Appl. Comput. Harmon. Anal. 59 (2022) 85–116

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Loss landscapes and optimization in over-parameterized non-linear systems and neural networks

Chaoyue Liu [a], Libin Zhu [b,c], Mikhail Belkin [c,*]

[a] *Department of Computer Science and Engineering, The Ohio State University, United States of America*
[b] *Department of Computer Science and Engineering, University of California, San Diego, United States of America*
[c] *Halıcıoğlu Data Science Institute, University of California, San Diego, United States of America*

ARTICLE INFO

Article history:
Received 9 June 2021
Received in revised form 24 December 2021
Accepted 26 December 2021
Available online 10 January 2022
Communicated by David Donoho

Keywords:
Deep learning
Non-linear optimization
Over-parameterized models
PL* condition

ABSTRACT

The success of deep learning is due, to a large extent, to the remarkable effectiveness of gradient-based optimization methods applied to large neural networks. The purpose of this work is to propose a modern view and a general mathematical framework for loss landscapes and efficient optimization in over-parameterized machine learning models and systems of non-linear equations, a setting that includes over-parameterized deep neural networks. Our starting observation is that optimization landscapes corresponding to such systems are generally not convex, even locally around a global minimum, a condition we call *essential non-convexity*. We argue that instead they satisfy PL*, a variant of the Polyak-Łojasiewicz condition [32,25] on most (but not all) of the parameter space, which guarantees both the existence of solutions and efficient optimization by (stochastic) gradient descent (SGD/GD). The PL* condition of these systems is closely related to the condition number of the tangent kernel associated to a non-linear system showing how a PL*-based non-linear theory parallels classical analyses of over-parameterized linear equations. We show that wide neural networks satisfy the PL* condition, which explains the (S)GD convergence to a global minimum. Finally we propose a relaxation of the PL* condition applicable to "almost" over-parameterized systems.

1. Introduction

A singular feature of modern machine learning is a large number of trainable model parameters. Just in the last few years we have seen state-of-the-art models grow from tens or hundreds of millions parameters to much larger systems with hundreds billion [6] or even trillions parameters [14]. Invariably these models are trained by gradient descent based methods, such as Stochastic Gradient Descent (SGD) or Adam [19]. Why are these local gradient methods so effective in optimizing complex highly non-convex systems? In the past few years an emerging understanding of gradient-based methods have started to focus on the insight

* Corresponding author.
E-mail address: mbelkin@ucsd.edu (M. Belkin).

https://doi.org/10.1016/j.acha.2021.12.009
1063-5203/

that optimization dynamics of "modern" over-parameterized models with more parameters than constraints are very different from those of "classical" models when the number of constraints exceeds the number of parameters.

The goal of this paper is to provide a modern view of the optimization landscapes, isolating key mathematical and conceptual elements that are essential for an optimization theory of over-parameterized models.

We start by characterizing a key difference between under-parameterized and over-parameterized landscapes. While both are generally non-convex, the nature of the non-convexity is rather different: under-parameterized landscapes are (generically) locally convex in a sufficiently small neighborhood of a local minimum. Thus classical analyses apply, if only locally, sufficiently close to a minimum. In contrast, over-parameterized systems are "essentially" non-convex systems, not even convex in arbitrarily small neighborhoods around global minima.

Thus, we cannot expect the extensive theory of convexity-based analyses to apply to such over-parameterized problems. In contrast, we argue that these systems typically satisfy the Polyak-Łojasiewicz condition [32,25], or more precisely, its slightly modified variant that we call the PL* condition, on most (but not all) of the parameter space. This condition ensures existence of solutions and convergence of GD and SGD, if it holds in a ball of sufficient radius. Importantly, we show that sufficiently wide neural networks satisfy the PL* condition around their initialization point, thus guaranteeing convergence. In addition, we show how the PL* condition can be relaxed without significantly changing the analysis. We conjecture that *in practice many large but not over-parameterized systems behave as if they were over-parameterized along the stretch of their optimization path from the initialization to the early stopping point.*

In a typical supervised learning task, given a training dataset of size n, $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d, y \in \mathbb{R}$, and a parametric family of models $f(\mathbf{w}; \mathbf{x})$, e.g., a neural network, one aims to find a model with parameter vector $\mathbf{w}^*$, that fits the training data, i.e.,

$$f(\mathbf{w}^*; \mathbf{x}_i) \approx y_i, \quad i = 1, 2, \ldots, n. \tag{1}$$

Mathematically, training such a model is equivalent to solving, exactly or approximately, a system of n equations.[1] Aggregating the predictions in the map $\mathcal{F}$ and the labels in the vector $\mathbf{y}$ (and suppressing the dependence on $\mathbf{x}_i$ in the notation) we write:

$$\mathcal{F}(\mathbf{w}) = \mathbf{y}, \quad \text{where } \mathbf{w} \in \mathbb{R}^m, \ \mathbf{y} \in \mathbb{R}^n, \quad \mathcal{F}(\cdot) : \mathbb{R}^m \to \mathbb{R}^n. \tag{2}$$

Here $(\mathcal{F}(\mathbf{w}))_i := f(\mathbf{w}; \mathbf{x}_i)$.

The system in Eq. (2) is solved through minimizing a certain loss function $\mathcal{L}(\mathbf{w})$, e.g., the square loss

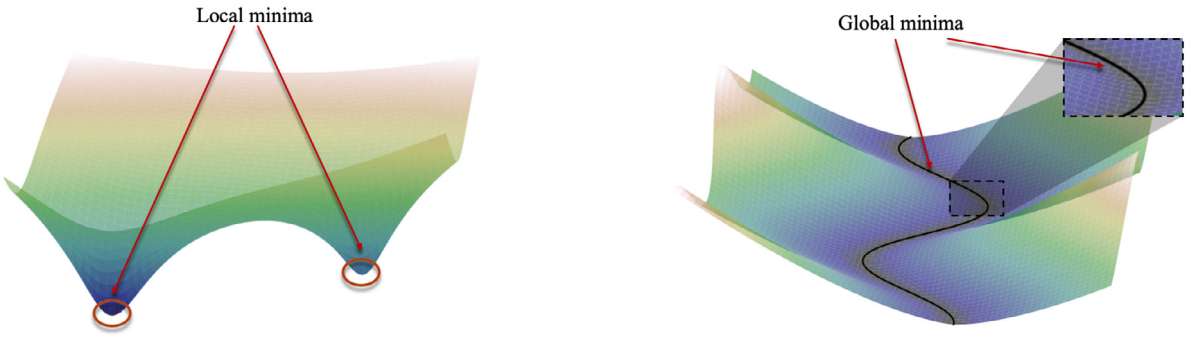
$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2 = \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{w}; \mathbf{x}_i) - y_i)^2$$

constructed so that the solutions of Eq. (2) are global minimizers of $\mathcal{L}(\mathbf{w})$. This is a non-linear least squares problem, which is well studied under classical under-parameterized settings (see [29], Chapter 10). An exact solution of Eq. (2) corresponds to interpolation, where a predictor fits the data exactly.

As we discuss below, for over-parameterized systems ($m > n$), we expect exact solutions to exist.

Essential non-convexity. Our starting point, discussed in detail in Section 3, is the observation that the loss landscape of an over-parameterized system is generally not convex in any neighborhood of any global minimizer. This is different from the case of under-parameterized systems, where the loss landscape is

[1] The same setup works for multiple outputs. For examples for multi-class classification problems both the prediction $f(\mathbf{w}; \mathbf{x}_i)$ and labels $\mathbf{y}_i$ (one-hot vector) are c-dimensional vectors, where c is the number of classes. In this case, we are in fact solving $n \times c$ equations. Similarly, for multi-output regression with c outputs, we have $n \times c$ equations.



(a) Loss landscape of under-parameterized models (b) Loss landscape of over-parameterized models

Fig. 1. Panel (a): Loss landscape is locally convex at local minima. Panel (b): Loss landscape incompatible with local convexity as the set of global minima is not locally linear.

globally not convex, but still typically locally convex, in a sufficiently small neighborhood of a (locally unique) minimizer. In contrast, the set of solutions of over-parameterized systems is generically a manifold of positive dimension [11] (and indeed, systems large enough have no non-global minima [22,28,37]). Unless the solution manifold is linear (which is not generally the case) the landscape cannot be locally convex. The contrast between over and under-parametrization is illustrated pictorially in Fig. 1.

The non-zero curvature of the curve of global minimizers at the bottom of the valley in Fig. 1(b) indicates the essential non-convexity of the landscape. In contrast, an under-parameterized landscape generally has multiple isolated local minima with positive-definite Hessian of the loss, where the function is locally convex. Thus we posit that

Convexity is not the right framework for analysis of over-parameterized systems, even locally.

Without assistance from local convexity, what alternative mathematical framework can be used to analyze loss landscapes and optimization behavior of non-linear over-parameterized systems?

In this paper we argue that such a framework is provided by the Polyak-Łojasiewicz condition [32,25], or, more specifically, by a variant that we call the PL* condition (independently introduced as local PL condition in [30]). We say that a non-negative function $\mathcal{L}$ satisfies μ-PL* on a set $\mathcal{S} \subset \mathbb{R}^m$ for $\mu > 0$, if

$$\|\nabla \mathcal{L}(\mathbf{w})\|^2 \geq \mu \mathcal{L}(\mathbf{w}), \quad \forall\ \mathbf{w} \in \mathcal{S}. \tag{3}$$

We will now outline some key reasons why the PL* condition provides a general framework for analyzing over-parameterized systems. In particular, we show that it is satisfied by the loss functions of sufficiently wide neural networks, although the corresponding loss functions are certainly non-convex.

PL* condition $\implies$ existence of solutions and exponential convergence of (S)GD. The first key point (see Section 5), is that if $\mathcal{L}$ satisfies the μ-PL* condition in a ball of radius $O(1/\mu)$ then $\mathcal{L}$ has a global minimum in that ball (corresponding to a solution of Eq. (2)). Furthermore, (S)GD initialized at the center of such a ball[2] converges exponentially to a global minimum of $\mathcal{L}$. Thus to establish both existence of solutions to Eq. (2) and efficient optimization, it is sufficient to verify the PL* condition in a ball of a certain radius.

Analytic form of PL* condition via the spectrum of the tangent kernel. Let $D\mathcal{F}(\mathbf{w})$ be the differential of the map $\mathcal{F}$ at $\mathbf{w}$, viewed as a $n \times m$ matrix. The tangent kernel of $\mathcal{F}$ is defined as an $n \times n$ matrix

$$K(\mathbf{w}) = D\mathcal{F}(\mathbf{w})\, D\mathcal{F}^T(\mathbf{w}). \tag{4}$$

[2] The constant in $O(1/\mu)$ is different for GD and SGD.

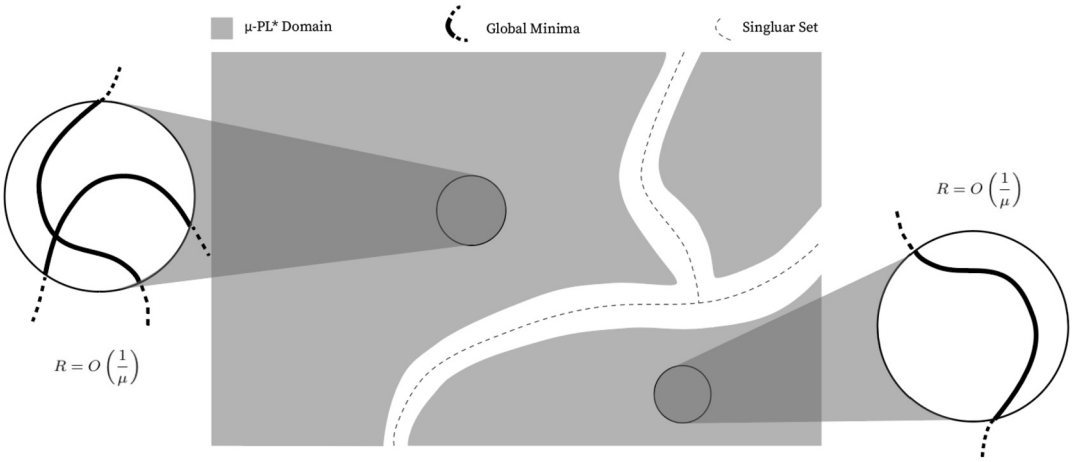


Fig. 2. The loss function $\mathcal{L}(\mathbf{w})$ is μ-PL* inside the shaded domain. Singular set corresponds to parameters $\mathbf{w}$ with degenerate tangent kernel $K(\mathbf{w})$. Every ball of radius $R = O(1/\mu)$ within the shaded set intersects with the set of global minima of $\mathcal{L}(\mathbf{w})$, i.e., solutions to $\mathcal{F}(\mathbf{w}) = \mathbf{y}$.

It is clear that $K(\mathbf{w})$ is a positive semi-definite matrix. It can be seen (Section 4) that the square loss $\mathcal{L}$ is μ-PL* at $\mathbf{w}$, where

$$\mu = \lambda_{\min}(K(\mathbf{w})), \tag{5}$$

is the smallest eigenvalue of the kernel matrix. Thus verification of the PL* condition reduces to analyzing the spectrum of the tangent kernel matrix associated to $\mathcal{F}$.

It is worthwhile to compare this to the standard analytical condition for convexity, requiring that the Hessian of the loss function, $H_{\mathcal{L}}$, is positive definite. While these spectral conditions appear to be similar, the similarity is superficial as $K(\mathbf{w})$ contains first order derivatives, while the Hessian is a second order operator. Hence, as we discuss below, the tangent kernel and the Hessian have very different properties under coordinate transformations and in other settings.

Why PL* holds across most of the parameter space for over-parameterized systems. We will now discuss the intuition why μ-PL* should be expected to hold for a sufficiently small μ over a domain containing most (but not all) of the parameter space $\mathbb{R}^m$. The intuition is based on simple parameter counting. For the simplest example consider the case $n = 1$. In that case the tangent kernel is a scalar and $K(\mathbf{w}) = \|D\mathcal{F}(\mathbf{w})\|^2$ is singular if and only if $\|D\mathcal{F}(\mathbf{w})\| = 0$. Thus, by parameter counting, we expect the singular set $\{\mathbf{w} : K(\mathbf{w}) = 0\}$ to have co-dimension m and thus, generically, consist of isolated points. Because of the Eq. (5) we generally expect most points sufficiently removed from the singular set to satisfy the PL* condition. By a similar parameter counting argument, the singular set of $\mathbf{w}$, such that $K(\mathbf{w})$ is not full rank will of co-dimension $m - n + 1$. As long as $m > n$, we expect the surroundings of the singular set, where the PL* condition does not hold, to be "small" compared to the totality of the parameter space. Furthermore, the larger the degree of the model over-parameterization $(m - n)$ is, the smaller the singular set is expected to be. This intuition is represented graphically in Fig. 2.

Note that under-parameterized systems are always rank deficient and have $\lambda_{\min}(K(\mathbf{w})) \equiv 0$. Hence such systems never satisfy PL*.

Wide neural networks satisfy PL*. We show that wide neural networks, as special cases of over-parameterized models, are PL*, using the technical tools developed in Section 4. This property of the neural networks is a key step to understand the success of the gradient descent based optimization, as seen in Section 5. To show the PL* condition for wide neural networks, a powerful, if somewhat crude, tool is provided by the remarkable recent discovery [17] that tangent kernels (so-called NTK) of wide neural networks with linear

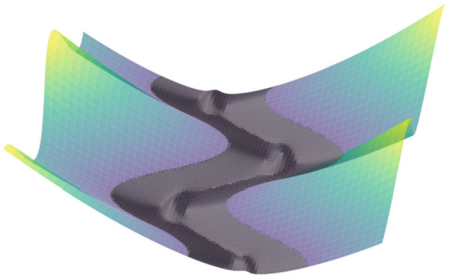


Fig. 3. The loss landscape of "almost over-parameterized" systems. The landscape looks over-parameterized except for the grey area where the loss is small. Local minima of the loss are contained there. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

output layer are nearly constant in a ball $\mathcal{B}$ of a certain radius around the ball with a random center. More precisely, it can be shown that the norm of the Hessian tensor $\|\mathbf{H}_{\mathcal{F}}(\mathbf{w})\| = O^*(1/\sqrt{m})$, where m is the width of the neural network and $\mathbf{w} \in \mathcal{B}$ [24].

Furthermore (see Section 4.1):

$$|\lambda_{\min}(K(\mathbf{w})) - \lambda_{\min}(K(\mathbf{w}_0))| < O\left(\sup_{\mathbf{w}\in\mathcal{B}} \|\mathbf{H}_{\mathcal{F}}(\mathbf{w})\|\right) = O(1/\sqrt{m}).$$

Bounding the kernel eigenvalue at the initialization point $\lambda_{\min}(K(\mathbf{w}_0))$ from below (using the results from [12,13]) completes the analysis.

To prove the result for general neural networks (note that the Hessian norm is typically large for such networks [24]) we observe that they can be decomposed as a composition of a network with a linear output layer and a coordinate-wise non-linear transformation of the output. The PL* condition is preserved under well-behaved non-linear transformations which yields the result.

Relaxing the PL* condition: when the systems are almost over-parameterized. Finally, a framework for understanding large systems cannot be complete without addressing the question of transition between under-parameterized and over-parameterized systems. While neural network models such as those in [6, 14] have billions or trillions of parameters, they are often trained on very large datasets and thus have comparable number of constraints. Thus in the process of optimization they may initially appear to be over-parameterized, while no truly interpolating solution may exist. Such an optimization landscape is shown in Fig. 3. While an in-depth analysis of this question is beyond the scope of this work, we propose a version of the PL* condition that can account for such behavior by postulating that the PL* condition is satisfied for values of $\mathcal{L}(\mathbf{w})$ above a certain threshold and that the optimization path from the initialization to the early stopping point lies in the PL* domain. Approximate convergence of (S)GD can be shown under this condition, see Section 6.

1.1. Related work

Loss landscapes of over-parameterized machine learning models, especially deep neural networks, have recently been studied in a number of papers including [11,22,28,37,20,26]. The work [11] suggests that the solutions of an over-parameterized system typically are not discrete and form a lower dimensional manifold in the parameter space. In particular [22,28] show that sufficiently over-parameterized neural networks have no strict local minima under certain mild assumptions. Furthermore in [20] it is proved that, for sufficiently over-parameterized neural networks, each local minimum (if it exists) is "path connected" to a global minimum where the loss value on the path is non-increasing. While the above works mostly focus on the properties of the minima of the loss landscape, in this paper we on the landscape within neighborhoods of these minima, which can often cover the whole optimization path of gradient-based methods.

There has also been a rich literature that particularly focuses on the optimization of wide neural networks using gradient descent [33,12,13,21,3,2,10,18,30], and SGD [1,38], especially after the discovery of the constancy of neural tangent kernels (NTK) for certain neural networks [17]. Many of these analyses are based on the observation that the constancy of the tangent kernel implies that the training dynamic of these networks is approximately that of a linear model [21]. This type of analysis can be viewed within our framework of Hessian norm control.

The Polyak-Łojasiewicz (PL) condition has attracted interest in connection with convergence of neural networks and other non-linear systems as it allows to prove some key properties of convexity with respect to the optimization in non-convex settings. For example, the work [8] proved that a composition of strongly convex function with a one-layer leaky ReLU network satisfies the PL condition. The paper [4] proved fast convergence of stochastic gradient descent with constant step size for over-parameterized models, while [30] refined this result showing SGD convergence within a ball with a certain radius. In related work [15] shows that the optimization path length can be upper bounded.

Finally, it is interesting to point out the connections with the contraction theory for differential equations explored in [36].

2. Notation and standard definitions

We use bold lowercase letters, e.g., $\mathbf{v}$, to denote vectors; capital letters, e.g., W, to denote matrices; and bold capital letters, e.g., $\mathbf{W}$, to denote tuples of matrices or higher order tensors. We denote the set $\{1, 2, \cdots, n\}$ as $[n]$. Given a function $\sigma(\cdot)$, we use $\sigma'(\cdot)$ and $\sigma''(\cdot)$ to denote its first and second respectively. For vectors, we use $\|\cdot\|$ to denote the Euclidean norm and $\|\cdot\|_\infty$ for the ∞-norm. For matrices, we use $\|\cdot\|_F$ to denote the Frobenius norm and $\|\cdot\|_2$ to denote the spectral norm (i.e., 2-norm). We use $D\mathcal{F}$ to represent the differential map of $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^n$. Note that $D\mathcal{F}$ is represented as a $n \times m$ matrix, with $(D\mathcal{F})_{ij} := \frac{\partial \mathcal{F}_i}{\partial w_j}$. We denote Hessian of the function $\mathcal{F}$ as $\mathbf{H}_\mathcal{F}$, which is an $n \times m \times m$ tensor with $(\mathbf{H}_\mathcal{F})_{ijk} = \frac{\partial^2 \mathcal{F}_i}{\partial w_j \partial w_k}$, and define the norm of the Hessian tensor to be the maximum of the spectral norms of each of its Hessian components, i.e., $\|\mathbf{H}_\mathcal{F}\| = \max_{i\in[n]} \|H_{\mathcal{F}_i}\|_2$, where $H_{\mathcal{F}_i} = \partial^2 \mathcal{F}_i / \partial \mathbf{w}^2$. When necessary, we also assume that $\mathbf{H}_\mathcal{F}$ is continuous. We also denote the Hessian matrix of the loss function as $H_\mathcal{L} := \partial^2 \mathcal{L} / \partial \mathbf{w}^2$, which is an $m \times m$ matrix. For a symmetric matrix A, $\lambda_{\min}(A)$ denotes the smallest eigenvalue of A.

In this paper, we consider the problem of solving a system of equations of the form Eq. (2), i.e., finding $\mathbf{w}$, such that $\mathcal{F}(\mathbf{w}) = \mathbf{y}$. This problem is solved by minimizing a loss function $\mathcal{L}(\mathcal{F}(\mathbf{w}), \mathbf{y})$, such as the square loss $\mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2$, with gradient-based algorithms. Specifically, the gradient descent method starts from the initialization point $\mathbf{w}_0$ and updates the parameters as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t), \quad \forall t \in \mathbb{N}. \tag{6}$$

We call the set $\{\mathbf{w}_t\}_{t=0}^\infty$ the optimization path, and put $\mathbf{w}_\infty = \lim_{t\to\infty} \mathbf{w}_t$ (assuming the limit exists).

Throughout this paper, we assume the map $\mathcal{F}$ is Lipschitz continuous and smooth as defined below.

Definition 1 *(Lipschitz continuity).* We say the map $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^n$ is $L_\mathcal{F}$-Lipschitz continuous, if

$$\|\mathcal{F}(\mathbf{w}) - \mathcal{F}(\mathbf{v})\| \le L_\mathcal{F} \|\mathbf{w} - \mathbf{v}\|, \quad \forall \mathbf{w}, \mathbf{v} \in \mathbb{R}^m. \tag{7}$$

Remark 1. In supervised learning cases, $(\mathcal{F}(\mathbf{w}))_i = f(\mathbf{w}; \mathbf{x}_i)$. If we denote L_f as the Lipschitz constant of the machine learning model $f(\mathbf{w}; \mathbf{x})$ w.r.t. the parameters $\mathbf{w}$, then $\|\mathcal{F}(\mathbf{w}) - \mathcal{F}(\mathbf{v})\|^2 = \sum_i (f(\mathbf{w}; \mathbf{x}_i) - f(\mathbf{v}; \mathbf{x}_i))^2 \le n L_f^2 \|\mathbf{w} - \mathbf{v}\|^2$. One has $L_\mathcal{F} \le \sqrt{n} L_f$.

A direct consequence of the $L_\mathcal{F}$-Lipschitz condition is that $\|D\mathcal{F}(\mathbf{w})\|_2 \le L_\mathcal{F}$ for all $\mathbf{w} \in \mathbb{R}^m$. Furthermore, it is easy to see that the tangent kernel has bounded spectral norm:

Proposition 1. *If the map $\mathcal{F}$ is $L_{\mathcal{F}}$-Lipschitz, then its tangent kernel K satisfies $\|K(\mathbf{w})\|_2 \leq L_{\mathcal{F}}^2$.*

Definition 2 *(Smoothness).* We say the map $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^n$ is $\beta_{\mathcal{F}}$-smooth, if

$$\|\mathcal{F}(\mathbf{w}) - \mathcal{F}(\mathbf{v}) - D\mathcal{F}(\mathbf{v})(\mathbf{w} - \mathbf{v})\|_2 \leq \frac{\beta_{\mathcal{F}}}{2}\|\mathbf{w} - \mathbf{v}\|^2, \quad \forall \mathbf{w}, \mathbf{v} \in \mathbb{R}^m. \tag{8}$$

Jacobian matrix. Given a map $\Phi : \mathbf{z} \in \mathbb{R}^n \mapsto \Phi(\mathbf{z}) \in \mathbb{R}^n$, the corresponding Jacobian matrix J_Φ is defined as following: each entry

$$(J_\Phi)_{ij} := \frac{\partial \Phi_i}{\partial z_j}. \tag{9}$$

3. Essential non-convexity of loss landscapes of over-parameterized non-linear systems

In this section we discuss the observation that over-parameterized systems give rise to landscapes that are *essentially* non-convex – there typically exists no neighborhood around any global minimizer where the loss landscape is convex. This is in contrast to under-parameterized systems where such a neighborhood typically exists, although it can be small.

For over-parameterized system of equations, the number of parameters m is larger than the number of constraints n. In this case, the system of equations generically has exact solutions, which form a *continuous* manifold (or several *continuous* manifolds), typically of dimension $m - n > 0$ so that none of the solutions are isolated. A specific result in this direction is given in Appendix A. It can be informally stated as the following

Proposition 2 *(Informal). A feedforward neural network of sufficient width does not have isolated global minima. That is, given a global minimum $\mathbf{w}^*$, there are other global minima in any neighborhood of $\mathbf{w}^*$.*

It is important to note that such continuous manifolds of solutions generically have non-zero curvature,[3] due to the non-linear nature of the system of equations. This results in the lack of local convexity of the loss landscape, i.e., the loss landscape is nonconvex in any neighborhood of a solution (i.e., a global minimizer of $\mathcal{L}$). This can be seen from the following argument. Suppose that the loss function landscape of $\mathcal{L}$ is convex within a ball $\mathcal{B}$ that intersects the set of global minimizers $\mathcal{M}$. The minimizers of a convex function within a convex domain form a convex set, thus $\mathcal{M} \cap \mathcal{S}$ must also be convex. Hence $\mathcal{M} \cap \mathcal{S}$ must be a convex subset within a lower dimensional linear subspace of $\mathbb{R}^m$ and cannot have curvature (either extrinsic ot intrinsic). This geometric intuition is illustrated in Fig. 1b, where the set of global minimizers is of dimension one.

To provide an alternative analytical intuition (leading to a precise argument) consider an over-parameterized system $\mathcal{F}(\mathbf{w}) : \mathbb{R}^m \to \mathbb{R}^n$, where $m > n$, with the square loss function

$$\mathcal{L}(\mathbf{w}) := \mathcal{L}(\mathcal{F}(\mathbf{w}), \mathbf{y}) = \frac{1}{2}\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2. \tag{10}$$

The Hessian matrix of the loss function takes the form

$$H_{\mathcal{L}}(\mathbf{w}) = \underbrace{D\mathcal{F}(\mathbf{w})^T \frac{\partial^2 \mathcal{L}}{\partial \mathcal{F}^2}(\mathbf{w}) D\mathcal{F}(\mathbf{w})}_{A(\mathbf{w})} + \underbrace{\sum_{i=1}^{n} \left(\mathcal{F}(\mathbf{w}) - \mathbf{y}\right)_i H_{\mathcal{F}_i}(\mathbf{w})}_{B(\mathbf{w})}, \tag{11}$$

[3] Both extrinsic, the second fundamental form, and intrinsic Gaussian curvature.

where $H_{\mathcal{F}_i}(\mathbf{w}) \in \mathbb{R}^{m\times m}$ is the Hessian matrix of ith output of $\mathcal{F}$ with respect to $\mathbf{w}$.

Let $\mathbf{w}^*$ be a solution to Eq. (2). Since $\mathbf{w}^*$ is a global minimizer of $\mathcal{L}$, $B(\mathbf{w}^*) = 0$. We note that $A(\mathbf{w}^*)$ is a positive semi-definite matrix of rank at most n with at least $m - n$ zero eigenvalues.

Yet, in a neighborhood of $\mathbf{w}^*$ there typically are points where $B(\mathbf{w})$ has rank m. As we show below, this observation, together with a mild technical assumption on the loss, implies that $H_{\mathcal{L}}(\mathbf{w})$ cannot be positive semi-definite in any ball around $\mathbf{w}^*$ and hence the loss landscape is not locally convex. To see why this is the case, consider an example of a system with only one equation ($n = 1$). The loss function takes the form as $\mathcal{L}(\mathbf{w}) = \frac{1}{2}(\mathcal{F}(\mathbf{w}) - y)^2, y \in \mathbb{R}$ and the Hessian of the loss function can be written as

$$H_{\mathcal{L}}(\mathbf{w}) = D\mathcal{F}(\mathbf{w})^T D\mathcal{F}(\mathbf{w}) + (\mathcal{F}(\mathbf{w}) - y)H_{\mathcal{F}}(\mathbf{w}). \tag{12}$$

Let $\mathbf{w}^*$ be a solution to $\mathcal{F}(\mathbf{w}^*) = y$ and suppose that $D\mathcal{F}(\mathbf{w}^*)$ does not vanish. In the neighborhood of $\mathbf{w}^*$, there exist arbitrarily close points $\mathbf{w}^* + \boldsymbol{\delta}$ and $\mathbf{w}^* - \boldsymbol{\delta}$, such that $\mathcal{F}(\mathbf{w}^* + \boldsymbol{\delta}) - y > 0$ and $\mathcal{F}(\mathbf{w}^* - \boldsymbol{\delta}) - y < 0$. Assuming that the rank of $H_{\mathcal{F}}(\mathbf{w}^*)$ is greater than one, and noting that the rank of $D\mathcal{F}(\mathbf{w})^T D\mathcal{F}(\mathbf{w})$ is at most one in this example, it is easy to see that either $H_{\mathcal{L}}(\mathbf{w}^* + \boldsymbol{\delta})$ or $H_{\mathcal{L}}(\mathbf{w}^* - \boldsymbol{\delta})$ must have negative eigenvalues, which rules out local convexity at $\mathbf{w}^*$.

A more general version of this argument is given in the following:

Proposition 3 *(Local non-convexity). Consider an over-parameterized system $\mathcal{F} : \mathbb{R}^m \to \mathbb{R}^n$, and the square loss function Eq. (10). Let $\mathbf{w}^*$ be a global minimizer of the loss function, i.e., $\mathcal{L}(\mathbf{w}^*) = 0$. Suppose that $D\mathcal{F}(\mathbf{w}^*) \neq \mathbf{0}$ and, for at least one $i \in [n]$, $\operatorname{rank}(H_{\mathcal{F}_i}(\mathbf{w}^*)) > 2n$. Then $\mathcal{L}(\mathbf{w})$ is not convex in any neighborhood of $\mathbf{w}^*$.*

Remark 2. Note that in general we do not expect $D\mathcal{F}$ to vanish at $\mathbf{w}^*$ as there is no reason why a solution to Eq. (2) should be a critical point of $\mathcal{F}$.

Remark 3. For a general (non-square) loss function $\mathcal{L}(\mathbf{w})$, the assumption in Proposition 3 that $D\mathcal{F}(\mathbf{w}^*) \neq 0$ is replaced by the condition $\left(\frac{d}{d\mathbf{w}}\frac{\partial \mathcal{L}}{\partial \mathcal{F}}\right)(\mathbf{w}^*) \neq 0$.

A full proof of Proposition 3 for a general loss function can be found in Appendix B.

Comparison to under-parameterized systems. For under-parameterized systems, local minima are generally isolated, as illustrated in Fig. 1a. Since $H_{\mathcal{L}}(\mathbf{w}^*)$ is generically positive definite when $\mathbf{w}^*$ is an isolated local minimizer, by the continuity of $H_{\mathcal{L}}(\cdot)$, positive definiteness holds in the neighborhood of $\mathbf{w}^*$. Therefore, $\mathcal{L}(\mathbf{w})$ is locally convex around $\mathbf{w}^*$.

4. Over-parameterized non-linear systems are PL* on most of the parameter space

In this section we argue that loss landscapes of over-parameterized systems satisfy the PL* condition across most of their parameter space. We begin by defining *uniform conditioning* of a system, an analytical condition closely related to PL*.

Definition 3 *(Uniform conditioning).* We say that $\mathcal{F}(\mathbf{w})$ is μ-uniformly conditioned ($\mu > 0$) on $\mathcal{S} \subset \mathbb{R}^m$ if the smallest eigenvalue of its tangent kernel $K(\mathbf{w})$ (defined in Eq. (4)) satisfies

$$\lambda_{\min}(K(\mathbf{w})) \geq \mu, \ \forall \mathbf{w} \in \mathcal{S}. \tag{13}$$

The following important connection shows that uniform conditioning of the system is sufficient for the corresponding square loss landscape to satisfy the PL* condition.

Theorem 1 *(Uniform conditioning ⇒ PL* condition). If $\mathcal{F}(\mathbf{w})$ is μ-uniformly conditioned, on a set $\mathcal{S} \subset \mathbb{R}^m$, then the square loss function $\mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2$ satisfies μ-PL* condition on $\mathcal{S}$.*

Proof.

$$\frac{1}{2}\|\nabla \mathcal{L}(\mathbf{w})\|^2$$
$$= \frac{1}{2}(\mathcal{F}(\mathbf{w}) - \mathbf{y})^T K(\mathbf{w})(\mathcal{F}(\mathbf{w}) - \mathbf{y}) \geq \frac{1}{2}\lambda_{\min}(K(\mathbf{w}))\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2 = \lambda_{\min}(K(\mathbf{w}))\mathcal{L}(\mathbf{w}) \geq \mu\mathcal{L}(\mathbf{w}). \quad \square$$

We will now provide some intuition for why we expect $\lambda_{\min}(K(\mathbf{w}))$ to be separated from zero over most but not all of the optimization landscape. The key observation is that

$$\operatorname{rank} K(\mathbf{w}) = \operatorname{rank}\left(D\mathcal{F}(\mathbf{w})\, D\mathcal{F}^T(\mathbf{w})\right) = \operatorname{rank} D\mathcal{F}(\mathbf{w})$$

Note that kernel $K(\mathbf{w})$ is an $n \times n$ positive semi-definite matrix by definition. Hence the *singular set* $\mathcal{S}_{sing}$, where the tangent kernel is degenerate, can be written as

$$\mathcal{S}_{sing} = \{\mathbf{w} \in \mathbb{R}^m \,|\, \lambda_{\min}(K(\mathbf{w})) = 0\} = \{\mathbf{w} \in \mathbb{R}^m \,|\, \operatorname{rank} D\mathcal{F}(\mathbf{w}) < n\}.$$

Generically, $\operatorname{rank} D\mathcal{F}(\mathbf{w}) = \min(m, n)$. Thus for over-parameterized systems, when $m > n$, we expect $\mathcal{S}_{sing}$ to have positive codimension and to be a set of measure zero. In contrast, in under-parametrized settings, $m < n$ and the tangent kernel is always degenerate, $\lambda_{\min}(K(\mathbf{w})) \equiv 0$ so such systems *cannot* be uniformly conditioned according to the definition above. Furthermore, while Eq. (13) provides a sufficient condition, it's also in a certain sense necessary:

Proposition 4. *If $\lambda_{\min}(K(\mathbf{w}_0)) = 0$ then the system $\mathcal{F}(\mathbf{w}) = \mathbf{y}$ cannot be PL* for all $\mathbf{y}$ on any set $\mathcal{S}$ containing $\mathbf{w}_0$.*

Proof. Since $\lambda_{\min}(K(\mathbf{w}_0)) = 0$, the matrix $K(\mathbf{w}_0)$ has a non-trivial null-space. Therefore we can choose $\mathbf{y}$ so that $K(\mathbf{w}_0)(\mathcal{F}(\mathbf{w}_0) - \mathbf{y}) = 0$ and $\mathcal{F}(\mathbf{w}_0) - \mathbf{y} \neq 0$. We have

$$\frac{1}{2}\|\nabla \mathcal{L}(\mathbf{w}_0)\|^2 = \frac{1}{2}(\mathcal{F}(\mathbf{w}_0) - \mathbf{y})^T K(\mathbf{w}_0)(\mathcal{F}(\mathbf{w}_0) - \mathbf{y}) = 0$$

and hence the PL* condition cannot be satisfied at $\mathbf{w}_0$. $\square$

Thus we see that only over-parameterized systems can be PL*, if we want that condition to hold for any label vector $\mathbf{y}$.

By parameter counting, it is easy to see that the codimension of the singular set $\mathcal{S}_{sing}$ is expected to be $m - n + 1$. Thus, on a compact set, for μ sufficiently small, points which are not μ-PL* will be found around $\mathcal{S}_{sing}$, a low-dimensional subset of $\mathbb{R}^m$. This is represented graphically in Fig. 4. Note that the more over-parameterization we have, the larger the codimension of $\mathcal{S}_{sing}$ is expected to be.

To see a particularly simple example of this phenomenon, consider the case when $D\mathcal{F}(\mathbf{w})$ is a random matrix with Gaussian entries.[4] Denote by

$$\kappa = \frac{\lambda_{\max}(K(\mathbf{w}))}{\lambda_{\min}(K(\mathbf{w}))}$$

[4] In this the matrix "$D\mathcal{F}(\mathbf{w})$" does not have to correspond to an actual map $\mathcal{F}$, rather the intention is to illustrate how over-parameterization leads to better conditioning.

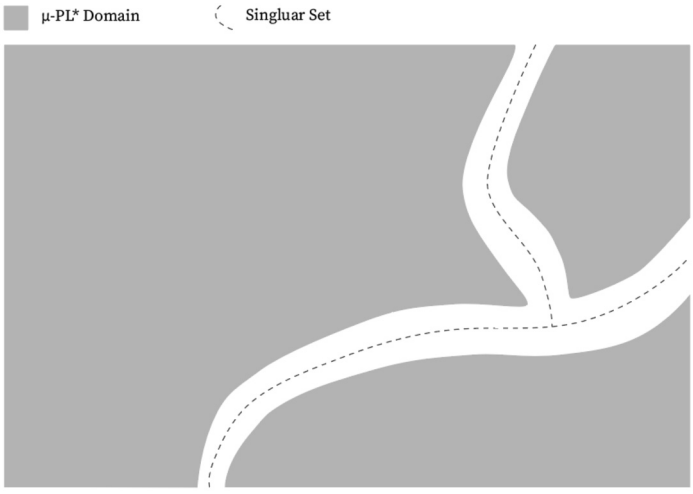


Fig. 4. μ-PL* domain and the singular set. We expect points away from the singular set to satisfy μ-PL* condition for sufficiently small μ.

the condition number of $K(\mathbf{w})$. Note that by definition $\kappa \geq 1$. It is shown in [9] that (assuming $m > n$)

$$\mathbb{E}(\log \kappa) < 2 \log \frac{m}{m-n+1} + 5.$$

We see that as the amount of over-parameterization increases κ converges in expectation (and also can be shown with high probability) to a small constant. While this example is rather special, it is representative of the concept that over-parameterization helps with conditioning. In particular, as we discuss below in Section 4.2, wide neural networks satisfy the PL* with high probability within a random ball of a constant radius.

We will now provide techniques for proving that the PL* condition holds for specific systems and later in Section 4.2 we will show how these techniques apply to wide neural networks, demonstrating that they are PL*. In Section 5 we discuss the implications of the PL* condition for the existence of solutions and convergence of (S)GD, in particular for deep neural networks.

4.1. Techniques for establishing the PL* condition

While we expect typical over-parameterized systems to satisfy the PL* condition at most points, directly analyzing the smallest eigenvalue of the corresponding kernel matrix is often difficult. Below we describe two methods for demonstrating the PL* condition of a system holds in a ball of a certain radius. First method is based on the observation that is well-conditioned in a ball provided it is well-conditioned at its center and the Hessian norm is not large compared to the radius of the ball. Interestingly, this "Hessian control" condition holds for a broad class of non-linear systems. In particular, as discussed in Sec 4.2, wide neural networks with linear output layer have small Hessian norm. In Appendix C, we use a simple family of neural models to illustrate the intuition behind the small Hessian norm of neural networks.

The second approach to demonstrating conditioning is by noticing that it is preserved under well-behaved transformations of the input or output or when composing certain models.

Combining these methods yields a general result on PL* condition for wide neural networks in Section 4.2.

Uniform conditioning through Hessian spectral norm. We will now show that controlling the norm of the Hessian tensor for the map $\mathcal{F}$ leads to well-conditioned systems. The idea of the analysis is that the change of the tangent kernel of $\mathcal{F}(\mathbf{w})$ can be bounded in terms of the norm of the Hessian of $\mathcal{F}(\mathbf{w})$. Intuitively, this is analogous to the mean value theorem, bounding the first derivative of $\mathcal{F}$ by its second derivative. If

the Hessian norm is sufficiently small, the change of the tangent kernel and hence its conditioning can be controlled in a ball $B(\mathbf{w}_0, R)$ with a finite radius, as long as the tangent kernel matrix at the center point $K(\mathbf{w}_0)$ is well-conditioned.

Theorem 2 *(PL* condition via Hessian norm). Given a point $\mathbf{w}_0 \in \mathbb{R}^m$, suppose the tangent kernel matrix $K(\mathbf{w}_0)$ is strictly positive definite, i.e., $\lambda_0 := \lambda_{\min}(K(\mathbf{w}_0)) > 0$. If the Hessian spectral norm $\|\mathbf{H}_{\mathcal{F}}\| \leq \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}\sqrt{nR}}$ holds within the ball $B(\mathbf{w}_0, R)$ for some $R > 0$ and $\mu > 0$, then the tangent kernel $K(\mathbf{w})$ is μ-uniformly conditioned in the ball $B(\mathbf{w}_0, R)$. Hence, the square loss $\mathcal{L}(\mathbf{w})$ satisfies the μ-PL* condition in $B(\mathbf{w}_0, R)$.*

Proof. First, let's consider the difference between the tangent kernel matrices at $\mathbf{w} \in B(\mathbf{w}_0, R)$ and at $\mathbf{w}_0$.

By the assumption, we have, for all $\mathbf{v} \in B(\mathbf{w}_0, R)$, $\|\mathbf{H}_{\mathcal{F}}(\mathbf{v})\| < \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}\sqrt{nR}}$. Hence, for each $i \in [n]$, $\|H_{\mathcal{F}_i}(\mathbf{w})\|_2 < \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}\sqrt{nR}}$. Now, consider an arbitrary point $\mathbf{w} \in B(\mathbf{w}_0, R)$. For all $i \in [n]$, we have:

$$D\mathcal{F}_i(\mathbf{w}) = D\mathcal{F}_i(\mathbf{w}_0) + \int_0^1 H_{\mathcal{F}_i}(\mathbf{w}_0 + \tau(\mathbf{w} - \mathbf{w}_0))(\mathbf{w} - \mathbf{w}_0) d\tau. \tag{14}$$

Since τ is in $[0, 1]$, $\mathbf{w}_0 + \tau(\mathbf{w} - \mathbf{w}_0)$ is on the line segment $S(\mathbf{w}_0, \mathbf{w})$ between $\mathbf{w}_0$ and $\mathbf{w}$, which is inside of the ball $B(\mathbf{w}_0, R)$. Hence,

$$\|D\mathcal{F}_i(\mathbf{w}) - D\mathcal{F}_i(\mathbf{w}_0)\| \leq \sup_{\tau \in [0,1]} \|H_{\mathcal{F}_i}(\mathbf{w}_0 + \tau(\mathbf{w} - \mathbf{w}_0))\|_2 \cdot \|\mathbf{w} - \mathbf{w}_0\| \leq \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}\sqrt{nR}} \cdot R = \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}\sqrt{n}}.$$

In the second inequality above, we used the fact that $\|H_{\mathcal{F}_i}\|_2 < \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}\sqrt{nR}}$ in the ball $B(\mathbf{w}_0, R)$. Hence,

$$\|D\mathcal{F}(\mathbf{w}) - D\mathcal{F}(\mathbf{w}_0)\|_F = \sqrt{\sum_{i \in [n]} \|D\mathcal{F}_i(\mathbf{w}) - D\mathcal{F}_i(\mathbf{w}_0)\|^2} \leq \sqrt{n} \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}\sqrt{n}} = \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}}.$$

Then, the spectral norm of the tangent kernel change is bounded by

$$\begin{aligned}
\|K(\mathbf{w}) - K(\mathbf{w}_0)\|_2 &= \|D\mathcal{F}(\mathbf{w})D\mathcal{F}(\mathbf{w})^T - D\mathcal{F}(\mathbf{w}_0)D\mathcal{F}(\mathbf{w}_0)^T\|_2 \\
&= \|D\mathcal{F}(\mathbf{w})\left(D\mathcal{F}(\mathbf{w}) - D\mathcal{F}(\mathbf{w}_0)\right)^T + \left(D\mathcal{F}(\mathbf{w}) - D\mathcal{F}(\mathbf{w}_0)\right)D\mathcal{F}(\mathbf{w}_0)^T\|_2 \\
&\leq \|D\mathcal{F}(\mathbf{w})\|_2 \|D\mathcal{F}(\mathbf{w}) - D\mathcal{F}(\mathbf{w}_0)\|_2 + \|D\mathcal{F}(\mathbf{w}) - D\mathcal{F}(\mathbf{w}_0)\|_2 \|D\mathcal{F}(\mathbf{w}_0)\|_2 \\
&\leq L_{\mathcal{F}} \cdot \|D\mathcal{F}(\mathbf{w}) - D\mathcal{F}(\mathbf{w}_0)\|_F + \|D\mathcal{F}(\mathbf{w}) - D\mathcal{F}(\mathbf{w}_0)\|_F \cdot L_{\mathcal{F}} \\
&\leq 2L_{\mathcal{F}} \cdot \frac{\lambda_0 - \mu}{2L_{\mathcal{F}}} \\
&= \lambda_0 - \mu.
\end{aligned}$$

In the second inequality above, we used the $L_{\mathcal{F}}$-Lipschitz continuity of $\mathcal{F}$ and the fact that $\|A\|_2 \leq \|A\|_F$ for a matrix A.

By triangular inequality, we have, at any point $\mathbf{w} \in B(\mathbf{w}_0, R)$,

$$\lambda_{\min}(K(\mathbf{w})) \geq \lambda_{\min}(K(\mathbf{w}_0)) - \|K(\mathbf{w}) - K(\mathbf{w}_0)\|_2 \geq \mu. \tag{15}$$

Hence, the tangent kernel is μ-uniformly conditioned in the ball $B(\mathbf{w}_0, R)$.

By Theorem 1, we immediately have that the square loss $\mathcal{L}(\mathbf{w})$ satisfies the μ-PL* condition in the ball $B(\mathbf{w}_0, R)$. □

Below in Section 4.2, we will see that wide neural networks with linear output layer indeed have small Hessian norms.

Conditioning of transformed systems. We now discuss why the conditioning of a system $\mathcal{F}(\mathbf{w}) = \mathbf{y}$ is preserved under transformations of the domain or range of $\mathcal{F}$, as long as the original system is well-conditioned and the transformation has a bounded inverse.

Remark 4. Note that the even if the original system had a small Hessian norm, there is no such guarantee for the transformed system.

Consider a transformation $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ that, composed with $\mathcal{F}$, results in a new transformed system $\Phi \circ \mathcal{F}(\mathbf{w}) = \mathbf{y}$. Put

$$\rho := \inf_{\mathbf{w} \in B(\mathbf{w}_0, R)} \frac{1}{\|J_\Phi^{-1}(\mathbf{w})\|_2}, \tag{16}$$

where $J_\Phi(\mathbf{w}) := J_\Phi(\mathcal{F}(\mathbf{w}))$ is the Jacobian of Φ evaluated at $\mathcal{F}(w)$. We will assume that $\rho > 0$.

Theorem 3. *If a system $\mathcal{F}$ is μ-uniformly conditioned in a ball $B(\mathbf{w}_0, R)$ with $R > 0$, then the transformed system $\Phi \circ \mathcal{F}(\mathbf{w})$ is $\mu\rho^2$-uniformly conditioned in $B(\mathbf{w}_0, R)$. Hence, the square loss function $\frac{1}{2}\|\Phi \circ \mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2$ satisfies the $\mu\rho^2$-PL* condition in $B(\mathbf{w}_0, R)$.*

Proof. First, note that,

$$K_{\Phi \circ \mathcal{F}}(\mathbf{w}) = \nabla(\Phi \circ \mathcal{F})(\mathbf{w}) \nabla(\Phi \circ \mathcal{F})^T(\mathbf{w}) = J_\Phi(\mathbf{w}) K_{\mathcal{F}}(\mathbf{w}; X) J_\Phi(\mathbf{w})^T.$$

Hence, if $\mathcal{F}(\mathbf{w})$ is μ-uniformly conditioned in $B(\mathbf{w}_0, R)$, i.e. $\lambda_{\min}(K_{\mathcal{F}}(\mathbf{w})) \geq \mu$, we have for any $\mathbf{v} \in \mathbb{R}^n$ with $\|\mathbf{v}\| = 1$,

$$\begin{aligned} \mathbf{v}^T K_{\Phi \circ \mathcal{F}}(\mathbf{w}) \mathbf{v} &= (J_\Phi(\mathbf{w})^T \mathbf{v})^T K_{\mathcal{F}}(\mathbf{w}) (J_\Phi(\mathbf{w})^T \mathbf{v}) \\ &\geq \lambda_{\min}(K_{\mathcal{F}}(\mathbf{w})) \|J_\Phi(\mathbf{w})^T \mathbf{v}\|^2 \\ &\geq \lambda_{\min}(K_{\mathcal{F}}(\mathbf{w})) / \|J_\Phi^{-1}(\mathbf{w})\|_2^2 \geq \mu\rho^2. \end{aligned}$$

Applying Theorem 1, we immediately obtain that $\frac{1}{2}\|\Phi \circ \mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2$ satisfies $\mu\rho^2$-PL* condition in $B(\mathbf{w}_0, R)$. □

Remark 5. A result analogous to Theorem 3 is easy to obtain for a system with transformed input $\mathcal{F} \circ \Psi(\mathbf{w}) = \mathbf{y}$. Assume the transformation map $\Psi : \mathbb{R}^m \to \mathbb{R}^m$ that applies on the input of the system $\mathbf{w}$ satisfies

$$\rho := \inf_{\mathbf{w} \in B(\mathbf{w}_0, R)} \frac{1}{\|J_\Psi^{-1}(\mathbf{w})\|_2} > 0. \tag{17}$$

If $\mathcal{F}$ is μ-uniformly conditioned with respect to $\Psi(\mathbf{w})$ in $B(\mathbf{w}_0, R)$, then an analysis similar to Theorem 3 shows that $\mathcal{F} \circ \Psi$ is also $\mu\rho^2$-uniformly conditioned in $B(\mathbf{w}_0, R)$.

Conditioning of composition models. Although composing different large models often leads to non-constant tangent kernels, the corresponding tangent kernels can also be uniformly conditioned, under certain conditions. Consider the composition of two models $h := g \circ f$, where $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $g : \mathbb{R}^{d'} \to \mathbb{R}^{d''}$. Denote $\mathbf{w}_f$ and $\mathbf{w}_g$ as the parameters of model f and g respectively. Then, the parameters of the composition model

h are $\mathbf{w} := (\mathbf{w}_g, \mathbf{w}_f)$. Examples of the composition models include "bottleneck" neural networks, where the modules below or above the bottleneck layer can be considered as the composing (sub-)models.

Let's denote the tangent kernel matrices of models g and f by $K_g(\mathbf{w}_g; \mathcal{Z})$ and $K_f(\mathbf{w}_f; \mathcal{X})$ respectively, where the second arguments, $\mathcal{Z}$ and $\mathcal{X}$, are the datasets that the tangent kernel matrices are evaluated on. Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, denote $f(\mathcal{D})$ as $\{(f(\mathbf{x}_i), y_i)\}_{i=1}^n$.

Proposition 5. *Consider the composition model $h = g \circ f$ with parameters $\mathbf{w} = (\mathbf{w}_g, \mathbf{w}_f)$. Given a dataset $\mathcal{D}$, the tangent kernel matrix of h takes the form:*

$$K_h(\mathbf{w}; \mathcal{D}) = K_g(\mathbf{w}_g; f(\mathcal{D})) + J_g(f(\mathcal{D}))K_f(\mathbf{w}_f; \mathcal{D})J_g(f(\mathcal{D}))^T,$$

where $J_g(f(\mathcal{D})) \in \mathbb{R}^{d'' \times d'}$ is the Jacobian of g w.r.t. f evaluated on $f(\mathcal{D})$.

From the above proposition, we see that the tangent kernel of the composition model h can be decomposed into the sum of two positive semi-definite matrices, hence the minimum eigenvalue of K_h can be lower bounded by

$$\lambda_{\min}(K_h(\mathbf{w})) \geq \lambda_{\min}(K_g(\mathbf{w}_g; f(\mathcal{D}))). \tag{18}$$

We note that g takes the outputs of f as inputs which depends on $\mathbf{w}_f$, while for the model f the inputs are fixed. Hence, if the model g is uniformly conditioned at all the inputs provided by the model f, we can expect the uniform conditioning of the composition model h.

We provide a simple illustrative example for how a "composition control" can be used to show that a "bottleneck" neural network satisfies the PL* condition in Appendix D.

4.2. Wide neural networks satisfy PL* condition

In this subsection, we show that wide neural networks satisfy a PL* condition, using the techniques we developed in the last subsection.

An L-layer (feedforward) neural network $f(\mathbf{W}; \mathbf{x})$, with parameters $\mathbf{W}$ and input $\mathbf{x}$, is defined as follows:

$$\begin{aligned} &\alpha^{(0)} = \mathbf{x}, \\ &\alpha^{(l)} = \sigma_l\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right), \quad \forall l = 1, 2, \cdots, L+1, \\ &f(\mathbf{W}; \mathbf{x}) = \alpha^{(L+1)}. \end{aligned} \tag{19}$$

Here, m_l is the width (i.e., number of neurons) of l-th layer, $\alpha^{(l)} \in \mathbb{R}^{m_l}$ denotes the vector of l-th hidden layer neurons, $\mathbf{W} := \{W^{(1)}, W^{(2)}, \ldots, W^{(L)}, W^{(L+1)}\}$ denotes the collection of the parameters (or weights) $W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ of each layer, and σ_l is the activation function of l-th layer, e.g., *sigmoid*, *tanh*, linear activation. We also denote the width of the neural network as $m := \min_{l \in [L]} m_l$, i.e., the minimal width of the hidden layers. In the following analysis, we assume that the activation functions σ_l are twice differentiable. Admittedly, this assumption excludes the popular ReLU activation function, which is not differentiable. However, it is likely that similar results can be obtained in for ReLU activation (see, e.g., [12,38] for related analyses in that setting).

Remark 6. The above definition of neural networks does not include convolutional (CNN) and residual (ResNet) neural networks. In Appendix E, we show that both CNN and ResNet also satisfy the PL* condition. Please see the definitions and analysis there.

We study the loss landscape of wide neural networks in regions around randomly chosen points in parameter space. Specifically, we consider the ball $B(\mathbf{W}_0, R)$, which has a fixed radius $R > 0$ (we will see later, in Section 5, that R can be chosen to cover the whole optimization path) and is around a random parameter point $\mathbf{W}_0$, i.e., $W_0^{(l)} \sim \mathcal{N}(0, I_{m_l \times m_{l-1}})$ for $l \in [L+1]$. Note that such a random parameter point $\mathbf{W}_0$ is a common choice to initialize a neural network. Importantly, the tangent kernel matrix at $\mathbf{W}_0$ is generally strictly positive definite, i.e., $\lambda_{\min}(K(\mathbf{W}_0)) > 0$. Indeed, this is proven for infinitely wide networks as long as the training data is not degenerate (see Theorem 3.1 of [12] and Proposition F.1 and F.2 of [13]). As for finite width networks, with high probability w.r.t. the initialization randomness, its tangent kernel $K(\mathbf{W}_0)$ is close to that of the infinite network and the minimum eigenvalue $\lambda_{\min}(K(\mathbf{W}_0)) = O(1)$.

Using the techniques in Section 4.1, the following theorem shows that neural networks with sufficient width satisfies a PL* condition over a ball of any fixed radius around $\mathbf{W}_0$, as long as the tangent kernel $K(\mathbf{W}_0)$ is strictly positive definite.

Theorem 4 *(Wide neural networks satisfy PL* condition). Consider the neural network $f(\mathbf{W}; \mathbf{x})$ in Eq. (19), and a random parameter setting $\mathbf{W}_0$ such that $W_0^{(l)} \sim \mathcal{N}(0, I_{m_l \times m_{l-1}})$ for $l \in [L+1]$. Suppose that the last layer activation σ_{L+1} satisfies $|\sigma'_{L+1}(z)| \geq \rho > 0$ and that $\lambda_0 := \lambda_{\min}(K(\mathbf{W}_0)) > 0$. For any $\mu \in (0, \lambda_0 \rho^2)$, if the width of the network*

$$m = \tilde{\Omega}\left(\frac{nR^{6L+2}}{(\lambda_0 - \mu\rho^{-2})^2}\right), \tag{20}$$

then the square loss function satisfies μ-PL* condition over the ball $B(\mathbf{w}_0, R)$.

Remark 7. In fact, it is not necessary to require $|\sigma'_{L+1}(z)|$ to be greater than ρ for all z. The theorem still holds as long as $|\sigma'_{L+1}(z)| > \rho$ is true for all values z actually achieved by the output neuron before activation.

This theorem tells that while the loss landscape of wide neural networks is nowhere convex (as seen in Section 3), it can still can be described by the PL* condition at many points, in line with our general discussion.

Proof of Theorem 4. We divide the proof into two distinct steps based on representing an arbitrary neural network as a composition of network with a linear output layer and an output non-linearity $\sigma_{L+1}(\cdot)$. In Step 1 we prove the PL* condition for the case of a network with a linear output layer (i.e., $\sigma_{L+1}(z) \equiv z$). The argument relies on the fact that wide neural networks with linear output layer have small Hessian norm in a ball around initialization. In Step 2 for general networks we observe that an arbitrary neural network is simply a neural network with a linear output layer from Step 1 with output transformed by applying $\sigma_{L+1}(z)$ coordinate-wise. We obtain the result by combining Theorem 3 with Step 1.

Step 1. Linear output layer: $\sigma_{L+1}(z) \equiv z$. In this case, $\rho = 1$ and the output layer of the network has a linear form, i.e., a linear combination of the units from the last hidden layer.

As was shown in [24], for this type of network with sufficient width, the model Hessian matrix has arbitrarily small spectral norm (a *transition to linearity*). This is formulated in the following theorem:

Theorem 5 *(Theorem 3.2 of [24]: transition to linearity). Consider a neural network $f(\mathbf{W}; \mathbf{x})$ of the form Eq. (19). Let m be the minimum of the hidden layer widths, i.e., $m = \min_{l \in [L]} m_l$. Given any fixed $R > 0$, and any $\mathbf{W} \in B(\mathbf{W}_0, R) := \{\mathbf{W} : \|\mathbf{W} - \mathbf{W}_0\| \leq R\}$, with high probability over the initialization, the Hessian spectral norm satisfies the following:*

$$\|H_f(\mathbf{W})\| = \tilde{O}\left(R^{3L}/\sqrt{m}\right). \tag{21}$$

In Eq. (21), we explicitly write out the dependence of Hessian norm on the radius R, according to the proof in [24].

Directly plugging Eq. (21) into the condition of Theorem 2, letting $\epsilon = \lambda_0 - \mu$ and noticing that $\rho = 1$, we directly have the expression for the width m.

Step 2. General networks: $\sigma_{L+1}(\cdot)$ is non-linear. Wide neural networks with non-linear output layer generally do not exhibit transition to linearity or near-constant tangent kernel, as shown in [24]. Despite that, these wide networks still satisfy the PL* condition in the ball $B(\mathbf{W}_0, R)$. Observe that this type of network can be viewed as a composition of a non-linear transformation function σ_{L+1} with a network $\tilde{f}$ which has a linear output layer:

$$f(\mathbf{W}; \mathbf{x}) = \sigma_{L+1}(\tilde{f}(\mathbf{W}; \mathbf{x})). \tag{22}$$

By the same argument as in Step 1, we see that $\tilde{f}$, with the width as in Eq. (20), is $\frac{\mu}{\rho^2}$-uniformly conditioned.

Now we apply our analysis for transformed systems in Theorem 3. In this case, the transformation map Φ becomes a coordinate-wise transformation of the output given by

$$\Phi(\cdot) = \operatorname{diag}(\underbrace{\sigma_{L+1}(\cdot), \sigma_{L+1}(\cdot), \cdots, \sigma_{L+1}(\cdot)}_{n}), \tag{23}$$

and the norm of the inverse Jacobian matrix, $\|J_\Phi^{-1}(\mathbf{w})\|_2$ is

$$\|J_\Phi^{-1}(\mathbf{w})\|_2 = \frac{1}{\min_{i\in[n]} \left|\sigma'_{L+1}(\tilde{f}(\mathbf{w}; \mathbf{x}_i))\right|} \leq \rho^{-1}. \tag{24}$$

Hence, the $\frac{\mu}{\rho^2}$-uniform conditioning of $\tilde{f}$ immediately implies μ-uniform conditioning of f, as desired.

5. PL* condition in a ball guarantees existence of solutions and fast convergence of (S)GD

In this section, we show that fast convergence of gradient descent methods is guaranteed by the PL* condition in a ball with appropriate size. We assume the map $\mathcal{F}(\mathbf{w})$ is $L_\mathcal{F}$-Lipschitz continuous and $\beta_\mathcal{F}$-smooth on the local region $\mathcal{S}$ that we considered. In what follows $\mathcal{S}$ will typically be a Euclidean ball $B(\mathbf{w}_0, R)$, with an appropriate radius R, chosen to cover the optimization path of GD or SGD.

First, we define the (non-linear) condition number, as follows:

Definition 4 *(Condition number).* Consider a system in Eq. (2), with a loss function $\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathcal{F}(\mathbf{w}), \mathbf{y})$ and a set $\mathcal{S} \subset \mathbb{R}^m$. Suppose the loss $\mathcal{L}(\mathbf{w})$ is PL* on $\mathcal{S}$ and let $\mu > 0$ be the largest number such that μ-PL* holds on $\mathcal{S}$. Define the condition number $\kappa_{\mathcal{L},\mathcal{F}}(\mathcal{S})$:

$$\kappa_{\mathcal{L},\mathcal{F}}(\mathcal{S}) := \frac{\sup_{\mathbf{w}\in\mathcal{S}} \lambda_{\max}(H_\mathcal{L}(\mathbf{w}))}{\mu}, \tag{25}$$

where $H_\mathcal{L}(\mathbf{w})$ is the Hessian matrix of the loss function. The condition number for *the square loss* (used throughout the paper) will be written as simply $\kappa_\mathcal{F}(\mathcal{S})$, omitting the subscript $\mathcal{L}$.

Remark 8. In the special case of a linear system $\mathcal{F}(\mathbf{w}) = A\mathbf{w}$ with square loss $\frac{1}{2}\|A\mathbf{w} - \mathbf{y}\|^2$, both the Hessian $H_\mathcal{L} = A^T A$ and the tangent kernel $K(\mathbf{w}) = A\,A^T$ are constant matrices. As AA^T and $A^T A$ have the same set of non-zero eigenvalues, the largest eigenvalue $\lambda_{\max}(H_\mathcal{L})$ is equal to $\lambda_{\max}(K)$. In this case, the condition number $\kappa_\mathcal{F}(\mathcal{S})$ reduces to the standard condition number of the tangent kernel K,

$$\kappa_\mathcal{F}(\mathcal{S}) = \frac{\lambda_{\max}(K)}{\lambda_{\min}(K)}. \tag{26}$$

Since $\mathcal{F}$ is $L_{\mathcal{F}}$-Lipschitz continuous and $\beta_{\mathcal{F}}$ smooth by assumption, we directly get the following by substituting the definition of the square loss function into Eq. (25).

Proposition 6. *For the square loss function $\mathcal{L}$, the condition number is upper bounded by:*

$$\kappa_{\mathcal{F}}(\mathcal{S}) \leq \frac{L_{\mathcal{F}}^2 + \beta_{\mathcal{F}} \cdot \sup_{\mathbf{w}\in\mathcal{S}} \|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|}{\mu}. \tag{27}$$

Remark 9. It is easy to see that the usual condition number $\kappa(\mathbf{w}) = \lambda_{\max}(K(\mathbf{w}))/\lambda_{\min}(K(\mathbf{w}))$ of the tangent kernel $K(\mathbf{w})$, is upper bounded by $\kappa_{\mathcal{F}}(\mathcal{S})$.

Now, we are ready to present an optimization theory based on the PL* condition. First, let the arbitrary set $\mathcal{S}$ be a Euclidean ball $B(\mathbf{w}_0, R)$ around the initialization $\mathbf{w}_0$ of gradient descent methods, with a reasonably large but finite radius R. The following theorem shows that satisfaction of the PL* condition on $B(\mathbf{w}_0, R)$ implies the existence of at least one global solution of the system in the same ball $B(\mathbf{w}_0, R)$. Moreover, following the original argument from [32], the PL* condition also implies fast convergence of gradient descent to a global solution w^* in the ball $B(\mathbf{w}_0, R)$.

Theorem 6 *(Local PL* condition ⇒ existence of a solution + fast convergence). Suppose the system $\mathcal{F}$ is $L_{\mathcal{F}}$-Lipschitz continuous and $\beta_{\mathcal{F}}$-smooth. Suppose the square loss $\mathcal{L}(\mathbf{w})$ satisfies the μ-PL* condition in the ball $B(\mathbf{w}_0, R) := \{\mathbf{w} \in \mathbb{R}^m : \|\mathbf{w} - \mathbf{w}_0\| \leq R\}$ with $R = \frac{2L_{\mathcal{F}}\|\mathcal{F}(\mathbf{w}_0)-\mathbf{y}\|}{\mu}$. Then we have the following:*

(a) Existence of a solution: There exists a solution (global minimizer of $\mathcal{L}$) $\mathbf{w}^* \in B(\mathbf{w}_0, R)$, such that $\mathcal{F}(\mathbf{w}^*) = \mathbf{y}$.

(b) Convergence of GD: Gradient descent with a step size $\eta \leq 1/(L_{\mathcal{F}}^2 + \beta_{\mathcal{F}}\|\mathcal{F}(\mathbf{w}_0) - \mathbf{y}\|)$ converges to a global solution in $B(\mathbf{w}_0, R)$, with an exponential (a.k.a. linear) convergence rate:

$$\mathcal{L}(\mathbf{w}_t) \leq \left(1 - \kappa_{\mathcal{F}}^{-1}(B(\mathbf{w}_0, R))\right)^t \mathcal{L}(\mathbf{w}_0), \tag{28}$$

where the condition number $\kappa_{\mathcal{F}}(B(\mathbf{w}_0, R)) = \frac{1}{\eta\mu}$.

The proof of the theorem is deferred to Appendix F.

It is interesting to note that the radius R of the ball $B(\mathbf{w}_0, R)$ takes a finite value, which means that the optimization path $\{\mathbf{w}_t\}_{t=0}^{\infty} \subset B(\mathbf{w}_0, R)$ stretches at most a finite length and the optimization happens only within a region of definite radius around the initialization point. Hence, the conditioning of the tangent kernel and the satisfaction of the PL* condition outside of this ball are irrelevant to this optimization and are not required.

Indeed, this radius R has to be of order $\Omega(1)$. From the $L_{\mathcal{F}}$-Lipschitz continuity of $\mathcal{F}(\mathbf{w})$, it follows that there is no solution within the distance $\|\mathcal{F}(\mathbf{w}_0) - \mathbf{y}\|/L_{\mathcal{F}}$ from the initialization point. Any solution must have a Euclidean distance away from $\mathbf{w}_0$ at least $R_{min} = \|\mathcal{F}(\mathbf{w}_0) - \mathbf{y}\|/L_{\mathcal{F}}$, which is finite. This means that the parameter update $\Delta\mathbf{w} = \mathbf{w}^* - \mathbf{w}_0$ cannot be too small in terms of the Euclidean distance. However, due to the large number m of the model parameters, each individual parameter w_i may only undergo a small change during the gradient descent training, i.e., $|w_i - w_{0,i}| = O(1/\sqrt{m})$. That, indeed, this has been observed theoretically for wide neural networks (see, e.g., the discussion in [24]).

Below, we make an extension of the above theory: from (deterministic) gradient descent to stochastic gradient descent (SGD).

Convergence of SGD. In most practical machine learning settings, including typical problems of supervised learning, the loss function $\mathcal{L}(\mathbf{w})$ has the form

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \ell_i(\mathbf{w}).$$

For example, for the square loss $\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \ell_i(\mathbf{w})$, where $\ell_i(\mathbf{w}) = \frac{1}{2}(\mathcal{F}_i(\mathbf{w}) - y_i)^2$. Here the loss ℓ_i corresponds simply to the loss for ith equation. Mini-batch SGD updates the parameter $\mathbf{w}$, according to the gradient of s individual loss functions $\ell_i(\mathbf{w})$ at a time:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \sum_{i \in \mathcal{S} \subset [n]} \nabla \ell_i(\mathbf{w}_t), \forall t \in \mathbb{N}.$$

We will assume that each element of the set S is chosen uniformly at random at every iteration.

We now show that the PL* condition on $\mathcal{L}$ also implies exponential convergence of SGD within a ball, an SGD analogue of Theorem 6. Our result can be considered as a local version of Theorem 1 in [4] which showed exponential convergence of SGD, assuming PL condition holds in the entire parameter space. See also [30] for a related result.

Theorem 7. *Given $0 < \delta < 1$, assume each $\ell_i(\mathbf{w})$ is β-smooth and $\mathcal{L}(\mathbf{w})$ satisfies the μ-PL* condition in the ball $B(\mathbf{w}_0, R)$ with $R = \frac{2n\sqrt{2\beta\mathcal{L}(\mathbf{w}_0)}}{\mu\delta}$. Then, with probability $1-\delta$, SGD with mini-batch size $s \in \mathbb{N}$ and step size $\eta \leq \frac{n\mu}{n\beta(n^2\beta+\mu(s-1))}$ converges to a global solution in the ball $B(\mathbf{w}_0, R)$, with an exponential convergence rate:*

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t)] \leq \left(1 - \frac{\mu s \eta}{n}\right)^t \mathcal{L}(\mathbf{w}_0). \tag{29}$$

The proof is deferred to Appendix G.

5.1. Convergence for wide neural networks

Using the optimization theory developed above, we can now show convergence of (S)GD for sufficiently wide neural networks. As we have seen in Section 4.2, the loss landscapes of wide neural networks $f(\mathbf{W}; \mathbf{x})$ defined in Eq. (19) obey the PL* condition over a ball $B(\mathbf{W}_0, R)$ of an arbitrary radius R. We will now show convergence of (S)GD for these models. Since the argument is essentially the same, we state the main result for GD and note the difference with SGD in Remark 10.

As before (Section 4.2) write $f(\mathbf{W}; \mathbf{x}) = \sigma_{L+1}(\tilde{f}(\mathbf{W}; \mathbf{x}))$ where $\tilde{f}(\mathbf{W}; \mathbf{x})$ is a neural network with a linear output layer and we denoted $\mathcal{F}(\mathbf{W}) = (f(\mathbf{W}; \mathbf{x}_1), \ldots, f(\mathbf{W}; \mathbf{x}_n))$, $\tilde{\mathcal{F}}(\mathbf{W}) = (\tilde{f}(\mathbf{W}; \mathbf{x}_1), \ldots, \tilde{f}(\mathbf{W}; \mathbf{x}_n))$ and $\mathbf{y} = (y_1, \ldots, y_n)$. We use tilde, e.g., $\tilde{O}(\cdot)$ to suppress logarithmic terms in Big-O notation.

We further assume $\sigma_{L+1}(\cdot)$ is L_σ-Lipschitz continuous and β_σ-smooth.

Theorem 8. *Consider the neural network $f(\mathbf{W}; \mathbf{x})$ and its random initialization $\mathbf{W}_0$ under the same condition as in Theorem 4. If the network width satisfies $m = \tilde{\Omega}\left(\frac{n}{\mu^{6L+2}(\lambda_0 - \mu\rho^{-2})^2}\right)$, then, with an appropriate step size, gradient descent converges to a global minimizer in the ball $B(\mathbf{W}_0, R)$, where $R = O(1/\mu)$, with an exponential convergence rate:*

$$\mathcal{L}(\mathbf{W}_t) \leq (1 - \eta\mu)^t \mathcal{L}(\mathbf{W}_0). \tag{30}$$

Proof. The result is obtained by combining Theorem 4 and 6, after setting the ball radius $R = 2L_{\mathcal{F}}\|\mathcal{F}(\mathbf{W}_0) - \mathbf{y}\|/\mu$ and choosing the step size $0 < \eta < \frac{1}{L_{\mathcal{F}}^2 + \beta_{\mathcal{F}}\|\mathcal{F}(\mathbf{W}_0) - \mathbf{y}\|}$.

It remains to verify that all of the quantities $L_{\mathcal{F}}$, $\beta_{\mathcal{F}}$ and $\|\mathcal{F}(\mathbf{W}_0) - \mathbf{y}\|$ are of order $O(1)$ with respect to the network width m. First note that, with the random initialization of $\mathbf{W}_0$ as in Theorem 4, it is shown in

[17] that with high probability, $f(\mathbf{W}_0;\mathbf{x}) = O(1)$ when the width is sufficiently large under mild assumption on the non-linearity functions $\sigma_l(\cdot)$, $l = 1, ..., L+1$. Hence $\|\mathcal{F}(\mathbf{W}_0) - \mathbf{y}\| = O(1)$.

Using the definition of $L_{\mathcal{F}}$, we have

$$\begin{aligned} L_{\mathcal{F}} &= \sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \|D\mathcal{F}(\mathbf{W})\| \\ &\le L_\sigma \left(\|D\tilde{\mathcal{F}}(\mathbf{W}_0)\| + R\sqrt{n} \cdot \sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \|\mathbf{H}_{\tilde{\mathcal{F}}}(\mathbf{W})\| \right) \\ &= \sqrt{\|K_{\tilde{\mathcal{F}}}(\mathbf{W}_0)\|} L_\sigma + L_\sigma R\sqrt{n} \cdot \tilde{O}\left(\frac{1}{\sqrt{m}}\right) = O(1), \end{aligned}$$

where the last inequality follows from Theorem 5, and $\|K_{\tilde{\mathcal{F}}}(\mathbf{W}_0)\| = O(1)$ with high probability over the random initialization.

Finally, $\beta_{\mathcal{F}}$ is bounded as follows:

$$\begin{aligned} \beta_{\mathcal{F}} &= \sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \|\mathbf{H}_{\mathcal{F}}(\mathbf{W})\| \\ &= \sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \|H_{f_k}(\mathbf{W})\| \\ &= \sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \left\| \sigma''_{L+1}\left(\tilde{f}_k(\mathbf{W})\right) D\tilde{f}_k(\mathbf{W}) D\tilde{f}_k(\mathbf{W})^T + \sigma'_{L+1}\left(\tilde{f}_k(\mathbf{W})\right) H_{\tilde{f}_k}(\mathbf{W}) \right\| \\ &\le \beta_\sigma \left(\sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \|D\tilde{f}_k(\mathbf{W})\| \right)^2 + L_\sigma \sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \|H_{\tilde{f}_k}(\mathbf{W})\| \\ &\le \beta_\sigma \cdot O(1) + L_\sigma \cdot \tilde{O}\left(\frac{1}{\sqrt{m}}\right) = O(1), \end{aligned}$$

where $k = \underset{i\in[n]}{\operatorname{argmax}} \left(\sup_{\mathbf{W}\in B(\mathbf{W}_0,R)} \|H_{f_i}(\mathbf{W})\| \right)$. □

Remark 10. Using the same argument, a result similar to Theorem 8 but with a different convergence rate,

$$\mathcal{L}(\mathbf{W}_t) \le (1 - \eta s \mu)^t \mathcal{L}(\mathbf{W}_0), \tag{31}$$

and with difference constants, can be obtained for SGD by applying Theorem 7.

Note that (near) constancy of the tangent kernel is not a necessary condition for exponential convergence of gradient descent or (S)GD.

6. "Almost" over-parameterized systems: relaxation of PL* to PL$^*_\epsilon$ condition

In certain situations, for example mildly under-parameterized cases, the PL* condition may not hold exactly, since an exact solution for system $\mathcal{F}(\mathbf{w}) = \mathbf{w}$ may not exist. In practice, however, the algorithms are rarely run a until exact convergence. Most of the time, *early stopping* is employed, i.e., the algorithm is stopped once it achieves a certain loss $\epsilon > 0$. To account for that, we define PL$^*_\epsilon$, a relaxed variant of PL* condition, which still implies fast convergence of gradient-based methods up to loss ϵ. While we propose a mathematical framework for "almost" over-parameterization, making connections to specific systems, such is neural networks, is left as a direction for future research.

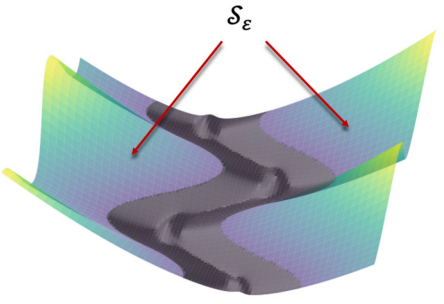


Fig. 5. The loss landscape of under-parameterized systems. In the set $\mathcal{S}_\epsilon$, where the loss is larger than ϵ, PL* still holds. Beyond that, the loss landscape can be arbitrary, which is the grey area, and PL* doesn't hold any more.

Definition 5 *(PL$^*_\epsilon$ condition).* Given a set $\mathcal{S} \subset \mathbb{R}^m$ and $\epsilon > 0$, define the set $\mathcal{S}_\epsilon := \{\mathbf{w} \in \mathcal{S} : \mathcal{L}(\mathbf{w}) \geq \epsilon\}$. A loss function $\mathcal{L}(\mathbf{w})$ is μ-PL$^*_\epsilon$ on $\mathcal{S}$, if the following holds:

$$\frac{1}{2}\|\nabla \mathcal{L}(\mathbf{w})\|^2 \geq \mu \mathcal{L}(\mathbf{w}), \ \forall \mathbf{w} \in \mathcal{S}_\epsilon. \tag{32}$$

Intuitively, the PL$^*_\epsilon$ condition is the same as PL* condition, except that the loss landscape can be arbitrary wherever the loss is less than ϵ. This is illustrated in Fig. 5.

Below, following a similar argument above, we show that a *local* PL$^*_\epsilon$ condition guarantees fast convergence to an approximation of a global minimizer. Basically, the gradient descent terminates when the loss is less than ϵ.

Theorem 9 *(Local PL$^*_\epsilon$ condition ⇒ fast convergence). Assume the loss function $\mathcal{L}(\mathbf{w})$ (not necessarily the square loss) is β-smooth and satisfies the μ-PL$^*_\epsilon$ condition in the ball $B(\mathbf{w}_0, R) := \{\mathbf{w} \in \mathbb{R}^m : \|\mathbf{w}-\mathbf{w}_0\| \leq R\}$ with $R = \frac{2\sqrt{2\beta\mathcal{L}(\mathbf{w}_0)}}{\mu}$. We have the following:*

(a) Existence of a point which makes the loss less than ϵ: There exists a point $\mathbf{w}^* \in B(\mathbf{w}_0, R)$, such that $\mathcal{L}(\mathbf{w}^*) < \epsilon$.

(b) Convergence of GD: Gradient descent with the step size $\eta = 1/\sup_{\mathbf{w}\in B(\mathbf{w}_0,R)} \|H_{\mathcal{L}}(\mathbf{w})\|_2$ after $T = \Omega(\log(1/\epsilon))$ iterations satisfies $\mathcal{L}(\mathbf{w}_T) < \epsilon$ in the ball $B(\mathbf{w}_0, R)$, with an exponential (also known as linear) convergence rate:

$$\mathcal{L}(\mathbf{w}_t) \leq \left(1 - \kappa^{-1}_{\mathcal{L},F}(B(\mathbf{w}_0, R))\right)^t \mathcal{L}(\mathbf{w}_0), \ \textit{for all} \ \ t \leq T, \tag{33}$$

where the condition number $\kappa_{\mathcal{L},F}(B(\mathbf{w}_0, R)) = \frac{1}{\eta\mu}$.

Proof. If $\mathcal{L}(\mathbf{w}_0) < \epsilon$, we let $\mathbf{w}^* = \mathbf{w}_0$ and we are done.

Suppose $\mathcal{L}(\mathbf{w}_0) \geq \epsilon$. Following the similar analysis to the proof of Theorem 6, as long as $\mathcal{L}(\mathbf{w}_t) \geq \epsilon$ for $t \geq 0$, we have

$$\mathcal{L}(\mathbf{w}_t) \leq (1 - \eta\mu)^t \mathcal{L}(\mathbf{w}_0). \tag{34}$$

Hence there must exist a minimum $T > 0$ such that

$$\mathcal{L}(\mathbf{w}_T) < \epsilon. \tag{35}$$

It's not hard to see that $\epsilon \leq (1 - \eta\mu)^T \mathcal{L}(\mathbf{w}_0)$, from which we get $T = \Omega(\log(1/\epsilon))$.

And obviously, $\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_T$ are also in the ball $B(\mathbf{w}_0, R)$ with $R = \frac{2\sqrt{2\beta\mathcal{L}(\mathbf{w}_0)}}{\mu}$. □

Following similar analysis, when a PL^*_ϵ condition holds over a ball, SGD converges exponentially to an approximation of the global solution.

Theorem 10. *Assume each $\ell_i(\mathbf{w})$ is γ-smooth and $\mathcal{L}(\mathbf{w})$ satisfies the μ-PL^*_ϵ condition in the ball $B(\mathbf{w}_0, R)$ with $R = \frac{2n\sqrt{2\gamma}\sqrt{\mathcal{L}(\mathbf{w}_0)}}{\mu\delta}$ where $\delta > 0$. Then, with probability $1-\delta$, SGD with mini-batch size $s \in \mathbb{N}$ and step size $\eta^*(s) = \frac{n\mu}{n\gamma(n^2\gamma+\mu(s-1))}$ after at most $T = \Omega(\log(1/\epsilon))$ iterations satisfies $\min\left(\mathbb{E}[\mathcal{L}(\mathbf{w}_T)], \mathcal{L}(\mathbf{w}_T)\right) < \epsilon$ in the ball $B(\mathbf{w}_0, R)$, with an exponential convergence rate:*

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t)] \leq \left(1 - \frac{\mu s \eta^*(s)}{n}\right)^t \mathcal{L}(\mathbf{w}_0), \ t \leq T. \tag{36}$$

Proof. If $\mathcal{L}(\mathbf{w}_0) < \epsilon$, we let $T = 0$ and we are done.

Suppose $\mathcal{L}(\mathbf{w}_0) \geq \epsilon$. By the similar analysis in the proof of Theorem 7, with the μ-PL^*_ϵ condition, for any mini-batch size s, the mini-batch SGD with step size $\eta^*(s) := \frac{n\mu}{n\lambda(n^2\lambda+\mu(s-1))}$ has an exponential convergence rate:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t)] \leq \left(1 - \frac{\mu s \eta^*(s)}{n}\right)^t \mathcal{L}(\mathbf{w}_0), \tag{37}$$

for all t where $\mathcal{L}(\mathbf{w}_t) \geq \epsilon$.

Hence there must exist a minimum $T > 0$ such that either $L(\mathbf{w}_T) < \epsilon$ or $\mathbb{E}[\mathcal{L}(\mathbf{w}_T)] < \epsilon$. Supposing $\mathbb{E}[\mathcal{L}(\mathbf{w}_T)] < \epsilon$ at time step T, we have $\epsilon \leq \left(1 - \frac{\mu s \eta^*(s)}{n}\right)^T \mathcal{L}(\mathbf{w}_0)$, therefore $T = \Omega(\log(1/\epsilon))$.

And it's easy to check that with probability $1-\delta$, the optimization path $\mathbf{w}_0, \mathbf{w}_1, ..., \mathbf{w}_T$ is covered by the ball $B(\mathbf{w}_0, R)$ with $R = \frac{2n\sqrt{2\gamma}\sqrt{\mathcal{L}(\mathbf{w}_0)}}{\mu\delta}$. □

While the global landscape of the loss function $\mathcal{L}(\mathbf{w})$ can be complex, the conditions above allow us to find solutions within a certain ball around the initialization point $\mathbf{w}_0$.

7. Concluding thoughts and comments

In this paper we propose a general framework for understanding generically non-convex landscapes and optimization of over-parameterized systems using the PL^* condition. We argued that a PL^* condition generally holds on much of, but not necessarily all of, the parameter space, and shown that this is sufficient for the existence of solutions and convergence of gradient-base methods to global minimizers. In contrast, it is not possible for loss landscapes of under-parameterized systems $\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2$ to satisfy PL^* for any $\mathbf{y}$. We conclude with a number of comments and observations.

Linear and non-linear systems. A remarkable property of over-parameterized non-linear systems discussed in this work is their strong resemblance to linear systems with respect to optimization by (S)GD, even as their dynamics remain nonlinear. In particular, optimization by gradient-based methods and proximity to global minimizers is controlled by non-linear condition numbers, similarly to classical analyses of linear systems. The key difference is that while for linear systems the condition number is constant, in the non-linear case we need a uniform bound in a domain containing the optimization path. Furthermore, increasing the degree of over-parameterization generally improves conditioning just like it does for linear systems (cf. [9] and the discussion in [31]). In particular, this suggests that the effectiveness of optimization should improve, up to a certain limit, with increased over-parameterization.

We note that in contrast to over-parameterized systems, optimization properties of non-linear systems in the *under-parameterized* regime appear very different from those of linear systems.

Transition over the interpolation threshold: from local convexity to essential non-convexity. Recognizing the power of over-parameterization has been a key insight stemming from the practice of deep learning. Transition to over-parameterized models with increasing number of parameters – over the interpolation threshold – leads to a qualitative change in a range of system properties. Statistically, over-parameterized systems enter a new interpolating regime, where increasing the number of parameters, even indefinitely to infinity, can improve generalization [5,34]. From the optimization point of view, over-parameterized system are generally easier to solve. There has been significant effort (continued in this work) toward understanding effectiveness of local methods in this setting [33,12,26,2,18,10,30].

In this paper we note another aspect of this transition, to the best of our knowledge not addressed in the existing literature: transition from local convexity to *essential non-convexity* (see Section 3 for a detailed discussion), where the loss function is generally not convex in any neighborhood of a global minimum. This observation has important consequences, pointing away from the standard machinery of convex analysis and toward PL* condition as we have argued in this work. Interestingly, our analyses suggest that this loss of local convexity is of little consequence for optimization, at least as far as gradient-based methods are concerned.

Transition from over-parameterization to under-parameterization along the optimization path. As discussed above, transition over the interpolation threshold occurs when the number of parameters in a variably parameterized system exceeds the number of constraints (corresponding to data points in typical ML scenarios). Over-parameterization does not refer to the number of parameters as such but to the difference between the number of parameters and the number of constraints. While some learning models are very large, with billions or even trillions parameters, they are often trained on equally large datasets and are thus not necessarily over-parameterized. Yet, it is still tempting to view these models through the lens of over-parameterization. While precise technical analysis is beyond the scope of this paper, we conjecture that transition between effective over-parameterization to under-parameterization happens along the optimization trajectory. Initially, the system behaves as over-parameterized but as the optimization process continues, it fails to reach zero. Mathematically it can be represented as our PL^*_ϵ condition. We speculate that for many realistic large models trained on big data the full optimization path lies within the PL^*_ϵ domain and hence, functionally, the analyses in this paper apply.

Condition numbers and optimization methods. In this work we concentrate on optimization by gradient descent and SGD. Yet, for linear systems of equations and in many other settings, the importance of conditioning extends far beyond one specific optimization technique [7]. We expect this to be case in the over-parameterized non-linear setting as well. To give just one example, we expect that accelerated methods, such as the Nesterov's method [27] and its stochastic gradient extensions in the over-parameterized case [23,35] to have faster convergence rates for non-linear systems in terms of the condition numbers defined in this work.

Equations on manifolds. In this paper we consider systems of equations $\mathcal{F}(\mathbf{w}) = \mathbf{y}$ defined on Euclidean spaces and with Euclidean output. A more general setting is to look for solutions of arbitrary systems of equations defined by a map between two Riemannian manifolds $\mathcal{F} : \mathcal{M} \to \mathcal{N}$. In that case the loss function $\mathcal{L}$ needs to be defined on $\mathcal{N}$. The over-parameterization corresponds to the case when dimension $\dim(\mathcal{M}) > \dim(\mathcal{N})$. While analyzing gradient descent requires some care on a manifold, most of the mathematical machinery, including the definitions of the PL* condition and the condition number associated to $\mathcal{F}$, is still applicable without significant change. In particular, as we discussed above (see Remark 5 in Section 4.1), the condition number is preserved under "well-behaved" coordinate transformations. In contrast, this is not the case for the Hessian and thus manifold optimization analyses based on *geodesic convexity* require knowledge about specific coordinate charts, such as those given by the exponential map.

We note that manifold and structural assumptions on the weight vector $\mathbf{w}$ is a natural setting for addressing many problems in inference. In particular, the important class of convolutional neural networks

is an example of such a structural assumption on $\mathbf{w}$, which is made invariant to certain parallel transforms. Furthermore, there are many settings, e.g., robot motion planning, where the output of a predictor, $\mathbf{y}$, also belongs to a certain manifold.

Acknowledgments

We thank Raef Bassily and Siyuan Ma for many earlier discussions about gradient methods and Polyak-Łojasiewicz conditions and Stephen Wright for insightful comments and corrections. The authors gratefully acknowledge support from National Science Foundation and Simons Foundation through the Collaboration on the Theoretical Foundations of Deep Learning (awards DMS-2031883 and #814639) as well as National Science Foundation IIS-1815697 and a Google Faculty Research Award.

Appendix A. Wide neural networks have no isolated local/global minima

In this section, we show that, for feedforward neural networks, if the network width is sufficiently large, there is no isolated local/global minima in the loss landscape.

Consider the following feedforward neural networks:

$$f(\mathbf{W};\mathbf{x}) := W^{(L+1)}\phi_L(W^{(L)}\cdots\phi_1(W^{(1)}\mathbf{x})). \tag{38}$$

Here, $\mathbf{x}\in\mathbb{R}^d$ is the input, L is the number of hidden layers of the network. Let m_l be the width of l-th layer, $l\in[L]$, and $m_0 = d$ and $m_{L+1} = c$ where c is the dimension of the network output. $\mathbf{W} := \{W^{(1)}, W^{(2)}, \cdots, W^{(L)}, W^{(L+1)}\}$, with $W^{(l)}\in\mathbb{R}^{m_{l-1}\times m_l}$, is the collection of parameters, and ϕ_l is the activation function at each hidden layer. The minimal width of hidden layers, i.e., width of the network, is denoted as $m := \min\{m_1,\cdots,m_L\}$. The parameter space is denoted as $\mathcal{M}$. Here, we further assume the loss function $\mathcal{L}(\mathbf{W})$ is of the form:

$$\mathcal{L}(\mathbf{W}) = \sum_{i=1}^{n} l(f(\mathbf{W};\mathbf{x}_i),\mathbf{y}_i), \tag{39}$$

where loss $l(\cdot,\cdot)$ is convex in the first argument, and $(\mathbf{x}_i,\mathbf{y}_i)\in\mathbb{R}^d\times\mathbb{R}^c$ is one of the n training samples.

Proposition 7 *(No isolated minima). Consider the feedforward neural network in Eq. (38). If the network width $m\geq 2c(n+1)^L$, given a local (global) minimum $\mathbf{W}^*$ of the loss function Eq. (39), there are always other local (global) minima in any neighborhood of $\mathbf{W}^*$.*

The main idea of this proposition is based on some of the intermediate-level results of the work [20]. Before starting the proof, let's review some of the important concepts and results therein. To be consistent with the notation in this paper, we modify some of their notations.

Definition 6 *(Path constant).* Consider two parameters $\mathbf{W},\mathbf{V}\in\mathcal{M}$. If there is a continuous function $h_{\mathbf{W},\mathbf{V}}:[0,1]\to\mathcal{M}$ that satisfies $h_{\mathbf{W},\mathbf{V}}(0) = \mathbf{W}$, $h_{\mathbf{W},\mathbf{V}}(1) = \mathbf{V}$ and $t\mapsto\mathcal{L}(h_{\mathbf{W},\mathbf{V}}(t))$ is constant, we say that $\mathbf{W}$ and $\mathbf{V}$ are *path constant* and write $\mathbf{W}\leftrightarrow\mathbf{V}$.

Path constantness means the two parameters $\mathbf{W}$ and $\mathbf{V}$ are connected by a continuous path of parameters that is contained in a level set of the loss function $\mathcal{L}$.

Lemma 1 *(Transitivity).* $\mathbf{W}\leftrightarrow\mathbf{V}$ *and* $\mathbf{V}\leftrightarrow\mathbf{U}\Rightarrow\mathbf{W}\leftrightarrow\mathbf{U}$.

Definition 7 *(Block parameters).* Consider a number $s \in \{0, 1, \cdots\}$ and a parameter $\mathbf{W} \in \mathcal{M}$. If

1. $W_{ji}^{(1)} = 0$ for all $j > s$;
2. $W_{ij}^{(l)} = 0$ for all $l \in [L]$ and $i > s$, and for all $l \in [L]$ and $j > s$;
3. $W_{ij}^{(L+1)} = 0$ for all $j > s$,

we call $\mathbf{W}$ an s-upper-block parameter of depth L.

Similarly, if

1. $W_{ji}^{(1)} = 0$ for all $j \leq m_1 - s$;
2. $W_{ij}^{(l)} = 0$ for all $l \in [L]$ and $i \leq m_l - s$, and for all $l \in [L]$ and $j \leq m_{l-1} - s$;
3. $W_{ij}^{(L+1)} = 0$ for all $j \leq m_L - s$,

we call $\mathbf{W}$ an s-lower-block parameter of depth L. We denote the sets of the s-upper-block and s-lower-block parameters of depth L by $\mathcal{U}_{s,L}$ and $\mathcal{V}_{s,L}$, respectively.

The key result that we use in this paper is that every parameter is path constant to a block parameter, or more formally:

Proposition 8 *(Path connections to block parameters [20]). For every parameter $\mathbf{W} \in \mathcal{M}$ and $s := c(n+1)^L$ (n is the number of training samples), there are $\overline{\mathbf{W}}, \underline{\mathbf{W}} \in \mathcal{M}$ with $\overline{\mathbf{W}} \in \mathcal{U}_{s,L}$ and $\underline{\mathbf{W}} \in \mathcal{V}_{s,L}$ such that $\mathbf{W} \leftrightarrow \overline{\mathbf{W}}$ and $\mathbf{W} \leftrightarrow \underline{\mathbf{W}}$.*

The above proposition says that, if the network width m is large enough, every parameter is path connected to both an upper-block parameter and a lower-block parameter, by continuous paths contained in a level set of the loss function.

Now, let's present the proof of Proposition 7.

Proof of Proposition 7. Let $\mathbf{W}^* \in \mathcal{M}$ be an arbitrary local/global minimum of the loss function $\mathcal{L}(\mathbf{W})$. According to Proposition 8, there exists an upper-block parameter $\overline{\mathbf{W}} \in \mathcal{U}_{s,L} \subset \mathcal{M}$ and $\underline{\mathbf{W}} \in \mathcal{V}_{s,L} \subset \mathcal{M}$ such that $\mathbf{W} \leftrightarrow \overline{\mathbf{W}}$ and $\mathbf{W} \leftrightarrow \underline{\mathbf{W}}$. Note that $\overline{\mathbf{W}}$ and $\underline{\mathbf{W}}$ are distinct, because $\mathcal{U}_{s,L}$ and $\mathcal{V}_{s,L}$ do not intersect except at zero due to $m \geq 2s$. This means there must be parameters distinct from $\mathbf{W}^*$ that are connected to $\mathbf{W}^*$ via a continuous path contained in a level set of the loss function. Note that all the points (i.e., parameters) along this path have the same loss value as $\mathcal{L}(\mathbf{W}^*)$, hence are local/global minima. Therefore, $\mathbf{W}^*$ is not isolated (i.e., there are other local/global minima in any neighborhood of $\mathbf{W}^*$). □

Appendix B. Proof of Proposition 3

Proof. The Hessian matrix of a general loss function $\mathcal{L}(\mathcal{F}(\mathbf{w}))$ takes the form

$$H_{\mathcal{L}}(\mathbf{w}) = D\mathcal{F}(\mathbf{w})^T \frac{\partial^2 \mathcal{L}}{\partial \mathcal{F}^2}(\mathbf{w}) D\mathcal{F}(\mathbf{w}) + \sum_{i=1}^{n} \frac{\partial \mathcal{L}}{\partial \mathcal{F}_i}(\mathbf{w}) H_{\mathcal{F}_i}(\mathbf{w}).$$

Recall that $H_{\mathcal{F}_i}(\mathbf{w})$ is the Hessian matrix of i-th output of $\mathcal{F}$ with respect to $\mathbf{w}$.

We consider the Hessian matrices of $\mathcal{L}$ around a global minimizer $\mathbf{w}^*$ of the loss function, i.e., solution of the system of equations. Specifically, consider the following two points $\mathbf{w}^* + \boldsymbol{\delta}$ and $\mathbf{w}^* - \boldsymbol{\delta}$, which are

in a sufficiently small neighborhood of the minimizer $\mathbf{w}^*$. Then the respective Hessian at these two points obey

$$H_{\mathcal{L}}(\mathbf{w}^*+\boldsymbol{\delta}) = \underbrace{D\mathcal{F}(\mathbf{w}^*+\boldsymbol{\delta})^T \frac{\partial^2 \mathcal{L}}{\partial \mathcal{F}^2}(\mathbf{w}^*+\boldsymbol{\delta}) D\mathcal{F}(\mathbf{w}^*+\boldsymbol{\delta})}_{A(\mathbf{w}^*+\boldsymbol{\delta})} + \sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i H_{\mathcal{F}_i}(\mathbf{w}^*+\boldsymbol{\delta}) + o(\|\boldsymbol{\delta}\|),$$

$$H_{\mathcal{L}}(\mathbf{w}^*-\boldsymbol{\delta}) = \underbrace{D\mathcal{F}(\mathbf{w}^*-\boldsymbol{\delta})^T \frac{\partial^2 \mathcal{L}}{\partial \mathcal{F}^2}(\mathbf{w}^*-\boldsymbol{\delta}) D\mathcal{F}(\mathbf{w}^*-\boldsymbol{\delta})}_{A(\mathbf{w}^*-\boldsymbol{\delta})} - \sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i H_{\mathcal{F}_i}(\mathbf{w}^*+\boldsymbol{\delta}) + o(\|\boldsymbol{\delta}\|).$$

Note that both the terms $A(\mathbf{w}^*+\boldsymbol{\delta})$ and $A(\mathbf{w}^*-\boldsymbol{\delta})$ are matrices with rank at most n, since $D\mathcal{F}$ is of the size $n \times m$.

By the assumption, at least one component $H_{\mathcal{F}_k}$ of the Hessian of $\mathcal{F}$ satisfies that the rank of $H_{\mathcal{F}_k}(\mathbf{w}^*)$ is greater than $2n$. By the continuity of the Hessian, we have that, if magnitude of $\boldsymbol{\delta}$ is sufficiently small, then the ranks of $H_{\mathcal{F}_k}(\mathbf{w}^*+\boldsymbol{\delta})$ and $H_{\mathcal{F}_k}(\mathbf{w}^*-\boldsymbol{\delta})$ are also greater than $2n$.

Hence, we can always find a unit length vector $\mathbf{v} \in \mathbb{R}^m$ s.t.

$$\mathbf{v}^T A(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} = \mathbf{v}^T A(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v} = 0, \tag{40}$$

but

$$\mathbf{v}^T H_{\mathcal{F}_k}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} \neq 0, \quad \mathbf{v}^T H_{\mathcal{F}_k}(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v} \neq 0. \tag{41}$$

Consequently the vector $\langle \mathbf{v}^T H_{\mathcal{F}_1}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v}, \ldots, \mathbf{v}^T H_{\mathcal{F}_n}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} \rangle \neq \mathbf{0}$ and $\langle \mathbf{v}^T H_{\mathcal{F}_1}(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v}, \ldots, \mathbf{v}^T H_{\mathcal{F}_n}(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v} \rangle \neq \mathbf{0}$.

With the same $\mathbf{v}$, we have

$$\mathbf{v}^T H_{\mathcal{L}}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} = \sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} + o(\|\boldsymbol{\delta}\|), \tag{42}$$

$$\mathbf{v}^T H_{\mathcal{L}}(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v} = -\sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v} + o(\|\boldsymbol{\delta}\|). \tag{43}$$

In the following, we show that, for sufficiently small $\boldsymbol{\delta}$, $\mathbf{v}^T H_{\mathcal{L}}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v}$ and $\mathbf{v}^T H_{\mathcal{L}}(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v}$ can not be non-negative simultaneously, which immediately implies that $H_{\mathcal{L}}$ is not positive semi-definite everywhere in the close neighborhood of $\mathbf{w}^*$, hence $\mathcal{L}$ is not locally convex at $\mathbf{w}^*$.

Specifically, with the condition $\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \neq \mathbf{0}$, for Eq. (42) and Eq. (43) we have the following cases:

Case 1: If $\sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} < 0$, then directly $\mathbf{v}^T H_{\mathcal{L}}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} < 0$ if $\boldsymbol{\delta}$ is small enough which completes the proof.

Case 2: Otherwise if $\sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} > 0$, by the continuity of each $H_{\mathcal{F}_i}(\cdot)$, we have

$$\begin{aligned}
&-\sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^*-\boldsymbol{\delta})\mathbf{v} \\
&= -\sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^*+\boldsymbol{\delta})\mathbf{v} \\
&\quad + \sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T (H_{\mathcal{F}_i}(\mathbf{w}^*+\boldsymbol{\delta}) - H_{\mathcal{F}_i}(\mathbf{w}^*-\boldsymbol{\delta}))\mathbf{v}
\end{aligned}$$

$$= -\sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{\partial \mathcal{L}}{\partial \mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^* + \boldsymbol{\delta})\mathbf{v} + O(\epsilon)$$

$$< 0,$$

when $\boldsymbol{\delta}$ is small enough.

Note. If $\sum_{i=1}^{n} \left(\frac{d}{d\mathbf{w}} \left(\frac{d\mathcal{L}}{d\mathcal{F}}(\mathbf{w}^*) \right) \boldsymbol{\delta} \right)_i \mathbf{v}^T H_{\mathcal{F}_i}(\mathbf{w}^* + \boldsymbol{\delta})\mathbf{v} = 0$, we can always adjust $\boldsymbol{\delta}$ a little so that it turns to case 1 or 2.

In conclusion, with certain $\boldsymbol{\delta}$ which is arbitrarily small, either $\mathbf{v}^T H_{\mathcal{L}}(\mathbf{w}^* + \boldsymbol{\delta})\mathbf{v}$ or $\mathbf{v}^T H_{\mathcal{L}}(\mathbf{w}^* - \boldsymbol{\delta})\mathbf{v}$ has to be negative which means $\mathcal{L}(\mathbf{w})$ is not convex, at some point in the neighborhood of $\mathbf{w}^*$. □

Appendix C. Small Hessian is a feature of certain large models

Here we show that the conditions we have placed on the Hessian spectral norm are not restrictive. In fact, they are mathematical freebies as long as the model has certain structure and is large enough. For example, if a neural network has a linear output layer and is wide enough, its Hessian spectral norm can be arbitrarily small, see [24].

In the following, let's consider an illustrative example. Let the model f be a linear combination of m independent sub-models,

$$f(\mathbf{w}; \mathbf{x}) = \frac{1}{s(m)} \sum_{i=1}^{m} v_i \alpha_i(\mathbf{w}_i; \mathbf{x}), \tag{44}$$

where $\mathbf{w}_i$ is the trainable parameters of the sub-model α_i, $i \in [m]$, and $\frac{1}{s(m)}$ is a scaling factor that depends on m. The sub-model weights v_i are independently randomly chosen from $\{-1, 1\}$ and not trainable. The parameters of the model f are the concatenation of sub-model parameters: $\mathbf{w} := (\mathbf{w}_1, \cdots, \mathbf{w}_m)$ with $\mathbf{w}_i \in \mathbb{R}^p$. We make the following mild assumptions.

Assumption 1. For simplicity, we assume that the sub-models $\alpha_i(\mathbf{w}_i, \mathbf{x})$ have the same structure but different initial parameters $\mathbf{w}_{i,0}$ due to random initialization. We further assume each sub-model has a $\Theta(1)$ output, and is second-order differentiable and β-smooth.

Due to the randomness of the sub-model weights v_i, the scaling factor $\frac{1}{s(m)} = o(1)$ w.r.t. the size m of the model f (e.g, $\frac{1}{s(m)} = \frac{1}{\sqrt{m}}$ for neural networks [17,24]).

The following theorem states that, as long as the model size m is sufficiently large, the Hessian spectral norm of the model f is arbitrarily small.

Theorem 11. *Consider the model f defined in Eq. (44). Under Assumption 1, the spectral norm of the Hessian of model f satisfies:*

$$\|H_f\| \leq \frac{\beta}{s(m)}. \tag{45}$$

Proof. An entry $(H_f)_{jk}$, $j, k \in \mathbb{R}^{m \times p}$, of the model Hessian matrix is

$$(H_f)_{jk} = \frac{1}{s(m)} \sum_{i=1}^{m} v_i \frac{\partial^2 \alpha_i}{\partial w_j \partial w_k} =: \frac{1}{s(m)} \sum_{i=1}^{m} v_i (H_{\alpha_i})_{jk}, \tag{46}$$

with H_{α_i} being the Hessian matrix of the sub-model α_i. Because the parameters of f are the concatenation of sub-model parameters and the sub-models share no common parameters, in the summation of Eq. (46),

there must be at most one non-zero term (non-zero only when w_j and w_k belong to the same model α_i. Thus, the Hessian spectral norm can be bounded by

$$\|H_f\| \leq \frac{1}{s(m)} \max_{i\in[m]} |v_i| \cdot \|H_{\alpha_i}\| \leq \frac{1}{s(m)} \cdot \beta. \quad \square \tag{47}$$

Appendix D. An illustrative example of composition models

Consider the model $h = g \circ f$ as a composition of 2 random Fourier feature models $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ with:

$$f(\mathbf{u}; x) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \left(u_k \cos(\omega_k x) + u_{k+m} \sin(\omega_k x)\right), \tag{48}$$

$$g(\mathbf{a}; z) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \left(a_k \cos(\nu_k z) + a_{k+m} \sin(\nu_k z)\right). \tag{49}$$

Here we set the frequencies $\omega_k \sim \mathcal{N}(0, 1)$ and $\nu_k \sim \mathcal{N}(0, n^2)$, for all $k \in [m]$, to be fixed. The trainable parameters of the model are $(\mathbf{u}, \mathbf{a}) \in \mathbb{R}^{2m} \times \mathbb{R}^{2m}$. It's not hard to see that the model $h(x) = g \circ f(x)$ can be viewed as a 4-layer "bottleneck" neural network where the second hidden layer has only one neuron, i.e. the output of f.

For simplicity, we let the input x be 1-dimensional, and denote the training data as $x_1, ..., x_n \in \mathbb{R}$. We consider the case of both sub-models, f and g, are large, especially $m \to \infty$.

By Proposition 5, the tangent kernel matrix of h, i.e. K_h, can be decomposed into the sum of two positive semi-definite matrices, and the uniform conditioning of K_h can be guaranteed if one of them is uniformly conditioned, as demonstrated in Eq. (18). In the following, we show K_g is uniformly conditioned by making f well separated.

We assume the training data are not degenerate and the parameters $\mathbf{u}$ are randomly initialized. This makes sure the initial outputs of f, which are the initial inputs of g, are not degenerate, with overwhelming probability. For example, let $\min_{i\neq j} |x_i - x_j| \geq \sqrt{2}$ and initialize $\mathbf{u}$ by $\mathcal{N}(0, 100R^2)$ with a given number $R > 0$. Then, we have

$$f(\mathbf{u}_0; x_i) - f(\mathbf{u}_0; x_j) \sim \mathcal{N}\left(0, 200R^2 - 200R^2 e^{\frac{-|x_i - x_j|^2}{2}}\right), \ \forall i, j \in [n].$$

Since $\min_{i\neq j} |x_i - x_j| \geq \sqrt{2}$, the variance $Var := 200R^2 - 200R^2 e^{\frac{-|x_i - x_j|^2}{2}} > 100R^2$. For this Gaussian distribution, we have, with probability at least 0.96, that

$$\min_{i\neq j} |f(\mathbf{u}_0; x_i) - f(\mathbf{u}_0; x_j)| > 2R. \tag{50}$$

For this model, the partial tangent kernel K_g is the Gaussian kernel in the limit of $m \to \infty$, with the following entries:

$$K_{g,ij}(\mathbf{u}) = \exp\left(-\frac{n^2 |f(\mathbf{u}; x_i) - f(\mathbf{u}; x_j)|^2}{2}\right).$$

By the Gershgorin circle theorem, its smallest eigenvalue is lower bounded by:

$$\inf_{\mathbf{u}\in\mathcal{S}} \lambda_{\min}(K_g(\mathbf{u})) \geq 1 - (n-1) \exp\left(-\frac{n^2 \rho(\mathcal{S})^2}{2}\right),$$

where $\rho(\mathcal{S}) := \inf_{\mathbf{u}\in\mathcal{S}} \min_{i\neq j} |f(\mathbf{u}; x_i) - f(\mathbf{u}; x_j)|$ is the minimum separation between the outputs of f, i.e., inputs of g in $\mathcal{S}$. If $\rho(\mathcal{S}) \geq \sqrt{2\ln(2n-2)}/n$, then we have $\inf_{\mathbf{u}\in\mathcal{S}} \lambda_{\min}(K_g(\mathbf{u})) \geq 1/2$. Therefore, we see that the uniform conditioning of K_g, hence that of K_h, is controlled by the separation between the inputs of g, i.e., outputs of f.

Within the ball $B((\mathbf{u}_0, \mathbf{a}_0), R) := \{(\mathbf{u}, \mathbf{a}) : \|\mathbf{u} - \mathbf{u}_0\|^2 + \|\mathbf{a} - \mathbf{a}_0\|^2 \leq R^2\}$ with an arbitrary radius $R > 0$, the outputs of f are always well separated, given that the initial outputs of f are separated by $2R$ as already discussed above. This is because

$$|f(\mathbf{u}; x) - f(\mathbf{u}_0; x)| = \frac{1}{\sqrt{m}} \left| \sum_{k=1}^{m} \left((u_k - u_{0,k}) \cos(\omega_k x) + (u_{k+m} - u_{0,k+m}) \sin(\omega_k x) \right) \right| \leq \|\mathbf{u} - \mathbf{u}_0\| \leq R,$$

which leads to $\rho(B((\mathbf{u}_0, \mathbf{a}_0), R)) > R$ by Eq. (50). By choosing R appropriately, to be specific, $R \geq \sqrt{2\ln(2n-2)}/n$, the uniform conditioning of K_g is satisfied.

Hence, we see that composing large non-linear models may make the tangent kernel no longer constant, but the uniform conditioning of the tangent kernel can remain.

Appendix E. Wide CNN and ResNet satisfy PL* condition

In this section, we will show that wide Convolutional Neural Networks (CNN) and Residual Neural Networks (ResNet) also satisfy the PL* condition.

The CNN is defined as follows:

$$\begin{aligned} &\alpha^{(0)} = \mathbf{x}, \\ &\alpha^{(l)} = \sigma_l \left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} * \alpha^{(l-1)} \right), \ \forall l = 1, 2, \ldots, L, \\ &f(\mathbf{W}; \mathbf{x}) = \frac{1}{\sqrt{m_L}} \langle W^{(L+1)}, \alpha^{(L+1)} \rangle, \end{aligned} \tag{51}$$

where $*$ is the convolution operator (see the definition below) and $\langle \cdot, \cdot \rangle$ is the standard matrix inner product.

Compared to the definition of fully-connected neural networks in Eq. (19), the l-th hidden neurons $\alpha^{(l)} \in \mathbb{R}^{m_l \times Q}$ is a matrix where m_l is the number of channels and Q is the number of pixels, and $W^{(l)} \in \mathbb{R}^{K \times m_l \times m_{l-1}}$ is an order 3 tensor where K is the filter size except that $W^{(L+1)} \in \mathbb{R}^{m_L \times Q}$.

For the simplicity of the notation, we give the definition of convolution operation for 1-D CNN in the following. We note that it's not hard to extend to higher dimensional CNNs and one will find that our analysis still applies.

$$\left(W^{(l)} * \alpha^{(l-1)} \right)_{i,q} = \sum_{k=1}^{K} \sum_{j=1}^{m_{l-1}} W^{(l)}_{k,i,j} \alpha^{(l-1)}_{j,q+k-\frac{K+1}{2}}, \ i \in [m_l], q \in [Q]. \tag{52}$$

The ResNet is similarly defined as follows:

$$\begin{aligned} &\alpha^{(0)} = \mathbf{x}, \\ &\alpha^{(l)} = \sigma_l \left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)} \right) + \alpha^{(l-1)}, \ \ \forall l = 1, 2, \cdots, L+1, \\ &f(\mathbf{W}; \mathbf{x}) = \alpha^{(L+1)}. \end{aligned} \tag{53}$$

We see that the ResNet is the same as a fully connected neural network, Eq. (19), except that the activations $\alpha^{(l)}$ has an extra additive term $\alpha^{(l-1)}$ from the previous layer, interpreted as skip connection.

Remark 11. This definition of ResNet differs from the standard ResNet architecture in [16] that the skip connections are at every layer, instead of every two layers. One will find that the same analysis can be easily generalized to cases where skip connections are at every two or more layers. The same definition, up to a scaling factor, was also studied theoretically in [13].

By the following theorem, we have an upper bound for the Hessian spectrum norm of the CNN and ResNet, similar to Theorem 5 for fully-connected neural networks.

Theorem 12 *(Theorem 3.3 of [24]). Consider a neural network $f(\mathbf{W};\mathbf{x})$ of the form Eq. (51) or Eq. (53). Let m be the minimum of the hidden layer widths, i.e., $m = \min_{l\in[L]} m_l$. Given any fixed $R > 0$, and any $\mathbf{W} \in B(\mathbf{W}_0, R) := \{\mathbf{W} : \|\mathbf{W} - \mathbf{W}_0\| \le R\}$, with high probability over the initialization, the Hessian spectral norm satisfies the following:*

$$\|H_f(\mathbf{W})\| = \tilde{O}\left(R^{3L}/\sqrt{m}\right). \tag{54}$$

Using the same analysis in Section 4.2, we can get a similar result with Theorem 4 for CNN and ResNet to show they also satisfy PL* condition:

Theorem 13 *(Wide CNN and ResNet satisfy PL* condition). Consider the neural network $f(\mathbf{W};\mathbf{x})$ in Eq. (51) or Eq. (53), and a random parameter setting $\mathbf{W}_0$ such that each element of $W_0^{(l)}$ for $l \in [L+1]$ follows $\mathcal{N}(0,1)$. Suppose that the last layer activation σ_{L+1} satisfies $|\sigma'_{L+1}(z)| \ge \rho > 0$ and that $\lambda_0 := \lambda_{\min}(K(\mathbf{W}_0)) > 0$. For any $\mu \in (0, \lambda_0\rho^2)$, if the width of the network*

$$m = \tilde{\Omega}\left(\frac{nR^{6L+2}}{(\lambda_0 - \mu\rho^{-2})^2}\right), \tag{55}$$

then the μ-PL* condition holds for square loss in the ball $B(\mathbf{W}_0, R)$.

Appendix F. Proof for convergence under PL* condition

In Theorem 6, the convergence of gradient descent is established in the square loss case. In fact, similar results (with a bit modification) hold for general loss functions. In the following theorem, we provide the convergence under PL* condition for general loss functions. Then, we prove Theorem 6 and 14 together.

Theorem 14. *Suppose the loss function $\mathcal{L}(\mathbf{w})$ (not necessarily the square loss) is β-smooth and satisfies the μ-PL* condition in the ball $B(\mathbf{w}_0, R) := \{\mathbf{w} \in \mathbb{R}^m : \|\mathbf{w} - \mathbf{w}_0\| \le R\}$ with $R = \frac{2\sqrt{2\beta\mathcal{L}(\mathbf{w}_0)}}{\mu}$. Then we have the following:*

(a) Existence of a solution: There exists a solution (global minimizer of $\mathcal{L}$) $\mathbf{w}^* \in B(\mathbf{w}_0, R)$, such that $\mathcal{F}(\mathbf{w}^*) = \mathbf{y}$.

(b) Convergence of GD: Gradient descent with a step size $\eta \le 1/\sup_{\mathbf{w}\in B(\mathbf{w}_0,R)} \|H_{\mathcal{L}}(\mathbf{w})\|_2$ converges to a global solution in $B(\mathbf{w}_0, R)$, with an exponential (also known as linear) convergence rate:

$$\mathcal{L}(\mathbf{w}_t) \le \left(1 - \kappa^{-1}_{\mathcal{L},\mathcal{F}}(B(\mathbf{w}_0, R))\right)^t \mathcal{L}(\mathbf{w}_0), \tag{56}$$

where the condition number $\kappa_{\mathcal{L},\mathcal{F}}(B(\mathbf{w}_0, R)) = \frac{1}{\eta\mu}$.

Proof. Let's start with Theorem 14. We prove this theorem by induction. The induction hypothesis is that, for all $t \ge 0$, $\mathbf{w}_t$ is within the ball $B(\mathbf{w}_0, R)$ with $R = \frac{2\sqrt{2\beta\mathcal{L}(\mathbf{w}_0)}}{\mu}$, and

$$\mathcal{L}(\mathbf{w}_t) \leq (1-\eta\mu)^t \mathcal{L}(\mathbf{w}_0). \tag{57}$$

In the base case, where $t = 0$, it is trivial that $\mathbf{w}_0 \in B(\mathbf{w}_0, R)$ and that $\mathcal{L}(\mathbf{w}_t) \leq (1-\eta\mu)^0 \mathcal{L}(\mathbf{w}_0)$.

Suppose that, for a given $t \geq 0$, $\mathbf{w}_t$ is in the ball $B(\mathbf{w}_0, R)$ and Eq. (57) holds. As we show separately below (in Eq. (61) we have $\mathbf{w}_{t+1} \in B(\mathbf{w}_0, R)$. Hence we see that $\mathcal{L}$ is $(\sup_{\mathbf{w}\in B(\mathbf{w}_0,R)} \|H_{\mathcal{L}}(\mathbf{w})\|_2)$-smooth in $B(\mathbf{w}_0, R)$. By definition of $\eta = 1/\sup_{\mathbf{w}\in B(\mathbf{w}_0,R)} \|H_L(\mathbf{w})\|_2$ and, consequently, $\mathcal{L}$ is $1/\eta$-smooth. Using that we obtain the following:

$$\begin{aligned}\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) - \nabla\mathcal{L}(\mathbf{w}_t)(\mathbf{w}_{t+1} - \mathbf{w}_t) &\leq \frac{1}{2} \sup_{\mathbf{w}\in B(\mathbf{w}_0,R)} \|H_{\mathcal{L}}\|_2 \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\ &= \frac{1}{2\eta}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2. \end{aligned} \tag{58}$$

Taking $\mathbf{w}_{t+1} - \mathbf{w}_t = -\eta\nabla\mathcal{L}(\mathbf{w}_t)$ and by μ-PL* condition at point $\mathbf{w}_t$, we have

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) \leq -\frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 \leq -\eta\mu\mathcal{L}(\mathbf{w}_t). \tag{59}$$

Therefore,

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq (1-\eta\mu)\mathcal{L}(\mathbf{w}_t) \leq (1-\eta\mu)^{t+1} \mathcal{L}(\mathbf{w}_0). \tag{60}$$

To prove $\mathbf{w}_{t+1} \in B(\mathbf{w}_0, R)$, by the fact that $\mathcal{L}$ is β-smooth, we have

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_0\| &= \eta\|\sum_{i=0}^{t} \nabla\mathcal{L}(\mathbf{w}_i)\| \\ &\leq \eta \sum_{i=0}^{t} \|\nabla\mathcal{L}(\mathbf{w}_i)\| \\ &\leq \eta \sum_{i=0}^{t} \sqrt{2\beta(\mathcal{L}(\mathbf{w}_i) - \mathcal{L}(\mathbf{w}_{i+1}))} \\ &\leq \eta \sum_{i=0}^{t} \sqrt{2\beta\mathcal{L}(\mathbf{w}_i)} \\ &\leq \eta\sqrt{2\beta}\left(\sum_{i=0}^{t} (1-\eta\mu)^{i/2}\right)\sqrt{\mathcal{L}(\mathbf{w}_0)} \\ &\leq \eta\sqrt{2\beta}\sqrt{\mathcal{L}(\mathbf{w}_0)}\frac{1}{1-\sqrt{1-\eta\mu}} \\ &\leq \frac{2\sqrt{2\beta}\sqrt{\mathcal{L}(\mathbf{w}_0)}}{\mu} \\ &= R. \end{aligned} \tag{61}$$

Thus, $\mathbf{w}_{t+1}$ resides in the ball $B(\mathbf{w}_0, R)$.

By the principle of induction, the hypothesis is true.

Now, we prove Theorem 6, i.e., the particular case of square loss $\mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathcal{F}(\mathbf{w}) - \mathbf{y}\|^2$. In this case,

$$\nabla\mathcal{L}(\mathbf{w}) = (\mathcal{F}(\mathbf{w}) - \mathbf{y})^T D\mathcal{F}(\mathbf{w}). \tag{62}$$

Hence, in Eq. (61), we have the following instead:

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}_0\| &\le \eta\sum_{i=0}^{t}\|\nabla\mathcal{L}(\mathbf{w}_i)\| \\ &\le \eta\sum_{i=0}^{t}\|D\mathcal{F}(\mathbf{w}_i)\|_2\|\mathcal{F}(\mathbf{w}_i)-\mathbf{y})\| \\ &\le \eta L_{\mathcal{F}}\left(\sum_{i=0}^{t}(1-\mu/\beta_{\mathcal{F}})^{i/2}\right)\|\mathcal{F}(\mathbf{w}_0)-\mathbf{y})\| \\ &\le \eta L_{\mathcal{F}}\cdot\frac{2}{\eta\mu}\|\mathcal{F}(\mathbf{w}_0)-\mathbf{y})\| \\ &= \frac{2L_{\mathcal{F}}\|\mathcal{F}(\mathbf{w}_0)-\mathbf{y})\|}{\mu}. \end{aligned} \tag{63}$$

Also note that, for all $t>0$, $\|H_{\mathcal{L}}(\mathbf{w}_t)\|_2 \le L_{\mathcal{F}}^2+\beta_{\mathcal{F}}\cdot\|\mathcal{F}(\mathbf{w}_0)-\mathbf{y}\|$, since $\|\mathcal{F}(\mathbf{w}_t)-\mathbf{y}\| \le \|\mathcal{F}(\mathbf{w}_0)-\mathbf{y}\|$. Hence, the step size $\eta = 1/L_F^2+\beta_{\mathcal{F}}\cdot\|\mathcal{F}(\mathbf{w}_0)-\mathbf{y}\|$ is valid. □

Appendix G. Proof of Theorem 7

Proof. We first aggressively assume that the μ-PL* condition holds in the whole parameter space $\mathbb{R}^m$. We will find that the condition can be relaxed to hold only in the ball $B(\mathbf{w}_0,R)$.

Following a similar analysis to the proof of Theorem 1 in [4], by the μ-PL* condition, we can get that for any mini-bath size s, the mini-batch SGD with step size $\eta^*(s) := \frac{n\mu}{n\lambda(n^2\lambda+\mu(s-1))}$ has an exponential convergence rate for all $t>0$:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t)] \le \left(1-\frac{\mu s\eta^*(s)}{n}\right)^t\mathcal{L}(\mathbf{w}_0). \tag{64}$$

Moreover, the expected length of each step is bounded by

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|\right] &= \eta^*\mathbb{E}\left[\|\sum_{j=1}^{s}\nabla\ell_{i_t^{(j)}}(\mathbf{w}_t)\|\right] \\ &\le \eta^*\sum_{j=1}^{s}\mathbb{E}\left[\|\nabla\ell_{i_t^{(j)}}(\mathbf{w}_t)\|\right] \\ &\le \eta^*\sum_{j=1}^{s}\mathbb{E}\left[\sqrt{2\lambda\ell_{i_t^{(j)}}}\right] \\ &\le \eta^* s\mathbb{E}\left[\sqrt{2\lambda\mathcal{L}(\mathbf{w}_t)}\right] \\ &\le \eta^* s\sqrt{2\lambda\mathbb{E}[\mathcal{L}(\mathbf{w}_t)]} \\ &\le \eta^* s\sqrt{2\lambda}\left(1-\frac{\mu s\eta^*}{n}\right)^{t/2}\sqrt{\mathcal{L}(\mathbf{w}_0)}, \end{aligned}$$

where we use $\{i_t^{(1)},i_t^{(2)},...,i_t^{(s)}\}$ to denote a random mini-batch of the dataset at step t.

Then the expectation of the length of the whole optimization path is bounded by

$$\mathbb{E}\left[\sum_{i=0}^{\infty}\|\mathbf{w}_{i+1}-\mathbf{w}_i\|\right]=\sum_{i=0}^{\infty}\mathbb{E}\|\mathbf{w}_{i+1}-\mathbf{w}_i\|$$
$$=\sum_{i=0}^{\infty}\eta^* s\sqrt{2\lambda}\left(1-\frac{\mu s\eta^*}{n}\right)^{t/2}\sqrt{\mathcal{L}(\mathbf{w}_0)}=\frac{2n\sqrt{2\lambda}\sqrt{\mathcal{L}(\mathbf{w}_0)}}{\mu}.$$

By Markov's inequality, we have that, with probability at least $1-\delta$, the length of the path is shorter than R, i.e.,

$$\sum_{i=0}^{\infty}\|\mathbf{w}_{i+1}-\mathbf{w}_i\|\leq\frac{2n\sqrt{2\lambda}\sqrt{\mathcal{L}(\mathbf{w}_0)}}{\mu\delta}=R.$$

This means that, with probability at least $1-\delta$, the whole path is covered by the ball $B(\mathbf{w}_0,R)$, namely, for all t,

$$\|\mathbf{w}_t-\mathbf{w}_0\|\leq\sum_{i=0}^{t-1}\|\mathbf{w}_{i+1}-\mathbf{w}_i\|\leq R.$$

For those events when the whole path is covered by the ball, we can relax the satisfaction of the PL* condition from the whole space to the ball. □

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, Zhao Song, A convergence theory for deep learning via over-parameterization, in: International Conference on Machine Learning, 2019, pp. 242–252.
[2] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, Ruosong Wang, Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, in: International Conference on Machine Learning, 2019, pp. 322–332.
[3] Peter L. Bartlett, David P. Helmbold, Philip M. Long, Gradient descent with identity initialization efficiently learns positive-definite linear transformations by deep residual networks, Neural Comput. 31 (3) (2019) 477–502.
[4] Raef Bassily, Mikhail Belkin, Siyuan Ma, On exponential convergence of SGD in non-convex over-parametrized learning, preprint, arXiv:1811.02564, 2018.
[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal, Reconciling modern machine- learning practice and the classical bias-variance trade-off, Proc. Natl. Acad. Sci. 116 (32) (2019) 15849–15854.
[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, 2020, pp. 1877–1901.
[7] Peter Burgisser, Felipe Cucker, Condition: The Geometry of Numerical Algorithms, vol. 349, Springer Science & Business Media, 2013.
[8] Zachary Charles, Dimitris Papailiopoulos, Stability and generalization of learning algorithms that converge to global optima, in: International Conference on Machine Learning, PMLR, 2018, pp. 745–754.
[9] Zizhong Chen, Jack J. Dongarra, Condition numbers of Gaussian random matrices, SIAM J. Matrix Anal. Appl. 27 (3) (2005) 603–620.
[10] Lenaic Chizat, Edouard Oyallon, Francis Bach, On lazy training in differentiable programming, in: Advances in Neural Information Processing Systems, 2019, pp. 2933–2943.
[11] Yaim Cooper, Global minima of overparameterized neural networks, SIAM J. Math. Data Sci. 3 (2) (2021) 676–691.
[12] Simon S. Du, Xiyu Zhai, Barnabas Poczos, Aarti Singh, Gradient descent provably optimizes over-parameterized neural networks, in: International Conference on Learning Representations, 2018.
[13] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, Xiyu Zhai, Gradient descent finds global minima of deep neural networks, in: International Conference on Machine Learning, 2019, pp. 1675–1685.
[14] William Fedus, Barret Zoph, Noam Shazeer, Switch transformers: scaling to trillion parameter models with simple and efficient sparsity, preprint, arXiv:2101.03961, 2021.
[15] Chirag Gupta, Sivaraman Balakrishnan, Aaditya Ramdas, Path length bounds for gradient descent and flow, J. Mach. Learn. Res. 22 (68) (2021) 1–63.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[17] Arthur Jacot, Franck Gabriel, Clement Hongler, Neural tangent kernel: convergence and generalization in neural networks, in: Advances in Neural Information Processing Systems, 2018, pp. 8571–8580.
[18] Ziwei Ji, Matus Telgarsky, Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks, in: International Conference on Learning Representations, 2019.
[19] Diederik P. Kingma, Jimmy Ba, Adam: a method for stochastic optimization, in: ICLR, Poster, 2015.
[20] Johannes Lederer, No spurious local minima: on the optimization landscapes of wide and deep neural networks, preprint, arXiv:2010.00885, 2020.
[21] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, Jeffrey Pennington, Wide neural networks of any depth evolve as linear models under gradient descent, in: Advances in Neural Information Processing Systems, 2019, pp. 8570–8581.
[22] Dawei Li, Tian Ding, Ruoyu Sun, Over-parameterized deep neural networks have no strict local minima for any continuous activations, preprint, arXiv:1812.11039, 2018.
[23] Chaoyue Liu, Mikhail Belkin, Accelerating SGD with momentum for over-parameterized learning, in: The 8th International Conference on Learning Representations, ICLR, 2020.
[24] Chaoyue Liu, Libin Zhu, Mikhail Belkin, On the linearity of large non-linear models: when and why the tangent kernel is constant, in: Advances in Neural Information Processing Systems, vol. 33, 2020.
[25] Stanislaw Lojasiewicz, A topological property of real analytic subsets, in: Coll. du CNRS, Les Equations aux dérivées partielles, vol. 117, 1963, pp. 87–89.
[26] Song Mei, Andrea Montanari, Phan-Minh Nguyen, A mean field view of the landscape of two-layer neural networks, Proc. Natl. Acad. Sci. 115 (33) (2018) E7665–E7671.
[27] Yurii Nesterov, A method for unconstrained convex minimization problem with the rate of convergence O $(1/k^2)$, Dokl. Acad. Nauk USSR 269 (1983) 543–547.
[28] Quynh Nguyen, Mahesh Chandra Mukkamala, Matthias Hein, On the loss landscape of a class of deep neural networks with no bad local valleys, in: International Conference on Learning Representations, 2018.
[29] Jorge Nocedal, Stephen Wright, Numerical Optimization, Springer Science & Business Media, 2006.
[30] Samet Oymak, Mahdi Soltanolkotabi, Toward moderate overparameterization: global convergence guarantees for training shallow neural networks, IEEE J. Sel. Areas Inf. Theory 1 (1) (2020) 84–105.
[31] Tomaso Poggio, Gil Kur, Andrzej Banburski, Double descent in the condition number, preprint, arXiv:1912.06190, 2019.
[32] Boris Teodorovich Polyak, Gradient methods for minimizing functionals, Ž. Vyčisl. Mat. Mat. Fiz. 3 (4) (1963) 643–653.
[33] Mahdi Soltanolkotabi, Adel Javanmard, Jason D. Lee, Theoretical insights into the optimization landscape of over-parameterized shallow neural networks, IEEE Trans. Inf. Theory 65 (2) (2018) 742–769.
[34] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, M. Wyart, A jamming transition from under- to over-parametrization affects generalization in deep learning, J. Phys. A, Math. Theor. 52 (47) (Oct. 2019) 474001.
[35] Sharan Vaswani, Francis Bach, Mark Schmidt, Fast and faster convergence of SGD for overparameterized models and an accelerated perceptron, in: The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1195–1204.
[36] Patrick M. Wensing, Jean-Jacques Slotine, Beyond convexity—contraction and global convergence of gradient descent, PLoS ONE 15 (8) (2020) e0236661.
[37] Xiao-Hu Yu, Guo-An Chen, On the local minima free condition of backpropagation learning, IEEE Trans. Neural Netw. 6 (5) (1995) 1300–1303.
[38] Difan Zou, Yuan Cao, Dongruo Zhou, Quanquan Gu, Gradient descent optimizes overparameterized deep ReLU networks, Mach. Learn. 109 (3) (2020) 467–492.