On the Optimization and Generalization of Two-layer Transformers with Sign Gradient Descent

Bingrui Li[1], Wei Huang[2], Andi Han[2], Zhanpeng Zhou[4],
Taiji Suzuki[3,2], Jun Zhu[1], Jianfei Chen[1]

[1]Tsinghua University, [2]RIKEN AIP, [3]University of Tokyo, [4]Shanghai Jiao Tong University

Background

Setup

Optimization Dynamics

Generalization Properties

Transformers (TFs) & Adam

The transformer architecture has led to many state-of-the-art approaches in language and vision

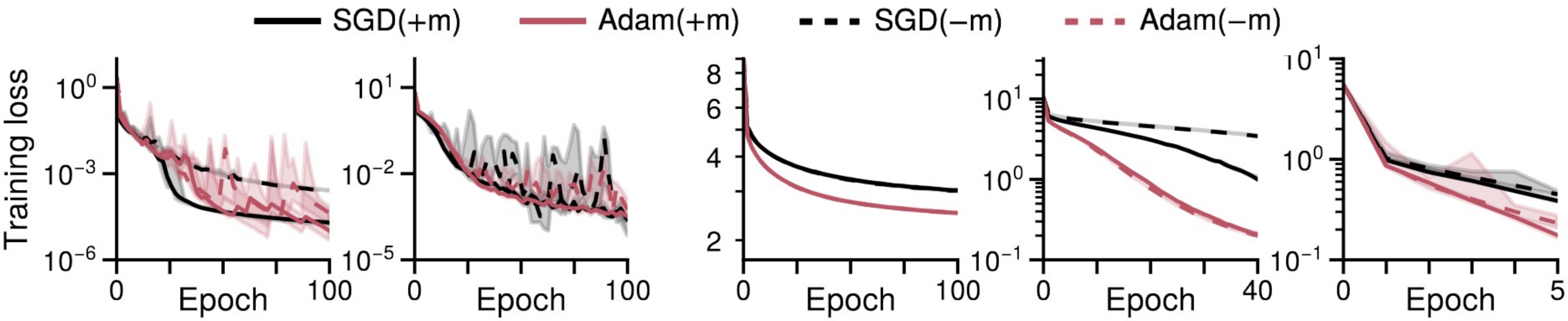
The Adam optimizer has achieved great success in **transformer optimization**, but gradient descent performs poorly.

[1] Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be, Kunstner et al., ICLR 2023

Transformers (TFs) & Adam

The transformer architecture has led to many state-of-the-art approaches in language and vision

The Adam optimizer has achieved great success in **transformer optimization**, but gradient descent performs poorly.



CNN, MNIST | CNN, CIFAR10 | Transformer, PTB | Transformer, WikiText2 | Transformer, SQuAD

Figure 1 in [1]

[1] Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be, Kunstner et al., ICLR 2023

Transformers (TFs) & Adam

In CNNs, SGD is comparable to Adam

In transformers, Adam performs better than SGD

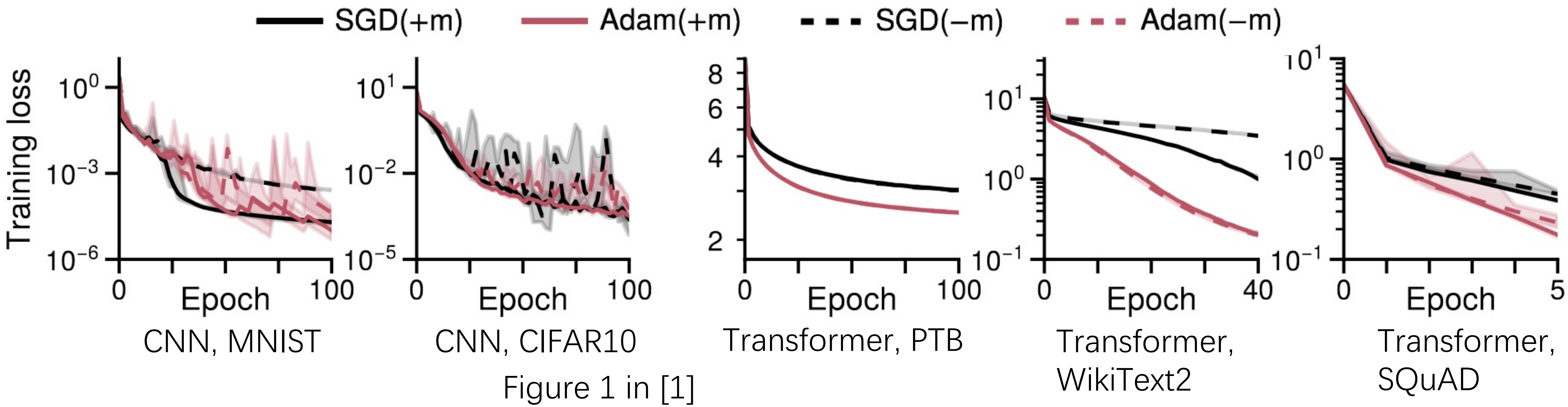


Figure 1 in [1]

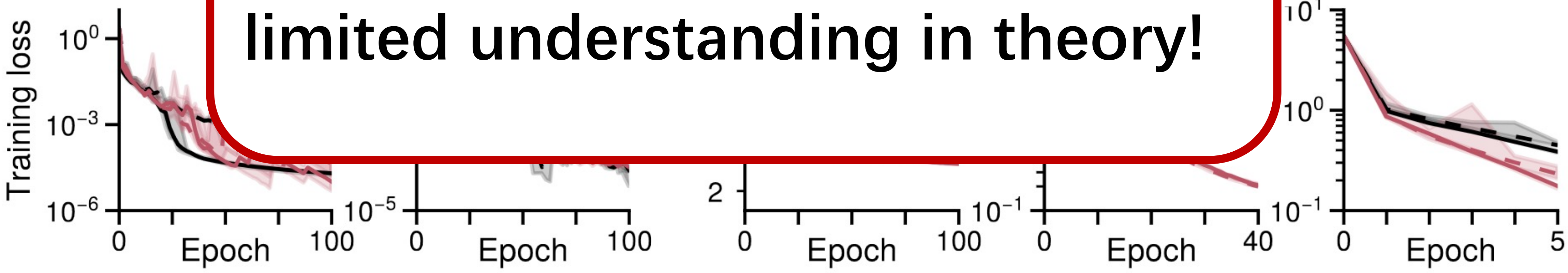
[1] Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be, Kunstner et al., ICLR 2023

Transformers (TFs) & Adam

In CNN … comp …

… Adam … SGD

Transformers + Adam,
great success in practice,
limited understanding in theory!



CNN, MNIST | CNN, CIFAR10 | Transformer, PTB | Transformer, WikiText2 | Transformer, SQuAD

Figure 1 in [1]

[1] Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be, Kunstner et al., ICLR 2023

Our Work

This talk: we give an end-to-end theory of learning **two-layer transformers** by **(Sign Gradient Descent) SignGD** on **signal-noise dataset.**

Optimization dynamics

Generalization properties

Background

Setup

Optimization Dynamics

Generalization Properties

Two-layer Transformers

one self-attention layer + one (fixed) linear layer

Matrix form:

$$f(\mathbf{W},\mathbf{X}) := F_1(\mathbf{W},\mathbf{X}) - F_{-1}(\mathbf{W},\mathbf{X})$$

$$F_j(\mathbf{W},\mathbf{X}) := \frac{1}{m_v}\sum_{l=1}^{2}\mathbf{1}_{m_v}^\top \mathbf{W}_{V,j}\mathbf{X}\mathrm{softmax}\left(\mathbf{X}^\top\mathbf{W}_K^\top\mathbf{W}_Q\mathbf{x}^{(l)}\right)$$

Scalar form (equivalently):

$$F_j(\mathbf{W},\mathbf{X}) = \frac{1}{m_v}\sum_{r\in[m_v]}\left[(s_{11}+s_{21})\left\langle \mathbf{w}_{V,j,r},\mathbf{x}^{(1)}\right\rangle + (s_{12}+s_{22})\left\langle \mathbf{w}_{V,j,r},\mathbf{x}^{(2)}\right\rangle\right]$$

Two-layer Transformers

one self-attention layer + one (fixed) linear layer

Matrix form:

$$f(\mathbf{W},\mathbf{X}) := F_1(\mathbf{W},\mathbf{X}) - F_{-1}$$

$$F_j(\mathbf{W},\mathbf{X}) := \frac{1}{m_v}\sum_{l=1}^{2}\mathbf{1}_{m_v}^{\top}\mathbf{W}_{V,j}$$

Advantage:
softmax attention ✅
gaussian initialization ✅
query-key parameterization ✅

Scalar form (equivalently):

$$F_j(\mathbf{W},\mathbf{X}) = \frac{1}{m_v}\sum_{r\in[m_v]}\left[(s_{11}+s_{21})\left\langle \mathbf{w}_{V,j,r},\mathbf{x}^{(1)}\right\rangle + (s_{12}+s_{22})\left\langle \mathbf{w}_{V,j,r},\mathbf{x}^{(2)}\right\rangle\right]$$

Data model: Signal-Noise dataset

Data model is inspired by image classification problems

For each data point (X, y),

- $\mathrm{X} = [x^{(1)}, x^{(2)}]^T \in \mathbb{R}^{2\times d}$ (2 tokens in $\mathbb{R}^d$), $y \sim \mathrm{Unif}(\{\pm 1\})$

signal patch:
$\mu = [1, 0, \ldots, 0]^T$,

noise patch:
ξ is sparse
and gaussian

Remark:
1. low SNR setting
2. context length is 2

Training algorithm: SignGD

Cross-entropy loss

$$L_S(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell[y_i \cdot f(\mathbf{W}, \mathbf{x}_i)],$$

where $l = \log(1 + \exp(-x))$ is the logistic function.

Sign gradient descent

Learning rate

$$\mathbf{w}_{V,j,r}^{(t+1)} = \mathbf{w}_{V,j,r}^{(t)} - \boxed{\eta}\, \mathrm{sgn}(\nabla_{\mathbf{w}_{V,j,r}} L_S(\mathbf{W}^{(t)})),$$

$$\mathbf{w}_{Q,s}^{(t+1)} = \mathbf{w}_{Q,s}^{(t)} - \eta\, \mathrm{sgn}(\nabla_{\mathbf{w}_{Q,s}} L_S(\mathbf{W}^{(t)})), \quad \mathbf{w}_{K,s}^{(t+1)} = \mathbf{w}_{K,s}^{(t)} - \eta\, \mathrm{sgn}(\nabla_{\mathbf{w}_{K,s}} L_S(\mathbf{W}^{(t)})),$$

SignGD: A Good Surrogate for Adam

1. SignGD is exactly Adam with β1 = β2 = ε = 0 in formulation

$$\mathbf{SignGD}(\mathbf{w}_{t-1}, \mathbf{g}_t) : \mathbf{w}_t = \mathbf{w}_{t-1} - \eta \cdot \operatorname{sgn}(\mathbf{w}_{t-1}),$$

$$\mathbf{Adam}(\mathbf{w}_{t-1}, \mathbf{m}_{t-1}, \mathbf{v}_{t-1}, \mathbf{g}_t) = \begin{cases} \mathbf{m}_t & = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t \\ \mathbf{v}_t & = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2 \\ \hat{\mathbf{m}}_t & = \mathbf{m}_t / (1-\beta_1^t) \\ \hat{\mathbf{v}}_t & = \mathbf{v}_t / (1-\beta_2^t) \\ \mathbf{w}_t & = \mathbf{w}_{t-1} - \eta \cdot \hat{\mathbf{m}}_t / \left(\sqrt{\hat{\mathbf{v}}_t} + \epsilon\right) \end{cases}$$

SignGD: A Good Surrogate for Adam

2. SignGD (with momentum), i.e., the Lion optimizer [2], can perform well on deep learning tasks

Table 5: One-shot evaluation averaged over three NLG and 21 NLU tasks. The results of GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022) are included for reference. The LLMs trained by Lion have better in-context learning ability. See Table 11 (in the Appendix) for detailed results on all tasks.

Task	1.1B Adafactor	1.1B Lion	2.1B Adafactor	2.1B Lion	7.5B Adafactor	7.5B Lion	6.7B GPT-3	8B PaLM
#Tokens	300B						300B	780B
Avg NLG	11.1	**12.1**	15.6	**16.5**	24.1	**24.7**	23.1	23.9
Avg NLU	53.2	**53.9**	56.8	**57.4**	61.3	**61.7**	58.5	59.4

[2] Symbolic Discovery of Optimization Algorithms, Chen et al., NeurIPS 2023

SignGD: A Good Surrogate for Adam

3. SignGD can recover the superior performance of Adam in the full-batch/deterministic setting [1]

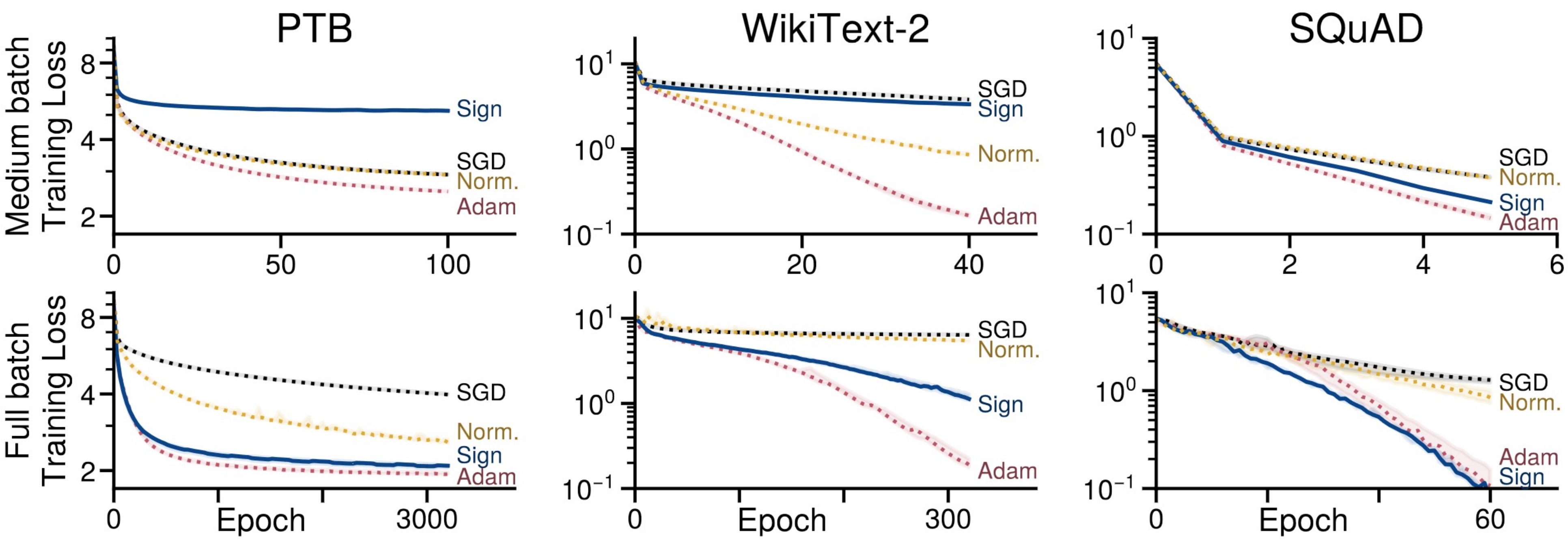


Figure 7 in [1]

[1] Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be, Kunstner et al., ICLR 2023

SignGD: A Good Surrogate for Adam

3. SignGD can recover the superior performance of Adam in the full-batch/deterministic setting [1]

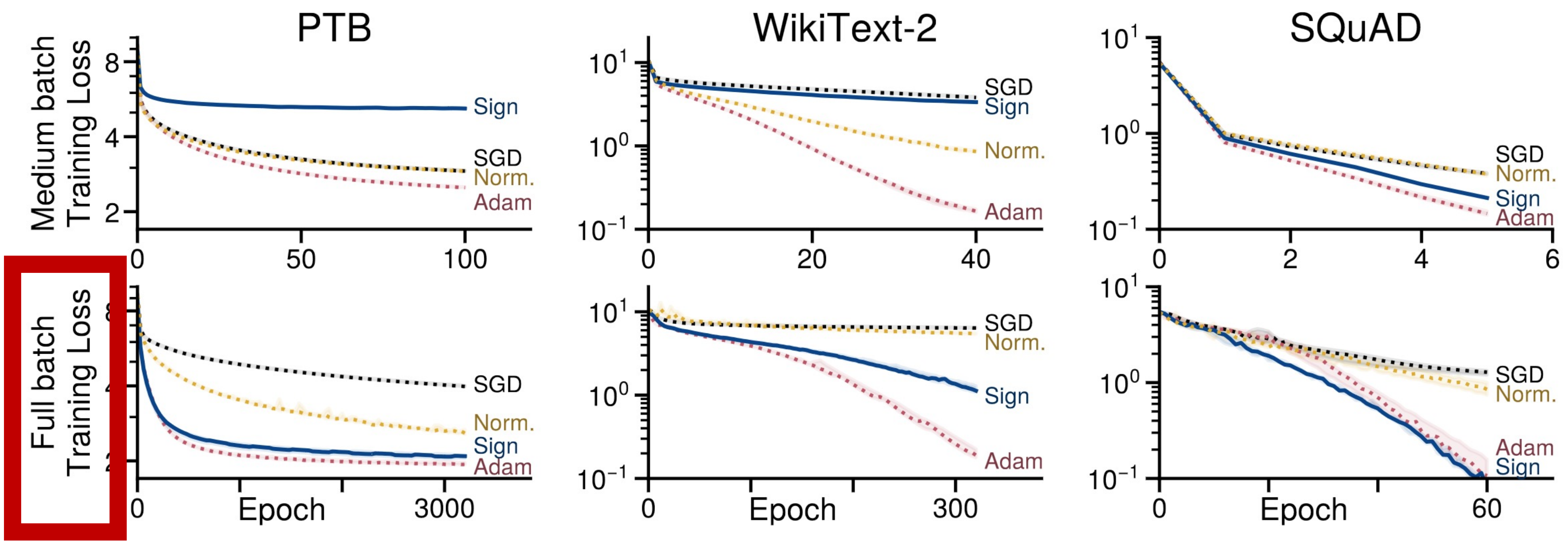


Figure 7 in [1]

[1] Noise Is Not the Main Factor Behind the Gap Between Sgd and Adam on Transformers, But Sign Descent Might Be, Kunstner et al., ICLR 2023

Background

Setup

Optimization Dynamics

Generalization Properties

$$F_j(\mathbf{W},\mathbf{X}) = \frac{1}{m_v}\sum_{r\in[m_v]} \left[(s_{11}+s_{21})\left\langle \mathbf{w}_{V,j,r}, \mathbf{x}^{(1)}\right\rangle + (s_{12}+s_{22})\left\langle \mathbf{w}_{V,j,r}, \mathbf{x}^{(2)}\right\rangle\right]$$

What do we focus on?

1. **data-parameter inner product:**

<p, d>

p = query, key, value

d = signal, noise

$< w_{V,j,r}^{(t)}, \mu >$ value signal

$< w_{V,j,r}^{(t)}, y_i\xi_i >$ value noise

$< w_{K,s}^{(t)}, y_i\xi_i >$ key noise

$< w_{K,s}^{(t)}, \mu >$ key signal

$< w_{Q,s}^{(t)}, y_i\xi_i >$ query noise

$< w_{Q,\boxed{s}}^{(t)}, \mu >$ query signal

r/s is the index of neurons

$$F_j(\mathbf{W},\mathbf{X}) = \frac{1}{m_v}\sum_{r\in[m_v]}\left[(s_{11}+s_{21})\left\langle \mathbf{w}_{V,j,r},\mathbf{x}^{(1)}\right\rangle + (s_{12}+s_{22})\left\langle \mathbf{w}_{V,j,r},\mathbf{x}^{(2)}\right\rangle\right]$$

What do we focus on?

2. **softmax outputs**

e.g. noise-signal softmax output(s)

$$s_{i,21}^{(t)} = \frac{\exp\left(\sum_{s\in[m_k]} \underbrace{\boxed{\langle \mathbf{w}_{Q,s}^{(t)}, \boldsymbol{\xi}_i\rangle}}_{\text{noise}} \underbrace{\boxed{\langle \mathbf{w}_{K,s}^{(t)}, y_i\boldsymbol{\mu}\rangle}}_{\text{signal}}\right)}{\exp\left(\sum_{s\in[m_k]}\langle \mathbf{w}_{Q,s}^{(t)}, \boldsymbol{\xi}_i\rangle\langle \mathbf{w}_{K,s}^{(t)}, y_i\boldsymbol{\mu}\rangle\right) + \exp\left(\sum_{s\in[m_k]}\langle \mathbf{w}_{Q,s}^{(t)}, \boldsymbol{\xi}_i\rangle\langle \mathbf{w}_{K,s}^{(t)}, \boldsymbol{\xi}_i\rangle\right)}.$$

Main Results: Four stage dynamics

Stage I. The **mean value noise** shifts early, then stabilizes.

Stage II. The **query & key noise** align their sign to each other.

Stage III. Majority voting determines the sign of **query & key signals**.

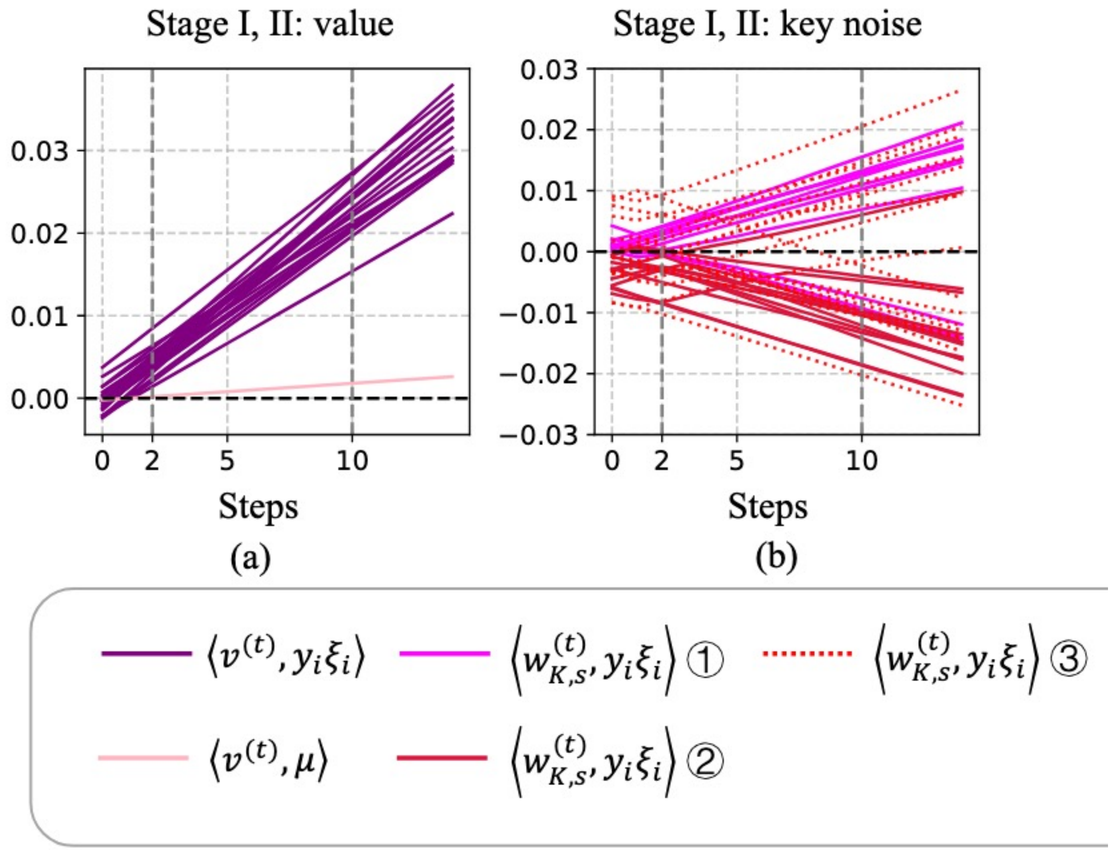
Stage IV. The **noise-signal softmax outputs** decay fast exponentially, then the **query & key noise** align their sign to signals.

Stage I: The mean value noise shifts early, then stabilizes. $(t = 0 \sim t = 2)$

Define mean value

$$v^{(t)} = \frac{1}{m_v} \sum_r w_{V,1,r}^{(t)} - w_{V,-1,r}^{(t)}$$

1. $< v^{(t)}, y_i \xi_i >$ increases monotonically, and stabilizes into a linear relationship with t.
2. other quantities stay close to initialization (mean value signal become neglectable)



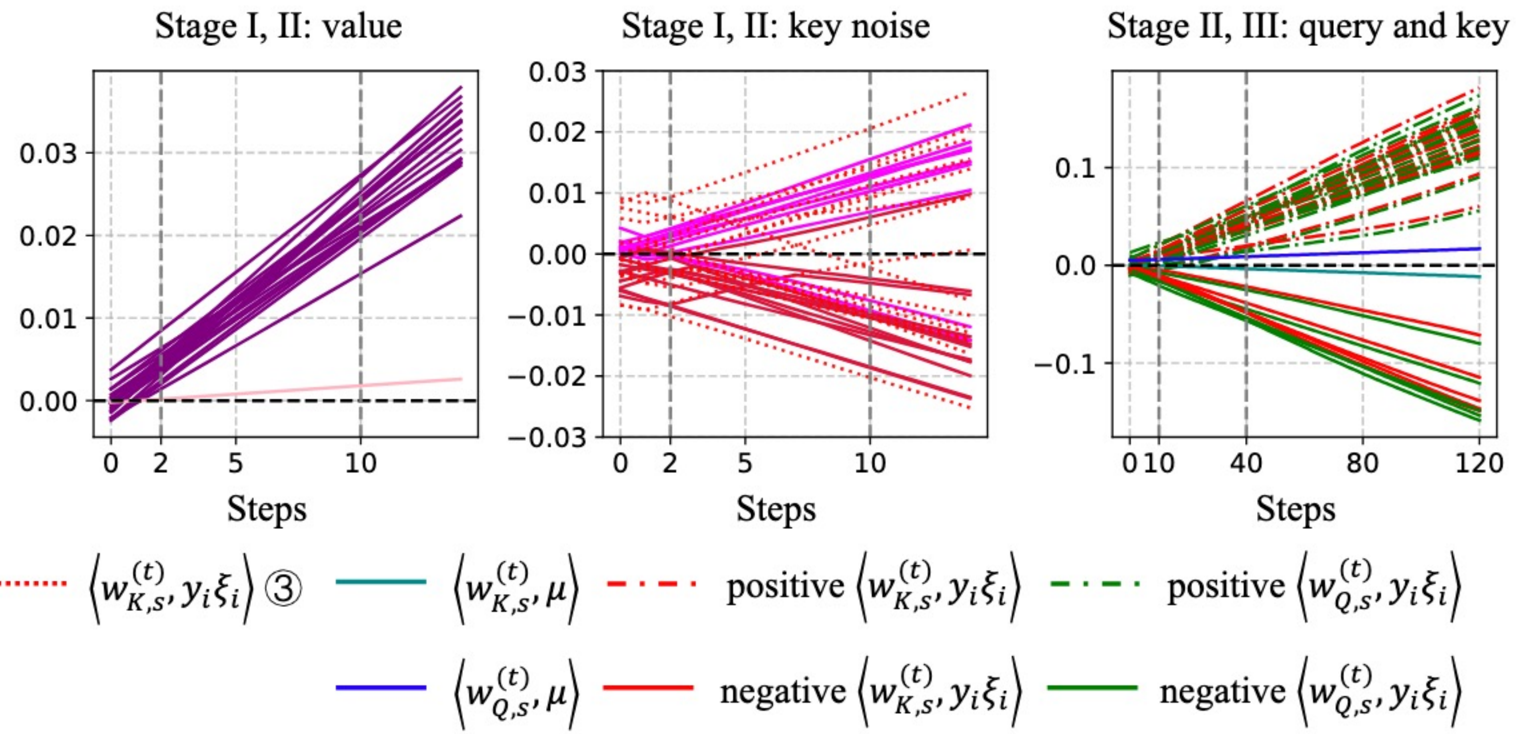
Stage II: The **query & key noise** align their sign to each other $(t = 2 \sim t = 10)$

1. For one neuron, sign alignment between $< w_{K,s}^{(t)}, y_i \xi_i >$ and $< w_{Q,s}^{(t)}, y_i \xi_i >$.

2. For all neurons (over s and i), the number of positive and negative neurons is nearly equal.

3. mean value noise continue to grow.

4. query and key feature stay close to initialization.



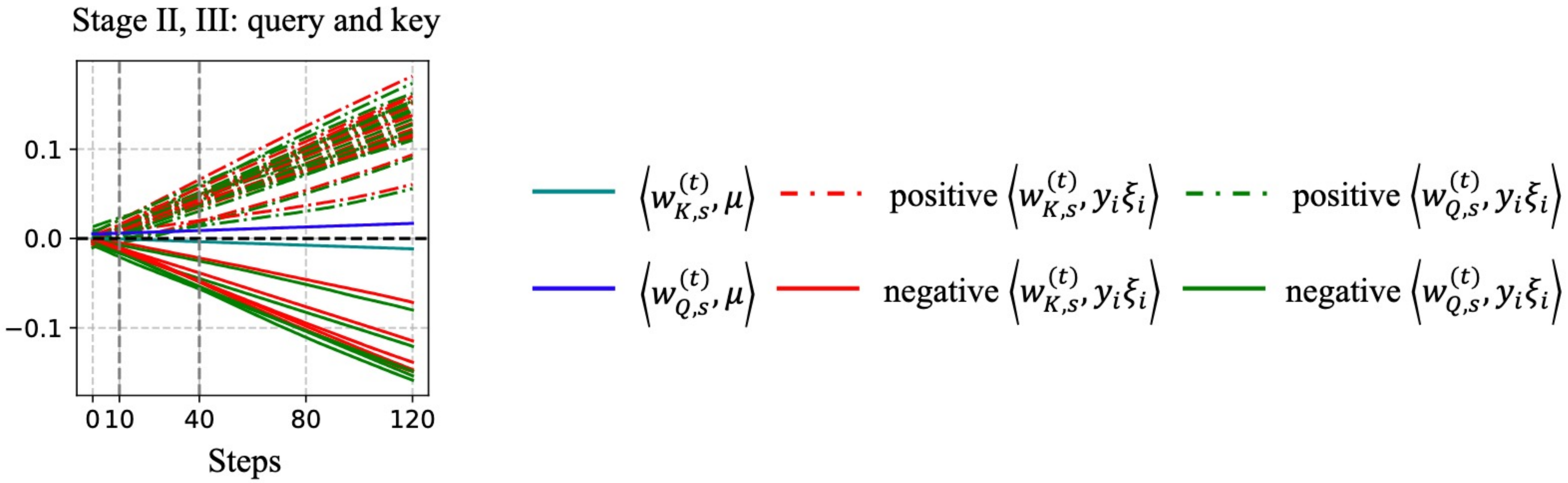
Stage II: sign alignment

Table 1: Sign alignment between query and key noise. $S^{(t)}_{K+,Q+}$ defined as $S^{(t)}_{K+,Q+} := \{(s,i) \in [m_k] \times [n] : \boxed{\langle \mathbf{w}^{(t)}_{K,s}, y_i \boldsymbol{\xi}_i \rangle > 0, \langle \mathbf{w}^{(t)}_{Q,s}, y_i \boldsymbol{\xi}_i \rangle > 0}\}$ represents the number of neurons and samples having positive query noise and positive key noise. The definitions for $S^{(t)}_{K+,Q-}$, $S^{(t)}_{K-,Q+}$, $S^{(t)}_{K-,Q-}$ are similar. Each element in the middle of the table represents the size of the intersection of the corresponding row set and the corresponding column set. For example, $|S^{(0)}_{K+,Q+} \cap S^{(t)}_{K+,Q+}| = 486$. The signs of query and key noise are independent at initialization but aligned at $t = 10$, which can be seen as an estimate of T_2^{SGN}.

init$(t=0)\backslash t=10$	$\lvert S^{(t)}_{K+,Q+}\rvert$	$\lvert S^{(t)}_{K+,Q-}\rvert$	$\lvert S^{(t)}_{K-,Q+}\rvert$	$\lvert S^{(t)}_{K-,Q-}\rvert$	Row sum
$\lvert S^{(0)}_{K+,Q+}\rvert$	486	1	0	25	512
$\lvert S^{(0)}_{K+,Q-}\rvert$	244	4	9	250	507
$\lvert S^{(0)}_{K-,Q+}\rvert$	223	10	4	221	458
$\lvert S^{(0)}_{K-,Q-}\rvert$	37	2	3	481	523
Column sum	990	17	16	977	2000

Stage III: Majority voting determines the sign of query & key signals. $(t = 10 \sim t = 40)$

1. The update direction of $< w_{Q,s}^{(t)}, \mu >$ is determined by $\sum_s < w_{K,s}^{(t)}, y_i \xi_i >$
2. $< w_{K,s}^{(t)}, \mu >$ is determined by $\sum_s < w_{Q,s}^{(t)}, y_i \xi_i >$, with an opposite sign to $< w_{Q,s}^{(t)}, \mu >$.
3. The dynamics of other quantities keep unchanged.

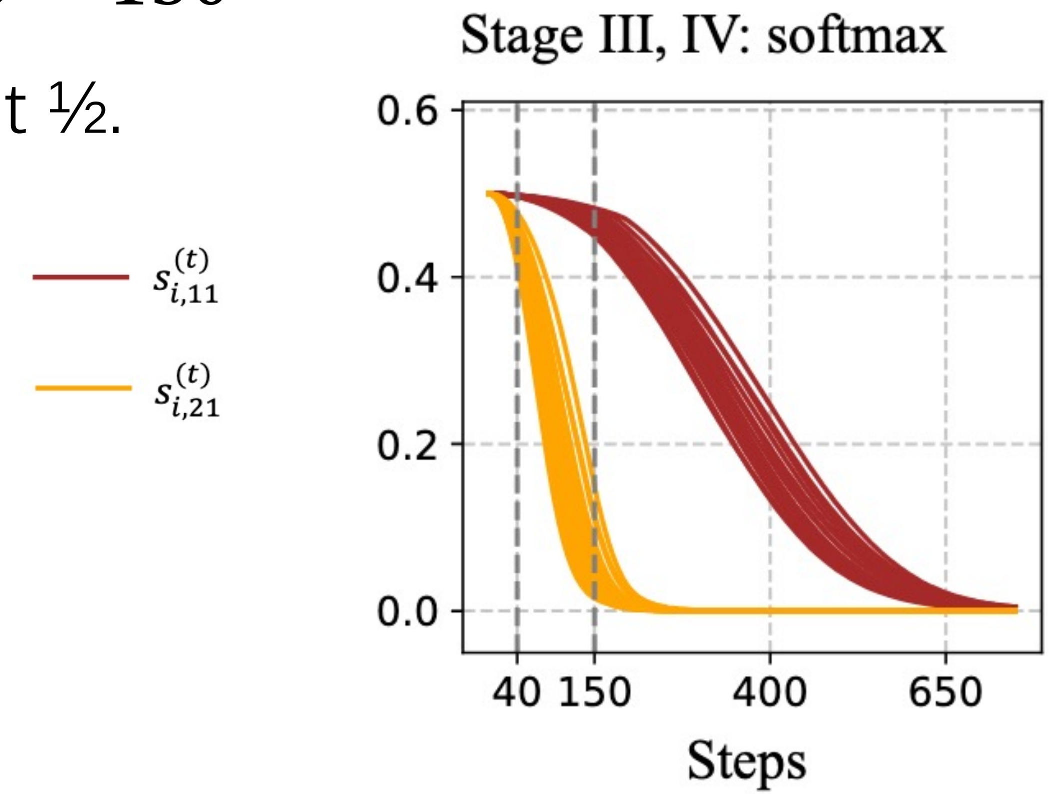


Stage IV: The **noise-signal softmax outputs** decay fast $(t = 40 \sim t = 2k)$

1. Before Stage IV, all softmax outputs are concentrated at ½.

2. In Stage IV, noise-feature softmax outputs $s_{i,21}^{(t)}$ first leave ½, decrease **exponentially**, and reach o(1) at $t = 150$

3. while $s_{i,11}^{(t)}$ is still concentrated at ½.



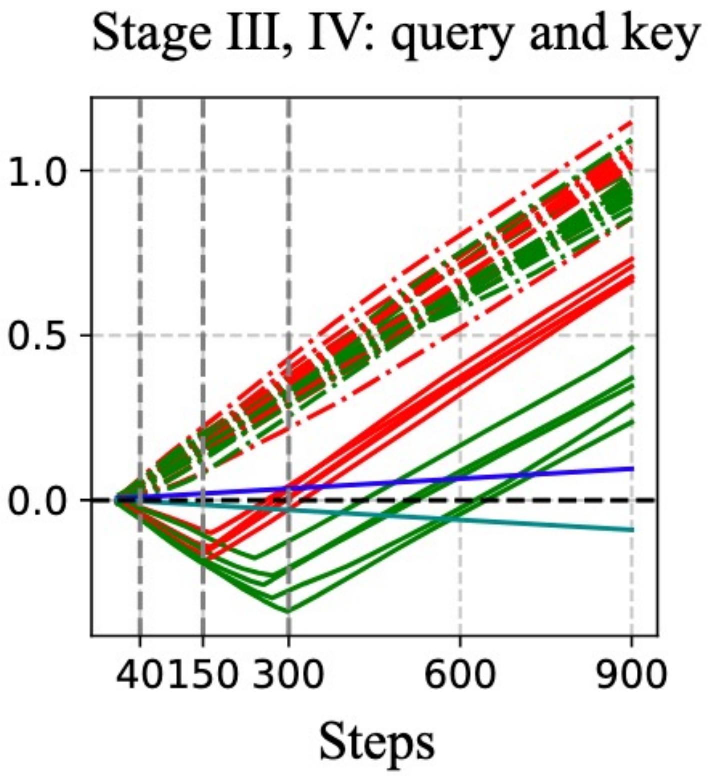
Stage IV: The **query & key noise** align their sign to signals. $(t = 40 \sim t = 2k)$

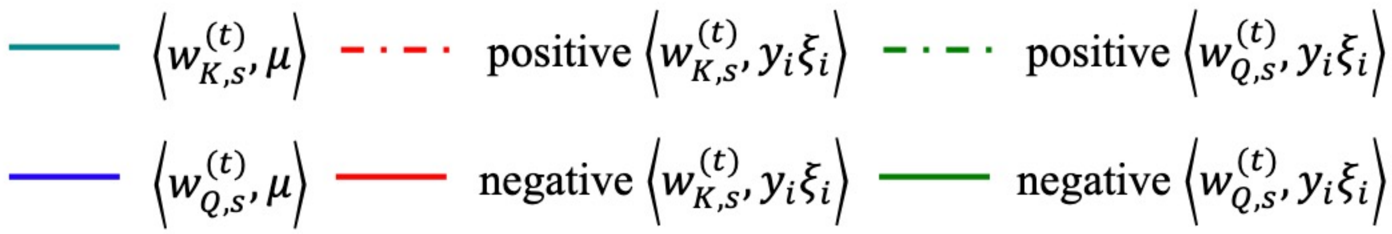
1. From Stage IV on, the signs of query & key signals are fixed (positive query signal).

2. When softmax outputs are small enough $(t = 150)$, all negative key noise begins to align with the positive query signal,

3. When negative key noise completes alignment $(t = 300)$ all negative query noise begins to align with the positive query signal.

4. The alignment of negative query noise completes before the end of this stage $(t = 600)$.





Main Results: Four stage dynamics

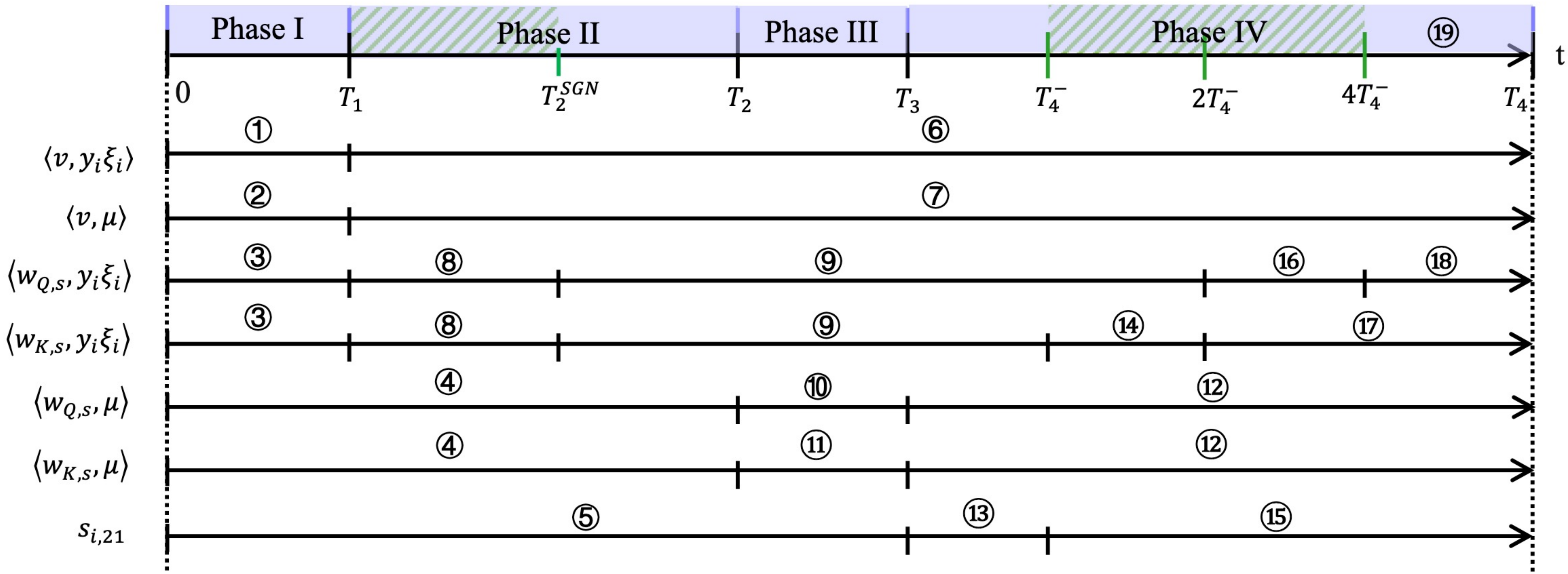


Figure 19 in the paper

Background

Setup

Optimization Dynamics

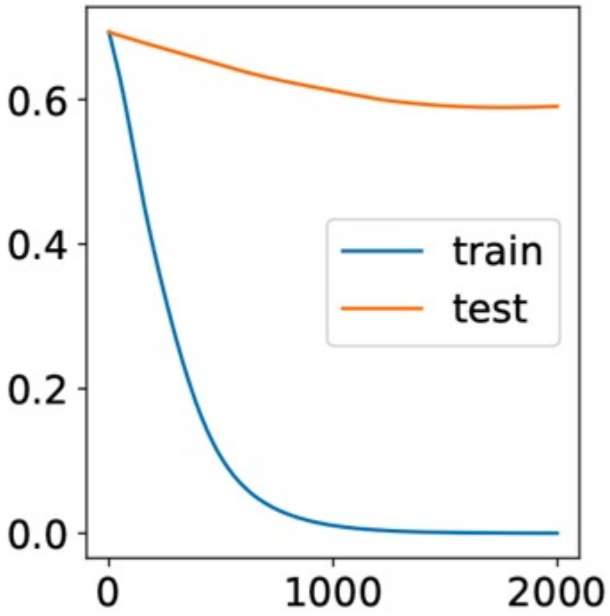
Generalization Properties

Fast Optimization but Poor Generalization

Theorem For any $\epsilon > 0$, there exists $T = O(\log(1/\epsilon))$, T_{attn} such that

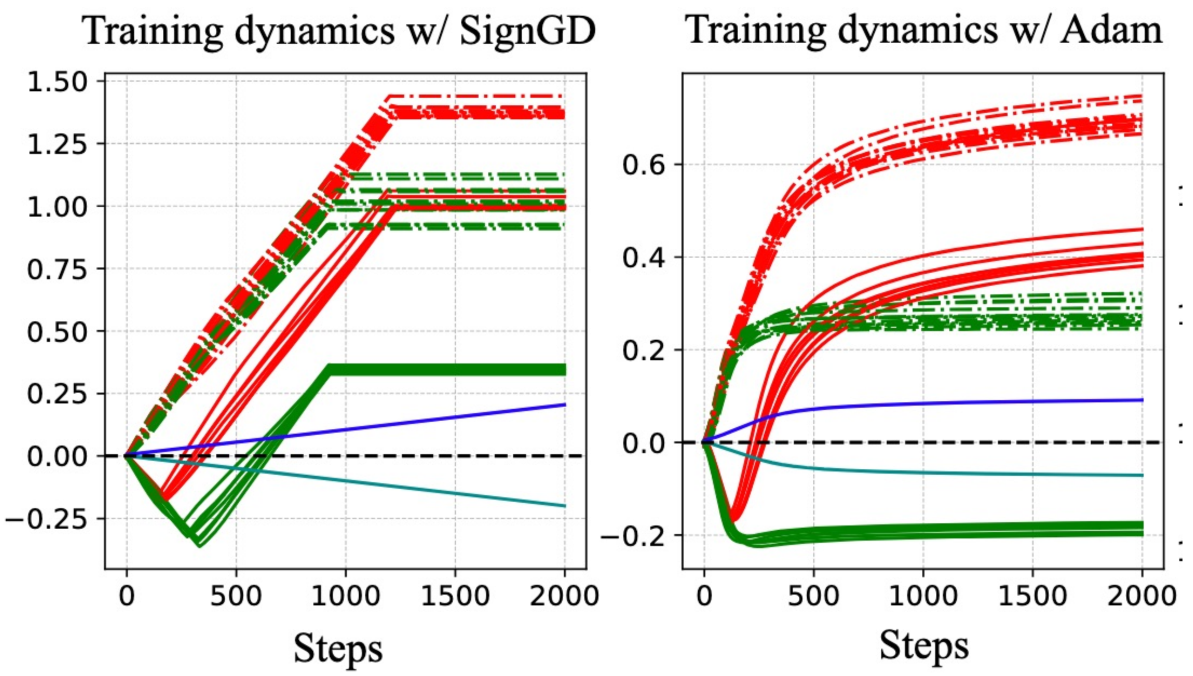
- **[Training loss]** The training loss converges to ϵ: $\mathrm{L_S}\big(\mathrm{W}^{(T)}\big) \leq \epsilon$.
- **[Test loss]** The trained transformer has a constant order test loss: $\mathrm{L_D}\big(\mathrm{W}^{(T)}\big) = \Theta(1)$.
- **[Noise memorization of query & key]** The value in attention layer memorizes noises in training data $< v^{(T)}, y_i\xi_i > = \Omega(1)$,$< v^{(T)}, \mu >= o(1)$
- **[Noise memorization of query & key]** The attention layer attends all to noise patch

$s_{i,21}^{(T_{\text{attn}})} = o(1)$, $s_{i,11}^{(T_{\text{attn}})} = o(1)$.

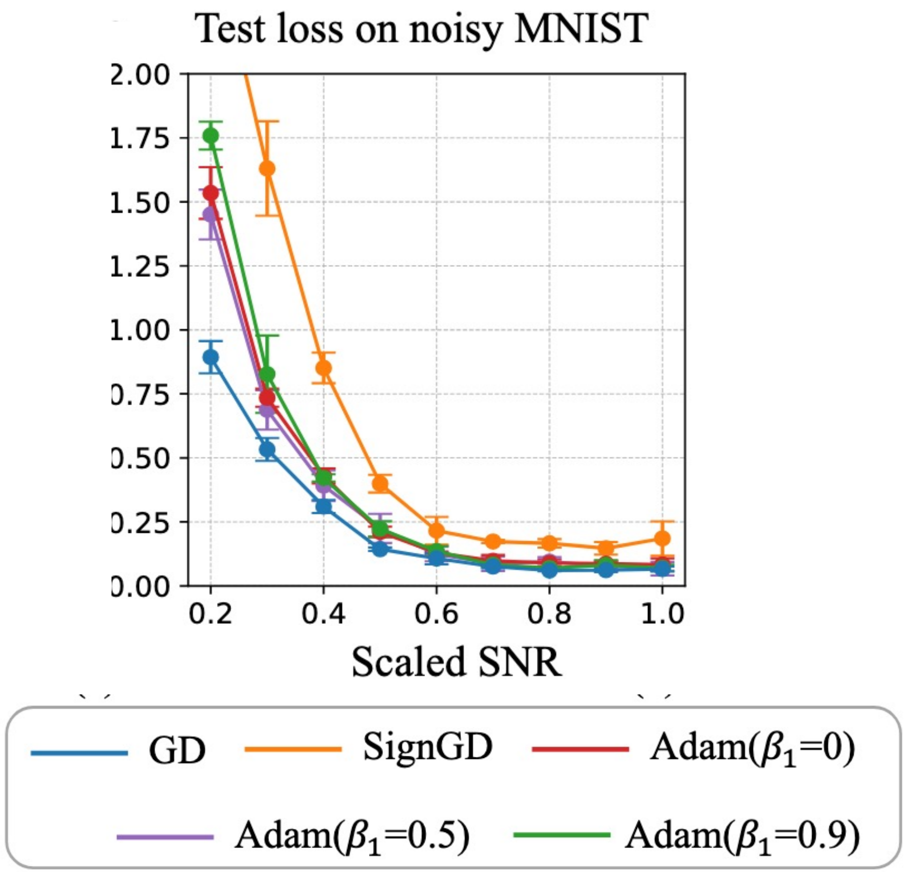


Adam vs SignGD: Similarity

1. training dynamics on synthetic data (**optimization**)

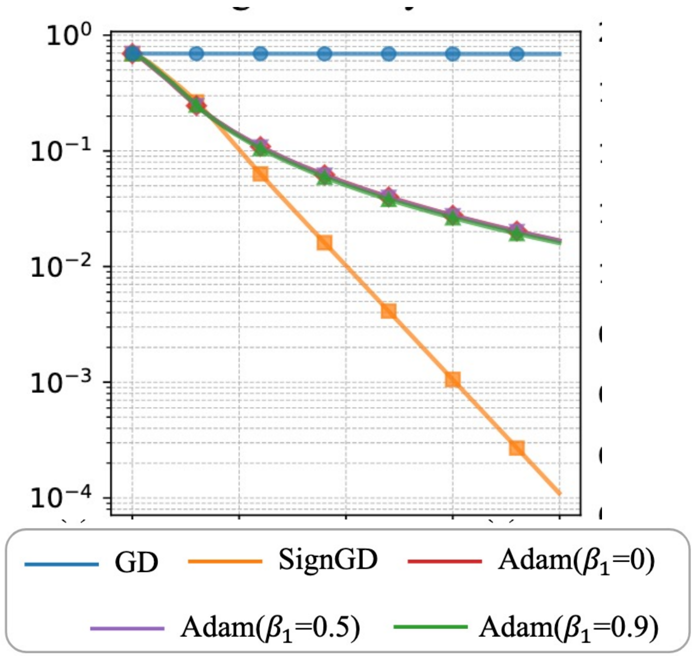


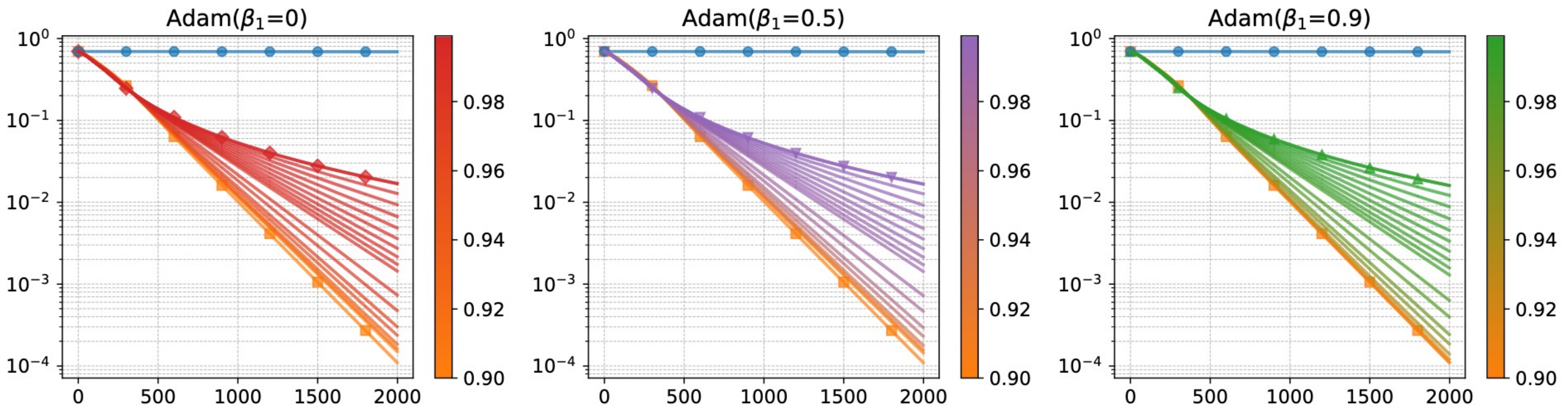
2. test loss on noisy MNIST data (**generalization**)



Adam vs SignGD: Disparity

1. training loss on synthetic data

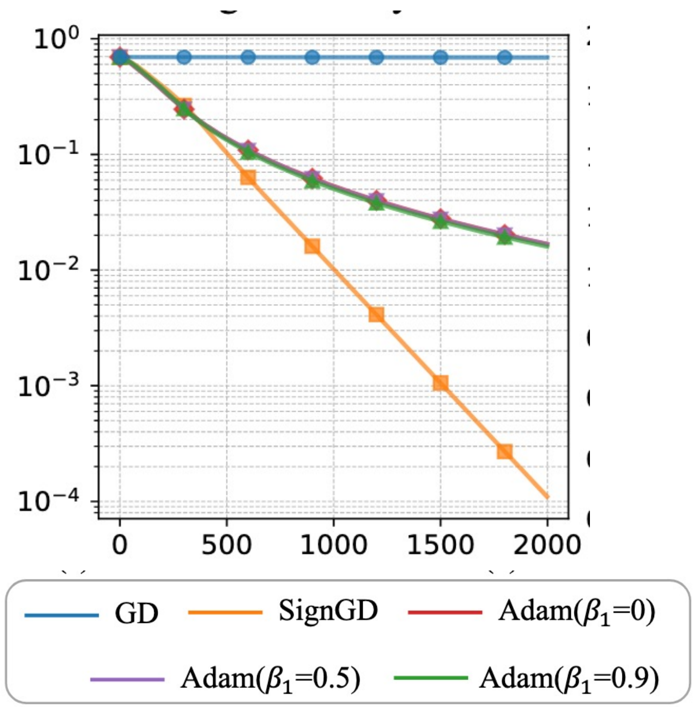




GD vs SignGD

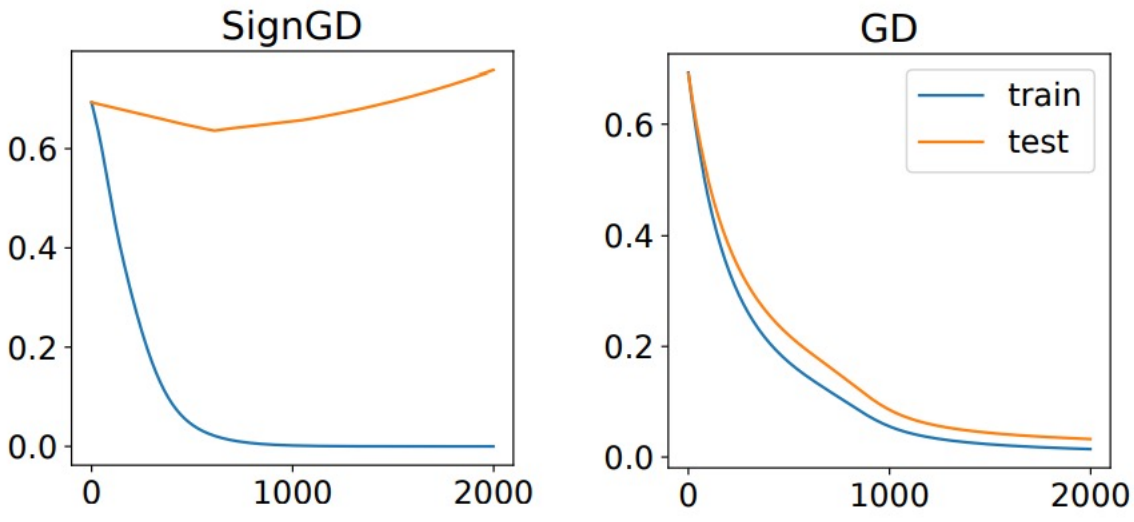
At the same signal-noise ratio,

SignGD trains faster



training loss on synthetic data

but GD generalizes better



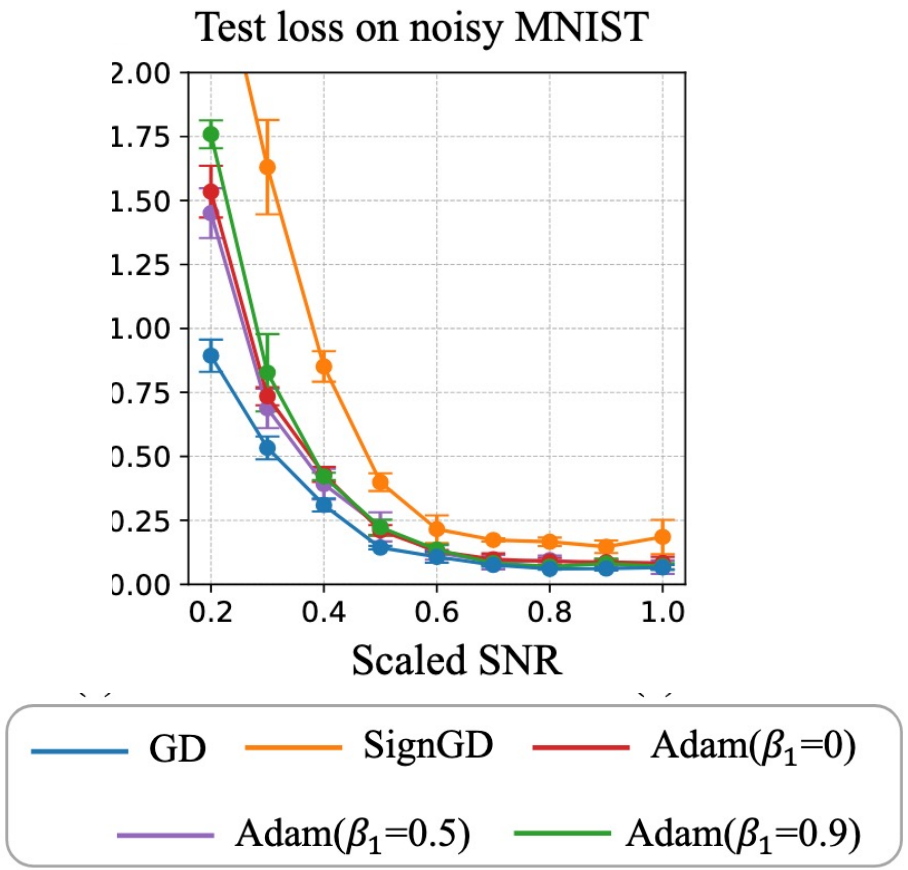
training loss and test loss
on synthetic data

SignGD and Adam require higher data quality than GD

GD vs SignGD

On noisy MNIST data

GD generalizes better
especially when noise is large



SignGD and Adam require higher data quality than GD

Thanks!

Summary

- We give an **end-to-end theory** of learning two-layer transformers by sign gradient descent
- We demonstrate the **four stages in the whole optimization dynamics** including rich alignment behaviors and exponential convergence of softmax outputs
- We prove the **fast convergence and poor generalization** results
- We provide evidence that Adam exhibits similar behaviors to SignGD, and SignGD and Adam require higher data quality

See more formal results in the paper!