Convergence Rates of Non-Convex Stochastic Gradient Descent Under a Generic Łojasiewicz Condition and Local Smoothness

Kevin Scaman, Cédric Malherbe, Ludovic dos Santos

Convergence Rates of Non-Convex Stochastic Gradient Descent Under a Generic Łojasiewicz Condition and Local Smoothness

Kevin Scaman [1 2] **Cédric Malherbe** [3] **Ludovic Dos Santos** [3]

Abstract

Training over-parameterized neural networks involves the empirical minimization of highly non-convex objective functions. Recently, a large body of works provided theoretical evidence that, despite this non-convexity, properly initialized over-parameterized networks can converge to a zero training loss through the introduction of the *Polyak-Łojasiewicz* condition. However, these analyses are restricted to quadratic losses such as *mean square error*, and tend to indicate fast exponential convergence rates that are seldom observed in practice. In this work, we propose to extend these results by analyzing stochastic gradient descent under more generic Łojasiewicz conditions that are applicable to any convex loss function, thus extending the current theory to a larger panel of losses commonly used in practice such as cross-entropy. Moreover, our analysis provides high-probability bounds on the approximation error under sub-Gaussian gradient noise and only requires the *local* smoothness of the objective function, thus making it applicable to deep neural networks in realistic settings.

1. Introduction

Large neural networks trained with gradient-like methods have proved successful in a wide variety of domains such as natural language processing (Devlin et al., 2019), computer vision (Krizhevsky et al., 2012) and reinforcement learning (Silver et al., 2016). However, the non-convex nature of the associated optimization problem makes the theoretical explanation of this success notoriously difficult. Recently, a lot of progress has been made in that direction by analyzing the convergence of gradient algorithms in a variety of specific settings (Jacot et al., 2018; Ji & Telgarsky, 2019; Nitanda & Suzuki, 2019; Li & Liang, 2018; Chizat & Bach, 2018; Song et al., 2018; Chizat & Bach, 2020; Li & Yuan, 2017). More specifically, in a series of works, Liu et al., 2020a; 2022 proposed a unified view on these results by observing that the optimization problem induced by over-parameterized models surprisingly satisfy a very simple assumption called PL* standing for Polyak-Łojasiewicz (Lojasiewicz, 1963). Indeed, using this assumption, they provide a theoretical explanation of the convergence to zero training loss for a wide variety of large networks such as ResNet (He et al., 2016) and ConvNets (Fukushima et al., 1983). However, despite their generality, their analysis requires the loss to be quadratic, such as *mean square error* (MSE), thus limiting the applicability of the results. Moreover, the direct extension of these results to *cross entropy* (CE) or the *logistic* loss poses significant challenges as their associated optimization problems do not satisfy the PL* condition. In this work, we extend these results to a wider class of loss functions by considering novel Łojasiewicz conditions and analyzing the convergence of *stochastic gradient descent* (SGD) under these assumptions. This analysis is then used to prove convergence to a zero training loss of SGD for neural networks satisfying a local smoothness and uniform conditioning assumption. As a byproduct, we show that the choice of the loss drastically impacts the convergence rates as shown in Table 1 summarizing our results. More precisely, our contribution can be summarized as follows:

1. The introduction of a novel assumption, called *Separable-Łojasiewicz** (SL*), which extend PL* as well as the *Kurdyka-Łojasiewicz** assumption (KL*) to a wider class of non-convex objective functions;
2. The derivation of novel convergence rates for GD and SGD under these assumptions holding with high probability and for a sub-Gaussian gradient noise;
3. We identify three different regimes, depending on the size of the noise and the dimension of the parameter space, leading to three different convergence rates;
4. We show that locally smooth neural networks satisfy

[1]DI ENS, École normale supérieure, CNRS, INRIA, PSL University [2]This work was done while the author was working at Huawei. [3]Huawei Noah's Ark. Correspondence to: Kevin Scaman <kevin.scaman@inria.fr>, Cédric Malherbe <cedric.malherbe@huawei.com>, Ludovic Dos Santos <ludovic.dos.santos@huawei.com>.

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

Table 1. Radius and time sufficient to reach a precision $\varepsilon > 0$ for SGD with high probability. $\kappa \geq 1$ is a measure of the capacity of the model at initialization (see Section 5.4). HL^2 stands for squared Hinge Loss, CE for Cross Entropy and CE^2 for squared Cross Entropy.

Loss function	MSE	HL^2	CE^2	CE	Logistic	Strongly Convex	Convex
Radius	$\Omega(1)$	$\Omega(1)$	$\Omega\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$\Omega\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$\Omega\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$\Omega(1)$	$\Omega(\varepsilon^{-\kappa})$
Time (GD)	$O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$O(\varepsilon^{-1})$	$O(\varepsilon^{-1})$	$O(\varepsilon^{-1})$	$O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$O(\varepsilon^{-1-2\kappa})$
Time (SGD)	$\widetilde{O}(\varepsilon^{-2})$	$\widetilde{O}(\varepsilon^{-2})$	$\widetilde{O}(\varepsilon^{-4})$	$\widetilde{O}(\varepsilon^{-4})$	$\widetilde{O}(\varepsilon^{-4})$	$\widetilde{O}(\varepsilon^{-2})$	$\widetilde{O}(\varepsilon^{-4-4\kappa})$

these assumptions in the over-parameterized regime, and provide the first results proving that these models can achieve zero training loss on any convex loss;

5. We provide numerical experiments showing that the theoretical assumptions we consider are satisfied for vision neural networks largely used in practical settings.

The rest of the paper is organized as follows. In Section 2 we present related works on smooth non-convex optimization and over-parameterized neural networks. Then, we introduce and motivate our framework in Section 3. In Section 4, we show how to derive generic convergence results for this framework using the SL^* and KL^* assumptions. In Section 5, we go on and show how to apply these results to the tuning of over-parameterized networks. Finally, in Section 6, we provide a numerical assessment and show that our novel assumptions are well aligned with what is observed in practice. All the proofs can be found in Appendix A.

2. Related Works

We detail here two lines of works that are closely related to our analysis: smooth non-convex optimization and the analysis of over-parameterized deep learning models.

Non-Convex Optimization. We start with the body of works that is dedicated to the generic analysis of gradient descent methods in non-convex landscapes. First, it has to be noticed that several universal lower bounds for this problem are provided in (Arjevani et al., 2019; Carmon et al., 2019). For smooth and deterministic settings, (Carmon et al., 2019) established that $\Omega(\varepsilon^{-1})$ gradient evaluations are necessary for finding a ε-stationary point (i.e. a point $\theta \in \mathbb{R}^d$ such that $\mathbb{E}\left[\|\nabla f(\theta)\|^2\right] \leq \varepsilon$); and showed that this rate is achieved by gradient descent. For smooth and stochastic settings, (Arjevani et al., 2019) showed that $\Omega(\varepsilon^{-2})$ noisy gradient evaluations are required to reach an ε-stationary point, proving that SGD is optimal with this worst case metric using the $O(\varepsilon^{-2})$ upper bound of (Ghadimi & Lan, 2013). However, all these works only prove convergence towards a stationary point, and not towards a global optimum (i.e. a point $\theta^\star$ such that $\theta^\star \in \operatorname{argmin}_\theta f(\theta)$) as witnessed when optimizing over-parameterized networks. Moreover, they do not exploit any assumption that specifically describe the behavior of functions that naturally appear during the training of neural networks. With regards to these works, our contribution is to introduce the *Separable-Łojasiewicz** (SL^*) and to consider the *Kurdyka-Łojasiewicz** (KL^*) assumption that describe the complexity of the optimization of non-convex functions and we provide novel convergence results under these assumptions.

Over-Parameterized Networks. The second line of works connected to our analysis studies the global convergence and generalization abilities of the gradient descent method for over-parameterized networks (Du et al., 2018; Sankararaman et al., 2020; Allen-Zhu et al., 2019; Li & Liang, 2018; Jacot et al., 2018; Arora et al., 2018; Chizat & Bach, 2018). These works generally consider different sets of assumptions on the activation functions, dataset and the size of the layers to derive convergence results. A first approach proved convergence to the global optimum of the loss function when the width of its layers tends to infinity (Jacot et al., 2018; Nitanda & Suzuki, 2019), using the linear-like behavior of the network around initialization. Other works (Li & Liang, 2018; Chizat & Bach, 2018; Ji & Telgarsky, 2019) focus on small networks with fixed architectures (i.e. two-layer networks), and use techniques such as optimal transport theory (Chizat & Bach, 2018), partial differential equations (Song et al., 2018; Chizat & Bach, 2020) or analysis of the dynamics of the algorithm (Li & Yuan, 2017). In classification settings, (Allen-Zhu et al., 2019) proved that $O(\varepsilon^{-2})$ iterations of SGD are required to reach a precision ε for two layers with a ReLU activation function. Independently, (Ji & Telgarsky, 2019) and (Chen et al., 2019) proved that $O(\varepsilon^{-1})$ steps of SGD are required to reach the same precision using different sets of assumptions with two layers. Finally, (Nitanda & Suzuki, 2019) proved that the $O(\varepsilon^{-1})$ iterations of GD are required to optimize a two-layer network with smooth activation functions, while (Du et al., 2018) managed to show that $O(\ln(1/\varepsilon))$ gradient iterations on a two-layer network with ReLU activation functions with additional assumptions on the dataset. These works are either for specific architectures (e.g. two or three layer networks), specific losses (e.g. MSE), or approximated algorithms (e.g. gradient flow instead of stochastic gradient descent). With regards to these works, our analysis aims at being more generic with regards to the assumptions on the loss function, network architecture and optimization

Algorithm 1 Stochastic gradient descent (SGD)

Input: number of iterations T, gradient step η, initial parameters θ_0
Output: optimized parameters θ_T
1: **for** $t = 0$ to $T-1$ **do**
2: Compute G_t, a noisy approximation of $\nabla f(\theta_t)$
3: $\theta_{t+1} = \theta_t - \eta G_t$
4: **end for**
5: **return** θ_T

algorithm. Finally, our assumption is closely connected to the Polyak-Łojasiewicz (PL) and Kurdyka-Łojasiewicz (KL) conditions (Lojasiewicz, 1963; 1993; Attouch et al., 2010; Noll, 2014; Karimi et al., 2016; Liu et al., 2020a) that lower bound the gradient norm by a function of the approximation error $f(\theta) - f(\theta^\star)$. However, these works do not consider the particular form of the SL* condition (see Assumption 4.4) as a product of two terms, and do not depend on the distance from initialization. As a consequence, the lower bound should hold on the whole space (which is not realistic for neural networks due to the presence of multiple stationary points) or only provide an asymptotic analysis. Moreover, these works assume that the optimum $\theta^\star \in \mathbb{R}^d$ exists, which is not necessarily true for the CE loss.

3. Non-Convex Optimization Setup

Throughout this paper, our objective is to analyze the convergence of stochastic gradient descent (SGD) described in Alg. (1) on non-convex functions that verify additional assumptions called *Łojasiewicz* conditions. More precisely, we consider the generic problem of finding the set of d-dimensional parameters $\theta \in \mathbb{R}^d$ which minimize the value of a function

$$\min_{\theta \in \mathbb{R}^d} f(\theta) \tag{1}$$

where f is a non-convex objective function, and is β-smooth on a ball $\mathcal{B}(\theta_0, R)$ around initialization for some $R > 0$. In the remainder of the paper, we will always denote as $\theta_0 \in \mathbb{R}^d$ the initialization point of SGD, as $\Delta = f(\theta_0) - \inf_{\theta \in \mathbb{R}^d} f(\theta)$ the maximum decrease of the function value, and as $\theta^\star \in \operatorname{argmin}_{\theta \in \mathbb{R}^d} f(\theta)$ a global optimizer when such an optimum exists. Moreover, we assume that the gradient noise is sub-Gaussian in order to obtain high-probability bounds on the approximation error $f(\theta_t) - \inf_{\theta \in \mathbb{R}^d} f(\theta)$.

Assumption 3.1 (Noise assumption)**.** Let $\sigma > 0$ and $(\mathcal{F}_t)_{t \geq 0}$ be the natural filtration associated to the iterates $(\theta_t)_{t \geq 0}$ of Alg. (1). We assume that $(G_t)_{t \geq 0}$ are random variables adapted to $(\mathcal{F}_{t+1})_{t \geq 0}$ such that $\mathbb{E}[G_t | \mathcal{F}_t] = \nabla f(\theta_t)$ and, for any $u \in \mathbb{R}^d$ s.t. $\|u\| = 1$, $\langle G_t - \nabla f(\theta_t), u \rangle$ is $\frac{\sigma}{\sqrt{d}}$-sub-Gaussian when conditioned on $\mathcal{F}_t$.

This relatively standard assumption (see e.g. Fang et al., 2019; Scaman & Malherbe, 2020) covers common noise distributions such as Gaussian or bounded noises, and the parameter σ is an upper bound on the standard deviation of the noise, i.e.

$$\mathbb{E}\left[\|G_t - \nabla f(\theta_t)\|^2\right] \leq \sigma^2 \,. \tag{2}$$

Moreover, for the training of neural networks, the objective function we wish to minimize is usually a loss of the form

$$\mathcal{L}(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(g_\theta(x_i), y_i) \,, \tag{3}$$

where $g_\theta : \mathcal{X} \to \mathcal{Y}$ is a neural network parameterized by some parameters $\theta \in \mathbb{R}^d$, $\ell : \mathcal{Y}^2 \to \mathbb{R}_+$ denotes the loss function and $(x_1, y_1), \dots, (x_n, y_n)$ is a collection of n observations with $x_i \in \mathcal{X}$ (e.g. images, sentences or input data usually preprocessed into vectors) and $y_i \in \mathcal{Y}$ (e.g. classes or regression vectors with $\mathcal{Y} = \mathbb{R}^l$). Under technical assumptions on the neural network, a recent line of works (Liu et al., 2020a; 2022) showed that, when $\ell(x, y) = \|x - y\|^2$ is the MSE loss, this objective function verifies the *PL** condition, i.e. $\forall \theta \in \mathcal{S}$,

$$\|\nabla \mathcal{L}(\theta)\|^2 \geq \mu \mathcal{L}(\theta) \,, \tag{4}$$

where $\mathcal{S}$ is typically a ball $\mathcal{B}(\theta_0, R)$ around initialization. Note that this condition is valid on a subset of the whole parameter space to avoid the presence of saddle points (Choromanska et al., 2015) for which we would have $\|\nabla \mathcal{L}(\theta)\| = 0$ while $\mathcal{L}(\theta) > 0$. This condition, along with the smoothness of the objective function, is sufficient to show convergence of the objective to its global optimum. In what follows, we extend this assumption to make it applicable to more losses of the deep learning literature, and provide high-probability convergence rates which are shown to hold during the training of a large number of neural networks. This extension is twofold: first, the KL* condition will provide results for most major loss functions of the deep learning literature, then the SL* condition will provide more general results applicable to any convex loss function.

Notations. For any $x \in \mathbb{R}^d$, we denote by $\|x\| = (\sum_{i=1}^d x_i^2)^{1/2}$ the standard ℓ_2-norm and by $\mathcal{B}(\theta_0, R) = \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\| \leq R\}$ the ball of radius $R > 0$ centered around $\theta_0 \in \mathbb{R}^d$. A real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if and only if $\forall (x_1, x_2) \in (\mathbb{R}^d)^2$, $\forall \lambda \in [0, 1]$, $f(\lambda x_1 + (1-\lambda) x_2) + \leq \lambda f(x_1) + (1 - \lambda) f(x_2)$, and L-Lipschitz if and only if $\forall (x_1, x_2) \in (\mathbb{R}^d)^2$, $\|f(x_1) - f(x_2)\| \leq L \cdot \|x_1 - x_2\|$. Additionally, a function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if and only if it is differentiable and its gradient is β-Lipschitz. We say that a real random variable X is σ-sub-Gaussian if, $\forall \lambda \in \mathbb{R}$, $\log \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \lambda^2 \sigma^2 / 2$. Finally and for simplicity, for any function $\phi : \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}$, we

Table 2. Details of the integral terms appearing in Proposition 4.2 for standard φ functions, where W_0 is the Lambert W function.

$\varphi(x)$	$-I_{\varphi^{-1}}(x)$	$-I_{\varphi^{-2}}(x)$	$I^{\dagger}_{\varphi^{-1}}(-x)$	$I^{\dagger}_{\varphi^{-2}}(-x)$
$\mathbb{1}\{x>0\}$	$1-x$	$1-x$	$(1-x)_+$	$(1-x)_+$
$\sqrt{x}$	$1-2\sqrt{x}$	$\log(1/x)$	$(1-x/2)^2_+$	$\exp(-x)$
x	$\log(1/x)$	$1/x$	$\exp(-x)$	$1/(1+x)$
$1-e^{-x}$	$-\log\left(e^x-1\right)$	$1/(e^x-1)-\log\left(e^x-1\right)$	$\log(1+e^{-x})$	$\log\left(1+1/W_0(e^x)\right)$

will denote as $I_{\phi,a}(x)=\int_a^x \phi(u)du$, as its pseudo-inverse $I^{\dagger}_{\phi,a}(y)=\sup\{x\in\mathbb{R}: I_{\phi,a}(x)\leq y\}$, and I_ϕ any primitive function of ϕ.

4. Theoretical Analysis of SGD under Łojasiewicz Conditions

In this section, we define the KL* and SL* conditions, and provide high-probability bounds on the convergence of SGD under these conditions. Note that KL* is a particular case of SL*, and thus all results for the former are natural consequences of proofs for the latter. However, for didactic purposes, we first provide the simpler setting of KL*, that is widely applicable and leads to a simpler theoretical analysis. The proofs of all lemmas and propositions can be found in Appendix A.

4.1. The KL* Condition

The linear dependency between $\|\nabla f(\theta)\|^2$ and $f(\theta)$ in the PL* condition of Eq. (4) is rather restrictive, and a natural generalization consists in allowing for non-linear dependencies between the two quantities. This was first considered for power functions by Lojasiewicz (1963), and later extended to arbitrary functions by Kurdyka (1998) and Attouch et al. (2010) with the Kurdyka-Łojasiewicz (KL) condition. Similarly to PL*, the KL* assumption presented below is a *local* extension of the KL condition that only requires the condition to hold in a limited subspace.

Assumption 4.1 (Kurdyka-Łojasiewicz*)**.** Let $\varphi:\mathbb{R}_+\to\mathbb{R}_+$ be a non-decreasing function. Then, the objective function f verifies the φ-KL* condition on $\mathcal{S}\subset\mathbb{R}^d$ if, $\forall\theta\in\mathcal{S}$,

$$\|\nabla f(\theta)\|\geq\varphi\big(f(\theta)\big)\,. \tag{5}$$

As we will see in Section 5, this condition is applicable to most losses of the deep learning literature. It ensures that gradients are large far from the optimum, and implicitly assumes that the objective function is equal to zero at optimum (i.e. $\inf_{\theta\in\mathbb{R}^d} f(\theta)=0$). Moreover, the μ-PL* condition of Liu et al. (2020a) is a particular case of φ-KL* with $\varphi(x)=\sqrt{\mu x}$, and the slope of φ near 0 will impact the convergence rate of SGD. Throughout the paper, we will consider the φ-KL* condition on balls $\mathcal{B}(\theta_0,R)$ centered on the initialization and that $\inf_{\theta\in\mathbb{R}d} f(\theta)=0$. Finally, we point out that the KL condition is usually written as $\kappa'(f(\theta)-f(\theta^\star))\|\nabla f(x)\|\geq 1$, where κ is a differentiable, concave and increasing function s.t. $\kappa(0)=0$. Here the choice of $\varphi(x)=1/\kappa'(x)$ is made for simplicity and readability of the analysis.

4.2. Convergence Rates of SGD Under KL*

We now provide non-asymptotic high-probability bounds on the approximation error of SGD under the φ-KL* condition.

Proposition 4.2. *Let $\delta\in[0,1]$, $\eta\in[0,1/\beta]$ and assume that f is β-smooth on $\mathcal{B}(\theta_0,R)$, f is φ-KL* on $\mathcal{B}(\theta_0,R)$, and the noisy gradients G_t verify Assumption 3.1. Then, with probability at least $1-\delta$, SGD achieves the error*

$$\min_{i\leq t} f(\theta_i)\leq\max\left\{I^{\dagger}_{\varphi^{-2},\Delta}\left(\frac{-\eta t}{2}\right),\; I^{\dagger}_{\varphi^{-1},\Delta}\left(\frac{-R+2\mathsf{C}_{\eta,t}}{2}\right)\right\}+\mathsf{A}_{\eta,t}\,, \tag{6}$$

where $\mathsf{A}_{\eta,t}=4\beta\mathsf{C}^2_{\eta,t}+2d^{-1/2}L\mathsf{C}_{\eta,t}$, $\mathsf{C}_{\eta,t}=\sigma\eta\sqrt{2t\log(6dt/\delta)}$, $L=\|\nabla f(\theta_0)\|+\beta R$, *and* $\Delta=f(\theta_0)-\inf_{\theta\in\mathbb{R}^d} f(\theta)$.

First, we point out that the functions $I^{\dagger}_{\varphi^{-2},\Delta}$ and $I^{\dagger}_{\varphi^{-1},\Delta}$ are provided for several common functions φ in Table 2. This bound contains three terms: 1) a convergence term that decreases when the number of iterations increases, 2) a radius term that requires the control radius R to be large enough to reach a given precision, and 3) a stochastic term $\mathsf{A}_{\eta,t}$ that behaves in $O(\eta\sqrt{t\log(t)})$ when $\sigma>0$ and will tend to 0 if $\eta=o(\sqrt{t\log t})$. When the gradient is exact (i.e. $\sigma=0$), we have $\mathsf{A}_{\eta,t}=\mathsf{C}_{\eta,t}=0$, and the approximation error simplifies to

$$\min_{i\leq t} f(\theta_i)\leq\max\left\{I^{\dagger}_{\varphi^{-2},\Delta}\left(\frac{-\eta t}{2}\right),I^{\dagger}_{\varphi^{-1},\Delta}\left(\frac{-R}{2}\right)\right\}. \tag{7}$$

Moreover, note that Proposition 4.2 controls the best approximation error before time t, instead of the final iterate $f(\theta_t)$. This is necessary, due to the fact that the function is smooth only on $\mathcal{B}(\theta_0,R)$, and the iterates of SGD may fall outside this ball after reaching a good approximation error, after which nothing can be said about the iterates as there are no assumptions on the regularity of the function

outside this ball. Optimizing over the gradient step provides the following convergence rates, with the notation $\mathcal{T}_\varepsilon = \min\{t \geq 0 \,:\, f(\theta_t) \leq \varepsilon\}$.

Corollary 4.3. *Let $\delta \in [0,1]$, $\varepsilon > 0$ and consider the same assumptions as in Proposition 4.2. Then, if*

$$R \geq -2\beta I_{\varphi^{-1},\Delta}\left(\frac{\varepsilon}{2}\right) + \sqrt{\frac{\varepsilon}{4\beta}}\,, \tag{8}$$

for a well-chosen gradient step η_ε, with probability at least $1-\delta$, the time for SGD to reach a precision $\varepsilon > 0$ is bounded by

$$\begin{aligned}\mathcal{T}_\varepsilon \;\leq\; & 1 - 2\beta I_{\varphi^{-2},\Delta}\left(\frac{\varepsilon}{2}\right) \\ & + h\left(\frac{\mathsf{C}_1 \beta\sigma^2 I_{\varphi^{-2},\Delta}\left(\frac{\varepsilon}{2}\right)^2}{\varepsilon}\right) \\ & + h\left(\frac{\mathsf{C}_2 L_\varepsilon^2 \sigma^2 I_{\varphi^{-2},\Delta}\left(\frac{\varepsilon}{2}\right)^2}{d\varepsilon^2}\right)\end{aligned} \tag{9}$$

where $h(x) = x\log(xd/\delta)$, $L_\varepsilon = \|\nabla f(\theta_0)\| + \sqrt{\varepsilon/2\beta} - 2\beta I_{\varphi^{-1},\Delta}\left(\frac{\varepsilon}{2}\right)$ and $\mathsf{C}_1, \mathsf{C}_2$ are fixed constants.

First, note that the ball around initialization should have a radius of order

$$R \geq \Omega\left(-I_{\varphi^{-1}}(\varepsilon/2)\right)\,, \tag{10}$$

in order to reach a precision ε, as the second term is negligible as ε tends to 0. Second, there are three different convergence rates, depending on the precision needed:

1. **Deterministic (short-term):** When the noise is small $\sigma \ll \sqrt{\varepsilon}/I_{\varphi^{-2}}(\varepsilon/2)$, the first term dominates and
$$\mathcal{T}_\varepsilon \leq O\left(-I_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right)\right) \tag{11}$$
2. **Stochastic (mid-term):** When the dimension is large $d \gg L_\varepsilon^2/\beta\varepsilon$, the second term dominates and
$$\mathcal{T}_\varepsilon \leq O\left(\frac{I_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right)^2}{\varepsilon}\log\left(\frac{-dI_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right)}{\delta\varepsilon}\right)\right)\,. \tag{12}$$
3. **Stochastic (long-term):** When the precision is small $\varepsilon \ll L_\varepsilon^2/\beta d$, the third term dominates and
$$\mathcal{T}_\varepsilon \leq O\left(\frac{I_{\varphi^{-1}}\left(\frac{\varepsilon}{2}\right)^2 I_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right)^2}{d\varepsilon^2}\log\left(\frac{-I_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right)}{\delta\varepsilon}\right)\right)\,. \tag{13}$$

While the third convergence rate will become dominant as $\varepsilon \to 0$, it is worth noting that this term is divided by the dimension, and will thus be negligible in practice for very high-dimensional problems and moderate target precision or number of iterations. The reason for this is that, as the dimension increases, the norm of the gradient noise will remain constant, while its projection along the true gradient $\langle G_t - \nabla f(\theta_t), \nabla f(\theta_t)\rangle$ will scale in $O(1/\sqrt{d})$, and thus become negligible as the dimension goes to infinity. However, note that this behavior is tightly connected to Assumption 3.1 and may not hold for less isotropic noises.

4.3. The SL* Condition

Our convergence results can be further extended to settings in which, instead of controlling the gradient norm on a ball $\mathcal{B}(\theta_0, R)$ around initialization, the lower bound on $\|\nabla f(\theta)\|$ also depends on the distance to initialization.

Assumption 4.4 (Separable-Łojasiewicz*)**.** Let $\theta_0 \in \mathbb{R}^d$, and $\phi, \psi : \mathbb{R} \to \mathbb{R}_+$ be two non-increasing functions. Then, the objective function f verifies the (ϕ, ψ, θ_0)-SL* condition if, $\forall \theta \in \mathbb{R}^d$,

$$\|\nabla f(\theta)\| \geq \phi\big(f(\theta_0) - f(\theta)\big)\,\psi\big(\|\theta - \theta_0\|\big)\,. \tag{14}$$

Moreover, we will say that a function is (ϕ, ψ, θ_0)-SL* on the ball $\mathcal{B}(\theta_0, R)$ if $\psi(x) = 0$ for any $x \geq R$. This assumption allows for more flexibility, and its particular form as a product of two terms is motivated by deep learning objectives (see Section 5). Finally, it is interesting to note that the φ-KL* condition on $\mathcal{B}(\theta_0, R)$ is a particular case of (ϕ, ψ, θ_0)-SL* where $\phi(x) = \varphi(f(\theta_0) - x)$ and $\psi(x) = \mathbb{1}\{x \leq R\}$.

Remark 4.5. The lack of dependence in $f(\theta^\star)$ in SL* compared to the standard PL and KL conditions allows to consider settings in which $\theta^\star$ does not exist (e.g. $f(\theta) = \ln(1 + \exp(-\theta))$ appearing in the cross entropy loss or even $f(\theta) = \theta$ for which $\inf_\theta f(\theta) = -\infty$).

4.4. Convergence Rates of SGD Under SL*

We now provide non-asymptotic high-probability bounds on the approximation error of SGD under the (ϕ, ψ, θ_0)-SL* condition on the ball $\mathcal{B}(\theta_0, R)$.

Proposition 4.6. *Let $\delta \in [0,1]$, $\eta \in [0, 1/\beta]$, and assume that f is L-Lipschitz and β-smooth on $B(\theta_0, R)$, f is (ϕ, ψ, θ_0)-SL* on $B(\theta_0, R)$, and the noisy gradients G_t verify Assumption 3.1. Then, with probability at least $1-\delta$, SGD achieves a decrease*

$$\min_{i \leq t} f(\theta_i) \leq f(\theta_0) - I^\dagger_{\chi,0}\left(\frac{\eta t}{2}\right) + \mathsf{A}_{\eta,t}\,, \tag{15}$$

where $\chi(x) = \phi(x)^{-2}(\psi \circ I^\dagger_{\psi,\mathsf{C}_{\eta,t}} \circ 2I_{\phi^{-1},0}(x))^{-2}$, $\mathsf{A}_{\eta,t} = 4\beta\mathsf{C}_{\eta,t}^2 + 2d^{-1/2}L\mathsf{C}_{\eta,t}$, and $\mathsf{C}_{\eta,t} = \sigma\eta\sqrt{2t\log(6dt/\delta)}$.

This result is more general than Proposition 4.2, and can be applied to prove convergence to a global minimum, or

decrease beyond a local minima. The key function appearing in the convergence rate χ is the product of two terms: 1) the function ϕ^{-2} that also appears in Proposition 4.2, and 2) the function $(\psi \circ I^{\dagger}_{\psi,\mathsf{C}_{\eta,t}} \circ 2I_{\phi^{-1},0}(x))^{-2}$ that will decrease the convergence rate of SGD as the iterates move away from the initialization, and $\psi(\|\theta_t - \theta_0\|)$ becomes small. The term $I^{\dagger}_{\psi,\mathsf{C}_{\eta,t}} \circ 2I_{\phi^{-1},0}(x)$ is indeed an upper bound on the distance from initalization $\|\theta_t - \theta_0\|$ upon reaching a decrease of $f(\theta_0) - f(\theta_t) = x$. In Section 5.4, we will see that Proposition 4.6 can be applied to prove the convergence of SGD to a global minima of the training loss when the loss function ℓ is convex. Finally, we point out that these results can be extended to the current iterates $f(\theta_t)$ instead of the minimum observed during the optimization $\min_{i \leq t} f(\theta_i)$ if the regularity assumptions (Lipschitz continuity, smoothness and SL* condition) hold on the whole space.

5. Application to Deep Learning

In this section, we show how to apply the theoretical results presented in Section 4 to derive global convergence of the training loss for locally smooth and over-parameterized neural networks.

5.1. Background

As discussed in Section 3, tuning over-parameterized neural networks typically requires the minimization of the empirical risk

$$\mathcal{L}(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(g_\theta(x_i), y_i)\,, \tag{16}$$

with SGD. In this case, a direct application of the chain rule gives that

$$\nabla \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \partial_\theta g_\theta(x_i)^\top \nabla_x \ell(g_\theta(x_i), y_i)\,, \tag{17}$$

and the norm of the gradient can be reformulated as a quadratic function involving the *Neural Tangent Kernel* of Jacot et al. (2018) defined below.

Definition 5.1 (Neural Tangent Kernel, adapted from Jacot et al., 2018)**.** Let $g_\theta : \mathcal{X} \to \mathbb{R}^l$ be a differentiable model. The *Neural Tangent Kernel* (NTK) $\kappa_\theta : \mathcal{X}^2 \to \mathbb{R}^{l \times l}$ of the model g at $\theta \in \mathbb{R}^d$ is the function defined by, $\forall x, y \in \mathcal{X}$,

$$\kappa_\theta(x, y) = \partial_\theta g_\theta(x) \times \partial_\theta g_\theta(y)^\top\,, \tag{18}$$

where $\partial_\theta g_\theta(x) \in \mathbb{R}^{l \times d}$ is the Jacobian matrix of $\theta \mapsto g_\theta(x)$.

As shown in e.g. Liu et al. (2020a), controlling the spectrum of the NTK directly implies a lower bound on the gradient norm, as

$$\|\nabla \mathcal{L}(\theta)\|^2 \geq \lambda_{\min}\left((\kappa_\theta(x_i, x_j))_{i,j \in [\![1,n]\!]}\right) \times \frac{1}{n^2} \sum_{i=1}^{n} \|\nabla_x \ell(g_\theta(x_i), y_i)\|^2\,, \tag{19}$$

where $\lambda_{\min}(\cdot)$ denotes the minimal eignenvalue of the NTK matrix. Thus, the gradient of the objective function can be lower bounded by a product of two terms: one related to the behavior of the gradient of the loss function ℓ, and another related to over-parameterization of the model at initialization through the NTK.

5.2. Assumptions on the Network and Losses

For clarity, we list here the assumptions used in our analysis. The first assumption relates to the local smoothness of the neural network on a ball $\mathcal{B}(\theta_0, R)$ around initialization.

Assumption 5.2 (Model Regularity)**.** We assume that, for any $x \in \mathcal{X}$, the model $\theta \mapsto g_\theta(x)$ is L_g-Lipschitz and β_g-smooth in $\mathcal{B}(\theta_0, R)$.

The above assumption is satisfied as soon as all the layers and activation functions of the model are smooth, by composition of locally smooth functions. However, it is interesting to note that, for a given model, increasing the radius R will often increase the smoothness (as deep learning models are generally not *globally* smooth). Nonetheless, recent results (see Theorem 3.2 of Liu et al. (2020b)) show that, for a large class of neural networks, and any radius $R > 0$, the smoothness of the model on $\mathcal{B}(\theta_0, R)$ tends to 0 as the number of neurons tend to $+\infty$ (in $1/\sqrt{m}$ where m is the number of neurons). As a consequence, for any desired precision $\varepsilon > 0$, we can for example fix $R = -2I_{\varphi^{-1},\Delta}\left(\frac{\varepsilon}{2}\right) + \sqrt{\frac{\varepsilon}{4}}$ (i.e. Eq. (8) where $\beta = 1$). Then, there is a number of neurons such that $\beta = 1$ in $\mathcal{B}(\theta_0, R)$, and the condition holds. Thus, we deduce that imposing a fixed smoothness on any given radius R is possible for standard neural networks with a large number of neurons. The next assumption relates the conditioning of the over-parameterized model around initialization, and is adapted from Liu et al. (2020a).

Assumption 5.3 (Uniform conditioning)**.** Let $\mu > 0$, $\mathcal{D} = (x_i, y_i)_{i \in \{1,n\}}$ a training dataset, and $g_\theta : \mathcal{X} \to \mathbb{R}^l$ a model. Then the smallest eigenvalue of the tangent kernel of g_θ satisfies, $\forall \theta \in \mathcal{B}(\theta_0, R)$,

$$\lambda_{\min}\left((\kappa_\theta(x_i, x_j))_{i,j \in [\![1,n]\!]}\right) \geq \mu\,. \tag{20}$$

This assumption provides a measure of the conditioning of the NTK matrix around initialization through the value of μ, and several rencent works developed techniques to bound this value. More precisely, Liu et al. (2020b) showed that

$|\lambda_{\min}(\kappa(\theta)) - \lambda_{\min}(\kappa(\theta_0))| = O(1/\sqrt{m})$ for any fully-connected network with smooth Lipschitz activation functions including MLP, ResNet and ConvNet where m denotes the width of the network. In other words, they show that the NTK minimum eigenvalues are nearly constant around initialization for smooth networks with large m. Moreover, they go on and show that the tools provided in (Du et al., 2019; 2018) allow to bound the NTK eigenvalue at initialization thus providing an interval for μ combined with their result for any smooth over-parameterized network. For further details on this topic, we refer the reader to (Liu et al., 2020a), Section 1. Finally, we consider that the loss function is sufficiently regular.

Assumption 5.4 (Loss Regularity)**.** The loss function $\ell(x, y)$ is L_ℓ-Lipschitz and β_ℓ-smooth w.r.t. its first input.

This assumption is necessary to ensure smoothness of the objective function $\mathcal{L}$ on the ball $\mathcal{B}(\theta_0, R)$, and is satisfied by most losses of the deep learning literature.

5.3. Convergence Rates for φ-KL* Loss Functions

As shown in Table 3, most losses of the literature verify a φ-KL* condition with φ ranging from linear to square root behaviors. When such a condition is verified, Eq. (19) implies that the objective function $\mathcal{L}$ is itself KL*. More specifically, two cases should be distinguished depending on the convexity of φ^2.

Lemma 5.5. *Under Assumptions 5.2, 5.3 and 5.4, if the loss function ℓ is φ-KL* on $\mathcal{Y}$ w.r.t. its first input, the objective $\mathcal{L}$ verifies the $\frac{\sqrt{\mu}}{n}\varphi$-KL* condition. Furthermore, if φ^2 is convex, then $\mathcal{L}$ verifies the $\sqrt{\frac{\mu}{n}}\varphi$-KL* condition.*

As the convergence rates in Corollary 4.3 depend on the integral of φ^{-2}, the additional multiplicative factor in $1/\sqrt{n}$ when φ^2 is not convex will induce a multiplicative factor n on the convergence rate. While such a multiplicative factor will significantly slow down convergence and may be prohibitive in practice, it is worth mentioning that the convergence rate is not changed. Applying Corollary 4.3 thus gives the following convergence rate without any additional assumption on the function φ.

Proposition 5.6. *Let $\delta \in [0, 1]$, $\varepsilon > 0$, and $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ any increasing function such that $\lim_{x\to+\infty} \omega(x) = +\infty$. Moreover, consider the same assumptions as in Lemma 5.5, and assume that the noisy gradients G_t verify Assumption 3.1. Then, there is a gradient step η_ε such that, if $R \geq -\omega(1/\varepsilon) I_{\varphi^{-1}}(\varepsilon/2))$, with probability at least $1 - \delta$, the time for SGD to reach a precision $\varepsilon > 0$ is bounded by*

$$\mathcal{T}_\varepsilon \leq O\left(\frac{I_{\varphi^{-1}}\left(\frac{\varepsilon}{2}\right)^2 I_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right)^2}{\varepsilon^2} \log\left(\frac{-I_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right)}{\delta\varepsilon}\right)\right). \tag{21}$$

First, note that, in most cases of interest, we have $I_{\varphi^{-1}}\left(\frac{\varepsilon}{2}\right) = O(\log(1/\varepsilon))$ and $I_{\varphi^{-2}}\left(\frac{\varepsilon}{2}\right) = O(\varepsilon^{-\alpha})$ where $\alpha > 0$, and thus in such a case

$$\mathcal{T}_\varepsilon \leq \widetilde{O}\left(I_{\varphi^{-2}}\left(\varepsilon\right)^2 \varepsilon^{-2}\right). \tag{22}$$

For example, for MSE, we have $I_{\varphi^{-2}}(\varepsilon/2) = O(\log(1/\varepsilon))$ and the convergence rate is in $\widetilde{O}(\varepsilon^{-2})$, while for cross entropy, we have $I_{\varphi^{-2}}(\varepsilon/2) = O(\varepsilon^{-1})$ and the convergence rate is in $\widetilde{O}(\varepsilon^{-4})$. Convergence rates for other common losses are available in Table 1. Note that, as shown in Lemma 5.5, the convexity of the function φ^2 will increase the speed of convergence. However, this effect only changes the constant factors but the limiting behavior remains the same.

5.4. Convergence Rates for Convex Loss Functions

As shown in Table 3, most common losses ℓ verify the φ-KL* condition on the whole input space. However, this condition is not necessarily implied by convexity, and a general convergence rate for convex losses cannot be deduced from Proposition 5.6. Fortunately, convexity does imply a lower bound in the gradient norm, as, for convex functions f, we have, for any $\theta \in \mathbb{R}^d$ and $\theta^\star \in \operatorname{argmin}_{\theta\in\mathbb{R}^d} f(\theta)$,

$$\|\nabla f(\theta)\| \geq \frac{f(\theta) - f(\theta^\star)}{\|\theta - \theta^\star\|}. \tag{23}$$

As a consequence, controlling the distance from initialization should also imply a Łojasiewicz condition. To understand why a new Łojasiewicz condition is required, let us first try to apply Corollary 4.3 by noting that, if the model is L_g-Lipschitz, the objective function $\mathcal{L}$ is then φ-KL* on $\mathcal{B}(\theta_0, R)$ where $\varphi(x) = x/L_g(\|\theta_0 - \theta^\star\| + R)$. Unfortunately, the condition on the radius thus becomes

$$R \geq 2L_g(R + \|\theta^\star - \theta_0\|)\ln(2\Delta/\varepsilon) + \sqrt{\beta\varepsilon/2}\,, \tag{24}$$

and as ε tends to 0, the condition will become impossible to verify. Fortunately, we can still use the SL* condition to derive meaningful convergence rates in such a setting.

Lemma 5.7. *Under Assumptions 5.2 and 5.3, if ℓ is convex w.r.t. its first input, the objective $\mathcal{L}$ verifies the SL* condition with $\phi(x) = (\Delta - x)_+$ and $\psi(x) = \kappa^{-1}\mathbb{1}\{x < R\}/(\|\theta_0 - \theta^\star\| + x)$, where $\kappa = L_g\sqrt{n/\mu}$ and $\theta^\star \in \operatorname{argmin}_\theta \mathcal{L}(\theta)$.*

Note that the multiplicative term $1/(\|\theta_0 - \theta^\star\| + x)$ in $\psi(x)$ will slow down convergence as the distance to initialization increases, and thus will lead to slower convergence rates. Equipped with this lemma one can directly apply Proposition 4.6 to obtain the following convergence result.

[1] The hinge loss and MAE are not smooth, and the gradient is not defined everywhere. These thus require a more general analysis than that provided in this work. For the sake of completeness, we nonetheless provide these losses to show the generality of the KL* condition.

Table 3. Description of standard losses and their associated KL* functions.

Name	Loss $\ell(x, y)$	KL* function $\varphi(x)$	φ^2 convex	ℓ smooth
MSE	$\|x - y\|^2$	$2\sqrt{x}$	✓	✓
Logistic loss	$\log(1 + e^{-xy})$	$1 - e^{-x}$	✗	✓
Cross entropy	$-\sum_i y_i \log(e^{x_i} / \sum_j e^{x_j})$	$1 - e^{-x}$	✗	✓
Squared CE	$\sum_i y_i \log(e^{x_i} / \sum_j e^{x_j})^2$	$\min\{x, \sqrt{x}\}$	✓	✓
Hinge loss[1]	$\sum_i \max\{0, 1 - x_i y_i\}$	$\mathbb{1}\{x > 0\}$	✗	✗
Squared hinge	$\sum_i \max\{0, 1 - x_i y_i\}^2$	$2\sqrt{x}$	✓	✓
MAE [1]	$\sum_i \|x_i - y_i\|$	1	✓	✗

Proposition 5.8. *Let $\delta \in [0, 1]$, $\varepsilon > 0$, and $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ any increasing function such that $\lim_{x \to +\infty} \omega(x) = +\infty$. Moreover, consider the same assumptions as in Lemma 5.7, and assume that the noisy gradients G_t verify Assumption 3.1. Then, there is a gradient step η_ε such that, if $R \geq \omega(1/\varepsilon)\varepsilon^{-\kappa}$, with probability at least $1 - \delta$, the time for SGD to reach a precision $\varepsilon > 0$ is bounded by*

$$\mathcal{T}_\varepsilon \leq \widetilde{O}\left(\varepsilon^{-4-4\kappa}\right) . \tag{25}$$

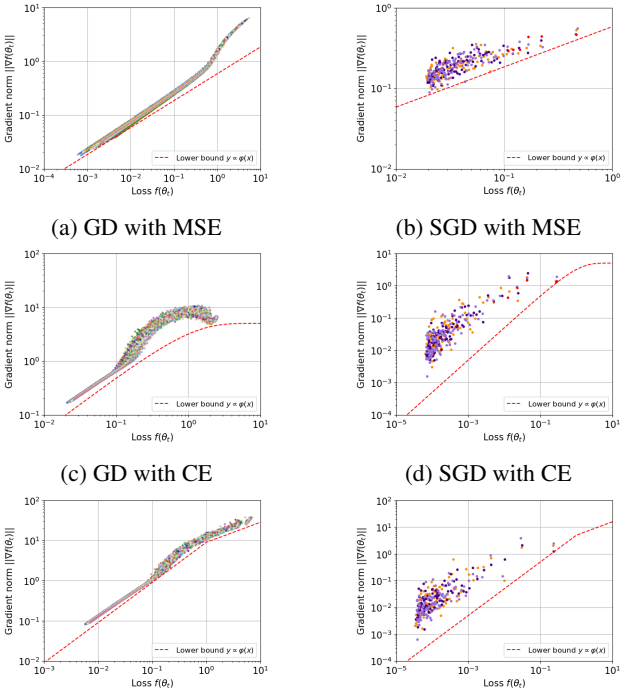
More precise bounds on the approximation error are available in Appendix A. While the convergence rate can be relatively slow when $\kappa \gg 1$, this result is nonetheless the first to prove convergence of SGD for arbitrary convex losses, thus showing the flexibility of the approach and robustness of SGD for the training of deep learning architectures.

6. Numerical experiments

The purpose of this section is to illustrate that the assumptions and results presented in the paper are satisfied in practice on a standard neural network used in vision for both the deterministic and stochastic settings.

Protocol and Metrics. To test our theoretical framework, we performed a series of experiments over the MNIST (LeCun et al., 2010) and CIFAR10 (Krizhevsky et al., 2009) classification datasets (denoted by $(x_i, y_i)_{i \leq n}$) using the ResNet-18 convolutional network (He et al., 2016) with smooth GELU activation functions (Hendrycks & Gimpel, 2016) denoted here by g_θ and containing more than 10^6 parameters. We considered the minimization of the empirical loss $f(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(g_\theta(x_i), y_i)$ using three different loss functions ℓ: (1) the mean squared error (MSE), (2) the cross entropy (CE) and (3) the squared cross entropy (CE2). To perform the optimization, we ran a mini-batch GD (deterministic) and SGD (stochastic) over the empirical loss using PyTorch (Paszke et al.) and the Pytorch image models library (Wightman, 2019).

We focus here on the MNIST dataset. For the deterministic setting we performed a gradient descent (GD) on the first batch of size $B = 128$ of the dataset at each epoch and run



(a) GD with MSE (b) SGD with MSE

(c) GD with CE (d) SGD with CE

(e) GD with CE2 (f) SGD with CE2

Figure 1. Gradient norm vs. overall train loss for 1000 (GD) and 3 (SGD) runs (one color per run) with ResNet-18 on MNIST using different loss functions.

this experiment 1000 times while for the stochastic setting we used the full dataset. We also point out that the same experiments have also been tested on the CIFAR dataset (provided in Appendix B due to space issue) and that the same conclusions hold.

The initial parameter θ_0 was set randomly using the default initialization method of PyTorch, the learning rate η was selected over a grid $[1, 0.1, 0.01]$ providing the best training loss at the end of the optimization consisting of $T = 100$ epochs. At the beginning of each epoch $t \in \{0, \dots, T-1\}$, we recorded the value of the gradient norm $\|\nabla f(\theta_t)\|$ and the loss $f(\theta_t)$ computed over the full dataset for SGD and first batch for GD. The profile of the runs iterates of the gradient norm vs. the loss recorded during the optimization are displayed in Figure 1.

Discussion. First, by looking at the values of the loss $f(\theta_t)$ recording during the optimization (x-axis), it is important to note that the model achieves almost zero training loss, suggesting that the model is over-parameterized enough to achieve global convergence as expected in the *lazy training* regime (Chizat et al., 2019; Geiger et al., 2020). We now investigate whether the assumptions formulated in the paper that can explain this phenomenon are satisfied. Looking at the profile of the gradient norm vs. the loss recorded during the optimization and the lower bounds $y \propto \varphi(f(\theta_t))$ provided in Table 3 and plotted here in red lines, the decreasing rate of each lower bound exactly matches the empirical slope, thus validating that the SL* assumption is satisfied in practice. More precisely, for the MSE the general slope is equal to 1/2 as opposed to other losses, illustrating the presence of the square root in φ in practical settings. Finally, observe that, without surprise, the points are more spread in the stochastic setting.

7. Conclusion

In this work, we introduced two novel Łojasiewicz conditions called KL* and SL* that extend PL* to a larger class of objective functions. We then provided high-probability bounds on the convergence rate of SGD under these assumptions, and used these results to prove the convergence to zero training loss of locally smooth and over-parameterized neural networks for a large panel of loss functions. In particular, our results provide the first convergence rates applicable to any convex loss function, thus extending the applicability of the Łojasiewicz approach to the analysis of neural network training. The derived convergence rates heavily depend on the behavior of the loss function near optimum, and could help practitioners decide which loss function to use during training. Promising research directions include: (1) investigating whether the convergence rates provided in the document are tight and optimal, (2) extending the bounds we obtain to the analysis of non-smooth over-parameterized networks and (3) investigating whether the KL* and SL* assumptions are satisfied in practice for a wider variety of networks, including natural language processing and reinforcement learning networks.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 2019.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, pp. 1–50, 2019.

Chen, Z., Cao, Y., Zou, D., and Gu, Q. How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360*, 2019.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3040–3050, 2018.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *NeurIPS 2019-33rd Conference on Neural Information Processing Systems*, 2019.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pp. 192–204. PMLR, 2015.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Fang, C., Lin, Z., and Zhang, T. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pp. 1192–1234. PMLR, 2019.

Fukushima, K., Miyake, S., and Ito, T. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5):826–834, 1983.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8580–8589, 2018.

Ji, Z. and Telgarsky, M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Kurdyka, K. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48 (3):769–783, 1998.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8168–8177, 2018.

Li, Y. and Yuan, Y. Convergence analysis of two-layer neural networks with relu activation. *Advances in Neural Information Processing Systems*, 30:597–607, 2017.

Liu, C., Zhu, L., and Belkin, M. Toward a theory of optimization for over-parameterized systems of nonlinear equations: the lessons of deep learning. *CoRR*, abs/2003.00307, 2020a.

Liu, C., Zhu, L., and Belkin, M. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33, 2020b.

Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

Lojasiewicz, S. Une propriété topologique des sous-ensembles analytiques réels: Les equations ´ aux dérivées partielles. *Editions du centre National de la Recherche Scientifique*, pp. 87—-89, 1963.

Lojasiewicz, S. Sur la géométrie semi- et sous-analytique. *Annales de l'Institut Fourier*, 43:1575—-1595, 1993.

Nitanda, A. and Suzuki, T. Refined generalization analysis of gradient descent for overparameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Noll, D. Convergence of non-smooth descent methods using the kurdyka–łojasiewicz inequality. *Journal of Optimization Theory and Applications*, 160(2):553–572, 2014.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library.

Sankararaman, K. A., De, S., Xu, Z., Huang, W. R., and Goldstein, T. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *International Conference on Machine Learning*, pp. 8469–8479. PMLR, 2020.

Scaman, K. and Malherbe, C. Robustness analysis of non-convex stochastic gradient descent using biased expectations. *Advances in Neural Information Processing Systems*, 33, 2020.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Song, M., Montanari, A., and Nguyen, P. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.

Wightman, R. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

A. Proofs of the theoretical section

A.1. Proofs of the analysis of stochastic gradient descent

As φ-KL* on $\mathcal{B}(\theta_0, R)$ is a particular case of (ϕ, ψ, θ_0)-SL*, we first provide a proof of the latter, and will then derive the convergence for φ-KL* as a consequence.

Proof of Proposition 4.6. Let $X_t = G_t - \nabla f(\theta_t)$ be the gradient noise at iteration $t \geq 0$. If $\eta \in [0, 1/\beta]$, using Lemma 3.4 in (Bubeck et al., 2015) for β-smooth functions, a direct calculation gives

$$\begin{aligned} f(\theta_{t+1}) - f(\theta_t) &\leq \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t \rangle + \tfrac{\beta}{2}\|\theta_{t+1} - \theta_t\|^2 \\ &= -\eta \langle \nabla f(\theta_t), \nabla f(\theta_t) + X_t \rangle + \tfrac{\beta}{2}\|\nabla f(\theta_t) + X_t\|^2 \\ &= -\eta \left(1 - \tfrac{\beta\eta}{2}\right) \|\nabla f(\theta_t)\|^2 + \tfrac{\beta\eta^2}{2}\|X_t\|^2 - \eta(1 - \beta\eta)\langle X_t, \nabla f(\theta_t)\rangle \\ &= -\eta \left(1 - \tfrac{\beta\eta}{2}\right) \|\nabla f(\theta_t)\|^2 + A_{t+1} - A_t \,, \end{aligned} \tag{26}$$

where $A_t = \sum_{i<t} \frac{\beta\eta^2}{2}\|X_i\|^2 - \eta(1-\beta\eta)\sum_{i<t}\langle X_i, \nabla f(\theta_i)\rangle$. Moreover, we have

$$\|\theta_t - \theta_0\| = \eta \left\| \sum_{i<t} (\nabla f(\theta_i) + X_i) \right\| \leq \eta \left\| \sum_{i<t} \nabla f(\theta_i) \right\| + B_t \,, \tag{27}$$

where $B_t = \eta \| \sum_{i<t} X_i \|$. By assumption, the noise terms A_t and B_t are sub-Gaussian, and we can control all of them with high probability using a union bound.

Lemma A.1. *Let $\delta \in [0,1]$ and $t \geq 0$. With probability $1-\delta$, we have, simultaneously, $\forall i \in \{1, \dots, t\}$,*

$$|A_i| \leq \mathsf{A}_{\eta,t}/2 \quad \textit{and} \quad B_i \leq \mathsf{C}_{\eta,t} \,, \tag{28}$$

where $\mathsf{A}_{\eta,t} = 4\beta\mathsf{C}_{\eta,t}^2 + 2d^{-1/2}L\mathsf{C}_{\eta,t}$ and $\mathsf{C}_{\eta,t} = \sigma\eta\sqrt{2t\log(6dt/\delta)}$.

Proof. First, note that all $(X_{j,k})_{j\in\{0,\dots,t-1\},k\in\{1,\dots,d\}}$ and $(\langle X_j, \nabla f(\theta_j)\rangle / \|\nabla f(\theta_j)\|)_{j\in\{0,\dots,t-1\}}$ are $\frac{\sigma}{\sqrt{d}}$-sub-Gaussian when conditioned on $\mathcal{F}_j$, and thus, for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda X_{j,k}} \mid \mathcal{F}_j\right] \leq e^{\lambda^2\sigma^2/2d} \,, \tag{29}$$

and

$$\mathbb{E}\left[e^{\lambda\langle X_j, \nabla f(\theta_j)\rangle} \mid \mathcal{F}_j\right] \leq e^{\lambda^2\|\nabla f(\theta_j)\|^2\sigma^2/2d} \leq e^{\lambda^2 L^2\sigma^2/2d} \,. \tag{30}$$

Thus, by Chernoff's bound, for any $i \in \{1, \dots, t\}$ and $s \geq 0$, we have

$$\mathbb{P}\left(\sum_{j<i} X_{j,k} \geq s\right) \leq \min_{\lambda>0} \frac{\mathbb{E}\left[\prod_{j<i} e^{\lambda X_{j,k}}\right]}{e^{\lambda s}} \leq \min_{\lambda>0} e^{t\lambda^2\sigma^2/2d - \lambda s} = e^{-s^2 d/2t\sigma^2} \,. \tag{31}$$

As the same result holds for $-\sum_{j<i} X_{j,k}$, we have, for any $\delta' \in [0,1]$,

$$\mathbb{P}\left(\left|\sum_{j<i} X_{j,k}\right| \geq \sigma\sqrt{2t\log(2/\delta')/d}\right) \leq \delta' \,. \tag{32}$$

Using the same argument, we have, for any $i \in \{1, \dots, t\}$,

$$\mathbb{P}\left(\left|\sum_{j<i} \langle X_j, \nabla f(\theta_j)\rangle\right| \geq L\sigma\sqrt{2t\log(2/\delta')}\right) \leq \delta' \,. \tag{33}$$

Finally, classical results for $\frac{\sigma}{\sqrt{d}}$-sub-Gaussian random variables state that $\mathbb{E}\left[e^{X_{i,k}^2 d/8\sigma^2}\right] \leq 2$, and thus

$$\mathbb{P}\left(X_{i,k}^2 \geq s\right) \leq \frac{\mathbb{E}\left[e^{X_{i,k}^2 d/8\sigma^2}\right]}{e^{sd/8\sigma^2}} \leq 2e^{-sd/8\sigma^2}\,, \tag{34}$$

and, for any $i \in \{1,\dots,t\}$ and $k \in \{1,\dots,d\}$,

$$\mathbb{P}\left(X_{i,k}^2 \geq 8\sigma^2 \log(2/\delta')/d\right) \leq \delta'\,. \tag{35}$$

Using the union bound, we now combine the $(2d+1)t$ equations in Eq. (32), Eq. (33) and Eq. (35), and obtain that, with probability at least $1-(2d+1)t\delta'$, simultaneously, for all $i \in \{1,\dots,t\}$ and $k \in \{1,\dots,d\}$,

$$\begin{array}{rcl} \left|\sum_{j<i} X_{j,k}\right| & \leq & \sigma\sqrt{2t\log(2/\delta')/d} \\ \left|\sum_{j<i}\langle X_j, \nabla f(\theta_j)\rangle\right| & \leq & L\sigma\sqrt{2t\log(2/\delta')/d} \\ X_{i,k}^2 & \leq & 8\sigma^2\log(2/\delta')/d\,. \end{array} \tag{36}$$

As a consequence, we have

$$\begin{array}{rcl} |A_i| & = & \left|\sum_{j<i}\frac{\beta\eta^2}{2}\|X_j\|^2 - \eta(1-\beta\eta)\sum_{j<i}\langle X_j, \nabla f(\theta_j)\rangle\right| \\ & \leq & \frac{\beta\eta^2}{2}\sum_{j<i,k\leq d} X_{j,k}^2 + \eta\left|\sum_{j<i}\langle X_j, \nabla f(\theta_j)\rangle\right| \\ & \leq & 4\beta\eta^2 t\sigma^2\log(2/\delta') + \eta L\sigma\sqrt{2t\log(2/\delta')/d}\,, \end{array} \tag{37}$$

and

$$B_i = \eta\left\|\sum_{i<t} X_i\right\| = \eta\sqrt{\sum_{k\leq d}\left(\sum_{i<t} X_{i,k}\right)^2} \leq \eta\sigma\sqrt{2t\log(2/\delta')}\,. \tag{38}$$

We obtain the desired result by taking $\delta' = \delta/3dt$ and reformulating the bounds with $\mathsf{C}_{\eta,t} = \sigma\eta\sqrt{2t\log(6dt/\delta)}$. □

Then, using Lemma A.1 to bound A_i and B_i for all $i < t$, with $f_i = f(\theta_0) - f(\theta_i) + A_i + \mathsf{A}_{\eta,t}/2$ and $d_i = \eta\sum_{j<i}\|\nabla f(\theta_j)\| + \mathsf{C}_{\eta,t}$, Eq. (26), Eq. (27), and the SL* condition becomes

$$f_{i+1} - f_i \geq \eta a\|\nabla f(\theta_i)\|^2 \tag{39}$$
$$d_{i+1} - d_i = \eta\|\nabla f(\theta_i)\| \tag{40}$$
$$\|\theta_i - \theta_0\| \leq d_i \tag{41}$$
$$\|\nabla f(\theta_i)\| \geq \phi\left(f_i\right)\psi\left(d_i\right) \tag{42}$$

where $a = 1-\beta\eta/2$. Using these equations, we can derive a bound on d_t as follows. First, we have

$$\frac{f_{i+1}-f_i}{\phi(f_i)} \geq \frac{\eta a\|\nabla f(\theta_i)\|^2}{\phi(f_i)} = \frac{a(d_{i+1}-d_i)\|\nabla f(\theta_i)\|}{\phi(f_i)} \geq a(d_{i+1}-d_i)\psi(d_i)\,. \tag{43}$$

Summing both sides for $i \in \{0, t-1\}$ and bounding these Riemann sum by their corresponding integrals (as ϕ^{-1} is non-decreasing and ψ is non-increasing) leads to

$$\int_{\mathsf{A}_{\eta,t}/2}^{f_i}\frac{du}{\phi(u)} \geq a\int_{\mathsf{C}_{\eta,t}}^{d_i}\psi(v)dv\,, \tag{44}$$

and thus, for all $i \leq t$,

$$\|\theta_i - \theta_0\| \leq I^{\dagger}_{\psi,\mathsf{C}_{\eta,t}}\left(a^{-1}I_{\phi^{-1},0}(f_i)\right)\,. \tag{45}$$

Finally, replacing d_t by this quantity gives

$$(f_{i+1} - f_i)\chi(f_i) \geq \eta a \|\nabla f(\theta_i)\|^2 \chi(f_i) \geq \frac{\eta a \|\nabla f(\theta_i)\|^2}{\phi(f_i)^2 \psi(d_i)^2} \geq \eta a \,, \tag{46}$$

and summing over all $i \in \{0, \dots, t-1\}$ and bounding this Riemann sum by its corresponding integral (as χ is non-decreasing) provides the desired result (noting that $a = 1 - \beta\eta/2 \in [1/2, 1]$). □

If f was L-Lipschitz and β-smooth on $\mathbb{R}^d$, then Proposition 4.6 with $\phi(x) = \varphi(\Delta - x)$ and $\psi(x) = \mathbb{1}\{x \geq R\}$ would prove Proposition 4.2. However, additional work is required due the fact that our assumptions on f only hold in $\mathcal{B}(\theta_0, R)$. Fortunately, as long as R is larger than the distance after which $\psi(x) = 0$, Proposition 4.6 still holds.

Lemma A.2. *Let $R = \min\{x \,:\, \psi(x) = 0\}$. If f is only $\tilde{L}$-Lipschitz and $\tilde{\beta}$-smooth on $\mathcal{B}(\theta_0, R)$, and $R_g \geq R + \mathsf{C}_{\eta,t}$, then Proposition 4.6 still hold by replacing $f(\theta_t)$ by $\min_{i \leq t} f(\theta_i)$.*

Proof. First, note that, if $I^\dagger_{\psi, \mathsf{C}_{\eta,t}} \circ a^{-1} I_{\phi^{-1}}(x) \geq R$, then $\chi(x) = 0$. As a consequence,

$$I^\dagger_\chi(a\eta t) \leq (I^\dagger_{\phi^{-1}} \circ a I_{\psi, \mathsf{C}_{\eta,t}})(R) \,. \tag{47}$$

If $\exists i \leq t$ such that $f(\theta_i) \leq f(\theta_0) - (I^\dagger_{\phi^{-1}} \circ a I_{\psi, \mathsf{C}_{\eta,t}})(R) + \mathsf{A}_{\eta,t}$, then Eq. (47) implies that

$$\min_{i \leq t} f(\theta_i) \leq f(\theta_0) - I^\dagger_\chi(a\eta t) + \mathsf{A}_{\eta,t} \,. \tag{48}$$

Otherwise, for all $i \leq t$, we have $f(\theta_i) > f(\theta_0) - (I^\dagger_{\phi^{-1}} \circ a I_{\psi, \mathsf{C}_{\eta,t}})(R) + \mathsf{A}_{\eta,t}$, and, by induction on $i \in \{0, \dots, t\}$, we have $\theta_i \in \mathcal{B}(\theta_0, R)$: First, $\theta_0 \in \mathcal{B}(\theta_0, R)$. Second, if, for all $j < i$, $\theta_j \in \mathcal{B}(\theta_0, R)$, then $\|\theta_i - \theta_0\| \leq \|\theta_{i-1} - \theta_0\| + \|\theta_i - \theta_{i-1}\| \leq R + \mathsf{C}_{\eta,t} \leq R_g$ by Lemma A.1, and thus Eq. (45) holds and $\theta_i \in \mathcal{B}(\theta_0, R)$. As a consequence, the function is β-smooth and L-Lipschitz along the iterates $(\theta_i)_{i \leq t}$, and the proof of Proposition 4.6 is applicable. □

We can now use Proposition 4.6 to prove Proposition 4.2 using $\phi(x) = \varphi(\Delta - x)$ and $\psi(x) = \mathbb{1}\{x \geq R\}$.

Proof of Proposition 4.2. First, note that $I_{\phi^{-1}}(x) = \int_0^x \varphi(\Delta - u)^{-1} du = -\int_\Delta^{\Delta - x} \varphi(u)^{-1} du = I_{\varphi^{-1}, \Delta}(\Delta - x)$, and, for $x < R$, $I^\dagger_{\psi, \mathsf{C}_{\eta,t}}(x) = x + \mathsf{C}_{\eta,t}$. Thus, for $-2I_{\varphi^{-1},\Delta}(\Delta - x) + \mathsf{C}_{\eta,t} \leq R$, we have

$$\chi(x) = \phi(x)^{-2} (\psi \circ I^\dagger_{\psi, \mathsf{C}_{\eta,t}} \circ 2 I_{\phi^{-1}})(x)^{-2} = \varphi(\Delta - x)^{-2} \,, \tag{49}$$

and $\chi(x) = +\infty$ otherwise. Hence, we have $I_\chi(x) = \int_0^x \varphi(\Delta - x)^{-2} du = -I_{\varphi^{-2},\Delta}(\Delta - x)$ if $x < \Delta - I^\dagger_{\varphi^{-1},\Delta}(-(R - \mathsf{C}_{\eta,t})/2)$, and $I_\chi(x) = +\infty$ otherwise, and

$$\begin{aligned} f(\theta_t) - f(\theta^\star) &\leq \Delta - I^\dagger_\chi(\eta t/2) + \mathsf{A}_{\eta,t} \\ &= \Delta - \min\{\Delta - I^\dagger_{\varphi^{-2},\Delta}(-\eta t/2), \Delta - I^\dagger_{\varphi^{-1},\Delta}(-(R - \mathsf{C}_{\eta,t})/2)\} + \mathsf{A}_{\eta,t} \\ &= \max\{I^\dagger_{\varphi^{-2},\Delta}(-\eta t/2), I^\dagger_{\varphi^{-1},\Delta}(-(R - \mathsf{C}_{\eta,t})/2)\} + \mathsf{A}_{\eta,t} \,, \end{aligned} \tag{50}$$

which gives the desired result using Lemma A.2. □

In order to prove Corollary 4.3, we will need the following Lemma.

Lemma A.3. *Let $A > 1$ and $t \geq \frac{A \ln(A)}{1 - 1/e}$. Then, we have $\frac{t}{\ln(t)} \geq A$.*

Proof. Let $g(t) = \ln(A) + \frac{t}{eA}$. We have $g(t) \geq \ln(t)$ by concavity of the logarithm and $t \geq \frac{A \ln(A)}{1 - 1/e}$ implies $t \geq A g(t) \geq A \ln(t)$. □

Proof of Corollary 4.3. The idea of the proof is to impose $\mathsf{A}_{\eta,t} \leq \varepsilon/2$ by chosing η, and then deriving the necessary conditions for t and R. As $\mathsf{A}_{\eta,t} = 4\beta\mathsf{C}_{\eta,t}^2 + 2d^{-1/2}L\mathsf{C}_{\eta,t}$ is a sum of two terms, we consider the two cases in which each term dominates, and choose η accordingly. We then take the minimum of these two terms and $1/\beta$, which gives

$$\eta_\varepsilon = \min\left\{\frac{1}{\beta}, \sqrt{\frac{\varepsilon}{32\sigma^2\beta t\log(6dt/\delta)}}, \frac{\varepsilon\sqrt{d}}{8L\sigma\sqrt{2t\log(6dt/\delta)}}\right\}. \tag{51}$$

This choice of η ensures that $\mathsf{A}_{\eta,t} \leq \varepsilon/2$ by imposing that $4\beta\mathsf{C}_{\eta,t}^2 \leq \varepsilon/4$ and $2d^{-1/2}L\mathsf{C}_{\eta,t} \leq \varepsilon/4$. Thus, imposing $I^\dagger_{\varphi^{-2},\Delta}(-\eta_\varepsilon t/2) \leq \varepsilon/2$ and $I^\dagger_{\varphi^{-1},\Delta}(-(R-2\mathsf{C}_{\eta,t})/2) \leq \varepsilon/2$ would immediately imply that $\min_{i<t} f(\theta_i) \leq \varepsilon$, and $\mathcal{T}_\varepsilon \leq t$. These two conditions give

$$t \geq -2I_{\varphi^{-2},\Delta}(\varepsilon/2)/\eta_\varepsilon \tag{52}$$

and

$$R \geq -2I_{\varphi^{-1},\Delta}(\varepsilon/2) + 2\mathsf{C}_{\eta,t} \tag{53}$$

First, as $4\beta\mathsf{C}_{\eta,t}^2 \leq \varepsilon/4$, we have $\mathsf{C}_{\eta,t} \leq \sqrt{\varepsilon/16\beta}$. Then, we consider each possible value of η_ε separately, and choose t accordingly:

1. If $\eta_\varepsilon = 1/\beta$, then we need $t \geq -2\beta I_{\varphi^{-2},\Delta}(\varepsilon/2)$.
2. If $\eta_\varepsilon = \sqrt{\frac{\varepsilon}{32\sigma^2\beta t\log(6dt/\delta)}}$, then we need

$$\sqrt{\frac{t}{\ln(6dt/\delta)}} \geq \sqrt{\frac{128\sigma^2\beta I_{\varphi^{-2},\Delta}(\varepsilon/2)^2}{\varepsilon}}, \tag{54}$$

   and using Lemma A.3, it is sufficient to take

$$t \geq h\left(\frac{128\sigma^2\beta I_{\varphi^{-2},\Delta}(\varepsilon/2)^2}{\varepsilon}\right), \tag{55}$$

   where $h(x) = x\log(6dx/\delta)/(1-1/e)$.
3. If $\eta_\varepsilon = \frac{\varepsilon\sqrt{d}}{8L\sigma\sqrt{2t\log(6dt/\delta)}}$, then we need

$$\sqrt{\frac{t}{\ln(6dt/\delta)}} \geq \frac{-16\sqrt{2}L\sigma I_{\varphi^{-2},\Delta}(\varepsilon/2)}{\varepsilon\sqrt{d}}, \tag{56}$$

   and using Lemma A.3, it is sufficient to take

$$t \geq h\left(\frac{512L^2\sigma^2 I_{\varphi^{-2},\Delta}(\varepsilon/2)^2}{d\varepsilon^2}\right), \tag{57}$$

   where $h(x) = x\log(6dx/\delta)/(1-1/e)$.

Finally, taking the maximum of these three values ensures that the condition is always satisfied, and thus leads to the desired result.

□

A.2. Proofs of the application to over-parameterized neural networks

Proof of Lemma 5.5. If φ^2 is convex, then we have

$$\begin{aligned}
\|\nabla\mathcal{L}(\theta)\|^2 &= \left\|\tfrac{1}{n}\textstyle\sum_i \partial_\theta g_\theta(x_i)^\top \nabla_x \ell(g_\theta(x_i), y_i)\right\|^2 \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \left\|\nabla_x \ell(g_\theta(x_i), y_i)\right\|^2 \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \varphi\big(\ell(g_\theta(x_i), y_i)\big)^2 \\
&\geq \tfrac{\mu}{n}\varphi\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(g_\theta(x_i), y_i)\big)^2 \\
&= \tfrac{\mu}{n}\varphi\big(\mathcal{L}(\theta)\big)^2,
\end{aligned} \tag{58}$$

where the first inequality is due to uniform conditioning, the second inequality due to the definition of φ, and the last from Jensen's inequality. Otherwise, we have

$$\begin{aligned}
\|\nabla\mathcal{L}(\theta)\|^2 &= \left\|\tfrac{1}{n}\textstyle\sum_i \partial_\theta g_\theta(x_i)^\top \nabla_x \ell(g_\theta(x_i), y_i)\right\|^2 \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \|\nabla_x \ell(g_\theta(x_i), y_i)\|^2 \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \varphi\big(\ell(g_\theta(x_i), y_i)\big)^2 \\
&\geq \tfrac{\mu}{n^2}\max_i \varphi\big(\ell(g_\theta(x_i), y_i)\big)^2 \\
&= \tfrac{\mu}{n^2}\varphi\big(\max_i \ell(g_\theta(x_i), y_i)\big)^2 \\
&\geq \tfrac{\mu}{n^2}\varphi\big(\mathcal{L}(\theta)\big)^2
\end{aligned} \tag{59}$$

,

□

Proof of Proposition 5.6. First, note that, as the model is L_g-Lipschitz and β_g-smooth on $\mathcal{B}(\theta_0, R)$, the objective function is also Lipschitz and smooth on this ball.

Lemma A.4. *The objective function $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(g_\theta(x_i), y_i)$ is $\tilde{L}$-Lipschitz and $\tilde{\beta}$-smooth on $\mathcal{B}(\theta_0, R)$, where $\tilde{L} = L_g L_\ell$ and $\tilde{\beta} = \beta_g L_\ell + \beta_\ell L_g^2$.*

Proof.

$$\begin{aligned}
\|\nabla\mathcal{L}(\theta)\| &= \|\tfrac{1}{n}\textstyle\sum_i \partial_\theta g_\theta(x_i)^\top \nabla_x \ell(g_\theta(x_i), y_i)\| \\
&\leq \tfrac{1}{n}\textstyle\sum_i \|\partial_\theta g_\theta(x_i)^\top \nabla_x \ell(g_\theta(x_i), y_i)\| \\
&\leq \tfrac{1}{n}\textstyle\sum_i L_g \|\nabla_x \ell(g_\theta(x_i), y_i)\| \\
&\leq L_g L_\ell
\end{aligned} \tag{60}$$

and, similarly,

$$\begin{aligned}
\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\| &= \|\tfrac{1}{n}\textstyle\sum_i (\partial_\theta g_\theta(x_i)^\top \nabla_x \ell(g_\theta(x_i), y_i) - \partial_\theta g_{\theta'}(x_i)^\top \nabla_x \ell(g_{\theta'}(x_i), y_i))\| \\
&\leq \tfrac{1}{n}\textstyle\sum_i \|(\partial_\theta g_\theta(x_i) - \partial_\theta g_{\theta'}(x_i))^\top \nabla_x \ell(g_\theta(x_i), y_i)\| \\
&\quad + \tfrac{1}{n}\textstyle\sum_i \|\partial_\theta g_{\theta'}(x_i)^\top (\nabla_x \ell(g_\theta(x_i), y_i) - \nabla_x \ell(g_{\theta'}(x_i), y_i))\| \\
&\leq \beta_g L_\ell \|\theta - \theta'\| + \beta_\ell L_g^2 \|\theta - \theta'\| .
\end{aligned} \tag{61}$$

□

Finally, we can apply Corollary 4.3 with $\tilde{\varphi}(x) = \sqrt{\frac{\mu}{n}}\varphi(x)$, which directly gives the desired result. □

Proof of Lemma 5.7. Using Eq. (23), we have

$$\begin{aligned}
\|\nabla\mathcal{L}(\theta)\|^2 &= \left\|\tfrac{1}{n}\textstyle\sum_i \partial_\theta g_\theta(x_i)^\top \nabla_x \ell(g_\theta(x_i), y_i)\right\|^2 \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \|\nabla_x \ell(g_\theta(x_i), y_i)\|^2 \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \frac{\ell(g_\theta(x_i), y_i)^2}{\|g_\theta(x_i) - g_{\theta^\star}(x_i)\|^2} \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \frac{\ell(g_\theta(x_i), y_i)^2}{L_g^2\|\theta - \theta^\star\|^2} \\
&\geq \tfrac{\mu}{n^2}\textstyle\sum_{i=1}^n \frac{\ell(g_\theta(x_i), y_i)^2}{L_g^2(\|\theta_0 - \theta^\star\| + \|\theta - \theta_0\|)^2} \\
&\geq \frac{\mu\mathcal{L}(\theta)^2}{nL_g^2(\|\theta_0 - \theta^\star\| + \|\theta - \theta_0\|)^2}
\end{aligned} \tag{62}$$

□

Proof of Proposition 5.8. Let $q \in [0, 1]$, $b = \kappa\|\theta_0 - \theta^\star\|$ and $c = \kappa$. Our approach is to apply Proposition 4.6 with $\phi(x) = (\Delta - x)_+$ and $\psi(x) = \frac{\mathbb{1}\{x \leq R\}}{b + cx}$. First, by Lemma A.4, the objective function is $\tilde{L}$-Lipschitz and $\tilde{\beta}$-smooth on $\mathcal{B}(\theta_0, R)$, where $\tilde{L} = L_g L_\ell$, $\tilde{\beta} = \beta_g L_\ell + \beta_\ell L_g^2$, and, using Lemma A.2, we can apply Proposition 4.6 if we replace R by $R - \mathsf{C}_{\eta,t}$ and $f(\theta_t)$ by $\min_{i \leq t} f(\theta_i)$. Using Lemma 5.7, we thus apply Proposition 4.6 to $\phi(x) = (\Delta - x)_+$ and

$\psi(x) = \frac{\mathbb{1}\{x \le R\}}{b+cx}$. We obtain that $I_{\phi^{-1}}(x) = -\log(1-x/\Delta)$ and $\psi \circ I^{\dagger}_{\psi,\mathsf{C}_{\eta,t}}(x) = (b+c\mathsf{C}_{\eta,t})^{-1}\exp(-cx)\mathbb{1}\{x \le R\}$. Hence, with $a = 1-\beta\eta/2$, we have, if $x < \Delta\left(1-\left(\frac{b+cR}{b+c\mathsf{C}_{\eta,t}}\right)^{-a/c}\right)$,

$$\begin{aligned}\chi(x) &= \phi(x)^{-2}(\psi \circ I^{\dagger}_{\psi,\mathsf{C}_{\eta,t}} \circ a^{-1} I_{\phi^{-1}}(x))^{-2} \\ &= (b+c\mathsf{C}_{\eta,t})^2(\Delta-x)_+^{-2}\exp(2cI_{\phi^{-1}}(x)/a) \\ &= (b+c\mathsf{C}_{\eta,t})^2(\Delta-x)_+^{-2}(1-x/\Delta)^{-2c/a} \\ &= \left(\tfrac{b+c\mathsf{C}_{\eta,t}}{\Delta}\right)^2 (1-x/\Delta)^{-2(1+c/a)}\,,\end{aligned} \tag{63}$$

and thus,

$$I^{\dagger}_{\chi}(x) = \Delta\left(1-\left[1+\frac{(1+2c/a)\Delta x}{(b+c\mathsf{C}_{\eta,t})^2}\right]^{-\frac{1}{1+2c/a}}\right). \tag{64}$$

As a consequence, we have that

$$\min_{i<t} f(\theta_i) \le \Delta \max\left\{\left(\frac{b+cR}{b+c\mathsf{C}_{\eta,t}}\right)^{-a/c}, \left[1+\frac{(1+2c/a)\Delta a\eta t}{(b+c\mathsf{C}_{\eta,t})^2}\right]^{-\frac{1}{1+2c/a}}\right\} + \mathsf{A}_{\eta,t}\,, \tag{65}$$

and thus, the conditions to reach a precision ε on the function value, with probability at least $1-\delta$, and before iteration t, are:

- $R \ge (b/c + \mathsf{C}_{\eta,t})\left(\frac{\Delta}{\varepsilon-\mathsf{A}_{\eta,t}}\right)^{c/a} + \mathsf{C}_{\eta,t}\,,$
- $\eta t \ge \frac{(b+c\mathsf{C}_{\eta,t})^2}{(1+2c/a)\Delta a}\left(\frac{\Delta}{\varepsilon-\mathsf{A}_{\eta,t}}\right)^{1+2c/a}.$

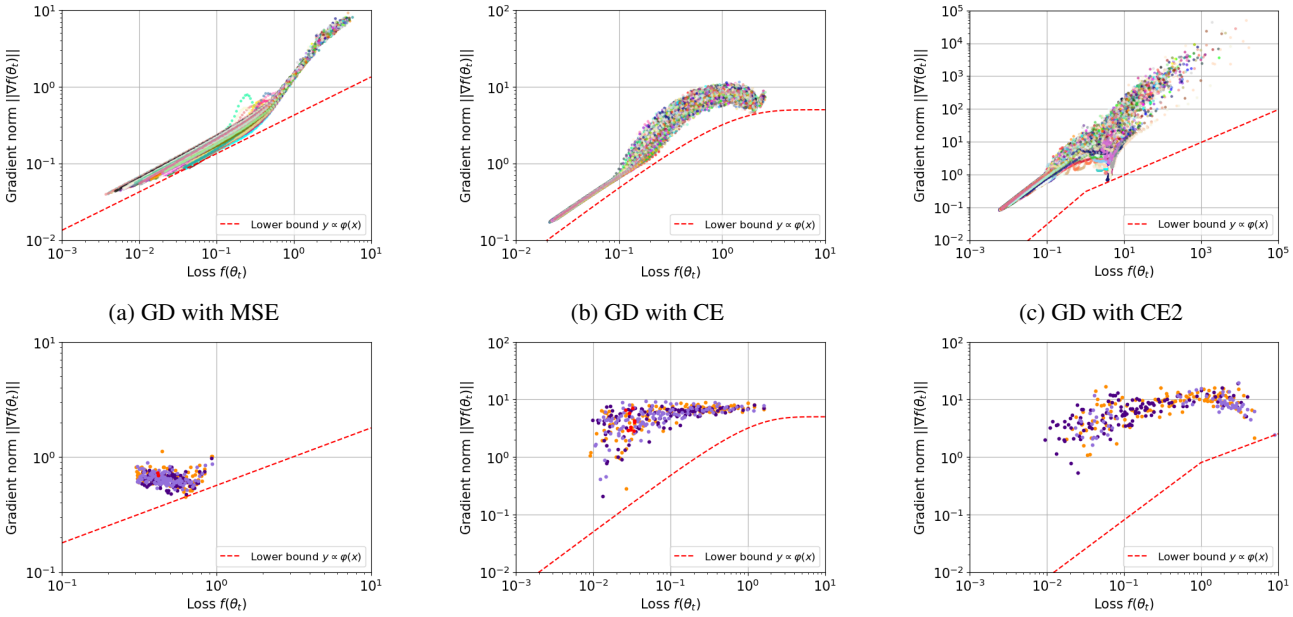
As in the proof of Corollary 4.3, we choose η in order to ensure that $\mathsf{A}_{\eta,t} = \varepsilon/2$. This gives $\eta_\varepsilon = O(\varepsilon/\sqrt{t\log(t)})$, which in turn implies that

- $R \ge \Omega\left(\varepsilon^{-c/a}\right)$,
- $\sqrt{t/\ln(t)} \ge O\left(\varepsilon^{-2-2c/a}\right)$,

and using Lemma A.3 and the fact that $a \to 1$ when $\eta \to 0$ gives the desired result. □

B. Numerical results for the deterministic (GD) and stochastic (SGD) setting on the CIFAR10 dataset

For the CIFAR10 dataset, Figure 2, we point out the same conclusions hold as for the MNIST dataset (see Section 6).



(a) GD with MSE (b) GD with CE (c) GD with CE2

(d) SGD with MSE (e) SGD with CE (f) SGD with CE2

Figure 2. Gradient norm vs. overall train loss for 1000 (GD) and 3 (SGD) runs (one color per run) with ResNet-18 on CIFAR10 using different loss functions.