Robust Training of Neural Networks Using Scale Invariant Architectures

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

Abstract

In contrast to SGD, adaptive gradient methods like ADAM allow robust training of modern deep networks, especially large language models. However, the use of adaptivity not only comes at the cost of extra memory but also raises the fundamental question: *can non-adaptive methods like* SGD *enjoy similar benefits*? In this paper, we provide an affirmative answer to this question by proposing to achieve both robust and memory-efficient training via the following general recipe: (1) modify the architecture and make it *scale invariant*, (2) train with SGD and weight decay, and optionally (3) clip the global gradient norm proportional to weight norm multiplied by $\sqrt{\frac{2\lambda}{\eta}}$, where η is learning rate and λ is weight decay. We show that this general approach is robust to rescaling of parameter and loss by proving that its convergence only depends logarithmically on the scale of initialization and loss, whereas the standard SGD might not even converge for many initializations. Following our recipe, we design a scale invariant version of BERT, called SIBERT, which when trained simply by vanilla SGD achieves performance comparable to BERT trained by adaptive methods like ADAM on downstream tasks.

1. Introduction

Neural architectures like transformers are the cornerstone for modern machine learning applications. However, training them is difficult and often results in training instability (Liu et al., 2020; Zhang et al., 2020b). To enable stable training, one typically requires adaptive and carefully tuned learning rates. However, the reason behind this issue is not very well-understood and lacks a formal treatment.

[1]Princeton University, the work is done when interning at Google Research New York [2]Google Research New York [3]Google DeepMind New York. Correspondence to: Zhiyuan Li <zhiyuanli@cs.princeton.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

In this paper, we hypothesize that a primary cause of such behavior is the k-homogeneous ($k \geq 2$) nature of the network i.e., property where network's output is scaled by s^k when its parameters are scaled by s. To illustrate our point, we consider the following instructive toy model.

Example 1.1. *Consider logistic regression with 1-dimensional non-separable data, $\{z_i, y_i\}_{i=1}^n \in (\mathbb{R} \times \{\pm 1\})^n$. The loss is defined as $L(x_1, , \ldots, x_{2k}) = \widetilde{L}(X) := -\sum_{i=1}^n \ln(1 + e^{-z_i y_i X})$ where $X = x_1 \ldots x_{2k}$ and $k \geq 2$.*

Since $\widetilde{L}$ is convex with bounded smoothness in X, there exists step size that are independent of any initialization that allow GD to converge to the optimal solution. In sharp contrast, the reparametrized loss $L(x_1, , \ldots, x_{2k})$ with $2k$-homogeneous structure does not enjoy this nice stability property — the learning rate has to be tuned according to the initialization. In particular, when $\eta \geq \frac{2}{|\nabla \widetilde{L}(X(0))|}(X(0))^{\frac{1}{k}-1}$ and $X(0) > X^*$ where $X^* > 0$ is the global minimizer, $X(t)$ will monotonically increase and explode, if all x_i are initialized to be the same.

We refer the reader to Appendix B for a formal justification of this example. In the above example, the success of optimization is very sensitive to the right choice of the learning rate that depends on the initialization. Furthermore, the training cannot recover once the norm explodes due to large gradient update.

In the above one-dimensional example it is still possible to find a small workable learning rate by extensive grid search that depends on the initial point, however, the situation can get worse when the k-homogeneous structure has an unbalanced initialization as below.

Example 1.2. *Consider solving low-rank matrix decomposition by Gradient Descent. Let $L(A, B) = \frac{1}{2}\left\|AB^\top - Y\right\|_2^2$ where $A, B \in \mathbb{R}^{d \times r}$ are both initialized i.i.d. gaussian with covariance $\sigma_A^2 \gg \sigma_B^2 \approx \sigma_A^{-2}$, $Y \in \mathbb{R}^d$ and $d \gg r$.*

Solving this optimization problem requires A and B learning the column and row space of Y respectively, but the

unbalanced initialization will force the learning rate to be small enough such that B does not explode and, thus, A is almost frozen. To see this, note in the standard convergence analysis of GD, we need LR smaller than $2/\left\|\nabla^2 L\right\|$ to ensure the Descent Lemma holds, i.e.*, loss decreases in a single step. Here we have that the smoothness w.r.t A (fixing B) is $\lambda_{max}(BB^T)$ and the smoothness w.r.t. B (fixing A) is $\lambda_{max}(AA^T)$. Thus, LR can be at most $O(\frac{1}{\sigma_A^2})$, but the gradient of A is only of magnitude $O(\sigma_B)$, resulting in A learning the column space slowly. On the other hand, when $d = r = 1$ and $Y = 0$, choosing $\eta > \frac{4}{\|\nabla^2 L(A(0),B(0))\|}$ will cause* GD *to explode (Lewkowycz et al., 2020).*

Similar issues can exist in deep neural networks as the k-homogeneous structure is quite common. For instance, (Liu et al., 2020) identified the gradient norm varies with depth and that no single learning rate is globally optimal for all layers. To this end, one has to resort to adaptive methods like ADAM to handle the k-homogeneous structure of deep networks and allow for its robust training. However, this not only comes at the expense of higher memory, but also raises the key question of our interest:

Can non-adaptive methods like SGD enjoy fast and robust convergence without training instability?

Answering this question, requires us to first define our notion of robustness. In this paper, we primarily aim for three aspects of robustness by preventing: explosion of parameters (e.g. due to frequent large gradient updates), slow progress in training (e.g. due to loss plateaus) and loss explosion or spikes (e.g. due to possibly infrequent large magnitude updates). In this paper, we propose a simple yet powerful general approach for achieving such fast and robust convergence. At a high level, our recipe for robust training includes three key ingredients:

1. *Designing architectural scale invariance which allows for improved training stability and prevents explosion of the parameters*. We show that by using scale invariance in the architecture (i.e., making the network 0-homogeneous), one can effectively control the gradient updates when the parameter norm is large.

2. *Using SGD with weight decay for training, wherein enabling weight decay improves training efficiency under rescaling of loss and initialization*. While scale invariance prevents explosion of parameters, the training convergence has strong dependence on initialization scale and learning rate, which can make training inefficient in face of parameter and initialization rescaling. Use of SGD with weight decay circumvents this issue.

3. *Using a novel Relative Global Clipping to prevent spikes in training loss and improve overall convergence speed*. Although scale invariance in the architecture already guarantees the training stability, it does not prevent severe non-monotonic loss explosion. By using a new global clipping approach, we show that one can prevent such loss explosions effectively.

We show that this surprisingly simple training recipe can not only improve the memory efficiency over adaptive methods but also achieves robust training. In light of the above background, we list our main contributions below.

- In Section 3, we propose a new general recipe for memory efficient, robust training using (1) scale invariant architecture; (2) SGD+WD for training and (3) a novel clipping rule, called Relative Global Clipping, for clipping the updates. Following this recipe, we design a new variant of BERT called Scale Invariant BERT (SIBERT).

- In Sections 4.1 and 4.2, we prove the convergence rate to the approximate first order point for GD and SGD for scale invariant loss. We show that SGD+WD matches the standard rates, even without the knowledge about the smoothness of loss and is robust to the scale of initialization or loss.

- In Section 4.3, we show SGD+WD with Relative Global Clipping has better parameter norm convergence via a novel analysis. With assumptions that the clipping does not bring too much bias in expected gradients, we show similar convergence result to SGD+WD.

- In our empirical analysis in Section 5, we demonstrate that SIBERT trained using simple SGD can achieve performance comparable to standard BERT trained with ADAM. Furthermore, we also verify our theoretical claims. To our knowledge, this is the first time a BERT-like model has been effectively trained using vanilla SGD.

2. Related Work & Background

The literature on adaptive methods and scale invariance in neural networks is vast, so we only discuss works that are most relevant to our paper.

Adaptive Methods & Clipping Methods. Adaptive learning rates have long been studied (Polyak, 1987). In machine learning, adaptive learning rates have been popularized by ADAGRAD, which particularly benefits from sparse stochastic gradients (Duchi et al., 2011). Inspired by ADAGRAD, several adaptive methods, like ADAM, RMSPROP

and its variants have been proposed in the deep learning community (Kingma & Ba, 2015; Tieleman & Hinton, 2012; Reddi et al., 2019; You et al., 2020; Shazeer & Stern, 2018). These approaches have been crucial in the success of many deep learning applications (Vaswani et al., 2017; Devlin et al., 2018; Raffel et al., 2019). Several works have studied the benefits of adaptive methods in deep learning settings (e.g. (Liu et al., 2020; Zhang et al., 2020b)). However, as mentioned earlier, these benefits come at the cost of computational and memory efficiency. Anil et al. (2019) proposed a variant of ADAGRAD requiring fewer parameters for adaptivity, but still requires momentum. ADAFACTOR (Shazeer & Stern, 2018) removes momentum and uses much fewer adaptivity parameters, but for large models, ADAFACTOR still needs momentum to ensure training stability (Chowdhery et al., 2022). Our approach is also related to normalized and projected gradient descent, which has been studied for quasi-convex and non-convex settings (e.g. see (Hazan et al., 2015; Levy, 2016; Huang et al., 2017)). However, these methods have seen very limited success.

Clipping based optimization methods, especially gradient clipping, are widely used in deep learning applications to improve training stability or ensure privacy (Pascanu et al., 2013; Chen et al., 2020; Zhang et al., 2020a). These approaches typically use a constant threshold to clip the gradients before the update. However, choosing this threshold is difficult and requires careful tuning. Adaptive variants of clipping methods partially alleviate this issue and are closely related to adaptive methods (Zhang et al., 2020b); however, they again incur additional computation and memory costs.

Scale Invariance in deep networks. Various normalization schemes are the main source of scale invariance in deep learning, *e.g.*, BatchNorm (Ioffe & Szegedy, 2015), LayerNorm (Ba et al., 2016), Weight Normalization (Salimans & Kingma, 2016), GroupNorm (Wu & He, 2018), InstanceNorm (Ulyanov et al., 2016). Scale invariance from normalization allows GD and SGD to converge to stationary points from any initialization and with any learning rate, in $O(T^{-1/2})$ and $\widetilde{O}(T^{-1/4})$ rates respectively (Arora et al., 2018). The interplay between SGD, scale invariance and WD has also been well studied. It was shown that the effect of WD for normalized networks can be replaced by LR schedules (Hoffer et al., 2018; Zhang et al., 2018). Li & Arora (2019) formally builds the equivalence between SGD+WD and SGD with an exponential increasing LR schedule for scale invariant loss. Van Laarhoven (2017) first proposed the notion of effective LR, $\eta/\left\|\boldsymbol{x}\right\|_2^2$, for normalized networks, and showed that the unique stationary value of $\left\|\boldsymbol{x}\right\|_2^4$ is proportional to λ/η, where η is LR and λ is WD. Li et al. (2020) proved that the parameter norm always converges to the above value by modeling SGD as Stochastic Differential Equation. Wan et al. (2020) proved the parameter norm converges to the same value directly for SGD+WD, but only in expectation.

2.1. Preliminary

In this section we present the definition of scale invariant functions and some of their useful properties. For $\boldsymbol{x} \in \mathbb{R}^d$, we define $\overline{\boldsymbol{x}} := \frac{\boldsymbol{x}}{\left\|\boldsymbol{x}\right\|_2}$. We say a function is $\mathcal{C}^k$ iff it is k-times continuously differentiable.

Definition 2.1. Given a cone $U \subset \mathbb{R}^d$, we say a function $f : U \to \mathbb{R}$ is *(positively) k-homogeneous* or *of homogeneity of degree k* iff for any $c > 0$ and $x \in U$, $f(c\boldsymbol{x}) = c^k f(\boldsymbol{x})$. We say a function is *scale invariant* iff it is 0-homogeneous.

Now we present some useful properties of the derivatives of homogeneous functions.

Theorem 2.2 (Euler's Homogeneous Function Theorem)**.** *For any k-homogeneous $\mathcal{C}^1$ function f, it holds that $\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} \rangle = k f(\boldsymbol{x})$.*

Lemma 2.3. *For any k-homogeneous $\mathcal{C}^l$ function f, $\nabla^l f$ is $k - l$ homogeneous.*

Lemma 2.4 (Equivalent Scaling)**.** *The properties below hold (and generalize to stochastic loss):*

1. *For any loss L, LR η, WD λ and initialization $\boldsymbol{x}(0)$, rescaling $(L, \eta, \lambda, \boldsymbol{x}(0)) \to (cL, \eta/c, c\lambda, \boldsymbol{x}(0))$ doesn't change GD iterate $\boldsymbol{x}(t)$ for any $t \geq 0$.*

2. *For any scale invariant loss L, LR η, WD λ and initialization $\boldsymbol{x}(0)$, rescaling $(L, \eta, \lambda, \boldsymbol{x}(0)) \to (L, c^2\eta, \lambda/c^2, c\boldsymbol{x}(0))$ doesn't change the direction of GD iterate $\overline{\boldsymbol{x}}(t)$ for any $t \geq 0$. (see Lemma 2.4 in (Li & Arora, 2019))*

3. Methods

In this section, we provide a more detailed description of our recipe for robust and memory-efficient network training, which includes three building blocks: (1) scale invariant architecture (Section 3.1), (2) SGD with Weight Decay (Section 3.2) and optionally (3) the *Relative Global Clipping* (Section 3.3 and Algorithm 1).

Algorithm 1 $\sqrt{C}$-Clipped SGD + WD

Input: Total steps T, Scale invariant loss $\{L_t\}_{t\geq 1}^T$, initialization $\boldsymbol{x}(0)$, LR η, WD λ, clipping factor $C > 1$ ($C = \infty \Leftrightarrow$ no clipping).
for $t = 0$ **to** $T - 1$ **do**
  $N_t \leftarrow \min\left\{\sqrt{\frac{2C\lambda}{\eta}}\left\|\boldsymbol{x}(t)\right\|_2, \left\|\nabla L_t(\boldsymbol{x}(t))\right\|_2\right\}$.
  $\boldsymbol{x}(t+1) \leftarrow (1 - \eta\lambda)\boldsymbol{x}(t) - \eta N_t \frac{\nabla L_t(\boldsymbol{x}(t))}{\left\|\nabla L_t(\boldsymbol{x}(t))\right\|_2}$.
end for

3.1. Designing Scaling Invariant Architectures

We first revisit an approach for introducing scale invariance in neural networks, which is presented in (Li & Arora, 2019). Viewing the neural network computation as a directed graph, the high level idea is to ensure same homogeneity degree of different edges reaching a node. For example in a RESNET block, the output from an affine transform is added back to the input z from the previous layer yielding $z + \text{Aff}(z)$. Now if we scale all the network parameters by c, both z and $\text{Aff}(z)$ should have the same degree of homogeneity and scale as c^k. Otherwise the network is no longer homogeneous and, hence, cannot be scale invariant.

In this paper, we apply the above design philosophy to develop a scale invariant version of BERT (Devlin et al., 2018) — a transformer based model. A transformer has two main building blocks that need to be made scale invariant – residual block and Attention (Vaswani et al., 2017). For residual block, Li & Arora (2019) already demonstrated how to make both the PreNorm and PostNorm version of RESNET scale invariant (see Appendix of their paper for more details). In this paper, we use their PreNorm variant (see Figure 5). Furthermore, we design a novel scale invariant version of Attention block in transformer, as described below.

Scale Invariant Attention: Recall the standard self attention block computes the following for a given input $Q, K, V \in \mathbb{R}^{n \times d_{model}}$:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QW^Q(KW^K)^\top}{\sqrt{d_k}})VW^V.$$

Here $W^Q, W^K \in \mathbb{R}^{d_{model} \times d_k}$ and $W^V \in \mathbb{R}^{d_{model} \times d_v}$ are affine transformations and, hence, are all 1-homogeneous transformations. The Softmax function computes row wise softmax normalization. It is easy to see that standard attention is not homogeneous as softmax is itself not homogeneous.

We design a novel Scale Invariant Attention (SI Attention) in the following way: (also see Figure 7)

$$\text{SI-Attention}(Q, K, V) = \text{N}(\text{ReLU}(QW^Q(KW^K)^\top)VW^V,$$

where N denotes the row-wise normalization by sum, *i.e.*, $[\text{N}(A)]_{ij} = \frac{a_{ij}}{\sum_j a_{ij}}$ and $\text{ReLU}(A)$ denote the element-wise max between matrix A and 0. Notably we replace the softmax with a ReLU activation followed by normalization. Both ReLU and normalization are homogeneous operations; thus, making the overall attention score computation $(\text{N}(\text{ReLU}(ZQK^\top Z^\top)))$ scale invariant to the concatenation of all parameters $\boldsymbol{x}$, assuming Q, K, V are already positive homogeneous to $\boldsymbol{x}$. Due to space constraints, the full design of Scale Invariant BERT (SIBERT) is relegated to Appendix A.

3.2. Training Algorithm: SGD + WD

Although scale invariance can prevent parameter divergence after a large gradient update by eliminating the positive feedback between gradient and parameter norm, it alone does not ensure SGD trains the network in a robust and efficient way. This is because, as shown in (Arora et al., 2018), the parameter norm monotonically increases when SGD is used to optimize a scale invariant loss. As a result, once the norm becomes too large (e.g due to large gradient in some step) the training can slow down drastically as the effective LR $\frac{\eta}{\|\boldsymbol{x}_t\|_2^2}$ is too small; thus, preventing effective recovery from even minor training instabilities.

To tackle this issue we propose to use Weight Decay(WD) as a way to reduce the parameter norm; thereby, allowing the network to recover from slow training induced by infrequent updates of large norm. Under mild assumptions that the expectation of squared norm of stochastic gradient does not vary too much on the unit sphere, (Li et al., 2020; Wan et al., 2020) show that the parameter norm will stabilize in $O(\frac{1}{\eta\lambda})$ steps and the learning dynamics is equivalent to one on unit sphere with effective learning rate proportional to $\Theta(\sqrt{\lambda\eta})$.

Leveraging the advantage of quick norm convergence, we show that the convergence of SGD+WD is insensitive to the following three operations: loss rescaling (A1), initialization rescaling (A2) and re-parametrization (A3), meaning the same convergence rate (independent of scaling c) can be achieved, in up to $\frac{|\log c|}{\lambda\eta}$ more steps. (See formal statement in Theorems 4.1 and 4.5 This property reduces the effort of hyperparameter tuning and also makes training more robust when switching between different codebases and frameworks, which is likely to have different default scaling or parametrization. Also note by scale invariance of loss L, (A2) is equivalent to (A3).

(A1). $L \to cL$, for any $c > 0$.

(A2). $\boldsymbol{x}(0) \to c\boldsymbol{x}(0)$, for any $c > 0$.

(A3). $(L, \boldsymbol{x}(0)) \to (L', c\boldsymbol{x}(0))$, where L' is defined as $L'(\boldsymbol{x}) := L(\frac{\boldsymbol{x}}{c})$ for any $c > 0$.

As a comparison, previous work (Arora et al., 2018) showed that GD converges to ϵ approximate stationary point of a scale invariant loss in $O(\frac{1}{\epsilon^2})$ and SGD converges in $\widetilde{O}(1/\epsilon^4)$ steps with any initialization. However, the constant in $O(\cdot)$ scales linearly or inversely to the above scalings (c in (A1-3)). This is far from satisfying, and indeed their experiments show that either large or small LR could substantially slow-down the training progress.

3.3. Relative Global Clipping

Gradient clipping is a widely used effective strategy to stabilize neural network training. However, often the clipping

threshold need to be tuned based on the optimization problem and the specific gradient distribution. Furthermore, simply using a constant threshold can severely degrade the performance (Zhang et al., 2020b). Thus, it is unclear how the clipping threshold needs to be set for SGD+WD on scale invariant functions such that it is insensitive to rescaling of loss and reparametrization, *e.g.*, (A1-3).

To this end, we propose a clipping strategy named *Relative Global Clipping* which allows consistent and robust training behavior for SGD+WD on scale invariant loss under the aforementioned operations. In particular, we propose to set the clipping threshold as $\sqrt{\frac{2C\lambda}{\eta}}\left\|\boldsymbol{x}\right\|_2$, where $C \geq 1$ is a hyperparamer with default value $\sqrt{C} = 2$. The high level design idea is that (1) the clipping rule should be invariant to the scalings $(L, \eta, \lambda) \to (cL, \eta/c, c\lambda)$ and $(\boldsymbol{x}, \eta, \lambda) \to (c\boldsymbol{x}, c^2\eta, \lambda/c^2)$ for any $c > 0$, to which SGD+WD is invariant (see Lemma 2.4); (2) the clipping rule should only remove the extremely large gradients and should not trigger too often to ensure that gradient after clipping remains almost unbiased.

Intuitively, the derivation of Relative Global Clipping involves the following line of reasoning: Suppose the norm of the stochastic gradient $\left\|\nabla L_\gamma(\boldsymbol{x})\right\|_2$ is constant, say σ, for all data and every parameter $\boldsymbol{x}$ on the unit sphere. In this case, we expect our clipping strategy to not be triggered since there are no extremely high stochastic gradients. Since L_γ is scale invariant, Theorem 2.2 implies that $\langle \nabla L_\gamma(\boldsymbol{x}), \boldsymbol{x}\rangle = 0$. That is,

$$\begin{aligned}\left\|\boldsymbol{x}(t+1)\right\|_2^2 =&(1-\eta\lambda)^2\left\|\boldsymbol{x}(t)\right\|_2^2 + \eta^2\left\|\nabla L_\gamma(\boldsymbol{x}(t))\right\|_2^2\\ =&(1-\eta\lambda)^2\left\|\boldsymbol{x}(t)\right\|_2^2 + \eta^2\sigma^2/\left\|\boldsymbol{x}(t)\right\|_2^2. \quad (1)\end{aligned}$$

It is not difficult to show the iteration (1) has a unique stationary point, $\left\|\boldsymbol{x}(t)\right\|_2^2 = \sqrt{\frac{2\eta}{\lambda(2-\eta\lambda)}}\sigma$(Van Laarhoven, 2017). In other words, at norm equilibrium, it holds

$$\left\|\nabla L_\gamma(\boldsymbol{x}(t))\right\|_2 = \frac{\sigma}{\left\|\boldsymbol{x}(t)\right\|_2} = \sqrt{\frac{\lambda(2-\eta\lambda)}{\eta}}\left\|\boldsymbol{x}(t)\right\|_2. \quad (2)$$

The above calculation suggests the clipping threshold should be at least $\sqrt{\frac{2\lambda}{\eta}}\left\|\boldsymbol{x}(t)\right\|_2$. [1] Furthermore, it is not difficult to check that the clipping threshold $\sqrt{\frac{2\lambda}{\eta}}\left\|\boldsymbol{x}(t)\right\|_2$ is indeed invariant to the above mentioned scalings $(L, \eta, \lambda) \to (cL, \eta/c, c\lambda)$ and $(\boldsymbol{x}, \eta, \lambda) \to (c\boldsymbol{x}, c^2\eta, \lambda/c^2)$. For each hyperparameter $C > 1$, the behavior of SGD+WD is consistent for different scalings (A1-3) and it also improves the norm convergence (reducing undesirable spikes in norm while training) for SGD+WD (see Theorem 4.8). Under mild assumptions that such clipping does not introduce too much bias in gradients, we show that our recipe enables convergence to approximate stationary points. Furthermore, the rate only depends *logarithmically* on the initialization and loss scale, as shown in the following section.

[1] We drop $-\eta\lambda$ for convenience. This doesn't lead to any practical difference as $\eta\lambda$ is typically very small, *e.g.* less than10^{-4}.

4. Theoretical Analysis

In this section, we provide theoretical analysis of the convergence of SGD+WD to approximate first order stationary points for scale invariant functions. We first start with the key highlights of our theoretical analysis for SGD+WD:

1. Parameter norm converges to $\Theta((\frac{\lambda}{\eta})^{\frac{1}{4}})$ in $T_1 = \widetilde{O}(\frac{1}{\eta\lambda})$ steps with high probability where T_1 is a function of loss L, initial norm $\left\|\boldsymbol{x}(0)\right\|_2$, LR η and WD λ. Moreover, $T_1(L, \left\|\boldsymbol{x}(0)\right\|_2, \eta, \lambda)$ changes most by $\frac{\ln|c|}{\eta\lambda}$ for operation (A1-3).
2. After step T_1, convergence to first order approximate stationary point happens and the rate only depends on $\eta\lambda$ and is unaffected by operations (A1-3).

Properties (1) and (2) suggest our results are more robust to initialization scale (by only having logarithmic dependence on it), showing the advantage of using scale invariant functions while matching the standard convergence rates for non-convex functions. Note that the standard notion of approximate stationary point, *i.e.* $\boldsymbol{x}$ with small gradient norm of $\left\|\nabla L(\boldsymbol{x})\right\|_2$ is not useful for scale invariant loss, as one can simply scale up the initialization $\boldsymbol{x}(0)$ to infinity and the gradient norm thus scales inversely. A more reasonable notion of 'stationary point' is that the direction of $\boldsymbol{x}$, denoted by $\overline{\boldsymbol{x}} := \frac{\boldsymbol{x}}{\left\|\boldsymbol{x}\right\|_2}$, has small gradient norm, as first introduced in (Arora et al., 2018). We will use this definition of approximate stationary point throughout the paper. In the section we also assume L is a C^2 and scale invariant function and $\rho := \max_{\left\|\boldsymbol{x}\right\|=1}\left\|\nabla^2 L(\boldsymbol{x})\right\|$.

4.1. Convergence of GD +WD

We first present the convergence result in the deterministic case, *i.e.*, Gradient Descent over $L(\boldsymbol{x}) + \frac{\lambda}{2}\left\|\boldsymbol{x}\right\|_2^2$.

$$\text{GD+WD:}\quad \boldsymbol{x}(t+1) = (1-\eta\lambda)\boldsymbol{x}(t) - \eta\nabla L(\boldsymbol{x}(t)) \quad (3)$$

Theorem 4.1 (GD+WD). *For $\eta\lambda \leq \frac{1}{2}$, let $\boldsymbol{x}(t)$ be defined by GD (3), and $T_0 = \left\lceil \frac{1}{2\eta\lambda}\left(\left|\ln\frac{\left\|\boldsymbol{x}(0)\right\|_2^2}{\rho\pi^2\eta}\right| + 3\right)\right\rceil$. We have*

$$\min_{t=0,\ldots,T_0}\left\|\nabla L(\overline{\boldsymbol{x}}(t))\right\|_2^2 \leq 8\pi^4\rho^2\lambda\eta. \quad (4)$$

This bound matches the standard $O(\frac{1}{\sqrt{T}})$ convergence rate to first order stationary point for non-convex functions. Remarkably, for a given training budget T, once we can set

$\eta\lambda$ to be $\frac{D}{T}$ where D is a constant (*e.g.* 10), the convergence becomes robust to the choice the hyperparameters due to just a logarithmic dependence on them. In particular, GD+WD can work with any scaling of L (which affects the smoothness on unit sphere, ρ), LR η and initial norm $\|\boldsymbol{x}(0)\|_2$, as long as $\frac{\|\boldsymbol{x}(0)\|_2^2}{\rho\pi^2\eta} \in [e^{-D}, e^D]$. This is in sharp contrast to GD on standard loss as it requires knowledge about the smoothness to set the optimal LR.

However, one weakness of the above result is that with a fixed $\eta\lambda$, longer training does not guarantee further convergence. The intuition is that once the iterate converge in direction and the gradient vanishes, Weight Decay will dominate the dynamics and thus the norm approaches 0, which increases the sharpness. When the sharpness gets larger than $2/\eta$, the dynamics become unstable and results in divergence. This phenomena is first observed in Li et al. (2020) and verified by Lobacheva et al. (2021) in practical settings. This behavior can also be viewed as a special case of Edge of Stability as described in Cohen et al. (2020).

Proof Sketch of Theorem 4.1. Scale invariant functions do not have bounded smoothness at 0 making it a challenge to use standard convergence analysis. Our key insight is that for scale invariant loss function, even with a fixed LR η, GD can tune its effective LR $\frac{\eta}{\|\boldsymbol{x}(t)\|_2^2}$ by changing the norm. Thus once GD passes the area of the suitable norm, the smoothness of scale invariant loss function is upper bounded by $\frac{\rho}{r^2}$ outside the ball with radius r centered at 0.

More concretely our proof consists of 2 steps. In the first step we show that GD+WD iterates pass an area of suitable norm ($\approx \sqrt{\rho\eta}$). For large initial norm, WD could bring the norm to correct scaling in log time and then converge (Theorem D.2). If the initial norm is too small and the direction is not approximately stationary, then the large gradient due to the small norm will increase the parameter norm drastically in a single step (Lemma D.1), and again Weight Decay can bring the norm down in log steps. In the second step we show that, once the norm reaches this suitable value, the descent lemma (Lemma 4.2) starts to hold and the convergence analysis is standard. □

Lemma 4.2. *Let $\boldsymbol{x}(t), \boldsymbol{x}(t+1)$ be defined as (3), we have*

$$L(\boldsymbol{x}(t)) - L(\boldsymbol{x}(t+1)) \geq \eta\left(\frac{1}{1-\eta\lambda} - \frac{\rho\eta}{2\|\boldsymbol{x}(t)\|_2^2(1-\eta\lambda)^2}\right)\|\nabla L(\boldsymbol{x}(t))\|_2^2.$$

When $\eta\lambda \leq \frac{1}{2}$, the above can be simplified into

$$L(\boldsymbol{x}(t)) - L(\boldsymbol{x}(t+1)) \geq \eta\left(1 - \frac{2\rho\eta}{\|\boldsymbol{x}(t)\|_2^2}\right)\|\nabla L(\boldsymbol{x}(t))\|_2^2.$$

Remark 4.3. One might wonder why the upper bounds on loss and gradient norm do not appear in Theorem 4.1. This is because we are working on a compact domain (the unit sphere) and twice-differentiability implies those bounds implicitly. (See Lemmas C.3 and C.4)

4.2. Convergence of SGD+WD

Below we present our convergence analysis for SGD+WD.

Setting: Let Γ be an index set and $L_\gamma : \mathbb{R}^d/\{\mathbf{0}\} \to \mathbb{R}$ be a scale invariant loss function for each $\gamma \in \Gamma$. We denote $\mathbb{E}_\gamma L_\gamma$ by L. We assume the largest possible stochastic gradient norm is finite, *i.e.*, $M := \sup_{\gamma\in\Gamma} \max_{\|\boldsymbol{x}\|=1} \|\nabla L_\gamma(\boldsymbol{x})\|$. SGD is defined as (5).

$$\text{SGD+WD: } \boldsymbol{x}(t+1) = (1-\eta\lambda)\boldsymbol{x}(t) - \eta\nabla L_{\gamma_t}(\boldsymbol{x}(t)), \quad (5)$$

where $\gamma_t \in \Gamma$ are i.i.d. random variables. We further assume there exists constants $\underline{\sigma}$ and $\overline{\sigma}$, such that $\underline{\sigma}^2 \leq \mathbb{E}\|\nabla L_\gamma(\boldsymbol{x})\|_2^2 \leq \overline{\sigma}^2$, for any $\|\boldsymbol{x}\|_2 = 1$. We finally need the following condition on $\eta\lambda$ to bound convergence.

Condition 4.4. $\frac{\underline{\sigma}^2}{M^2} \geq 3e^{4\eta\lambda}\sqrt{\lambda\eta\max\{\ln\frac{2T^2}{\delta}, 1\}}$.

The Condition 4.4 is useful for proving norm convergence in high probability. In practice, typically $\eta\lambda$ is very small. Our experiments use $\eta = 0.0008$ and $\lambda = 0.01$. Hence $e^{4\eta\lambda} \approx 1$, and Condition 4.4 essentially requires the gradient norm square cannot exceed its average multiplied by $1/\sqrt{\eta\lambda} \approx 350$, which is reasonable for most iterates.

Theorem 4.5 (SGD+WD)**.** *Let $\boldsymbol{x}(t)$ be defined by* SGD (5)*. For $\eta\lambda \leq 0.1$, under Condition 4.4, with probability $1 - 5\delta$,*

$$\forall T_1 \leq t \leq T-1, \quad \frac{\underline{\sigma}^2}{2} \leq \frac{2\lambda}{\eta}\|\boldsymbol{x}(t)\|_2^4 \leq 4\overline{\sigma}^2, \quad (6)$$

and

$$\begin{aligned}&\frac{1}{T-T_1}\sum_{t=T_1}^{T-1}\|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2 \\ \leq&\frac{\pi^2\rho\overline{\sigma}}{(T-T_1)\sqrt{2\eta\lambda}} + 4\sqrt{\eta\lambda}\frac{\rho\overline{\sigma}^3}{\underline{\sigma}^2} \\ &+\sqrt{\frac{\ln\frac{2}{\delta}}{T-T_1}}4\frac{\pi\rho M\overline{\sigma}}{\underline{\sigma}} + \sqrt{\frac{\ln\frac{2}{\delta}}{T-T_1}}4\sqrt{\lambda\eta}\frac{M^2\rho\overline{\sigma}}{\underline{\sigma}^2},\end{aligned} \quad (7)$$

where $T_1 = \frac{1}{4\eta\lambda}\max\left\{\ln\frac{M^2\eta\lambda}{\overline{\sigma}^2} + \left|\ln\frac{2e^4M^2}{\|\boldsymbol{x}(0)\|_2^4\eta^{-2}}\right|, 8\right\}$.

The proof of this theorem is presented in Appendix E. Similar to our earlier result for GD this bound matches the standard $O(T^{-1/4})$ convergence rate of SGD for non-convex functions by setting $T = \widetilde{O}(\frac{1}{\eta\lambda})$. Further, it only has a logarithmic dependence on the initialization scale $\|\boldsymbol{x}(0)\|_2$, and

enjoys robustness to initialization scale as discussed earlier for GD. We further extend this result to the case where the scale invariant loss has multiple scale invariant parameter groups in Appendix G.

We next present our analysis for SGD with clipping.

4.3. Convergence of SGD with Relative Global Clipping

Now we will present our analysis for the clipped SGD. Recall the clipped SGD update from Algorithm 1 has the following norm dynamics.

Norm dynamics of clipped SGD:

$$\|\boldsymbol{x}(t+1)\|_2^2 = (1-\eta\lambda)^2 \|\boldsymbol{x}(t)\|_2^2 + \eta^2 \min\left\{ \frac{\|\nabla L_\gamma(\overline{\boldsymbol{x}}(t))\|_2^2}{\|\boldsymbol{x}(t)\|_2^2}, \frac{2\lambda C}{\eta} \|\boldsymbol{x}(t)\|_2^2 \right\}.$$

To present our bound we need the following definitions.

Definition 4.6 (C-clipped mean)**.** Given a distribution P on $\mathbb{R}_{\geq 0}$ and constant $C > 1$, we define $F_{P,C}(\mu) = \mathbb{E}_{t\sim P}[\min\{t, C\mu\}]$, and define the *$C$-clipped mean* of P, $\mu_{P,C}$ as the largest positive real number satisfying that $F_{P,C}(C\mu_{P,C}) = \mu_{P,C}$. Such a definition is valid because $F_{P,C}(0) = 0$ and thus 0 is always a solution.

For convenience, we also define $G_{P,C}(\mu) := F_{P,C}(C\mu) - \mu$ and $M_{P,\frac{1}{C}}$ is defined as the $\frac{1}{C}$ median of P, that is, $M_{P,C} := \sup\left\{M \geq 0 \mid \mathbb{P}_{t\sim P}[t \geq M] \geq \frac{1}{C}\right\}$. Since the cumulative density function $\mathbb{P}_{t\sim P}[t \geq M]$ is left continuous in M, it holds that $\mathbb{P}_{t\sim P}[t \geq M_{P,C}] \geq \frac{1}{C}$.

Let $P_{\boldsymbol{x}}$ denote the distribution of $\|\nabla L_\gamma(\boldsymbol{x})\|_2^2$. Below is a mild assumption saying $P_{\boldsymbol{x}}$ is universally well-concentrated from below in the sense that the mean of the smallest $(1-\frac{1}{C})$ part of $P_{\boldsymbol{x}}$ is at least a constant fraction of the C-clipped mean of $P_{\boldsymbol{x}}$. Since $\mu_{P_{\boldsymbol{x}},C} \leq \mu_{\boldsymbol{x}}$, the assumption below holds whenever $\alpha_C \mu_{\boldsymbol{x}} \leq \mathbb{E}_{t\sim P_{\boldsymbol{x}}}[t\mathbb{1}[t < M_{P_{\boldsymbol{x}},\frac{1}{C}}]]$.

Assumption 4.7. $\exists \alpha_C > 0$, such that for all $\boldsymbol{x} \neq 0$, $\alpha_C \cdot \mu_{P_{\boldsymbol{x}},C} \leq \mathbb{E}_{t\sim P_{\boldsymbol{x}}}[t\mathbb{1}[t < M_{P_{\boldsymbol{x}},\frac{1}{C}}]]$.

We further define $\underline{\mu}_C := \min_{\|\boldsymbol{x}\|_2=1} \mu_{P_{\boldsymbol{x}},C}$ and $\overline{\mu}_C := \max_{\|\boldsymbol{x}\|_2=1} \mu_{P_{\boldsymbol{x}},C}$ and have the following theorem:

Theorem 4.8 ($\sqrt{C}$-Clipped SGD+WD)**.** *Let $\boldsymbol{x}(t)$ be defined by $\sqrt{C}$-Clipped* SGD *+WD (Algorithm 1). Under Assumption 4.7, for $\eta\lambda = O(\min\{1, \frac{\alpha_C}{C \ln T/\delta^2}\})$, with probability $1 - 5\delta$, we have*

$$\forall T' \leq t \leq T-1, \quad \frac{\underline{\mu}_C}{2} \leq \frac{2\lambda}{\eta} \|\boldsymbol{x}(t)\|_2^4 \leq 2\overline{\mu}_C. \tag{8}$$

and

$$\begin{aligned} &\frac{1}{T-T'} \sum_{t=T'}^{T-1} \left\langle \nabla L(\overline{\boldsymbol{x}}(t)), \widetilde{\nabla L}(\boldsymbol{x}(t)) \right\rangle \\ \leq &\frac{\pi^2 \rho \sqrt{\overline{\mu}_C}}{(T-T')\sqrt{2\eta\lambda}} + 4\sqrt{\eta\lambda} \frac{\rho \overline{\mu}_C^{\frac{3}{2}}}{\underline{\mu}_C} \\ &+ \sqrt{\frac{\ln \frac{2}{\delta}}{T-T'}} 8 \frac{\pi \rho \overline{\mu}_C^2}{\underline{\mu}_C} + \sqrt{\frac{\ln \frac{2}{\delta}}{T-T'}} 16\sqrt{\lambda\eta} \frac{\rho \overline{\mu}_C^3}{\underline{\mu}_C^2}. \end{aligned} \tag{9}$$

where $T' = \frac{1}{\alpha_C \eta\lambda} \max\left\{\ln \frac{R_0^2}{\overline{\mu}_C}, \ln \frac{\underline{\mu}_C}{R_0^2}\right\} + O(1)$ *and* $\widetilde{\nabla L}(\boldsymbol{x}) := \mathbb{E}\left[\nabla L_\gamma(\overline{\boldsymbol{x}}) \min\left\{\sqrt{\frac{2C\lambda}{\eta}} \frac{\|\boldsymbol{x}\|_2^2}{\|\nabla L_\gamma(\overline{\boldsymbol{x}})\|_2}, 1\right\}\right]$.

The proof of this theorem is presented in Appendix F. Note that with clipping Theorem 4.8 shows that the norm convergence (8) is more robust as it doesn't need to make any assumption about the maximum gradient norm M, unlike Theorem 4.5. Indeed, from the definition of C-clipped mean, for each $\boldsymbol{x}$, we can allow all the gradients with norm larger than $C \cdot \mu_{P_{\boldsymbol{x}},C}$ to become infinity, and yet not affect the norm convergence, as $\mu_{P_{\boldsymbol{x}},C}$ and the condition in Assumption 4.7 do not change.

Under the additional assumption that $\left\langle \nabla L(\overline{\boldsymbol{x}(t)}), \widetilde{\nabla L}(\boldsymbol{x}(t)) \right\rangle = \Omega(\|\nabla L(\boldsymbol{x}(t))\|_2^2)$, we can use Equation (9) to show convergence to stationary points. This is a reasonable assumption if the clipping frequency is low, *e.g.*, it's 1.5% in our experiments for SIBERT.

5. Experiments

We now conduct a comprehensive empirical study in order to demonstrate the following key aspects of our recipe: (i) yields competitive training performance using significantly low memory footprint, (ii) training becomes highly robust to initialization scale, and (iii) provides better convergence of norm with clipping.

Experimental Setup. We consider the standard task of pretraining a transformer model and fine-tuning it on benchmark datasets, following Devlin et al. (2018). We compare its performance with SIBERT, a scale invariant version of BERT as described in Sec. 3.1. For both these models, we use their base size versions unless specified otherwise. For SIBERT, the scale invariant portion is trained using SGD+WD with a piecewise constant LR schedule and WD of $1e-2$. We use LAMB optimizer for the non-scale invariant parts. The initial LR for SGD is $8e-4$ without warmup and is divided by 10 at step 600k and 900k. Default training is for $1M$ steps. For LAMB we use a linear decay schedule with initial learning rate $8e-4$ and a linear warmup of 10k steps.

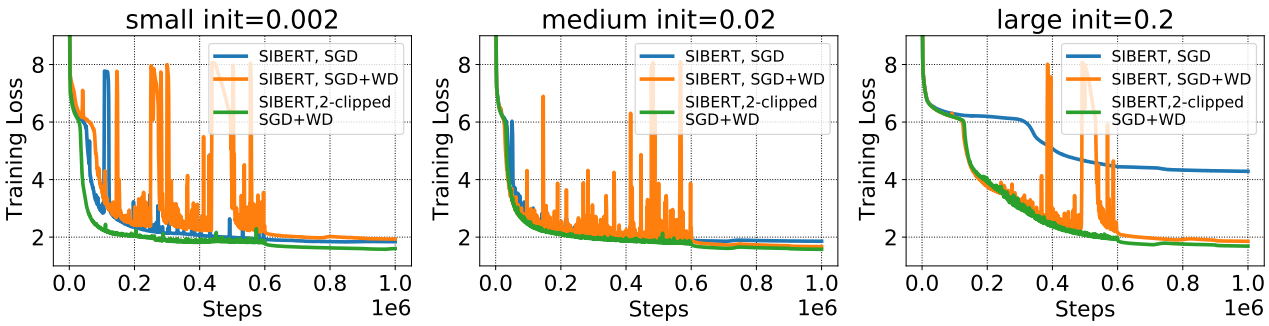

Figure 1: SGD+WD optimizes the scale invariant training loss of SIBERT robustly for all initialization scales, and thus for loss scalings and different learning rates (with $\lambda\eta$ fixed). Here the default initialization for parameters in SIBERT encoder is a truncated normal distribution with standard deviation equal to 0.02 (the same as BERT).

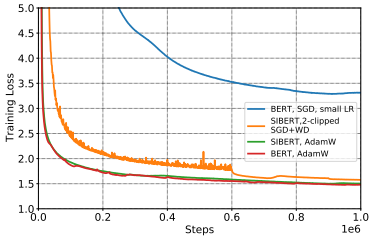

Figure 2: Our recipe (SIBERT, SGD+WD and Relative Global Clipping) significantly improves the optimization performance compared to the baseline, BERT trained by SGD with small LR. The final training loss is close to BERT trained by ADAM.

Performance. We begin by establishing that proposed SIBERT with SGD+WD training performs competitively. In this regard, we first look at pretraining loss between standard training of BERT with ADAM and our SIBERT trained by SGD+WD with or without clipping (the clipping factor is set as $\sqrt{C} = 2$). From Figure 2, one can see that our training curve closely follows that of BERT trained by ADAM, but without the need for extra memory for keeping track of first and second order momentum. If we use SGD on standard BERT architecture, then either we have to use small learning rates, which slows down training, or the loss diverges. This further highlights the importance of the scale invariant architecture, which improves training stability by eliminating the k-homogeneous structure. To our knowledge, this is the first work that shows effective training of BERT-like model using simple SGD (even without any momentum).

Next, we compare the downstream performance on three benchmark datasets (SQuADv1.1 (Rajpurkar et al., 2016), SQuADv2 (Rajpurkar et al., 2018) and MNLI (Williams et al., 2018)). We tried to follow standard setup, e.g. BERT is finetuned by ADAM. However for SIBERT we had to use LAMB, as ADAM is very sensitive to the scale. We observe comparable performance and when trained longer it can even outperform conventional BERT.

Table 1: Downstream Performance of SIBERT trained by SGD+WD +clipping is close to that of BERT trained ADAM- which uses $3X$ memory for saving (auxiliary) parameters than SGD. The gap is further reduced by doubling the training budget of SIBERT.

		MNLI Acc	SQuAD1 F1	SQuAD2 F1	Pretraining Loss
Base	BERT	**84.4**	**90.3**	78.8	**1.479**
	SIBERT	81.1	88.1	74.8	1.672
	+ clipping	82.6	89.3	76.8	1.58
	+ 2x training	83.3	**90.3**	**80.0**	1.495
Large	BERT	**86.8**	**92.4**	**84.1**	**1.181**
	SIBERT	83.7	90.6	79.3	1.404
	+ clipping	85.3	91.6	81.3	1.322
	+ 2x training	86.4	**92.4**	83.1	1.194

Training Stability: Insensitivity to the scale of initialization. To showcase ease of optimization offered by our recipe, we consider different initialization scales spanning two orders of magnitude. The results for the pretraining task in Figure 1 show good convergence across the board for our approach, whereas SGD on its own struggles even with the scale invariant architecture.

Further note that these experiments simultaneously showcase robustness to rescaling of loss, parameterization, or LR. This is because in a scale invariant model trained by SGD+WD (+clipping), it holds that all of following scalings are equivalent: $(c_1 L, c_2 \boldsymbol{x}(0), c_3\eta, c_4\lambda) \longleftrightarrow (L, \frac{c_2}{\sqrt{c_1 c_3}}\boldsymbol{x}(0), \eta, c_3 c_4 \lambda)$ for any $c_1, c_2, c_3, c_4 > 0$.

Training Stability: Improvement in parameter norm convergence. Finally, we look at parameter norms dur-

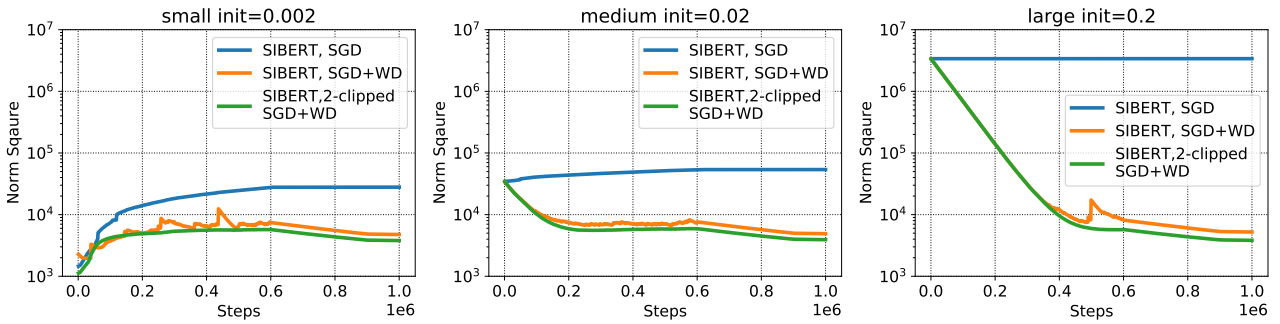

Figure 3: The robust optimization performance of SGD+WD over the scale invariant training loss of SIBERT originates from its ability to fast adjust the parameter norm. In contrast, when the initial norm is too large, SGD w.o. WD optimizes slowly. Relative Global Clipping reduces the spikes in the norm curve, which verifies our theoretical result Theorem 4.8 that clipping leads to better norm convergence. Here, only the norm of the scale invariant part, *i.e.*, the encoder part is plotted.

ing training in experiments. We observe that even when starting from very different initialization scale, SGD+WD (+clipping) quickly brings parameter norm to desired ranges. In contrast, SGD struggles when initial norm and learning rate are not aligned - see the rightmost plot with large initialization in Figure 3. This shows that our recipe has the ability to quickly adapt to different initialization scales, in-line with our theoretical result (Theorem 4.8) showing better norm convergence of SGD+WD (+clipping).

6. Conclusion

In this paper, we presented a simple yet effective method to robustly train transformers with non-adaptive methods such as SGD. By designing novel scale invariant architecture and using a tailored optimization procedure — which makes our optimization scheme truly *architecture aware* — we provably achieve robust training of neural networks with substantially low memory footprint when compared to adaptive methods. We believe designing neural architecture and the optimizer jointly is an exciting research direction and will yield even better training procedures in the future.

References

Anil, R., Gupta, V., Koren, T., and Singer, Y. Memory efficient adaptive optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Arora, S., Li, Z., and Lyu, K. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations*, 2018.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Chen, X., Wu, Z. S., and Hong, M. Understanding gradient clipping in private SGD: A geometric perspective. *CoRR*, abs/2006.15429, 2020. URL https://arxiv.org/abs/2006.15429.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Hazan, E., Levy, K., and Shalev-Shwartz, S. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1594–1602, 2015.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Hoffer, E., Banner, R., Golan, I., and Soudry, D. Norm matters: efficient and accurate normalization schemes in deep networks. *arXiv preprint arXiv:1803.01814*, 2018.

Huang, L., Liu, X., Lang, B., and Li, B. Projection based weight normalization for deep neural networks. *ArXiv*, abs/1710.02338, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Levy, K. Y. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Li, Z. and Arora, S. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2019.

Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020.

Liu, L., Liu, X., Gao, J., Chen, W., and Han, J. Understanding the difficulty of training transformers. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 5747–5763. Association for Computational Linguistics, 2020.

Lobacheva, E., Kodryan, M., Chirkova, N., Malinin, A., and Vetrov, D. P. On the periodic behavior of neural network training with batch normalization and weight decay. *Advances in Neural Information Processing Systems*, 34, 2021.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/pascanu13.html.

Polyak, B. T. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of ADAM and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29:901–909, 2016.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2):26–31, 2012.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

van Handel, R. Probability in high dimension. 2016.

Van Laarhoven, T. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wan, R., Zhu, Z., Zhang, X., and Sun, J. Spherical motion dynamics: Learning dynamics of neural network with normalization, weight decay, and sgd. *arXiv preprint arXiv:2006.08419*, 2020.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018.

Wu, Y. and He, K. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

You, Y., Li, J., Reddi, S. J., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C. Large batch optimization for deep learning: Training BERT in 76 minutes. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Zhang, G., Wang, C., Xu, B., and Grosse, R. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2018.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S. J., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

A. Design Details of Scale Invariant BERT

Definition A.1. For a module with n inputs and m outputs, we say the module is $(a_1, ...a_n; b_1, ..., b_m)$-homogeneous if the m outputs are b_i-homogeneous to the network parameters whenever the n inputs are a_i-homogeneous to the network parameters. A model is scale invariant iff its output is $(; 0)$-homogeneous. (A complete model doesn't take any input from another module)

Following (Li & Arora, 2019), we view the computation graph as a directed acyclic graph, where each module is a node and each tensor (including inputs, intermediate computation results and final output) as an edge. Each edge can be viewed as a function of parameters, and we can decide the homogeneity by doing induction over the computation graph by its topological order. In detail, we know the jth output edge of some $(a_1, \ldots, a_n; b_1, ldots, b_n)$- homogeneous module is b_j homogeneous if for each $1 \leq i \leq n$, the ith input edge is a_i-homogeneous. For convenience, we allow a_i,b_i to be functions of free variable x, meaning the module is $(a1(x), \ldots, a_n(x); b_1(x), \ldots, b_m(x))$-homogeneous for every $x \in \mathbb{R}$.

In Table 2, we summarize the homogeneity of building blocks in our design.

Overview of SIBERT structure: Our SIBERT has two main parts — encoder and classification head, which is the same to standard BERT. We only make encoder part scale invariant and train it by SGD+WD. We leave the classification head not scale invariant and train it by LAMB. Note the classification head is only used in pretraining and is not used in the downstream task.

(2;2)-homogeneous encoder layer: As mentioned in Appendix A, residual block and attention are the two main building blocks that needs to be made scale invariant. Following Li & Arora (2019), we choose to use PreNorm structure for residual block and make it $(2; 2)$-homogeneous. We also replace GeLU (Hendrycks & Gimpel, 2016) in BERT by ReLU for homogeneity. Since ReLU is $(1; 1)$ homogeneous, we omit ReLU from the design, without affecting the final scale invariance.

Table 2: Homogeneity of building blocks of SIBERT.

Symbol	Module	Homogeneity
I	Input	(0;1)
B	Adding Bias	(1;1)
N	Layer Normalization (no affine)	(x;0)
L	Linear Layer	(x;x+1)
Embed	Embedding Layer	(x;x+1)
NA	Layer Normalization with affine	(x;1)
FF	2-layer feedforward network	(0;2)
ATTN	Scale Invariant Attention	(x,x,x;x+2)
Encoder	Our Encoder Layer	(2;2)

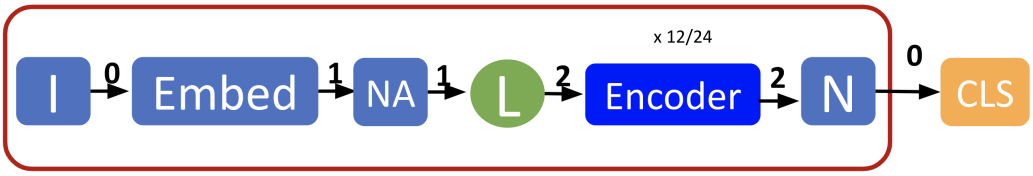

Figure 4: Encoder and Classification Head (CLS). 'x12/24' means to stack 12 our $(2; 2)$-homogeneous encoder layer for base SIBERT (or 24 for large SIBERT)

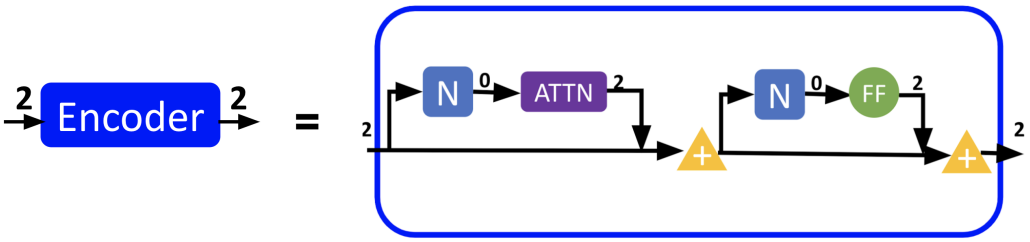

Figure 5: The $(2;2)$-homogeneous encoder layer. 'ATTN' denotes our Scale Invariant Attention (see Figure 7). 'FF' denotes the 2-layer feedforward structure, which is $(0;2)$-homogeneous.

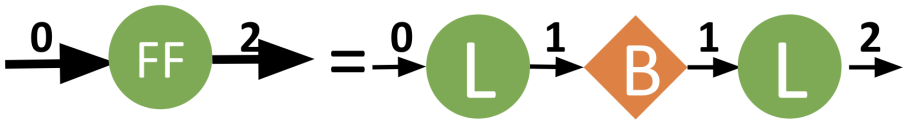

Figure 6: The $(0;2)$-homogeneous FeedForward layer

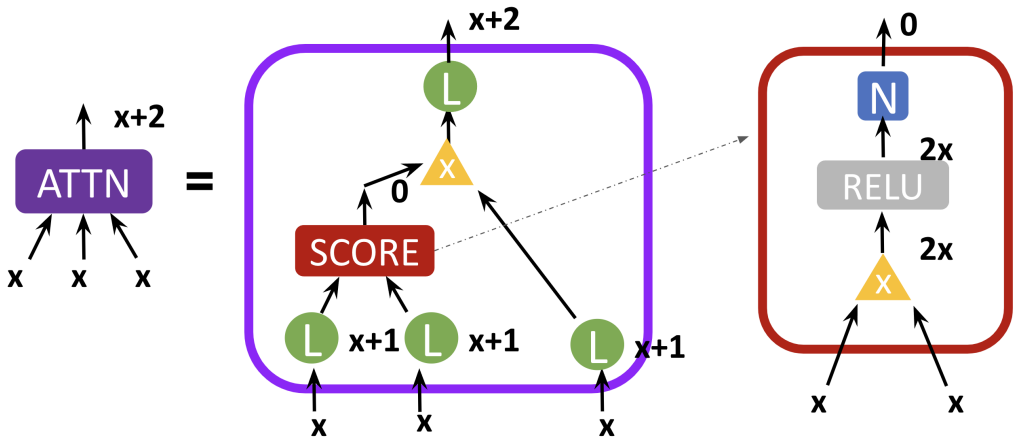

Figure 7: The $(x,x,x;x+2)$-homogeneous Attention, which is defined as Multi-Head-SI-Attention$(Q,K,V) = \sum_i \mathrm{N}(\mathrm{ReLU}(QW_i^Q(KW_i^K)^\top)VW_i^VW_i^O$, where $W_i^Q, W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_k\times d_v}$ and $W_i^O \in \mathbb{R}^{d_v\times d_{model}}$. That is, if Q,K,V are k-homogeneous functions of parameter $\boldsymbol{x}$, then Multi-Head-SI-Attention(Q,K,V) is $k+2$-homogeneous, for any $k\in\mathbb{R}$. We also call it *Scale Invariant Attention* because its attention score is scale invariant.

B. Introduction examples analysis

In the first example, since the data is non-separable, the global optimum X^* must be finite and, thus, $|\nabla\widetilde{L}(X)|$ is positive and monotone increases among all $X > X^* > 0$. For simplicity, assume $X^* > 0$ and $x_1 = \cdots = x_{2k} > (X^*)^{\frac{1}{2k}}$ at initialization (and thus at any iteration t). It holds that $x_i(t+1) = x_i(t) - \eta\frac{X(t)}{x_i(t)}\nabla\widetilde{L}(X(t)) = x_i(t)\left(1-\eta\frac{X(t)}{x_i^2(t)}\nabla\widetilde{L}(X(t))\right)$, where $X(t) = \Pi_{j=1}^{2k}x_j(t)$. This implies $X(t+1) = X(t)\left(1-\eta\frac{X(t)}{\sqrt[k]{X(t)}}\nabla\widetilde{L}(X(t))\right)^{2k} \geq 0$. Thus we conclude if $\eta \geq \frac{2}{|\nabla\widetilde{L}(X(0))|}(X(0))^{\frac{1}{k}-1}$ and $X(0) > X^*$, $X(t)$ will increase monotonically and explode.

C. Useful Lemmas

C.1. Scale Invariance

Lemma C.1 (Smoothness). *For any $\boldsymbol{v}, \boldsymbol{x} \in \mathbb{R}^d$ with $\langle \boldsymbol{x}, \boldsymbol{v} \rangle = 0$, suppose L is scale-invariant and twice differentiable with $\rho := \max_{\|\boldsymbol{x}\|_2=1} \left\|\nabla^2 L(\boldsymbol{x})\right\|$, we have*

$$L(\boldsymbol{x}+\boldsymbol{v}) - L(\boldsymbol{x}) \leq \langle \boldsymbol{v}, \nabla L(\boldsymbol{x}) \rangle + \frac{\rho \left\|\boldsymbol{v}\right\|_2^2}{2 \left\|\boldsymbol{x}\right\|_2^2}.$$

Proof of Lemma C.1. Define $\gamma(s) = \boldsymbol{x} + s\boldsymbol{v}$, then we have $L(\gamma(0)) = L(\boldsymbol{x})$ and $L(\gamma(1)) = L(\boldsymbol{x}+\boldsymbol{v})$. Taking Taylor expansion of $F(s) = L(\gamma(s))$ at $s = 0$, we have

$$F(1) - F(0) = F'(0) + \frac{F''(s^*)}{2}, \quad \text{for some } s^* \in [0,1]. \tag{10}$$

Note $F'(0) = \langle \gamma'(0), \nabla L(\gamma(0)) \rangle = \langle \nabla L(\boldsymbol{x}), \boldsymbol{v} \rangle$ and

$$F''(s^*) = \gamma'(s^*) \nabla^2 L(\gamma(s^*)) \gamma'(s^*) \leq \frac{\rho}{\left\|\gamma(s^*)\right\|_2^2} \left\|\gamma'(s^*)\right\|_2^2, \tag{11}$$

where the last inequality uses the fact that L is scale invariant. The proof is completed by noting that $\left\|\gamma(s^*)\right\|_2 \geq \left\|\gamma(0)\right\|_2 = \left\|\boldsymbol{x}\right\|_2^2$ and that $\gamma'(s^*) = \boldsymbol{v}$.

□

Lemma C.2 (Smoothness, Multi-group). *For any $\boldsymbol{v}, \boldsymbol{x} \in \mathbb{R}^d$ with $\langle \boldsymbol{x}_k, \boldsymbol{v}_k \rangle = 0$ for all $k \in [K]$, suppose L is multi-group scale invariant (see Definition G.1), we have*

$$L(\boldsymbol{x}+\boldsymbol{v}) - L(\boldsymbol{x}) \leq \langle \boldsymbol{v}, \nabla L(\boldsymbol{x}) \rangle + \frac{\rho}{2} \sum_{k=1}^{K} \frac{\left\|\boldsymbol{v}_i\right\|_2^2}{\left\|\boldsymbol{x}_i\right\|_2^2}.$$

Proof of Lemma C.2. We first prove for the case where $\left\|\boldsymbol{x}_k\right\|_2 = 1, \forall k \in [K]$. Similar to the proof of Lemma C.1, it suffices to show that the smoothness of L is at most ρ along the line joining $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{v}$. This holds because $\forall s \in [0,1], k \in [K]$, $\left\|\boldsymbol{x}_i + s\boldsymbol{v}_i\right\|_2 \geq \left\|\boldsymbol{x}_i\right\|_2$ by assumption that $\langle \boldsymbol{x}_k, \boldsymbol{v}_k \rangle = 0$ for all $k \in [K]$.

Now we turn to the general case. Define $\widehat{\boldsymbol{x}} = [\frac{\boldsymbol{x}_1^\top}{\|\boldsymbol{x}_1\|_2}, \ldots, \frac{\boldsymbol{x}_K^\top}{\|\boldsymbol{x}_K\|_2}]^\top$ and $\boldsymbol{v}' = [\frac{\boldsymbol{v}_1^\top}{\|\boldsymbol{x}_1\|_2}, \ldots, \frac{\boldsymbol{v}_K^\top}{\|\boldsymbol{x}_K\|_2}]^\top$. Since L is multi-group scale invariant, we have $L(\boldsymbol{x}) = L(\widehat{\boldsymbol{x}})$ and $L(\boldsymbol{x}+\boldsymbol{v}) = L(\widehat{\boldsymbol{x}}+\boldsymbol{v}')$. The proof is completed by applying the previous argument on $\widehat{\boldsymbol{x}}$ and $\boldsymbol{v}'$. □

Lemma C.3. *If L is scale invariant, $\left\|\nabla L(\boldsymbol{x})\right\|_2 \leq \frac{\pi}{\|\boldsymbol{x}\|_2} \sup_{\|\boldsymbol{x}\|=1} \left\|\nabla^2 L(x)\right\|_2$.*

Proof of Lemma C.3. It suffices to prove the above bound for all $\boldsymbol{x}$ with $\left\|\boldsymbol{x}\right\|_2 = 1$. Let $\boldsymbol{x}^*$ be any local minimizer of L on $\mathbb{S}^{d-1}$ and $\gamma : [0,1] \to \mathbb{S}^{d-1}$ be the geodesic curve satisfying that $\gamma(0) = \boldsymbol{x}^*$ and $\gamma(1) = \boldsymbol{x}$. We know the length of $\{\gamma(t)\}_{t=0}^1 \leq \pi$ and thus

$$\left\|\nabla L(\boldsymbol{x})\right\| = \left\| \int_{t=0}^1 \nabla^2 L(\gamma(t)) \frac{d\gamma(t)}{dt} dt \right\| \leq \int_{t=0}^1 \left\|\nabla^2 L(\gamma(t))\right\|_2 \left\| \frac{d\gamma(t)}{dt} \right\|_2 dt \leq \rho \cdot \pi \tag{12}$$

□

Lemma C.4. *If L is scale invariant, $\sup_{\boldsymbol{x},\boldsymbol{x}'} L(\boldsymbol{x}) - L(\boldsymbol{x}') \leq \frac{\pi^2}{2} \sup_{\|\boldsymbol{x}\|=1} \left\|\nabla^2 L(x)\right\|_2$.*

Proof of Lemma C.4. Similar to the proof of Lemma C.3. □

C.2. Probablity

Definition C.5. A random variable $X \in \mathbb{R}$ is said to be *sub-Gaussian* with variance proxy σ^2 (denoted by $X \sim \mathsf{subG}(\sigma^2)$) if its moment generating function satisfies

$$\mathbb{E}[\exp(sX)] \leq \exp(\frac{\sigma^2 s^2}{2}), \forall s \in \mathbb{R}. \tag{13}$$

In this manuscript, we also also use the following notion of *conditional subgaussian*. We say a random variable $X \in \mathbb{R}$ is said to be *sub-Gaussian* with variance proxy σ^2 conditioned on event $\mathcal{E}$ (denoted by $X \sim \mathsf{subG}(\sigma^2, \mathcal{E})$) if its moment generating function satisfies

$$\mathbb{E}[\exp(sX)\mathbb{1}[\mathcal{E}]] \leq \exp(\frac{\sigma^2 s^2}{2}), \forall s \in \mathbb{R}. \tag{14}$$

Lemma C.6 (Chernoff Bound with Conditioning)**.** *Let $X \sim \mathsf{subG}(\sigma^2, \mathcal{E})$. Then for any $t > 0$, it holds that*

$$\mathbb{P}[X > t \wedge \mathcal{E}] \leq \exp(-\frac{t^2}{2\sigma^2}), \quad \textit{and} \quad \mathbb{P}[X < -t \wedge \mathcal{E}] \leq \exp(-\frac{t^2}{2\sigma^2}) \tag{15}$$

When $\mathbb{P}[\mathcal{E}] = 1$, we get the standard Chernoff bound. Let $X \sim \mathsf{subG}(\sigma^2)$. Then for any $t > 0$, it holds that

$$\mathbb{P}[X > t] \leq \exp(-\frac{t^2}{2\sigma^2}), \quad \textit{and} \quad \mathbb{P}[X < -t] \leq \exp(-\frac{t^2}{2\sigma^2}) \tag{16}$$

Proof of Lemma C.6. For any $s > 0$, we have

$$\mathbb{P}[X > t \wedge \mathcal{E}] = \mathbb{P}[e^{sX} \geq e^{st} \wedge \mathcal{E}] \leq e^{-st}\mathbb{E}[e^{sX}\mathbb{1}[\mathcal{E}]] = \exp(-st + \frac{\sigma^2 s^2}{2}). \tag{17}$$

The proof is completed by picking $s = \frac{t}{\sigma^2}$.

□

We will use $(\Omega, \Sigma, \mathbb{P})$ to note the probability space and $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ to denote the filtration

Lemma C.7 (Azuma Inequality with Conditioning)**.** *Let $\mathcal{E}_t \in \mathcal{F}_t$ and $\mathcal{E}_{t+1} \subset \mathcal{E}_t$ for all $t \geq 0$. Let $\{X_t\}_{t\geq 1}$ be a martingale difference sequence and $\mathsf{subG}(\sigma_t^2, \mathcal{E}_{t-1})$ conditioned on $\mathcal{F}_{t-1}$,* i.e.*, $\mathbb{E}[\exp(sX_t)\mathbb{1}[\mathcal{E}_{t-1}] \mid \mathcal{F}_{t-1}] \leq \exp(\frac{s^2\sigma_t^2}{2})$ for all $t \geq 0$. Then $\sum_{i=1}^T X_i$ is $\mathsf{subG}(\sum_{t=0}^{T-1}\sigma_t^2, \mathcal{E}_{T-1})$.*

Proof. We will prove by induction on T. When $T = 1$, the statement is true by assumption. Now suppose the statement holds for $T-1$, we have for any $s > 0$

$$\begin{aligned}
\mathbb{E}[\exp(s\sum_{i=1}^{T} X_i)\mathbb{1}[\mathcal{E}_{T-1}]] =& \mathbb{E}[\exp(s\sum_{i=1}^{T-1} X_i)\mathbb{1}[\mathcal{E}_{T-1}]\mathbb{E}[\exp(sX_T)\mathbb{1}[\mathcal{E}_{t-1}] \mid \mathcal{F}_{T-1}]] \\
\leq& \mathbb{E}[\exp(s\sum_{i=1}^{T-1} X_i)\mathbb{1}[\mathcal{E}_{T-1}]\exp(\frac{s^2\sigma_{T-1}^2}{2})] \\
\leq& \mathbb{E}[\exp(s\sum_{i=1}^{T-1} X_i)\mathbb{1}[\mathcal{E}_{T-2}]]\exp(\frac{s^2\sigma_{T-1}^2}{2})
\end{aligned}$$

Thus we have

$$\mathbb{E}[\exp(s\sum_{i=1}^{T} X_i)\mathbb{1}[\mathcal{E}_{T-1}]] \leq \exp(\frac{s^2\sum_{t=0}^{T-1}\sigma_t^2}{2}). \tag{18}$$

□

C.3. Others

Lemma C.8. $\forall t \in \mathbb{N}, k \in \mathbb{N}^+, 0 < x < 1$,

$$\sum_{\tau=0}^{t}(1-x)^{k\tau} \le \frac{e^{kx}}{kx}$$

Proof of Lemma C.8.

$$\sum_{\tau=0}^{t}(1-x)^{k\tau} \le \sum_{\tau=0}^{\infty}(1-x)^{k\tau} \le \sum_{\tau=0}^{\infty} e^{-kx\tau} = \frac{1}{1-e^{-kx}} \le \frac{e^{kx}}{kx},$$

where the last step is because $e^x \ge 1 + x, \forall x \in \mathbb{R}$. □

D. Omitted Proofs for the Convergence of GD

Proof of Lemma 4.2. This is a special case of Lemma C.1 with $\boldsymbol{x} = (1-\eta\lambda)\boldsymbol{x}(t)$ and $\boldsymbol{v} = -\eta\nabla L(\boldsymbol{x}(t))$. Here we use the assumption that L is scale invariant, ∇L is -1-homogeneous. By Lemma 2.3, which means $\nabla L(\boldsymbol{x}) = \frac{\nabla L(\boldsymbol{x}(t))}{1-\eta\lambda}$. □

The following lemma deals with the case where $\|\boldsymbol{x}(0)\|_2^2 < \pi^2\rho\eta$.

Lemma D.1. *Let* $I = \{T' \in \mathbb{N} \mid \forall 0 \le t \le T',\ \|\boldsymbol{x}(t)\|_2^2 \le \pi^2\rho\eta \ \wedge\ \|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2 > 8\pi^4\rho^2\lambda\eta\}$. *Suppose* $0 \in I$ *and* $T = \max I$. *Then* $T \le \frac{1}{6\lambda\eta}$ *and* $\|\boldsymbol{x}(T+1)\|_2^2 \le \frac{2(\pi^2\rho\eta)^2}{\|\boldsymbol{x}(0)\|_2^2}$.

Proof of Lemma D.1. For any $t \le T$, we have

$$\begin{aligned}
\|\boldsymbol{x}(t+1)\|_2^2 - \|\boldsymbol{x}(t)\|_2^2 =& ((1-\lambda\eta)^2 - 1)\|\boldsymbol{x}(t)\|_2^2 + \eta^2\|\nabla L(\boldsymbol{x}(t))\|_2^2 && (19)\\
\ge& -2\lambda\eta\|\boldsymbol{x}(t)\|_2^2 + \frac{\eta^2\|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2}{\|\boldsymbol{x}(t)\|_2^2} && (20)\\
\ge& -2\pi^2\rho\lambda\eta^2 + 8\pi^2\rho\lambda\eta^2 && (21)\\
=& 6\pi^2\rho\lambda\eta^2. && (22)\\
& && (23)
\end{aligned}$$

Thus $6\pi^2\rho\lambda\eta^2 \cdot T \le \|\boldsymbol{x}(T)\|_2^2 - \|\boldsymbol{x}(0)\|_2^2 < \|\boldsymbol{x}(T)\|_2^2 \le \pi^2\rho\eta$, which implies that $T < \frac{1}{6\lambda\eta}$. Moreover,

$$\begin{aligned}
\|\boldsymbol{x}(T+1)\|_2^2 =& (1-\eta\lambda)^2\|\boldsymbol{x}(T)\|_2^2 + \eta^2\|\nabla L(\boldsymbol{x}(T))\|_2^2 && (24)\\
\le& \|\boldsymbol{x}(T)\|_2^2 + \frac{\eta^2\|\nabla L(\overline{\boldsymbol{x}}(T))\|_2^2}{\|\boldsymbol{x}(T)\|_2^2} && (25)\\
\le& \|\boldsymbol{x}(T)\|_2^2 + \frac{\eta^2\|\nabla L(\overline{\boldsymbol{x}}(T))\|_2^2}{\|\boldsymbol{x}(0)\|_2^2} && (26)\\
\le& \pi^2\rho\eta + \frac{\rho^2\pi^2\eta^2}{\|\boldsymbol{x}(0)\|_2^2} && (27)\\
\le& \frac{2(\pi^2\rho\eta)^2}{\|\boldsymbol{x}(0)\|_2^2} && (28)
\end{aligned}$$

□

Theorem D.2 (convergence rate of GD+WD)**.** *Suppose* $\eta\lambda \le \frac{1}{2}$. *Let* $\boldsymbol{x}(t)$ *be the* t*-th iterate of GD* (3)*, and* $T_0 = \left\lceil \frac{1}{2\eta\lambda} \ln \frac{2\|\boldsymbol{x}(0)\|_2^2}{\rho\pi^2\eta} \right\rceil$. *If* $\|\boldsymbol{x}(0)\|_2^2 \ge \pi^2\rho\eta$*, we have*

$$\min_{t=0,\dots,T_0} \|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2 \le 8\pi^4\rho^2\lambda\eta. \tag{29}$$

Proof of Theorem D.2. We first claim there's $0 \le t \le T_0$, such that $\|\boldsymbol{x}(t)\|_2^2 < \pi^2\rho\eta$.

Otherwise, by Lemma 4.2, for $t = 0, \dots, T_0$, we have $L(\boldsymbol{x}(t)) - L(\boldsymbol{x}(t+1)) \le \frac{\eta}{2}\|\nabla L(\boldsymbol{x}(t))\|_2^2$. Note that $\|\boldsymbol{x}(t+1)\|_2^2 - (1-\eta\lambda)^2\|\boldsymbol{x}(t)\|_2^2 = \eta^2\|\nabla L(\boldsymbol{x}(t))\|_2^2$.

Therefore,

$$\|\boldsymbol{x}(T_0)\|_2^2 - (1-\eta\lambda)^{2T_0}\|\boldsymbol{x}(0)\|_2^2 = \sum_{t=0}^{T_0-1}\eta^2(1-\eta\lambda)^{2(T_0-t)}\|\nabla L(\boldsymbol{x}(t))\|_2^2 \tag{30}$$

$$\le \sum_{t=0}^{T_0-1}\eta^2\|\nabla L(\boldsymbol{x}(t))\|_2^2 \tag{31}$$

$$\le \frac{\eta}{2}(L(\boldsymbol{x}(0)) - L(\boldsymbol{x}_{T_0-1})) \tag{32}$$

$$\le \frac{\eta\pi^2\rho}{2} \tag{33}$$

By the definition of T_0, we have $(1-\eta\lambda)^{2T_0}\|\boldsymbol{x}(T_0)\|_2^2 \le e^{-2\eta\lambda T_0}\|\boldsymbol{x}(0)\|_2^2 \le \frac{\eta\pi^2\rho}{2}$. Thus $\|\boldsymbol{x}(T_0)\| \le \pi^2\rho\eta$.

Without loss of generality, we let T be the smallest integer such that $\|\boldsymbol{x}(T)\|_2^2 < \pi^2\rho\eta$. By assumption, $T \ge 1$. Therefore $\|\boldsymbol{x}(T-1)\|_2^2 \ge \pi^2\rho\eta$. Because $\|\boldsymbol{x}(T)\|_2^2 = (1-\eta\lambda)^2\|\boldsymbol{x}(T-1)\|_2^2 + \eta^2\|\nabla L(\boldsymbol{x}(T-1))\|_2^2$, we have

$$\|\nabla L(\overline{\boldsymbol{x}}(T-1))\|_2^2 = \|\nabla L(\boldsymbol{x}(T-1))\|_2^2\|\boldsymbol{x}(T-1)\|_2^2 \le \eta^{-2}\left(\|\boldsymbol{x}(T)\|_2^2 - (1-\eta\lambda)^2\|\boldsymbol{x}(T-1)\|_2^2\right)\|\boldsymbol{x}(T-1)\|_2^2. \tag{34}$$

Note that $\|\boldsymbol{x}(T)\|_2^2 < \pi^2\rho\eta$ and $\frac{\|\boldsymbol{x}(T)\|_2^2}{(1-\lambda\eta)^2} \ge \|\boldsymbol{x}(T-1)\|_2^2 \ge \pi^2\rho\eta$, we conclude

$$\|\nabla L(\overline{\boldsymbol{x}}(T-1))\|_2^2 \le \eta^{-2}\left(\|\boldsymbol{x}(T)\|_2^2 - (1-\eta\lambda)^2\|\boldsymbol{x}(T-1)\|_2^2\right)\frac{\|\boldsymbol{x}(T)\|_2^2}{(1-\lambda\eta)^2} \tag{35}$$

$$\le \frac{1-(1-\lambda\eta)^2}{\eta^2(1-\lambda\eta)^2}(\pi^2\rho\eta)^2 \tag{36}$$

$$\le 8\lambda\eta\pi^4\rho^2, \tag{37}$$

which completes the proof. □

Combining Lemma D.1 and Theorem D.2 removes the initial condition in Theorem D.2, and completes the proof of Theorem 4.1.

E. Omitted Proofs for Convergence Rate of SGD

We will use $(\Omega, \Sigma, \mathbb{P})$ to note the probability space and $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ to denote the filtration where $\mathcal{F}_t := \sigma(\{\gamma_i \mid 0 \le i \le t\})$ is the σ-algebra generated by $\gamma_0, \dots, \gamma_t$.

Lemma E.1. $\|\nabla L_\gamma(\boldsymbol{x})\|_2^2 - \mathbb{E}\|\nabla L_\gamma(\boldsymbol{x})\|_2^2 \sim \mathsf{subG}(\frac{M^4}{4\|\boldsymbol{x}\|_2^4})$.

Proof. Lemma E.1 Note $0 \le \|\nabla L_\gamma(\boldsymbol{x})\|_2^2 \le \frac{M^2}{\|\boldsymbol{x}\|_2^2}$. The proof is immediate by Hoeffding Lemma (see Lemma 3.6 in (van Handel, 2016)). □

Given a integer $T \ge 0$, let $\mathcal{E}_T$ be the event that $\forall 0 \le t' \le t \le T-1$,

$$\left|\sum_{\tau=t'}^{t}(1-\eta\lambda)^{4(t-\tau)}\left(\|\nabla L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 - \mathbb{E}[\|\nabla L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 \mid \overline{\boldsymbol{x}}(\tau)]\right)\right| \le e^{4\eta\lambda}\cdot\frac{M^2}{4}\sqrt{\frac{1}{\lambda\eta}\ln\frac{2T^2}{\delta}}. \tag{38}$$

Lemma E.2. *For any* $0 \le t' \le t \le T-1$,

$$\sum_{\tau=t'}^{t}(1-\eta\lambda)^{4(t-\tau)}\left(\|\nabla L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 - \mathbb{E}[\|\nabla L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 \mid \boldsymbol{x}(\tau)]\right) \sim \mathsf{subG}(\frac{e^{8\eta\lambda}M^4}{32}) \tag{39}$$

Thus we have $\mathbb{P}[\mathcal{E}_T] \ge 1-\delta$ *by Lemma C.6.*

Proof of Lemma E.2. Note that $\sum_{\tau=t'}^{t}(1-\eta\lambda)^{8(t-\tau)}\frac{M^4}{4} \le \frac{e^{8\eta\lambda}}{32}$ by Lemma C.8. Thus by Azuma Inequality and Lemma E.1, we have that the martingale

$$\sum_{\tau=t'}^{t}(1-\eta\lambda)^{4(t-\tau)}\left(\|\nabla L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 - \mathbb{E}[\|\nabla L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 \mid \boldsymbol{x}(\tau)]\right)$$

is $\frac{e^{8\eta\lambda}}{32}$-subgaussian.

By Lemma C.6, we have for any $\forall 0 \le t' \le t \le T-1$, Equation (117) holds with probability at least $\frac{\delta}{T^2}$. The proof is completed by applying union bound. □

Lemma E.3 (Norm Lower Bound)**.** *Under Condition 4.4 and additionally assume* $\eta\lambda \le \frac{1}{2}$. *On* $\mathcal{E}_T$, *it holds that for any* $t \ge 0$,

$$\eta^{-2}\|\boldsymbol{x}(t)\|_2^4 \ge \frac{1-\eta\lambda}{2\eta\lambda}(1-e^{-4t\eta\lambda(1-\eta\lambda)})\underline{\sigma}^2 - \frac{1}{2}(1-\eta\lambda)^2M^2e^{4\eta\lambda}\sqrt{\frac{1}{\lambda\eta}\ln\frac{2T^2}{\delta}} \tag{40}$$

When $\frac{\underline{\sigma}^2}{12\eta\lambda} \ge \frac{M^2}{2}e^{4\eta\lambda}\sqrt{\frac{1}{\lambda\eta}\ln\frac{2T^2}{\delta}}$, *the above condition is simplified into the following: on* $\mathcal{E}_T$ *for any* $\frac{1}{\eta\lambda} \le t \le T$,

$$\eta^{-2}\|\boldsymbol{x}(t)\|_2^4 \ge \frac{5(1-\eta\lambda)^2\underline{\sigma}^2}{12\eta\lambda} - \frac{(1-\eta\lambda)^2\underline{\sigma}^2}{6\eta\lambda} = \frac{(1-\eta\lambda)^2\underline{\sigma}^2}{4\eta\lambda}, \tag{41}$$

In the above inequality, we also used the fact that $1-e^{-4(1-\eta\lambda)} \ge \frac{5}{6}$, *which is implied by* $\eta\lambda \le 0.5$.

Proof of Lemma E.3. Since L_γ is scale invariant, by Theorem 2.2, we have

$$\|\boldsymbol{x}(t+1)\|_2^2 = (1-\eta\lambda)^2\|\boldsymbol{x}(t)\|_2^2 + \eta^2\frac{\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2}{\|\boldsymbol{x}(t)\|_2^2}. \tag{42}$$

Squaring both sides of Equation (42), we have

$$\|\boldsymbol{x}(t+1)\|_2^4 = (1-\eta\lambda)^4\|\boldsymbol{x}(t)\|_2^4 + 2(1-\eta\lambda)^2\eta^2\|\nabla L_\gamma(\overline{\boldsymbol{x}}(t))\|_2^2 + \frac{\eta^4\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^4}{\|\boldsymbol{x}(t)\|_2^4}. \tag{43}$$

Thus

$$\begin{aligned}
\eta^{-2}\|\boldsymbol{x}(t+1)\|_2^4 \ge& 2\sum_{\tau=0}^{t}(1-\eta\lambda)^{4(t-\tau)+2}\|\nabla L_{\gamma_\tau}(\boldsymbol{x}(\tau))\|_2^2 \\
\ge& 2\sum_{\tau=0}^{t}(1-\eta\lambda)^{4(t-\tau)+2}\mathbb{E}\|\nabla L_{\gamma_\tau}(\boldsymbol{x}(\tau))\|_2^2 \\
&+2\sum_{\tau=0}^{t}(1-\eta\lambda)^{4(t-\tau)+2}\left(\|\nabla L_{\gamma_\tau}(\boldsymbol{x}(\tau))\|_2^2 - \mathbb{E}\|\nabla L_{\gamma_\tau}(\boldsymbol{x}(\tau))\|_2^2\right).
\end{aligned} \tag{44}$$

We also have

$$\sum_{\tau=0}^{t}(1-\eta\lambda)^{4(t-\tau)} \geq \sum_{\tau=0}^{t} e^{-4(t-\tau)\eta\lambda(1-\eta\lambda)} = \frac{1-e^{-4t\eta\lambda(1-\eta\lambda)}}{1-e^{-4\eta\lambda(1-\eta\lambda)}} \geq \frac{1-e^{-4t\eta\lambda(1-\eta\lambda)}}{4\eta\lambda(1-\eta\lambda)}. \tag{45}$$

Therefore, it holds that for any $t \geq 0$, conditioned on $\mathcal{E}_T$,

$$\eta^{-2}\left\|\boldsymbol{x}(t)\right\|_2^4 \geq \frac{1-\eta\lambda}{2\eta\lambda}(1-e^{-4t\eta\lambda(1-\eta\lambda)})\underline{\sigma}^2 - \frac{1}{2}(1-\eta\lambda)^2 M^2 e^{4\eta\lambda}\sqrt{\frac{1}{\lambda\eta}\ln\frac{2T^2}{\delta}} \tag{46}$$

□

Lemma E.4 (Norm upper bound)**.** *Under Condition 4.4 and additionally assume $\eta\lambda \leq 0.1$. Let $T_0 = \lceil \frac{1}{\eta\lambda} \rceil$. Let t^* be the earliest step t in $\{0,\dots,T_0-1\}$ that $\eta^{-2}\left\|\boldsymbol{x}(t)\right\|_2^4 \geq \frac{e^8(1-\eta\lambda)^2\underline{\sigma}^2}{4\eta\lambda}$ and we denote $t^* = T_0$ if this doesn't happen in $\{0,\dots,T_0-1\}$. For the case $t^* = T_0$, we have $\eta^{-2}\left\|\boldsymbol{x}(T_0)\right\|_2^4 \leq \frac{(1-\eta\lambda)^2\underline{\sigma}^2}{4\eta\lambda}$. On $\mathcal{E}_T$, for any $t \geq t^*$,*

$$\eta^{-2}\left\|\boldsymbol{x}(t+1)\right\|_2^4 \leq e^{-4\lambda\eta(t-t^*)}\max\left\{2M^2 e^{\left|\ln\frac{2e^4M^2}{\|\boldsymbol{x}(0)\|_2^4\eta^{-2}}\right|}, e^4\frac{\underline{\sigma}^2}{\eta\lambda}.\right\} + \frac{\overline{\sigma}^2}{\eta\lambda}. \tag{47}$$

Thus, there exists $T_1 = T_0 + \frac{1}{4\eta\lambda}\max\left\{\ln\frac{M^2\eta\lambda}{\overline{\sigma}^2} + \left|\ln\frac{2e^4M^2}{\|\boldsymbol{x}(0)\|_2^4\eta^{-2}}\right|, 4\right\}$, such that $\forall t \geq T_1$, $\eta^{-2}\left\|\boldsymbol{x}(t+1)\right\|_2^4 \leq \frac{2\overline{\sigma}^2}{\eta\lambda}$.

Proof of Lemma E.4. If $t^* < T_0$, it holds that conditioned on $\mathcal{E}_T$, for any $t^* \leq t < T_0$,

$$\eta^{-2}\left\|\boldsymbol{x}_t\right\|_2^4 \geq (1-\eta\lambda)^{4(t-t^*)}\eta^{-2}\left\|\boldsymbol{x}(t^*)\right\|_2^4 \geq (1-\eta\lambda)^{4(T_0-1)}\eta^{-2}\left\|\boldsymbol{x}(t^*)\right\|_2^4 \geq \frac{(1-\eta\lambda)^2\underline{\sigma}^2}{4\eta\lambda} \tag{48}$$

Therefore, for any $t \geq t^*$, we have

$$\begin{aligned}
&\eta^{-2}\left\|\boldsymbol{x}(t+1)\right\|_2^4 \\
=&(1-\eta\lambda)^4\eta^{-2}\left\|\boldsymbol{x}(t)\right\|_2^4 + 2(1-\lambda\eta)^2\left\|\nabla L_\gamma(\overline{\boldsymbol{x}}(t))\right\|_2^2 + \frac{\left\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\right\|_2^4}{\left\|\boldsymbol{x}(t)\right\|_2^4\eta^{-2}} \\
=&(1-\eta\lambda)^{4(t+1-t^*)}\eta^{-2}\left\|\boldsymbol{x}(t^*)\right\|_2^4 + \underbrace{2\sum_{\tau=t^*}^{t}(1-\eta\lambda)^{4(t-\tau)+2}\mathbb{E}[\left\|\nabla L_{\gamma_\tau}(\boldsymbol{x}(\tau))\right\|_2^2 \mid \boldsymbol{x}(\tau)]}_{\text{(A)}} \\
&+ \underbrace{2\sum_{\tau=t^*}^{t}(1-\eta\lambda)^{4(t-\tau)+2}\left(\left\|\nabla L_{\gamma_\tau}(\boldsymbol{x}(\tau))\right\|_2^2 - \mathbb{E}[\left\|\nabla L_{\gamma_\tau}(\boldsymbol{x}(\tau))\right\|_2^2 \mid \boldsymbol{x}(\tau)]\right)}_{\text{(B)}} \\
&+ \underbrace{\sum_{\tau=t^*}^{t}(1-\eta\lambda)^{4(t-\tau)}\frac{\left\|\nabla L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\right\|_2^4}{\left\|\boldsymbol{x}(\tau)\right\|_2^4\eta^{-2}}}_{\text{(C)}}.
\end{aligned} \tag{49}$$

Below we will upper-bound the terms (A), (B) and (C) on $\mathcal{E}_T$ respectively.

(A). By Lemma C.8, we have

$$\text{(A)} \leq 2\sum_{\tau=t^*}^{t}(1-\eta\lambda)^{4(t-\tau)+2}\overline{\sigma}^2 \leq \frac{(1-\eta\lambda)^2 e^{4\eta\lambda}}{2\eta\lambda}\overline{\sigma}^2 \leq \frac{e^{0.2}}{2\eta\lambda}\overline{\sigma}^2, \tag{50}$$

where in the last step we used $\eta\lambda \leq 0.1$ and $e^x(1-x) \leq 1$ for any $0 \leq x \leq 1$.

(B). By the definition of event $\mathcal{E}_T$, we have

$$\text{(B)} \le (1-\eta\lambda)^2 \frac{M^2}{2} e^{4\eta\lambda} \sqrt{\frac{1}{\lambda\eta} \ln \frac{2T^2}{\delta}} \le \frac{(1-\eta\lambda)^2}{6\eta\lambda} \underline{\sigma}^2 \tag{51}$$

(C). Combining the above analysis and Lemma E.3, we know conditioned on $\mathcal{E}_T$, for any $t \ge t^*$, it holds $\|\boldsymbol{x}(t)\|_2^4 / \eta^2 \ge \frac{(1-\eta\lambda)^2 \underline{\sigma}^2}{4\eta\lambda}$.
Therefore, by Lemma C.8, we have

$$\text{(C)} \le \frac{4\eta\lambda M^4}{\underline{\sigma}^2} \sum_{\tau=t^*}^{t} (1-\eta\lambda)^{4(t-\tau)-2} \le \frac{e^{4\eta\lambda} M^4}{(1-\eta\lambda)^2 \underline{\sigma}^2} \tag{52}$$

Under Condition 4.4, we can further upper bound (C) by $\frac{\underline{\sigma}^2}{9\eta\lambda e^{4\eta\lambda}(1-\eta\lambda)^2} \le \frac{\underline{\sigma}^2}{9\times\frac{8}{9}\times\frac{7}{8}\eta\lambda} = \frac{\underline{\sigma}^2}{7\eta\lambda}$, where we used the fact that $\eta\lambda \le 0.1$.

What is left to do is to upper bound $\eta^{-2}\|\boldsymbol{x}(t^*)\|_2^4$. We proceed by discussing the following three cases respectively:

- $t^* = 0$. Then $\eta^{-2}\|\boldsymbol{x}(t^*)\|_2^4 = \eta^{-2}\|\boldsymbol{x}(0)\|_2^4$.
- $1 \le t^* \le T_0 - 1$. In this case, we have

  $$\eta^{-1}\|\boldsymbol{x}_{t^*-1}\|_2^2 \ge (1-\eta\lambda)^{2(t^*-1)}\eta^{-1}\|\boldsymbol{x}(0)\|_2^2 \ge e^{-4(T_0-1)\eta\lambda}\eta^{-1}\|\boldsymbol{x}(0)\|_2^2 \ge e^{-4}\|\boldsymbol{x}(0)\|_2^2\,\eta^{-1}.$$

  Thus it holds that

$$\begin{aligned}
\eta^{-1}\|\boldsymbol{x}(t^*)\|_2^2 =& (1-\eta\lambda)^2\eta^{-1}\|\boldsymbol{x}_{t^*-1}\|_2^2 + \frac{\left\|\nabla L_{\gamma_{t^*-1}}(\overline{\boldsymbol{x}}(t^*-1))\right\|_2^2}{\|\boldsymbol{x}_{t^*-1}\|_2^2\,\eta^{-1}} \\
\le& (1-\eta\lambda)^2\sqrt{\frac{e^8(1-\eta\lambda)^2\underline{\sigma}^2}{4\eta\lambda}} + e^4\frac{M^2}{\|\boldsymbol{x}(0)\|_2^2\,\eta^{-1}} \\
\le& 2\max\{\sqrt{\frac{e^8\underline{\sigma}^2}{4\eta\lambda}}, e^4\frac{M^2}{\|\boldsymbol{x}(0)\|_2^2\,\eta^{-1}}\}
\end{aligned} \tag{53}$$

- $t^* = T_0$. Then we have $\eta^{-2}\|\boldsymbol{x}(t^*)\|_2^4 \le \frac{(1-\eta\lambda)^2\underline{\sigma}^2}{4\eta\lambda}$.

Taking maximum over three cases, we have

$$\eta^{-2}\|\boldsymbol{x}(t^*)\|_2^4 \le \max\left\{2e^4M^2e^{\left|\ln\frac{2e^4M^2}{\|\boldsymbol{x}(0)\|_2^4\eta^{-2}}\right|}, e^8\frac{\underline{\sigma}^2}{\eta\lambda}.\right\} \tag{54}$$

Plugging (54) back into (49), we got for any $t \ge t^*$

$$\begin{aligned}
&\eta^{-2}\|\boldsymbol{x}(t+1)\|_2^4 \\
=&(1-\eta\lambda)^{4\eta\lambda(t+1-t^*)}\eta^{-2}\|\boldsymbol{x}(t^*)\|_2^4 + (A) + (B) + (C) \\
\le& e^{-4\lambda\eta(t-t^*)}\max\left\{2M^2e^{\left|\ln\frac{2e^4M^2}{\|\boldsymbol{x}(0)\|_2^4\eta^{-2}}\right|}, e^4\frac{\underline{\sigma}^2}{\eta\lambda}.\right\} + \frac{\overline{\sigma}^2}{\eta\lambda},
\end{aligned} \tag{55}$$

where we used the fact that $(0.5e^{0.2} + \frac{1}{6} + \frac{1}{7} \approx 0.9202 < 1)$ in the last step.

Therefore there exists $T_1 = T_0 + \frac{1}{4\eta\lambda}\max\left\{\ln\frac{M^2\eta\lambda}{\overline{\sigma}^2} + \left|\ln\frac{2e^4M^2}{\|\boldsymbol{x}(0)\|_2^4\eta^{-2}}\right|, 4\right\}$, such that for all $t \ge T_1$, $\eta^{-2}\|\boldsymbol{x}(t)\|_2^4 \le \frac{2\overline{\sigma}^2}{\eta\lambda}$.

□

Theorem 4.5 (SGD+WD). *Let $\boldsymbol{x}(t)$ be defined by* SGD (5)*. For $\eta\lambda \le 0.1$, under Condition 4.4, with probability $1-5\delta$,*

$$\forall T_1 \le t \le T-1, \quad \frac{\underline{\sigma}^2}{2} \le \frac{2\lambda}{\eta}\left\|\boldsymbol{x}(t)\right\|_2^4 \le 4\overline{\sigma}^2, \tag{6}$$

and

$$\begin{aligned}
&\frac{1}{T-T_1}\sum_{t=T_1}^{T-1}\left\|\nabla L(\overline{\boldsymbol{x}}(t))\right\|_2^2 \\
\le&\frac{\pi^2\rho\overline{\sigma}}{(T-T_1)\sqrt{2\eta\lambda}} + 4\sqrt{\eta\lambda}\frac{\rho\overline{\sigma}^3}{\underline{\sigma}^2} \\
+&\sqrt{\frac{\ln\frac{2}{\delta}}{T-T_1}}4\frac{\pi\rho M\overline{\sigma}}{\underline{\sigma}} + \sqrt{\frac{\ln\frac{2}{\delta}}{T-T_1}}4\sqrt{\lambda\eta}\frac{M^2\rho\overline{\sigma}}{\underline{\sigma}^2},
\end{aligned} \tag{7}$$

where $T_1 = \frac{1}{4\eta\lambda}\max\left\{\ln\frac{M^2\eta\lambda}{\overline{\sigma}^2} + \left|\ln\frac{2e^4M^2}{\|\boldsymbol{x}(0)\|_2^4\eta^{-2}}\right|, 8\right\}$.

Proof. By Lemma C.1, we have

$$L(\boldsymbol{x}(t+1)) - L(\boldsymbol{x}_t) \le -\frac{\eta}{1-\eta\lambda}\frac{\langle\nabla L(\overline{\boldsymbol{x}}(t)), \nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\rangle}{\left\|\boldsymbol{x}(t)\right\|_2^2} + \frac{\rho\eta^2\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2}{2(1-\eta\lambda)^2\left\|\boldsymbol{x}(t)\right\|_2^4} \tag{56}$$

Summing up for $t = T_1$ to $T-1$, we have

$$\begin{aligned}
&\sum_{t=T_1}^{T-1}\eta\left\|\nabla L(\overline{\boldsymbol{x}}(t))\right\|_2^2\left\|\boldsymbol{x}(t)\right\|_2^{-2} = \sum_{t=T_1}^{T-1}\eta\left\|\nabla L(\boldsymbol{x}(t))\right\|_2^2 \\
\le&(1-\eta\lambda)\left(L(\boldsymbol{x}_{T_1}) - L(\boldsymbol{x}_T)\right) + \underbrace{\sum_{t=T_1}^{T-1}\frac{\rho\eta^2\mathbb{E}[\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2 \mid \boldsymbol{x}(t)]}{2(1-\eta\lambda)\left\|\boldsymbol{x}(t)\right\|_2^4}}_{\text{(A)}} \\
&+\underbrace{\sum_{t=T_1}^{T-1}\frac{\eta\langle\nabla L(\overline{\boldsymbol{x}}(t)), \nabla L(\overline{\boldsymbol{x}}(t)) - \nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\rangle}{\left\|\boldsymbol{x}(t)\right\|_2^2}}_{\text{(B)}} \\
&+\underbrace{\sum_{t=T_1}^{T-1}\frac{\rho\eta^2\left(\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2 - \mathbb{E}[\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2 \mid \boldsymbol{x}(t)]\right)}{2(1-\eta\lambda)\left\|\boldsymbol{x}(t)\right\|_2^4}}_{\text{(C)}}
\end{aligned} \tag{57}$$

Below we will give high-probability bounds for (A), (B) and (C) respectively. For convenience, we will use $A(t), B(t), C(t)$ to denote the tth term in (A), (B) and (C).

Claim E.4.1. $\mathcal{E}_T \Longrightarrow \forall T_1 \le t \le T,\ A(t) \le 2\sqrt{2}\rho\eta\lambda\frac{\overline{\sigma}^2}{\underline{\sigma}^2}$

Claim E.4.2. (B) $= \sum_{t=T_1}^{T-1} B(t)$ *is* $\mathsf{subG}((T-T_1)\frac{4\pi^2\lambda\eta\rho^2M^2}{\underline{\sigma}^2}, \mathcal{E}_T)$

Claim E.4.3. (C) $= \sum_{t=T_1}^{T-1} C(t)$ *is* $\mathsf{subG}((T-T_1)\frac{4\rho^2\lambda^2\eta^2M^4}{\underline{\sigma}^4}, \mathcal{E}_T)$

Here Claim E.4.1 follows from that $2(1-\eta\lambda) \ge \sqrt{2}$ and Lemma E.3. Note by the choice of T_1, we can upper and lower bound $\left\|\boldsymbol{x}(t)\right\|_2$ by Lemmas E.3 and E.4, that is $\frac{\underline{\sigma}^2}{4\eta\lambda} \le \eta^{-2}\left\|\boldsymbol{x}(t)\right\|_2^2 \le \frac{2\overline{\sigma}^2}{\eta\lambda}$. Thus Claims E.4.2 and E.4.3 is a direct consequence of Lemma C.7.

Thus we conclude w.p. $1-5\delta$,

$$\begin{aligned}\sqrt{\frac{\lambda\eta}{2\overline{\sigma}^2}}\frac{1}{T-T_1}\sum_{t=T_1}^{T-1}\|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2 \leq& \frac{L(\boldsymbol{x}(T_1))-\min_{\boldsymbol{x}} L(\boldsymbol{x})}{T-T_1}+2\sqrt{2}\rho\eta\lambda\frac{\overline{\sigma}^2}{\underline{\sigma}^2}\\ &+\sqrt{\frac{8\lambda\eta\ln\frac{2}{\delta}}{T-T_1}}\frac{\pi\rho M}{\underline{\sigma}}+\sqrt{\frac{8\ln\frac{2}{\delta}}{T-T_1}}\lambda\eta\frac{M^2\rho}{\underline{\sigma}^2},\end{aligned} \tag{58}$$

rearranging it and applying Lemma C.4, we get

$$\begin{aligned}\frac{1}{T-T_1}\sum_{t=T_1}^{T-1}\|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2 \leq& \frac{\pi^2\rho\overline{\sigma}}{(T-T_1)\sqrt{2\eta\lambda}}+4\sqrt{\eta\lambda}\frac{\rho\overline{\sigma}^3}{\underline{\sigma}^2}\\ &+\sqrt{\frac{\ln\frac{2}{\delta}}{T-T_1}}\frac{4\pi\rho M\overline{\sigma}}{\underline{\sigma}}+\sqrt{\frac{\ln\frac{2}{\delta}}{T-T_1}}4\sqrt{\lambda\eta}\frac{M^2\rho\overline{\sigma}}{\underline{\sigma}^2}.\end{aligned} \tag{59}$$

By Condition 4.4, we have $\frac{\sigma^2}{M^2}\geq 3\sqrt{\lambda\eta\ln\frac{2}{\delta}}$, and thus we have

$$\frac{1}{T-T_1}\sum_{t=T_1}^{T-1}\|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2 \leq \frac{\pi^2\rho\overline{\sigma}}{(T-T_1)\sqrt{2\eta\lambda}}+4\sqrt{\eta\lambda}\frac{\rho\overline{\sigma}^3}{\underline{\sigma}^2}+\frac{4}{3}\sqrt{\frac{1}{(T-T_1)\eta\lambda}}\pi\rho\underline{\sigma}+\sqrt{\frac{1}{T-T_1}}\frac{4\rho\overline{\sigma}}{3}. \tag{60}$$

□

F. Omitted Proofs for Convergence of SGD with Relative Global Clipping

Norm dynamics of clipped SGD:

$$\|\boldsymbol{x}(t+1)\|_2^2=(1-\eta\lambda)^2\|\boldsymbol{x}(t)\|_2^2+\eta^2\min\left\{\frac{\|\nabla L_\gamma(\overline{\boldsymbol{x}}(t))\|_2^2}{\|\boldsymbol{x}(t)\|_2^2},\frac{2\lambda C}{\eta}\|\boldsymbol{x}(t)\|_2^2\right\}. \tag{61}$$

Lemma F.1 (General Properties of $G_{P,C}$)**.** *For any $C>1$ and measure P supported on $\mathbb{R}^{\geq 0}$, it holds that*

1. *$G_{P,C}$ is continuous and concave;*
2. *$\sup_{\mu\geq 0}G_{P,C}(\mu)=G_{P,C}(\frac{1}{C}M_{P,\frac{1}{C}})$;*
3. *$\frac{1}{C}M_{P,\frac{1}{C}}\leq\mu_{P,C}\leq\mu_P$, where μ_P is the expectation of P.*

Proof of Lemma F.1. (1). Note $\min\{x,\cdot\}$ is a continuous and concave function for any x, we know $G_{P,C}$ is a concave function. (2). When $G_{P,C}$ is differentiable, we have $G'_{P,C}(\mu)=CF'_{P,C}(C\mu)-1$. Let $G'_{P,C}(\mu)=0$ implies that $F'_{P,C}(C\mu)=\frac{1}{C}$. Note $F'_{P,C}(C\mu)=\mathbb{P}_{t\sim P}[t>F_{P,C}]$, we know $G'_{P,C}(\frac{1}{C}M_{P,\frac{1}{C}})=0$. By concavity, $\sup_{\mu\geq 0}G_{P,C}(\mu)=G_{P,C}(\frac{1}{C}M_{P,\frac{1}{C}})$. This argument can be easily generalized to non-differentiable case by using $G_{P,C}(\mu)$ must be larger than $G_{P,C}(\mu\pm\delta)$ for infinitesimal δ. (3). First note that $F_{P,C}(M_{P,\frac{1}{C}})=\mathbb{E}_{t\sim P}[\min\{t,M_{P,\frac{1}{C}}\}]\geq M_{P,\frac{1}{C}}\cdot\mathbb{P}_{t\sim P}[t\geq M_{P,\frac{1}{C}}]=\frac{1}{C}M_{P,\frac{1}{C}}$. In other words, $G_{P,C}(\frac{1}{C}M_{P,\frac{1}{C}})\geq 0$.

Now suppose $\frac{1}{C}M_{P,\frac{1}{C}}>\mu_{P,C}$. If $G_{P,C}(\frac{1}{C}M_{P,\frac{1}{C}})=0$, then by definition, $\frac{1}{C}M_{P,\frac{1}{C}}\leq\mu_{P,C}$. If $G_{P,C}(\frac{1}{C}M_{P,\frac{1}{C}})>0$, by concavity, $G_{P,C}(\mu_{P,C})>0$, contradiction! □

Theorem F.2. *[Classifications of solutions of $F_{P,C}(C\mu)=\mu$]*

1. *If $\mathbb{P}[x=0]<1-\frac{1}{C}$, then $F_{P,C}(C\mu)=\mu$ has exact two solutions which are 0 and $\mu_{P,C}>0$;*
2. *If $\mathbb{P}[x=0]=1-\frac{1}{C}$, then $F_{P,C}(C\mu)=\mu$ for all $0\leq\mu\leq\frac{1}{C}M_{P,C}$ and $\mu_{P,C}=\frac{1}{C}M_{P,C}$;*

3. *If* $\mathbb{P}[x=0] > 1-\frac{1}{C}$*, then* $F_{P,C}(C\mu)=\mu$ *has only one solution which is* $\mu_{P,C}=0$.

Proof. Suppose there are two solutions $0<\mu_1<\mu_2$. By concavity, we have $\forall 0\le\mu\le\mu_2$, $G_{P,C}(\mu)=0$. Thus $0=G_{P,C}(0)+G_{P,C}(\mu_2)=2g(\frac{\mu_2}{2})$, which implies that

$$\mathbb{E}_{t\sim P}[\min\{t,C\mu_2\}]=2\mathbb{E}_{t\sim P}[\min\{t,\frac{C\mu_2}{2}\}]=\mathbb{E}_{t\sim P}[\min\{2t,C\mu_2\}], \tag{62}$$

that is, $\mathbb{P}_{t\sim P}[t\ge C\mu_2\vee t=0]=1$. Thus for any $0\le\mu\le\mu_2$, we have $G_{P,C}(\mu)=C\mu\mathbb{P}[x\ge C\mu_2]-\mu=0$, which implies $\mu_2=\frac{1}{C}M_{P,\frac{1}{C}}$ and $\mathbb{P}[x=0]=1-\frac{1}{C}$! □

Lemma F.3. *Under Assumption 4.7, it holds that* $G_{P,C_{\boldsymbol{x}}}(\frac{1}{C}M_{P_{\boldsymbol{x}},\frac{1}{C}})\ge\alpha_C\mu_{P_{\boldsymbol{x}},C}$ *for all* $\boldsymbol{x}\neq 0$.

Proof of Lemma F.3. By definition,

$$G_{P,C_{\boldsymbol{x}}}(\frac{1}{C}M_{P_{\boldsymbol{x}},\frac{1}{C}})=\mathbb{E}_{t\sim P_{\boldsymbol{x}}}[t\mathbb{1}[t<M_{P_{\boldsymbol{x}},C}]]+(\mathbb{P}_{t\sim P_{\boldsymbol{x}}}[t\ge M_{P_{\boldsymbol{x}},C}]-\frac{1}{C})\cdot M_{P_{\boldsymbol{x}},C}. \tag{63}$$

By the definition of the $\frac{1}{C}$-median, the second term is non-negative. The proof is completed by applying Assumption 4.7. □

Lemma F.4 (Lower and upped bounds for $G_{P_{\boldsymbol{x}},C}$)**.** *Under Assumption 4.7, it holds that*

1. $G_{P_{\boldsymbol{x}},C}(\mu)\ge\alpha_C\mu$*, for* $0\le\mu\le\frac{\mu_{P_{\boldsymbol{x}},C}}{2}$*;*
2. $G_{P_{\boldsymbol{x}},C}(\mu)\ge\alpha_C(\mu_{P_{\boldsymbol{x}},C}-\mu)$*, for* $\frac{\mu_{P_{\boldsymbol{x}},C}}{2}\le\mu\le\mu_{P_{\boldsymbol{x}},C}$*;*
3. $G_{P_{\boldsymbol{x}},C}(\mu)\le-\alpha_C(\mu-\mu_{P_{\boldsymbol{x}},C})$*, for* $\mu\ge\mu_{P_{\boldsymbol{x}},C}$.

Proof of Lemma F.5. By Lemma F.3, Assumption 4.7 implies that $G_{P,C_{\boldsymbol{x}}}(\frac{1}{C}M_{P_{\boldsymbol{x}},\frac{1}{C}})\ge\alpha_C\mu_{P_{\boldsymbol{x}},C}$ for all $\boldsymbol{x}\neq 0$. Further note that $G_{P,C_{\boldsymbol{x}}}(0)=G_{P,C_{\boldsymbol{x}}}(\mu_{P_{\boldsymbol{x}}},C)=0$. The claims (a), (b) and (c) are immediate by concavity of $G_{P,C_{\boldsymbol{x}}}$. □

The above inequalities also directly imply the following version using $\underline{\mu}_C$ and $\overline{\mu}_C$ as thresholds.

Lemma F.5 (Uniform Lower and upped bounds for $G_{P_{\boldsymbol{x}},C}$)**.** *Under Assumption 4.7, it holds that for* $\|\boldsymbol{x}\|_2=1$*,*

1. $G_{P_{\boldsymbol{x}},C}(\mu)\ge\alpha_C\mu$*, for* $0\le\mu\le\frac{\underline{\mu}_C}{2}$*;*
2. $G_{P_{\boldsymbol{x}},C}(\mu)\ge\alpha_C(\underline{\mu}_C-\mu)$*, for* $\frac{\underline{\mu}_C}{2}\le\mu\le\underline{\mu}_C$*;*
3. $G_{P_{\boldsymbol{x}},C}(\mu)\le-\alpha_C(\mu-\overline{\mu}_C)$*, for* $\mu\ge\overline{\mu}_C$.
4. $G_{P_{\boldsymbol{x}},C}(\mu)\ge\frac{\alpha_C\mu}{4}$*, for* $0\le\mu\le\frac{4\underline{\mu}_C}{5}$*; (4. follows from Property 1. and 2.)*

For convenience, we define $R_t:=\frac{2\lambda}{\eta}\|\boldsymbol{x}(t)\|_2^2$, $g_t:=\|\nabla L_{\gamma_t}(\boldsymbol{x}(t))\|_2^2$, $\widehat{g}_t:=\min\{CR_t,g_t\}$, $\widetilde{g}_t:=R_t\widehat{g}_t=\min\{CR_t^2,\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2\}$ and $\overline{g}_t:=\frac{\widehat{g}_t}{R_t}=\min\{C,\frac{\|\nabla L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2}{R_t}\}$. Thus we have $\mathbb{E}[\widehat{g}_t\mid\boldsymbol{x}(t)]=\mu_{P_{\boldsymbol{x}(t)},C}$. We further define $\beta_l:=1-2\lambda^2\eta^2+\eta^4\lambda^4-4\eta\lambda\alpha_C(1-\eta\lambda)^2=1-4\eta\lambda\alpha_C+O(\eta^2\lambda^2)$ and $\beta_u:=1-2\lambda^2\eta^2+\eta^4\lambda^4-4\eta\lambda\alpha_C(1-\eta\lambda)^2+4C^2\eta^2\lambda^2=1-4\eta\lambda\alpha_C+O(\eta^2\lambda^2)$.

Given an integer $T\ge 0$, let $\mathcal{E}_T^1$ be the event that $\forall 0\le t'\le t\le T$,

$$\left|\sum_{s=t'}^{t}{\beta_l}^{t-s}\left(\widetilde{g}_s-\mathbb{E}[\widetilde{g}_s\mid\boldsymbol{x}(s)]\right)\mathbb{1}\left[R_s^2\le\underline{\mu}_C\right]\right|\le\sqrt{C}\underline{\mu}_C\sqrt{\frac{1}{1-{\beta_l}^2}\ln\frac{2T^2}{\delta}}.$$

Let $\mathcal{E}_T^2$ be the event that $\forall 0\le t'\le t\le T$,

$$\left|\sum_{s=t'}^{t}{\beta_l}^{t-s}\left(\widetilde{g}_s-\mathbb{E}[\widetilde{g}_s\mid\boldsymbol{x}(s)]\right)\mathbb{1}\left[R_s^2\le 2\overline{\mu}_C\right]\right|\le 2\sqrt{C}\overline{\mu}_C\sqrt{\frac{1}{1-{\beta_l}^2}\ln\frac{2T^2}{\delta}}.$$

Let $\mathcal{E}_T^3$ be the event that $\forall 0 \le t' \le t \le T$,

$$\left|\sum_{s=t'}^{t} \overline{g}_s - \mathbb{E}[\overline{g}_s \mid \boldsymbol{x}(s)]\right| \le C\sqrt{T \ln \frac{2T^2}{\delta}}.$$

Lemma F.6. $\mathbb{P}[\mathcal{E}_T^i] \ge 1-\delta$, *for* $i = 1, 2, 3$.

Proof of Lemma F.6. Note the sequence in $\mathcal{E}_T^i$ are martingales whose differences are uniformly bounded ($\underline{\mu}_C, \overline{\mu}_C$ and C). The lemma follows directly from Hoeffding Inequality and Azuma Inequality. □

Theorem F.7 (Norm lower bound with clipping: Warm Start)**.** *Suppose Assumption 4.7 holds, with probability at least* $1-\delta$ *(or whenever* $\mathcal{E}_T^1$ *holds), if* $R_t^2 \ge \frac{3}{4}\underline{\mu}_C$*, then for any* $t' \ge t$*, we have*

$$R_{t'}^2 \ge \left(1 - \frac{{\beta_l}^{t'-t}}{4} - O(\sqrt{\eta\lambda}) - \sqrt{\frac{2C}{\alpha_C}\eta\lambda \ln\frac{T^2}{\delta}}(1+O(\eta\lambda))\right)\underline{\mu}_C \tag{64}$$

Proof. We first claim for any $t \le t' \le T$, conditioned on $\mathcal{E}_T^1$, it holds that $R_{t'}^2 \ge \frac{\underline{\mu}_C}{2}$. Below we prove by contradiction. If not, let t' be the smallest step such that $R_{t'}^2 < \frac{\underline{\mu}_C}{2}$. We let t^* be the largest step between t and t' such that $R_{t^*}^2 \ge \underline{\mu}_C$ ($t^* = t-1$ is no such t^* exists) Thus if $t^* \ge t$ then $R_{t^*+1}^2$ is at least $(1-\eta\lambda)^4 R_t^2 = (1-O(\eta\lambda))\underline{\mu}_C$. Otherwise $t^* = t$ and it implies that $R_{t^*+1}^2 = R_t^2 = (\frac{3}{4} - O(\sqrt{\eta\lambda}))\underline{\mu}_C$. By the definition, we know for any $t^*+1 \le s \le t'$, $R_s^2 \le \underline{\mu}_C$.

Similar to Equation (43), we have

$$\begin{aligned}
R_{s+1}^2 =& R_s^2(1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2\widetilde{g}_s + 4\eta^2\lambda^2\widetilde{g}_t^2 && (65)\\
\ge& R_s^2((1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2 + 4C^2\eta^2\lambda^2) && (66)\\
&+4\eta\lambda(1-\eta\lambda)^2(\mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2) + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) && (67)
\end{aligned}$$

Thus for any s such that $\overline{\mu}_C \le R_s^2 \le 2\overline{\mu}_C$, by Lemma F.5, it holds that

$$G_{P_{\overline{\boldsymbol{x}}(s)},C}(R_s^2) = \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2 \le \alpha_C(\underline{\mu}_C - R_s^2). \tag{68}$$

Thus,

$$\begin{aligned}
R_{s+1}^2 \ge& R_s^2(1 - 2\eta^2\lambda^2 + \eta^4\lambda^4) && (69)\\
&+4\eta\lambda\alpha_C(1-\eta\lambda)^2(\underline{\mu}_C - R_s^2) + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) && (70)\\
=& \beta_l R_s^2 + 4\eta\lambda\alpha_C(1-\eta\lambda)^2\underline{\mu}_C + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]). && (71)
\end{aligned}$$

That is,

$$\begin{aligned}
& R_{s+1}^2 - \frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\underline{\mu}_C}{1-\beta_l} && (72)\\
\ge& \beta_l(R_s^2 - \frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\underline{\mu}_C}{1-\beta_l}) && (73)\\
&+4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) && (74)
\end{aligned}$$

Applying the above inequality for $s = t^* + 1, \ldots, t' - 1$, we have

$$R_{t'}^2 \geq \underbrace{{\beta_l}^{t'-t^*-1}\left(R_{t^*+1}^2 - \frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\underline{\mu}_C}{1-\beta_l}\right)}_{(A)} \tag{75}$$

$$+ \underbrace{\frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\underline{\mu}_C}{1-\beta_l}}_{(B)} \tag{76}$$

$$+ \underbrace{4\eta\lambda(1-\eta\lambda)^2\sum_{s=t^*+1}^{t'}{\beta_l}^{t-s}\left(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]\right)\mathbb{1}\left[R_s^2 \leq \underline{\mu}_C\right]}_{(C)}. \tag{77}$$

For term (B), we have $1-\beta_u = 4\eta\lambda\alpha_C(1-\eta\lambda)^2(1+O(\eta\lambda))$ and thus $(B) = \underline{\mu}_C(1+O(\eta\lambda))$. Since $R_{t^*+1} \geq \frac{3}{4}\underline{\mu}_C$, it holds that $(A) \geq -{\beta_l}^{t'-t^*-1}(\frac{1}{4}+O(\sqrt{\lambda\eta}))\underline{\mu}_C \geq -(\frac{1}{4}+O(\sqrt{\lambda\eta}))\underline{\mu}_C$. Since $\mathcal{E}_T^1$ holds, we have

$$|(C)| \leq 4\eta\lambda(1-\eta\lambda)^2\cdot\sqrt{C}\underline{\mu}_C\sqrt{\frac{1}{1-{\beta_l}^2}\ln\frac{2T^2}{\delta}} = \underline{\mu}_C\sqrt{\frac{2C}{\alpha_C}\eta\lambda\ln\frac{T^2}{\delta}}(1+O(\eta\lambda)) \tag{78}$$

Thus there's some constant ι, such for $\eta\lambda \leq \min\{\iota, \frac{\alpha_C}{64C\ln T^2/\delta}\}$, $(A)+(B)+(C) \geq (\frac{6-\sqrt{2}}{8} - O(\sqrt{\eta\lambda}))\underline{\mu}_C \geq \frac{\underline{\mu}_C}{2}$. This leads to a contradiction to the definition of t'. Thus for any $t \leq t' \leq T$, conditioned on $\mathcal{E}_T^1$, it holds that $R_{t'}^2 \geq \frac{\underline{\mu}_C}{2}$. Furthermore, if $t^* \neq t$, then $R_{t^*+1} \geq (1-O(\sqrt{\eta\lambda}))\underline{\mu}_C$. Thus $(A) \geq -O(\sqrt{\eta\lambda})\underline{\mu}_C$. Otherwise if $t^* = t$, then $(A) \geq -{\beta_l}^{t'-t}(\frac{1}{4}+O(\sqrt{\lambda\eta}))\underline{\mu}_C$. Combine the bounds in these two cases, we conclude that

$$R_{t'}^2 \geq \left(1 - \frac{{\beta_l}^{t'-t}}{4} - O(\sqrt{\eta\lambda}) - \sqrt{\frac{2C}{\alpha_C}\eta\lambda\ln\frac{T^2}{\delta}}(1+O(\eta\lambda))\right)\underline{\mu}_C \tag{79}$$

□

Theorem F.8 (Norm upper bound with clipping: Warm Start)**.** *Suppose Assumption 4.7 holds, with probability at least $1-\delta$ (or whenever $\mathcal{E}_T^2$ holds), if $R_t^2 \leq \frac{3}{2}\overline{\mu}_C$, then for any $t' \geq t$, we have*

$$R_{t'}^2 \leq \left(1 + \frac{{\beta_l}^{t'-t}}{2} + O(\sqrt{\eta\lambda}) + \sqrt{\frac{2C}{\alpha_C}\eta\lambda\ln\frac{T^2}{\delta}}(1+O(\eta\lambda))\right)\overline{\mu}_C \tag{80}$$

Proof. We first claim for any $t \leq t' \leq T$, conditioned on $\mathcal{E}_T^2$, it holds that $R_{t'}^2 \leq 2\overline{\mu}_C$. Below we prove by contradiction. If not, let t' be the largest step such that $R_{t'}^2 > 2\overline{\mu}_C$. We let t^* be the largest step between t and t' such that $R_{t^*}^2 \leq \overline{\mu}_C$ ($t^* = t-1$ is no such t^* exists) Thus if $t^* \geq t$ then $R_{t^*+1}^2$ is at most $(1+2C\eta\lambda)^2R_t^2 = (1+2C\eta\lambda)^2\overline{\mu}_C$. Otherwise $t^* = t$ and it implies that $R_{t^*+1}^2 = R_t^2 \leq \frac{3}{2}\overline{\mu}_C$. By the definition, we know for any $t^*+1 \leq s \leq t'$, $R_s^2 \geq \overline{\mu}_C$.

Similar to Equation (43), we have

$$R_{s+1}^2 \leq R_s^2(1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2\widetilde{g}_s + 4\eta^2\lambda^2\widehat{g}_s^2 \tag{81}$$

$$\leq R_s^2((1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2 + 4\eta^2\lambda^2C^2) \tag{82}$$

$$+4\eta\lambda(1-\eta\lambda)^2(\mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2) + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) \tag{83}$$

Thus for any s such that $\overline{\mu}_C \leq R_s^2$, by Lemma F.5, it holds that

$$G_{P_{\overline{\boldsymbol{x}}(s)},C}(R_s^2) = \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2 \geq \alpha_C(\overline{\mu}_C - R_s^2). \tag{84}$$

Thus,

$$R_{s+1}^2 \leq R_s^2(1-2\eta^2\lambda^2+\eta^4\lambda^4+4\eta^2\lambda^2C^2) \tag{85}$$
$$+4\eta\lambda\alpha_C(1-\eta\lambda)^2(\overline{\mu}_C-R_s^2)+4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s-\mathbb{E}[\widetilde{g}_s\mid \boldsymbol{x}(s)]) \tag{86}$$
$$=\beta_u R_s^2+4\eta\lambda\alpha_C(1-\eta\lambda)^2\overline{\mu}_C+4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s-\mathbb{E}[\widetilde{g}_s\mid \boldsymbol{x}(s)]). \tag{87}$$

That is,

$$R_{s+1}^2-\frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\overline{\mu}_C}{1-\beta_u} \tag{88}$$
$$\leq\beta_u(R_s^2-\frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\overline{\mu}_C}{1-\beta_u}) \tag{89}$$
$$+4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s-\mathbb{E}[\widetilde{g}_s\mid \boldsymbol{x}(s)]) \tag{90}$$

Applying the above inequality for $s=t^*+1,\dots,t'-1$, we have

$$R_{t'}^2 \leq \underbrace{{\beta_u}^{t'-t^*-1}\left(R_{t^*+1}^2-\frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\overline{\mu}_C}{1-\beta_u}\right)}_{(A)} \tag{91}$$
$$+\underbrace{\frac{4\eta\lambda\alpha_C(1-\eta\lambda)^2\overline{\mu}_C}{1-\beta_u}}_{(B)} \tag{92}$$
$$+\underbrace{4\eta\lambda(1-\eta\lambda)^2\sum_{s=t^*+1}^{t'}{\beta_u}^{t-s}\left(\widetilde{g}_s-\mathbb{E}[\widetilde{g}_s\mid \boldsymbol{x}(s)]\right)\mathbb{1}\left[R_s^2\leq 2\overline{\mu}_C\right]}_{(C)}. \tag{93}$$

For term (B), we have $1-\beta_u=4\eta\lambda\alpha_C(1-\eta\lambda)^2(1+O(\eta\lambda))$ and thus $(B)=\overline{\mu}_C(1+O(\eta\lambda))$. Since $R_{t^*+1}\leq\frac{3}{2}\overline{\mu}_C$, it holds that $(A)\leq {\beta_u}^{t'-t^*-1}(\frac{1}{2}+O(\sqrt{\lambda\eta}))\overline{\mu}_C\leq(\frac{1}{2}+O(\sqrt{\lambda\eta}))\overline{\mu}_C$. Since $\mathcal{E}_T^2$ holds, we have

$$|(C)|\leq 8\eta\lambda(1-\eta\lambda)^2\cdot\sqrt{C}\overline{\mu}_C\sqrt{\frac{1}{1-{\beta_u}^2}\ln\frac{2T^2}{\delta}}=2\overline{\mu}_C\sqrt{\frac{2C}{\alpha_C}\eta\lambda\ln\frac{T^2}{\delta}}(1+O(\eta\lambda)) \tag{94}$$

Thus there's some constant ι, such for $\eta\lambda\leq\min\{\iota,\frac{\alpha_C}{64C\ln T^2/\delta}\}$, $(A)+(B)+(C)\leq(\frac{6+\sqrt{2}}{4}+O(\sqrt{\eta\lambda}))\overline{\mu}_C\leq 2\overline{\mu}_C$. This leads to a contradiction to the definition of t'. Thus for any $t\leq t'\leq T$, conditioned on $\mathcal{E}_T^1$, it holds that $R_{t'}^2\geq 2\overline{\mu}_C$. Furthermore, if $t^*\neq t$, then $R_{t^*+1}\leq(1+O(\sqrt{\eta\lambda}))\overline{\mu}_C$. Thus $(A)\leq O(\sqrt{\eta\lambda})\overline{\mu}_C$. Otherwise if $t^*=t$, then $(A)\leq {\beta_u}^{t'-t}(\frac{1}{2}+O(\sqrt{\lambda\eta}))\overline{\mu}_C$. Combine the bounds in these two cases, we conclude that

$$R_{t'}^2\leq\left(1+\frac{{\beta_l}^{t'-t}}{2}+O(\sqrt{\eta\lambda})+\sqrt{\frac{2C}{\alpha_C}\eta\lambda\ln\frac{T^2}{\delta}}(1+O(\eta\lambda))\right)\overline{\mu}_C \tag{95}$$

□

Theorem F.9 (Norm Convergence of clipped SGD)**.** *Suppose Assumption 4.7 holds, for $\eta\lambda=O(\min\{1,\frac{\alpha_C}{C\ln T/\delta^2}\})$, with probability $1-3\delta$ (when $\mathcal{E}_T^1$,$\mathcal{E}_T^2$ and $\mathcal{E}_T^3$ happens), there is a $T'=\frac{\max\left\{\ln\frac{R_0^2}{\overline{\mu}_C},\ln\frac{\mu_C}{R_0^2}\right\}+O(1)}{\alpha_C\eta\lambda}$, such that for all $T'\leq t\leq T$, we have*

$$\frac{\mu_C}{2}\leq R_t^2\leq 2\overline{\mu}_C. \tag{96}$$

More concretely, we have

$$R_t^2 \in [(1-\beta_l^{t-T'})\underline{\mu}_C - \widetilde{O}(\sqrt{\lambda\eta}),\quad \overline{\mu}_C(1+\beta_u^{t-T'}) + \widetilde{O}(\sqrt{\lambda\eta})]. \tag{97}$$

Proof of Theorem F.9. We will prove the desired inequality always holds when $\mathcal{E}_T^i$ holds, for $i=1,2,3$. We have already proved the result for the case where $\frac{3}{4}\underline{\mu}_C \le R_t^2 \le \frac{3}{2}\overline{\mu}_C$ in Theorems F.7 and F.8. Now we turn to the case where $R_0^2 \ge \frac{3}{2}\overline{\mu}_C$ and $R_0^2 \le \frac{1}{2}\underline{\mu}_C$. Our goal is to prove with high probability, that $R_t^2 \in [\frac{3}{4}\underline{\mu}_C, \frac{3}{2}\overline{\mu}_C]$ for at least some $t < T'$.

Below we first show $\exists 0 < t < T'$, $R_t^2 \le \frac{3}{2}\overline{\mu}_C$. Otherwise, similar to Equation (81),

$$\begin{aligned}
R_{s+1}^2 \le& R_s^2(1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2\widetilde{g}_s + 4\eta^2\lambda^2\widetilde{g}_s^2 && (98)\\
\le& R_s^2((1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2 + 4\eta^2\lambda^2C^2) && (99)\\
&+4\eta\lambda(1-\eta\lambda)^2(\mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2) + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) && (100)
\end{aligned}$$

Thus for any s such that $\frac{3}{2}\overline{\mu}_C \le R_s^2$, by Lemma F.5, it holds that

$$G_{P_{\overline{\boldsymbol{x}}(s)},C}(R_s^2) = \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2 \ge \alpha_C(\overline{\mu}_C - R_s^2) \ge -\frac{\alpha_C}{3}R_s^2. \tag{101}$$

Thus,

$$\begin{aligned}
R_{s+1}^2 \le& R_s^2(1-2\eta^2\lambda^2+\eta^4\lambda^4+4\eta^2\lambda^2C^2) && (102)\\
&-\frac{4}{3}\eta\lambda\alpha_C(1-\eta\lambda)^2R_s^2 + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) && (103)\\
=& R_s^2\left(1-2\eta^2\lambda^2+\eta^4\lambda^4+4\eta^2\lambda^2C^2 - \frac{4}{3}\eta\lambda\alpha_C(1-\eta\lambda)^2 + 4\eta\lambda(1-\eta\lambda)^2(\overline{g}_s - \mathbb{E}[\overline{g}_s \mid \boldsymbol{x}(s)])\right) && (104)
\end{aligned}$$

Note that $\overline{g}_s \le C$, we have

$$\ln R_{s+1}^2 - \ln R_s^2 \le -\frac{4}{3}\eta\lambda\alpha_C + \eta\lambda(\overline{g}_s - \mathbb{E}[\overline{g}_s \mid \boldsymbol{x}(s)]) + O(\eta^2\lambda^2) \tag{105}$$

Since we assume $\forall 0 \le t \le T'$, $R_t^2 \ge \frac{3}{2}\overline{\mu}_C$, conditioned on $\mathcal{E}_T^3$, we have

$$\ln\frac{3}{4} + \ln\overline{\mu}_C - \ln R_0^2 \le \ln R_{T'}^2 - \ln R_0^2 \le -\frac{4T}{3}\eta\lambda\alpha_C + C\eta\lambda\sqrt{T\ln\frac{2T^2}{\delta}} + O(\eta^2\lambda^2T), \tag{106}$$

which is in contradiction with the definition of $T' = \frac{\max\left\{\ln\frac{R_0^2}{\overline{\mu}_C}, \ln\frac{\underline{\mu}_C}{R_0^2}\right\}+O(1)}{\alpha_C\eta\lambda}$.

Now we show $\exists 0 < t < T'$, $R_t^2 \ge \frac{3}{4}\underline{\mu}_C$. Otherwise, similar to Equation (81),

$$\begin{aligned}
R_{s+1}^2 =& R_s^2(1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2\widetilde{g}_s + 4\eta^2\lambda^2\widetilde{g}_t^2 && (107)\\
\ge& R_s^2((1-\eta\lambda)^4 + 4\eta\lambda(1-\eta\lambda)^2 + 4C^2\eta^2\lambda^2) && (108)\\
&+4\eta\lambda(1-\eta\lambda)^2(\mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2) + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) && (109)
\end{aligned}$$

Thus for any s such that $R_s^2 \le \frac{4}{5}\underline{\mu}_C$, by Lemma F.5, it holds that

$$G_{P_{\overline{\boldsymbol{x}}(s)},C}(R_s^2) = \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)] - R_s^2 \ge \frac{\alpha_C}{4}R_s^2. \tag{110}$$

Thus,

$$\begin{aligned}
R_{s+1}^2 \ge& R_s^2(1-2\eta^2\lambda^2+\eta^4\lambda^4) && (111)\\
&+\eta\lambda\alpha_C(1-\eta\lambda)^2R_s^2 + 4\eta\lambda(1-\eta\lambda)^2(\widetilde{g}_s - \mathbb{E}[\widetilde{g}_s \mid \boldsymbol{x}(s)]) && (112)\\
=& R_s^2\left(1-2\eta^2\lambda^2+\eta^4\lambda^4 + \eta\lambda\alpha_C(1-\eta\lambda)^2 + 4\eta\lambda(1-\eta\lambda)^2(\overline{g}_s - \mathbb{E}[\overline{g}_s \mid \boldsymbol{x}(s)])\right) && (113)
\end{aligned}$$

Note that $\overline{g}_s \le C$, we have

$$\ln R_{s+1}^2 - \ln R_s^2 \ge \eta\lambda\alpha_C + \eta\lambda(\overline{g}_s - \mathbb{E}[\overline{g}_s \mid \boldsymbol{x}(s)]) + O(\eta^2\lambda^2) \tag{114}$$

Since we assume $\forall 0 \le t \le T'$, $R_t^2 \ge \frac{3}{2}\overline{\mu}_C$, conditioned on $\mathcal{E}_T^3$, we have

$$\ln \overline{\mu}_C - \ln R_0^2 \ge \ln R_{T'}^2 - \ln R_0^2 \ge T\eta\lambda\alpha_C - C\eta\lambda\sqrt{T\ln\frac{2T^2}{\delta}} + O(\eta^2\lambda^2 T), \tag{115}$$

which is in contradiction with the definition of $T' = \frac{\max\left\{\ln\frac{R_0^2}{\overline{\mu}_C}, \ln\frac{\underline{\mu}_C}{R_0^2}\right\}+O(1)}{\alpha_C\eta\lambda}$.

□

Proof of Theorem 4.8. The proof of Algorithm 1 is almost identical to that of Theorem 4.5, except replacing M by $2\overline{\mu}_C$, $\overline{\sigma}$ by $\overline{\mu}_C$, $\underline{\sigma}$ by $\underline{\mu}_C$ since the clipped stochastic gradient has smaller maximum norm, maximum covariance and smaller covariance. □

G. Convergence of SGD for multi-group scale invariant functions

In this section we extend our results to the multi-group scale invariant setting, which is quite common in practice, *e.g.* a feedforward network with normalization after each layer. By Definition G.1, multi-group scale invariant function is also scale invariant. However, it violates the assumption that the smoothness and the expectation of stochastic gradient norm square is lower bounded on unit sphere (indeed the loss function is not defined at everywhere on unit sphere), and thus needs to be treated separately. A simple example would be $L(\boldsymbol{x}, \boldsymbol{y}) = L(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}, \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2})$, the loss L is undefined at any point where $\|x\|_2 = 1$ and $\boldsymbol{y} = \boldsymbol{0}$. Yet our analysis for single scale invariant parameter group can still extend to this case, with a similar assumption that the expected gradient norm square is lower bounded.

Let $d_1, \ldots, d_K$ be positive integers with $d = \sum_{k=1}^K d_k$. For $\boldsymbol{x} \in \mathbb{R}^d = \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_K}$, we use s_k to denote $\sum_{i\le k} d_i$ and $\boldsymbol{x}_k$ to denote the vector $[x_{s_{k-1}}, \ldots, x_{s_k-1}]^\top$. For convenience, we define $\nabla_k f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_k}$ for any $1 \le k \le K$.

Definition G.1. Given $d_1, \ldots, d_K$ and a cone $U \subset \mathbb{R}^d$, we say a function $f : U \to \mathbb{R}$ is *multi-group scale invariant* iff $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) = f(c_1\boldsymbol{x}_1, \ldots, c_K\boldsymbol{x}_K)$ for any $\boldsymbol{x} \in U$ and $c_k > 0$ for $1 \le k \le K$.

Setting: Similarly, we assume there exists constants $\underline{\sigma}_k$ and $\overline{\sigma}_k$, such that $\underline{\sigma}_k^2 \le \mathbb{E}\|\nabla_k L_\gamma(\boldsymbol{x})\|_2^2 \le \overline{\sigma}_k^2$, for any $\boldsymbol{x}$ such that $\|\boldsymbol{x}_k\|_2 = 1$. In this subsection, we define $\rho := \max_{\|\boldsymbol{x}_k\|_2=1,\forall k} \lambda_{\max}(\nabla^2 L(\boldsymbol{x}))$.

Condition G.2. $\frac{\underline{\sigma}_k^2}{M_k^2} \ge 3e^{4\eta\lambda}\sqrt{\lambda\eta\max\{\ln\frac{2T^2}{\delta}, 1\}}$.

Theorem G.3 (SGD+WD, Multi-group Scale Invariance)**.** *With probability* $1-(K+2)\delta$*,*

$$\begin{aligned}
&\frac{\sqrt{\lambda\eta/2}}{\sum_{k=1}^K \underline{\sigma}_k}\frac{1}{T-T_1}\sum_{t=T_1}^{T-1}\|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2 \\
\le&\frac{\pi^2\rho}{T-T_1} + 2\sqrt{2}\rho\eta\lambda\sum_{k=1}^K\frac{\overline{\sigma}_k^2}{\underline{\sigma}_k^2} \\
&+\sqrt{\frac{8\lambda\eta\ln\frac{2}{\delta}}{T-T_1}}\pi\rho\sum_{k=1}^K\frac{M_k}{\underline{\sigma}_k} + \sqrt{\frac{8\ln\frac{2}{\delta}}{T-T_1}}\lambda\eta\rho\sum_{k=1}^K\frac{M_k^2}{\underline{\sigma}_k^2},
\end{aligned} \tag{116}$$

where $T_1 = \frac{1}{4\eta\lambda}\max_k\left\{\ln\frac{M_k^2\eta\lambda}{\overline{\sigma}_k^2} + \left|\ln\frac{2e^4M_k^2}{\|\boldsymbol{x}_k(0)\|_2^4\eta^{-2}}\right|, 8\right\}$.

Following the same strategy, we can prove the multi-group counterpart of norm convergence result, Lemma E.2. Given a integer $T \ge 0$, let $\mathcal{E}_{T,k}$ be the event that $\forall 0 \le t' \le t \le T-1$,

$$\left|\sum_{\tau=t'}^t(1-\eta\lambda)^{4(t-\tau)}\left(\|\nabla_k L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 - \mathbb{E}[\|\nabla_k L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 \mid \overline{\boldsymbol{x}}(\tau)]\right)\right| \le e^{4\eta\lambda}\cdot\frac{M_k^2}{4}\sqrt{\frac{1}{\lambda\eta}\ln\frac{2T^2}{\delta}}. \tag{117}$$

Lemma G.4. *For any* $0 \le t' \le t \le T-1$, $1 \le k \le K$

$$\sum_{\tau=t'}^{t}(1-\eta\lambda)^{4(t-\tau)}\left(\|\nabla_k L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 - \mathbb{E}[\|\nabla_k L_{\gamma_\tau}(\overline{\boldsymbol{x}}(\tau))\|_2^2 \mid \boldsymbol{x}(\tau)]\right) \sim \mathsf{subG}(\frac{e^{8\eta\lambda}M_k^4}{32}) \tag{118}$$

Thus we have $\mathbb{P}[\mathcal{E}_{T,k}] \ge 1-\delta$ *by Lemma C.6.*

The following theorem is a restatement of Lemmas E.3 and E.4 in the context of multi-group scale invariance.

Lemma G.5. *Under Condition G.2, there exists* $T_1 = \frac{1}{4\eta\lambda}\max_k\left\{\ln\frac{M_k^2\eta\lambda}{\overline{\sigma}_k^2} + \left|\ln\frac{2e^4M_k^2}{\|\boldsymbol{x}_k(0)\|_2^4\eta^{-2}}\right|, 8\right\}$, *such that* $\forall t \ge T_1$, $\frac{\sigma_k^2}{4\eta\lambda} \le \eta^{-2}\|\boldsymbol{x}(t)\|_2^4 \le \frac{2\overline{\sigma}_k^2}{\eta\lambda}$, *conditioned on* $\cup_{k=1}^K\mathcal{E}_{T,k}$.

The proof of Theorem G.3 is a natural generalization of Theorem 4.5.

Proof of Theorem G.3. Setting $\boldsymbol{x} = (1-\eta\lambda)\boldsymbol{x}(t)$ in Lemma C.2, we have

$$L(\boldsymbol{x}(t+1)) - L(\boldsymbol{x}_t) \le -\frac{\eta}{1-\eta\lambda}\langle\nabla L(\boldsymbol{x}(t)), \nabla L_{\gamma_t}(\boldsymbol{x}(t))\rangle + \sum_{k=1}^K\frac{\rho\eta^2\|\nabla_k L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2}{2(1-\eta\lambda)^2\|\boldsymbol{x}_k(t)\|_2^4} \tag{119}$$

For convenience we define $\widehat{\boldsymbol{x}} = [\frac{\boldsymbol{x}_1^\top}{\|\boldsymbol{x}_1\|_2}, \dots, \frac{\boldsymbol{x}_K^\top}{\|\boldsymbol{x}_K\|_2}]^\top$. Summing up for $t = T_1$ to $T-1$, we have

$$\begin{aligned}
&\sum_{t=T_1}^{T-1}\eta\|\nabla L(\overline{\boldsymbol{x}}(t))\|_2^2\|\boldsymbol{x}(t)\|_2^{-2} = \sum_{t=T_1}^{T-1}\eta\|\nabla L(\boldsymbol{x}(t))\|_2^2 \\
\le&(1-\eta\lambda)\left(L(\boldsymbol{x}_{T_1}) - L(\boldsymbol{x}_T)\right) + \underbrace{\sum_{t=T_1}^{T-1}\sum_{k=1}^K\frac{\rho\eta^2\mathbb{E}[\|\nabla_k L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2 \mid \boldsymbol{x}(t)]}{2(1-\eta\lambda)\|\boldsymbol{x}_k(t)\|_2^4}}_{\text{(A)}} \\
&+ \underbrace{\sum_{t=T_1}^{T-1}\sum_{k=1}^K\frac{\eta\langle\nabla_k L(\widehat{\boldsymbol{x}}(t)), \nabla_k L(\widehat{\boldsymbol{x}}(t)) - \nabla_k L_{\gamma_t}(\widehat{\boldsymbol{x}}(t))\rangle}{\|\boldsymbol{x}_k(t)\|_2^2}}_{\text{(B)}} \\
&+ \underbrace{\sum_{t=T_1}^{T-1}\sum_{k=1}^K\frac{\rho\eta^2\left(\|\nabla_k L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2 - \mathbb{E}[\|\nabla_k L_{\gamma_t}(\overline{\boldsymbol{x}}(t))\|_2^2 \mid \boldsymbol{x}(t)]\right)}{2(1-\eta\lambda)\|\boldsymbol{x}_k(t)\|_2^4}}_{\text{(C)}}
\end{aligned} \tag{120}$$

Below we will give high-probability bounds for (A), (B) and (C) respectively. For convenience, we will use $A(t), B(t), C(t)$ to denote the tth term in (A), (B) and (C).

Claim G.5.1. $\cup_{k=1}^K\mathcal{E}_{T,k} \Longrightarrow \forall T_1 \le t \le T,\ A(t) \le 2\sqrt{2}\rho\eta\lambda\sum_{k=1}^K\frac{\overline{\sigma}_k^2}{\sigma_k^2}$

Claim G.5.2. $(\mathrm{B}) = \sum_{t=T_1}^{T-1}B(t)$ *is* $\mathsf{subG}(4\pi^2\lambda\eta\rho^2(T-T_1)\left(\sum_{k=1}^K\frac{M_k}{\sigma_k}\right)^2, \cup_{k=1}^K\mathcal{E}_{T,k})$

Claim G.5.3. $(\mathrm{C}) = \sum_{t=T_1}^{T-1}C(t)$ *is* $\mathsf{subG}(4\rho^2\lambda^2\eta^2(T-T_1)\left(\sum_{k=1}^K\frac{M_k^2}{\sigma_k^2}\right)^2, \cup_{k=1}^K\mathcal{E}_{T,k})$

Here Claim G.5.1 follows from that $2(1-\eta\lambda) \ge \sqrt{2}$ and Lemma E.3. Note by the choice of T_1, we can upper and lower bound $\|\boldsymbol{x}(t)\|_2$ by Lemma G.5, that is $\frac{\sigma_k^2}{4\eta\lambda} \le \eta^{-2}\|\boldsymbol{x}_k(t)\|_2^2 \le \frac{2\overline{\sigma}_k^2}{\eta\lambda}$. Thus Claims G.5.2 and G.5.3 is a direct consequence of Lemma C.7.

Thus by Chernoff bound (Lemma C.6), with probability at least $1-(K+2)\delta$, Equation (116) holds. □