Robust Training of Neural Networks Using Scale Invariant Architectures

Zhiyuan Li[1], Srinadh Bhojanapali[2], Manzil Zaheer[3],
Sashank J. Reddi[2], Sanjiv Kumar[2]

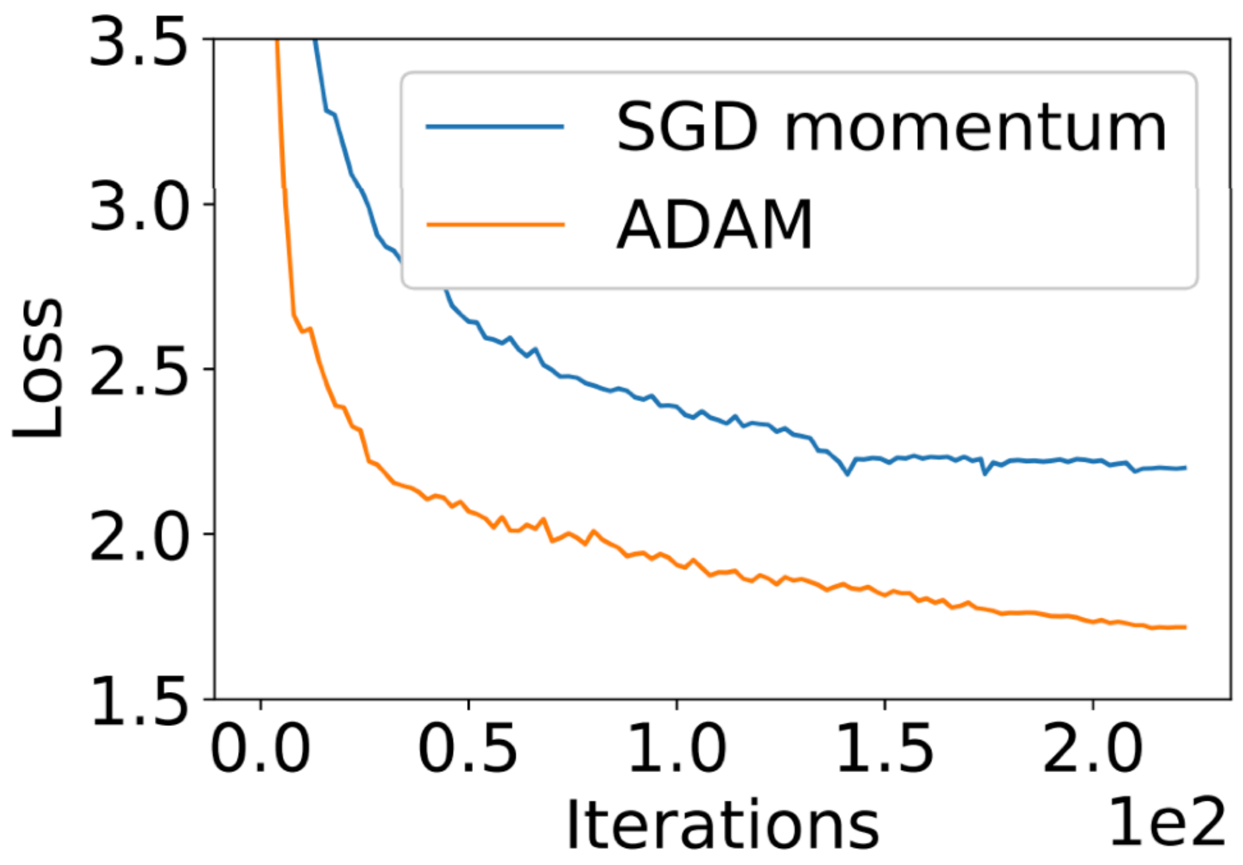
ICML, 2022

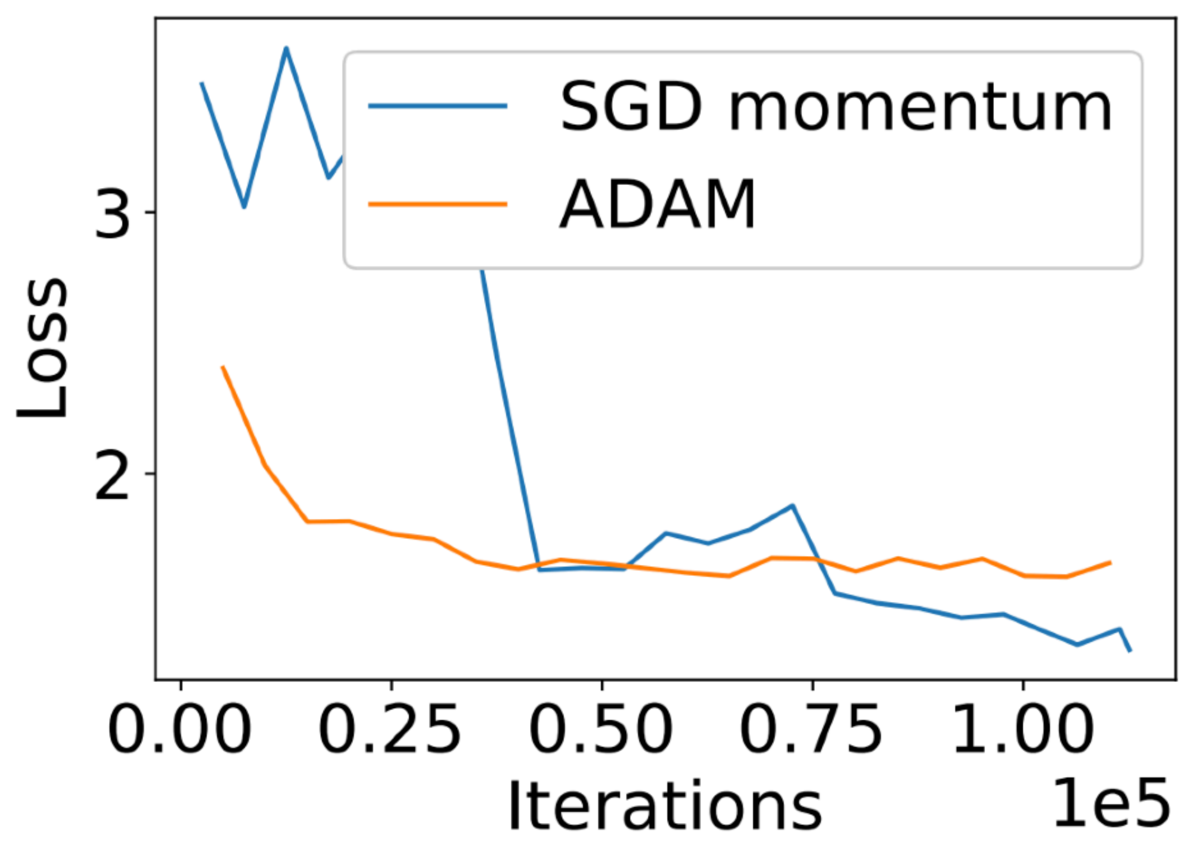
[1] Princeton University, [2] Google Research New York, [3] Google Deepmind New York

Background

- Training Transformers is difficult and often unstable.
- Training Transformer requires adaptive methods (e.g., ADAM), unlike ResNets, which can be trained by SGD.

BERT





ResNet

[Zhang et al.,19]

- What complicates Transformers training?

What Complicates Transformer Training?

More likely to get extremely large gradient

- [Zhang et al.,19]: Heavy-tailed gradient noise from both architecture (Attention) and dataset (text).

  $\Longrightarrow$ Clipping and adaptive LR improves convergence by avoiding huge updates.

- However, heavy-tailed noise may not be the entire answer.
- [You et al., 20, Chen et al.,21]: ADAM consistently outperforms SGD even in **large-batch** or **full-batch** setting in NLP, while the gap is much smaller in vision.

- **Question**: What are other possible issues? Can non-adaptive methods like SGD enjoy fast and robust convergence?

Outline

Our Hypothesis

- k-homogeneity in Architecture

Our Recipe

1. Design Scale Invariant Architecture
2. SGD + WD
3. A Novel Clipping Rule

Experiments

- Train Scale Invariant BERT with SGD

Our Hypothesis: k-homogeneity in Architecture

- **Definition:** $f(x)$ is k-homogeneous $\Longleftrightarrow f(cx) = c^k f(x), \forall c > 0$ (x=param, f(x)= output)
- **Lemma:** f is k-homo $\Longleftrightarrow$ $\nabla^l f$ is $(k - l)$-homo.
- **Descent Lemma:** Learning Rate(LR) $< 2/\text{smoothness}$ $\implies$ GD decreases loss L
- **Observation 1:** $k \geq 3$ $\implies$ model $f(x)$ has unbounded smoothness, so does loss L

  $\implies$ success of LR (for GD) is sensitive to the initialization

Our Hypothesis: k-homogeneity in Architecture

- **Observation 1:** $k \geq 3 \implies$ model $f(x)$ has unbounded smoothness, so does loss L

  $\implies$ success of LR η (for GD) is sensitive to the initialization

1d logistic regression with non-separable data

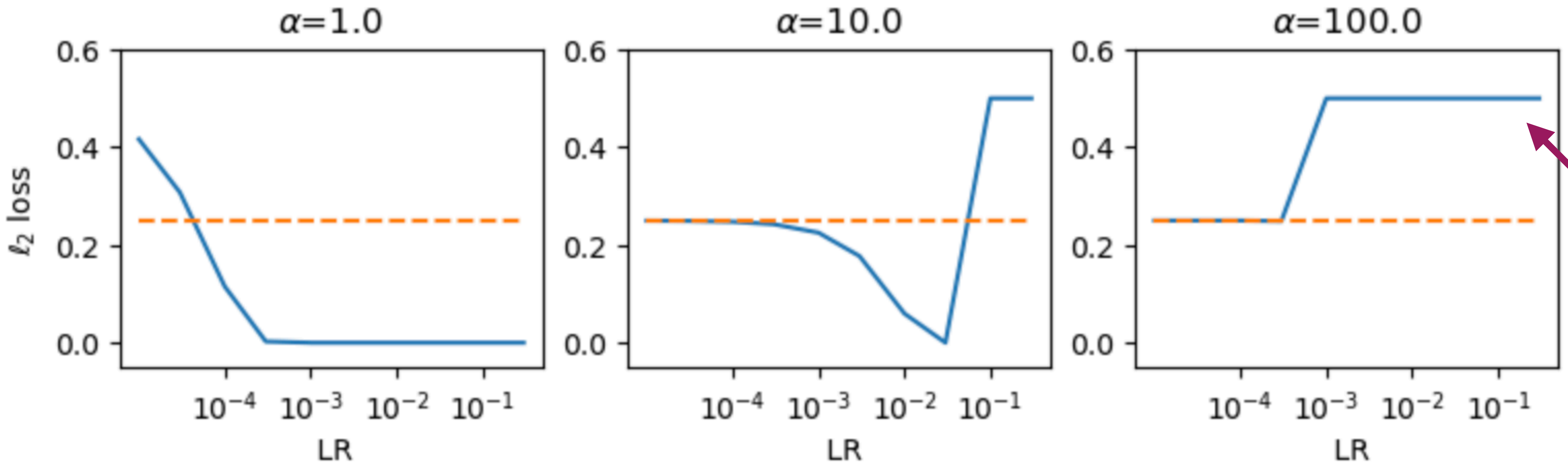
- **Ex 1:** $\tilde{L} : \mathbb{R} \to \mathbb{R}$, convex with bounded smoothness, minimizer $X^* > 0$.
  - There is a sufficiently small LR, that GD on $\tilde{L}$ converges for all init
  - Let $X = x_1 \ldots x_{2k}, k \geq 2, L(x_1, , \ldots, x_{2k}) = \tilde{L}(X) \implies L$ has unbounded smoothness
  - GD on L diverges if LR $\geq \dfrac{2(X(0))^{\frac{1}{k}-1}}{|\nabla \tilde{L}(X(0))|}$, $x_i(0)$ are the same and $X(0) \geq X^*$.

Our Hypothesis: k-homogeneity in Architecture

- **Observation 2**: Even fine-tuned LR cannot learn from unbalanced initialization efficiently.

- **Ex 2: low-rank matrix factorization,** $L(A,B) = \frac{1}{2}\|AB^\top - Y\|_F^2$, $A, B \in \mathbb{R}^{d\times r}$ and $d > r$.
  - For simplicity, assume $r = 1, d = 2$, and $A(0) = \begin{pmatrix} \alpha \\ 0 \end{pmatrix}, B(0) = \begin{pmatrix} \alpha^{-1} \\ 0 \end{pmatrix}, Y = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$

optimal loss when training B only



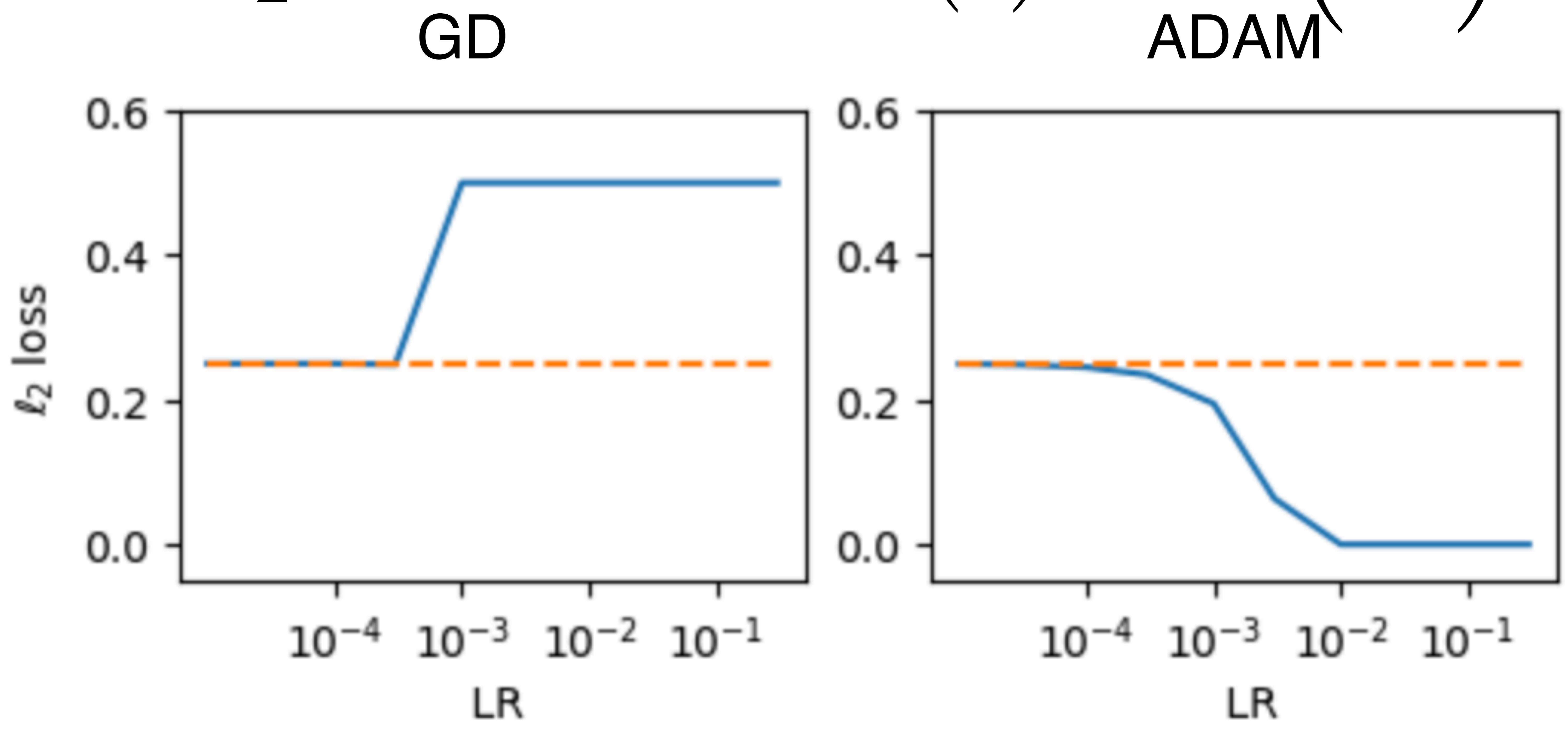
10k steps

Diverge

Our Hypothesis: k-homogeneity in Architecture

- **Ex 2**: $L(A,B) = \frac{1}{2}\|AB^\top - Y\|_F^2,\ A(0) = \begin{pmatrix}\alpha \\ 0\end{pmatrix}, B(0) = \begin{pmatrix}\alpha^{-1} \\ 0\end{pmatrix}, Y = \begin{pmatrix}0.5 & 0.5 \\ 0.5 & 0.5\end{pmatrix}$

$\alpha = 100$



ADAM solves the unbalanced weight issue, but costs 3x memory storing parameters.

Can non-adaptive methods like SGD enjoy fast and robust convergence?

Outline

Our Hypothesis

- k-homogeneity in Architecture

Our Recipe

1. Design Scale Invariant Architecture
2. SGD + WD (no momentum, no warm-up)
3. A Novel Clipping Rule: Relative Global Clipping

Experiments

- Train Scale Invariant BERT with SGD

Outline

Our Hypothesis

- k-homogeneity in Architecture

Our Recipe

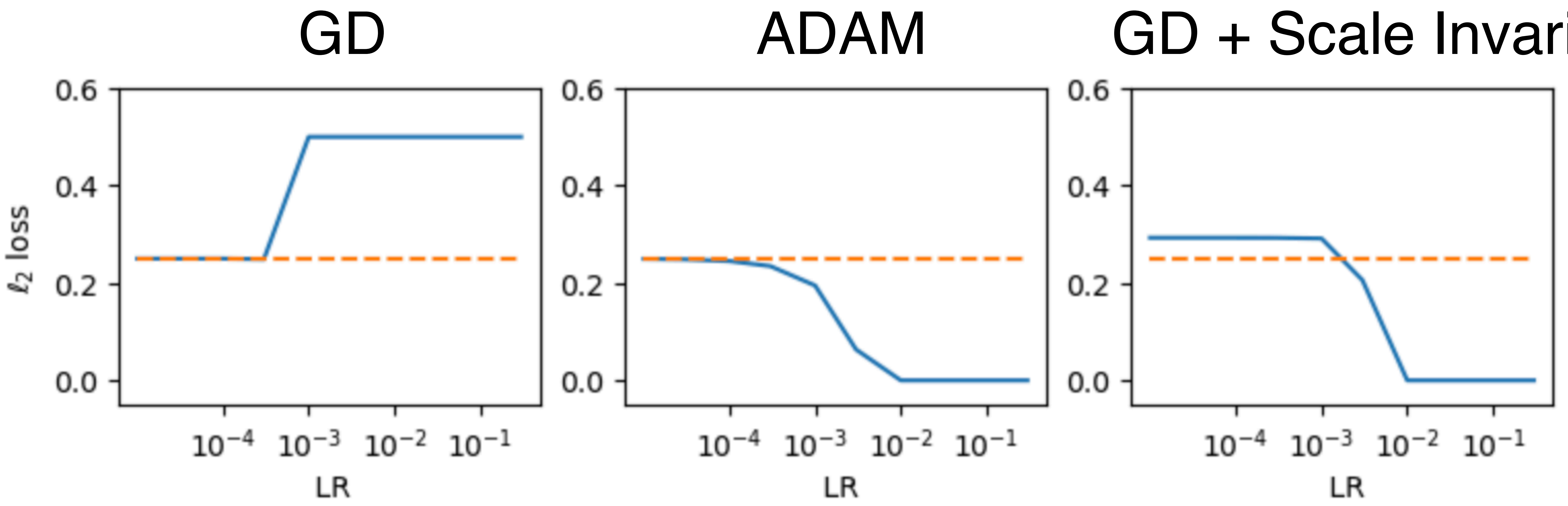
1. Design Scale Invariant Architecture $\Longrightarrow$ Increase training stability
2. SGD + WD (no momentum, no warm-up)
3. A Novel Clipping Rule: Relative Global Clipping

Experiments

- Train Scale Invariant BERT with SGD

Ingredient 1: Scale Invariant Architecture

- Scale Invariant $\Longleftrightarrow$ 0-homo $\Longleftrightarrow f(x) = f(cx), \forall c > 0$
- Thus 0-homo model output $f \Longrightarrow$ 0-homo loss function $L \Longrightarrow$ (-2)-homo Hessian $\nabla^2 L$
- **Euler's Theorem:** L is k-homo $\Longrightarrow \langle x, \nabla L(x) \rangle = kL(x)$
  - L is scale invariant (0-homo) $\Longrightarrow \langle x, \nabla L(x) \rangle = 0 \Longrightarrow \|x - \eta \nabla L(x)\| \geq \|x\|$
- **How scale invariance removes training instability**:
- Too large LR $\Longrightarrow$ Loss $\uparrow$, Norm $\uparrow \Longrightarrow$ Hessian $\downarrow \Longrightarrow$ Optimization resumes!



$$L(A, B) = \frac{1}{2} \left\| \frac{AB^\top}{\|AB^\top\|_F} - Y \right\|_F^2$$

$\alpha = 100$

Ingredient 1: Scale Invariant Architecture

- VGG and ResNets are (nearly) scale invariant, with BatchNorm or other normalization.
- But Transformer is not, even with layernorm. (Attention!!)

- We designed a scale invariant variant of BERT — **SIBERT**.
- **Key features** (to make encoder scale invariant):
  1. Scale Invariant Attention Score: $\mathbf{p} = \text{sofmax}(\mathbf{q}) \rightarrow p_i = \max(q_i, 0) / \sum_i \max(q_i, 0)$
  2. Architecture Change *(PostNorm→ PreNorm)*
  3. Activation Change *(GeLU→ ReLU)*

Existing Analysis for SGD on Scale Invariant Loss

- Scale Invariance $\iff L(cx) \equiv L(x), \forall c > 0$
- Let $\bar{x} = \dfrac{x}{\|x\|}, \rho = \sup_x \lambda_{\max}(\nabla^2 L(\bar{x}))$.
- **Goal**: Find ~~parameter x~~ direction $\bar{x}$ with small gradient.

- **Thm**[Arora, **L** & Lyu, 19]: For GD w. any fixed LR and init, $\min_{0\le t\le T} \|\nabla L(\bar{x}(t))\|^2 \le O(T^{-1})$

  For SGD w. any fixed LR and init, $\min_{0\le t\le T} \mathbb{E}\|\nabla L(\bar{x}(t))\|^2 \le \tilde{O}(T^{-0.5})$

- But $O(\cdot)$ hides poly dependence on scale initialization…
- Too large norm $\Longrightarrow$ small 'effective LR' , $\dfrac{\eta}{|x|^2} \Longrightarrow$ Optimization stucks!

Outline

Our Hypothesis

- k-homogeneity in Architecture

Our Recipe

1. Design Scale Invariant Architecture $\Longrightarrow$ Increase Training stability
2. SGD + WD (no momentum, no warm-up) $\Longrightarrow$ Increase training efficiency under rescaling of loss and initialization
3. A Novel Clipping Rule: Relative Global Clipping

Experiments

- Train Scale Invariant BERT with SGD

Ingredient 2: SGD + WD

- Use Weight Decay to shrink weight per step and accelerate when stucking

$$x(t+1) = (1-\eta\lambda)x(t) - \eta\nabla L(x(t))$$

Weight Decay(WD)

- Or increase LR multiplicatively per step, like $\eta_t = \eta \cdot 1.001^t$.

- Two methods are mathematically equivalent with $\eta_t = \eta \cdot (1-\eta\lambda)^{-2t}$. [**L**&Arora,21]

Convergence Results for GD + WD on Scale Invariant Loss

- GD+WD: $$x(t+1) = (1-\eta\lambda)x(t) - \eta \nabla L(x(t))$$
- **Thm**(GD+WD): For $\eta\lambda \leq 0.5$, $T_0 \lesssim \frac{1}{2\eta\lambda}\left|\ln \frac{\|x(0)\|_2^2}{\eta}\right|$, we have
$$\min_{t=0,\ldots,T_0} \|\nabla L(\bar{x}(t))\|_2^2 \leq O(\lambda\eta)\,. \qquad \min_t \|\nabla L(\bar{x}(t))\|^2 = O(T^{-1})$$
- Proof sketch:
  1. $\|\bar{x}(t)\|_2 \to \Theta(\sqrt{\rho\eta})$. (balance of 2 forces: GD→ norm ↑, WD → norm ↓)
  2. **Descent Lemma** + standard analysis:
$$L(x(t)) - L(x(t+1)) \geq \eta\left(1 - \frac{2\rho\eta}{\|x(t)\|_2^2}\right)\|\nabla L(x(t))\|_2^2.$$

Convergence Results for SGD +WD

- SGD+WD: $x(t+1) = (1-\eta\lambda)x(t) - \eta \nabla L_{\gamma_t}(x(t))$ γ_t — data/batch at step t
- **Assumption:** $\underline{\sigma}^2 \le \mathbb{E}_\gamma \|\nabla L_\gamma(\bar{x})\|^2 \le \bar{\sigma}^2$.
- **Thm**(SGD+WD): For $\lambda\eta \lesssim \frac{\underline{\sigma}^4}{M^4}(\ln \frac{T}{\delta^2})^{-1}$, where $M = \sup_{x,\gamma} \|\nabla L_\gamma(\bar{x})\|$, w.p. $1-5\delta$,

$$\forall T_1 \le t \le T-1, \quad \frac{\underline{\sigma}^2}{2} \le \frac{2\lambda}{\eta}\|x(t)\|_2^4 \le 4\bar{\sigma}^2$$

where $T_1 = \frac{1}{\eta\lambda} O\left(\ln|\eta\lambda| + \left|\ln \eta/\|x(0)\|_2^2\right|\right)$, and $\min_t \|\nabla L(\bar{x}(t))\|^2 = O(T^{-0.5})$

$$\frac{1}{T-T_1}\sum_{t=T_1}^{T-1} \|\nabla L(\bar{x}(t))\|_2^2 \le \tilde{O}\left(\frac{1}{(T-T_1)\sqrt{\eta\lambda}} + \sqrt{\eta\lambda}\right)$$

Outline

Our Hypothesis

- k-homogeneity in Architecture

Our Recipe

1. Design Scale Invariant Architecture $\Longrightarrow$ Increase Training stability
2. SGD + WD (no momentum, no warm-up) $\Longrightarrow$ Increase training efficiency under rescaling of loss and initialization
3. Relative Global Clipping $\Longrightarrow$ Reduce spikes in training loss and param norm

Experiments

- Train Scale Invariant BERT with SGD

Ingredient 3: Global Relative Clipping

- Analysis of SGD+WD only works for sufficiently small $\eta\lambda$.
- Gradient norm is heavy-tailed $\Longrightarrow$ norm and loss oscillates.
- **Goal:** clip only when necessary so that stochastic gradient is almost unbiased.

  $\Longrightarrow$ Clipping should not be triggered when $\|\nabla L_\gamma(\bar{x})\| \equiv \sigma$ for all x, γ

  $\Longrightarrow \|x(t)\|_2^2 \to \sqrt{\frac{2\eta}{\lambda(2-\eta\lambda)}}\sigma$, and $\|\nabla L_\gamma(x(t))\|_2 = \sqrt{\frac{\lambda(2-\eta\lambda)}{\eta}}\|x(t)\|_2.$

- **Global Relative Clipping**: Clip grad norm to $\sqrt{\frac{2C\lambda}{\eta}}\|x(t)\|_2.$ (C>1, Default =2)
- **Thm**(Informal): Global Relative Clipping $\Longrightarrow$ better norm convergence.

Outline

Our Hypothesis

- k-homogeneity in Architecture

Our Recipe

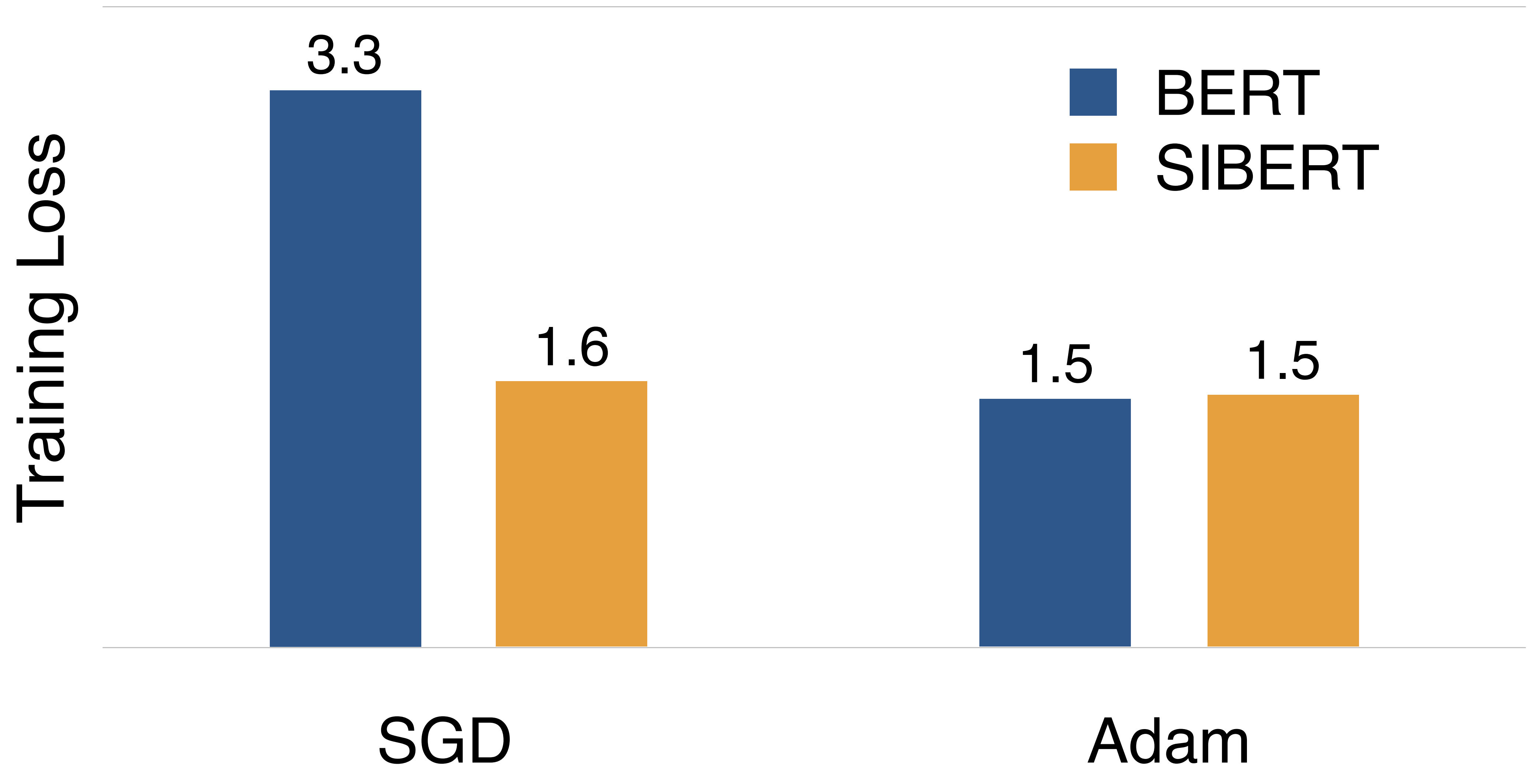
1. Design Scale Invariant Architecture
2. SGD + WD
3. A Novel Clipping Rule

Experiments

- Train Scale Invariant BERT with SGD

Performance of SIBERT

- **Dataset**: Wiki+Books. **Model**: Base size BERT.
- We use Global Relative Clipping (C=2) and WD for SGD on SIBERT .

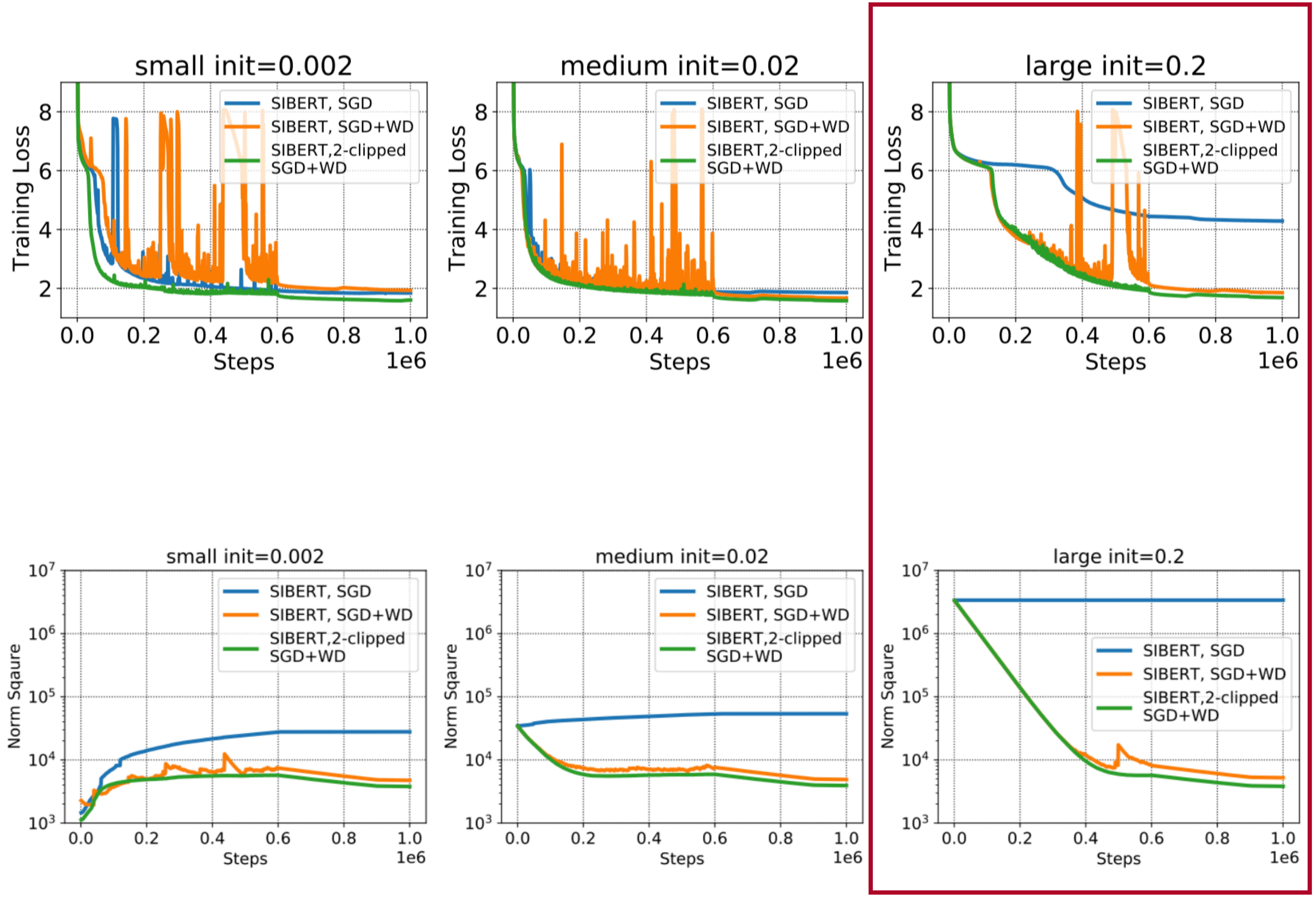


Downstream Task

		MNLI Acc	SQuAD1 F1	SQuAD2 F1
Base	BERT + ADAM	**84.4**	**90.3**	78.8
	SIBERT + SGD	82.6	89.3	76.8
	+ 2x training	83.3	**90.3**	**80.0**

larger is better

Robustness of SIBERT to Scalings



SGD+WD performs consistently for different initialization scales.

Clipping removes spikes in train curve and yields better convergence.
(occur for only ~1% steps)

Without WD, the performance of SGD can be significantly affected by scaling of initialization.

Conclusion

- We hypothesis the training instability of Transformers is related to homogeneity structure in the network.

- Our recipe for fast and robust training via non-adaptive methods
  1. Design scale invariant architecture (BERT $\rightarrow$SIBERT)
  2. SGD + WD
  3. Relative Global Clipping

Thank You !