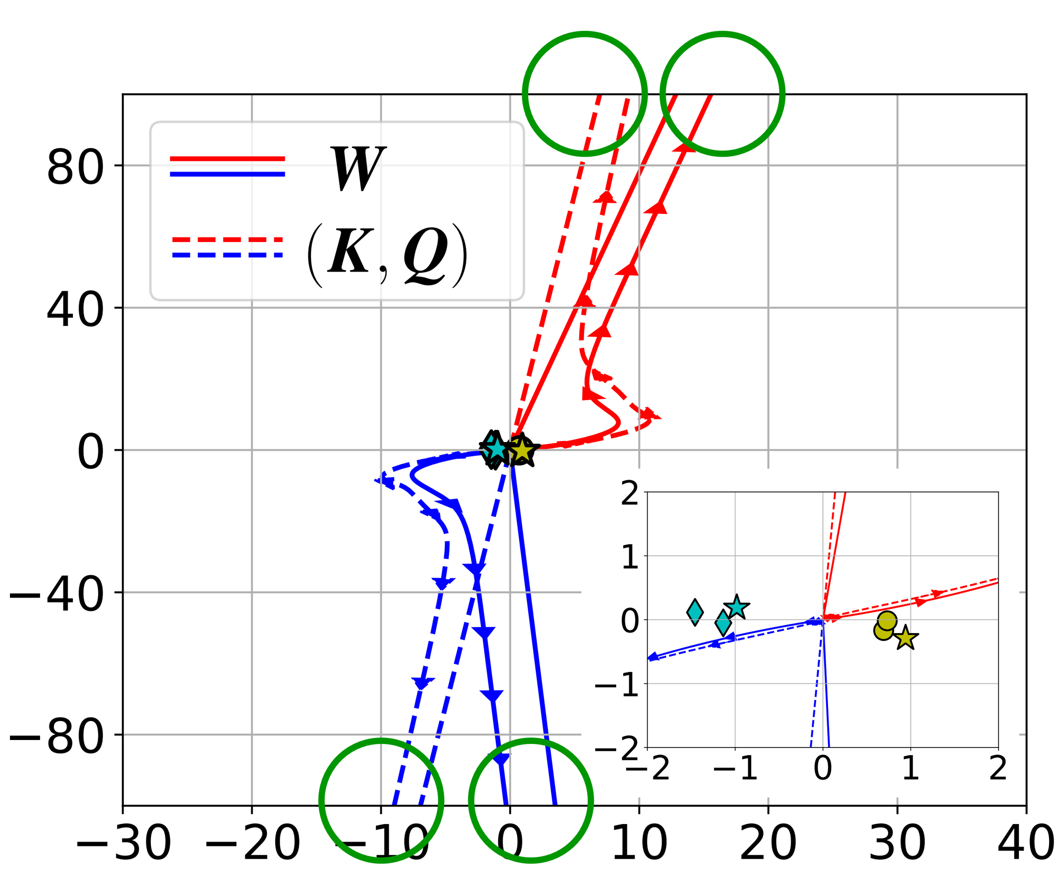
Assumptions: Over any bounded interval [a,b]:

1. $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing; and
2. ℓ' is Lipschitz continuous and bounded.

Theorem I: Convergence of Gradient Descent

Under Assumptions 1&2 and $h(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$, GD-W with a fixed η and proper starting point satisfy $\lim_{k\to\infty} \|\boldsymbol{W}(k)\|_F = \infty$ and $\lim_{k\to\infty} \boldsymbol{W}(k)/\|\boldsymbol{W}(k)\|_F = \boldsymbol{W}^{mm}/\|\boldsymbol{W}^{mm}\|_F$.

- **Regularized path**: Under Assumptions 1&2 and $h(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$, RP-KQ satisfies $\lim_{R\to\infty} \frac{\boldsymbol{K}_R \boldsymbol{Q}_R^\top}{R} = \frac{\boldsymbol{W}_\star^{mm}}{\|\boldsymbol{W}_\star^{mm}\|_F}$.



- Arrows: GD trajectories.
- Lines: the SVM directions mapped to $\boldsymbol{z}$, e.g., $\boldsymbol{W}\boldsymbol{z}$.

Implicit Bias of Attention with Nonlinear Head

Q: *What is the implicit bias and the form of $\boldsymbol{W}(k)$ when the GD solution is composed by multiple tokens?*

General SVM

$$\boldsymbol{W}(k) \approx \boldsymbol{W}^{fin} + \|\boldsymbol{W}(k)\|_F \cdot \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}$$

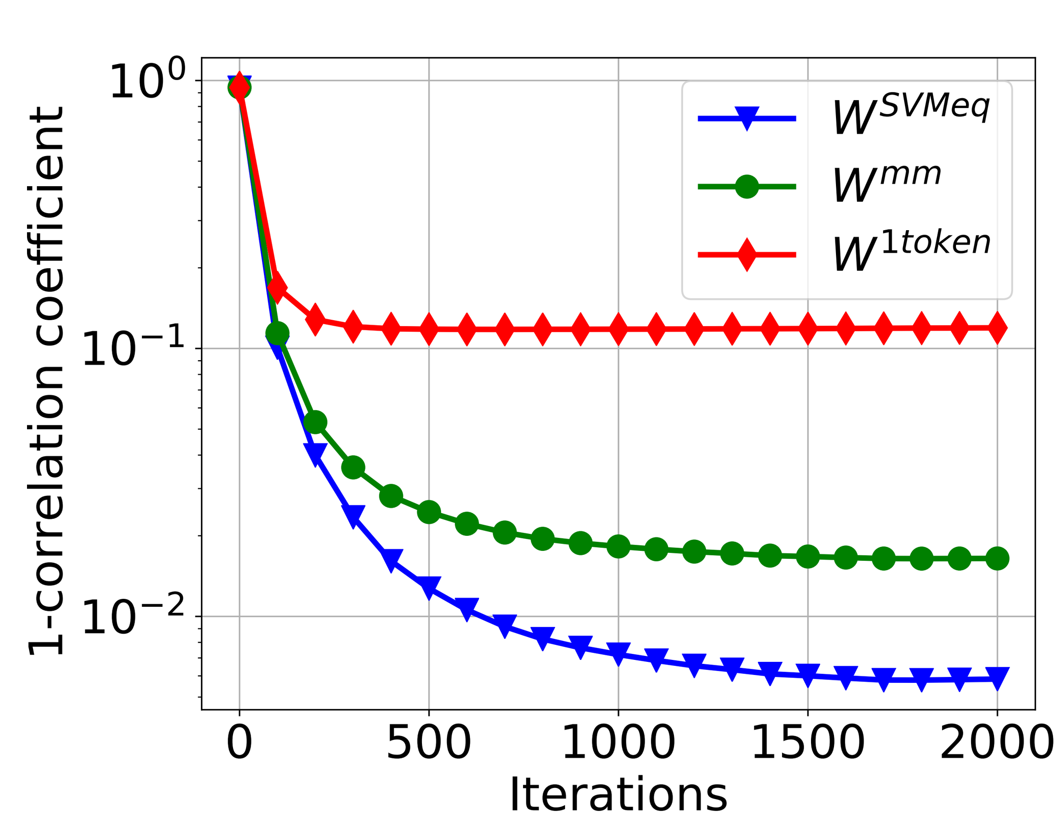
Finite component ($\boldsymbol{W}^{fin}$):

$$(\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W}^{fin} \boldsymbol{z}_i = \log(s_{it}/s_{i\tau}) \quad \forall t, \tau \in \mathcal{O}_i, \ i \in [n].$$

Directional component ($\boldsymbol{W}^{mm}$):

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

- $\mathcal{O}_i, i \in [n]$: Sets of relevant tokens.
- $s_{it}, t \in \mathcal{O}_i$: Assigned softmax probabilities.



$$\boldsymbol{W}^{SVMeq} = \boldsymbol{W}^{fin} + C \cdot \boldsymbol{W}^{mm}$$

where $C = \arg\max \langle \boldsymbol{W}^{SVMeq}, \boldsymbol{W}^{GD} \rangle$.